

Validating Dynamic General Equilibrium Model Forecasts

NaiChia Li*and Terry Roe†

May 2006

Abstract

The maintained hypotheses embodied in structural general equilibrium models calibrated to data have tended to make economists and policy makers insecure regarding their empirical foundation. Advances in dynamic general equilibrium (DGE) theory and its empirical application have exacerbated this insecurity since the forecasts provide by these models brings questions of validation to the forefront. Here, methods are developed to measure the magnitude of bias in DGE forecasts that are simple to implement. We adopted the concordance correlation measure, and introduced a time function method to assess the bias in DGE forecasts. A time-series confidence interval method is also introduced to formally judge the “good” forecasts from the “bad”. A calibrated DGE model is used to illustrate them. The time function method allows for the choosing of a functional form and an upper bound on forecast error. The time-series confidence interval method allows the DGE results to be evaluated by the standard of the rival time series models. If the DGE results are as good as time-series forecasts, the DGE model is a superior framework because of its advantage in providing not only “good” forecasts, but also insights into the economic structure generating the results. To illustrate these methods, we calibrate to Taiwanese data for the year 1988 a multi-sector Ramsey-based DGE model. The model is shown to forecast various dimension of the economy with surprising good, but varying accuracy. The proposed validation measures show effectiveness in distinguishing among diverse model parameter values and detecting model improvements. The measures are also statistically meaningful and require no arbitrary probabilistic assumptions on the distribution of either the results or the data.

*Ph. D student, Dept. Applied Economics, University of Minnesota, lixx0219@umn.edu

†Professor, Dept. of Applied Economics, University of Minnesota, troe@umn.edu

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 the need for validation is even more pressing as the state of the art has advanced from static computable general equilibrium to inter-temporal models.

This paper advances new methods to evaluate the results of calibrated models. Standards are suggested to statistically compare results from calibrated dynamic general equilibrium (DGE) models with data. We adopted the concordance correlation and introduced a time function method to evaluate the DGE model’s forecasts. A time-series confidence interval method is also introduced to formally judge the “good” forecasts from the bad. Through these methods, the performance of the model can be better evaluated and the long over due insight into the question of whether the results fit the data to a predetermined level of confidence is provided. The method and standards developed are simple to implement and require no arbitrary probabilistic assumptions on the distribution of both results and data.

This paper is organized by first providing an overview of the applied general equilibrium models, the type of insights they are hoped to provide, and why better validation methods would seem to be of growing importance. We next briefly describe a DGE model and possible causes of model bias. The methodology section discusses and lays out the validation methods introduced. We then use the DGE model results to illustrate the application of these methods.

2 Overview of Applied General Equilibrium Modeling

Applied general equilibrium (GE) analysis has been described by Kehoe and Prescott (1995, p.1) as “... the numerical implementation of general equilibrium models calibrated to data.” The conceptual framework is typically microeconomics based. The relatively large number of parameters to measure requires researchers to use a combination of approaches which is perhaps one of the more controversial aspects of this domain of query. Some parameters maybe obtained in the form of statistical estimates while other estimates rely on theory so that in an extreme case a single year’s observation is used to estimate a factor elasticity based on cost-share data. The model is often calibrate to data so that it mimics the world as closely as possible along a limited, but clearly specified, number of dimensions.

Two general types of analyses are commonly performed. One type involves a test of theory. For example, Gibson (2006) constructs a dynamic GE model of producers of heterogeneous efficiency and changes in producer-level investments

in trade relationships. A solution of the model for the case of trade reform results in an increase in total factor productivity (TFP) that, when compared to the data on Mexican TFP, provides an explanation for how trade reform increase factor productivity. Failure of a model to explain the data can also be viewed as a success in the sense that the maintained hypotheses are rejected, thus creating a paradox that may only be explained by further developments in theory. Implicit in this process is some form of validation.¹

Another common type of analysis 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 the forecasts of the model, but because the model is "structural" in the sense that it is based on macro-economic foundations, a clear explanation of the causes underlying the modeled economy's economic adjustments and corresponding magnitudes provide a basis to pursue a dialogue with policy makers². The confidence that policy makers, let alone other economists, place in the forecasts and the dialogue motivated by the structural nature of the model would seem to depend on how well the model replicates observed facts. However, with exceptions mentioned below, little effort has been made to validate GE models, and the 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 models are essential if policy makers are to have confidence in the results produced by these models. 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. They found that it performed well in capturing the changes that actually occurred. More recently, Kehoe (2003) evaluated the performance of several GE models used during the 1990s to predict the impact of NAFTA. He finds that these models drastically underestimated the impact of the agreement on North American trade, and they failed to capture much of the relative impacts on different sectors.

In the family of dynamic stochastic general equilibrium (or real business cycle (RBC)) models, evaluation efforts focus on moment comparison of the

¹In their article 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.

