Macroeconomic Dynamics, **19**, 2015, 167–188. Printed in the United States of America.
doi:10.1017/S1365100513000321

ARE UNIT ROOT TESTS USEFUL IN THE DEBATE OVER THE (NON)STATIONARITY OF HOURS WORKED?

Amélie Charles
Audencia Nantes School of Management

Olivier Darné
LEMNA, University of Nantes

Fabien Tripier
CLERSE, University of Lille 1
and
CEPII

The performance of unit root tests on simulated series is compared, using the business-cycle model of Chang et al. [*Journal of Money, Credit and Banking* 39(6), 1357–1373 (2007)] as a data-generating process. Overall, Monte Carlo simulations show that the efficient unit root tests of Ng and Perron (NP) [*Econometrica* 69(6), 1519–1554 (2001)] are more powerful than the standard unit root tests. These efficient tests are frequently able (i) to reject the unit-root hypothesis on simulated series, using the best specification of the business-cycle model found by Chang et al., in which hours worked are stationary with adjustment costs, and (ii) to reduce the gap between the theoretical impulse response functions and those estimated with a Structural VAR model. The results of Monte Carlo simulations show that the hump-shaped behavior of data can explain the divergence between unit root tests.

Keywords: Unit Root Test, DSGE Models, Hours Worked, Structural VAR

1. INTRODUCTION

Economists use econometrics to identify key statistical properties of the data, which are afterward incorporated into theoretical models. For an econometric tool to be useful for this purpose, it must pass a "natural economic test"[1]: it should be possible to reidentify the statistical properties of the data that were identified by this econometric tool when the theoretical model is used as the data-generating process (DGP hereafter). Theoretical models are widely used as DGPs by researchers on the business cycle to assess the performance of econometric methods, as in Erceg

We would like to thank the editor, William A. Barnett, and the anonymous referee for the constructive comments that improved the paper. The usual disclaimers apply. Address correspondence to: Olivier Darné, IEMN-IAE, Chemin de la censive du Terte, BP 52231, 44322 Nantes Cedex 3, France; e-mail: olivier.darne@univ-nantes.fr.

et al. (2005) and Chari et al. (2008) for structural VAR, in Lindé (2005) to compare the full information maximum likelihood approach and the generalized method of moments, in An and Schorfheide (2007) for Bayesian methods, in Canova and Sala (2009) for methods based on impulse response functions, and in Gorodnichenko and Ng (2010) for methods of moments. To the best of our knowledge, models of the business cycle have not yet been used to assess the performance of unit root tests.

Applications of unit root tests to financial and macroeconomics series have challenged conventional economic theory and stimulated the development of new theories in numerous fields, such as economic fluctuations [Nelson and Plosser (1982)].[2] The debate over the stationarity of hours worked was sparked by Gali's (1999) results on the effects of technological shocks,[3] which contradict the technology-driven business cycle theory. Gali's (1999) results are based on a structural VAR (SVAR hereafter) model à la Blanchard and Quah (1989) that uses the hours worked in first difference. Gali (1999) motivates this specification by appealing to the outcome of standard augmented Dickey and Fuller (1981) (ADF hereafter) tests.[4] Among the responses to Gali's (1999) findings,[5] Christiano et al. (2004) obtained results opposite to those of Gali (1999) by using the hours worked in level, and not the first difference of the series in the SVAR. Like Gali (1999), Christiano et al. (2004) motivate the specification of the SVAR by the outcome of a stationarity test. Whelan (2009) also obtained results that contradict those obtained by Gali (1999) using different tests and data. In response to these mixed results, one strand of the literature suggests abandoning the use of standard unit root tests.[6] These results can be explained by a well-known shortcoming of unit root tests,[7] which is that the properties of unit root tests are generally weak for the sample size of typical macroeconomic time series (about 100–200 observations), such as the hours worked series.

Previous studies on the stationarity of hours worked suffer from two further drawbacks. First, they consider few and relatively "old" standard unit root tests (especially ADF) and do not include the recent developments of efficient unit root tests, especially those of Elliott et al. (1996) (ERS hereafter) and Ng and Perron (2001) (NP hereafter). These efficient tests allow the elimination of the deterministic components that are included in the test regression of the standard unit root tests (a constant mean in the hours worked series) to bring about a gain in efficiency of the unit root tests by increasing their power [Schmidt and Phillips (1992)]. Second, when several tests are used, their performances are not compared in the business-cycle model framework. However, if observed data are viewed as one realization of an economic model, it is essential that the unit root tests used perform well when this economic model is used to generate data.[8] To show the usefulness of the unit root tests in the debate over (non)stationarity of hours worked, we compare the performance of several tests (ADF, ERS, and NP) using a business cycle model to generate data.[9]

Here, we adopt the model proposed by Chang et al. (2007), which has several important attractive features. It (i) allows either stationary or nonstationary hours worked, (ii) considers whether there are adjustment costs of labor, and (iii) has been estimated with Bayesian methods to account for certain facts about the business cycle that pertain to output and labor. We use the four model specifications estimated by Chang et al. (2007) to assess the sensitivity of test performances to the choice of the DGP. For each specification, we simulate the model for various sample sizes (100, 200, 500, 1,000) and evaluate the size and power properties of the various unit root tests.

We show that the performance of the unit root tests is very sensitive to the specification of the model, i.e., the structure of shocks as well as the existence of adjustment costs. Even if the ADF and NP tests give similar (incorrect) properties for the DGP with stationary hours and no adjustment costs of labor, the NP tests dominate the ADF test when the adjustment of labor is costly. This result indicates the need to assess the performance of tests rigorously before applying them to observed data. It also raises the issue of how to specify the model, given the effect that the specification can have on the evaluation of tests. In the model of Chang et al. (2007), adjustment costs are a powerful propagation mechanism that induces hump-shaped responses of hours worked to shocks, with a quicker return to the steady state level. Monte Carlo simulations show a similar difference in performance between ADF and NP unit root tests for ARMA processes with hump-shaped behavior. Because adjustment costs are widely supported by quantitative macroeconomic studies, notably by Chang et al. (2007), these results lead us to prefer the model specification with adjustment costs and therefore to recommend the NP tests rather than the ADF test. Therefore, we investigate the implications of specifying the model in this way for the SVAR methodology.

The SVAR methodology has been discussed extensively in the literature [e.g., Faust and Leeper (1997); Cooley and Dwyer (1998)] and criticized for its inability to identify the correct impulse response functions (IRFs hereafter) when a business cycle model is DGP [e.g., Ercerg et al. (2004); Dupaigne et al. (2007); Ravenna (2007); Chari et al. (2008)]. Chari et al. (2008) demonstrate that the bias in the estimated IRFs is larger when the VAR is specified with hours worked in first difference rather than in level of the series. Our contribution is to improve the specification of the VAR in the SVAR methodology. To demonstrate our improvement in the specification, we simulate output and hours series with a small sample size (200 quarters) for a specification of the model that uses stationary hours and labor adjustment costs. This specification is held to be more consistent with the empirical facts than other specifications [Chang et al. (2007)]. We apply unit root tests to series of simulated hours worked. Then, and depending on the outcomes of tests, we specify an empirical VAR in first difference or in level to estimate IRFs using the long-run restrictions. The NP tests indicate that hours worked are stationary more frequently than the ADF test; hence, the empirical VAR is more frequently specified in level and the estimated bias of IRFs is smaller when the NP tests are used, rather than the ADF test.

Finally, we compare the coverage ratios[10] of the IRF of hours to the technological shocks using our pretest procedure and the agnostic procedure proposed

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

by Pesavento and Rossi (2006), which is based on a SVAR in quasi-difference. Because the estimated bias is lower for the SVAR in level than in quasi-difference, the coverage ratio turns out to be higher with the pretest procedure than with the agnostic procedure. This finding is in opposition to that of Pesavento and Rossi (2006). This difference can be explained by (i) the DGP, in that like Pesavento and Rossi (2006), we consider a VARMA model, but our is based on an estimated DSGE model; (ii) the structural restrictions being in the short run in Pesavento and Rossi (2006), whereas we consider long-run restrictions.

The remainder of the paper is organized as follows. Section 2 describes the methodology that we use. Section 3 presents the results, and discusses (i) the effect of the results on the persistence mechanism and hump-shaped behavior and (ii) the implications of the results for SVAR methodology. Section 4 concludes.

