

## **Validating Dynamic General Equilibrium Model Forecasts**

NaiChia Li

Research Associate, Dept. Applied Economics, University of Minnesota

Lixx0219@umn.edu

612-625-9715 (Office)

612-625-2729 (Fax)

Terry Roe

Professor, Dept. Applied Economics, University of Minnesota

troe@umn.edu

612-625-6706 (Office)

612-625-2729 (Fax)

## **Validating DGE Model Forecasts**

Contact Author:

NaiChia Li

Research Associate, Dept. Applied Economics, University of Minnesota

[Lixx0219@umn.edu](mailto:Lixx0219@umn.edu)

612-625-9715 (Office)

612-625-2729 (Fax)

## **Abstract**

The maintained hypotheses of micro-foundation based general equilibrium models often required to estimate a relatively large number of parameters. This has tended to make economists and policy makers insecure in their use to forecast and to provide a structural explanation of an economy's behavior. Advances in the empirical application of neoclassical growth models have exacerbated this insecurity and brought questions of validation to the forefront. Current methods, such as forecast accuracy and association measures, are adopted to evaluate the magnitude of bias in the forecasts of such models. Two innovative methods are also designed to provide statistically meaningful validation and require no arbitrary probabilistic assumptions on the distribution of either the forecasts or the data. To illustrate these methods, a multi-sector Ramsey-based model is calibrated to Taiwanese data for the year 1988 and used to provide forecast through 2003. The methods developed suggest the model forecasts various dimensions of the economy surprisingly well. (JEL Codes: C40, O41, O47. Keywords: Dynamic general equilibrium models, Forecast validation.)

# 1 Introduction

The reluctance to validate calibration models has been one of the main critiques leading to some skepticism as to their value in explaining and predicting economic events. Kehoe et. al. (1995, p.116) state that “Only by showing a model can replicate, and to some extent, predict principal developments that take place in the economy that it intends to represent can we justify the effort put into a large-scale quantitative model.” We agree with this view, and feel that advances in the empirical application of neoclassical growth models have exacerbated the need of validation and brought questions of validation to the forefront.

To address the long-due need of validation, this paper adopts current forecast accuracy measures and proposes innovative methods to statistically compare forecasts from calibrated dynamic general equilibrium (DGE) models with data. A concordance correlation statistic (Lin (1989)) is applied to measure the deviation of forecasts from the real growth path. A loss function method is introduced to provide a time interval over which forecast bias stay within a tolerance level pre-determined by researchers. A time-series confidence interval method is also introduced to compare the DGE model’s forecasts with that of a time series - reduced form models. The methods developed are simple to implement and require no arbitrary probabilistic assumptions on the distribution of both forecasts and data.

The paper is organized by first providing an overview of applied general equilibrium models, the type of insights they are hoped to provide, and why better validation methods are of growing importance. We then briefly describe a DGE model that is fit to Taiwanese data, and later used to illustrate the methods developed. The following sections of the paper develop and illustrate the validation methods adopted and introduced.

## 2 Overview of Applied General Equilibrium Modeling

Kehoe and Prescott (1995, p.1) describe applied general equilibrium (GE) analysis as “.. the numerical implementation of general equilibrium models calibrated to data.” A controversial aspect is the methods employed to estimate a relatively large number of parameters, and the lack of generally acceptable procedures to validate the parameters and structural features of the

model. Some parameters maybe obtained using statistical methods while other estimates can be based a single year's observation. The calibrated model is often fit to data in such a way that a solution reproduces the data of modeled economy.

Two general types of analyses are commonly performed. One involves a test of theory. For example, Gibson (2006) constructs a DGE model of producers of heterogeneous efficiency and differences in producer-level investments in trade relationships. A solution for the case of trade reform results in an increase in total factor productivity (TFP) that, when compared to other studies of Mexican TFP, provides an explanation for how trade reform increase factor productivity. In this type of analysis, failure of a model to explain the data can be viewed as a success in the sense that the maintained hypotheses are rejected. This thus creates a paradox that may only be explained by further investigation.<sup>1</sup>

Another common type is to assess the welfare effects of policy change. Recent examples are Diao et al (2002), Anderson and Martin (2005) and Bouet et al (2005). Emphasis is not only placed on model forecasts. The structural features of the model are used to provide a clear explanation of the causes underlying the modeled economy's economic adjustments and corresponding magnitudes. This provides a basis to pursue a dialogue with policy makers<sup>2</sup>.

The confidence that policy makers, economists and others place in the forecasts and dialogue motivated by the structural nature of the model would seem to depend on the model's replication of observed facts. However, with exceptions mentioned below, little effort has been made to validate these models. Some early practitioners tended to dismissed the need for validation. For example, Whalley (1986, 1988), among the first to develop a calibrated multisectoral GE model, contends that these models are not intended to forecast the values of economic variables, but rather to provide useful insights to policy makers so that they may undertake more informed and desirable policy actions.

Kehoe and his associates express an opposing view, a view which we share. Kehoe (2003) suggests that ex-post performance evaluations of applied GE

---

<sup>1</sup>On the empirical foundations of calibration, Hansen and Heckman (1996) note that calibration of GE models is the econometrician's estimation stage while validation is equivalent to their testing stage.

<sup>2</sup>However, all too frequently, emphasis is placed on model forecasts either because the author is uninformed of the model's economics, or the model structure is too complicated to discern the importance of key Stolper Samuelson and Rybczynski like effects.

models are essential if policy makers are to have confidence in the results they produced. Further, he suggests that evaluations also help make applied GE analysis a scientific discipline in which there are well-defined puzzles with clear successes and failures for competing theories. Kehoe et al (1995) perform an ex-post validation of the Spanish model developed in 1986 to analyze the impact on the Spanish economy of reforms implemented in the 1980s. Adopting correlation measures, they found that it performed well in capturing the changes that actually occurred.

