An identification-robust test for time-varying parameters in the dynamics of energy prices *

Jean-Thomas Bernard † Université Laval

Jean-Marie Dufour ‡ McGill University

Lynda Khalaf § Carleton University

Maral Kichian ¶ Bank of Canada

June 2010

* The authors thank Eva Ortega, three anonymous referees and the Editor Herman K. van Dijk for several helpful suggestions and comments, as well as Robert Pindyck for kindly providing the data set. This work was supported by the Bank of Canada Research Fellowship Program, the Canada Research Chair Program (Econometrics, Université de Montréal, and Environment, Université Laval), the Chair on the Economics of Electrical Energy (Université Laval), the Institut de Finance Mathématique de Montréal (IFM2), the Alexander-von-Humboldt Foundation (Germany), the Canadian Network of Centres of Excellence (program on Mathematics of Information Technology and Complex Systems [MITACS]), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de Recherche sur la Société et la Culture (Québec), and the Fonds de Recherche sur la Nature et les Technologies (Québec). The views in this paper are our own and do not necessarily reflect those of the Bank of Canada.

† Département d'économique and Groupe de Recherche en économie de l'énergie de l'environement et des ressources naturelles (GREEN), Université Laval. Mailing address: Pavillon J.-A.-De Sève, St. Foy, Québec, Canada, G1K 7P4. Tel: (418) 656 5123; FAX: (418) 656 7412. Email: jtber@ecn.ulaval.ca

‡ William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

§ Economics Department, Carleton University, Centre interuniversitaire de recherche en économie quantitative (CIREQ), and Groupe de recherche en économie de l'énergie, de l'environnement et des ressources naturelles (GREEN), Université Laval. Mailing address: Economics Department, Carleton University, Loeb Building 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada. Tel (613) 520-2600-8697; FAX: (613)-520-3906. email: Lynda_Khalaf@carleton.ca

¶ Canadian Economic Analysis Department, Bank of Canada, Ottawa, Canada. Tel: (613) 782 7460; FAX: (613) 782 7163. Email: mkichian@bankofcanada.ca.

ABSTRACT

We test for the presence of time-varying parameters (TVP) in the long-run dynamics of energy prices for oil, natural gas and coal, within a standard class of mean-reverting models. We also propose residual-based diagnostic tests and examine out-of-sample forecasts. In-sample LR tests support the TVP model for coal and gas but not for oil, though companion diagnostics suggest that the model is too restrictive to conclusively fit the data. Out-of-sample analysis suggests a random-walk specification for oil price, and TVP models for both real-time forecasting in the case of gas and long-run forecasting in the case of coal.

Keywords : structural change; time-varying parameter; energy prices; coal; gas; crude oil; unidentified nuisance parameter; exact test; Monte Carlo test; Kalman filter; normality test.

Journal of Economic Literature classification: C22, C52, C53, Q40.

1. Introduction

It is widely recognized that fluctuations in energy prices have important and lasting effects on the economies of industrialized countries.[1] At the same time, interpreting and predicting the behavior of energy prices remain challenging problems. In addition to domestic and international supply as well as demand conditions, non-market-related features (such as regulations, technological advances and geopolitical considerations) are difficult to characterize. Fully articulated structural models are difficult to build and may be unreliable. Instead, analysts have proposed simple time-series reduced forms for various purposes, notably: (i) testing and validating non-renewable resources models such as the Hotelling rule, or (ii) forecasting. The associated literature is very large and statistical support is claimed for many different models.[2] In this paper, we focus on the class of trend models with time-varying parameters (TVP) proposed by Pindyck (1999) to derive long-run forecasts for oil, natural gas and coal prices.

An important feature of Pindyck's models is the inclusion of time-varying trend parameters to reflect alternative assumptions on demand shifts, resource depletion and technological change. Using a simple Hotelling model, Pindyck argues that long-run energy prices should revert to an *unobservable* trending long-run marginal cost, with continuous random changes in the level and slope of the trend. Pindyck further proposed a family of econometric models that integrate the latter feature. Alternative versions of these models were estimated and out-of-sample forecasts were computed, using Kalman filter techniques and annual data from 1870 to 1996 for crude oil and bituminous coal, and from 1919 to 1996 for natural gas. Pindyck's general specifications for the latter data assume a Gaussian AR(1) process for log-prices, with drift and trend, where the drift and trend coefficients follow themselves uncorrelated Gaussian AR(1) processes. Such models are

[1] Work on the linkages between energy prices and financial markets or the macroeconomy is abundant. For critical discussions, see Hamilton (2003), Hamilton and Herrera (2004), Barsky and Kilian (2004), Kilian (2008*b*, 2008*c*, 2008*a*), Kilian (2009), Kilian and Park (2009), and Kilian, Rebucci and Spatafora (2009).