²However, all too frequently, emphasis is placed on model forecasts either because the author is uninformed of the model's economics, or model structure is too complicated to discern the importance of key Stolper Samuelson and Rybczynski like effects, or the experiment being performed is itself poorly designed.

data and results. A large body of studies focus on understanding the impacts of parameter uncertainty on model forecasts. Watson (1993), differing from the rest, established a R-Square like statistic to measure the goodness-of-fit of the model by augmenting the model results with a minimized random component for a perfect fit between the data and results. He concludes that model performance was adversely affected by parameter values.

The 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 models that provide inter-temporal forecasts. In this environment, it seems even more compelling to question how well the model fits the data. If the model is fit to a point on an economy's transition path using current data, then it can be used to forecast backwards, and validated by contrasting these forecasts with time series data; if fit to data at an earlier point in time, forward in-sample forecasts can be used for validation. 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 Brief Overview of DGE Framework (Ramsey 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. For more model detail and empirical findings for the case of Taiwan, see Li and Roe (2004). The economy is small and consists of 3 aggregated sectors: agriculture, industry and service sector, indexed as $i = m, a, s$, respectively. All firms in the economy employ two economy-wide factors, the services of 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 η . The 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, augment's 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 house-

hold members and discounted by the time preference rate ρ ,

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

where θ 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 π 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 λ_s is the share of total expenditure allocated to the service good.

3.1 Equilibrium

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

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

production 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 inter-temporal equations are specified that satisfy the definition of equilibrium,

computer code written, and numerically solutions obtained yielding values (4), (5) and (6) over an interval of time t^3 . With model's forecasts as a departure point, the key question is: how to assess the model's accuracy in predicting these values?

Clearly, general equilibrium models are, by their very nature, prone to specification and estimation biases. Moreover, the estimated parameters should not be treated as one unifying group since their effect on model results are dissimilar. The methods for estimating some parameters, such as productivity growth rates x and η , 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. 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 are non-stochastic and hence they do not take into account the effects of idiosyncratic shocks on an economy⁴. Instead, the "typical" model is designed to capture the long-term growth trend exhibited by the data, a trend that might be depicted by using time series filtering methods.

Thus, the practitioners of GE modeling must face a number of trade-offs, and recognize, as the Nobel Laureate Aumann (1985) noted, a model is a caricature or metaphor of an environment. Instead of asking whether it is right or wrong, the fundamental question is how useful is it? This usefulness, we maintain, depends in part on the model's capacity to replicate an economy's behavior.

4 Assessing DGE Results

If the probability structure of the observed data series, the structure of the model series, and probability structure of the bias (or error) between them are known, then hypothesis testing can be performed. This ideal obviously cannot be accomplished without the knowledge of the probability structures on parameters, results, and data. The challenge is to establish sensible standards of judging the performance of the model without imposing arbitrary probabilistic assumptions on results or data.

Before we begin the discussion of the methods for the comparison of model results to data, it should be recognized that macroeconomic series are stochastic, and they tend to feature non-stationary time-series behavior. The error struc-

³The 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.

⁴An exception of course is the real business cycle literature pioneered by Parente and Prescott ().

ture (or bias) measured as the difference between the model's forecasts (4)-(6), and the data are thus expected to also feature non-stationary time-series behavior. A further complicating factor is that the data generating processes (DGP) of the data and biases are all unknown. This unknown probability structure complicates the process of meaningful statistical inference and hypothesis testing. If the joint distribution of model results and data were known, the distribution of bias is implied and inference and hypothesis testing of whether the bias is significantly different from zero can be performed. The statistic validation of levels of model's predictions is then attainable. It is however not possible without assumptions on joint probability structure of results and data.

Many measures have been developed to assess the discrepancies between forecasts and data. These forecast error or distance measures include percentage errors and root mean squared error in various forms and Theil's U distance. The use of them for validating DGE models have been reported in Li & Roe (2004). The formulas and the application of them are provided again in the Appendix.

4.1 Measures of Association and Reproducibility

The most commonly used measure is the Pearson's correlation, which measures the linear association between two random variables. Lin (1989) argued that a limiting feature of this measure is the inability to detect the difference between the mean of the forecast and the mean of the data, i.e., deviation from a 45 degree line constructed on a plain with data on one axis and forecast on the other. A calibrated model may yield a trend close to the trend of the data, thus yielding a high Pearson's correlation coefficient, but the two series may differ greatly in their respective means. 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 σ_1^2 and σ_2^2 are the variances of variables 1 and 2, respectively, and σ_{12} is their covariance. Their respective means are given by μ_1 and μ_2 . Of course, in estimation, 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 2 compared series. Consequently, the concordance correlation coefficient is more powerful in detecting the deviation from a perfect fit of the 2 series, the 45-degree line. Like any other correlation coefficient, the concordance correlation coefficient is bounded between 0 and 1. If the coefficient is close to one, it lends supports to model's ability in replicating the past growth path.