The overall lack of effort to validate GE models may relate to their static nature. However, recent developments in the application and extension of neoclassical growth theory of the Ramsey-Cass-Koopmans type, examples of which are Echevarria (1997), Elbasha and Roe (1996), Diao et al (1998), and Roe et al (2003), yield empirical DGE models that provide inter-temporal forecasts. In this environment, it seems even more compelling to question the fit of the model to data. If the model is fit to a point on an economy's transition path, then the model's forward or backwards forecasts should, in principle, be confronted to time series data for validation purposes. The need to develop more systematic and rigorous validation procedures now seems even more important for the testing of theory and providing policy makers and others confidence in the analysts ability to engage in an informed dialogue.

### **3 A three sector inter-temporal model**

A three sector growth model similar to Echevarria (1997) and Roe et al (2003) is sufficiently general yet parsimonious to best show our proposed validation measures. We briefly sketch the model and then use the results to illustrate our validation procedures<sup>3</sup>.

The economy is small and consists of 3 aggregated sectors: agriculture, industry and service, indexed  $i = m, a, s$ , respectively. All firms in the economy employ two economy-wide factors, labor and capital, while in addition, agriculture employs land which is specific to the sector. The supply of workers is presumed to grow at a constant rate  $n$ . Technologies are neoclassical and constant-return-to-scale. Labor augmenting exogenous technological change, denoted  $A(t)$ , is economy-wide and grows at a constant rate  $x$ . In addition, agriculture experiences land augmenting technological change at rate  $\eta$ . The

---

<sup>3</sup>See Li and Roe (2004) for model detail and empirical findings.

agricultural and manufacturing outputs are traded internationally at given prices  $p_a$  and  $p_m$ , respectively, while the service output is only traded in the domestic economy, and hence its price, denoted  $p_s$ , is endogenous. The manufacturing sector's output is both a consumption good and a capital good which, net of depreciation, augments the economy's stock of capital. The stock of capital is the model's state variable.

The typical household is presumed to obtain utility from the consumption sequence  $\{c_m, c_a, c_s\}_{t=0}^{t=\infty}$  where account is taken of the future addition of household members and discounted by the time preference rate  $\rho$ ,

$$U = \int_{t=0}^{t=\infty} \frac{u(c_m, c_a, c_s)^{1-\theta} - 1}{1-\theta} e^{(n-\rho)t} dt \quad (1)$$

where  $\theta$  is the inverse of the household's intertemporal substitution. The the felicity function  $u(c_m, c_a, c_s)$  is homothetic, and in the empirical analysis, taken to be of Cobb-Douglas form.

Households own all of the economy's resources, leaving savings  $\dot{K}$  at each instant in time to be the difference between returns to factors of production and expenditures,

$$\dot{K} = wL + rK + \pi T - E \quad (2)$$

The wage rate is denoted by  $w$ ,  $r$ , and  $\pi$  denote the return to capital  $K$  and land  $T$  respectively, and  $E$  denotes total expenditure. Since foreign liabilities are not allowed, national assets are the economy's value of capital stock and the value of land  $P_L T$ , where  $P_L$  is the price of land. The present value Hamiltonian implies the Euler equation governing the household's allocation between consumption and savings over time.

$$\frac{\dot{c}}{c} = \frac{1}{\theta} \left[ (r - \rho) - \lambda_s \frac{\dot{p}_s}{p_s} \right] \quad (3)$$

where  $\lambda_s$  is the share of total expenditure allocated to the service good.

### 3.1 Equilibrium

Given initial conditions  $\{p_m, p_a, p_s(0), k(0), \tilde{T}(0), L(0)\}$ , (i.e., a point on the economy's transition path) a competitive equilibrium is sequence of positive prices

$$\{p_s^*, \hat{w}^*, r^*, P_L^*\}_{t=0}^{t=\infty} \quad (4)$$

allocations per worker by firms

$$\{y_m^*, y_a^*, y_s^*, k_m^*, k_a^*, k_s^*, l_m^*, l_a^*, l_s^*\}_{t=0}^{t=\infty} \quad (5)$$

and per worker household allocation to consumption and investment

$$\{k^*, c_m^*, c_a^*, c_s^*\}_{t=0}^{t=\infty} \quad (6)$$

such that firms maximize profit subject to their technologies with given prices and zero profits and households maximize (1) subject to the budget constraint (2). Here,  $y_i$  is output per economy-wide worker,  $k_i$  is capital stock employed in sector  $i$  per worker employed in that sector, and  $l_i$  is the share of workers employed in the  $i - th$  sector.

The model's primitives are calibrated to a set of parameter values, some of which draw upon econometric estimates, others either taken from previous studies or input-output data supplied by Taiwanese authorities. The temporal and intra-temporal equations are specified to satisfy the definition of equilibrium, computer code written, and numerical solutions obtained yielding values (4), (5) and (6) over an interval of time  $t^4$ . With model's forecasts as a departure point, the challenge is to assess the model's accuracy in predicting these values.

Clearly, general equilibrium models are, by their vary nature, prone to specification and estimation biases. The methods for estimating some parameters, such as productivity growth rates  $x$  and  $\eta$ , are somewhat controversial in the literature, and almost always prone to considerable variation over time. Values of other parameters, such as consumption shares, are unlikely to remain fixed through time and their time-variant behavior may require more attention in the calibration process, or the need to specify utility functions that are non-homothetic. Values of some parameters, such as factor shares or the population growth rate  $n$ , are observable and they tend to be relatively constant over time. Invariably, the modeled economy takes some variables as unchanging and exogenous. The economy modeled here treats that the rest of the world is as though it is in a steady state equilibrium so that world prices are constant over time. Most applied dynamic GE models

---

<sup>4</sup>The terminal period is often a point when the economy gets arbitrarily close to its long-run equilibrium. If the equations of motion are non-autonomous, a terminal period can be chosen arbitrarily and the equations of motion solved "backwards" to the initial period. This model is easily solved using Mathematica to code the Time Elimination method.

are non-stochastic and hence they do not take into account the effects of idiosyncratic shocks on an economy<sup>5</sup>. Instead, the "typical" model is designed to capture the long-term growth trend exhibited by the data. This point is discussed later in the paper.