[2] For recent references, statistical results and critical discussions, see Ahrens and Sharma (1997), Berck and Roberts (1996), Cortazar and Naranjo (2006), Cortazar and Schwartz (2003), Gibson and Schwartz (1990), Lee, List and Strazicich (2006), Moshiri and Foroutan (2006), Pindyck (1999, 2001), Postali and Picchetti (2006), Regnier (2007), Sadorsky (2006), Schwartz (1997), Schwartz and Smith (2000), Slade (1982, 1988), Tabak and Cajueiro (2007). For a recent survey related to the Hotelling rule, see Livernois (2009) and the references therein.

parsimonious yet flexible, allowing both random walks with drift and/or changing trend lines where prices revert to a possibly moving mean. The forecast exercises conducted by Pindyck yield mixed results, but on balance, the class of models considered appears to be quite promising.

Pindyck (1999) did not provide statistical tests for the proposed class of models. In particular, the time-varying parameter specification was not tested statistically. It is worth noting that a zero-variance restriction on the processes postulated by Pindyck for the drift and trend coefficients leads to their time invariant counterpart. Given the sample sizes at hand, the decision to use a TVP model, as opposed to a more common autoregressive or fixed coefficient trend model [such as those used by Slade (1982)] may have non-negligible finite-sample statistical consequences. The fact remains that shifts in trends are empirically documented with energy prices.[3]

TVP models, which capture continuous and unpredictable shifts in slopes and trends through random coefficients, have obvious appeal from an economic perspective relative to data-driven change-point methods [see, for example, Lee et al. (2006)]. TVP models however raise computational difficulties. Despite the fact that TVP likelihoods can easily be evaluated using Kalman filtering, such functions are typically ill-behaved for empirically relevant parameter values. Even though sophisticated numerical recipes and global maximizers are readily available, it is well known that maximization may be difficult to achieve in this context. In particular, irregularities can be linked to parameter space regions where the variances of the stochastic coefficients are "small" or "close-to-zero". Unfortunately, if the variances are actually zero, the corresponding t-ratios do not have a regular asymptotic distribution; it is thus hard to assess *how small is small*.[4]

Statistical tests for TVP are particularly challenging, for at least three reasons. First, under the null hypothesis of no parameter variation, the parameters describing the distribution of the TVP processes are not identified. Second, the no-variation hypothesis sets the parameters describing the TVP processes on the boundary of their permissible domain (the so-called "nesting-at-boundary"

[3] The above-cited work provide ample evidence on this feature; see especially Lee et al. (2006) and the recent review of Livernois (2009).

[4] See Stock and Watson (1998) for an early reference to this problem in the special case where the TVP specification follows a unit-root.

problem).[5] The third difficulty stems from the coefficient on the lagged price, which shows up as a nuisance parameter in the no-variation test problem and is subject to usual unit-root type issues.

The above features can cause test sizes to deviate severely from their nominal levels. Usual chi-square critical points can easily lead to spurious rejections even with fairly large data sets, because the regularity conditions underlying classical asymptotics fail. *Identification-robust* TVP tests are as yet unavailable.[6] Furthermore, while the Kalman-filter makes likelihood-based inference possible, if not always tractable, the associated specifications may provide a poor fit or bad forecasts when the underlying parametric assumptions are not compatible with available data. Careful residual-based diagnostics are thus required, in particular to assess departures from normality. Residual-based normality tests which are *robust to estimation effects* are unavailable for TVP models.[7]

