International Journal of Economics and Finance; Vol. 10, No. 8; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Forecasting a Mix of Temporal and Non-Temporal Economic Variables with a Mixture-of-Experts Neural Network

Dat-Dao Nguyen[1] & Dennis Kira [2]

[1] Department of Accounting & Information Systems, California State University – Northridge, USA.

[2] Department of Supply Chain & Business Technology Management, Concordia University, Montreal (Quebec), Canada

Correspondence: Dat-Dao Nguyen, Department of Accounting & Information Systems, California State University – Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8372, USA.

Received: June 11, 2018 Accepted: July 3, 2018 Online Published: July 10, 2018

doi:10.5539/ijef.v10n8p141 URL: https://doi.org/10.5539/ijef.v10n8p141

Abstract

This study investigates a versatile forecasting technique using an integrated system of Artificial Neural Networks (ANN) and Genetic Algorithms (GA) in a mixture-of-experts architecture to solve a general economic forecasting problem involving a mix of temporal and non-temporal variables. Using Klein Model I as a context and previous estimations from traditional methods as benchmarks, the study provides evidence on the effectiveness and efficiency of this integrated system. ANN helps overcome the imposition of assumptions on the behaviors of related variables, the specification of exact relationships, and the difficulty in nonlinear estimations of the economic model. GA helps overcome the sub-optimality of the tedious trial-and-error process in network building. The flexibility of the mixture-of experts network architecture offers many alternative configurations to capture the peculiarities of variables in context before aggregating intermediate estimations into the final result. The integrated system has shown its ability in processing effectively the mixture of economic variables, and producing efficient estimations and forecasts.

Keywords: economic models, forecasting, Genetic Algorithms, multivariate time series, mixture-of-experts neural networks

1. Introduction

General business forecasting problems, particularly those dealing with socio-economic variables, usually involve many temporal and non-temporal interactions. Very often, the value of an economic variable is not only related to its predecessors in time but also to the current and past values of other variables. Hence macroeconomic models have to incorporate various interrelated variables in the economy.

Most of econometric models have difficulty in providing accurate estimates/ forecasts due to the complexity of the economic system, the impossibility of validation with controlled experiments on the economy, and the existence of non-quantifiable factors in economic activities (Moody, 1995). Also, many assumptions have been imposed on the behaviors of related variables in the modeling process (Cromwell et al., 1994). In addition, one may encounter the complexity of estimation when dealing with nonlinear models (Mills, 1991).

This study focuses on the overcoming of these constraints in traditional modeling and forecasting methods with the implementation of an integrated system of Artificial Neural Network - ANN (Rosenblatt, 1959) and Genetic Algorithm - GA (Holland, 1975) in a modular network architecture (Jacobs et al., 1991a, 1991b). Such a mixture network system would be able to handle effectively a general family of business forecasting problems, i.e., forecasting with a mixture of temporal and non-temporal variables, in which an econometric model should be a useful context. An ANN has been proved to be a universal function approximator (Funahashi, 1989; Cybenko, 1989; Hornik et al., 1989) in learning nonlinear relationships inherent in the data without *a priori* functional form and imposed assumptions on the behavior of data. With a mixture of network architecture, one can also partition the problem space into domains and assign them to modular ANNs to learn the related peculiar patterns. GA can explore a large number of alternatives in the problem space – specifically, all possible ANN architectures - in order to avoid sub-optimality (Goldberg, 1989).

The paper is organized as follows. Section 2 reviews issues in forecasting of a mixture of temporal and

non-temporal variables presenting in an economic system, in which Klein Model I serves as an example. Section 3 discusses the use of ANN in forecasting. Section 4 proposes a mixture of ANNs for effective forecasting. Section 5 presents findings of this study and discussions. Finally, Section 6 concludes the paper with some remarks for future applications and practices.

2. Forecasting an Economic System

An economic system involving a mixture of temporal and non-temporal variables, could be a representative of a general family of business forecasting problems (Tong, 2011). An econometric model is a set of simultaneous equations to describe the working of an economy as a whole or one of its sectors (Ruud, 2000).

2.1 Structural Equations of an Economic System

Equations of an econometric model usually contain information on the following variables (Judge et al., 1985):

- *Endogenous* or jointly determined variables have outcome values determined through the joint interaction with other variables within the system.

- *Exogenous variables* affect the outcome of the endogenous variables, but whose values are determined outside the system. Exogenous variables are assumed to condition the outcome values of the endogenous variables but are not reciprocally affected because no feedback relation is assumed.

- *Lagged endogenous variables* can be placed in the same category as the exogenous variables since the observed values are predetermined for the current period. The exogenous variables and lagged endogenous variables that may involve any length of lag are called predetermined variables.

- *Non-observable random errors*, also called random shocks or disturbances.

2.2 Nonlinearity and Dynamics of Economic Variables

Economic variables change over time so that the linearity of an economic model is a strong assumption. Therefore, major concerns in forecasting is in how to capture the nonlinearity and the dynamics of economic events in economic modelling.

Nonlinearity in economic models could exist in the variables and/or in the parameters. In such cases, a traditional method is to find a method, such as Box-Cox transformation, to convert the model into a linear specification. But there are intrinsically nonlinear models, which cannot be linearly transformed. The estimation of these models is based on minimizing or maximizing an objective function such as the sum of squared errors or the likelihood function (Judge et al., 1985). However, with the current optimization methods, one may still encounter estimation complexity when dealing with nonlinear optimization problems (Mills, 1991).

Dynamics in forecasting is usually captured with dynamic regression models (Pankratz, 1991), in which an output is linearly related to current and past values of one or more inputs. An alternative approach is simultaneous equation modeling taking into account the relationship among a set of macroeconomic time series (Sims, 1980). In this multivariate perspective, a given time series may be influenced not only by certain exogenous events occurring at a particular point in time but also by contemporaneous, lagged, and leading values of a second variable or many other variables (Cromwell et al., 1994).

Judge et al. (1985) note that it is not unusual that parameters entering in a regression model simply reflect one's uncertainty on which model would adequately represent the relationship among the variables.

2.3 Klein Model I Revisited

Econometric models reported in literature range from including less than ten endogenous variables to more than one hundred endogenous variables (Bodkin et al., 1991). The classic model Klein Model I of US interwar economy from 1921 to 1941 (Klein, 1950), has been used as an example of modeling and estimation in econometrics. This model has three behavioral equations and three identities. All endogenous variables are in 1934 dollars, and all relationships are strictly linear. The model specification and variable descriptions are summarized in the following. For simplicity, time subscripts are omitted unless they indicate the lagged effects.

- Consumption Equation:

$$C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (Wp_t + Wg_t) + u_{1t} \quad (1)$$

where C is consumption, Wp is private wage bill, Wg is government wage bill, and P is non-wage income (profits).

- Investment Equation:

$$I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + u_{2t} \quad (2)$$

where I is net investment, P is profits, and K_{t-1} is stock of capital at the beginning of the year.

- Private Wages: (Demand of Labor)

$$Wp_t = \gamma_0 + \gamma_1(Y_t + T_t - Wg_t) + \gamma_2(Y_{t-1} + T_{t-1} - Wg_{t-1}) + \gamma_3 t + u_{3t} \quad (3)$$

where Y is output, T is taxes, and t is time trend (year minus 1931).

- Equilibrium Demand:

$$Y_t + T_t = C_t + I_t + G_t \quad (4)$$

- Income:

$$Y_t = Wp_t + Wg_t + P_t \quad (5)$$

- Capital Stock:

$$K_t = K_{t-1} + I_t \quad (6)$$

Therefore, the system has six endogenous variables *C, I, Wp, P, K, Y,* and four exogenous variables *T, Wg, G* and *t*.