Lin showed that for a bivariate normal series, $\hat{\rho}_c$ is a consistent estimator of ρ_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. This structure is not likely to be normal for either model forecasts or the data. The problem is addressed by applying a bootstrap technique to compute the statistic.

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, and Theil's U distance. 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 and their application is reported in Li & Roe (2004). The next section focuses on the more recently developed methods.

5 Methodology

Although forecast error and correlation measures are informative for model's forecasting performance, we now address the question: "How close is close Enough?"

5.1 Polynomial Time Function of Bias

This procedure provides a measure of how far beyond the last time period of data the model results fall within a predetermined error of tolerance. The procedure is to express the difference between model results and data as a loss function. Define bias as some 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 . The bias for each series, \mathbf{b}_{ht} , is express as a function 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 ε_{ht} is the non-systematic or error component. The function $F_h(t)$ can be viewed as the "best" fit to the norm, \mathbf{b}_{ht} .

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

In the example provided later, the percentage error is chosen as the norm $\|Y_{ht} - \hat{Y}_{ht}\|$ and $F_h(t)$ is assumed to be a polynomial function in time. The bias then is fitted as a polynomial function of time by adopting least square technique with serial correlation adjustments to residuals. In this example, two maximum tolerances for the bias are tried, 5% and 10%.

5.2 Time-Series Confidence Intervals Method

The idea of this procedure is to confront model results by the standard of rival forecasts using modern time series methods. The method aims to provide empirical evidences to assess DGE model performance with the assistance of time series techniques.

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. The best ARIMA process is chosen according to the various criterion available to evaluate the fit. Since forecasts from time-series modeling are considered as the best possible alternative available, their confidence intervals are used as the standard to assess the prediction accuracy of the DGE model results. If predictions of a DGE model lie within the intervals obtained for time-series forecasts, this suggests that predictions are no worse than the best alternative available, and moreover, the DGE has the advantage of a theoretic framework to explain the growth dynamics that underlie the forecasts.

The procedure is brief summary here. The actual application is presented in the application section of the paper. First, the transform series (i.e., transformed by differencing, detrending, or other methods) is fit to using an ARMA model.

$$\begin{aligned}\Phi(B)\tilde{Y}_{ht} &= \Theta(B)w_{ht} \\ \Phi(B) &= 1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p \\ \Theta(B) &= 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q\end{aligned}$$

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. The lag operator works as:

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

All the available historical data are utilized up until the calibration 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 results are considered “good” if the DGE forecasts fall within confidence intervals of their ARIMA forecasts counterparts. More details of the application of this method is presented later in the application section.

6 An Application Example of Taiwanese Economy

In this example, the DGE model described above is calibrated to Taiwanese economy to depict the 1988 point on the country’s transition path. Details are given in Li and Roe (2004). The model was solved numerically for the sequences (4), (5) and (6). Validation entails comparing these results to the observed variables from 1988 to 2003.

