COMPARING EVALUATION METHODOLOGIES FOR STOCHASTIC DYNAMIC GENERAL EQUILIBRIUM MODELS

Eva Ortega

Banco de España - Servicio de Estudios
Documento de Trabajo nº 9820

COMPARING EVALUATION METHODOLOGIES FOR STOCHASTIC DYNAMIC GENERAL EQUILIBRIUM MODELS

Eva Ortega *

* This paper is based on Chapter 4 of my dissertation at the European University Institute in florence, Italy. I would like to thank Fabio Canova, Tryphon Kollintzas, Marta Regúlez and Mark Salmon for very helpful comments and suggestions. All remaining errors are my own.

Banco de España - Servicio de Estudios
Documento de Trabajo nº 9820

In publishing this series the Banco de España seeks to disseminate studies of interest that will help acquaint readers better with the Spanish economy.

The analyses, opinions and findings of these papers represent the views of their authors; they are not necessarily those of the Banco de España.

The Banco de España is disseminating some of its principal reports via INTERNET and INFOVÍA.

The respective WWW server addresses are:
http://www.bde.es and http://www.bde.inf.

ISSN: 0213-2710
ISBN: 84-7793-634-X
Depósito legal: M. 40131-1998
Imprenta del Banco de España

ABSTRACT

This paper "tests" the performance of the approaches of Watson (1993), DeJong, Ingram and Whiteman (1996), Canova and De Nicoló (1995) and Ortega (1998) for evaluating stochastic dynamic general equilibrium models using Monte Carlo techniques. It asks: Do the different model evaluation methodologies effectively improve an informal approach as in the typical calibration exercise? Are they only valid under limited assumptions, for evaluating the fit over a particular set of statistics or a particular model?

The Monte Carlo experiments evaluate the ability of each methodology to accept a model when it is equal to the actual DGP and to reject it when it is at odds with the actual DGP. In a sense, we are treating each methodology as a test for an economic model and compute its "size" and "power", respectively. We find that all four methodologies outperfrom the informal approach since they substantially reduce the risk of rejecting the true DGP and are able to discriminate more clearly between the DGP and models different to it.

1 Introduction

Stochastic dynamic general equilibrium (SDGE) models have become in recent years the central paradigm for the analysis of macroeconomic fluctuations. After the influential work of Brock and Mirman (1972), Lucas (1976), Sargent (1979), Lucas and Sargent (1981) and many others, economists started moving from using reduced form structural models to the use of stochastic dynamic general equilibrium models in order to formulate the economic questions of interest. As Gali (1995) puts it, though early applications of the latter to business cycle models (e.g., Kydland and Prescott (1982)) were generally restricted to model economies for which technology shocks were the only sources of fluctuations, and where built-in classical assumptions guaranteed the optimality of equilibrium allocations, the flexibility of the SDGE paradigm has been illustrated in a number of recent papers which have developed model economies characterized by the presence of non-classical features (e.g., imperfect competition as in Rotemberg and Woodford (1991), Gali (1994) or Ubide (1995)) and/or alternative sources of fluctuations (e.g. shocks to government spending as in Christiano and Eichenbaum (1992) or Baxter and King (1993)). These efforts to enrich the basic framework have been conducted with the objective of improving its empirical relevance and performance.

How to obtain and assess the quantitative implications of stochastic dynamic general equilibrium models gave rise to the development of the calibration methodology, which obtains quantitative predictions from a fully articulated, artificial model economy very often with the aim of comparing them to the particular set of observed stylized facts the model wanted to help explain. See Canova and Ortega (1996) for a detailed analysis of the calibration methodology. However, classical pieces in the calibration literature (e.g., Kydland and Prescott (1982) or (1991)) are typically silent on the metric one should use to evaluate the quality of the approximation of the model to the actual data. The approach favored by most calibrators is to glare over the exact definition of the metric used and informally assess the properties of the model by simple comparison of selected statistics of data simulated from the model to those of actual data.

Recent research in macroeconomics and time series econometrics has provided a number of alternative methodologies to formally evaluate the success of a model in replicating the stylized facts it wanted to explain, and to compare the success of alternative model specifications. Examples are, among many others, Gregory and Smith (1991), Watson (1993), DeJong, Ingram and Whiteman (1996), Canova (1994, 1995), Canova and De Nicolò (1995), Diebold, Ohanian and Berkowitz (1995) and Ortega (1998). Each of these model evaluation procedures has been proposed as an alternative to the common informal assessment of a SDGE model. However, it is a difficult task for the researcher to choose among these model evaluation methodologies.

Each of them summarizes the information given by the model in a different way, with a different set of statistics, and some of them base their assessment of the model on the distribution of different elements: DeJong, Ingram and Whiteman (1996) use the posterior distribution of the parameters as well as that of the actual data statistics, Canova (1994, 1995) uses the distribution of the shocks impinging on the economy and of model parameters, Canova and de Nicolò (1995) use also the distribution of actual data statistics, Ortega (1998) uses the distribution of the spectral density matrix estimator, ... This diversity makes it possible that alternative methodologies assess as very different the success of a model in reproducing the same stylized facts of the actual data. A comparison under uniform conditions of the performance of these new methodologies is called for.

This paper "tests" the performance of a selection of these methodologies using Monte Carlo techniques. It evaluates the ability of each methodology to accept a model when it is equal to the actual DGP and to reject it when it is at odds with the actual DGP. In a sense, we are treating each methodology as a test for stochastic dynamic general equilibrium macroeconomic models and compute its "size" and "power", respectively.

In the next section we describe the experimental design under which we will assess each methodology. We present the benchmark model we will take as the actual DGP in the Monte Carlo experiments as well as two alternative models. All of them are versions of the King, Plosser and Rebelo (1988) one-sector real business cycle model. Section 3 assesses both the benchmark and the two alternative models with respect to actual US data informally as it has been done in standard calibration exercises, i.e. based on a simple look at summary statistics from both actual and model data. We try to capture with a simple rule the overall measure of fit of a model under this "informal approach".

How do the different evaluation methodologies suggest to improve the informal evaluation of a model? And, more importantly, do they effectively lead to more accurate model evaluation than the informal approach? Are they only valid under limited assumptions, for evaluating a particular set of statistics or a particular model? Sections 4 to 7 answer these questions for four different methodologies: Watson (1993), DeJong, Ingram and Whitemann (1996), Canova and De Nicolò (1995) and Ortega (1998). For each of them we answer the first question explaining briefly what do they exactly consist on and illustrating them with the assessment of the benchmark and alternative models with respect to actual US data for 1964Q1–1995Q3. The fit of the three models is similar under all methodologies. We also find that results are sensitive to whether model parameters are allowed to vary or not.

The main contribution of this paper is the answer to the last two questions. For each of the four methodologies studied, as well as for the informal approach's rule, we conduct perform a Monte Carlo experiment which "tests" their performance as model evaluation methodologies

under uniform conditions.

We find that the rule we define for the informal approach risks of not accepting even the true DGP and, at best, is only able to reject those models that remarkably differ from the true DGP if the subjective degree of divergence between model and actual data the naive calibrator is willing to tolerate is not appropriately chosen. Watson's approach is found a reasonably accurate methodology. Using a rough measure (only exact for the spectral density distance approach of Ortega (1998)) of the "size" and "power" of each methodology, Watson (1993) appears less accurate than DeJong, Ingram and Whiteman (1996) and Canova and De Nicolò (1995). Among these two approaches, the latter achieves a better "size" at the cost of a lower "power", which is still enough to correctly rank the models according to their discrepancy with the true DGP. The spectral density distance approach (i.e. that of Ortega (1998)) is the one which obtains the smaller size and the larger power against models very different to the DGP, but shows no power against real business cycle models similar to the DGP because they generate very similar spectral density matrices to the DGP ones at business cycle frequencies. We find that all four methodologies outperform the informal approach since they substantially reduce the risk of rejecting the true DGP and are able to discriminate more clearly between the DGP and models different to it. Section 8 concludes.

2 The experiment

To compare different model evaluation methodologies under uniform conditions we design the following Monte Carlo experiment: at each replication, we simulate several times realizations for the shocks impinging on the SDGE model which we want to evaluate and we draw parameter vectors from their corresponding distribution if they are not fixed but stochastic, then compute the statistics of interest from the simulated series each time and assess the model by evaluating how close these statistics are to those generated with the actual DGP (which is known in the Monte Carlo experiment). Model statistics are simulated after generating DGP statistics so that the random numbers do not overlap.

When model parameters are stochastic, the uncertainty the researcher associates to the parameters is included in the statistics of interest as in DeJong, Ingram and Whiteman (1996), Canova (1994)-(1995) and Canova and de Nicolò (1995) at the cost of a Monte Carlo error since parameters are drawn randomly from their distribution. When model parameters are fixed but the statistics are estimated for model series derived from a realization of the exogenous stochastic disturbances (simulated from their corresponding distribution), we are introducing both a Monte Carlo error (in the simulation of the shock) and an estimation error with respect to using the theoretical statistics implied by the model. Standard calibration exercises typically

follow this approach and compute average model statistics across many simulations. The spectral density distance approach of Ortega (1998) also simulates times series from the model and estimates their statistics. In all cases, errors are reduced if the fit is assessed averaging somehow across many simulations of the model. Watson (1993) uses the theoretical values of model statistics with fixed parameters, in which case no repeated simulation of the model is required.

When illustrating how each methodology works using actual US data, we will simulate 1000 times the model economy (except for Watson's approach). But when coming to our Monte Carlo experiment for comparing evaluation methodologies, the cost of using such a large number of simulations to obtain a measure of fit of the model per replication is too high[1], so we opt for running 100 simulations per replication. We replicate 100 times the experiment just described. Considering that at each replication we are computing 100 times the simulated statistics, 100 replications is not such a small number and increasing it would be unfeasible in terms of computer time for some of the methodologies compared.

When the simulated model and the actual DGP are the same any model evaluation methodology should always accept the null hypothesis H_0: model = DGP, against H_1: model $\neq$ DGP, except for a predetermined arbitrary α% of the times. The rejection frequency of H_0 across Monte Carlo replications is the empirical "size" of the methodology. When the simulated model differs from the actual DGP then the methodology should reject such H_0 in favor of H_1 in (100-α)% of the replications. The actual percentage rejection of H_0 now is the empirical "power" of the methodology. We would also like the methodologies to reject *more* those models which generate statistics that lay further away from those generated by the actual DGP, i.e. to provide an indication of how far a model is from the actual DGP with respect to alternative models. This point is particularly important for a researcher interested in discriminating between models when none of them is likely to be exactly the actual DGP, which is very often the case for SDGE models.

The first problem we have to solve is the selection of the actual DGP for the Monte Carlo experiments. We want a model which, on the one hand, highlights the usefulness of all the methodologies studied and, on the other hand, allows to derive easily alternative models that differ from it in different degrees so that we can measure how each methodology captures the degree of divergence between each alternative model and the actual DGP. Some of the methodologies we compare (Watson (1993), DeJong, Ingram and Whiteman (1996) and Canova (1994)-(1995)) are specific for assessing calibrated models, Canova (1994)-(1995), Canova and De Nicolò (1995), DeJong, Ingram and Whiteman (1996) and Ortega (1998) apply to simulated

[1]It takes a Pentium around 8 hours to asseas the fit of each of the three models we evaluate using the DeJong, Ingram and Whiteman (1996) methodology, not much less using Canova and De Nicolò (1995) methodology, and around 12 hours using the spectral density distance approach.

SDGE models and Ortega (1998) is constructed for assessing multivariate dynamic models whose solution has to be approximated. Hence, we select a calibrated SDGE model whose solution is not exact but approximated and which can generate simulated statistics of interest sufficiently different from each other when generated under alternative versions of the model.

A very well known prototype of such model is the basic real business cycle model explained in detail in King, Plosser and Rebelo (KPR) (1988): the one-sector neoclassical model of capital accumulation where work effort is a choice variable and economic fluctuations are initiated by impulses to technology. Such model has the further advantage that some of the methodologies we compare in this paper have been applied to versions of this model so we can have in some cases direct means of comparison in the literature.

Next we have to select with respect to which particular features of the data we want to assess the fit of the model. We choose to focus on few multivariate statistics: the relative standard deviation between per capita consumption (C) and output (Y) and the contemporaneous correlations between C and Y, hours (H) and output, and hours and average labor productivity (AP). When the model evaluation methodology is performed in the frequency domain the standard deviation is replaced by the power spectrum and the correlation by the coherence at selected frequencies. The reason for selecting these multivariate relationships is that they include statistics which are typically successfully captured by the different versions of the model (corr(Y,C)), some others which typically are not (corr(H,AP) is typically too high and std(C)/std(Y) too low in the model with respect to the actual data) and some which vary a lot across model specifications (corr(Y,H)).

The rest of this section presents briefly the model used as the actual DGP in the Monte Carlo experiments (Model 1) and its statistical implications and two versions of this model (Model 2 and Model 3) which will be used to check the power of the various evaluation procedures.

2.1 The models

Model 1 has technology shocks as the only source of economic fluctuations, is the most commonly studied one-country real business cycle model and has been used in the literature for illustrating some of the model evaluation methodologies we are comparing in this paper (Watson (1993) and DeJong, Ingram and Whiteman (1996)). Model 2 includes government spending shocks. The addition of this further shock alters the dynamics while keeping similarities with the benchmark model. Model 3 allows for government spending shocks only and generates very different model dynamics. In what follows, we first present the model structure that encompasses Model 1, 2 and 3 and its solution and then we specify the parameterization for each model as well as their implications for our multivariate statistics.