2. METHODOLOGY

This section presents the models that are used to generate data, the unit root tests, the SVAR methodology, and the Monte Carlo study.

2.1. Business-Cycle Models as a Data-Generating Process

We now describe the model briefly and present the various specifications that are suggested by Chang et al. (2007) and are used to generate data. The model is real and the economy is perfectly competitive. Households consume, accumulate physical capital, and supply production factors (labor and physical capital) to firms. Households maximize the expected intertemporal utility function

$$\mathrm{E}_t \left\{ \sum_{s=0}^{\infty} \beta^{t+s} \left[\ln C_{t+s} - \frac{\left(H_{t+s}/B_{t+s}\right)^{1+1/\nu}}{1+1/\nu} \right] \right\}, \tag{1}$$

where $0 < \beta < 1$ is the subjective discount factor, ν the Frisch elasticity of labor supply, C_t the household consumption, H_t the household hours worked, B_t a preference shock on the disutility of labor, and t the period. The representative household faces the budget constraint

$$W_t H_t + R_t K_t = C_t + K_{t+1} - (1-\delta) K_t, \tag{2}$$

where δ is the rate at which physical capital depreciates, W_t the wage rate, R_t the rate at which physical capital is rented, and K_t the stock of physical capital held by the household. Firms combine physical capital and labor to produce the final good according to

$$Y_t = \left(A_t H_t^d\right)^{\alpha} \left(K_t^d\right)^{1-\alpha} \left[1 - \varphi \left(\frac{H_t^d}{H_{t-1}^d} - 1 \right)^2 \right], \tag{3}$$

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

where $0 < \alpha < 1$ is the elasticity parameter of the production function, A_t is the technological shock common to all firms, H_t^d and K_t^d are the demand for inputs, and $\varphi \geq 0$ measures the size of the adjustment costs of labor. The model description is closed with the shock processes

$$\ln A_t = \gamma + \ln A_{t-1} + \varepsilon_{a,t}, \;\; \varepsilon_{a,t} \sim \text{iid}\,(0, \sigma_a), \tag{4}$$

$$\ln B_t = \rho_b \ln B_{t-1} + (1 - \rho_b) \ln B_0 + \varepsilon_{b,t}, \;\; \varepsilon_{b,t} \sim \text{iid}\,(0, \sigma_b), \tag{5}$$

where $\gamma > 0$ is the deterministic component of the drift of technological shocks and $0 < \rho_b \leq 1$ denotes the persistence of shocks to the household's utility function.

2.2. The Unit Root Tests

Dickey and Fuller (1981) developed the ADF unit root test for testing the hypothesis that a univariate time series contains a unit root against the alternative hypothesis that it is level stationary or trend stationary. For our case of interest, a constant mean in the hours worked series [Gali (1999); Whelan (2009)], the test regression is defined by

$$\Delta y_t = \alpha + \beta_0 y_{t-1} + \sum_{j=1}^{k} \beta_j \Delta y_{t-j} + \varepsilon_t, \tag{6}$$

where $\{\varepsilon_t\}$ is a sequence of independent normal random variables with mean zero and variance σ^2; i.e., $\varepsilon_t \sim \text{IN}(0; \sigma^2)$. The ADF t-test is performed by testing the null hypothesis $\beta_0 = 0$ against the alternative $\beta_0 < 0$.

Some studies show that the elimination of deterministic components (here the constant mean) may result in the unit root tests being more efficient by increasing their power. ERS develop a unit root test based on a quasi-difference detrending of the series. They suggest using the Dickey–Fuller generalized least squares (DF-GLS) test with the regression

$$\Delta \tilde{y}_t = \beta_0 \tilde{y}_{t-1} + \sum_{j=1}^{k} \beta_j \Delta \tilde{y}_{t-j} + \varepsilon_t, \tag{7}$$

where $\tilde{y}_t$ is the locally detrended series y_t. The DF-GLS t-test is performed by testing the null hypothesis $\beta_0 = 0$ against the alternative $\beta_0 < 0$. The local detrending series is defined by

$$\tilde{y}_t = y_t - \hat{\psi}' z_t,$$

where z_t is equal to 1 for the constant mean case, and $\hat{\psi}'$ is the GLS estimator obtained by regressing $\tilde{y}$ on $\hat{z}$, where $\tilde{y} = (y_1, (1 - \bar{\alpha} B)\, y_2, \ldots, (1 - \bar{\alpha} B)\, y_T)$, $\bar{z} = (z_1, (1 - \bar{\alpha} B)\, z_2, \ldots, (1 - \bar{\alpha} B)\, z_T)'$, and $\bar{\alpha} = 1 + c/T$. ERS recommend using $\bar{c} = -7$ for the constant mean case. They also consider a point-optimal

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

test of the unit root null hypothesis $\alpha = 1$ against the alternative $\alpha = \bar{\alpha}$ (see the Appendix).

NP also propose efficient unit root tests based on the regression (7). Their tests, called M-GLS tests (see the Appendix), are modifications of the Phillips and Perron (1988) test, which is a nonparametric approach to correcting residual autocorrelation by modifying the Dickey–Fuller test statistics: first, correcting the size distortions [as suggested by Perron and Ng (1996)], and second, improving the power [as suggested by Elliott et al. (1996)].

2.3. The SVAR Methdolology

Chari et al. (2007), among others, provide a general description of the SVAR methodology with long-run restrictions. Stationary data X_t are described by the following empirical VAR with p lags:

$$X_t = B_1 X_{t-1} + B_2 X_{t-2} + \cdots B_p X_{t-p} + v_t. \tag{8}$$

v_t are the canonical innovations, with $E v_t v_t' = \Omega$, and B_i are matrices of autoregressive coefficients for $i = 1, \ldots, p$. Equation (8) is inverted to get the Wold decomposition

$$X_t = v_t + C_1 v_{t-1} + C_2 v_{t-2} + \cdots, \tag{9}$$

where the C's satisfy $I = (I - B_1 L - B_2 L^2 \ldots - B_p L^p)(I + C_1 L + C_2 L^2 + \ldots)$ for all values of L. The model with structural innovations is defined as follows:

$$X_t = A_0 \epsilon_t + A_1 \epsilon_{t-1} + A_2 \epsilon_{t-2} + \cdots, \tag{10}$$

with $A_0 \epsilon_t = v_t$ and $A_j = C_j A_0$, $j \geq 1$. The identifying restrictions for SVAR are $E \epsilon_t \epsilon_t' = I$ and the (1,1) element of $\sum_{j=0}^{\infty} A_j$, or equivalently, $[\sum_{j=0}^{\infty} C_j] A_0$, is equal to 0. This gives a system of four equations and four unknowns.

2.4. The Monte Carlo Design

The model is calibrated using the outcome of the estimations of Chang et al. (2007, Table 2, p. 1366) for the four specifications given in Table 1. The model is simulated using the programs provided by the authors.[11] All experiments are based on 30,000 replications. We consider separately each specification of the model that is used to generate data.