This paper makes three main contributions. First, we propose and apply finite-sample tests for TVP within the class of models proposed by Pindyck (1999). We also show - in a Monte Carlo study - that using standard asymptotic critical points for these tests can lead to severe size distortions. Second, we propose finite-sample residual-based tests to assess the underlying normality assumption; we also show that diagnostic tests which sidestep parameter estimation uncertainty have sizes that deviate arbitrarily from their nominal levels. Third, we examine in-sample and out-of-sample fit for the class of models at hand. Specifically, we complement our in-sample analysis with a forecasting exercise. In this case, in addition to a long-run analysis similar to Pindyck's, we consider a

[5] Indeed, setting the coefficients on the autoregressive terms to one and the variances to zero in the processes considered by Pindyck (1999) for the intercept and trend, leads to a constant coefficient autoregressive model with a linear trend and drift. The latter model is a special case of the TVP model under consideration, yet the unit root as well as the zero variances lie on the boundary - rather than the interior - of the model parameter space.

[6] A test is considered robust to identification if its significance level is controlled - at least asymptotically - regardless of identification, that is irrespective of whether identification holds, holds only weakly, or does not hold. See the reviews of Stock, Wright and Yogo (2002) and Dufour (2003). These authors use the term "robust to weak instruments" to designate procedures whose validity is not affected by a set of instruments that does not allow one to identify structural parameters. Since we consider here a setup where instrumental variables are not explicitly required, we shall employ the term "identification-robust" which appears sufficiently general to cover the kind of situation studied in this paper.

[7] A diagnostic test is considered robust to estimation effects if the associated significance level is the same - at least asymptotically - irrespective of whether disturbances or residuals are used to construct the test statistic. See Godfrey (1996, section 2) for an asymptotic definition, and Dufour, Farhat, Gardiol and Khalaf (1998), Dufour and Khalaf (2002), Dufour, Khalaf and Beaulieu (2003), Dufour, Khalaf, Bernard and Genest (2004), Khalaf and Kichian (2005), Bernard, Idoudi, Khalaf and Yelou (2007), and Dufour, Khalaf and Beaulieu (2010), for a finite sample perspective. In this paper, and in contrast with the latter finite sample motivated works, residuals from non-linear models are in question, which raises more pernicious nuisance-parameter dependence problems.

continuously-updated one-step-ahead approach.

The proposed TVP and normality tests rely on exact simulation-based test procedures, applicable – even with small samples – to highly irregular problems for which standard techniques are not valid. Specifically, we apply the *maximized Monte Carlo* (MMC) test technique [Dufour (2006)], which is based on comparing the maximal p-value of the test (over the nuisance parameters that are identified under the null hypothesis, obtained by simulation) with the significance level. Consequently, level control is ensured by construction.[8] Empirically, the proposed tests allow us to select, within the suggested family of models, specifications that are statistically justified for crude oil, coal, and natural gas prices.

In our forecasting exercise, for both long-run and real-time exercises, we produce and analyze forecasts based on model averaging. The usual model selection practice has recently been somewhat outrun by the concept of model averaging.[9] In particular, model averaging seems particularly useful in accounting for breaks and instabilities, a fact to be seriously considered for the problem at hand.[10] Model-averaged approaches to analyze trends in energy prices are rare.[11] To the best of our knowledge, model-averaging has not been considered so far in order to assess the Hotelling rule. While developing a formal inferential procedure to assess out-of-sample fit with and without averaging is beyond the scope of this paper, we believe our analysis for the data set at hand is empirically worthwhile given the absence of relevant results.

Empirical results can be summarized as follows. Our in-sample Likelihood Ratio (LR) testing exercise rejects the fixed-coefficient model in favour of Pindyck's model for coal and gas, though not for oil. However, independence and normality tests suggest that Pindyck's model is too restrictive to conclusively fit the data. Our out-of-sample exercise selects a random-walk specification for long-run and real-time forecasting for oil, and Pindyck's model for both real-time forecasting in the case

[8] When the null distribution of test statistic depends on nuisance parameters, an α-level is guaranteed in finite samples [see Lehmann (1986)] when the largest p-value (over all values of the nuisance parameters consistent with the null hypothesis) is referred to α.

[9] See *e.g.* Hansen (2007, 2008, 2009*a*).

[10] See Clark and McCracken (2009) or Hansen (2009*a*), on models with discrete breaks.

[11] We thank an anonymous referee for pointing this out.

of gas and long-run forecasting in the case of coal. Furthermore, in the case of natural gas price, we also show that accounting for the post eighties market deregulation tends to improve both time invariant and TVP specifications.