This system could be represented in "reduced form" with respect to the endogenous variables (Theil & Boot, 1962). Given the assumption on linearity of the system, the reduced form can be specified as follows

$$y_t = Ay_{t-1} + Bx_t + Cx_{t-1} + u^*_t \quad (7)$$

where

$$y = [C\ P\ Wp\ I\ Y\ K]^T \quad (8)$$

$$x = [Wg\ T\ G\ t\ ]^T \quad (9)$$

$$u^* = [u^*C\ u^*P\ u^*Wp\ u^*I\ u^*Y\ u^*K]^T \quad (10)$$

In the reduced form, each endogenous variable in year t is described linearly in terms of the same variable lagged one year (Ay_{t-1}), the exogenous variables in the same year (Bx_t), the exogenous variables lagged one year (Cx_{t-1}), and the reduced-form disturbances u. Since *C, Wp* and *I* do not occur in lagged form, the corresponding columns in the coefficient matrix consist of zeros.

Klein Model I has been estimated by the various traditional methods. These methods address either single equations or the whole system of equations (Klein, 1950; Theil, 1971; Greene, 2011). These traditional estimations serve as benchmarks for comparison with those from proposed techniques in this study.

2.4 Single Equation Estimations of Klein Model I

2.4.1 Single-Equation Method of Least Squares

This method treats each equation independently of all others in the system. Klein noted that one had to make arbitrary choice of dependent variables for each of the three equations (Klein, 1950). It does not account for simultaneous and contemporaneous effects as one takes the values of other endogenous variables as predetermined in the calculation of the equation of interest. At best, it may serve as a sensitivity analysis given the predetermined values of other endogenous and exogenous variables of the system.

2.4.2 Two-Stage Least Squares (2SLS)

Estimation of one system equation at a time is called limited-information method. It neglects information contained in other equations. The Ordinary Least Squares method in single equation estimation cannot be applied with overidentified equations. One has to use Two-Stage Least Squares (2SLS) as an alternative.

Consider a system of simultaneous equations, the nonzero terms in the *j*th equation are

$$y_j = Y_j\gamma_j + X_j\beta_j + \varepsilon_j \quad (11)$$

In the first stage, Ordinary Least Squares prediction Y_j^* is obtained from a regression of Y_j on X. Then the 2SLS estimator is obtained by Ordinary Least Squares regression of y_j on Y_j^* and Xj (Greene, 2011).

2.4.3 Limited Information Maximum Likelihood

In Limited Information Maximum Likelihood estimation, one takes into account the absence of certain variables from a particular system equation (Theil, 1971). Using the reduced form of the system, the joint density of endogenous variables is formulated and maximized subject to the constraints that relate the structure to the reduced form (Klein, 1950).

2.5 System Estimations of Klein Model I

2.5.1 Three-Stage Least Squares (3SLS)

Three-Stage Least Squares (3SLS) method uses Generalized Least Squares estimation to the system estimation, each of which has first been estimated with 2SLS.

In the first stage, the reduced form of the system is estimated. Using Ordinary Least Squares method, this results in Y_j* for each equation. Then the fitted values of the endogenous variables are used to get 2SLS estimations of all the equations in the system. Also residuals of each equation are used to estimate the cross-equation variances and covariances Σ*. In the last stage, Generalized Least Squares parameters are obtained for the system (Greene, 2011).

2.5.2 Full Information Maximum Likelihood

This method assumes that each of the three variables *C*, *Wp*, and *I* is non-autocorrelated, and there is no correlation between the disturbances in any of the structural equations. The estimators treat all equations and all parameters jointly in formulating the likelihood function to be maximized subject to all restrictions imposed by the structure. Estimation with Full Information Maximum Likelihood was reported in Klein's monograph (Klein, 1950).

Estimated parameters for the three equations for *C*, *Wp*, and *I* obtained from different methods of limited- and full-information estimations are reported in Greene (2011). In this study, the comparison across methods is based on residuals of related estimations reported in Klein (1950), SAS/ETS (SAS, 1984).

3. Estimation and Forecasting with Artificial Neural Networks

3.1 Function Approximation with Artificial Neural Networks

An Artificial Neural Network (ANN) topology consists of nodes as autonomous processing units connected by directed arcs and arranged into layers. Every node, other than input node, computes its output S as a function of the weighted sum of inputs directed to it from other nodes,

$$S_i = \Sigma^n_j \; w_{i,\,j} \; u_j \tag{12}$$

$$u_i = f(S_i) \tag{13}$$

where $f(.)$ is a transfer function, usually a nonlinear, bounded and piecewise differentiable function, such as the sigmoid function

$$f(x) = 1/(1 + e^{-x}) \tag{14}$$

Such an ANN produces a response, which is the superposition of *n* sigmoid functions, where *n* is the number of hidden nodes, to map a complex function. As one adds more hidden layers, ANN will be able to map higher order functions (Haykin, 2009; Graupe, 2013; Schmidhuber, 2014).

The ability of ANN in function approximation is due to its capability of learning the underlying functional relationship from the data itself, therefore, minimizing the necessary *a priori* non-sample information. A multi-layer network can produce a mapping between inputs and outputs consistent with any underlying functional relationship regardless of its true functional form. It eliminates the need for unjustified *a priori* restrictions, such as the Gauss Markov assumption, frequently used to facilitate estimation in regression analysis. In traditional statistics, the appropriateness of the Ordinary Least Squares method is an empirical question, therefore the test of assumptions is a routine part of any application. In contrast, whether these assumptions hold or not, the ANN still yields a similar solution since the image of any underlying mapping can always be projected into a perfectly flexible mapping.

It has been shown that standard multi-layer networks using arbitrary transfer functions can approximate any Borel measurable function to any desired degree of accuracy (Hassoun, 1995; Steeb, 2005). The similarity between ANN techniques and traditional methods in statistics and econometrics has been investigated in the literature (Cheng & Titterington, 1994; Ripley, 1994; Hwang et al., 1994; Kuan & White, 1994; Dreyfus, 2005).

3.2 First Attempt of Forecasting Klein Model I with ANN

The first attempt to capture nonlinear relationships among economic variables of a structural system with ANN was undertaken by Caporaletti et al. (1994) with an in-sample estimation of Klein's Model I. Three ANNs are constructed and trained, each of which is used to forecast one of three endogenous variables of the model, i.e., consumption, investment, and private wage bills. Each ANN has thirteen input nodes corresponding to seven predetermined variables plus six exogenous variables of the model. The hidden layer contains eight nodes. The

output layer has a single node corresponding to a particular endogenous variable. The authors conduct ex-post forecasts and find that results are significantly better than those from traditional estimation methods.

This attempt has some shortcomings. First, with a single output node the network does not account for the contemporaneous and simultaneous effects of endogenous variables. As such, it has a similar drawback of the traditional single equation estimation method. In a simultaneous equation system, the appropriate estimation should be based on a multivariate approach instead.

Then, current values of endogenous variables in this setting are considered as inputs of the network. In addition, there is no feedback to account for the dynamics of the system. As such, this network cannot estimate and forecast a particular endogenous variable without the need of predetermined, current as well as lagged, values of all other endogenous variables.

At last, this network architecture does not handle a mixture of non-temporal and temporal variables. As such, one cannot effectively account for the contemporaneous and lagged effects of related variables of an economic system.