The economy is populated by a large number of identical infinitely-lived agents. All variables are expressed in per capita terms. Preferences of the representative agent are given by:

$$U \equiv E_0 \sum_{t=0}^{\infty} \frac{\beta^t}{1-\sigma} C_t^{1-\sigma} v(L_t) \tag{1}$$

where C_t is private consumption of the single good by the representative agent and L_t is leisure, β is the discount factor and σ the coefficient of relative risk aversion. Leisure choices are constrained by:

$$0 \leq L_t + H_t \leq 1 \tag{2}$$

where the total endowment of time is normalized to 1 and H_t represents the proportion of time devoted to market activities.

The single final good is produced with a Cobb-Douglas technology with constant returns to scale:

$$Y_t = A_t (K_t)^{1-\alpha} (X_t H_t)^{\alpha} \tag{3}$$

where K_t is the capital input, α is the share of labor in GDP, and X_t is labor-augmenting Harrod-neutral technological progress with deterministic growth rate equal to θ_x, i.e. $X_t = \theta_x X_{t-1}$ with $\theta_x \geq 1$. X_t represents permanent technical change while temporary changes in technology are represented by variation in total factor productivity according to

$$lnA_t = \rho_A lnA_{t-1} + \epsilon_{At}$$

where $\epsilon_{At} \sim N(0, \sigma_A^2)$.

Capital goods are accumulated according to:

$$K_{t+1} = (1 - \delta_K) K_t + I_t \tag{4}$$

There is an output tax whose revenues are used to finance an exogenous path of per capita government expenditures G_t and lump sum transfers TR_t. These expenditures are assumed not to affect the economy's production possibilities nor the representative agent's marginal utility. The government budget constraint is

$$G_t + TR_t = \tau Y_t \tag{5}$$

where G_t follows the stochastic process:

$$lnG_t = \rho_G lnG_{t-1} + \epsilon_{Gt}$$

where $\epsilon_{Gt} \sim N(0, \sigma_G^2)$. Innovations to total factor productivity, ϵ_{At}, and to government spending, ϵ_{Gt}, are assumed to be independently distributed.

The economy-wide resource constraint is given by:

$$Y_t - G_t - C_t - I_t \geq 0 \tag{6}$$

All variables except hours and leisure are assumed to grow in the steady state at the same rate as the technological progress, θ_x-1, so that business cycle dynamics are separated from growth by associating the latter to that deterministic trend common to all drifting variables. Technology, preferences and government behavior are restricted following King, Plosser and Rebelo (1990) so that the suboptimal (because of distorting taxes) competitive equilibrium solution is compatible with steady state growth. The equilibrium solution is characterized by the optimality conditions for the individual's maximization problem together with the government constraint.

To characterize the local dynamics around the steady state path, i.e. what happens to the economy when it faces alternative sequences of exogenous shocks, we follow KPR (1990) and express the transformed optimality conditions in terms of detrended variables: we take ratios of the original per capita drifting variables with respect to the labor augmenting technological progress so that the economy is transformed from steady state growth to stationarity. The modified optimality conditions are then approximated with a log-linear expansion around the steady state.

Time series for consumption (C), output (Y), hours (H) and average labor productivity (AP) are generated from the approximate optimality conditions, once the free parameters and time series for the innovations to exogenous processes of the model are given. The statistics of interest (standard deviations and correlations or spectra and coherences) of simulated data are computed after extracting from the raw simulated time series a linear trend, in the same fashion as actual data statistics.

Table 1 shows the parameter values for each of the three model specifications we consider. They only differ in the parameterization of the exogenous processes so that Model 1 (the actual DGP in the Monte Carlo experiments) has only technology shocks, Model 2 has both technology and government spending shocks and Model 3 has only shocks to the exogenous government spending process.

Parameter values are taken not only from KPR (1988) but also from literature related. We have estimated θ_x as KPR (1988) do, one plus the average quarterly rate of growth of real per capita output, but with an updated data set (1964Q1–1995Q3 instead of 1948Q1–1986Q4). α, δ_K and β are taken from KPR (1988). σ=2 is the standard value calibrated models use for multiplicatively separable momentary utility. We impose government budget balance in the steady state by assuming a constant tax rate (τ) equal to a constant government spending output share (sg) and zero transfers. We have taken a value for τ and sg which lays in between

the one suggested by KPR (1988) of 30% and that used by Baxter and King (1993) of 20% for the case of steady state balanced budget (Aiyagari, Christiano and Eichenbaum (1992) suggest a government spending share of 17.7%). ρ_A and σ_A are the standard values used in the literature for the persistence of technology disturbances and the standard deviation of technology innovations, respectively. The persistence of government spending process (ρ_G) and the standard deviation of its innovations (σ_G) are from Aiyagari, Christiano and Eichenbaum (1992).

The first three columns of Table 2 show the statistics of interest for each of the three models (see Stadler (1994) for a good summary of the basic implications of real business cycle models).

Positive temporary technology shocks increase output and hence consumption, generating positive and large consumption-output contemporaneous correlation in Models 1 and 2. Consumption increases to a lesser extent since agents seek to smooth it over time (so that the relative standard deviation of consumption with respect to output is lesser than one in all three models) and this increases the capital stock. Temporary productivity shocks shift the production function and hence the labor demand curve. The marginal product of labor is also increased, but since the utility function is specified so that the income and substitution effects of a real wage change cancel each other, the labor supply curve does not shift. Then, under Model 1 and 2 the shift in labor demand increases real wages[2] and the hours worked. Because of the intertemporal substitution of leisure these models generate high positive hours-output contemporaneous correlation and significantly positive hours-average product of labor one (this last observation is referred to in the literature as the "productivity puzzle").

Government spending does not enter directly the agent's utility function (nor the production function) and hence shocks to government expenditure do not have a substitution effect but only a wealth effect. That is the reason why when we added them to technology shocks as the sources of business cycle fluctuations (Model 2) the statistics displayed in Table 2 do not change that much (correlations are slightly reduced). Things change, though, when government spending shocks are the only source of fluctuations (Model 3). The rise in government spending financed by taxes results in a negative wealth effect that shifts the labor supply curve while the labor demand one remains unchanged. This produces an increase in hours worked (contemporaneous correlation with output of 1) and a decrease in real wages (contemporaneous hours-average labor productivity of -1). Government consumption crowds out consumption through this negative wealth effect, resulting in a contemporaneous consumption-output correlation of -1.

[2]The Cobb-Douglas specification of the production function implies that the marginal product of labor (real wage in competitive equilibrium) will move quite closely with the average product of labor, or productivity. Therefore, the increase in real wages is translated into an increase in AP.

3 An informal evaluation

Table 2 shows also the statistics of the actual data. Data is from OECD Quarterly National Accounts, in constant 1985 US$Bln, and from National Government OECD Series (Department of Labor) in thousands of people, all seasonally adjusted. The sample period is 1964Q1–1995Q3. Y is GNP, C is personal final consumption expenditure, H is total civilian employment times average weekly hours of all private workers on nonfarm payrolls. Variables are transformed into per capita terms dividing them by civilian noninstitutional population of 16 and more years old excluding armed forces (source: Department of Labor, National Government OECD Series). AP is Y/H. To maintain a close relationship between the model and the actual data we linearly detrend the logs of all the series but for H (and AP is detrended by substracting log(H) from the detrended log(Y)) before computing the four statistics we are interested in. The statistics differ slightly from those reported in other works for two reasons: because we have used an updated data set and also because the detrending method chosen has an impact on these statistics (see Canova (1997)).

Informal evaluation of how the three models reproduce the relationships between output, consumption, hours and average productivity observed for US data would consist on casual inspection of columns 1, 2 or 3 of Table 2 compared to the last column. The conclusion would be that Model 1 and 2 are similarly successful in reproducing the observed facts although they predict too little consumption variability and too much hours-productivity contemporaneous correlation. Instead, Model 3 would be rejected as a good explanation of the observed facts since it totally misses the positive high consumption-output correlation and predicts an hours-productivity correlation of -1.

Very often SGDE models are judged successful or rejected according to similar informal criteria. An informal evaluator would consider the model more successful than others the larger the number of model statistics which are similar to the actual data ones and the smaller the divergence between actual and simulated statistics. To put this "formally" we arbitrarily choose the following rule: reject a model if at least 3 out of the 4 statistics we are interested in explaining differ in absolute value from the observed ones by more than x%. We perform the Monte Carlo experiment outlined in Section 2 on this rule to evaluate how reliable such an informal criterium is to accept or reject a model. We compute the rejection frequencies of the null hypothesis H_0: Model i = DGP, against H_1: Model $i \neq$ DGP, for i=1, 2 and 3. At each Monte Carlo replication, we simulate 100 times time series of the usual length (127 observations as we had for actual US data) from a model and use the informal evaluation's rule to compare the average (across simulations) of the 4 statistics we are interested in to the DGP statistics. Since Model 1 is taken as the DGP, DGP statistics are the theoretical statistics of

Model 1 (calculated using simulated series of 10,000 observations), which are kept fixed across Monte Carlo replications and across experiments. If the null hypothesis H_0: Model 1 = DGP is rejected 0% of the times and the rejection frequencies of H_0: Model 2 = DGP and of H_0: Model 3 = DGP are 100%, the informal approach would be a perfect model evaluation methodology (0% *size* and 100% *power*) since it would always be able to recognise which are the correct and incorrect models. Obviously, these rejection frequencies will depend on the x% the informal evaluator is willing to consider as a significant divergence between actual and model statistics.

Over 100 replications, we find that Model 1 is rejected 0% of the times when the rule is: reject if the divergence exceeds 50% of the absolute value of the actual data statistic for 3 or more out of 4 statistics. Such a "permissive" rule leads to reject H_0: Model 2 = DGP also 0% of the times while rejects H_0: Model 3 = DGP 100% of the times. Being so permissive, the informal approach will always consider as equal to the DGP models which are not equal but similar to the DGP (such as Model 2). However, reducing the accepted divergence to 10% of the value of the actual data statistics, the rule becomes "too strict", in the sense that although it succeeds to reject models different to the DGP (Model 2 and Model 3) 100% of the times, it also rejects the true model (Model 1) 100% of the times. This is because with short time series and using the average statistics (across simulations) induces a large enough error which makes some simulated statistics differ in more than 10% from the true ones [3].

Summarizing, the rule we have defined to capture the typical informal evaluation approach risks of not accepting even the true DGP and, at best, is only able to reject those models that

[3]The following table shows the average across Monte Carlo replications of the divergence between model simulated statistics (averages across 100 simulations of the model) and actual DGP statistics (theoretical statistics of Model 1), measured in percentage of the values of actual DGP statistics. We are actually computing for each statistic the x% divergence which, on average, should be allowed by the informal approach rule in order to accept the H_0: Model i = DGP. The results of the first column indicate that the divergence between the values of the theoretical statistics implied by Model 1 (actual DGP) and those computed averaging statistics from short series generated also from Model 1 is very large.

Statistic	Model 1	Model 2	Model 3
std(C)/std(Y)	28%	21%	14%
corr(C,Y)	4%	5%	211%
corr(H,Y)	57%	56%	92%
corr(H,AP)	423%	398%	135%

An informal evaluator with a rule not accepting less than 28% divergence between model and actual data statistics for 3 or more out of 4 cases would not be able to accept the true DGP (i.e. the decision rule would have a huge "size"). Whereas one would need to accept a percentage divergence as high as 92% for 3 out of 4 statistics to accept models having very different equilibrium dynamics than the true DGP such as Model 3. It is important to note, however, that these percentages indicate also a high variability of the correlations involving labor series, especially corr(H,AP), and that another arbitrary model evaluation rule which takes into account the variance of the statistics would have a more acceptable performance.

remarkably differ from the true DGP if the subjective degree of divergence between model and actual data the informal evaluator is willing to tolerate is not appropriately chosen. The results of this experiment strongly advices SDGE modellers not to rely on averaging, across several simulations of the model, the values of the statistics of short simulated time series, as it is often found in the literature of calibrated models.

How do more formal evaluation methodologies suggest to improve the informal evaluation? And, more importantly, do they effectively lead to more accurate model evaluation than the informal approach? Are they only valid under limited assumptions, for evaluating a particular set of statistics or a particular model? The following sections answer these questions in three steps. First, we describe briefly each methodology. Second, we illustrate how they work by evaluating our three models with respect to actual US data. Finally, we check their performance as model evaluation methodologies with the Monte Carlo experiment described, in the same fashion as we have done to the informal approach.

4 Watson's measures of fit

Watson (1993) suggests a way to evaluate calibrated models by asking how much error should be added to a model generated series, $x_t^* = g(z_t, \gamma)$ (where γ are the parameters of the model and z_t are exogenous stochastic disturbances) so that its spectral density equals the spectral density of the corresponding actual data series y_t. The error $u_t^* = y_t - x_t^*$ includes both model error ($u_t = y_t - x_t$) and the error of approximating with x_t^* the exact model solution $x_t = f(z_t, \gamma)$ since it is most of the times not obtainable analytically. The choice of the spectral density function of y_t as the set of stylized facts of the data to be matched by the model has clear advantage over selecting relative standard deviations and correlations when we are interested in evaluating the business cycle properties of a model, because we can focus easily on only those frequencies associated with business cycle fluctuations.

Watson provides an R^2-type measure of fit between the model and the data based on the ratio of the spectral density of the error $A_{u^*}(\omega)$ to that of the actual data $A_y(\omega)$ for a particular frequency ω or for a frequency range $[\omega_1, \omega_2]$. The size of the ratio is evaluated informally (i.e. whether it is greater than one, between zero and one or close to zero). This ratio is a lower bound: when it is large the model poorly fits the data but when it is small it does not necessarily follow that the model is appropriate. Note that in this approach, γ and z_t are fixed, and A_{x^*} and A_y are assumed to be measured without error.