In Section 2, we describe the class of proposed models and the test method used. Section 3 documents and discusses our empirical in-sample results. In section 4, we report our forecasting analysis. We conclude in Section 5. The Appendix reports the results of a small Monte Carlo experiment which document the unreliability of usual asymptotic approximations.

2. Model and test methods

Pindyck (1999) considers a basic Hotelling model for a depletable resource produced in a competitive market. Under the assumption of constant marginal cost of extraction c and isoelastic demand with unitary elasticity, the price level is given by

$$P_t = c + [(ce^{rt}/(e^{rcR_0/A} - 1)] \tag{2.1}$$

where R_0 is the initial stock of the depletable resource, A is a demand shifter, and r is the interest rate. This implies that the slope of the price trajectory is given by

$$dP_t/dt = rce^{rt}/(e^{rcR_0/A} - 1)\,, \tag{2.2}$$

so changes in demand, extraction costs, and reserves all affect this slope. For example, an increase in A causes the slope to increase, while increases in c or R_0 reduce the slope. In addition, increases in c or A raise the price level, whereas an increase in R_0 leads to a decrease in this level. If, as Pindyck (1999) argues, these factors fluctuate in a continuous and unpredictable manner over time, then long-run energy prices should revert to a trend which itself fluctuates in the same fashion.

A class of models which integrates the above features is the generalized Ornstein-Uhlenbeck process. Pindyck (1999) proposes a discretized version of this model as a suitable econometric framework for analyzing long-run energy prices.[12] This leads to the following AR(1)-type dynamic model:

$$P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t\,, \quad t = 1,\,\ldots\,,T\,, \tag{2.3}$$

where P_t refers to the logarithm of the real price of an energy product and the coefficients ϕ_{1t} and ϕ_{2t} follow the stochastic processes

$$\phi_{1t} = c_3 \phi_{1,t-1} + v_{1t}\,, \tag{2.4}$$

$$\phi_{2t} = c_4 \phi_{2,t-1} + v_{2t}\,. \tag{2.5}$$

The processes for ϕ_{1t} and ϕ_{2t} are unobservable, continuously evolving parameters which reflect long-run marginal costs including scarcity rent, in the underlying structural model. Formally, ε_t, v_{1t}, and v_{2t}, $t = 1,\,\ldots\,,T$, are assumed to be independently and identically normally distributed with zero means and covariances, and variances σ^2_ϵ, $\sigma^2_{v_1}$, and $\sigma^2_{v_2}$, respectively. The lag structure and reliance on linear trends and uncorrelated unobservable components are dictated by the length of the sample: the series considered by Pindyck on the U.S. extend from 1870 to 1996 for crude oil and bituminous coal, and from 1919 to 1996 for natural gas.

Assuming normality of ε_t, v_{1t}, and v_{2t}, Pindyck proposes that Kalman filtering be applied to obtain paths for the state variables ϕ_{1t} and ϕ_{2t}. This means that, starting with initial values for model parameters and state variables, the filter computes at each period new values for the state variables to reflect new information on the observable series. Once the paths of the state variables are determined, the model can be estimated by maximum likelihood. The reader is referred to Kim and Nelson (1999, Chapter 3) for details of the Kalman filtering procedure and the associated likelihood functions. Following Pindyck (1999), we estimate and test this model separately for each price series considered, and over a long time span, as discussed in section 3.

[12]Formally, Pindyck suggests the quadratic trend model $P_t = c_1 + \phi_{1t} + c_5 t + c_6 t^2 + \phi_{2t} t + c_2 P_{t-1} + \epsilon_t\,, \quad t = 1,\,\ldots\,,T$, yet because of sample size restrictions, the author estimates $P_t = c_1 + \phi_{1t} + \phi_{2t} t + c_2 P_{t-1} + \epsilon_t\,, \quad t = 1,\,\ldots\,,T$. Given our emphasis on testing the time-invariant counterpart of this model (we provide further justifications below), we consider (2.3).