## 3.2 Empirical results

The model is fit to the Taiwanese economy and solved numerically for the sequences (4), (5) and (6), starting in 1988. Details are given in Li and Roe (2004). Validation for our purposes here entails comparing these results to the observed data from 1988 to 2003.

As is typically the case, the DGE model has a larger number of endogenous variables than the available counter-part time series to evaluate them. Thus, our comparison between data and forecasts is only conducted for 6 series, GDP, industry output, agriculture output, service output, service price, and household saving. Selected graphs of four of these series are showed in Figure 1.

Figure 1 Here

>From the graphs alone, a number observations are worth mentioning. First, while the observed macroeconomic series are stochastic, the model is not designed and should not be expected to capture the effects of idiosyncratic shocks, such as business cycles or deviations from the "fundamental" growth component embodied in the data. Hodrick and Prescott (1997, p. 2) make this point in their analysis of business cycles when they state "the maintained hypothesis, based upon growth theory considerations, is that the growth component of aggregate economic time series varies smoothly over time." Following this logic, it is preferable in many cases to validate model forecasts with "smoothed" data.<sup>6</sup>

Second, the forecasts tend to exhibit a persistent bias (or error), measured as the difference between the model's forecasts and data, which grows with time. This non-stationarity in bias can be caused by the time variant nature

---

<sup>5</sup>An exception of course is the real business cycle literatuer pioneered by Kydland & Prescott (1982).

<sup>6</sup>Validation with de-randomized or smoothed data has the disadvantage of allowing the choice of filter and filter parameters to be made in such a way as to best fit model forecasts.

of some parameters that is not captured by the model. Third, the model's performance differs noticeably among various aspects of the economy. The model results clearly underestimate the overall GDP, in part because it tends to underestimate service output and price. The results on industry output however appear to fit the data exceedingly well. Depending on the purposes of the analysis, the researcher may have a preference for greater accuracy in forecasting some variables relative to others. This concern is dealt with below.

## 4 Methodology for assessing DGE results

Hypothesis testing can be performed when the probability structure of the observed data series, the model series, and the bias are known. However, the data generating processes (DGP), forecasts, and the bias are typically unknown in this type of analysis. The unknown probability structure complicates the process of meaningful statistical inference and hypothesis testing. The challenge in validating DGE models is thus to establish sensible standards for judging the performance of the model without imposing arbitrary probabilistic assumptions on results or the data.

Many forecast accuracy measures have been developed to assess the discrepancies between forecasts and data. These error or distance measures include absolute percentage error, symmetric absolute percentage error, root mean square error, normalized mean square error, Theil's U distance, and others. The main difference among them is the form of the loss function adopted. The formulas of some of these measures are provided in the Appendix.

### 4.1 Measures of association

A commonly used measure of the linear association between two random variables is the Pearson's correlation coefficient. A limiting feature of Pearson's correlation is its inability to detect the difference between the levels of the forecast and the data, i.e., deviation from a 45 degree line constructed on a plain with data and forecasts on the axes. Lin's (1989) concordance correlation coefficient overcomes this shortcoming. The concordance correlation coefficient is defined as:

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

where  $\sigma_1^2$  and  $\sigma_2^2$  are the variances of variables 1 and 2, respectively, and  $\sigma_{12}^2$  is their covariance. Their respective means are given by  $\mu_1$  and  $\mu_2$ . Of course, when employed, the sample counterparts are used.

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2} \quad (7)$$

The concordance correlation differs from the Pearson's measure in accounting for the discrepancies in the means of the compared series. Consequently, the concordance correlation coefficient is useful in detecting the deviation from the fit of the 2 series, the 45-degree line. Like any other correlation coefficient, the concordance correlation coefficient is bounded between 0 and 1 in absolute value. If the coefficient is close to one, it lends supports to model's ability in replicating the past growth path. A disadvantage is that  $\hat{\rho}_c$ s, and other error measures also in this case, are likely to vary depending upon the time span of the forecast. For the case of concordance correlation, dates farther into the future tend to give rise to a larger deviation  $(\bar{Y}_1 - \bar{Y}_2)^2$  and hence a lower estimated association.

Lin (1989) showed that for a bivariate normal series,  $\hat{\rho}_c$  is a consistent estimator of  $\rho_c$  and has desirable asymptotically normal properties. This enables the hypothesis testing on the correlation coefficient. In our case, the complication, as mentioned above, is that the probabilistic structure is unknown. Without knowledge of the probabilistic nature of the series, statistics and their inference based on normality assumptions can be misleading. Bootstrap procedures can partially remedy this problem.<sup>7</sup>

The concordance correlation coefficient, equation (7), and the Pearson's correlation coefficient are presented in Table 1 for six of the data series for

---

<sup>7</sup>The bootstrap method is often used to replace an unknown distribution with a known distribution to permit the drawing of statistical inferences from probability calculation upon the known distribution. Here, we adopt the popular non-parametric bootstrap by assuming the sample  $\{x_1, x_2, \dots, x_n\}$  from an unknown distribution can be successfully represented by  $f(x) = \begin{cases} 1/n, & \text{if } x \in \{x_1, x_2, \dots, x_n\} \\ 0, & \text{otherwise} \end{cases}$ . A Word of caution. The benefit of bootstrapping however heavily depends on how well the known distribution closely approximates the unknown distribution. Also, problems, such as small sample size in our case, cannot be solved by applying bootstrap techniques.

the period 1988 to 2003. The results suggest that the concordance correlation can better detect reproducibility between two series than can Pearson's correlation. For example, the model's forecast of the aggregate price of service yields a Pearson's coefficient of  $\rho = 0.91$ , suggesting a relatively high correlation between the forecast and the data. However, the Concordance correlation is only .07. This result suggests that the "level" of the forecast service price departs rather substantially from the data. For the case of industry and service output, both measures show relatively high correlation and good forecast of the levels of these two variables, on average over the covered period.