Watson's measures of fit (for a single frequency or for a frequency range) are univariate but can be easily extended to a multivariate evaluation of a model in the same fashion as in Canova and Ortega (1996), so that we can evaluate how well our three models reproduce the

multivariate relationships between Y, C, H and AP observed in US data for 1964Q1-1995Q3. Table 3 reports the results of such evaluation. All statistics reported in Table 3 are averages across business cycle frequencies, i.e. those associated to cycles 2 to 8 years long.

We have estimated the spectral density matrix for the linearly detrended 4-variable actual data set as well as that implied by each model[4]. Spectral density matrices are estimated using a Bartlett window (see Priestley (1981), ch.9.5) with a sample size-dependent bandwidth parameter $M = 1 + 3 \times T^{1/3}$ so that we capture the optimal rate of convergence of the Bartlett window of $O(T^{1/3})$ (see Andrews (1991)). Figures 1 and 2 display spectra and coherences for actual US data and for the model series generated under Model 1, 2 and 3, and for all frequencies.

We compute Watson's measure of fit for each of our four series

$$R_j = \frac{\int_{\omega \in Z} A_{u\bullet}(\omega)_{jj} d\omega}{\int_{\omega \in Z} A_y(\omega)_{jj} d\omega}, \quad j = 1, 2, 3, 4$$

where $A_y(\omega)_{jj}$ and $A_{u\bullet}(\omega)_{jj}$ are the actual data and the error spectral densities, respectively, for series j and where the ω frequencies included in the Z interval are business cycle frequencies. R_j is calculated under two different identification schemes: one which minimizes the trace of $A_{u\bullet}$ with equal weight to its 4 components, and a second one which minimizes the spectral density of the error associated to output (when there is only one source of fluctuations: technology shocks in Model 1 or government expenditure shocks in Model 3) or to output and hours (when there are two types of shocks in the model, i.e. Model 2). The measures of fit for the consumption-output, hours-output and hours-average product of labor coherences (i.e. the frequency domain equivalents to corr(C,Y), corr(H,Y) and corr(H,AP), respectively) are calculated as the ratio between the average coherence of model series across BC frequencies and that of actual data coherence, since it is hard to interpret what the coherence between approximation errors means. Instead, our measure is expected to be closer to 1 when the observed coherence is explained by the model. Note that, by construction, such a statistics it is not affected by the identification scheme chosen.

Table 3 shows, consistently with the informal inspection of the relative standard deviation and correlations of Table 2, that the fit of Models 1 and 2 is very similar and much more satisfactory than that of Model 3 (only government spending shocks). It changes across identification

[4]Instead of deriving the theoretical spectral density of each model as in Watson (1993), we have estimated it for the 4 variables simulated using the parameter values of Table 1 and simulating the model only once. We have simulated time series 1000 observations long so that their estimated model spectra are sufficiently close to their theoretical values. Sensitivity analysis has been performed on the effect of the model series length, i.e. all the statistics in Table 3 have been computed for the case in which model spectral density is estimated from series 500, 200 and 100 observations long. The main result is that the measures of fit statistics increase in general the shorter the model series length, indicating a worse fit.

schemes and seems to be better when equal weights are given to all 4 approximation errors. The advantage of Watson's approach over the informal evaluation one is that it allows us to know the percentage of the spectral density of each actual series that the model is missing, which ranges from 2.7% for Y to 31% for AP in Model 1 (2.6% and 28% in Model 2) but from 91% for Y to 105% for C in Model 3. The coherence between C and Y is particularly well captured by Models 1 and 2 (only 5% and 2% higher than in actual US data respectively) and not that bad by Model 3 (27% higher in the model), while the hours-AP coherence is particularly badly captured in all three models. In general, Watson's measures of fit lead to prefer Model 2 to Model 1 (they have lower values) and to reject clearly Model 3 as possible explanations of the business cycle relationships between Y, C, H and AP observed in the US in 1964Q1-1995Q3.

4.1 Evaluating Watson's approach

Next we face Watson's methodology with a "test" similar to the one faced by the decision rule we constructed to approximate the informal evaluation approach. We keep the theoretical spectral density matrix of the model fixed across replications but estimate the actual data spectral density matrix at each replication of the Monte Carlo experiment. Actual data being generated from the DGP (Model 1), we simulate once series of a length usually found in practice (we actually take 127 observations as we had for the US data) using the parameter values of the first column in Table 1 and estimate with a Bartlett window their spectral density matrix. Then, the 7 measures of fit corresponding to BC frequencies are calculated minimizing the spectral density matrix of the approximation errors according to one of the two identification schemes explained above. Table 4 summarizes the empirical distribution of the 7 measures of fit across the 1000 replications[5] for each model being evaluated and for each identification scheme, with the mean, the standard deviation, and the 5%, 50% and 95% percentiles.

The median measure of fit across Monte Carlo replications indicates an error of 0.7% (cohe(H,Y)) to 7.4% (sp(C)) for Model 1, when 0% should be expected. This can be considered the "size" of Watson's model evaluation methodology according to our Monte Carlo experiment. As we pointed out when evaluating the naive calibrator's rule, apart from the obvious Monte Carlo error, this error may come mainly from the fact that we are comparing model spectra estimated for short time series simulated from the DGP to the theoretical model spectra. However, the divergence using Watson's approach is considerably smaller than that

[5] We choose 1000 replications for two reasons: first, because increasing the number of replications would have had a big cost in terms of computing time since we are computing the Monte Carlo distribution of the Watson measure of fit twice (one per identification scheme) for each of our three models and, second, because there cannot be less replications if we want the results to be comparable to other methodologies (for which we perform only 100 replications but statistics are simulated 100 times per replication).

we observed for the informal approach[6].

When assessing Model 2, the median measure of fit ranges from 2% (cohe(H,Y)) to 12% (cohe(H,AP)). This can be considered a measure of the "power" of Watson's approach versus models which are known to be close to the actual DGP. Such values indicate a worse fit (are further away from 0% for the spectra and from 100% for the coherences) for Model 2 than for Model 1, as we would expect. In both cases, the measures of fit increase when the identification scheme weights differently the errors (we obtain a range of median measures of fit of 0.7% to 38% for Model 1 and of 2% to 69% for Model 2). Although the "size" gets worse, the "power" to discriminate between the actual DGP and other models gets better. The bottom part of Table 4 shows a high "power" against models very different to the DGP. The median measures of fit indicate an error ranging from 16% to more than 100% under both identification schemes.

For comparison purposes with other evaluation criteria, we construct the following summary measure of the overall goodness of fit: we average across the 7 measures of fit (4 for power spectra and 3 for coherences at business cycle frequencies) the difference between their median value across Monte Carlo replications and that expected if the model was the true DGP (0% in the first 4 cases and 100% in the last 3). The resulting values are:

Identification scheme:	Model 1	Model 2	Model 3
Equal weights	3.87%	6.13%	70.29%
Different weights	16.97%	26.57%	73.71%

The first column can be associated to the "size" of the Watson (1993) evaluation methodology for calibrated models, whereas the 2nd and 3rd are measures of its "power". The table shows that, once the statistics of interest are selected, Watson's approach is substantially more accurate (better "size" and not much worse "power") when equal weights are given to all errors.

To summarize, extending Watson (1993) to evaluate calibrated models along multivariate

[6]We have performed a further sensitivity analysis on this issue and computed Table 4 for the case in which model spectral density matrices are also estimated for short time series simulated from the corresponding model instead of taking their theoretical values. That is, instead of simulating model series of 1000 observations, we have estimated the spectral density matrix from model series of length 500, 200 and 100.

For model series of 100 observations, the range of the median measure of fit rises to 4% to 21% when testing H_0: Model 1 = DGP (the range is equal under both identification schemes but most values are lower when equal weights are taken) but is 4.3% to 8% (8% to 54% when weighting only Y and H approximation errors) when testing H_0: Model 2 = DGP. The error induced using statistics estimated for short simulated time series leads to prefer Model 2 to the true DGP which is Model 1. However, the median mesures of fit when testing H_0: Model 3 = DGP for simulated series of 100 observations (ranging from 17% to 104%, or to 111% when weighting only Y errors) lead to reject Model 3 as clearly as when using the theoretical model spectral density matrix.

dimensions is a reasonably accurate evaluation methodology. Not only the error associated by the Watson's measures of fit to testing the correct model is reasonable (small "size"), but also these measures are able to evaluate "how different" from the DGP alternative models are: the "power" is higher the more different the spectral density of the model is from that estimated for the actual data (in Figure 2 the spectral density matrix of Model 3 is more different from Model 1 than that of Model 2). Using a standard 5% significance level, Watson's approach (with an identification scheme of equal weights) would successfully accept only the true model (Model 1). It would also indicate that the error that should be added to Model 2 to match the DGP is less than a 10th of what Model 3 would need.

5 DeJong, Ingram and Whiteman's approach

DeJong, Ingram and Whiteman (DIW) (1996) propose a bayesian evaluation methodology for calibrated models which takes into account the uncertainty present in the statistics of both actual and simulated data to measure the fit of the model to the data. They suggest representing the actual data with a VAR and computing the distribution of the statistics of interest by drawing VAR parameters from their posterior distribution. In constructing distributions of simulated model statistics, DIW consider only parameter uncertainty, and not the stochastic nature of the exogenous shocks as Canova and De Nicolò (1995). They use subjectively specified prior distributions (generally normal) for the parameters of the model whose location is set at the value typically calibrated in the literature while the dispersion is free. By enabling the specification of a sequence of increasingly diffuse priors over the parameter vector, the authors illustrate whether the uncertainty in the model's parameters can mitigate differences between the model and the actual data, so that the measure of dispersion can be used in order to (informally) minimize the distance between actual and simulated distributions of the statistics of interest.

DIW suggest two statistics aimed at synthetically measuring the degree of overlap among actual and model statistics distributions. The first one is the Confidence Interval Criterion (CIC), which is defined as

$$CIC_{ij} = \frac{1}{1-\alpha}\int_a^b P_j(s_i)ds_i \tag{7}$$

where s_i, $i = 1, \ldots, n$, are the statistics of interest, $a = \frac{\alpha}{2}$ and $b = 1 - a$ are the quantiles of $D(s_i)$, the distribution of the statistic in the actual data, $P_j(s_i)$ is the distribution of model statistic where j is the diffusion index of the prior on the parameter vector and $1 - \alpha = \int_a^b D(s_i)ds_i$. For CIC close to $\frac{1}{1-\alpha}$ the two distributions overlap substantially. If $CIC > 1$, $D(s_i)$ is diffuse relative to $P_j(s_i)$, i.e. the data is found to be relatively uninformative regarding

s_i. For CIC close to zero, the fit of the model is poor, either because the overlap is small or because P_j is very diffuse. The second statistics DIW propose helps distinguishing among these two possible interpretations. The Difference of Means statistic is analogous to a t-statistic for the mean of $P_j(s_i)$ in the $D(s_i)$ distribution, i.e.

$$d_{ji} = \frac{EP_j(s_i) - ED(s_i)}{\sqrt{varD(s_i)}} \tag{8}$$

Large values of d_{ji} indicate that the location of $P_j(s_i)$ is quite different from the location of $D(s_i)$.

By providing a distribution rather than a single number, DIW methodology gives a more comprehensive characterization of actual and model statistics relative to the informal approach and also to Watson's, although these distributions rely on the subjective priors given by the researcher. The two measures of fit (CIC and d) are complementary and give a good summary of the fit of the model, which allows for comparison across models, e.g. the smaller is the average d across statistics of interest the better the fit.

The results of applying DIW methodology to assess the fit of our three models with respect to actual US data are summarized in Table 6. Prior distributions for parameters used for each model are shown in Table 5. DeJong, Ingram and Whiteman (1996) illustrate their methodology with the simplest version of the King, Plosser and Rebelo (1988) model, so our choice for the prior distributions is similar to theirs, although some parameters (especially the ones related to the exogenous government spending process) have been chosen using as reference Baxter and King (1993) and Aiyagari, Christiano and Eichenbaum (1992). We take 1000 draws of the parameter vector and compute at each draw the statistics std(C)/std(Y), corr(C,Y), corr(H,Y) and corr(H,AP) implied by the model[7]. For each model we characterize the statistics' distributions with the 5%, 50%, 95% percentiles, the mean and the standard deviation. Actual data statistics are computed fitting a VAR to linearly detrended logs of US data for 1964Q1-1995Q3 and randomizing its coefficients so that the statistics are computed for 1000 draws from the VAR coefficients distributions. The first lines in Table 6 summarize the distribution of actual data statistics.

We assess the fit of each model computing both the percentage of the simulated statistics laying between the 5% and 95% percentiles of the actual statistics distribution (i.e. CIC measure with $\alpha = 10\%$) and the standarized differences of means (d-statistic). The CIC and difference of means measures suggest a reasonable fit for Models 1 and 2 but, contrary to

[7]Instead of taking the theoretical values of the statistics, and for consistency with the statistical treatment of the actual data, we simulate time series for Y, C, H and AP 10,000 observations long and compute the statistics for their linearly detrended logs, so that these statistics are sufficiently close to their theoretical values.

Watson's measure of fit, somehow better for the former with a higher overlap of distributions (average CIC of 0.87 versus 0.83 in Model 2) and with simulated statistics centered closer to actual ones in the case of Model 1 (smaller d-statistics). Model 1 statistics are less volatile than those observed for US data, while statistics from Model 2 and 3 are more volatile. This suggests that the standard deviation of government spending shocks has been left too volatile. Probably, reducing the standard deviation of its distribution would improve the fit of Model 2. The CIC and d-statistic for Model 3 clearly indicate a very bad fit. DeJong, Ingram and Whiteman (1996) obtain a worse fit than here for Model 1. There are two main reasons for this divergence. First, they evaluate 10 statistics of which std(C)/std(Y) and corr(C,Y) are the ones better captured by the model. Second, their actual data statistics differ from ours (different time period -1959Q1 to 1992Q2- and different detrending method -extracting a common time trend from consumption, investment and output-) especially those related to H. They measure H using average weekly hours of all workers instead of using per capita total hours (their measure of hours times employment divided by total population) as we do, following King, Plosser and Rebelo (1988)[8].