6.1 Results from the Taiwan model

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

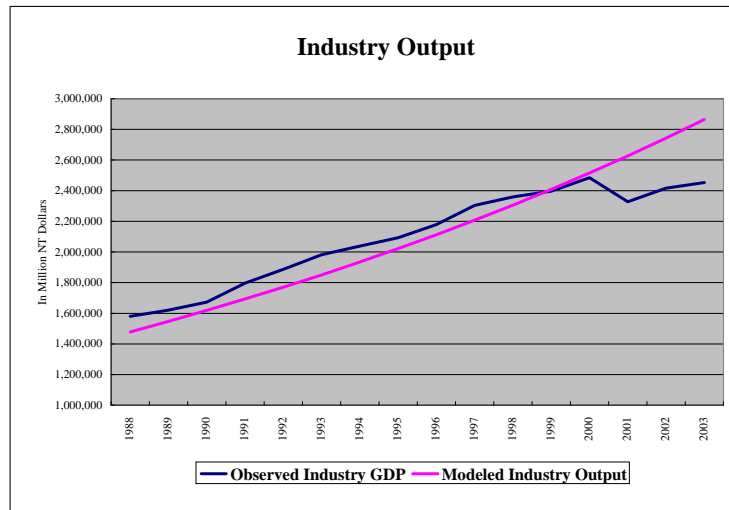


Figure 1A

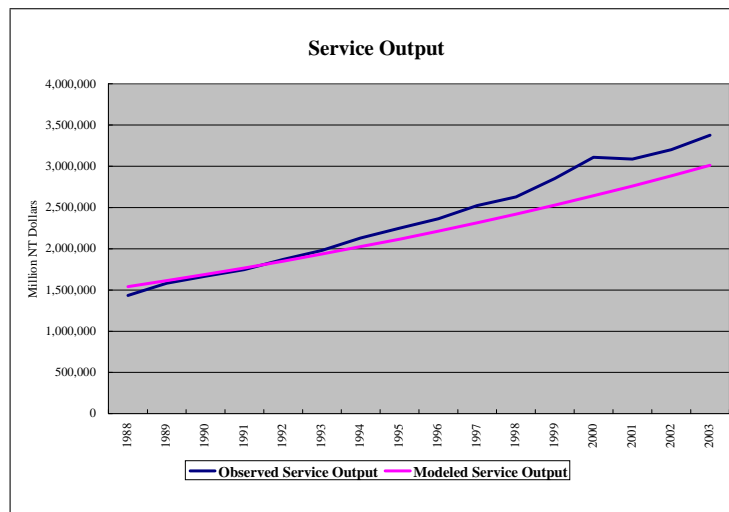


Figure 1B

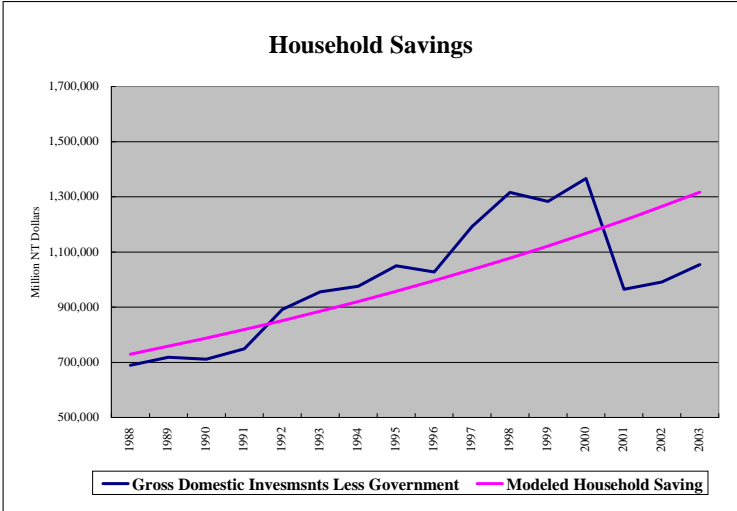


Figure 1C

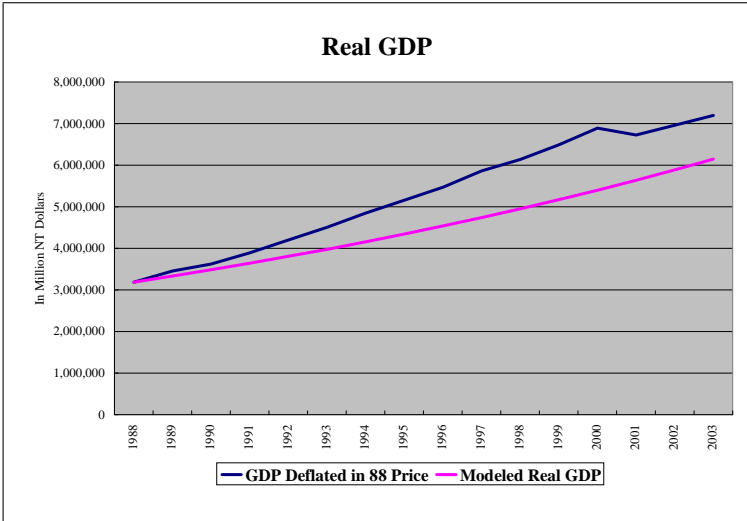


Figure 1D

From the graphs alone, a couple of noticeable observations are worth mentioning. First, the data exhibit a non-stationary bias overtime. As stated previously, non-stationarity can be caused by the time variant nature of some parameters which is not captured by the model. Also, forecasts tend to exhibit a persistent bias so that the difference between model forecasts and data grow with time. Second, the model's performance differs noticeably among various aspects of the economy. The model results clearly underestimate the overall

GDP through underestimated service output and price. The results on industry output however fit the data exceedingly well. This is the reason why a multivariate validation measure is desirable as stated previously. Again, this can be remedied in part since GDP can serve as a overall measure for model performance.

6.1.1 Descriptive Statistics on Model Bias

Except for "eyeballing" the results, descriptive statistics on the bias for each series may be a good place to start. The bias (norm) here is measured as percentage error $(\hat{Y}_{ht} - Y_{ht})/\hat{Y}_{ht}$. In Table 1, bootstrapped⁵ means and variances of the discrepancies, or the bias, between the results and data are reported. Without knowledge of the probabilistic nature of the series, statistics and inference relied on normality assumptions can be misleading. Bootstrap techniques can partially remedy the situation by acquiring sampling distribution of the statistics of interests and increase the reliability of statistic inference. Thus, in the example provided later in the paper, bootstrap technique was applied to all the statistics reported whenever statistic inference is sought.