1. The specifications of the DSGE described in Table 1 are used to generate simulated macroeconomic data of length T. The sample sizes considered are $T = 100$, 200, 500, and 1,000.
2. Unit root tests are applied to simulated data for hours worked to compute their properties. We base the choice of lag length on the sequential procedure proposed by Ng and Perron (1995) for the ADF test and we use the modified Akaike information criteria suggested by Ng and Perron (2001) for efficient unit root tests.[12] The observed unit-root test statistics are compared to their finite-sample 5% critical values given

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

TABLE 1. Specifications of the DGPs

No.	Specification	Parameter values
1	Stationary hours worked	$\alpha = 0.652$; $\beta = 0.995$; $\gamma = 0.004$; $\delta = 0.023$; $\nu = 0.527$; $\rho_B = 0.951$
	Without adjustment cost	$\alpha_A = 0.011$; $\sigma_B = 0.006$; $\ln A_0 = 5.708$; $\ln B_0 = 3.176$; $\varphi = 0$
2	No stationary hours worked	$\alpha = 0.654$; $\beta = 0.995$; $\gamma = 0.004$; $\delta = 0.024$; $\nu = 0.474$; $\rho_B = 1.000$
	Without adjustment cost	$\alpha_A = 0.011$; $\sigma_B = 0.006$; $\ln A_0 = 5.717$; $\ln B_0 = 3.166$; $\varphi = 0$
3	Stationary hours worked	$\alpha = 0.658$; $\beta = 0.995$; $\gamma = 0.004$; $\delta = 0.023$; $\nu = 0.433$; $\rho_B = 0.800$
	With adjustment cost	$\alpha_A = 0.011$; $\sigma_B = 0.034$; $\ln A_0 = 5.748$; $\ln B_0 = 3.171$; $\varphi = 11.36$
4	No stationary hours worked	$\alpha = 0.661$; $\beta = 0.995$; $\gamma = 0.004$; $\delta = 0.024$; $\nu = 1.153$; $\rho_B = 1.000$
	With adjustment cost	$\alpha_A = 0.011$; $\sigma_B = 0.012$; $\ln A_0 = 5.754$; $\ln B_0 = 3.194$; $\varphi = 8.054$

Source: Table 2 of Chang et al. (2007).

in (i) the original papers on the unit root tests, (ii) MacKinnon (1991) and Vougas (2007) for the small finite sample, and (iii) our computations.

3. Simulated data from DSGE models for output and hours are used to estimate SVAR with long-run restrictions. If the test indicates that hours are stationary, the hours series is introduced in level in the SVAR; otherwise, the hours in first difference is introduced. For each test, we compute the moments of IRFs.

3. RESULTS

We now present the results for the performance of unit root tests and the SVAR predictions from the Monte Carlo experiments in Section 3.1 and discuss the role of persistence mechanisms in Section 3.2. An illustration with observed data is provided in Section 3.3. Finally, Section 3.4 shows the implications for SVAR methodology.

3.1. The Performance of Unit Root Tests

Table 2 displays the results for the DGP where hours worked are stationary. Table 3 reports the results for the DGPs where the hours worked are nonstationary, without (Panel A) and with (Panel B) adjustment costs. The power of unit root tests is given in Table 2 and the size of unit root tests is presented in Table 3. For the DGPs with nonstationary hours worked (Table 3), the unit root tests show good size, whatever the sample sizes, and without and with adjustment costs.

TABLE 2. Reject rates of unit-root test statistics—DGP: stationary hours worked

Sample	MZ_α	MZ_t	DF-GLS	PT	MPT	ADF
			Panel A: Without adjustment costs (1)			
$T = 1{,}000$	0.9326	0.9281	0.9341	0.9242	0.9316	1.0000
$T = 500$	0.7984	0.8014	0.8320	0.7762	0.7953	0.9601
$T = 200$	0.4503	0.4580	0.4089	0.3952	0.4299	0.3688
$T = 100$	0.1859	0.1605	0.1460	0.1432	0.1643	0.1326
			Panel B: With adjustment costs (3)			
$T = 1{,}000$	0.9196	0.9154	0.9314	0.9125	0.9202	1.000
$T = 500$	0.8579	0.8620	0.8659	0.8422	0.8577	0.8642
$T = 200$	0.7286	0.7326	0.6821	0.6776	0.7148	0.6535
$T = 100$	0.5168	0.4874	0.3977	0.4369	0.4894	0.3535

Notes: (1) and (3) denote specifications 1 and 3 in Table 1. MZ_α, MZ_t, and MPT denote the Ng and Perron (2001) tests; DF-GLS and PT denote the Elliot et al. (1996) tests; and ADF denotes the augmented Dickey and Fuller (1981) test.

For the DGPs with stationary hours worked (Table 2), the major issue concerns stationary hours worked when the samples are small ($T = 100$ and 200). Such sample sizes are typical for macroeconomic series. In this case, significant differences appear between tests and interestingly also between model specifications. Overall, the NP tests (MZ_α, MZ_t, and MPT) exhibit higher power than the other tests that we studied, particularly the ADF test, but with some differences according to the model specification. For the model without adjustment costs (Panel A) with $T = 200$, the NP tests reject the unit root hypothesis at a rate of 45% (especially for MZ_α and MZ_t) against 37% for the ADF test ($T = 200$) and 40% for ERS

TABLE 3. Reject rates of unit-root test statistics—DGP: nonstationary hours worked

Sample	MZ_α	MZ_t	DF-GLS	PT	MPT	ADF
			Panel A: Without adjustment costs (2)			
$T = 1{,}000$	0.0656	0.0615	0.0683	0.0639	0.0649	0.0620
$T = 500$	0.0642	0.0649	0.0658	0.0598	0.0624	0.0568
$T = 200$	0.0599	0.0603	0.0491	0.0486	0.0535	0.0590
$T = 100$	0.0550	0.0455	0.0403	0.0401	0.0466	0.0549
			Panel B: With adjustment costs (4)			
$T = 1{,}000$	0.0541	0.0501	0.0536	0.0530	0.0534	0.0390
$T = 500$	0.0547	0.0559	0.0535	0.0515	0.0538	0.0391
$T = 200$	0.0615	0.0616	0.0424	0.0500	0.0550	0.0389
$T = 100$	0.0722	0.0614	0.0390	0.0528	0.0609	0.0446

Notes: (2) and (4) denote specifications 2 and 4 in Table 1. MZ_α, MZ_t, and MPT denote the Ng and Perron (2001) tests; DF-GLS and PT denote the Elliot et al. (1996) tests; and ADF denotes the augmented Dickey and Fuller (1981) test.

tests (DF-GLS and PT). This slight difference does not warrant a preference for the NP tests over the standard unit root test. The conclusion is different for the model with adjustment costs (Panel B). In this case, the NP tests reject the unit root hypothesis at a higher rate in small samples, and improve the ADF test to almost 10% for $T = 200$ (71% against 65%, respectively) and 20% for $T = 100$ (48% against 35%, respectively). Note that the ERS tests are slightly less powerful than the NP tests.[13]

In light of the foregoing, it would seem that the efficient unit root tests, especially the NP tests, are more powerful than the standard unit root test. This indicates that the NP tests should be preferred to the ADF test in this framework, given that the model with adjustment costs is more consistent with empirical facts than the model without adjustment costs, as shown by Chang et al. (2007).

3.2. The Persistence Mechanisms

How well a test performs, given the specification of the model, is a function of the amplification and propagation mechanisms of the model in question. Adjustment costs are well known to propagate the effects of shocks in the economy. Agents smooth the adjustment of labor to reduce total costs. Given that adjustment costs increase the persistence of shocks in the economy, it is surprising that the NP tests reject the unit root hypothesis more frequently for the DGP with adjustment costs. This result can be explained by the fact that with adjustment costs, shocks to the household's utility function are less persistent ($\rho_b = 0.80$) than without adjustment costs ($\rho_b = 0.95$). If we simultaneously considered very persistent shocks to the household's utility function (i.e., $\rho_b = 0.95$) and the persistence induced by labor adjustment costs, the NP tests would fail to reject the unit root hypothesis.[14] However, Chang et al. (2007) show that labor adjustment costs result in a reduction in the persistence of shocks that are due to variations in the supply of labor (measured by ρ_b) in the model.