5.1 Evaluating DeJong, Ingram and Whiteman's approach

To evaluate the DIW methodology we conduct the same Monte Carlo experiments we have used before and test the following three hypotheses H_0: Model 1 = DGP, H_0: Model 2 = DGP and H_0: Model 3 = DGP. At each replication we generate a distribution of model statistics using 100 draws from the corresponding prior distributions for the parameters, simulating at each draw long time series for Y, C, H and AP and computing the statistics of their linearly

[8]This is an important difference, since we are including the evolution of both employment and hours in our measure of H whereas they only include that of hours worked by employees. A well known fact of the US economy is that about two thirds of the variation in total hours worked appears to be due to movements into and out of employment, while the remainder is due to adjustments in hours worked by employees. This fact has inspired a large number of real business cycle models which include nonconvexities in labor supply so that changes in total hours are brought by changing employment only (see Hansen (1985)) or changing both employment and hours per worker (see Cho and Cooley (1994)). The contemporaneous correlation between total hours and output reported in the literature for the US varies depending on the time period considered and the detrending method: using the Hodrick-Prescott filter Kydland and Prescott (1982) report a corr(H,Y) of 0.85, Hansen (1985) of 0.76, Cho and Cooley (1995) of 0.87 while King, Plosser and Rebelo (1988) report a contemporaneous corr(H,Y) of 0.07 extracting a common trend from output, consumption and investment. However, King, Plosser and Rebelo (1988) argue that this correlation rises considerably by splitting the sample into subperiods: they report that it averages 0.77 across subsamples. They interpret this sensitivity to the sample period as a suggestion that their detrending method may not have removed a low frequency component in output.

detrended logs. The distribution of actual DGP statistics is constructed similarly drawing parameter values from Model 1 priors but it is kept fixed across replications and for testing all three hypothesis.

Table 7 displays the median and standard deviation (across 100 Monte Carlo replications) of the 5%, 50% and 95% percentiles of the simulated distributions and, more importantly for evaluating the DIW methodology, medians and standard deviations of CIC (including the average CIC across statistics), d-statistic and the standarized difference of medians. For completeness, we have also computed the percentage of replications for which the difference of the medians exceeds 2 standard deviations of the actual statistic.

The overlap of DGP and Model 1 distributions is almost perfect (the median CIC across Monte Carlo replications is almost 1 in all 4 cases) and quite good but worse for Model 2, as expected. The d-statistic and the standarized difference of medians suggest that the worse fit of Model 2 is due to the fact that the mean and median of the DGP and model distributions are different, although the degree of overlap is high. This is especially the case for the corr(H,AP), as shown by the rejection frequency for the difference of medians (40% in Model 2 versus 1% in Model 1). That rejection frequency is 100% for all statistics but for std(C)/std(Y) under H_0: Model 3 = DGP. The methodology also reveals that among the four statistics, it is the relative standard deviation of C to Y the one that differs less between Models 3 and 1, both in the degree of overlap and in the location of the distributions. But the divergence is still clearly high.

Repeating what we have done with Watson's methodology, we construct a summary measure of the degree of rejection of a particular hypothesis, a measure which roughly captures the "size" and the "power" of the DIW methodology. For this purpose we choose the average of the differences between the four median CICs and their corresponding expected value if the model was the true DGP, i.e. equal to 1. It does not make much sense to include the difference between the d-statistic or standarized difference of medians and their expected values since they are measured in standard deviations of the actual mean or median, and hence not comparable. The values of the summary measure are:

Model 1	Model 2	Model 3
1.25%	9%	90.5%

According to this ad-hoc summary measure, the DIW methodology seems to be much more accurate as a model evaluation methodology than Watson's, showing a smaller "size" (1.25%) and higher "power" especially against Model 3.

6 Canova and De Nicolò approach

Canova and De Nicolò (1995) extend Canova (1994)-(1995) model evaluation methodology to a multivariate framework and compute measures of overlap of actual and model statistics in the same spirit as DIW but with some differences[9].

Canova (1994)-(1995) takes the actual data statistics as fixed numbers and uses the uncertainty of simulated data to provide a measure of fit for the model. In addition to allowing the realization of the exogenous disturbance to vary, he also allows for parameter variability in measuring the dispersion of simulated statistics. As in DIW, parameters are considered uncertain not so much because of sample variability, but because there are many estimates of the same parameter obtained in the literature since estimation techniques, samples and frequency of the data tend to differ. Canova proposes to calibrate the parameter vector to an interval selected on the basis of these estimates, rather than to a particular value or than centering an arbitrarily diffuse prior normal to a particular value, as in DIW. Once the empirical distribution of the statistics of interest is constructed (simulating the model repeatedly by drawing parameter vectors from the postulated distribution and by drawing realizations of the exogenous stochastic process from some given distribution), one can then compute either the size of calibration tests (using the actual statistic as a critical value for the simulated distribution) or the percentiles where the actual statistic lies.

Canova and De Nicolò (CDN) (1995) consider as DIW the uncertainty present in the statistics of both actual and model simulated data to measure the fit of the model to the data. CDN use a parametric bootstrap algorithm to construct distributions for the statistics of the actual data. In constructing distributions of simulated statistics, CDN take into account both the uncertainty in exogenous processes and parameters following Canova (1994)-(1995), while DIW only consider parameter uncertainty. To assess the degree of overlap of the two distributions, CDN choose a particular contour probability for one of the two distributions and ask how much of the other distribution is inside the contour. In other words, the fit of the model is examined very much in the style of the Monte Carlo literature: a good fit is indicated by a high probability covering of the two regions. To describe the features of the two distributions, they also repeat the exercise varying the chosen contour probability, say, from 50% to 75%, 90%, 95% and 99%. As in the DIW approach, actual data and simulated data are used symmetrically, in the sense that in the CDN approach one can either ask whether the actual data could be generated by the model, or viceversa, whether simulated data are consistent with the distribution of the observed sample. As the DIW approach, CDN provides a more comprehensive evaluation of a calibrated model than Watson's or than the informal approach.

[9]See Canova and Ortega (1996) for a detailed explanation and comparison of these approaches.

We have evaluated our three models with respect to US data using the CDN approach in Tables 9 and 10. For consistency with the analysis of other methodologies, we have computed separately the distributions of each statistic as in Canova (1994)-(1995) but considered distributions of both actual and model simulated statistics as in Canova and De Nicolò (1995). The distribution of actual data statistics has been constructed bootstrapping 1000 times the VAR fitted for the same US data used in previous sections (1964Q1-1995Q3). Statistics describing the bootstrap distribution of actual moments (5%, 50%, and 95% percentiles) are displayed in the top part of Table 9. The distribution is similar to that obtained using DIW (see Table 6) although std(C)/std(Y) is smaller and less volatile while corr(H,AP) is higher and more volatile. For each of the three models, we have simulated time series of the same sample size of actual data from the model 1000 times taking each time a draw from the parameters' prior distributions described in Table 8 and a different random realization for the exogenous disturbances. These distributions are constructed just as the empirical based distributions used to evaluate Baxter and Crucini (1993) in Canova and Ortega (1996). We have used existing estimates of these parameters in the literature or, when there are none, we have chosen a-priori an interval on the basis of theoretical considerations and imposed a uniform distribution on it. In comparison with the DIW approach, we drop the Normality assumption in many cases. At each draw we compute the 4 statistics for the simulated series, instead of taking their theoretical counterparts as in DIW. We are introducing two new sources of error with respect to DIW: a Monte Carlo error for the fact of taking random realizations of the shocks series and an estimation error for computing the simulated statistics from short time series for Y, C, H and AP. Once the two distributions are constructed, we report the following measures of overlap: percentage of the distribution of simulated statistics into the 50%, 90% and 95% one-sided, 90% and 95% two-sided confidence intervals of the bootstrap distribution of actual statistics, and viceversa.

Model 1 statistics are smaller than actual US data ones (except for corr(H,AP)) but more volatile. They lay substantially inside the actual distributions but the distributions are not equally centered (a lower percentage of the simulated distributions lay inside the two-sided actual distributions) whereas the actual statistics, being higher in values, lay in higher percentage inside the two-sided confidence intervals of simulated statistics which include higher values than the one-sided confidence intervals. The introduction of government spending shocks makes simulated statistics of Model 2 even more volatile and their medians even lower than Model 1 ones (except for corr(H,Y)) so that the percentage of simulated statistics inside the two-sided confidence intervals of actual statistics gets smaller and the other way around for actual statistics into simulated distributions. In sum, the two distributions lay further apart while the coverage is still high overall. Model 3 statistics are less volatile than those of Model

1 or Model 2 (the volatility allowed in Table 8 for government spending shocks is smaller than that of technology shocks) but their median values lay so far from the actual ones that the two distributions hardly overlap when considering two-sided confidence intervals. For example, 100% of the actual corr(H,Y) are smaller than the median value of 0.99 implied by Model 3, and 100% of simulated corr(C,Y) and corr(H,AP) are smaller than the actual ones (the median values implied by the model are -0.94 and -0.97, respectively).

Overall, the CDN methodology conducts an even more thorough analysis of the model and actual statistics distributions than the DIW approach. It discriminates better between the first two models, even though it gives a similar picture of the fit: Model 1 appears preferable to Model 2 as whith the DIW methodology. This is because Model 2 statistics are more volatile and centered further away from the actual ones. Nevertheless, the fit of Model 2 remains fair. As it happens with other model evaluation criteria, the CDN approach gives a bad fit to Model 3.

6.1 Evaluating Canova and De Nicolò approach

The results of the Monte Carlo experiment evaluating the performance of CDN are summarized in Table 11. As with the DIW approach, we keep the distribution of actual DGP moments fixed across replications and for evaluating all three models. Such distribution has been generated simulating 100 times the DGP (Model 1) using a single realization of the technology shock process and one draw of the parameter vector from the prior distributions reported in column 1 of Table 8 at each iteration, and computing the statistics each time. At each replication, distributions of model statistics have been constructed by simulating 100 times the corresponding model, and one-sided and two-sided measures of overlap between actual and simulated distributions of each statistic have been computed. Table 11 displays their medians and standard deviations across the 100 Monte Carlo replications performed.

The last column of Table 11 presents the theoretical value of each measure of overlap which should be found if the model was the true DGP. The difference between the empirical values and the theoretical ones is an indication of the "size" of the CDN methodology when testing H_0: Model 1 = DGP, and of its "power" when testing H_0: Model 2 = DGP or H_0: Model 3 = DGP. Finally, we define a summary measure of the percentage rejection of each H_0 averaging across the 40 measures of overlap the difference between their median values across replications and their expected true values. These summary measures are the following:

Model 1	Model 2	Model 3
0.55%	7.8%	53.4%

It is remarkable how accurately the CDN methodology recognises the true DGP: all 40 statistics presented in Table 11 for Model 1 are almost equal to their theoretical values. In fact the "size" measure is lower than using Watson or DIW methodologies. And this is particularly remarkable with respect to DIW, since as explained above we are introducing both a Monte Carlo and an estimation error when computing simulated statistics. This better "size" comes at one cost: the "power" against alternative models is in fact reduced. However, we still find that Model 3 is clearly rejected while the rejection of Model 2 is somewhat marginal, i.e. CDN manages to correctly rank the alternative models different to the DGP.

7 Spectral density distance approach

The last approach we evaluate is the one presented in Ortega (1998), where an asymptotic test is derived for the hypothesis that the quadratic standarized distance between the spectral density matrices of simulated and actual data is zero or less than an arbitrary prespecified bound. It is especially suitable for assessing the performance of models at a certain frequency range, such as business cycle models.

The test statistic proposed explicitly acknowledges that the solution paths generated by the model for the variables of interest are only approximations to the true model solution. Watson (1993) also recognises that there is an approximation error but, contrary to his approach, Ortega (1998) takes it into account to derive a formal test of the distance between the model and the observed data. Diebold, Ohanian and Berkowitz (1995) propose a measure of distance to evaluate how well the model matches the spectral density matrix of the actual data, too, but they assume that model spectra can be obtained without error. On the contrary, Ortega (1998) compares actual to simulated data by treating them as samples from an unknown DGP and hence both spectral density matrices are estimated with error (the former because of sampling error, and the latter because of the approximation error). As in the DIW and CDN methodologies, the test proposed treats symmetrically actual and simulated data by taking into account the uncertainty existing in both data sets. While not excluding the possibility of stochastic parameters in the model, the uncertainty considered in the model derives from the fact that there exists an approximation error. The main differences between this methodology and that of DIW and CDN are, first, that both sets of statistics are estimated in a classical instead of a Bayesian way and, second, that model and actual data are compared using test statistics with known asymptotic distributions.