Bootstrap Analysis for Percentage Errors						
	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%)

Table 1

The descriptive statistics of model bias further provide confirmation of the model's accuracy for predictions on various aspects of the economy. The model depicts the past growth path of industry output exceptionally well. The average percentage error over 16 prediction years was less than 1% and not statistically different from 0. The model's underestimation in service price and output is clear with average percentage error around 5% and 7% respectively. Consequently, the model's underestimation in total GDP is around 13%. Although

⁵Bootstrap technique is often resort to replace an unknown distribution with a know distribution in a probability calculation. In this study, we simulated a large number of bootstrap samples and compute the sample statistics for each bootstrapped sample. We then compute the nonparametric sample statistics from the bootstrapped samples.

the variation of model predictions on saving is larger, the mean percentage error is not statistically different from 0. The model's performance seems to be the least impressive in agriculture output with sizable overestimation⁶.

The concordance correlation coefficient, equation (7), and the Pearson's correlation coefficient are presented in Table 2. The results suggest that the concordance correlation can better detect reproducibility between two series than can Pearson's correlation. The underestimation in service price and GDP is much more accurately reported by the concordance correlation, where Pearson's correlation failed to detect differences in the mean of the respective series. These statistics once again confirm model's diverse performance on different variables. Concordance correlation coefficients suggest the model's better forecasts on industry and service output and underestimation on service price and GDP.

Bootstrapped Concordance Correlation Between Observed & Predicted Variables				
	Pearson's Correlation (mean)	Confidence Interval	Concordance Correlation (mean)	Confidence Interval
Industry Output	0.95	(0.95, 0.99)	0.89	(0.78, 0.96)
Agriculture Output	-0.59	(-0.85, -0.17)	-0.17	(-0.26, -0.05)
Service Output	1.00	(0.99, 1.00)	0.91	(0.84, 0.95)
Service Price	0.91	(0.84, 0.97)	0.07	(0.10, 0.04)
Saving	0.69	(0.41, 0.98)	0.66	(0.41, 0.82)
GDP	0.99	(0.98, 1.00)	0.74	(0.60, 0.84)

Table 2

The forecast error and distance measures and concordance correlation can be performed on "normalized" data as well. The "normalization" process⁷ 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 forecast removes the initial "scale" errors of calibration. The rationale is, in dynamic models, the data may allow for consistent intra-temporal calibration⁸, but may not be of sufficient quality to allow for consistent inter-temporal calibration of the model. Clearly, the normalization procedure yields better validation results. In the concordance case, the mean difference ($\bar{Y}_1 - \bar{Y}_2$) in (7) is reduced leading to a higher Concordance coefficient.

⁶These results suggest that we may have over estimated agriculture's TFP and, in the case of the service sector, under estimated its TFP and its labor's share in output.

⁷The 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 $t = 1988, \dots, 2003$.

⁸That is, 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.

We next turn to the question regarding acceptable errors levels that remains unanswered. It is subject to researchers to determine whether the forecast error and correlation measures are good or not. A 4% error can be excellent in the eyes of one researcher, but not so good in another. Is a .95 correlation good enough? Is a 3% error good enough? This is the reason why more objective standards, the time function and time-series confidence interval methods, are of great necessity.

6.2 Time Function of the Bias

6.2.1 Smoothing If Necessary

As previously mentioned, in contrast to RBC models, DGE models depict the long-term growth trend, not the short-term shocks that interpreted as deviations from trend. In this case, to compare forecast with data, the data may be "de-randomize" or filtered (, i.e., smoothed). The rationale is not to penalize the model for what it is not intended to capture. Validation with de-randomized or smoothed data however has the disadvantage of allowing the choice of filter and filter parameters in such a way the data can be altered or adulterated to be more consistent with model forecasts. With this negative caveat in mind, we de-randomized data in order to estimate the time functions $F_h(t)$ in (8). There are many parametric and non-parametric smoothing techniques that can be used for this purpose. Two parametric smoothing methods are summarized in the following section. In this study, we use the Hodrick-Prescott (HP) filter.

The HP filter was developed to estimate the unobservable non-stationary trend growth component of a series, from which the unobservable stationary random component of the series can be calculated. The "smoothness" in the growth component is assumed to be the sum of squares of its second difference. How smooth the growth component is depends on a positive parameter, λ , which penalizes volatility in the growth component. The choice of an optimal λ depends on the variances of the random component and the second difference of the series. A value of 100 for λ is commonly suggested for annual data. An alternative approach is to obtain a least squares estimate of the time trend of growth. The procedure is to regress data on a time trend, linear, quadratic, exponential, or in any other form, with serial correlation adjustment in the residuals. The fitted value from the least square model then is used as the de-randomized smoothed series for comparison results. Because most of the Taiwanese macro aggregates are identified as I(2) processes, instead of trend stationary processes, the HP filter is chosen for de-randomizing the data in this paper. Smoothed series on the selected variables are shown in Figure A1 in the appendix.