Table1 Here

A calibrated model may yield a trend close to the trend of the data, but the two series may differ in a constant portion in their respective means (such as two parallel lines). In this case, the penalty term  $(\bar{Y}_1 - \bar{Y}_2)^2$  in concordance correlation, can mask the success of the model in correctly forecast the growth rates of the economy. To avoid this situation, concordance correlation may be performed on "normalized" data. The "normalization" process<sup>8</sup> scales the model forecasts and data so that they share the same initial value of unity. A time t comparison between the normalized data and the normalized forecasts removes the initial "scale" errors of calibration. The rationale is, in dynamic models, the data may allow for consistent intra-temporal calibration<sup>9</sup>, but may not be of sufficient quality to allow for consistent inter-temporal calibration of the model. In the normalized case, the mean difference  $(\bar{Y}_1 - \bar{Y}_2)$  in (7) is reduced leading to a higher Concordance coefficient.

## 4.2 Polynomial Time Function of Bias

We next turn to the question of "How close is close enough?" What is the maximum acceptable forecast error?. A 4% error in the forecasts may be acceptable for the needs of one analyst, but not for another. In the time function procedure discussed here, researchers can choose an error tolerance

---

<sup>8</sup>The analyses proceeds by obtaining the accuracy measures and correlation between  $\hat{y}_{ht} = \hat{Y}_{ht}/\hat{Y}_h(0)$  and  $y_{ht} = Y_{ht}/Y_h(0)$  for the sixteen year period  $t = 1988, \dots, 2003$ .

<sup>9</sup>The calibration to the "static equations" which describe equilibrium at a point in time, taking as given variables determined by the equations of motion, may reproduce the static aspects of the model exactly. However, the equations of motion are typically more difficult to fit to the data precisely.

suitable to their research interests. The procedure provides an estimate of how far beyond the last time period of data the model results fall within the predetermined tolerance. The procedure is to express the difference between model forecasts and data as a loss function. Define bias as a norm (or distance measure) equal to  $\|Y_{ht} - \hat{Y}_{ht}\|$ , where  $Y_{ht}$  and  $\hat{Y}_{ht}$  are the observed and predicted values respectively for a series  $h$  at time  $t$ .<sup>10</sup> The norm (or bias) for each series,  $b_{ht}$ , is express as a function,  $F_h(t)$ , of time,  $t$ ,

$$\|Y_{ht} - \hat{Y}_{ht}\| = \mathbf{b}_{ht} = F_h(t) + \varepsilon_{ht} \quad (8)$$

$$lb \leq F_h(t) \leq ub$$

$$T = F_h^{-1}(lb, ub)$$

where  $\varepsilon_{ht}$  is the non-systematic error component. The function  $F_h(t)$  can be viewed as the "best" fit to express the norm,  $b_{ht}$ , as a function of time. Although constrained by the behavior of the bias, the function  $F_h(t)$  should be such as to allow for the derivation of a "meaningful" root, e.g., to be invertible and an one-to-one mapping.

Next, set the LHS of the equation equal to a chosen bound, say  $\delta$ , then solve the equation for time,  $T = F_h^{-1}(\delta)$ .  $T$  is now the number of years into the future the discrepancies between forecasts and data are within the tolerance for bias established by the researcher. The value of  $\delta$ , can be viewed as the maximum tolerance that the researcher is willing to accept based upon the form of  $F_h(t)$  and the behavior of bias.

The empirical application of (8) proceeds as follows. Percentage errors between the smoothed data<sup>11</sup>, as described above, and results were used as the measure of the model's forecast errors. These percentage errors are

---

<sup>10</sup>To not penalize the model for what it is not intended to capture, the observed data can be "de-randomized" or "smoothed" before it is compared to model forecasts.

<sup>11</sup>There are many smoothing techniques that can be used to de-randomize or smooth the data used to estimate the time functions  $F_h(t)$  in (8). The Hodrick-Prescott (HP) filter is often used to estimate the unobservable non-stationary trend growth component of a series, from which the unobservable stationary random component can be calculated. The "smoothness" in this component is assumed to be the sum of squares of its second difference. The "smoothness" the growth component depends on a positive parameter,  $\lambda$ , which penalizes volatility in the growth component. The choice of an optimal  $\lambda$  depends on the variances of the random component and the second difference of the series. A value of 100 is commonly suggested for annual data. Because the Taiwanese macro aggregates are identified as I(2) processes, instead of trend stationary processes, the HP filter is used here. The smoothed series for the selected variables are shown in Figure A1 in the appendix.

then regressed on a polynomial function<sup>12</sup> of time, utilizing least square with serially adjusted residuals. The most appropriate degree of the polynomial is quadratic as determined by t-tests, AIC, and BIC statistics. The bias functions are reported in Table A1 in the Appendix.

Table 3 presents the number years into the future where model results are estimated to fall within the +/- 5% and +/- 10% error region of the real "smoothed" data. The numbers suggest that if the model is calibrated to 1988 data, the DGE model's forecast GDP for the next 1 and 3 years are within the 5% and 10% error region respectively. Those numbers are 9 and 11 years for industry output.