Our study experiments a network architecture that has the ability to account for the simultaneous and contemporaneous effects of the variables in an economic model. Using recurrent network design, the proposed network also accounts for the dynamics of the system. As such, ANNs can effectively provide not only *ex-post* estimations but also *ex-ante* forecasts of an economic system as well.

3.3 Temporal Pattern Recognition with ANN

An ANN, if it is configured appropriately, does have the ability of recognizing and storing the temporal nature of patterns. This study experiments a combination of the static representation of temporal information and storing temporal patterns in a recurrent network.

In the static representation, a sequence of incoming temporal data is represented simultaneously in the network with an input node for the value of an economic variable corresponding to each time lag. For instance, if the variable X has three lags X_{t-1}, X_{t-2}, X_{t-3}, then three input nodes needed to capture these lagged values.

In dynamic forecast, the predicted values of economic variables of concern are used in next period forecasting. Applying to ANN, one can store and generate temporal patterns via recurrent connection. In this configuration, the output just produced by the network is fed back to the input level to represent the state of the network at the preceding moment in time. Also, nodes can be created to keep some residue of the previous signals and allow slow decay of historical information.

Jordan (1986) proposes an architecture in which the value of output layer is fed back to a context unit to create the memory traces. Both input units and context units activate the hidden units to produce the next network output. A context unit retains the past value of its input with an exponential decay. It can be considered as a lowpass filter to create an output that is a weighted average of some of its recent past inputs.

$$y(n) = \sum^{n}_{i=0} x(n)\tau^{n-1} \tag{15}$$

where $0 \leq \tau \leq 1$ is a time constant to control the degree to which past values are factored in. The time constant could be set to $1 - 1/D$, where $D > 0$ represents the memory depth, i.e., how long a given value fed to the context unit is remembered.

Literature has reported the performance of recurrent ANN versus Vector Autoregression (VAR) and asserted the comparable ability of ANN in multivariate time series forecasting (Nguyen & Kira, 1997; Moshiri & Cameron, 2000; Aydin & Cavdar, 2015). The study reported herein extends previous works with forecasting a mix of temporal as well non-temporal economic data of an economic model.

3.4 Mixture-of-Experts Architecture

In complex situations, one needs a system of networks in which many specialized networks are integrated or interacted which each other in logical or real parallelism. The mixture approach is to build complex models out of simple parts.

Function approximation with ANN is traditionally based on a superposition of simple basic functions such as logistic functions. Instead of using solely superposition, one can also use the principle of divide-and-conquer to split an input space into smaller regions, which can be fitted with simpler functions by a set of function approximators called expert networks (Jacobs et al., 1991,1 991b; Jordan & Jacobs, 1995a, 1995b).

The assumption is that data can be well described by a collection of functions, each of which is defined over a relatively local region of the input space (Jordan & Jacobs, 1995a, 1995b). The expert networks could be

arranged in modular and/or hierarchical systems. They offer the ability of solving a complex problem by dividing it into a set of sub-problems, each of which may be simpler to solve than the original one. With the assumption that a particular type of network - an "expert" - is appropriate in a region of the input space, the network architecture requires a mechanism that identifies the experts or a mixture of experts that most likely produce the correct output from given associated inputs. This is accomplished with an auxiliary network, called a gating network, to provide the weight of contribution to various experts.

$$O = \Sigma^{n}_{i} \; w_i S_i \tag{16}$$

where O is the final output, S_i is modular/intermediate estimation calculated from Equations 12-13.

Various network training algorithms have been proposed to take advantage of the modularity of mixture-of-experts systems (Masoudnia & Ebrahimpour, 2014).

3.5 Genetic Algorithms in ANN Optimization

Genetic Algorithms – GA (Holland, 1975; Goldberg, 1989) have been applied to the optimization of ANN. They are implemented to search for either a set of optimal network weights and/or an optimal network architecture.

GAs have been used to search for optimal interconnection weights in the weight space of a multiplayer, feedforward network without using any gradient information (Montana & Davis, 1989). Unlike the backpropagation using gradient method, a GA can avoid local minimum traps while performing a global search for best set of connection weights. Literature has reported on the superiority of a set of network weights selected by GA (Whitley et al., 1990; Sexton et al., 1998).

GAs have also been used to search in the space of all possible ANN architectures. Schaffer et al. (1990) propose the use of GA to evolve ANN architecture. Their method of representing a network architecture in a string allows for the possibility of including or excluding a hidden node/layer and changing network learning parameters during the evolutionary process. The method of optimizing network architecture with GA has been investigated by many other researchers (Davis, 1991). A neural genetic network behaves similar to nonlinear, nonparametric stepwise regression without any *a priori* assumption on the functional form of the relationship among data (Reeves & Rowe, 2002).

4. Mixture of Neural Networks in Estimation and Forecasting of Klein Model I

This study experiments the mixture-of-experts network architecture for estimation and forecast a mixture of temporal and non-temporal variables. The proposed network is able to account for not only the nonlinearity and dynamics but also the simultaneous and contemporaneous effects of the variables in an economic system.

For comparative purposes on estimation and forecasting of Klein Model I, relative performances of previous estimations versus one of the proposed system are evaluated on ex-post forecast for the period 1921-1941. Data for estimation are taken from Klein (1950). Then, the Klein's Model I framework is used to train and validate the *ex-ante* forecast ability of ANN on a moving window scheme from 1950 to 1994. Data are taken from National Income and Products Accounts of the United States 1929-1994 (U.S. Department of Commerce, 1998). Within this time horizon, a moving window frame is implemented. In each window, 20 annual periods are used for estimation, 5 subsequent periods for validation, and the next 5 periods for testing.

If one relaxes the linear restriction on the relationships among variables of the Klein Model I, then the reduced form of the Klein Model I (Theil & Boot, 1962; Intriligator, 1978) can be specified as:

$$y_t = f\,(y_{t-1},\; z_t) \tag{17}$$

where

$$\boldsymbol{y}_t = [\; P \;\; Y \;\; K \;\; C \;\; Wp \;\; I]^T \tag{18}$$

$$\boldsymbol{y}_{t-1} = [P_{t-1} \;\; Y_{t-1} \;\; K_{t-1} \;\; C_{t-1} \;\; Wp_{t-1} \;\; I_{t-1}]^T \tag{19}$$

$$z_t = [Wg \;\; T \;\; G \;\; Wg_{\,t-1} \;\; T_{t-1}]^T \tag{20}$$

Since the system has a group of temporal variables and a group of non-temporal variables, ANN needs two modules to learn the specific patterns of each type of variables. ANN also has a gating network to aggregate modular estimations into final results and to account for the simultaneous and contemporaneous effect of endogenous variables. The proposed mixture-of-experts network estimates Klein Model I with two-stage and modular architectures.

4.1 Two-Stage Estimation

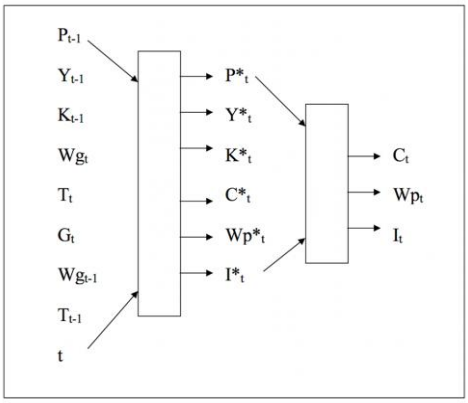

Figure 1. Two-stage ANN Estimation of Klein Model 1

Since the endogenous variables are contemporaneously related, it is not accurate to estimate them with a single equation approach. The relationship of endogenous variables and other variables of the system are estimated herein in the instrumental stage. Although these variables are estimated simultaneously, their contemporaneous effect has not been taken into account. Consequently, these instrumental estimations will be mapped to their actual values to account for this contemporaneous effect in the final stage.