To clarify this point, we make a distinction between the *endogenous persistence*, associated with adjustment costs, and the *exogenous persistence*, associated with the persistence of the exogenous shocks to the supply of labor. In their procedure for estimating the business cycle model, Chang et al. (2007) proposed that there is an inverse relation between the two forms of persistence. A high value for φ, which measures the size of adjustment costs, is associated with a low value of ρ_b, which measures the autocorrelation of the shocks due to variation in the supply of labor [see Table 1 and Chang et al. (2007, Fig. 2, p. 1367)]. Figure 1 shows the sharp contrast in the IRFs of hours worked between the two specifications (with and without adjustment costs).[15] The model without adjustment costs generates monotonic responses of labor to a stationary supply shock, but these responses last for a very long time, whereas the model with adjustment costs generates hump-shaped responses of labor with a quicker return to the steady-state level. Hump-shaped behavior of series is a major issue in the literature on the business cycle [see, e.g., Cogley and Nason (1995)]. Further, Chang et al. (2007) conclude

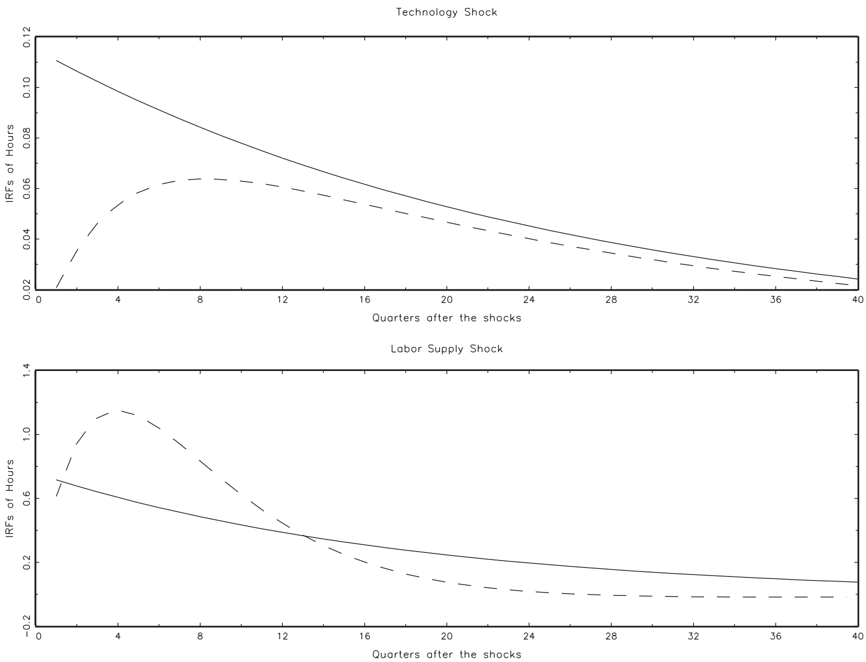

FIGURE 1. IRFs of hours worked to shocks with (dashed lines) and without (solid lines) labor adjustment costs for the model.

that the model with adjustment costs and stationary hours has the best fit among the four specifications. Given these findings, our results suggest that we should use the efficient unit root tests proposed by Ng and Perron (2001) because they are more powerful than the ADF test when simulated series are hump-shaped.

To confirm this intuition concerning the effect of hump-shaped behavior on the ADF test, we performed another Monte Carlo study. We simulated an ARMA(1,2) process with hump-shaped behavior, i.e., $y_t = 0.80y_{t-1} + \varepsilon_t + 0.65\varepsilon_{t-1} + 0.60\varepsilon_{t-2}$, where ε_t is i.i.d.. We also simulated an AR(1) model as a benchmark, defined as $y_t = 0.85y_{t-1} + \varepsilon_t$. The AR model presents the same mean-reversion behavior as the ARMA model. Table 4 gives the power of unit root tests, and Figure 2 plots the IRFs. The sample sizes, the number of replications, and the choices of lag length for the unit root tests are based on the same procedures as were used in the previous Monte Carlo experiment.

For the AR(1) process (Panel A), all the unit root tests show good power, even for small sample sizes. Note that the ADF shows lower power than the efficient unit root tests when $T = 100$. For the ARMA(1,2) process with hump-shaped behavior (Panel B), all the unit root tests have high power for large sample sizes ($T = 1,000$ and 500). When $T = 200$, the efficient unit root tests have good power, whereas the ADF test shows a loss of power (with a rate of rejection of 90% for the NP and ERS tests, against 50% for the ADF test). More interestingly,

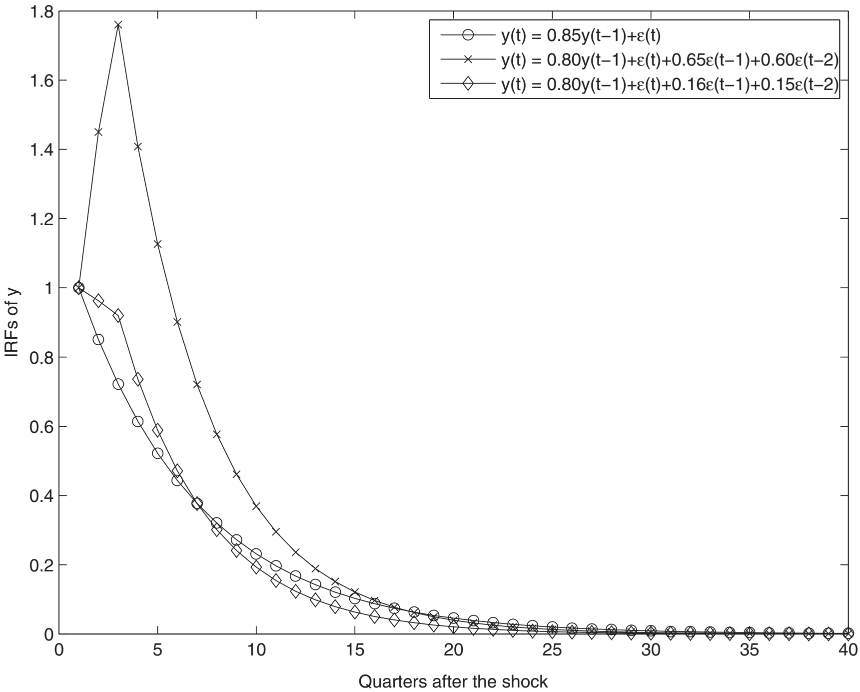

FIGURE 2. IRFs for three ARMA processes.

the ADF test shows low power for $T = 100$, with a rate of rejection of 17% against 60% for the efficient tests. These results show that the ADF test is affected by the hump-shaped behavior, whereas the NP tests have good power. Note that the ADF test is also more biased than the NP tests by the presence of a MA component in the ARMA model without hump-shaped behavior (Panel C).

3.3. Illustration with Observed Data

Our Monte Carlo experiments indicate that there are strong differences between the various unit root tests on the (non)stationarity of hours worked. It is crucial to see whether the results still differ when observed data are used instead of simulated data. To this end, we applied the ERS and NP efficient tests to the three data sets used in Chang et al. (2007) and to the hours series proposed by Francis and Ramey (2009). We obtained results for the (non)stationarity of the hours worked different from those obtained by Chang et al. (2007) using ADF tests (see Table 5). For two of the three series of Chang et al. (2007), the unit root hypothesis is rejected by the efficient unit root tests, whereas this hypothesis is never rejected for the three series according to the ADF test. For the two series of Francis and Ramey (2009), the unit root hypothesis is not rejected, as found by Francis and Ramey (2005, 2009).[16]

TABLE 4. Reject rates of unit-root test statistics—DGP: AR(1) and ARMA(1,2) models

Sample	MZ_α	MZ_t	DF-GLS	PT	MPT	ADF
			Panel A: AR(1) with $\rho = 0.85$			
$T = 1{,}000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 500$	0.9998	0.9997	0.9999	0.9996	0.9996	1.0000
$T = 200$	0.9490	0.9507	0.9496	0.9430	0.9440	0.9958
$T = 100$	0.8033	0.7767	0.7528	0.7682	0.7756	0.6154
			Panel B: ARMA(1,2) with $\rho = 0.80$, $\theta_1 = 0.65$, and $\theta_2 = 0.60$			
$T = 1{,}000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 500$	0.9999	0.9998	1.0000	0.9998	0.9998	0.9943
$T = 200$	0.9433	0.9442	0.9424	0.9335	0.9350	0.5069
$T = 100$	0.6627	0.6291	0.5512	0.6266	0.6330	0.1698
			Panel C: ARMA(1,2) with $\rho = 0.80$, $\theta_1 = 0.16$ and $\theta_2 = 0.15$			
$T = 1{,}000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 500$	1.0000	1.0000	1.0000	0.9998	0.9998	0.9956
$T = 200$	0.9628	0.9633	0.9662	0.9585	0.9592	0.5885
$T = 100$	0.8460	0.8146	0.7936	0.8070	0.8119	0.2393

Notes: The AR(1) model is defined as $y_t = \rho y_{t-1} + \varepsilon_t$, and the ARMA(1,2) model as $y_t = \rho y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$, with ε_t i.i.d. MZ_α, MZ_t, and MPT denote the Ng and Perron (2001) tests; DF-GLS and PT denote the Elliot et al. (1996) tests; and ADF denotes the augmented Dickey and Fuller (1981) test.