Table 2 Here

The time function method provides not only a measure of model performance, but also a procedure to adjust for the bias in forecasts. To illustrate, consider the case where the initial period values for 1988,  $\{p_m, p_a, p_s(0), k(0), \tilde{T}(0), L(0)\}$ , are up-dated to the most recently available data (e.g., 2005) and the model solved to obtain (4), (5) and (6) for 2006 to some future period. If the structure of the estimated bias for the years 1988-2003 remains unchanged, then the estimated functions,  $F_h(t)$  based on the previous solutions, can be used to evaluate and to adjust the resulting "out of sample" forecasts as well as to adjust for bias<sup>13</sup>.

### 4.3 Using the confidence intervals of time-series methods

This procedure confronts DGE model forecasts by the standard of forecasts obtained from its rival reduced-form time series methods. The procedure starts with the assumption that data series follow time series process. The historical data on key macroeconomic aggregates is modeled with different specification of ARIMA( $p, i, q$ ) process.<sup>14</sup> The best ARIMA process is chosen

---

<sup>12</sup>Although a polynomial function has multiple roots, it is obvious that only one of them is non-trivial in our case.

<sup>13</sup>An alternative procedure is to calibrate the model to year 2006 and solve it "backwards" to say 1970. Estimate  $F^h(t)$ , and use this structure to adjust model forecasts from 2006 onward.

<sup>14</sup>The data was also modeled with multivariate time series models such as VAR (vector autoregressive) and others. The best performance of the time series forecasts was given by univariate ARIMA models. Hence it is the method chosen and presented.

according to the various criterion available to evaluate the fit.<sup>15</sup> The confidence intervals given by best model are then used as the standard to assess the prediction accuracy of the DGE model forecasts. If predictions of a DGE model lie within the intervals obtained for time-series forecasts, this suggests these predictions are no worse than the best alternative available. Moreover, the DGE has a great advantage of a structural framework to explain the growth dynamics that underlie the forecasts.

The time series procedure is typical of ARIMA procedures. First, the transformed series (i.e., transformed by differencing, detrending, or other methods for stationarity) is fit to using an ARMA model.

$$\Phi(B)\tilde{Y}_{ht} = \Theta(B)w_{ht}$$

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q$$

where  $\tilde{Y}_{ht}$  is the transformed stationary series and  $w_{ht}$  is the white noise series for variable  $h$  over time,  $\Phi(B)$  and  $\Theta(B)$  are the autoregressive and moving-average linear lag operator<sup>16</sup>.

All the available historical data are utilized up to point in time for which the DGE is designed to replicate the economy, i.e., the point in time,  $T_c = t(0)$ . The best fit ARIMA model then produces forecast from  $T_c$  to correspond to the DGE results. The confidence intervals of the forecast from  $T_c$  to  $T$  is obtained and DGE modelers can use these confidence intervals to gauge their results. The DGE framework and results may be considered “good” if the DGE forecasts fall within confidence intervals of their ARIMA forecasts counterparts.

This approach is illustrated here by fitting the time series model to the annual data on GDP, industry output, service GDP, and household saving from 1961 to 1988 where, as above, 1988 is the initial year captured by the DGE model. Following standard practice in handling time series, first, each series was tested for stationarity. None of the series is found to be

---

<sup>15</sup>The criterion utilized are information criteria (AIC and BIC) and various tests on residual autocorrelation.

<sup>16</sup>The lag operator is:

$$\Phi(B)\tilde{Y}_{ht} = \tilde{Y}_{ht} - \phi_1 \tilde{Y}_{ht-1} - \phi_2 \tilde{Y}_{ht-2} - \dots - \phi_p \tilde{Y}_{ht-p}$$

stationary, as expected. They are then tested to determine whether they are unit-root or a trend-stationary processes. All the data series are identified as integrated processes with unit roots, which suggests differenced stationarity. After the proper degree of differencing, ARMA processes are fit to these series to model their behavior. The competing ARMA models are evaluated and selected with the aid of model selection information criteria and tests on residual autocorrelation.

The selected ARMA model is then used to forecast the respective series from the calibration year, 1988, to 2003. To form a comparable basis between the DGE and the ARMA, the ARMA forecasts of the 1988-2003 period are conducted without annual updating. The confidence intervals obtained for the ARMA's 1988 to 2003 forecasts are treated as the standard to which the DGE model results are compared. The ARMA model selected for the data series are reported in Table A2. Forecast comparison is provided with Figure 2A to 2D. Table 5 presents the number years into the future where model results are within these confidence intervals.

Figure 2 Here

As shown in the figures, DGE model results compare favorably, in most cases, in the sense that they are no worse than the forecasts of their time-series model rival. For industry output, both DGE and time-series models perform well. The underestimation of service GDP and hence economy GDP occurs in the case of both methods. The DGE results for values of service output lie well within the time-series forecast confidence intervals. As for agriculture output where the DGE performed most poorly, the time-series models also greatly overestimate observed values. The DGE forecasts for agriculture output lie within time-series forecast interval until the 9th year after calibration point. For household savings, the DGE forecasts are more accurate than the time-series forecasts. For the overall performance, though both methods underestimate GDP, the DGE results fall within the time-series confidence interval for two years into the future. These validation results provide some evidence in support of the DGE models capacity to forecast, and by implication, increased confidence in drawing upon the underlying structural reasons explaining why the economy evolved as it has.

Table3 Here

## 5 Concluding Remarks

The ability to address problems requiring an inter-temporal general equilibrium structure in which the microeconomic behavior of optimizing agents are featured is becoming ever more feasible. However, with some exception, the procedures to validate the results produced by these models seems to have lagged far behind their application. The lack of easily applied validation measures raises questions as to the reliability of inferences drawn, and the confidence policy makers can have in the analysis of policy alternatives, let alone the potentially rich dialogue they can potentially provide regarding the adjustment and growth process of real economies.