6.2.2 The Taiwan Example

Percentage errors between the smoothed data, as described above, and results were used as the measure of the model's forecast errors. These percentage

errors were regressed on a polynomial function of time utilizing least square with serially adjusted residuals. The most appropriate degree of the polynomial is quadratic as determined by t-tests, AIC, BIC and other relevant model selection criteria. The bias functions are reported in Table A1 in the Appendix.

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

Number of Years the Results within the Upper Bound		
	Predictions under	Predictions 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

Table 3

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 and the model solved to obtain (4), (5) and (6) for 2006 to some future period. If the structure of the estimated bias 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 bias⁹. For example, from Table A1, the percentage error for forecast for GDP in 2007, the first year, is to be -4.4% given the estimated time function. The forecast of GDP in 2007 then can be adjusted upwards to account for the expected underestimation.

6.3 Time Series Confidence Interval Method

In this section, we applied the time-series confidence interval methods mentioned above to the data on GDP, industry output, service GDP, and household saving from 1961 to 1988 are the data series to be modeled. Following standard practice in handling time series, first, each series was tested for stationarity. None of the

⁹An 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.

series is 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 standard t-tests, AIC, and BIC model selection criteria.

The selected ARMA model is then used to forecast the respective series from the calibration year, 1988, to 2003. 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.