4.2 Modular Estimation

In an economic system, some endogenous variables are affected by their lagged values. Also, the depth of lagged effects may vary across endogenous variables. In addition, some variables of the model may be affected by a certain exogenous variable *a priori*. Without modular estimation for each effect, it could be very difficult to approximate accurately the mixture of temporal and non-temporal variables. Consequently, the ANN should have different modules at the instrumental stage to capture these lagged effects or specified effects separately. Instrumental output results from modular estimations are aggregated at the final stage to account for the contemporaneous effect of all endogenous variables in the network outcome.

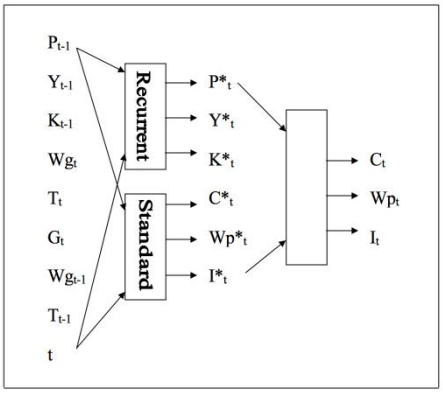

Figure 2. Modular ANN Estimation of Klein Model 1

Specifically, in the modular estimation of Klein Model I, the instrumental stage has two modules: a *recurrent module* to estimate P_t , Y_t , K_t taken into account their lagged effect, i.e., P_{t-1}, Y_{t-1}, K_{t-1}, and a *standard module* to estimate C_t, Wp_t, I_t. Then these instrumental estimations P^*_t, Y^*_t, K^*_t , C^*_t, Wp^*_t, and I^*_t are mapped to their actual values P_t, Y_t, K_t, C_t, Wp_t, and I_t to account for their contemporaneous effect.

From the initial structure of mixture-of-networks, GAs are used to select the optimal network topology at each stage and for each module in the ANN estimation. The fitness of each topology is evaluated on two criteria: the simplicity in terms of number of hidden layers and hidden nodes and relative performance in terms of its discrepancy between network outputs and desired targets.

In the following, the network configuration is represented as I-H1F-H2F-OF where I is the number of input nodes, H1 is the number of nodes in the first hidden layer, H2 is the number of hidden nodes in the second layer, O is the number of output nodes, and F is the transfer function choosing from a pool of logistic sigmoid functions (L), hyperbolic tangent functions (T), and linear functions (Lin). For instance, the notation of 9-7L-3T denotes a network configuration of 9 input nodes, one hidden layer with 7 nodes using sigmoid logistic transfer functions, and 3 output nodes using hyperbolic tangent transfer functions.

5. Findings and Discussion

5.1 Two-Stage ANN Estimation of the Klein Model I

At each stage of estimations (i.e., instrumental stage and final stage), the selected network is trained in 30 runs; each run lasts 1,000 epochs with different initial random weights at each stage of estimation, minimum and maximum errors are recorded. This results in two streams of data representing instrumental estimations, one with maximum error and the other with minimum error. These streams of instrumental estimations are used in final estimation of system equations. The GA selects a network configuration of 9-7L-6T for the instrumental stage and 6-6L-3T for the final stage. Table 1 reports the performance of network with minimum/maximum error at the instrumental stage and at the final stage. The rationale of this recording is to evaluate the case where the network is trained in just one run at each stage, what would be the boundary of errors in the best and the worst estimations from 30 runs.

Results in Table 1 show that the performance of the ANN is superior to those of traditional methods (Klein, 1950; SAS, 1984) and Caporaletti et al. (1994). The comparison is based on Sum of Squared Errors (SSE) in the estimation of each exogenous variable as well as Total SSE in the estimation of the whole system at different training times.

Caporaletti et al. (1994) use an ANN to estimate each endogenous variable of the system, one by one. This single equation estimation approach may not capture well the simultaneous and contemporaneous effects of other endogenous variables in the economic system as the system estimation approach used in this study.

5.2 Modular ANN Estimation of the Klein Model I

Using mixture-of-experts network architecture, the network configuration in this study has two modules: a recurrent module and a standard one. The recurrent module is refined to learn the lagged effect on related temporal variables. The standard module learns the inter-relationship of other variables in the system. Then instrumental estimations from these two modules are processed in the final stage with the mapping of instrumental estimations to desired targets of the system. This mapping accounts for the contemporaneous and simultaneous effects on the final estimation of the endogenous variables.

Table 1. Two-Stage ANN Estimation of the Klein Model I

	Total SSE	SSE		
		C	*Wp*	*I*
2SLS*	60.97260	21.92525	10.000496	29.04686
LIML*	85.90255	40.88414	10.021920	34.99649
3SLS*	73.60150	18.72696	10.920560	43.95398
ML**	56.26009	22.08910	10.218495	23.95249
CapoANN***	32.4987	9.5356	9.8813	13.0813
Two-Stage ANN				
Min**** - Min*****	18.72112	6.805794	12.80344	2.431763
Min**** - Max*****	28.53662	9.140769	16.64345	2.752003
Max**** - Min*****	23.33051	4.796138	13.92080	4.613574
Max**** - Max*****	33.87866	8.404020	20.55768	4.916960

Legend: * SAS (1984) SAS/ETS User's Guide, Version 5; ** Klein (1950); *** Caporaletti et al. (1995); **** Max/Min error on Instrumental ANN Estimation in 30 runs; ***** Max/Min error on Final Stage ANN Estimation in 30 runs.

Table 2. Modular ANN Estimation of the Klein Model I

	Total SSE	SSE		
		C	*Wp*	*I*
2SLS*	60.97260	21.92525	10.000496	29.04686
LIML*	85.90255	40.88414	10.021920	34.99649
3SLS*	73.60150	18.72696	10.920560	43.95398
ML**	56.26009	22.08910	10.218495	23.95249
CapoANN***	32.4987	9.5356	9.8813	13.0813
Modular ANN				
Min**** - Min*****	6.49236	2.548467	3.376657	.567239
Min**** - Max*****	10.71094	3.150886	6.57862	.963706
Max**** - Min*****	13.81735	5.598224	6.402702	1.816427
Max**** - Max*****	17.79338	4.621934	10.73451	2.436934

Legend: * SAS (1984) SAS/ETS User's Guide, Version 5; ** Klein (1950); *** Caporaletti et al. (1995); **** Max/Min error on Instrumental ANN Estimation in 30 runs; ***** Max/Min error on Final Stage ANN Estimation in 30 runs.

In the instrumental stage, GA selects the configuration of 9-6L-3T for the recurrent module and 9-7L-3T for the standard module. Each network module is trained in 30 runs; each run lasts 1,000 epochs with different initial random weights. It results in two streams of data representing instrumental estimations to be used in the final estimation of system equations. In the final stage, GA selects the configuration of 6-6T-3T for the network. The final network is trained in 30 runs, each run lasts 1000 epochs with different initial random weights. In each module, minimum and maximum errors of estimation are recorded to define the boundary of errors. Results from modular ANN estimation are reported in Table 2.

In all cases, the results obtained from modular ANN estimations are superior to those of two-stage ANN and traditional methods reported in the previous section. The reason for this improvement is that the temporal effect of lagged endogenous variables on the system is taken into account explicitly in modular estimation.