TABLE 5. Results of unit root tests on hours worked series

Data set	MZ_α	MZ_t	MPT	DF-GLS	PT	k^a	ADF^b	k^b
1	−12.40*	−2.47*	2.05*	−2.48*	2.09*	1	−2.80	4
2	−3.65	−1.34	6.71	−1.42	7.93	1	−2.55	4
3	−11.20*	−2.34*	2.30*	−2.43*	2.29*	1	−2.44	4
4	−0.43	−0.24	20.54	−1.53	21.94	11	−2.52	4
5	−3.48	−1.31	7.04	−2.27*	7.38	9	−3.38	4
Critical value at the 5% level	−8.10	−1.98	3.17	−1.98	3.17		−2.86	

Notes: * indicates rejection of the unit-root null hypothesis at the 5% level of significance. [a]the lag order k in the regression is selected by using the modified information criteria (MIC) proposed by Ng and Perron (2001). [b]the values of the ADF tests and lag order k are taken from Chang et al. (2007, footnote 7, p. 1363). The first three data sets were collected by Chang et al. (2007). Data set 1 was constructed by the Bureau of Labor Statistics and corresponds to the average weekly hours of all people in the nonfarm business sector. Data set 2 was constructed by Christiano et al. (2004) (LBMN, DRI-Global Insight). Data set 3 was constructed by Gali and Rabanal (2004) and corresponds to nonfarm business sector hours (LXNFH, Haver Analytics' USECON). Data sets 4 and 5 were constructed by Frances and Ramey (2009) and correspond to unadjusted and adjusted average weekly hours, respectively. MZ_α, MZ_t, and MPT denote the Ng and Perron (2001) tests; DF-GLS and PT denote the Elliot et al. (1996) tests; and ADF denotes the augmented Dickey and Fuller (1981) test.

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

3.4. Implications for SVAR Methodology

We now derive and present the implications of our results for the SVAR methodology with long-run restrictions. We restrict our attention to the model specification that has the best agreement with the empirical facts [Chang et al. (2007)], i.e., the specification with stationary hours and labor adjustment costs (specification 3) and with a small sample size ($N = 200$).

IRF population moments. The SVAR has been criticized by Chari et al. (2008) for being unable to provide useful predictions for business cycle theory. To show the restrictive feature of this method, they simulate a DSGE model, apply the SVAR methodology to simulated series, and show that for realistic sample sizes, the estimated IRFs are far from the actual IRFs. This leads the authors to conclude that the SVAR methodology has no practical application in business cycle theory.

We performed an exercise that is similar to that of Chari et al. (2008), with one major exception: we used the outcome of unit root tests to choose the SVAR specification (in level or in first difference). If the unit root test indicated that hours are stationary, the SVAR was specified with the hours series in level. Otherwise, if the test indicated that hours are not stationary, the SVAR was specified with the hours series in first difference. We compared the moments of IRFs according to the unit root test used: NP[17] or ADF. For each test, Figure 3 reports the true IRFs and the median IRFs, and Figure 4 displays the RMSE between the estimated IRFs and the true IRF.

The medians of population IRFs show significant differences according to the unit root test that is used. Several features of the theoretical IRFs are replicated when the NP test is used instead of the ADF test. First, the median IRF of output in response to the no-technological-shock condition is hump-shaped for the NP test, as in the DSGE model, whereas it is monotonically decreasing for the ADF test. Second, the median IRF of hours worked in response to the no-technological-shock condition is positive and hump-shaped for the NP test, as in the DSGE model, whereas it is negative for the ADF test. Third, the median IRF of hours worked in response to the technological shock returns to zero in the long run for the NP test, as in the DSGE model, whereas it is highly positive for the ADF test.

These findings show the advantages for the SVAR methodology of using efficient unit root tests, such as the NP test, rather than the standard ADF test. Nevertheless, there are also dimensions for which the results of the NP test are less satisfactory. For the NP test, the median IRF of hours worked in response to the technological shock is overestimated for the first quarters after the shocks and the long-run median IRF of output in response to the technological shock is underestimated. This last point is the only one for which the median IRF is closest to the DSGE model IRF when the ADF test is used, rather than the NP test. These conclusions are confirmed by Figure 4, which shows that the RMSE is always lower with the NP test than with the ADF test, except for the long-term IRF of output to the technological shock.

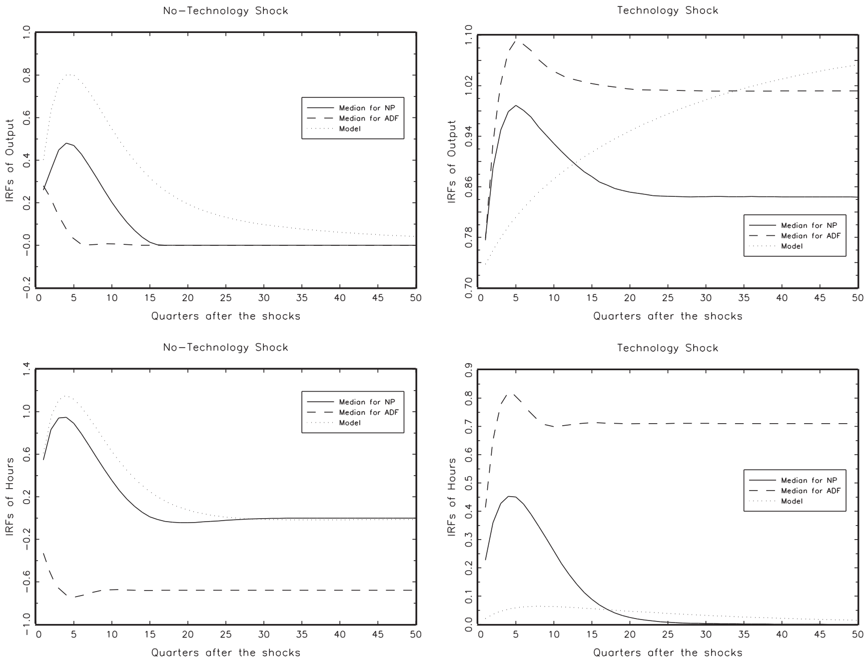

FIGURE 3. Model and SVAR responses of output and hours to no-technological and technological shocks. For SVAR, solid lines are the medians of IRFs using the NP test and dashed lines with squares using the ADF test. Dotted lines are for the model.

A comparison with agnostic procedures. This section compares our approach with the agnostic procedure of Pesavento and Rossi (2005, 2006). Pesavento and Rossi (2005, 2006) recommend not using procedures based on a pretest of series that determines the specification of the SVAR. Instead, they recommend estimating a quasi-difference SVAR and using unit root tests to construct confidence intervals for the IRFs associated with the SVAR in quasi-difference. Pesavento and Rossi (2006) demonstrate the higher performance of agnostic procedures than of pretest procedures for SVAR based on short-run identification assumptions (a VARMA model is used as a DGP to generate artificial data). Pesavento and Rossi (2005) apply this agnostic procedure to an empirical SVAR (historical output and hours worked) based on long-run identification assumptions and not on short-run identification assumptions. Therefore, Pesavento and Rossi (2005, 2006) do not provide a comparison of agnostic and pretest procedures in the case of SVAR with long-run restrictions. We use our DGP to provide such a comparison.