More specifically, the assessment of the fit of a model over a particular set of frequencies

(e.g. business cycle frequencies $[\omega_1,\omega_2]$) is based on testing the following null hypothesis

$$H_0 : \Lambda D(\omega;\gamma) = 0 \quad , \quad \forall\omega \in [\omega_1,\omega_2]$$

where Λ is a selection matrix which weights the elements in the measure of distance $D(\omega;\gamma)$. $D(\omega;\gamma)$ is defined as

$$D(\omega;\gamma) = S\text{vec}f(\omega;\gamma) = \text{vec}f^{\mathbf{y}}(\omega) - \text{vec}f^{\mathbf{x}}(\omega;\gamma) \tag{9}$$

where $f(\omega;\gamma)$ is the spectral density matrix of the vector $[y_t \;\; x_t(\gamma,z_t)]$. y_t and x_t are the 1xN vectors of actual and simulated data, respectively. x_t depends on the model parameter vector γ and the exogenous shocks series z_t. $f^{\mathbf{y}}(\omega)$ and $f^{\mathbf{x}}(\omega;\gamma)$ are the upper left and lower right submatrices of $f(\omega;\gamma)$. To test H_0, the following test statistic is proposed

$$fit([\omega_1,\omega_2];\gamma) = \sum_{\omega=\omega_1}^{\omega_2} \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right)}' \left(\Lambda\Sigma_D(\omega;\gamma)\Lambda'\right)^{-1} \sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma) \tag{10}$$

where $\Sigma_D(\omega;\gamma)$ is the covariance matrix of $D(\omega;\gamma)$ and $\hat{D}(\omega;\gamma)$ is the estimated distance. The asymptotic distribution and properties of $\hat{D}(\omega;\gamma)$ are derived from those of the spectral density matrix estimator $\hat{f}(\omega;\gamma)$. Appropriately choosing the spectral window function and the bandwidth parameter (see Ortega (1998)) we can derive the following asymptotic distribution of the *fit* test statistic for each frequency under H_0

$$fit(\omega;\gamma) \;\dot{\sim}\; \chi^2_{(N^2-Q)} \quad , \; \omega \neq 0, \pm\pi \tag{11}$$

where Q is the number of zero elements in the diagonal of Λ. Therefore,

$$fit([\omega_1,\omega_2];\gamma) \dot{\sim} \chi^2_{L(N^2-Q)}$$

where L is the number of frequencies included in $[\omega_1,\omega_2]$.

The main advantage of this methodology relative to others is that it can reject or accept a model in a strict statistic sense, because such a statement is made by comparing a test statistic to a known asymptotic distribution. On the other hand, it provides only one measure of fit per frequency, while Watson's approach provides one per statistic (as the informal approach), and CDN and DIW provide a wider variety of measures of fit.

Each time h the model is simulated, an $\hat{f}^{\mathbf{x}}_h(\omega;\gamma)$ is estimated keeping y_t fixed and using x_{ht}, for $h = 1,\ldots,H$. In practice, what we are interested in obtaining is the average across the H estimated distances, i.e. $\hat{D}(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H}\hat{D}_h(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H} S\,\text{vec}\hat{f}_h(\omega;\gamma)$. Given that x_{ht} are iid, that average keeps the same distribution and theoretical mean than $\hat{D}_h(\omega;\gamma)$, and $\Sigma_D(\omega;\gamma)$ becomes $\frac{1}{H}\Sigma_D(\omega;\gamma)$. The asymptotic distribution for $fit([\omega_1,\omega_2];\gamma)$ remains valid

as long as we premultiply $\hat{D}(\omega;\gamma)$ by $\sqrt{H}$ when constructing the test statistic. Then, H_0 will be rejected and the distance between the model and the actual data found significantly different from zero if $fit([\omega_1,\omega_2];\gamma)$ is greater than the critical value of a $\chi^2_{L(N^2-Q)}$, for a selected significance level α.

To assess the fit of our three models, we simulate 1000 times (H=1000) the model using the parameter vectors (γ) of Table 1 and compute $\hat{D}_h(\omega;\gamma)$ at each simulation by estimating the joint spectral density matrix for linearly detrended logs of the 4 actual US series and those simulated from the model. We have used a Quadratic Spectral window function and an optimal spectral window parameter estimate following Andrews (1991). Then we have taken the average $\hat{D}_h$ across simulations and computed $fit(\omega_1;\gamma)$, $fit(\omega_2;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$, where ω_1 (ω_2) are the frequencies associated with cycles 8 years (2 years) long. We have given equal weights to the elements in the spectral density submatrices $\hat{f}^{\mathbf{y}}(\omega)$ and $\hat{f}^{\mathbf{x}}(\omega;\gamma)$ (Q=0), therefore the asymptotic distributions of the three *fit* statistics are χ^2_{16}, χ^2_{16} and $\chi^2_{7\times16}$, respectively (the number of independent frequencies included in the $[\omega_1,\omega_2]$ interval depends on the value of the Andrews optimal bandwidth parameter which in turn depends on the parametric model fitted for the actual data, and is 7 in this case). To evaluate the $fit([\omega_1,\omega_2];\gamma)$ statistic, we have used the following Normal approximation typically used for χ^2_k distributions of k > 100 : $\sqrt{2\,\chi^2_k} \sim N(\sqrt{2k-1};1)$.

Following we report the *fit* statistics and the 90% and 95% critical values:

	Model 1	Model 2	Model 3	90% C.V.	95% C.V.
$fit(\omega_1;\gamma)$	7379	7614	29819	23.5	26.3
$fit(\omega_2;\gamma)$	7264	7537	29342	23.5	26.3
$fit([\omega_1,\omega_2];\gamma)$	11224	11588	45863	137.3	142.7

It turns out that none of the models is accepted in strict statistic sense as the US data DGP: the values of all test statistics are clearly grater than the critical values in all cases. As pointed out in Ortega (1998), the sample size of the data used is too short to assume the asymptotic normality of $\hat{D}(\omega;\gamma)$. Its distribution may be closer to a χ^2 and therefore how many $\hat{D}_h(\omega;\gamma)$ it aggregates matters, and the small sample distribution of $fit(\omega;\gamma)$ would also depend on H (and hence, the critical values too).

On the other hand, and consistently to other model evaluation methodologies, Model 1 and Model 2 are found to be almost equally closer to the actual data and Model 3 four times more distant. Just as the CDN and DIW methodologies (both allowing for random parameters in the models) and contrary to Watson's, Model 1 appears closer to the US data than Model 2.

7.1 Evaluating the spectral density distance approach

We finally perform the Monte Carlo experiment on this last methodology. At each replication, one realization of the 4 time series from the DGP (using Model 1 with fixed parameters) is compared to one of the 4 simulated series from the corresponding model, for each of the 100 times the model is simulated, and we compute the average estimated distance across the 100 simulations to calculate the $fit(\omega_1;\gamma)$, $fit(\omega_2;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$ statistics as explained above.

Table 12 summarizes the performance of the *fit* test statistics in testing H_0: Model i = DGP, for $i = 1$, 2 and 3. It displays the percentage rejection of each hypothesis when comparing the corresponding *fit* test statistic to its 90% and 95% critical values, and the 5%, 50%, 90%, 95% percentiles and the mean and standard deviation of each of the three *fit* statistics computed, across the 100 replications of the Monte Carlo experiment. Now the number of frequencies included in the business cycle interval is 15 instead of 7 (because the optimal bandwidth parameter has changed since the actual data now is not the US observed data but that simulated from the DGP -Model 1-) and hence the critical values for the $fit([\omega_1,\omega_2];\gamma)$ test statistic change.

The first two rows of statistics reported in Table 12 for testing each hypothesis are the empirical size (for H_0: Model 1 = DGP) and power (for H_0: Model 2 = DGP and H_0: Model 3 = DGP) of the spectral density distance methodology. A summary table comparable to the measures of the size and power we have computed for the other methodologies would be:

	Model 1	Model 2	Model 3
significance level 5%	0%	0%	100%
significance level 10%	0%	0%	100%

The small size found (0% versus theoretical 5% or 10%), is consistent with the Monte Carlo experiment on the small sample properties of the *fit* test statistic performed in Ortega (1998).

A superficial analysis would lead us to think that the power against models not too different from the DGP (Model 2) is null. The actual values of the *fit* statistic clearly indicate a worse fit for Model 2 than for Model 1 (as should be the case) but not bad enough to reject that the spectral density of Model 2 is equal to that of the DGP, contrary to the previous methodologies. Part of the reason can be in that the *fit* test is a single overall measure of fit, while Watson's approach and especially DIW and CDN approaches can capture the discrepancy between the model and the DGP along many dimensions (they compute several measures of fit for each statistic), and this property is kept even when aggregating their different measures of fit into a summary one as we present at the end of each section. However, recall that this methodology

tests the distance between model and actual spectral densities. Figures 1 and 2 clearly show that the frequency domain properties of Model 1 (our DGP) and Model 2 are almost indistinguishable at BC frequencies. When models, as it is the case for most SDGE models, are known to be false, we want the evaluation methodology to be able to capture how well they reproduce certain particular statistics. It turns out that two models different but close to each other as Model 2 to Model 1 may generate almost the same statistics we are interested in, in our case the multivariate spectral density matrix for Y, C, H and AP. We interpret the apparent inability for discriminating between Model 1 and Model 2 shown when evaluating both of them as exactly equal to the DGP (0% rejection) as the correct indication that both reproduce almost exactly the precise statistic of the DGP we want them to replicate. Hence we should be equally happy with both of them just as we would be equally unhappy if the DGP's spectral density differs equally from both. In fact, consistently with the results in Ortega (1998), the *fit* test is very powerful when the alternative hypothesis imply different spectral density matrices to the DGP (see Model 3 spectral properties in Figure 2).

The issue arising here is an important one: before using any model evaluation methodology, it should be checked whether discriminating models according to the statistics they focus on (in this case, the spectral densities at particular frequencies) is desirable. It is highly probable that the spectral densities at low frequencies (such as business cycle frequencies) of series with similar autorregressive structures with high persistence parameters will not significantly differ, as was shown in Ortega (1998). Many real business cycle model series follow that structure. On top of that, using a detrending method which does not totally remove very low frequency movements (as the linear detending method) concentrates relatively less spectral density at other frequencies, making harder to discriminate models according to their spectra at, say, the higher frequencies included in the business cycle range.

Probably, the real business cycle models we have chosen for performing our comparison exercise yield more observable differences between alternative models when evaluated using time domain statistics such as relative standard deviations and correlations than when using frequency domain statistics. A simple look at the discrepancies observed between the statistics simulated from Model 1 and 2 in Table 2 as compared to the difference between spectra and coherencies simulated from the same two models in Figures 1 and 2 confirms this point.

8 Conclusions

In this paper we compare under uniform conditions the performance of alternative methodologies recently proposed in the literature to evaluate stochastic dynamic general equilibrium models.

We have first described the approaches, emphasizing the differences among them and with the standard informal evaluation approach. Second, we have illustrated the methodologies of Watson (1993), DeJong, Ingram and Whiteman (1996), Canova and De Nicolò (1995) and the one based on spectral density distance presented in Ortega (1998), using three versions of a simple one-sector real business cycle model from King, Plosser and Rebelo (1988). Government shocks seem to add little or none explanatory power to technology shocks in a one-sector SDGE model for the US, and are certainly not enough to provide a reasonable fit when considered as the only source of fluctuations.

The main contribution of this paper is to conduct a Monte Carlo experiment on the four methodologies to "test" them and compare their performance as evaluation procedures for dynamic general equilibrium models. We have encountered several difficulties in undertaking this task, which are mainly related to four facts. First, the comparison is made on a multivariate level, which complicates the effort of summarizing the overall performance of each methodology. Second, some methodologies are constructed in the frequency domain (Watson's and the spectral density distance approaches) while others are built in the time domain (DIW and CDN). Third, DIW and CDN define distributions for the parameters of the model in different ways while the two other approaches take parameters as fixed. Fourth, Watson and DIW use the theoretical values of model statistics (DIW also for actual data statistics) while CDN and the spectral density distance approach estimate them. Despite of these difficulties, we have been able to compute rough measures of the "size" and "power" of each model evaluation methodology.

Our exercise highlights that there are differences between the methodologies along many dimensions, but looking at the summary comparison provided by the "size" and "power" measures it is the two approaches allowing for stochastic parameters (DIW and CDN) the ones that seem to achieve a better performance. Probably, the real business cycle models we have chosen for performing our comparison exercise yield more observable differences between models when evaluated using time domain statistics such as relative standard deviations and correlations than when using frequency domain statistics. A simple look at the discrepancies observed between the statistics simulated from Model 1 and 2 in Table 2 as compared to the difference between spectra and coherencies simulated from the same two models in Figures 1 and 2 confirms this point.

In fact, although the spectral density distance approach presented in Ortega (1998) is the one which obtains the smaller size and the larger power against models very different to the DGP (Model 3), it shows no power against Model 2. The spectral density matrices of Model 1 and 2 do not differ enough for the methodology to recognise them as different models when comparing the fit-statistic to standard significance levels. But the actual value of the statistic or its p-value associated are different: that of Model 2 is further from 0 than that of Model 1, showing their relative closeness to the DGP. Watson's approach, also in the frequency domain, has significantly worse size and worse power against Model 3, but captures some discrepancy between the spectral properties of Model 1 and 2.

The time domain approaches of DeJong, Ingram and Whiteman (1996) and Canova and De Nicolò (1995) appear more accurate than Watson's. Among this two approaches, CDN achieves a better "size" at the cost of a lower "power", which is still enough to correctly rank the models according to their discrepancy with the true DGP.

We find that all four methodologies outperform the informal approach since they substantially reduce the risk of rejecting the true DGP, are able to discriminate more clearly between the DGP and models very distant from it and all but the spectral density distance approach (for the reasons explained above) also have power against models whose DGP is slightly different to the true DGP.

References

[1] Aiyagari, R., Christiano, L. and M. Eichenbaum (1992) "The Output, Employment, and Interest Rate Effects of Government Consumption", *Journal of Monetary Economics*, 30, 73-86.

[2] Andrews, D.W. (1991), "Heteroskedasticity and autocorrelation consistent covariance matrix estimation", *Econometrica*, vol.59, No.3, 817-858.

[3] Baxter, M. and M. Crucini (1993) "Explaining Saving-Investment Correlations", *American Economic Review*, 83, 416-436.

[4] Baxter, M. and R. King (1993) "Fiscal Policy in General Equilibrium", *American Economic Review*, vol.83 (3), 315-334.