5.3 Modular ANN Forecasting of the Klein Model I

This study uses the variables defined in the Klein Model I to forecast the related endogenous variables for the period from 1950 to 1994. As the US economy grows dramatically, the level of macroeconomic variables in this period increases accordingly. It would be difficult for a network to deal with variables whose values increase to an unbound limit and spacing with big gaps. As an alternative, this study considers a more compact space and focuses on the growth rate of related endogenous variables. Consequently, related data in the period are transformed into first differences of their natural logarithmic values to capture their *growth rates*. In the following, the growth rates of consumption, private wages, and net investment are indicated as *DLC*, *DLWp*, *DLI*, instrumental estimates as *DLC**, *DLWp**, *DLI**, and final estimates as *DLC***, *DLWp***, *DLI***, respectively.

Table 3. Modular ANN Estimation and Forecasting of the Klein Model I

	SSE		
	DLC	*DLWp*	*DLI*
Period of 1950-79:			
Training (1950-69)	.000331	.000832	.017397
Testing (1970-74)	.001794	.001821	.096760
Forecasting (1975-79)	.008602	.001209	.267968
Period of 1955-84:			
Training (1955-74)	.000682	.001172	.008684
Testing (1975-79)	.000751	.002552	.074401
Forecasting (1980-84)	.004883	.010367	.758536
Period of 1960-89:			
Training (1960-79)	.000752	.001325	.019973
Testing (1980-84)	.006123	.007730	.082321
Forecasting (1985-89)	.012539	.024928	.727144
Period of 1965-94:			
Training (1965-84)	.001085	.001471	.028825
Testing (1985-89)	.002213	.002426	.136035
Forecasting (1990-94)	.014209	.012759	.164281

Following Klein (1950) and Klein-Goldberger (1955) concentrated on the sign of the forecast residual, the current analysis focuses on the ability of the ANN to pick up the future direction of related variables. The experiments have been conducted with available data from US Bureau of Census from 1950 to 1994, being divided into 30-year moving time windows. For each window, 20 yearly periods are used for training, the next 5 for testing, and the subsequent 5 for validation or forecasting. GA is used to select the appropriate configuration for each module. The best network is used to make forecasts for the next 5 out-of-sample periods of the time frame.

5.3.1 Period from 1950 to 1979

For this period, data from 1950-1969 are used for training, 1970-1974 for testing, and 1975-1978 for forecasting. GA selects a network configuration of 9-4L-3T for the recurrent module, 9-5L-3T for the standard module, and 6-4L-3T for the final stage.

For *DLC*, the network learns well as it captures correctly the changes in direction of this variable with a SSE of .000331. However, in the test/forecasting period, the network projects a slight fluctuation at a lower level when the related variable started fluctuating in an upward trend (Figure 3). The SSEs in test and forecast periods are .001794 and .008602, respectively. The network does not experience these high growth levels of DLC in an upward trend. Consequently, it provides moderate forecasts.

For *DLWp*, the network learns well the data patterns and follows closely the changes in direction of the variable with a SSE of .000832. In the test and forecast periods, the network picks up the changes in direction with SSEs of .001821 and .001209, respectively. One notes that as the network has learned the large fluctuation patterns in the training set, it is able to forecast a moderate level while following the future directions of the data (Figure 4).

For *DLI*, network forecasting picks up the changes in direction of the variable as it already learned the fluctuated patterns in the data. The SSEs for training and testing periods are .017397 and .09676, respectively. However, when the variable fluctuates at a wide level (in 1974-75), the network has not experienced this new pattern to make a close prediction. Therefore, it produces forecasts at moderate levels. The SSE for this forecast period is .267968 (Figure 5).

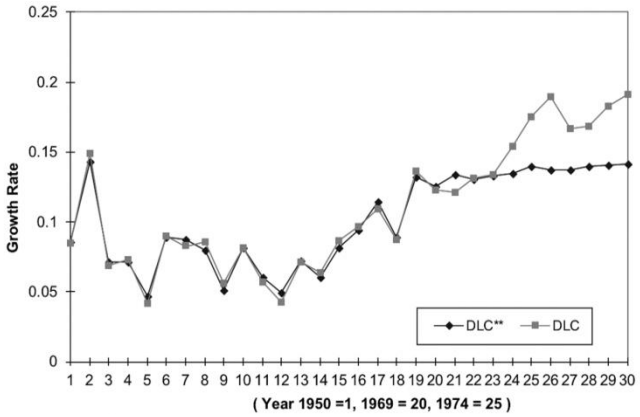

Figure 3. Estimation and Forecasting of *DLC* (1950-1980)

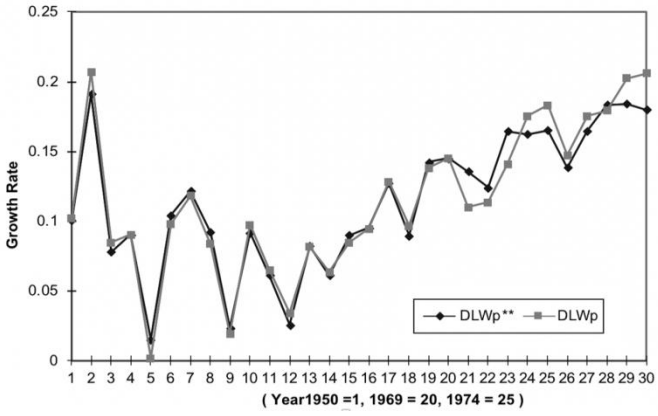

Figure 4. Estimation and Forecasting of *DLWp* (1950-1980)

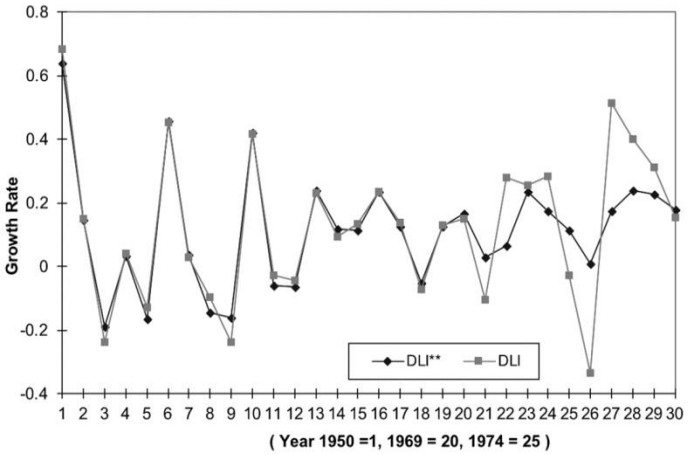

Figure 5. Estimation and Forecasting of *DLI* (1950-1980)

5.3.2 Period from 1955 to 1984

In this window, data from 1955-1974 are used for training, 1975-1979 for testing, and 1980-1984 for forecasting. GA selects a network configuration of 9-6T-3T for the recurrent module, 9-6T-3T for the standard module, and 6-7T-3T for the final stage.

For *DLC*, the network learns well the upward trend in the training set by following correctly the changes in direction of the variable with a SSE of .000682. It is able to pick up the patterns in the test period with a SSE of .000751. When the future data (1980-84) fluctuated in a new downward pattern, the network produced a dampening forecast at a moderate level (Figure 6). As the network has not experienced this pattern, it produces a SSE of .004883 for the forecast period.

For *DLWp*, after learning well the upward trend in the training set with a SSE of .001172, the network forecasts a slight fluctuation at moderate level when the future data (1980-84) fluctuate in a new pattern (Figure 7). SSEs for the test and forecast periods are .002552 and .010367, respectively.