Table 6 reports the coverage ratios for the agnostic and the pretest procedures using several unit root tests (ADF, ERS, and NP). The coverage ratios are higher

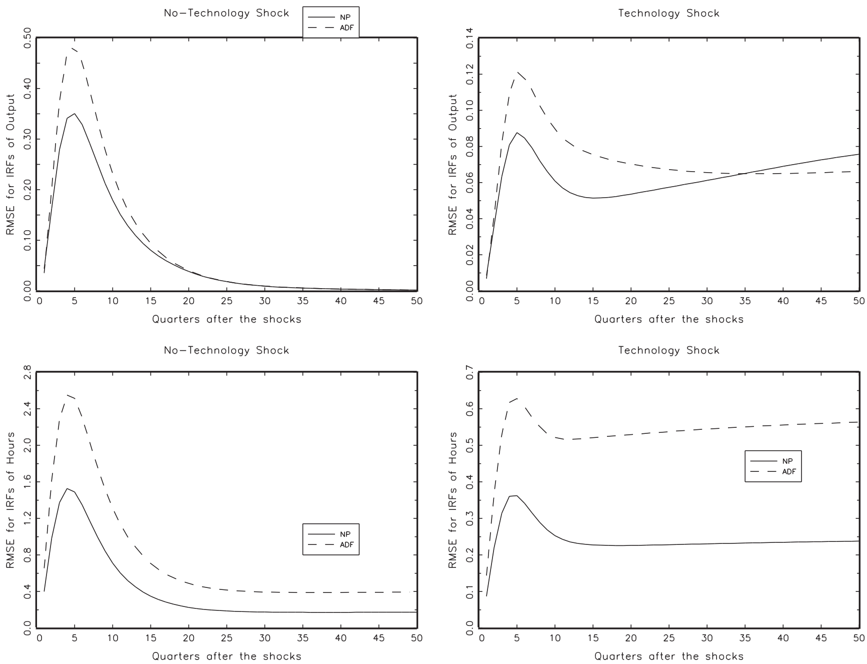

FIGURE 4. RMSE between the model and the SVAR responses of output and hours to no-technological and technological shocks. For SVAR, solid lines are the RMSEs of IRFs using the NP test and dashed lines using the ADF test.

for pretest procedures, especially with the NP test, than for agnostic procedures. The agnostic approach fails because it relies on estimation of the SVAR in quasi-difference, which strongly overestimates the IRF of hours to a technological shock. To highlight this property, we report in Figures 5 and 6 the median IRFs and the RMSE of IRFs for the three SVAR: in level, in first difference, and in quasi-difference. The estimated IRFs of hours to the technological shocks are far above the true IRF for the first difference and the quasi-difference cases. Therefore, the confidence interval rarely includes the true IRF and the coverage rate turns to be very low, especially for the first horizons. The estimated IRF is less overestimated for the specification in level than for the SVAR either in quasi-difference or in first difference. In fact, the coverage ratio of the SVAR in level is quite high. Because this specification is frequently selected when the NP test is used in the pretest procedure, the coverage ratio is higher for the pretest procedure than for the agnostic procedure (Table 6). It is worth mentioning that if the confidence intervals associated with a SVAR in level often include the true response, they also frequently include the zero value (in 61.4% of simulations), making it difficult to identify significantly the positive response of hours to a technological shock assumed in the DGP.

TABLE 6. Coverage ratios of IRFs of hours to technological shocks

Lag	Diff	Level	SR ADF	SR ERS	SR NP	PRE NP	PRE ADF	PRE ERS
1.000	0.001	0.728	0.000	0.000	0.000	0.522	0.507	0.500
2.000	0.001	0.740	0.001	0.001	0.001	0.530	0.515	0.508
3.000	0.001	0.748	0.001	0.004	0.003	0.536	0.521	0.512
4.000	0.001	0.750	0.007	0.014	0.012	0.540	0.520	0.516
5.000	0.002	0.754	0.107	0.155	0.126	0.541	0.524	0.518
6.000	0.005	0.775	0.310	0.431	0.346	0.558	0.537	0.534
7.000	0.011	0.800	0.320	0.442	0.355	0.575	0.556	0.550
8.000	0.013	0.839	0.328	0.452	0.364	0.605	0.582	0.579
9.000	0.012	0.876	0.341	0.461	0.372	0.631	0.607	0.604
10.000	0.011	0.909	0.355	0.466	0.379	0.657	0.628	0.629
11.000	0.011	0.953	0.390	0.482	0.390	0.686	0.655	0.656
12.000	0.010	0.971	0.425	0.501	0.413	0.703	0.665	0.672
13.000	0.011	0.984	0.485	0.525	0.441	0.709	0.671	0.679
14.000	0.011	0.989	0.566	0.550	0.469	0.714	0.673	0.683
15.000	0.011	0.988	0.625	0.579	0.497	0.714	0.671	0.683
16.000	0.011	0.988	0.688	0.611	0.525	0.714	0.671	0.683
17.000	0.011	0.987	0.744	0.634	0.556	0.712	0.669	0.681
18.000	0.011	0.981	0.785	0.662	0.581	0.709	0.663	0.678
19.000	0.011	0.979	0.825	0.677	0.603	0.709	0.662	0.678
20.000	0.010	0.974	0.856	0.694	0.625	0.706	0.658	0.676
21.000	0.010	0.970	0.876	0.715	0.644	0.703	0.654	0.673
22.000	0.010	0.972	0.896	0.731	0.656	0.703	0.656	0.673
23.000	0.010	0.970	0.903	0.747	0.677	0.703	0.655	0.673
24.000	0.010	0.966	0.915	0.758	0.692	0.701	0.653	0.670
25.000	0.010	0.962	0.925	0.768	0.704	0.696	0.649	0.666
26.000	0.010	0.956	0.929	0.777	0.717	0.694	0.643	0.663
27.000	0.010	0.951	0.935	0.786	0.732	0.691	0.639	0.661
28.000	0.010	0.945	0.936	0.794	0.742	0.687	0.634	0.657
29.000	0.009	0.940	0.939	0.800	0.752	0.684	0.630	0.654
30.000	0.009	0.933	0.942	0.808	0.760	0.680	0.625	0.650
31.000	0.009	0.923	0.942	0.814	0.767	0.673	0.618	0.644
32.000	0.009	0.918	0.940	0.816	0.773	0.671	0.614	0.642
33.000	0.009	0.914	0.941	0.819	0.775	0.669	0.611	0.640
34.000	0.009	0.909	0.940	0.822	0.776	0.666	0.606	0.638
35.000	0.009	0.905	0.938	0.827	0.780	0.663	0.603	0.635
36.000	0.009	0.903	0.938	0.830	0.786	0.661	0.600	0.633
37.000	0.008	0.896	0.938	0.832	0.790	0.657	0.594	0.630
38.000	0.008	0.893	0.937	0.833	0.793	0.656	0.592	0.629

Notes: Lags denotes the lags of IRFs; Level and Diff denote the SVAR in level and in first difference, respectively; SR ADF, SR ERS, and SR NP denote the agnostic procedure of Pesavento and Rossi (2006) from the ADF, ERS, and NP test statistics, respectively. PRE ADF, PRE ERS, and PRE NP denote the pretest procedures from the ADF, ERS, and NP unit root tests, respectively. ADF, ERS, and NP denote the augmented Dickey and Fuller (1981) test, the PT test of Elliot et al. (1996), and the MPT test of Ng and Perron (2001), respectively.

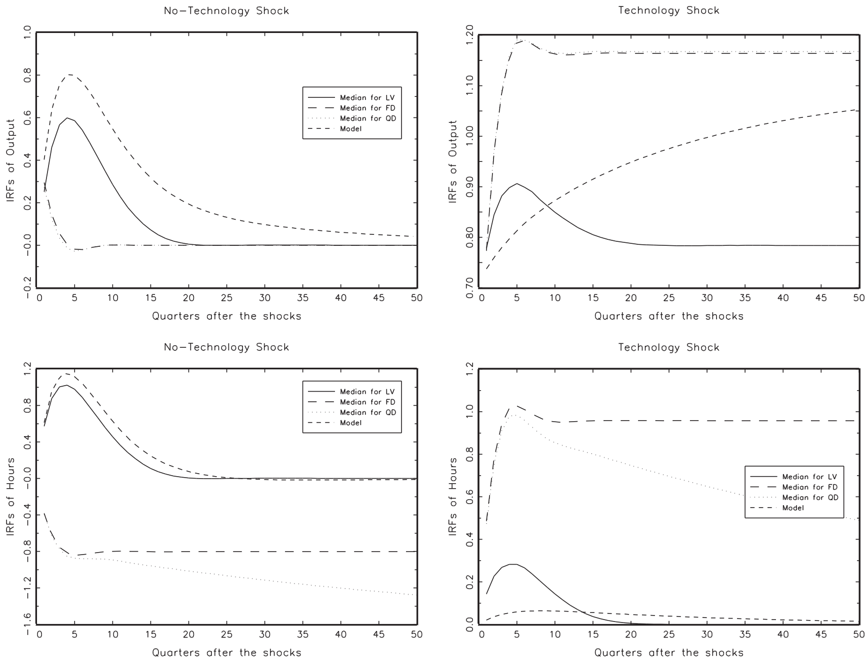

FIGURE 5. Model and SVAR responses of output and hours to no-technological and technological shocks. Solid lines are the medians of IRFs using the SVAR in level, long dashed lines using the SVAR in first difference, dotted lines using the SVAR in quasi-difference, and short dashed lines are for the true IRF.