[5] Brock, W.A. and L. Mirman (1972) "Optimal Economic Growth and Uncertainty: The Discounted Case", *Journal of Economic Theory*, 479-513.

[6] Canova, F. (1994) " Statistical Inference in Calibrated Models", *Journal of Applied Econometrics*, 9, S123-S144.

[7] Canova, F. (1995) "Sensitivity Analysis and Model Evaluation in Simulated Dynamic General Equilibrium Economies", *International Economic Review*, 36, 477-501.

[8] Canova, F. (1997) "Detrending and Business Cycle Facts", *Journal of Monetary Economics*, forthcoming.

[9] Canova, F. and G. De Nicoló (1995), "The Equity Premium and the Risk Free Rate: A Cross Country, Cross Maturity Examination", CEPR Working Paper 1119.

[10] Canova, F. and E. Ortega (1996), "Testing Calibrated General Equilibrium Models", Universitat Pompeu Fabra Economics Working Paper 166. Forthcoming in Mariano, R., Schuermann, T. and M. Weeks (eds.), *Simulation Based Inference in Econometrics: Methods and Applications*, Cambridge: Cambridge University Press.

[11] Christiano, L.J. and M. Eichenbaum (1992), " Current Business Cycle Theories and Aggregate Labor Market Fluctuations", *American Economic Review*, 82, 430-450.

[12] DeJong, D., Ingram, B. and C. Whiteman (1996), "A Bayesian Approach to Calibration", *Journal of Business and Economic Statistics*, 14, 1-9.

[13] Den Haan, W. and A. Marcet (1994), "Accuracy in Simulations", *Review of Economic Studies*, 61, 3-17.

[14] Diebold, F., Ohanian, L. and J. Berkowitz (1995), "Dynamic Equilibrium Economies: A Framework for Comparing Models and Data", NBER Technical Working Paper No.174.

[15] Gali, J. (1994), "Monopolistic Competition, Business Cycles and the Composition of Aggregate Demand?", *Journal of Economic Theory*, 63, 73-96.

[16] Gali, J. (1995), "Real Business Cycles with Involuntary Unemployment", CEPR Discussion Paper No.1206.

[17] Granger, C.W.J. (1964), *Spectral Analysis of Economic Time Series*, Princeton University Press.

[18] Gregory, A. and G. Smith (1991), " Calibration as Testing: Inference in Simulated Macro Models", *Journal of Business and Economic Statistics*, 9(3), 293-303.

[19] Gregory, A. and G. Smith (1993), " Calibration in Macroeconomics", in Maddala, G.S. (ed.), *Handbook of Statistics*, vol. 11, Amsterdam, North Holland.

[20] Hannan, E.J. (1970), *Multiple Time Series*, John Wiley and Sons.

[21] *Journal of Business and Economic Statistics*, January 1990.

[22] King, R., Plosser, C. and S. Rebelo (1988), " Production, Growth and Business Cycles: I and II", *Journal of Monetary Economics*, 21, 195-232 and 309-342.

[23] King, R., Plosser, C. and S. Rebelo (1990), " Production, Growth and Business Cycles: Technical Appendix", mimeo, University of Rochester.

[24] Kim, K. and A. Pagan (1995) "The Econometric Analysis of Calibrated Macroeconomic Models", in Pesaran, H. and M. Wickens, eds., *Handbook of Applied Econometrics*, vol.I, London: Blackwell Press.

[25] Kydland, F. and E. Prescott (1982), "Time To Build and Aggregate Fluctuations", *Econometrica*, 50, 1345-1370 .

[26] Kydland, F. and E. Prescott (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles", *The Scandinavian Journal of Economics*, 93(2), 161-178.

[27] Lucas, R.E., Jr. (1976), "Econometric Policy Evaluation: A Critique", in Brunner, K. and A. Meltzer (eds.), *Carnegie-Rochester Series on Public Policy*, North-Holland, vol.1, 19-46.

[28] Lucas, R.E., Jr. (1977), 'Understanding Business Cycles', in Brunner, K. and A. Meltzer (eds.), *Carnegie-Rochester Series on Public Policy*, No.5.

[29] Lucas, R.E., Jr. and T.J. Sargent (1981), *Rational Expectations and Econometric Practice*, Minneapolis: The University of Minnesota Press.

[30] Marcet, A. (1994), "Simulation Analysis of Stochastic Dynamic Models: Applications to Theory and Econometrics" in Sims, C. (ed.), *Advances in Econometrics. Sixth World Congress of the Econometric Society*, Cambridge: Cambridge University Press.

[31] Newey, W. and K. West (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-708.

[32] Ortega, E. (1998), "Assessing the Fit of Simulated Multivariate Dynamic Models", Documento de Trabajo, Servicio de Estudios, Banco de España.

[33] Pagan, A. (1994), "Calibration and Econometric Research: An Overview", *Journal of Applied Econometrics*, 9, S1-S10.

[34] Priestley, M.B. (1981), *Spectral Analysis and Time Series*, Academic Press.

[35] Rotemberg, J. and M. Woodford (1991), " Markups and the Business Cycle", in Blanchard, O.J. and S. Fisher (eds.) *NBER Macroeconomics Annual 1991*, Cambridge: MIT Press.

[36] Sargent, T. (1979), *Macroeconomic Theory*, New York: Academic Press.

[37] Sargent, T. (1987), *Dynamic Macroeconomic Theory*, Cambridge, Ma: Harvard University Press.

[38] Stadler, G.W. (1994), "Real Business Cycles", *Journal of Economic Literature*, vol.XXXII, 1750-1783.

[39] Ubide, A.J. (1995), "On International Business Cycles", Ph.D. Dissertation, European University Institute.

[40] Watson, M. (1993), "Measures of Fit for Calibrated Models", *Journal of Political Economy*, 101, 1011-1041.

Table 1: **Baseline parameter values**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor in Output (α)	0.58	0.58	0.58
Growth rate (θ_x)	1.0036	1.0036	1.0036
Depreciation Rate of Capital (δ_K)	0.025	0.025	0.025
Discount Factor (β)	0.9875	0.9875	0.9875
Steady State hours ($\bar{H}$)	0.20	0.20	0.20
Risk Aversion (σ)	2	2	2
Share of Government Spending in Output (sg)	0.25	0.25	0.25
Tax Rate (τ)	0.25	0.25	0.25
Persistence of Technology Disturbances (ρ_A)	0.9	0.9	0
Persistence of Government Spending Disturbances (ρ_G)	0	0.97	0.97
Standard Deviation of Technology Innovations (σ_A)	0.00852	0.00852	0
Standard Deviation of Government Spending Innovations (σ_G)	0	0.0036	0.0036

Table 2: **Actual data and simulated moments**

Statistic	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks	Actual Data
std(C)/std(Y)	.667 (.088)	.671 (.094)	.717 (.006)	.826
corr(C,Y)	.869 (.038)	.859 (.044)	-.999 (.0003)	.863 (.133)
corr(H,Y)	.776 (.097)	.765 (.108)	.999 (0)	.807 (.157)
corr(H,AP)	.374 (.171)	.344 (.182)	-.999 (.0001)	-.065 (.265)

Notes: Moments of both model simulated series and actual data are computed after linearly detrending the series.

Actual data are logs of US per capita real variables in $Mln and for the period 1964Q1-1995Q3. Newey and West (1987) consistent S.E. are reported for the correlation coefficients. Simulated statistics are average (std) across 100 simulations of the corresponding model where at each simulation different random series are used for the exogenous shocks and time series for Y, C, H and AP are generated of sample size equal to the actual data (127 observations). The random number generator is seeded at 0 before simulating each model. When persistence parameters or S.D. of innovations are 0, model simulations are run using $1x10^{-10}$ instead to avoid non-full rank matrix problems.

Table 3: **Watson's measures of fit. Averages across BC frequencies**

	sp(Y)	sp(C)	sp(H)	sp(AP)	cohe(C,Y)	cohe(H,Y)	cohe(H,AP)
Actual data statistics	.0004	.0003	.0002	.0001	.79	.64	.04
Model 1 statistics	.0004	.0001	.0001	.0001	.83	.87	.52
Measures of Fit for Model 1 (only A_t shocks)							
Equal Weight	.027	.22	.18	.31	1.05	1.35	13.12
Min error(Y)	.012	.46	.50	1	1.05	1.35	13.12
Model 2 statistics	.0004	.0001	.0001	.0001	.80	.85	.46
Measures of Fit for Model 2 (both A_t and G_t shocks)							
Equal Weight	.026	.22	.16	.28	1.02	1.31	11.64
Min errors(Y,H)	.13	.82	.26	1.61	1.02	1.31	11.64
Model 3 statistics	.00002	.00001	.00004	.00001	.997	.999	.998
Measures of Fit for Model 3 (only G_t shocks)							
Equal Weight	.91	1.05	.74	.97	1.27	1.55	25.21
Min error(Y)	.87	1.11	.78	1.10	1.27	1.55	25.21

See text for explanation

Table 4: **Monte Carlo on Watson's measures of fit**

Summary statistics	sp(Y)	sp(C)	sp(H)	sp(AP)	cohe(C,Y)	cohe(H,Y)	cohe(H,AP)
Testing H_0: Model 1 = DGP							
Identification: equal weights							
mean	.07	.15	.09	.15	1.005	1.05	1.16
std	.10	.23	.09	.23	.07	.15	.53
5%perc	.008	.023	.024	.023	.90	.91	.67
median	.034	.074	.062	.074	.99	1.007	.99
95%perc	.24	.55	.27	.55	1.14	1.34	2.13
Identification: min error(Y)							
mean	.065	.48	.44	.48	1.005	1.05	1.16
std	.10	.30	.20	.30	.07	.15	.53
5%perc	.006	.26	.22	.26	.90	.91	.67
median	.031	.38	.37	.38	.99	1.007	.99
95%perc	.24	1.04	.82	1.04	1.14	1.34	2.13
Testing H_0: Model 2 = DGP							
Identification: equal weights							
mean	.07	.15	.10	.15	.97	1.03	1.03
std	.10	.24	.11	.24	.07	.15	.47
5%perc	.009	.023	.027	.023	.87	.89	.59
median	.035	.073	.068	.073	.96	.98	.88
95%perc	.25	.57	.32	.57	1.11	1.31	1.89
Identification: min error(Y) and error(H)							
mean	.12	.79	.25	.79	.97	1.03	1.03
std	.11	.37	.15	.37	.07	.15	.47
5%perc	.05	.46	.12	.46	.87	.89	.59
median	.09	.69	.21	.69	.96	.98	.88
95%perc	.31	1.48	.53	1.48	1.11	1.31	1.89
Testing H_0: Model 3 = DGP							
Identification: equal weights							
mean	.90	1.06	.63	1.06	1.21	1.22	2.22
std	.03	.06	.05	.06	.09	.18	1.02
5%perc	.85	.98	.54	.98	1.09	1.05	1.28
median	.90	1.05	.64	1.05	1.20	1.16	1.92
95%perc	.95	1.17	.72	1.17	1.38	1.55	4.10
Identification: min error(Y)							
mean	.86	1.17	.70	1.17	1.21	1.22	2.22
std	.02	.05	.06	.05	.09	.18	1.02
5%perc	.81	1.10	.59	1.10	1.09	1.05	1.28
median	.86	1.16	.70	1.16	1.20	1.16	1.92
95%perc	.90	1.28	.79	1.28	1.38	1.55	4.10

Empirical distribution of Watson's Measures of Fit at business cycle frequencies over 1000 replications. Measures of Fit at each replication are computed as in Table 3.

Table 5: **Parameter distributions for the DIW methodology**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor in Output (α)	N(0.58,0.05)	N(0.58,0.05)	N(0.58,0.05)
Growth rate (θ_x)	1.0036	1.0036	1.0036
Depr. Rate of Capital (δ_K)	N(0.025,0.004)	N(0.025,0.004)	N(0.025,0.004)
Discount Factor (β)	N(0.988,0.001)	N(0.988,0.001)	N(0.988,0.001)
Steady State hours ($\bar{H}$)	N(0.20,0.02)	N(0.20,0.02)	N(0.20,0.02)
Risk Aversion (σ)	N(2,1)	N(2,1)	N(2,1)
Share of Government Spending in Output (sg)	N(0.25,0.05)	N(0.25,0.05)	N(0.25,0.05)
Persistence of Technology Disturbances (ρ_A)	N(0.9,0.25)	N(0.9,0.25)	0
Persistence of Government Spending Disturbances (ρ_G)	0	N(0.97,0.02)	N(0.97,0.02)
St.D. of Technology Innovations (σ_A)	N(0.00852,0.004)	N(0.00852,0.004)	0
St.D. of Government Spending Innovations (σ_G)	0	N(0.0036,0.002)	N(0.0036,0.002)

Table 6: **DeJong, Ingram and Whiteman methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
US Data, 1964Q1-1995Q3				
5% perc	.72	.77	.82	-.16
median	.85	.87	.89	.12
95% perc	1.11	.96	.96	.60
mean	.89	.81	.85	.09
std	.46	.29	.16	.39
Simulated statistics, Model 1				
5% perc	.71	.78	.004	-.40
median	.87	.89	.50	.04
95% perc	1.07	.95	.73	.33
mean	.88	.88	.45	.01
std	.11	.06	.23	.22
Evaluating Model 1				
CIC	1.07	1.07	0.26	1.09
Average(CIC)	.87			
d-statistic	-.03	.33	-2.48	-.23
Simulated statistics, Model 2				
5% perc	.67	.07	.13	-.01
median	.87	.87	.53	-.15
95% perc	1.25	.95	.99	-.45
mean	1.41	.79	.51	-.03
std	3.07	.23	.26	.23
Evaluating Model 2				
CIC	.99	.98	.27	1.09
Average(CIC)	.83			
d-statistic	1.12	-.10	-2.13	-.33
Simulated statistics, Model 3				
5% perc	.20	-.20	.76	.10
median	.74	-.90	.99	-.97
95% perc	2.02	-.99	1	-.99
mean	1.36	-.75	.95	-.84
std	3.11	.35	.11	.37
Evaluating Model 3				
CIC	.45	.006	.22	.097
Average(CIC)	.19			
d-statistic	1.03	-7.85	.63	-2.38

Notes: Actual data statistics are computed fitting a VAR to linearly detrended logs of US data for 1964Q1-1995Q3 and randomizing its coefficients so that the statistics are computed for 1000 draws from the VAR coefficients distributions.
Simulated statistics are computed for series of 10,000 observations simulated from each model 1000 times, using at each simulation a different draw from the prior distributions of the parameters in Table 5.