For *DLI*, after learning well the fluctuation in the training set with a SSE of .008684, ANN forecasts follow the future data pattern. However, the network had not experienced the large changes in the levels of the extreme variation in the forecast period (e.g., the changes in 1982 to 1984). Consequently, when future data start fluctuating widely (1980-84), the network produced forecasts at moderate levels (Figure 8). SSEs for the test and forecast periods are .074401 and .758536, respectively. The large errors in the forecast period are due to the large changes in levels of data that the network is unable to capture.

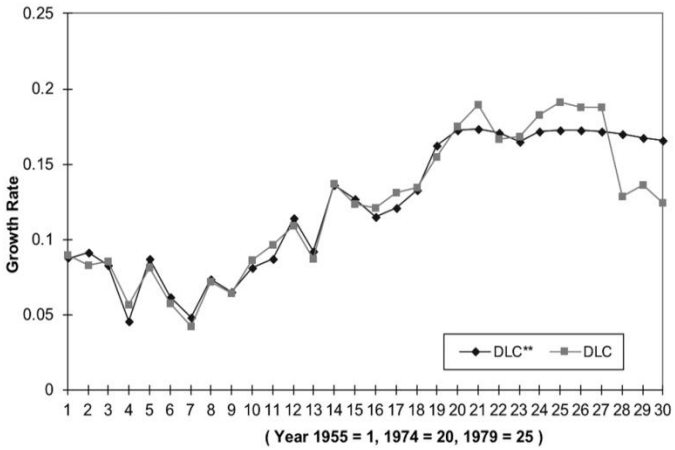

Figure 6. Estimation and Forecasting of *DLC* (1955-1985)

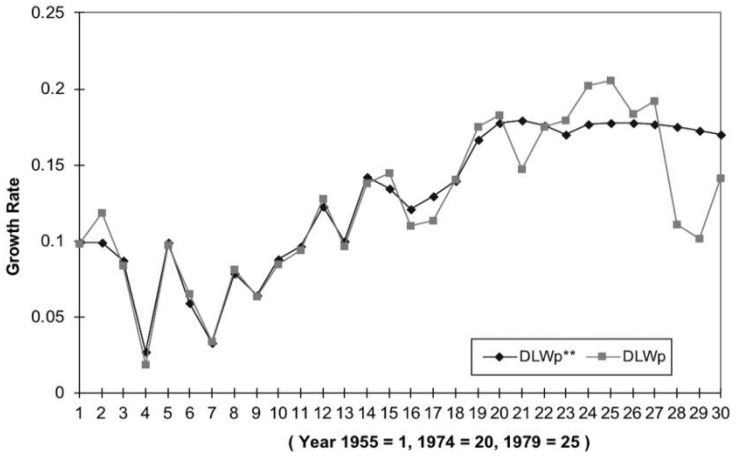

Figure 7. Estimation and Forecasting of *DLWp* (1955-1985)

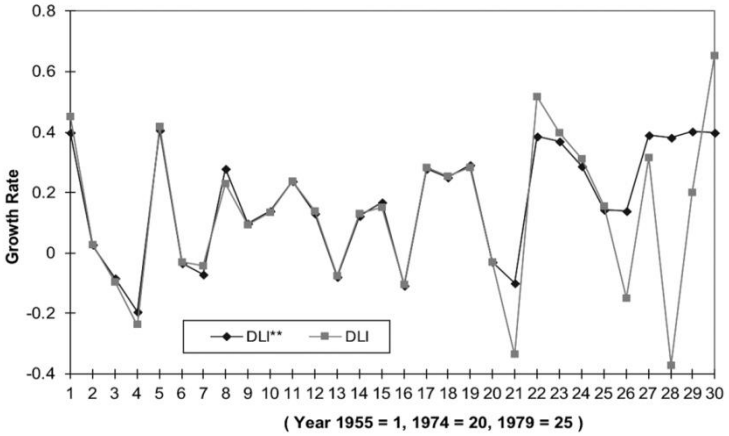

Figure 8. Estimation and Forecasting of *DLI* (1955-1985)

5.3.3 Period from 1960 to 1989

In this window, data from 1960-1979 are used for training, 1980-1984 for testing, and 1985-1989 for forecasting. GA selects a network configuration of 9-5T-3L for the recurrent module, 9-6T-3T for the standard module, and 9-3L-3T for the final stage.

For *DLC*, the network learns well the upward pattern of the training set with a SSE of .000752. When future data starts a downward trend (1985-1989), the network has not experienced such a large change in levels to produce closer forecasts. As a result, network forecasts follows the future directions at higher levels (Figure 9). SSEs for the test and forecast periods are .006123 and .012539, respectively.

For *DLWp*, the network learns well the upward pattern of the training set with a SSE of .001325. When future data starts a downward trend, ANN forecasts follow the trend but at higher levels (Figure 10). Similar to learning and forecasting DLC in this period, the network has not learned the large change in levels of future fluctuation in order to provide closer forecasts. SSEs for the test and forecast periods are .00773 and .024928, respectively.

For *DLI*, the network learns well the fluctuation in the training set with a SSE of .019973. When future data start dropping (1985) and then fluctuating at a lower level (1985-89), the network forecasts follow the trend but at a higher level (Figure 11). SSEs for the test and forecast periods are .082321 and .727144. The large error for the forecast period is due to the change in data patterns as they fluctuate at a moderate levels that the network is unable to follow closely.

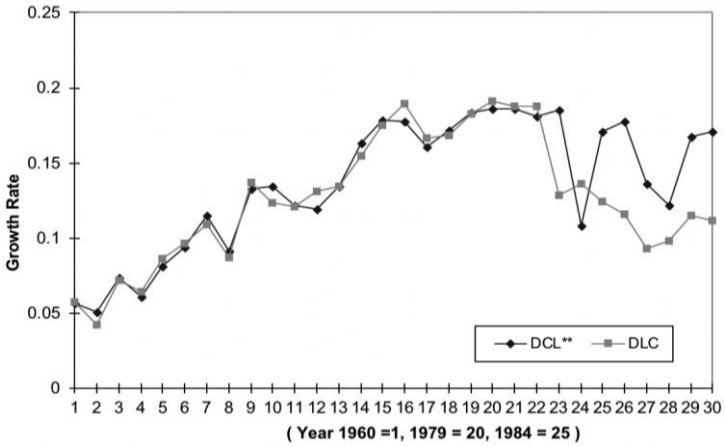

Figure 9. Estimation and Forecasting of *DLC* (1960-1990)

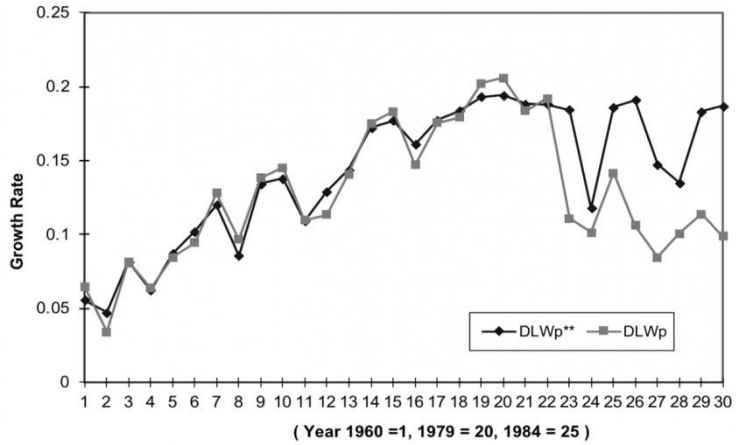

Figure 10. Estimation and Forecasting of *DLWp* (1960-1990)