4. CONCLUDING REMARKS

The mixed results of the unit root tests on the (non)stationarity of hours worked cast doubt on how far they can be useful for developing business cycle theory. In the work reported herein, we attempted to improve the contribution of unit root tests to economic theory by linking the process by which the quality of the tests is assessed to economic theory. From Monte Carlo simulations using data generated by a well-specified business cycle model, namely the Chang et al. (2007) model with labor adjustment costs, we showed that the efficient unit root tests proposed by Ng and Perron (2001) are more powerful than the standard ADF unit root test. This result can be explained by the fact that the labor adjustment costs generate hump-shaped behavior and reduce the persistence of shocks to the household's utility function. The effect of hump-shaped behavior on the ADF test is confirmed from Monte Carlo experiments on ARMA models. This finding suggests that the Ng and Perron tests should be preferred in this framework. Furthermore, we found that using the NP tests, rather than the ADF test, to choose the SVAR specification (in level or in first difference) for the hours worked narrows the gap between the theoretical IRFs and those estimated with a SVAR model.

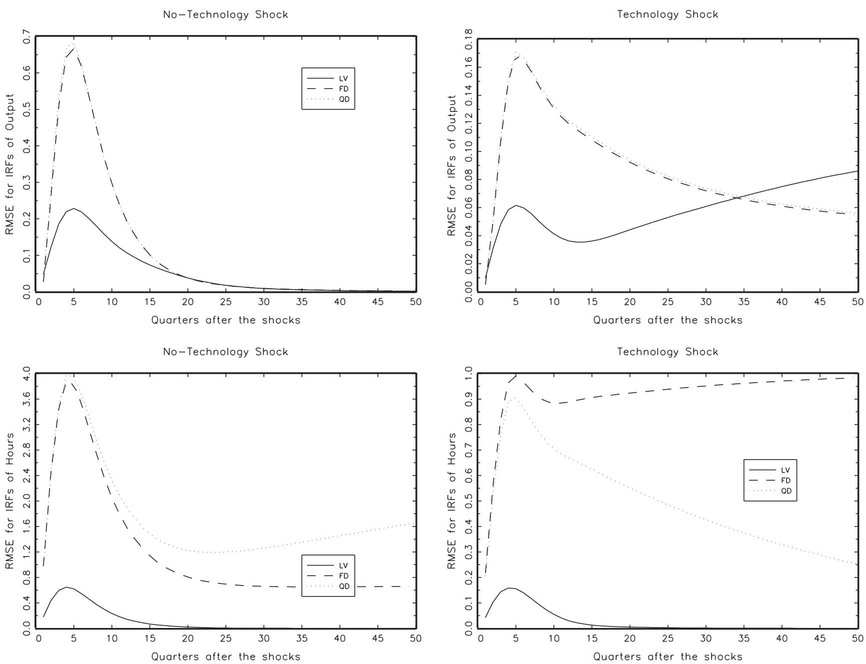

FIGURE 6. RMSE between the model and the SVAR responses of output and hours to no-technological and technological shocks. Solid lines are the medians of IRFs using the SVAR in level, dashed lines using the SVAR in first difference, and dotted lines using the SVAR in quasi-difference.

Naturally, our results remain specific to the choice of the DSGE model and our analysis could be conducted for other DSGE models than that of Chang et al. (2007), and for other series than hours. The key message of our paper is the interest of using theoretical models to generate artificial data and assess the performances of statistical tests applied to observed series, such as unit root tests. In further research, it should be interesting to study the exchange rate, for which stationarity is highly debated from both theoretical and empirical perspectives and numerous theoretical models have been developed.

NOTES

1. This expression is borrowed from Chari et al. (2008), who apply this “natural economic test” to the methodology of structural VAR with long-run restriction.
2. For example, the detection of a unit root in output by Nelson and Plosser (1982) legitimated the development of business cycle models with very persistent or nonstationary shocks to factors’ productivity. The first generation of real business cycle models considered a very persistent autoregressive process for the technological shock; see Kydland and Prescott (1982), Hansen (1985), and Prescott (1986). The effects of technological shocks have been modeled as a random walk, generally in multiple-shocks models as in King et al. (1991) and Christiano and Eichenbaum (1992). See Hansen (1997) for a discussion of this issue.

https://doi.org/10.1017/S1365100513000321 Published online by Cambridge University Press

3. Gali (1999) concludes that technological shocks play a minor role in the business cycle and that a positive technological shock induces a decrease in the number of hours worked.

4. Gali and Rabanal (2004) extend the set of tests to the KPSS test and confirm the findings of Gali (1999).

5. For example, Francis and Ramey (2005) and Fout and Francis (in press) develop real business cycle models consistent with a negative response of employment to a positive technological shock, whereas Chari et al. (2008) argue that SVAR is useless for developing business cycle theory.

6. Pesavento and Rossi (2006) propose an agnostic procedure using approximation based on local-to-unity asymptotic theory to overcome the choice between hours worked in first difference or in level; see Pesavento and Rossi (2005) for an application of this procedure to the effects of technological shocks. Gil-Alana and Moreno (2009) propose a method in a fractional integration framework that is also agnostic with respect to the order of integration of the variables. Fève and Guay (2009) suggest using a more clearly stationary variable in the SVAR instead of hours, namely the ratio of consumption to output, and show how to recover the responses of hours to shocks in a second step, independent of the specification of the series (in level or in first difference).

7. This shortcoming has been addressed by Campbell and Perron (1991), DeJong et al. (1992), and Haldrup and Jansson (2006).

8. For example, one issue with standard unit root tests as used in Chari et al. (2008) is that they are unable to reject the hypothesis that the hours series has a unit root, whereas the hours series in the model is highly persistent, but stationary.

9. Other unit root tests have been developed to overcome the limitations of the standard unit root tests, such as the presence of structural breaks [e.g., Perron (1989); Zivot and Andrews (1992)] or the presence of nonlinearity [e.g., Enders and Granger (1998); Caner and Hansen (2001)]. We do not use these tests because the DGPs do not show breaks and/or nonlinearity.

10. The coverage ratio measures the frequency with which the true IRF is inside the confidence intervals for each horizon.

11. The required programs are dsge.g, dsgemod.src, and dsgesim.src.

12. Ng and Perron (2001) show that the popular Akaike and Schwarz information criteria are not sufficiently flexible for unit root tests to select the appropriate number of lags in the regression (mainly when there are negative moving-average errors).

13. Note that we also consider the stationarity test of Kwiatkowski et al. (1992) in the Monte Carlo experiments. This test shows strong size distortions in small samples, especially for $T = 100$. The results of the KPSS test are available upon request.

14. If we impose $\rho_b = 0.95$ in specification 3 of the model, the rates of rejection are 65% for the NP test and 86% for the ADF ($T = 200$). The complete table is available upon request.

15. The stable roots of the matrix LAMBDA computed in the gensys procedure are -2.22×10^{-016} and 0.79 for specification 0 without adjustment costs and 0.92 and 0.80 for specification 2 with adjustment costs.

16. See Christiano et al. (2004, Figs. 2 and 3), Pesavento and Rossi (2005, Figs. 1 and 2), and Whelan (2009, Figs. 1 and 2), among others, to see how the IRF of hours to technological shocks changes with the specification of the SVAR (in level or in difference) according to the result of the unit root test.

17. In the remainder of the paper, we display the results of the modified point optimal (MPT) test proposed by Ng and Perron (2001). We obtained similar results with the others NP tests, namely the MZ_a and MZ_t test statistics.

REFERENCES

An, Sungbae and Frank Schorfheide (2007) Bayesian analysis of DSGE models. *Econometric Reviews* 26, 113–172.

Blanchard, Olivier J. and Danny Quah (1989) The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–673.

Campbell, John Y. and Pierre Perron (1991) Pitfalls and opportunities: What macroeconomists should know about unit roots. *NBER Macroeconomics Annual* 6, 141–201.