This paper advances three basic and relatively easy to implement methods. To illustrate the methods, A DGE model is fit to the Taiwan economy's for the year 1988, and solved to yield forecasts over the period 1988 to 2003. The methods suggested are then illustrated by using them to estimate the accuracy of the model's forecasts and establish boundaries for tolerable model errors. While not the main focus of the paper, the DGE model seems to forecast the evolution of the Taiwan economy surprisingly well, and the methods clearly show effectiveness in distinguishing model performances and suggest varying degrees of accuracy in predicted variables.

The methods discussed serve as a starting point for future validation efforts of applied DGE models. The methods advanced here can be used to help select those model parameterization and model specification that would provide a better fit to the data. It is also in our hope that, with relatively benign assumptions on probabilistic nature of the series, more robust statistic analysis can be carried out.

## References

- [1] Anderson, K. and W. Martin (2005),. "Agricultural Trade Reform and the Doha Development Agenda," *The World Economy*. 28(9):1301-1327.
- [2] Bouet, A, J. Bureau, Y. Decreux and S. Jean (2005), "Multilateral Agricultural Trade Liberalization: The Contrasting Fortunes of Developing Countries in the Doha Round," *The World Economy*.28(9):1329-1354.

- [3] Diao, X., Roe, T., and Yeldan E. (1998), "Fiscal Debt Management, Accumulation and Transitional Dynamics in a CGE model for Turkey." *Canadian Journal of Development Studies*, (2):.
- [4] Diao, X., A. Somwaru, T. Roe (2002), "A Global analysis of Agricultural Reform in WTO Member Countries." *Agricultural Policy Reform in the WTO: The Road Ahead*, U.S.D.A. Economic Research Service, Report No. 802. Washington DC:25-40.
- [5] Echevarria, Cristina (1997). "Changes in Sector Composition Associated with Economic Growth," *International Economic Review*, 38(2):431-452.
- [6] Elbasha, E. and T. Roe (1996), "On Endogenous Growth: The Implications of Environmental Externalities," *Journal of Environmental Economics and Management*, 31:240-268.
- [7] Gibson, M. J. (2006), "Productivity and the Dynamics of Trade Liberalization," Working Paper, Federal Reserve Bank of Minneapolis.
- [8] Hansen, L. P and J. J. Heckman (1996), "The Empirical Foundations of Calibration," *Journal of Economic Perspectives*, 10(1):87-104.
- [9] Kehoe, T. J. and E. C. Prescott (1995), "Introduction to the Symposium: The discipline of applied general equilibrium," *Economic Theory*, 6:1-11.
- [10] Kehoe, Timothy J., Clemete Polo, and Ferran Sancho (1995), "An Evaluation of the Performance of An Applied General Equilibrium Model of the Spanish Economy," *Economic Theory*, 6:115-141.
- [11] Kehoe, T. J. (2003), "An Evaluation of the Performance of Applied General Equilibrium models of the Impact of NAFTA," Federal Reserve Bank of Minneapolis, Research Department Staff Report 320.
- [12] Kydland, F., E., and E. C. Prescott (1982), "Time to Build and Aggregate Fluctuations," *Econometrica*, 50(6):1345-70.
- [13] Li, NaiChia and Terry Roe (2004), "Understanding Taiwanese Growth Experiences in A Ramsey Framework," 2004 Taipei International Conference on Growth and Development in Global Perspectives, Institute of Economics, Academia Sinica, Taipei, July 2004.

- [14] Roe, T, A. Somwaru and X. Diao (2003), "Do Direct Payments Have Intertemporal Effects on U.S. Agriculture?" in C. Moss and A. Schmitz, Editors, *Government Policy and Farmland Markets: The Maintenance of Farmer Wealth*, Iowa State press:115-140.
- [15] Lin I-Kuei L. (1989), "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 45(1):255-268.
- [16] Kim K. and A. R. Pagan (1995), "The Econometric Analysis of Calibrated Macroeconomic models," *Handbook of Applied Econometrics*, 1:356-390.
- [17] Watson, M. W. (1993), "Measures of Fit for Calibrated models," *Journal of Political Economy*, 101:1011-1041.
- [18] Whalley, J. (1986), "What Have We Learned from General Equilibrium Tax Policy Models," Working Paper 8625C, University of Western Ontario.
- [19] Whalley, J. (1988), "Lessons from General Equilibrium Models," In H. J. Arron, H. Galper, J. A. Pechman (ed.) *Uneasy Compromise: Problems of A Hybrid Income-Consumption Tax*, Washington: Brookings Institute:15-20.

## 6 Appendix: Formulas for Forecast Error Measures

### 6.1 Mean Absolute Percentage Error (MAPE)

For each data point, we calculated the absolute percentage error as  $\frac{|Y_{ht} - \hat{Y}_{ht}|}{Y_{ht}}$ , where  $\hat{Y}_{ht}$  is the predicted value for variable  $h$  at time  $t$ ,  $Y_{ht}$  is its corresponding observed value. For every series, the weighted mean absolute percentage error is computed using the following formula. Our modification of the usual MAPE is that the bias each year can be weighted differently. It may be desirable to put heavier weights on the years closer to the initial year since model's ability in predicting the immediate future is of greater relevance. Researchers are at liberty in assigning weights consistent with their research interests.

$$\sum_t w_t \frac{|Y_{ht} - \hat{Y}_{ht}|}{y_{ht}}$$

Since the model results include predictions over multiple variables, a weighted mean error is considered to include all series. The weights for each series can be determined to fit different research interests. In the sample application given later, the weights for series are in accordance to the variables' economic importance in GDP.