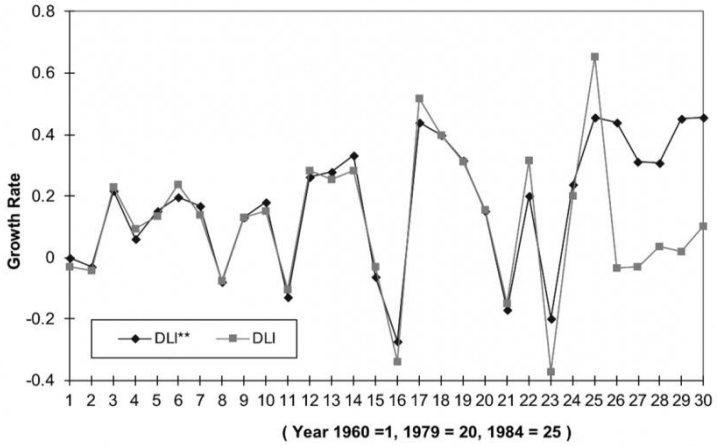

Figure 11. Estimation and Forecasting of *DLI* (1960-1990)

5.3.4 Period from 1965 to 1994

In this window, data from 1965-1984 are used for training, 1985-1989 for testing, and 1990-1994 for forecasting. GA had selected a network configuration of 9-5L-3T for the recurrent module, 9-3L-3T for the standard module, and 9-5L-3T for the final stage.

For *DLC*, the network learns well the upward trend in the training set with a SSE of .001085. When future data start a long downward trend, the network has not learned these patterns in order to predict accurately the future

level and, on some occasions, the changes of directions, e.g., in 1990 (Figure 12). SSEs for the test and forecast periods are .002213 and .014209, respectively.

For *DLWp*, the network learns well the upward trend in the training set with a SSE of .001471. When future data has a downward trend, the network does not predict accurately the level and change of directions from the pattern that it has learned (Figure 13). SSEs for the test and forecast periods are .002426 and .012759, respectively.

For *DLI*, the network learns well the fluctuation in the training set. Since the training set contains patterns of large fluctuations, the network forecast is able to follow the trend of future data, however in a large fluctuating pattern that it has learned (Figure 14). SSEs for the test and forecast periods are .136035 and .164281, respectively. The large error from these periods is due to the sudden change in levels of fluctuation.

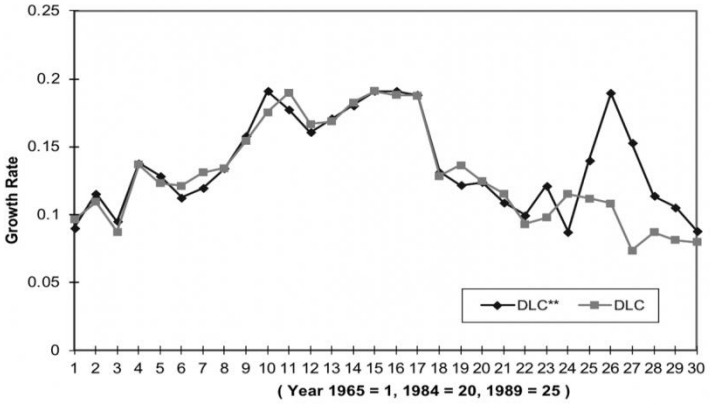

Figure 12. Estimation and Forecasting of DLC (1965-1995)

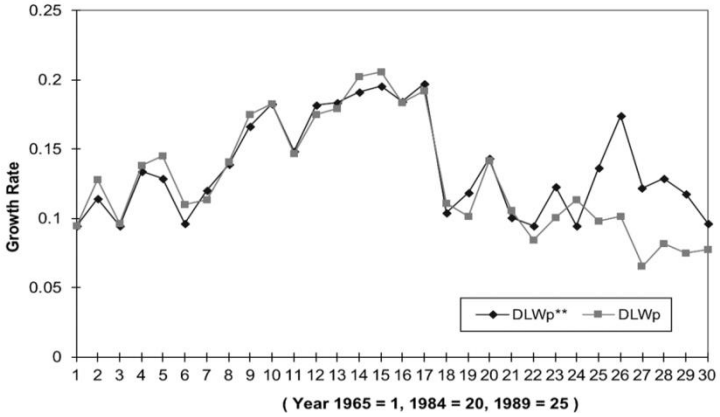

Figure 13. Estimation and Forecasting of DLWp (1965-1995)

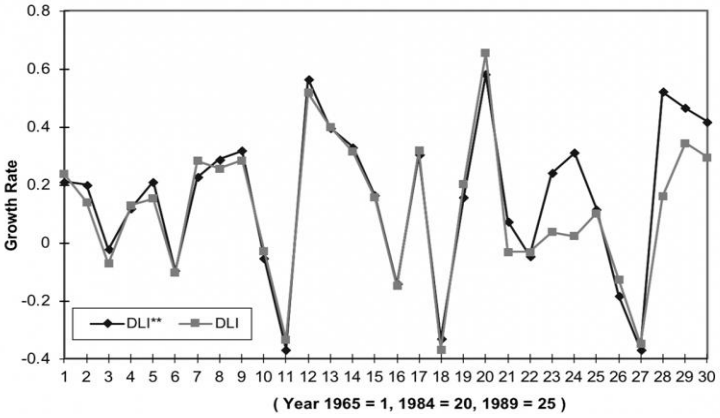

Figure 14. Estimation and Forecasting of DLC (1965-1995)

The following are some observations on the ANN behavior in learning patterns of the training set and producing forecasts on unseen data.

- The network does not learn and project the recent extreme trend. It tends to provide a moderate forecast in terms of directions and levels.
- If the network has been trained with data having an upward trend and related variable to be predicted fluctuates in a downward trend, network forecasts will be dampened at a middle level.
- If the network has not experienced drastic level changes in the training set, it produces a forecast following the trend but at a higher level for future downward change and lower level for future upward change.
- If the network is trained with the fluctuated pattern, its forecasts follow the future trend but at a moderate change in level. The larger the variation in the training set, the closer the ANN will follow the patterns in the forecast period in terms of directions and levels.
- The network cannot predict accurately a level outside the range of pattern it has learned from the training set. When it encounters such a case, it produces a forecast at an average level of the data in the training set.

Consequently, in order to improve its forecasting ability, a network should experience with the variation in trend (upward/downward, long/short fluctuations) and the possible highest and lowest levels of data patterns. Experiments illustrate that the more variations exist in the training set, the closer ANN follows with future fluctuations in terms of directions and levels. In any case, the ANN forecasting tends to be conservative, not following immediately the drastic upward or downward trend.

6. Concluding Remarks

From extensive experiments with the Klein Model I, an integrated ANN and GA in mixture-of-experts network architecture has provided evidences of effective alternatives to traditional estimation / forecasting techniques to handle a mix of temporal and non-temporal variables. One can use hierarchical networks to conduct instrumental estimations. One can also partition the problem space into domains and assign them to modular ANN to learn the related patterns.

The versatile technique of integrated mixture-of-experts ANN overcome the imposed assumptions on the behaviors of related variables, the specification of exact relationship, and the difficulty in nonlinear estimation of the economic model. The GA helps overcome the sub-optimality of the tedious trial-and-error process in network building. The flexible network architecture offers many alternative network configurations to capture the peculiarities of variables in a problem space before aggregating intermediate estimations into final results. The integrated system processes effectively the mixture of variables, and produces efficient estimations and forecasts.