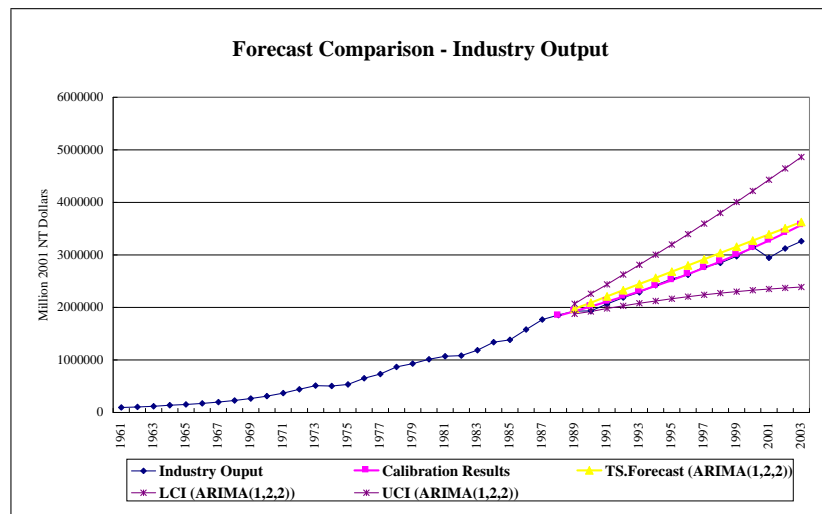


Figure 2A

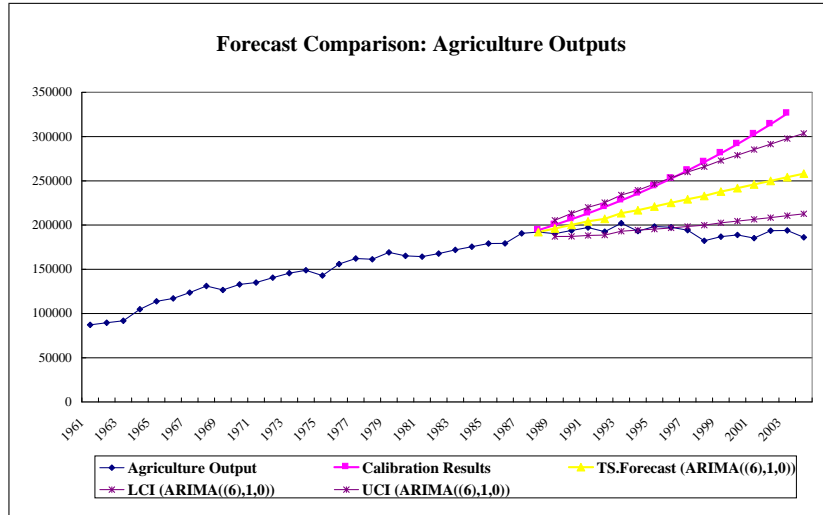


Figure 2B

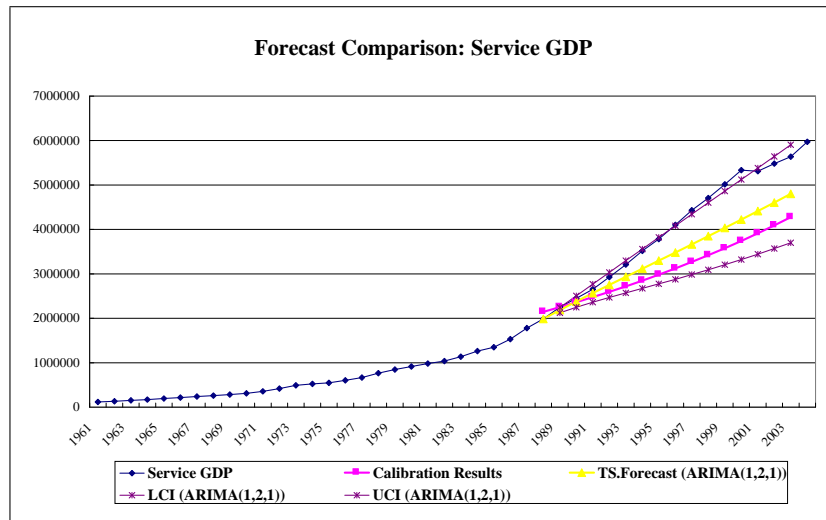


Figure 2C

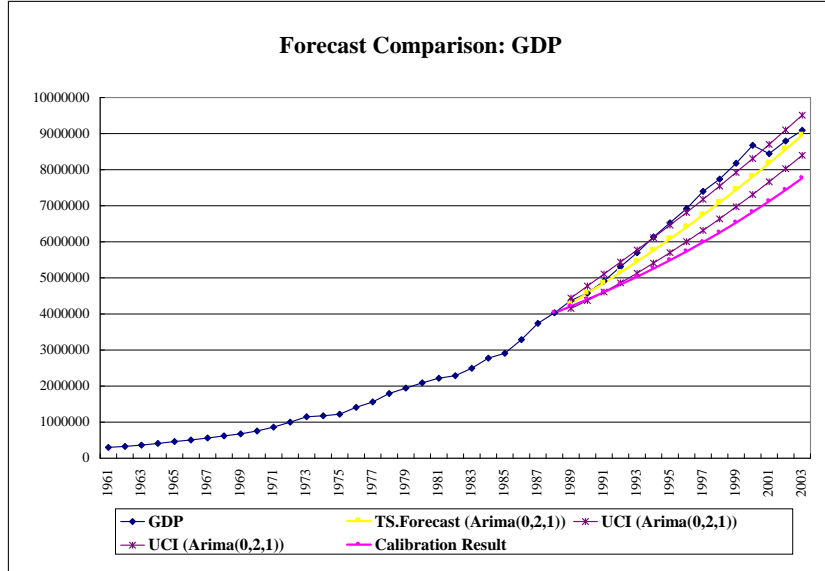


Figure 2D

As shown in the figures, DGE model results compare favorably, in most cases they are no worse than the forecasts of their time-series rivals. For industry output, both DGE and time-series models perform excellently. The underestimation in service GDP and hence in GDP is true with 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, only two years into the future, the DGE results fall within the time-series confidence interval. These results provide some evidence in support of the DGE models capacity to forecast, and by implication, explain the causes of the economy's evolution.

Number of Years Where Predictions Met the Time-Series Confidence Interval Standard	
Variables	Years
GDP	2 Years
Agriculture Output	8 Years
Industry Output	16 Years
Service GDP	16 Years
Household Saving	N/A

Table 4

7 Concluding Remarks

The felt need to address problems requiring a general equilibrium structure appears to grow as have the various methods for applying inter-temporal general equilibrium theory in which the microeconomic behavior of optimizing agents are featured. However, with some exception, the procedures to validate the results produced by these models seem to have lagged far behind their application. The lack of easily applied validation measures raises questions as to the reliability of inferences drawn from the models, and the confidence policy makers can have in them, let alone the potential rich dialogue of the adjustment process the structural model provides. This paper advances three basic and relatively easy to implement methods. A DGE model is fit to the Taiwan economy's growth path for the year 1988, and solved to yield forecasts over the period 1988 to 2003. The validation methods suggested are then illustrated by using them to validate or test 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.

This study is a starting point for future validation efforts on DGE forecasts. 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. We leave this task to another paper. It is 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] 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.
- [13] 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.
- [14] Lin I-Kuei L. (1989), "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 45(1):255-268.

- [15] Kim K. and A. R. Pagan (1995), "The Econometric Analysis of Calibrated Macroeconomic models," *Handbook of Applied Econometrics*, 1:356-390.
- [16] Watson, M. W. (1993), "Measures of Fit for Calibrated models," *Journal of Political Economy*, 101:1011-1041.
- [17] Whalley, J. (1986), "What Have We Learned from General Equilibrium Tax Policy Models," Working Paper 8625C, University of Western Ontario.
- [18] 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.

8 Appendix

8.1 Formulas for Forecast Error Measures

8.1.1 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.

$$\sum_t w_t \frac{|Y_{ht} - \hat{Y}_{ht}|}{y_{ht}} \quad (9)$$

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.

$$\sum_h w_h \left(\sum_t w_t \frac{|Y_{ht} - \hat{Y}_{ht}|}{y_{ht}} \right) \quad (10)$$

8.1.2 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.

$$\sqrt{\sum_h w_h \left(\frac{\sum_t w_t (Y_{ht} - \hat{Y}_{ht})^2}{\sum_t w_t \cdot y_{ht}^2} \right)} \quad (11)$$

Without the knowledge of the probabilistic nature of the series, it is unlikely that any statistic inference can be drawn from these measures.