Canova, Fabio and Luca Sala (2009) Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics* 56, 431–449.

Chang, Yongsung, Taeyoung Doh, and Frank Schorfheide (2007) Non-stationary hours in a DSGE model. *Journal of Money, Credit and Banking* 39(6), 1357–1373.

Chari, V.V., Patrick J. Kehoe, and Ellen R. McGrattan (2007) Are Structural VARs with Long-Run Restrictions Useful in Developing Business Cycle Theory? Staff report 364, Federal Reserve Bank of Minneapolis.

Chari, V.V., Patrick J. Kehoe, and Ellen R. McGrattan (2008) Are structural VARs with long-run restrictions useful in developing business cycle theory? *Journal of Monetary Economics* 55, 1337–1352.

Christiano, Larry J. and Martin Eichenbaum (1992) Current real business cycle theories and aggregate labor market fluctuations. *American Economic Review* 82, 430–450.

Christiano, Larry J., Martin Eichenbaum, and Robert Vigfusson (2004) What Happens after a Technology Shock? Working paper 9819, NBER.

Cogley, Timothy and James M. Nason (1995) Output dynamics in real-business-cycle models. *American Economic Review* 84, 492–511.

Cooley, Thomas F. and Mark Dwyer (1998) Business cycle analysis without much theory: A look at structural VARs. *Journal of Econometrics* 83, 57–88.

DeJong, David N., John C. Nankervis, N.E. Savin, and Charles H. Whiteman (1992) The power problems of unit root tests in time series with autoregressive errors. *Journal of Econometrics* 53, 323–343.

Dickey, David A. and Wayne A. Fuller (1981) Likelihood ratio statistics for autoregressive time series with unit root. *Econometrica* 49, 1057–1072.

Dupaigne, Martial, Patrick Fève, and Julien Matheron (2007) Some analytics on bias in DSVARs. *Economics Letters* 97, 32–38.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.

Erceg, Christopher J., Luca Guerrieri, and Christopher Gust (2005) Can long-run restrictions identify technology shocks? *Journal of the European Economic Association* 3, 1237–1278.

Faust, Jon and Eric M. Leeper (1997) When do long-run identifying restrictions give reliable results? *Journal of Business and Economic Statistics* 15, 345–353.

Fève, Patrick and Alain Guay (2009) The response of hours to a technology shock: A two-step structural VAR approach. *Journal of Money, Credit and Banking* 41, 987–1013.

Fout, Hamilton B. and Neville R. Francis (in press) Imperfect transmission of technology shocks and the business cycle consequences. *Macroeconomic Dynamics*.

Francis, Neville and Valerie A. Ramey (2005) Is the technology-driven real business cycle hypothesis dead? Shocks and aggregate fluctuations revisited. *Journal of Monetary Economics* 52, 1379–1399.

Francis, Neville and Valerie A. Ramey (2009). Measures of per capita hours and their implications for the technology-hours debate. *Journal of Money, Credit and Banking* 41, 1071–1097.

Galí, Jordi (1999) Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review* 89, 249–271.

Galí, Jordi and Paul Rabanal (2004) Technology Shocks and Aggregate Fluctuations: How Well Does the Real Business Cycle Model Fit Postwar U.S. Data? Working paper 10636, NBER.

Gil-Alana, Luis and Antonio Moreno (2009) Technology shocks and hours worked: A fractional integration perspective. *Macroeconomic Dynamics* 13(5), 580–604.

Gorodnichenko, Yuriy and Serena Ng (2010) Estimation of DSGE models when the data are persistent. *Journal of Monetary Economics* 57, 325–340.

Haldrup, Niels and Michael Jansson (2006) Improving size and power in unit root testing. In Terence C. Mills and Kerry Patterson (eds.), *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory*, pp. 252–277. Basingstoke, Hants., UK: Palgrave Macmillan.

Hansen, Gary D. (1985) Indivisible labor and the business cycle. *Journal of Monetary Economics* 16, 309–327.

Hansen, Gary D. (1997) Technological progress and aggregate fluctuations. *Journal of Political Economy* 16, 1005–1023.

King, Robert G., Charles I. Plosser, James H. Stock, and Mark W. Watson (1991) Stochastic trends and economic fluctuations. *American Economic Review* 81, 819–940.

Kwiatkowski, Denis, Peter C.B. Phillips, Peter Schmidt and Yongcheol Shin (1992) Testing the null hypothesis of stationary against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–178.

Kydland, Finn E. and Edward C. Prescott (1982) Time to build and aggregate fluctuations. *Econometrica* 50, 1345–1370.

Lindé, Jesper (2005) Estimating New-Keynesian Phillips curves: A full information maximum likelihood approach. *Journal of Monetary Economics* 52, 1135–1149.

MacKinnon, James G. (1991) Critical values for cointegration tests. In Robert F. Engle and Clive W. J. Granger (eds.), *Long-run Economic Relationships, Readings in Cointegration*, pp. 267–276. Oxford, UK: Oxford University Press.

Nelson, Charles R. and Charles I. Plosser (1982) Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10, 139–162.

Ng, Serena and Pierre Perron (1995) Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90, 268–281.

Ng, Serena and Pierre Perron (2001) Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69(6), 1519–1554.

Perron, Pierre and Serena Ng (1996) Useful modifications to unit root tests with dependent errors and their local asymptotic properties. *Review of Economic Studies* 63, 435–465.

Pesavento, Elena and Barbara Rossi (2005) Do technology shocks drive hours up or down? A little evidence from an agnostic procedure. *Macroeconomic Dynamics* 9, 478–488.

Pesavento, Elena and Barbara Rossi (2006) Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics* 21, 1135–1155.

Phillips, Peter C.B. and Pierre Perron (1988) Testing for unit root in time series regression. *Biometrika* 75, 347–353.

Prescott, Edward C. (1986) Theory ahead of business cycle measurement. *Carnegie Rochester Conference Series on Public Policy* 25, 11–44.

Ravenna, Federico (2007) Vector autoregressions and reduced form representations of DSGE models. *Journal of Monetary Economics* 54, 2048–2064.

Vougas, Dimitrios V. (2007) GLS detrending and unit root testing. *Economics Letters* 97, 222–229.

Whelan, Karl T. (2009) Technology shocks and hours worked: Checking for robust conclusions. *Journal of Macroeconomics* 31, 231–239.

APPENDIX: UNIT ROOT TESTS

A.1. POINT-OPTIMAL TEST OF ELLIOTT ET AL.

Elliott et al. (1996) consider a point-optimal test of the unit root null hypothesis $\alpha = 1$ against the alternative $\alpha = \bar{\alpha}$, given by

$$\mathrm{PT} = \left[S(\bar{\alpha}) - \bar{\alpha} S(1)\right] / s_{\mathrm{ar}}^2,$$

where $S(a)$ is given by $(y_a - z_a \psi)'(y_a - z_a \psi)$, and s_{ar} is the autoregressive spectral density estimator of the long-term variance. The value of $\bar{c}$ is chosen so that the asymptotic power of the test is 50% against the local alternative ($\bar{\alpha} = 1 + \bar{c}/T$). ERS advise $\bar{c} = -7$ for the constant-mean case.

A.2. M-GLS TESTS OF NG AND PERRON

The M-GLS tests proposed by Ng and Perron (2001) are defined as

$$\mathrm{MZ}_t = \left(T^{-1}\tilde{y}_T^2 - s_{\mathrm{ar}}^2\right)\left(4 s_{\mathrm{ar}}^2 T^{-2} \sum_{t=1}^{T} \tilde{y}_{t-1}^2\right)^{-1/2},$$

$$\mathrm{MZ}_a = \left(T^{-1}\tilde{y}_T^2 - s_{\mathrm{ar}}^2\right)\left(2 T^{-2} \sum_{t=1}^{T} \tilde{y}_{t-1}^2\right)^{-1},$$

where s_{ar} is the autoregressive spectral density estimator of the long-term variance. NP also consider a modified feasible point-optimal test,

$$\mathrm{MPT} = \left(\bar{c}^2 T^{-2} \sum_{t=1}^{T} \tilde{y}_{t-1}^2 - \bar{c} T^{-1} \tilde{y}_T\right) \Big/ s_{\mathrm{ar}}^2.$$