Table 7: **Monte Carlo on DIW methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
Testing H_0: Model 1 = DGP				
Simulated statistics, Model 1				
5% perc	.71 (.02)	.77 (.04)	.02 (.11)	-.41 (.08)
median	.87 (.01)	.89 (.006)	.50 (.02)	.03 (.03)
95% perc	1.06 (.03)	.95 (.007)	.72 (.02)	.30 (.07)
Evaluating Model 1				
CIC	1.01 (.03)	.99 (.04)	1.01 (.03)	1.02 (.03)
Average(CIC)	1.003 (.02)			
d-statistic	.001 (.09)	.0009 (.11)	.071 (.09)	-.058 (.08)
Diff.Medians	-.04 (.11)	-.06 (.09)	.02 (.10)	-.05 (.13)
Rej. freq of Diff.Medians	0%	0%	0%	1%
Testing H_0: Model 2 = DGP				
Simulated statistics, Model 2				
5% perc	.65 (.03)	.12 (.25)	.14 (.10)	.001 (.015)
median	.86 (.01)	.86 (.009)	.54 (.02)	.13 (.15)
95% perc	1.15 (2.69)	.94 (.008)	.99 (.06)	.32 (.46)
Evaluating Model 2				
CIC	.91 (.05)	.82 (.04)	.93 (.04)	.98 (.04)
Average(CIC)	.91 (.03)			
d-statistic	-.004 (.11)	-.54 (.27)	.28 (.11)	.25 (.13)
Diff.Medians	-.10 (.12)	-.42 (.14)	.16 (.09)	.52 (.66)
Rej. freq of Diff.Medians	0%	0%	0%	40%
Testing H_0: Model 3 = DGP				
Simulated statistics, Model 3				
5% perc	.18 (.07)	.07 (.18)	.72 (.09)	.024 (.06)
median	.73 (.06)	-.90 (.02)	.99 (.002)	-.97 (.006)
95% perc	1.89 (2.55)	-.999 (.0003)	1 (0)	-.999 (.001)
Evaluating Model 3				
CIC	.25 (.05)	0 (.003)	.06 (.02)	.07 (.02)
Average(CIC)	.097 (.014)			
d-statistic	.64 (.15)	-19 (.17)	2.14 (.05)	-5 (.04)
Diff.Medians	-1.25 (.56)	-27.36 (.36)	2.11 (.008)	-4.52 (.03)
Rej. freq of Diff.Medians	54%	100%	100%	100%

Medians (standard deviations) across 100 Monte Carlo replications of summary statistics of the simulated distributions and of DIW model evaluation statistics (CIC, *d*-statistic and related). See text.

Table 8: **Parameter distributions for the CDN methodology**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor (α)	U[0.5,0.75]	U[0.5,0.75]	U[0.5,0.75]
Growth rate (θ_z)	N(1.0036,0.001)	N(1.0036,0.001)	N(1.0036,0.001)
Depreciation Rate of Capital (δ_K)	U[0.02,0.03]	U[0.02,0.03]	U[0.02,0.03]
Discount Factor (β)	TruncN[0.9855,1.002]	TruncN[0.9855,1.002]	TruncN[0.9855,1.002]
St.St. Hours ($\bar{H}$)	U[0.2,0.35]	U[0.2,0.35]	U[0.2,0.35]
Risk Aversion (σ)	Truncχ^2(2)[0,10]	Truncχ^2(2)[0,10]	Truncχ^2(2)[0,10]
Share of G (sg)	U[0.2,0.3]	U[0.2,0.3]	U[0.2,0.3]
Persistence of Tech. Disturbances (ρ_A)	N(0.9,0.2)	N(0.9,0.2)	0
Persistence of G Disturbances (ρ_G)	0	U[0.95,0.9999]	U[0.95,0.9999]
Std of Technology Innovations (σ_A)	Truncχ^2(1)[0,0.0202]	Truncχ^2(1)[0,0.0202]	0
Std of G Innovations (σ_G)	0	Truncχ^2(1)[0,0.01]	Truncχ^2(1)[0,0.01]

Table 9: **Canova and De Nicoló methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
US Data				
5% perc	.60	.72	.83	-.33
median	.76	.89	.93	.15
95% perc	.98	.96	.97	.57
Simulated statistics, Model 1				
5% perc	.47	.48	.47	-.38
median	.65	.82	.79	.29
95% perc	.91	.94	.95	.72
Evaluating Model 1				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	77	77	94	35
90% one-sided C.I.	96	96	99	75
95% one-sided C.I.	98	98	99	83
90% two-sided C.I.	63	78	40	76
95% two-sided C.I.	74	84	47	85
% of actual statistics into simulated distributions'				
50% one-sided C.I.	13	22	2	69
90% one-sided C.I.	78	74	35	98
95% one-sided C.I.	89	86	55	99
90% two-sided C.I.	89	85	55	96
95% two-sided C.I.	93	93	65	98

Table 10: **Canova and De Nicoló methodology (cont.)**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
Simulated statistics, Model 2				
5% perc	.39	-.87	.50	-.94
median	.63	.76	.82	.11
95% perc	.93	.93	.98	.67
Evaluating Model 2				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	78	84	84	53
90% one-sided C.I.	95	98	91	82
95% one-sided C.I.	97	99	92	87
90% two-sided C.I.	55	56	40	65
95% two-sided C.I.	66	62	47	74
% of actual statistics into simulated distributions'				
50% one-sided C.I.	9	8	4	44
90% one-sided C.I.	78	61	83	96
95% one-sided C.I.	91	79	99	98
90% two-sided C.I.	90	79	99	98
95% two-sided C.I.	95	87	100	99
Simulated statistics, Model 3				
5% perc	.28	-.99	.96	-.99
median	.54	-.94	.99	-.98
95% perc	.96	-.71	1	-.86
Evaluating Model 3				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	81	100	2	100
90% one-sided C.I.	92	100	6	100
95% one-sided C.I.	96	100	8	100
90% two-sided C.I.	37	0	8	0
95% two-sided C.I.	45	0	12	0
% of actual statistics into simulated distributions'				
50% one-sided C.I.	1	0	100	0
90% one-sided C.I.	84	0	100	0
95% one-sided C.I.	94	0	100	0.1
90% two-sided C.I.	94	0	14	0.1
95% two-sided C.I.	96	0	31	0.1

Table 11: **Monte Carlo on the CDN methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)	
Testing H_0: Model 1 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	51 (5.1)	50 (5)	50 (5.1)	50 (5)	50
90% one-sided C.I.	90 (3.1)	91 (2.8)	90 (3.1)	90 (3)	90
95% one-sided C.I.	94 (2.4)	96 (2)	95 (2)	95 (2)	95
90% two-sided C.I.	88 (3.3)	91 (2.8)	90 (3)	90 (3)	90
95% two-sided C.I.	94 (2.5)	96 (2)	94 (2)	95 (2.2)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	49 (5.4)	50 (5)	50.5(5.2)	50 (4.8)	50
90% one-sided C.I.	90 (3.7)	88.5(3.3)	90.5(2.7)	90 (3)	90
95% one-sided C.I.	96 (2.5)	95 (2.3)	95 (2.1)	94 (2)	95
90% two-sided C.I.	91 (3)	88 (3)	90 (3)	89 (3)	90
95% two-sided C.I.	95 (2.1)	94 (2.6)	95 (2.2)	94 (2.3)	95
Testing H_0: Model 2 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	55 (5)	66 (4.8)	42 (4.9)	64 (4.7)	50
90% one-sided C.I.	89 (3.1)	94 (2.2)	79 (3.9)	93 (2.6)	90
95% one-sided C.I.	93 (2.5)	97 (1.7)	85 (3.5)	97 (1.7)	95
90% two-sided C.I.	81 (3.9)	73 (4.5)	81 (3.9)	75 (4.3)	90
95% two-sided C.I.	88 (3.2)	79 (4.1)	85 (3.6)	80 (3.9)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	44 (6)	28 (5)	58 (5)	33 (5.2)	50
90% one-sided C.I.	91 (4)	83 (3.3)	99 (1.7)	85 (4)	90
95% one-sided C.I.	97 (2)	92(2.3)	100(.4)	93 (3)	95
90% two-sided C.I.	95.5(2.4)	92 (3)	93.4(3.2)	93 (3.1)	90
95% two-sided C.I.	98 (1.3)	95.5(2.6)	97.3(2.2)	95.5(2.2)	95
Testing H_0: Model 3 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	66.5(5)	100 (0)	0 (.3)	100 (0)	50
90% one-sided C.I.	89 (3)	100 (0)	1 (.9)	100 (0)	90
95% one-sided C.I.	93 (2.6)	100 (0)	1 (1.2)	100 (0)	95
90% two-sided C.I.	55 (5.1)	0 (0)	1 (1.2)	0 (.1)	90
95% two-sided C.I.	63 (5)	0 (.04)	2 (1.5)	0 (.2)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	17 (7.6)	0 (0)	100 (0)	0 (0)	50
90% one-sided C.I.	92 (4)	0 (0)	100 (0)	0 (0)	90
95% one-sided C.I.	97 (2)	0 (0)	100 (0)	0 (.01)	95
90% two-sided C.I.	97 (2)	0 (0)	0.9(0.9)	0 (.01)	90
95% two-sided C.I.	98.7(1.2)	0 (0)	2 (4)	0 (.08)	95

Medians (S.D.) across 100 Monte Carlo replications of the CDN measures of percentage overlap between the distributions of actual and model statistics.

Table 12: **Monte Carlo on the spectral density distance methodology**

	$fit(\omega_1;\gamma)$	$fit(\omega_2;\gamma)$	$fit([\omega_1,\omega_2];\gamma)$
Testing H_0: Model 1 = DGP			
% rejection (90% C.I.)	0%	0%	0%
% rejection (95% C.I.)	0%	0%	0%
5% perc	1.55	1.47	31.45
median	3.42	3.29	52.8
90% perc	7.19	5.3	73.79
95% perc	7.56	5.43	91.72
mean	3.84	3.50	53.61
S.D.	1.99	1.48	17.73
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6
Testing H_0: Model 2 = DGP			
% rejection (90% C.I.)	0%	0%	0%
% rejection (95% C.I.)	0%	0%	0%
5% perc	2.23	2.67	49.09
median	5.8	5.16	76.7
90% perc	9.1	7.4	117.5
95% perc	10.6	8.13	123.78
mean	6.09	5.18	81.26
S.D.	2.76	1.96	26.6
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6
Testing H_0: Model 3 = DGP			
% rejection (90% C.I.)	100%	100%	100%
% rejection (95% C.I.)	100%	100%	100%
5% perc	338.3	407.4	5721.1
median	351.5	418.3	5836.2
90% perc	364	428	5908
95% perc	367.7	430	5921
mean	352	418.6	5834
S.D.	9.84	7.17	61.08
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6

Summary statistics of the empirical distribution across 100 Monte Carlo replications of the *fit* test statistics for frequencies associated to cycles 8 years long (ω_1), 2 years long (ω_2) and for averages across business cycle frequencies ($[\omega_1,\omega_2]$ interval).

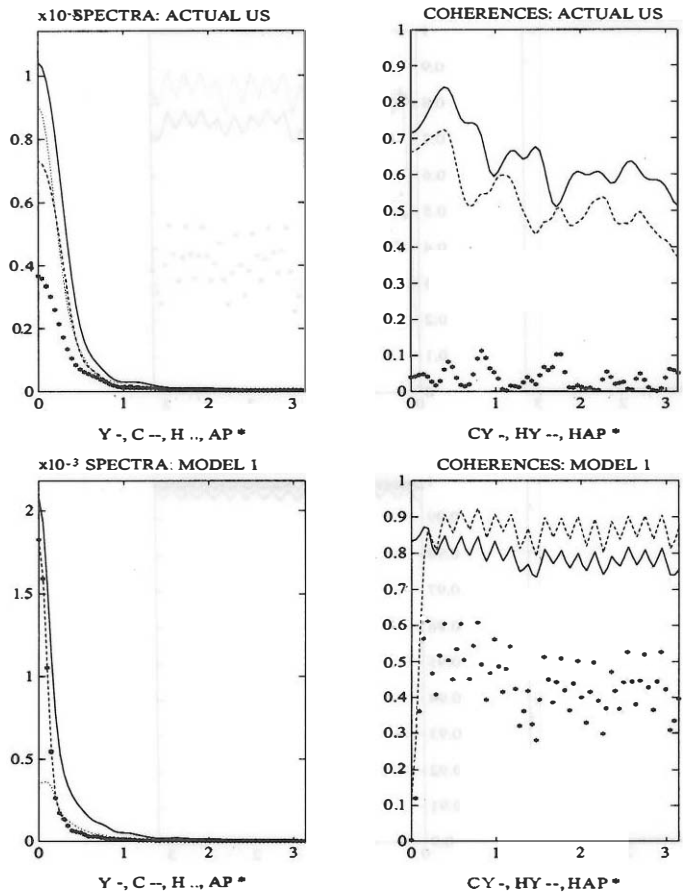

Figure 1: Spectra of Y(–),C(- -),H(..) and AP(*) and coherences of C,Y(–), H,Y(- -) and H,AP(*). **Actual US data** in the upper plots, simulated data from **Model 1** in the lower plots. A linear trend has been extracted from all series.