In a future work, seasonal patterns of temporal data would be explicitly recognized in an integrated ANN system. In that architecture, GA could be used to define a minimum number of input nodes (time lags) that is still capable to provide accurate forecasts. Such a design would alleviate the ANN technique from using a traditional statistical technique, such as the Box-Jenkins method (1976), to determine time lags, and consequently the memory, of a neural forecasting system.

References

Aydin, A. D., & Cavdar, S. C. (2015). Two Different Points of View through Artificial Intelligence and Vector Autoregressive Models for Ex Post and Ex Ante Forecasting. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2015/409361

Bodkin, R. G., Klein, L. R., & Marwah, K. (1991). *A History of Macroeconomic Model-Building*. Edward Elgar Publishing Ltd.

Box, G., & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-day.

Caporaletti, L. E., Dorsey, R. E., Johnson, J. D., & Powell, W. A. (1994). A Decision Support System for In-sample Simultaneous Equation Systems Forecasting Using Artificial Neural Systems. *Decision Support Systems, 11*, 481-495. https://doi.org/10.1016/0167-9236(94)90020-5

Cheng, B., & Titterington, D. M. (1994). Neural Networks: A Review From A Statistical Perspective. *Statistical Science, 9*(1), 2-54. https://doi.org/10.1214/ss/1177010638

Cromwell, J. B., Hannan, M. J., Labys, W. C., & Terraza, M. (1994). *Multivariate Tests for Time Series Models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series no. 07-100). CA: Sage.

Cybenko, G. (1989). Approximation by Superpositions of A Sigmoid Function. *Mathematics of Control, Signals, and Systems, 2*, 303-314. https://doi.org/10.1007/BF02551274

Davis, L. (Ed.) (1991). *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold.

Dreyfus, G. (2005). *Neural Networks Methodology and Applications*. Berlin: Springer.

Funahashi, K. (1989). On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks, 2*, 355-363. https://doi.org/10.1016/0893-6080(89)90003-8

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. MA: Addison-Wesley.

Graupe, D. (2013). *Principles of Artificial Neural Networks* (3rd Ed.). Singapore: World Scientific Publishing. https://doi.org/10.1142/8868

Greene, W. H. (2011). *Econometric Analysis* (7th ed.). New York: Pearson.

Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. Cambridge, MA: The MIT Press.

Haykin, S. (2009). *Neural Networks and Machine Learning* (3rd ed.). Pearson: New York.

Holland, J. (1975). *Adaptation in Neural and Artificial Systems*. Michigan: The University of Michigan Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multi-Layer Feedforward Networks Are Universal Approximators. *Neural Networks, 2*, 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Hwang, J. N., Lay, S. R., Maechler, M., & Martin, R. D. (1994). Regression Modeling in Back-Propagation and Projection Pursuit Learning. *IEEE Transactions on Neural Networks, 5*(3), 342-353. https://doi.org/10.1109/72.286906

Intriligator, M. D. (1978). *Econometric Models, Techniques, and Applications*. New Jersey: Prentice Hall.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991a). Task Decompositions through Competition in a Modular Connectionist Architecture. *Cognitive Science, 15*(2), 219-250. https://doi.org/10.1207/s15516709cog1502_2

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. (1991b). Adaptive Mixtures of Local Experts. *Neural Computation, 3*, 79-87. https://doi.org/10.1162/neco.1991.3.1.79

Jordan, M. I. (1986). *Serial Order: A Parallel Distributed Processing Approach*. UC San Diego, Institute for Cognitive Science Report 8604.

Jordan, M. T., & Jacobs, R. A. (1995a). Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation, 6*(2), 181-214. https://doi.org/10.1162/neco.1994.6.2.181

Jordan, M. T., & Jacobs, R. A. (1995b). Modular and Hierarchical Learning Systems. In M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: The MIT Press.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T. C. (1985). *The Theory and Practice of Econometrics* (2nd ed.). New York: John Wiley.

Klein, L. R. (1950). *Economic Fluctuation in the United States, 1921-1941*. New York: John Wiley.

Klein, L. R., & Goldberger, A. S. (1955). *An Econometric Model of the United States 1921-1952*. Amsterdam: North-Holland.

Kuan, C. M., & White, H. (1994). Artificial Neural Networks: An Econometric Perspective. *Econometric Reviews, 13*(1), 1-91. https://doi.org/10.1080/07474939408800273

Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of Experts: A Literature Survey. *Artificial Intelligence Review, 42*, 275-293. https://doi.org/10.1007/s10462-012-9338-y

Mills, T. C. (1991). *Time Series Techniques for Economists*. MA: The Cambridge University Press.

Montana, D. J., & Davis, L. D. (1989). Training Feedforward Networks Using Genetic Algorithms. *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, 1989.

Moody, J. (1995). Economic Forecasting: Challenges and Neural Network Solutions. Keynote talk presented at the *International Symposium on Artificial Neural Networks*, Hsinchu, Taiwan, December 1995.

Moshiri, S., & Cameron, N. (2000). Neural Network versus Econometric Models in Forecasting Inflation. *Journal of Forecasting, 19*(3), 201-217. https://doi.org/10.1002/(SICI)1099-131X(200004)19:3<201::AID-FOR753>3.0.CO;2-4

Nguyen, D. D., & Kira, D. (1997). Using Artificial Neural Networks and Vector Autoregressive Method in Multiple Time Series Forecasting. In *Proceedings of Decision Sciences Institute Annual Meeting*, San Diego, November 1997.

Prankatz, A. (1991). *Forecasting with Dynamic Regression Models*. New York: John Wiley. https://doi.org/10.1002/9781118150528

Reeves, C., & Rowe, J. (2002). *Genetic Algorithms: Principles and Perspectives*. Berlin: Springer

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of Royal Statistical Society, B, 56*(3), 409-456.

Rosenblatt, F. (1959). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review, 65*, 386-408. https://doi.org/10.1037/h0042519

Ruud, P. (2000). *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.

SAS Institute. (1984). *SAS/ETS User Guide, Version 5*.

Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *Neural Networks, 61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Sexton, R. S., Dorsey, R. E., & Johnson, J. D. (1998). Toward Global Optimization of Neural Networks: A Comparison of the Genetic Algorithm and Backpropagation. *Decision Support Systems, 22*, 171-185. https://doi.org/10.1016/S0167-9236(97)00040-7

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica, 48*(1), 1-48. https://doi.org/10.2307/1912017

Steeb, W. H. (2008). *The Nonlinear Workbook* (4th ed.). Singapore: World Scientific Publishing. https://doi.org/10.1142/6883

Theil, H. (1971). *Principles of Econometrics*. John Wiley: New York.

Theil, H., & Boot, J. C. G. (1962). The Final Form of Econometric Equation Systems. *Review of the International Statistical Institute, 30*, 136-152. https://doi.org/10.2307/1401895

Tong, H., Kumar, K., & Huang, Y. (2011). *Developing Econometrics.* West Sussex, UK: John Wiley. https://doi.org/10.1002/9781119954231

U.S. Department of Commerce, Bureau of Economic Analysis. (1998). *National Income and Product Accounts of the United States, 1929-94*. Washington DC: U.S. Government Printing Office.

Whitley, D., Starkweather, T., & Bogart, C. (1990). Genetic Algorithms and Neural Networks: Optimizing Connections and Connectivity. *Parallel Computing, 14*, 347-361. https://doi.org/10.1016/0167-8191(90)90086-O