$$\sum_h w_h \left( \sum_t w_t \frac{|Y_{ht} - \hat{Y}_{ht}|}{y_{ht}} \right)$$

#### 6.1.1 Theil's U Distance

Theil's U Statistics is a distance measure of the prediction accuracy suggested by Theil (1961). The formula for Theil's U adopted is provided in the following equation. Similar to the MAPE, researchers are at liberty to weight each year and each series differently according to their research interests. The value of Theil's U statistic is clearly bounded from below by zero and with no upper bound. The smaller the statistic is, the more accurate

predictions are.

$$\sqrt{\sum_h w_h \left( \frac{\sum_t w_t (Y_{ht} - \hat{Y}_{ht})^2}{\sum_t w_t \cdot y_{ht}^2} \right)} \quad (9)$$

## 6.2 Other Forecast Error Measures

### 6.2.1 Symmetric MAPE

$$\sum_h w_h \left( \sum_t w_t \frac{|Y_{ht} - \hat{Y}_{ht}|}{\frac{Y_{ht} + \hat{Y}_{ht}}{2}} \right)$$

### 6.2.2 Root Mean Squared Error

$$\sqrt{\sum_h w_h \left( \sum_t w_t (Y_{ht} - \hat{Y}_{ht})^2 \right)}$$

### 6.2.3 Some Applications of Error Measures

Table A1 reports the MAPE and Theil's U statistics measuring the forecast accuracy.

Table A1 Here.

Table A2 reports the bootstrapped absolute percentage error for each series available for validation.

Table A2 Here.

## 6.3 Additional Tables and Graphs from the application

### 6.3.1 Smoothed vs. Observed Macro variables

Figure A1 Here

### **6.3.2 The Estimated Time Function of Bias**

The residual covariance matrix was adjusted with assumption of AR(1) process for the residuals.

Table A3 Here

### **6.3.3 Time-Series modeling**

Model selection was based on information criteria such as AIC and BIC subject to the passing of residual autocorrelation tests.

Table A4 Here

## **7 Tables and Graphs**

Figure 1 Model Forecasts and Observed Data, Taiwan, 1988-2003

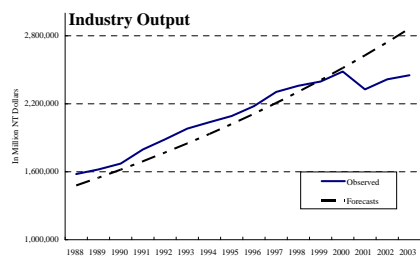


Figure 1A

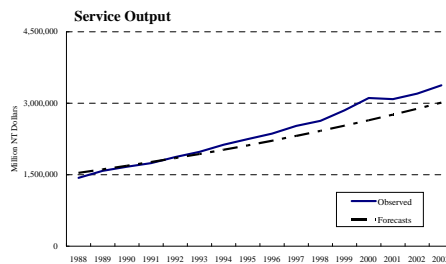


Figure 1B

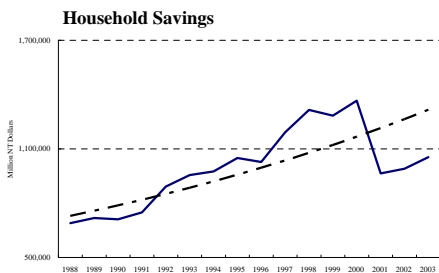


Figure 1C

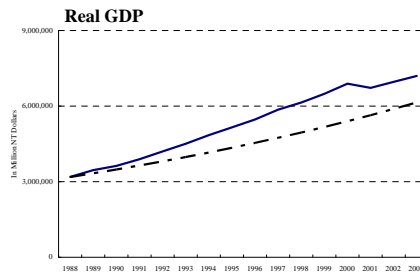


Figure 1D

<b>Bootstrapped Concordance Correlation Between Observed &amp; Predicted Variables</b>						
Taiwan 1988-2003						
	<b>Pearson's</b>	Lower 95%	Upper 95%	<b>Concordance</b>	Lower 95%	Upper 95%
	<b>Correlation</b>	Confidence	Confidence	<b>Correlation</b>	Confidence	Confidence
	<b>(mean)</b>	Interval	Interval	<b>(mean)</b>	Interval	Interval
<b>Industry Output</b>	<b>0.95</b>	0.95	0.99	<b>0.89</b>	0.78	0.96
<b>Agriculture Output</b>	<b>-0.59</b>	-0.85	-0.17	<b>-0.17</b>	-0.26	-0.05
<b>Service Output</b>	<b>1.00</b>	0.99	1.00	<b>0.91</b>	0.84	0.95
<b>Service Price</b>	<b>0.91</b>	0.84	0.97	<b>0.07</b>	0.10	0.04
<b>Saving</b>	<b>0.69</b>	0.41	0.98	<b>0.66</b>	0.41	0.82
<b>GDP</b>	<b>0.99</b>	0.98	1.00	<b>0.74</b>	0.60	0.84

Note: Statistics calculated with 10,000 bootstrap repetitions.

Table 1

<b>Number of Years the Results within the Chosen Bound</b>		
<b>(Taiwan 1988-2003)</b>		
Variables	Forecasts under	Forecasts under
	5% Error	10% Error
GDP	1 Year	3 Year
Industry Output	9 Year	11 Year
Agriculture Output	3 Year	4 Year
Service Output	2 Year	4 Year
Service Price	5 Year	8 Year
Household Saving	1 Year	2 Year

Note: Sample length is 16 years.