Below, we adapt Bootstrap techniques as a partial to remedy to the problem of probabilistic structure.

8.1.3 Other Forecast Error Measures

Symmetric MAPE

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

Root Mean Squared Error

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

8.2 Application of Forecast Error Measures

In preparing for the numbers, all years are given the same weight and, when weighted, the series are weighted by their importance in GDP. The unweighted MAPE is simply the average with equal weights assigned to each series.

MAPE and Theil's U Statistics for Percentage Errors		
	All-Year Average	First-5-Year Average
MAPE for All Variables W/O Agriculture Output	5.9%	4.0%
Weighted MAPE for Sectoral GDP	10.7%	4.3%
Weighted MAPE for Sectoral GDP and HH Saving	11.1%	4.9%
Average Theil's U	0.208	0.062
Weighted Average of Theil's U	0.171	0.053
Note: All average values are calculated excluding GDP.		

Table A1

MAPE measures the average absolute percentage error across predictions over all years and series. In the numbers reported here, the calculation was done by giving each year the same weight and by weighting the series differently. The MAPE cross series including all years ranges from 6% to 11% depending on how they are weighted. If only the first 5 years of predictions are considered, the MAPE ranges from 4% to 5% again depending again on the weighting process. For the weighted MAPE, the bias of all years is double the size of the bias for the first 5 years after calibration year. Our expectation of more accurate predictions for years immediate after the initial year is confirmed. Statistics on Theil's U measures, less than .03, are very small considering the statistic is without an upper bound. Accuracy in earlier years than later years is also confirmed with the results from Theil's U statistics.

8.3 Additional Tables and Graphs from the application

8.3.1 Smoothed vs. Observed Macro variables

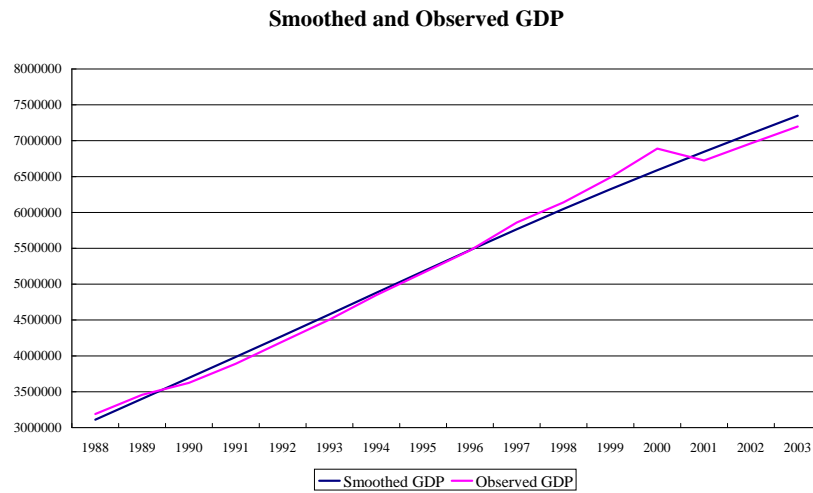


Figure A1a

Smoothed vs. Observed Household Saving

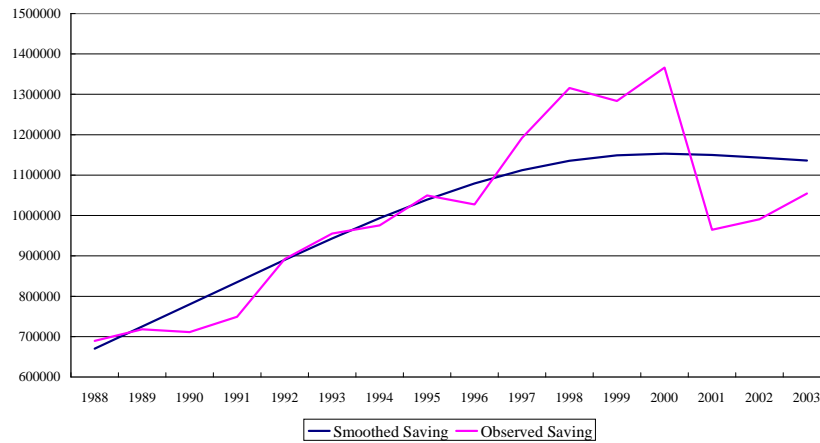


Figure A1b

8.3.2 The Estimated Time Function of Bias

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

Estimated Models on the Bias (Percentage Errors)			
	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	$B(T) = .00893 - .03622 * T + .001668 * T^2$

Table A2

8.3.3 Time-Series modeling

model selection was based on information criteria such as AIC and BIC.

Time-Series Process		
Variables	Process	Model
Industry Output	I(2)	ARMA(1,2)
Agriculture Outpu	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 A3