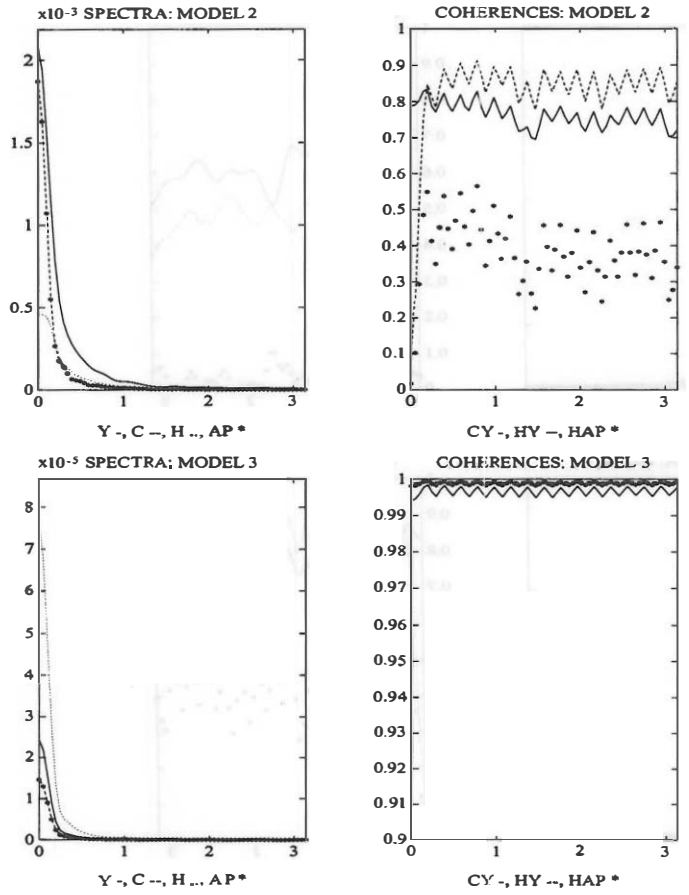

Figure 2: Spectra of Y(–),C(- -),H(..) and AP(*) and coherences of C,Y(–), H,Y(- -) and H,AP(*). Simulated data from **Model 2** in the upper plots, simulated data from **Model 3** in the lower plots. A linear trend has been extracted from all series.

WORKING PAPERS (1)

9525 **Aurora Alejano y Juan M.ª Peñalosa:** La integración financiera de la economía española: efectos sobre los mercados financieros y la política monetaria.

9526 **Ramón Gómez Salvador y Juan J. Dolado:** Creación y destrucción de empleo en España: un análisis descriptivo con datos de la CBBE.

9527 **Santiago Fernández de Lis y Javier Santillán:** Regímenes cambiarios e integración monetaria en Europa.

9528 **Gabriel Quirós:** Mercados financieros alemanes.

9529 **Juan Ayuso Huertas:** Is there a trade-off between exchange rate risk and interest rate risk? (The Spanish original of this publication has the same number.)

9530 **Fernando Restoy:** Determinantes de la curva de rendimientos: hipótesis expectacional y primas de riesgo.

9531 **Juan Ayuso and María Pérez Jurado:** Devaluations and depreciation expectations in the EMS.

9532 **Paul Schulstad and Ángel Serrat:** An Empirical Examination of a Multilateral Target Zone Model.

9601 **Juan Ayuso, Soledad Núñez and María Pérez-Jurado:** Volatility in Spanish financial markets: The recent experience.

9602 **Javier Andrés e Ignacio Hernando:** ¿Cómo afecta la inflación al crecimiento económico? Evidencia para los países de la OCDE.

9603 **Barbara Dluhosch:** On the fate of newcomers in the European Union: Lessons from the Spanish experience.

9604 **Santiago Fernández de Lis:** Classifications of Central Banks by Autonomy: A comparative analysis.

9605 **M.ª Cruz Manzano Frías y Sofía Galmés Belmonte:** Credit Institutions' Price Policies and Type of Customer: Impact on the Monetary Transmission Mechanism. (The Spanish original of this publication has the same number.)

9606 **Malte Krüger:** Speculation, Hedging and Intermediation in the Foreign Exchange Market.

9607 **Agustín Maravall:** Short-Term Analysis of Macroeconomic Time Series.

9608 **Agustín Maravall and Christophe Planas:** Estimation Error and the Specification of Unobserved Component Models.

9609 **Agustin Maravall:** Unobserved Components in Economic Time Series.

9610 **Matthew B. Canzoneri, Behzad Diba and Gwen Eudey:** Trends in European Productivity and Real Exchange Rates.

9611 **Francisco Alonso, Jorge Martínez Pagés y María Pérez Jurado:** Weighted Monetary Aggregates: an Empirical Approach. (The Spanish original of this publication has the same number.)

9612 **Agustín Maravall and Daniel Peña:** Missing Observations and Additive Outliers in Time Series Models.

9613 **Juan Ayuso and Juan L. Vega:** An empirical analysis of the peseta´s exchange rate dynamics.

9614 **Juan Ayuso Huertas:** Un análisis empírico de los tipos de interés reales *ex-ante* en España.

9615 **Enrique Alberola Ila:** Optimal exchange rate targets and macroeconomic stabilization.

9616 **A. Jorge Padilla, Samuel Bentolila and Juan J. Dolado:** Wage bargaining in industries with market power.

9617 **Juan J. Dolado and Francesc Marmol:** Efficient estimation of cointegrating relationships among higher order and fractionally integrated processes.

9618 **Juan J. Dolado y Ramón Gómez:** La relación entre vacantes y desempleo en España: perturbaciones agregadas y de reasignación.

9619 **Alberto Cabrero and Juan Carlos Delrieu:** Construction of a composite indicator for predicting inflation in Spain. (The Spanish original of this publication has the same number.)

9620 **Una-Louise Bell:** Adjustment costs, uncertainty and employment inertia.

9621 **M.ª de los Llanos Matea y Ana Valentina Regil:** Indicadores de inflación a corto plazo.

9622 **James Conklin:** Computing value correspondences for repeated games with state variables.

9623 **James Conklin:** The theory of sovereign debt and Spain under Philip II.

9624 **José Viñals and Juan F. Jimeno:** Monetary Union and European unemployment.

9625 **María Jesús Nieto Carol:** Central and Eastern European Financial Systems: Towards integration in the European Union.

9626 **Matthew B. Canzoneri, Javier Vallés and José Viñals:** Do exchange rates move to address international macroeconomic imbalances?

9627 **Enrique Alberola Ila:** Integración económica y unión monetaria: el contraste entre Norteamérica y Europa.

9628 **Víctor Gómez and Agustín Maravall:** Programs TRAMO and SEATS.

9629 **Javier Andrés, Ricardo Mestre y Javier Vallés:** Un modelo estructural para el análisis del mecanismo de transmisión monetaria: el caso español.

9630 **Francisco Alonso y Juan Ayuso:** Una estimación de las primas de riesgo por inflación en el caso español.

9631 **Javier Santillán:** Política cambiaria y autonomía del Banco Central.

9632 **Marcial Suárez:** Vocábula (Notas sobre usos lingüísticos).

9633 **Juan Ayuso and J. David López-Salido:** What does consumption tell us about inflation expectations and real interest rates?

9701 **Víctor Gómez, Agustín Maravall and Daniel Peña:** Missing observations in ARIMA models: Skipping strategy versus outlier approach.

9702 **José Ramón Martínez Resano:** Los contratos DIFF y el tipo de cambio.

9703 **Gabriel Quirós Romero:** Una valoración comparativa del mercado español de deuda pública.

9704 **Agustín Maravall:** Two discussions on new seasonal adjustment methods.

9705 **J. David López-Salido y Pilar Velilla:** La dinámica de los márgenes en España (Una primera aproximación con datos agregados).

9706 **Javier Andrés and Ignacio Hernando:** Does inflation harm economic growth? Evidence for the OECD.

9707 **Marga Peeters:** Does demand and price uncertainty affect Belgian and Spanish corporate investment?

9708 **Jeffrey Franks:** Labor market policies and unemployment dynamics in Spain.

9709 **José Ramón Martínez Resano:** Los mercados de derivados y el euro.

9710 **Juan Ayuso and J. David López-Salido:** Are *ex-post* real interest rates a good proxy for *ex-ante* real rates? An international comparison within a CCAPM framework.

9711 **Ana Buisán y Miguel Pérez:** Un indicador de gasto en construcción para la economía española.

9712 **Juan J. Dolado, J. David López-Salido and Juan Luis Vega:** Spanish unemployment and inflation persistence: Are there phillips trade-offs?

9713 **José M. González Mínguez:** The balance-sheet transmission channel of monetary policy: The cases of Germany and Spain.

9714 **Olympia Bover:** Cambios en la composición del empleo y actividad laboral femenina.

9715 **Francisco de Castro and Alfonso Novales:** The joint dynamics of spot and forward exchange rates.

9716 **Juan Carlos Caballero, Jorge Martínez y M.ª Teresa Sastre:** La utilización de los índices de condiciones monetarias desde la perspectiva de un banco central.

9717 **José Viñals y Juan F. Jimeno:** El mercado de trabajo español y la Unión Económica y Monetaria Europea.

9718 **Samuel Bentolila:** La inmovilidad del trabajo en las regiones españolas.

9719 **Enrique Alberola, Juan Ayuso and J. David López-Salido:** When may peseta depreciations fuel inflation?

9720 **José M. González Mínguez:** The back calculation of nominal historical series after the introduction of the european currency (An application to the GDP).

9721 **Una-Louise Bell:** A Comparative Analysis of the Aggregate Matching Process in France, Great Britain and Spain.

9722 **Francisco Alonso Sánchez, Juan Ayuso Huertas y Jorge Martínez Pagés:** El poder predictivo de los tipos de interés sobre la tasa de inflación española.

9723 **Isabel Argimón, Concha Artola y José Manuel González-Páramo:** Empresa pública y empresa privada: titularidad y eficiencia relativa.

9724 **Enrique Alberola and Pierfederico Asdrubali:** How do countries smooth regional disturbances? Risksharing in Spain: 1973-1993.

9725 **Enrique Alberola, José Manuel Marqués and Alicia Sanchís:** Unemployment persistence, Central Bank independence and inflation performance in the OECD countries. (The Spanish original of this publication has the same number.)

9726 **Francisco Alonso, Juan Ayuso and Jorge Martinez Pagés:** How informative are financial asset prices in Spain?

9727 **Javier Andrés, Ricardo Mestre and Javier Vallés:** Monetary policy and exchange rate dynamics in the Spanish economy.

9728 **Juan J. Dolado, José M. González-Páramo y José Viñals:** A cost-benefit analysis of going from low inflation to price stability in Spain.

9801 **Ángel Estrada, Pilar García Perea, Alberto Urtasun y Jesús Briones:** Indicadores de precios, costes y márgenes en las diversas ramas productivas.

9802 **Pilar Álvarez Canal:** Evolución de la banca extranjera en el período 1992-1996.

9803 **Ángel Estrada y Alberto Urtasun:** Cuantificación de expectativas a partir de las encuestas de opinión.

9804 **Soyoung Kim:** Monetary Policy Rules and Business Cycles.

9805 **Víctor Gómez and Agustín Maravall:** Guide for using the programs TRAMO and SEATS.

9806 **Javier Andrés, Ignacio Hernando and J. David López-Salido:** Disinflation, output and unemployment: the case of Spain.

9807 **Olympia Bover, Pilar García-Perea and Pedro Portugal:** A comparative study of the Portuguese and Spanish labour markets.

9808 **Víctor Gómez and Agustín Maravall:** Automatic modeling methods for univariate series.

9809 **Víctor Gómez and Agustín Maravall:** Seasonal adjustment and signal extraction in economic time series.

9810 **Pablo Hernández de Cos e Ignacio Hernando:** El crédito comercial en las empresas manufactureras españolas.

9811 **Soyoung Kim:** Identifying European Monetary Policy Interactions: French and Spanish System with German Variables.

9812 **Juan Ayuso, Roberto Blanco y Alicia Sanchís:** Una clasificación por riesgo de los fondos de inversión españoles.

9813 **José Viñals:** The retreat of inflation and the making of monetary policy: where do we stand?

9814 **Juan Ayuso, Graciela L. Kaminsky and David López-Salido:** A switching-regime model for the Spanish inflation: 1962-1997.

9815 **Roberto Blanco:** Transmisión de información y volatilidad entre el mercado de futuros sobre el índice Ibex 35 y el mercado al contado.

9816 **M.ª Cruz Manzano e Isabel Sánchez:** Indicadores de expectativas sobre los tipos de interés a corto plazo. La información contenida en el mercado de opciones.

9817 **Alberto Cabrero, José Luis Escrivá, Emilio Muñoz and Juan Peñalosa:** The controllability of a monetary aggregate in EMU.

9818 **José M. González Mínguez y Javier Santillán Fraile:** El papel del euro en el Sistema Monetario Internacional.

9819 **Eva Ortega:** The Spanish business cycle and its relationship to Europe.

9820 **Eva Ortega:** Comparing Evaluation Methodologies for Stochastic Dynamic General Equilibrium Models.

---

(1) Previously published Working Papers are listed in the Banco de España publications catalogue.

Queries should be addressed to: Banco de España
Sección de Publicaciones. Negociado de Distribución y Gestión
Telephone: 91 338 5180
Alcalá, 50. 28014 Madrid