Table 2

Figure 2: Forecast Comparison, DGE Model and ARIMA forecasts, Taiwan, 1988-2003

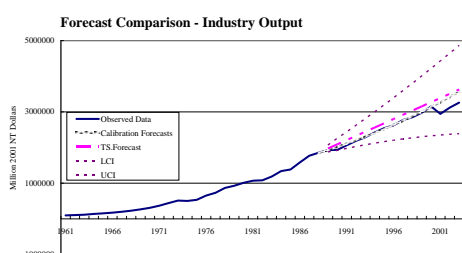


Figure 2A

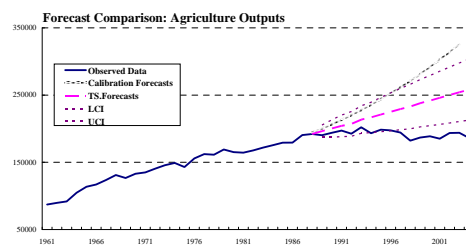


Figure 2B

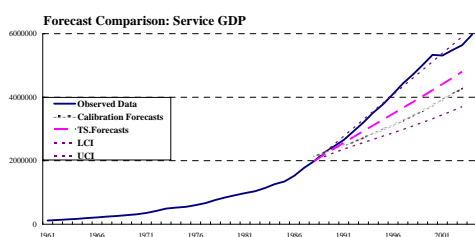


Figure 2C

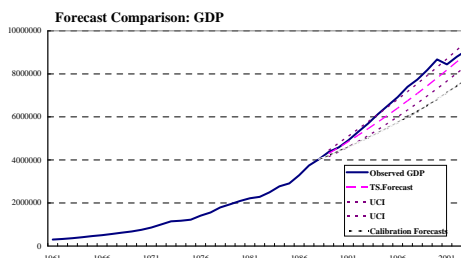


Figure 2D

**Number of Years Where Predictions Met the Time-Series Confidence Interval Standard (Taiwan: 1988-2003)**

Variables	Years
GDP	2 Years
Agriculture Output	8 Years
Industry Output	16 Years
Service GDP	16 Years
Household Saving	N/A

Note1: In the case of household saving, DGE model forecasts perform better time series forecasts.

Note2: Sample length is 16 years.

Table 3

MAPE and Average Theil's U Statistics Between Data and Forecasts (Taiwan 1988-2003)		
	Forecasts for All 16 Years	Forecasts for First 5 Years
MAPE for All Variables without Agriculture Output	5.9%	4.0%
Weighted MAPE for Sectoral GDP	10.7%	4.3%
Weighted MAPE for Sectoral GDP and Household Saving	11.1%	4.9%
Average Theil's U	0.21	0.06
Weighted Average of Theil's U	0.17	0.05

Note: To avoid double counting, all average values are calculated excluding GDP. Numbers reported here are not bootstrapped.

Table A1

Bootstrapped Percentage Errors Between Data and Forecasts (Taiwan 1988-2003)						
	GDP	Industry Output	Agriculture Output	Service Output	Service Price	Saving
Mean	-13.0%	-0.7%	32.6%	-5.2%	-6.9%	-1.4%
Conf. Interval	(-16.5%, -9.5%)	(-4.6%, 3.2%)	(14.5%, 50.7%)	(-8.4%, -1.9%)	(-10.7%, -3.0%)	(-6.6%, 9.5%)
Bootstrapped Mean	-13.0%	-0.7%	32.8%	-5.1%	-6.9%	-1.4%
Bootstrapped Conf. Interval	(-16.4%, -9.7%)	(-0.7%, 1.8%)	(16.1%, 48.7%)	(-8.1%, -2.3%)	(-10.3%, -3.5%)	(-6.1%, 8.9%)

Note: Bootstrapped statistics calculated with 5000 repetitions.

Table A2

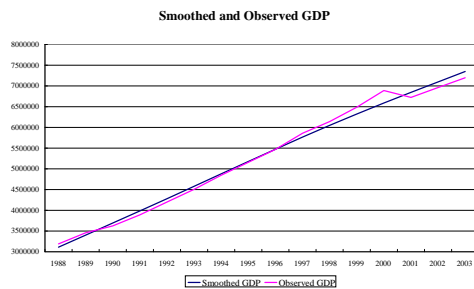


Figure A1a

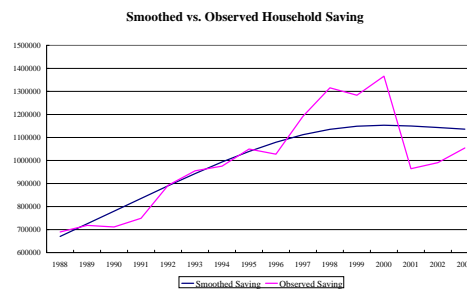


Figure A1b

<b>Estimated Models on the Bias (Percentage Errors)</b>			
Variables	AIC	BIC	Estimated Model Equation
Industry Output	-110.5	-108.9	$B(T) = .002731 - .00768 * T + .001445 * T^2$
Agriculture Output	-48.7	-47.2	$B(T) = .02859 + .001832 * T + .00427 * T^2$
Service Output	-91.8	-90.3	$B(T) = -.00713 - .02736 * T + .001059 * T^2$
Service Price	-105.3	-103.7	$B(T) = .001574 - .0081 * T - .0004 * T^2$
Saving	-86.9	-85.4	$B(T) = .005443 - .04612 * T + .003331 * T^2$
GDP	-83	-81.5	$B(T) = .00893 - .03622 * T + .001668 * T^2$

Note: Reported are the best OLS estimated time functions.

Table A3

<b>Time-Series Process</b>		
Variables	Process	Model
Industry Output	I(2)	ARMA(1,2)
Agriculture Output	I(1)	AR((6))
Service GDP	I(2)	ARMA(1,1)
Service Output	I(2)	MA(1)
Service Price	I(1)	ARMA(2,2)
Household Saving	I(1)	AR((3))
GDP	I(2)	MA(1)

Table A4