2.1. Testing for time-varying parameters

In view of assessing the statistical significance of TVP effects, the null hypothesis of interest is a simple mean-reverting model around a fixed trend line [the trending Ornstein-Uhlenbeck process given by equation (24) in Pindyck (1999)], *i.e.*,

$$P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t\,, \quad t = 1,\, \ldots\,,\, T\,. \tag{2.6}$$

It is clear that the models to be compared statistically are nested at the boundaries of certain parameters; formally,

$$\text{model (2.6)} \subseteq \text{model (2.3) when } \sigma^2_{v_1} \to 0,\ \sigma^2_{v_2} \to 0,\ \text{and } c_3 = c_4 = 1\,. \tag{2.7}$$

In other words, when $\sigma^2_{v_1} = 0$, $\sigma^2_{v_2} = 0$ and $c_3 = c_4 = 1$, then $\phi_{1t} = \phi_1$ and $\phi_{2t} = \phi_2$, $t = 1,\, \ldots\,,\, T$, leading to (2.6). As argued above, the zero variances as well as the unit root values for c_3 and c_4 lie on the boundary - rather than the interior - of the parameter space associated with (2.3). Further, it is easy to see that some parameters may not be identifiable under certain parameter configurations: for example, c_1 is not identified when $c_3 = 1$ and $\sigma^2_{v_1} = 0$, and it is "poorly" identified when we are close to these values; the same observation holds for c_5, when $c_4 = 1$ and $\sigma^2_{v_2} = 0$.

In this context, one cannot rely on estimated standard errors and standard limiting distributions, since their use for building tests and confidence sets is not justified even asymptotically. In particular, the distributions of some widely used test statistics, such as t-type and more generally Wald-type statistics, may be difficult (if not impossible) to bound under various null hypotheses, so that controlling the level of such tests may not be feasible. By contrast, the distributions of likelihood-ratio-type statistics appear to be more stable, so such tests provide a more appropriate basis for statistical inference.[13] Taking into account these observations, we consider the LR statistic:

$$LR = 2[L_{TVP} - L_{FCM}] \tag{2.8}$$

where L_{TVP} and L_{FCM} are, respectively, the maximum of the log-likelihood functions associated with (2.3) and (2.6).

For further reference, let $\hat{\varepsilon}_t$, $\hat{v}_{1t}$, and $\hat{v}_{2t}$, $t = 1, \ldots, T$, refer to post-estimation residuals associated with (2.3). In addition, denote the vector of parameters that define this model as

$$\omega = \left\{c_1, c_3, c_2, c_5, c_4, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{\epsilon}\right\} \tag{2.9}$$

and the vector of free parameters under the null hypothesis as

$$\theta = \left\{(c_1 + \phi_1)\,,\,(c_5 + \phi_2)\,,\,c_2,\,\sigma^2_{\epsilon}\right\}\,. \tag{2.10}$$

Furthermore, let $\hat{\omega}_{TVP}$ and $\hat{\theta}_{FCM}$ refer to the maximum-likelihood estimates of ω and θ calculated from the observed sample, imposing (2.3) and (2.6), respectively. Finally, denote the observed value of LR (*i.e.* the value obtained from the observed sample) as LR_0.

To obtain p-values for this statistic, we resort to the maximized Monte Carlo (MMC) test techniques [Dufour (2006)]. In what follows, we summarize the technique as it applies to our specific problem. Mainly, what we need is the possibility of simulating the relevant test statistic under the null hypothesis. Drawing from the null data-generating process under consideration requires setting a value for θ; the unidentified nuisance parameters (for example, c_3 and c_4 under the constant coefficient model) do not matter, because they simply do not appear in the null data-generating process. So given a specific value for θ, we define a p-value function, denoted $\hat{p}(LR_0|\theta)$, as follows.

(i) Generate a simulated sample from the null model by drawing from the normal distribution given the specific choice for θ. Reestimate the restricted and the alternative models (2.6) and

[13] See Andrews (2000, 2001), Dufour (1997, 2003) and Stock et al. (2002).

(2.3) given the simulated data, and using (2.8), compute the corresponding test statistic.

(iii) Repeat this process N times; this yields N simulated test statistics, denoted $LR_i(\theta), i = 1, \ldots, N$, from a data generating process (DGP) that satisfies the null hypothesis. In our notation, the presence of θ indicates that $LR_i(\theta)$ depends on data simulated given a specific choice for θ.

(iii) Compute the number $\hat{g}(LR_0 \,|\, \theta)$ of $LR_i(\theta)$ values which are not smaller than LR_0. Then

$$\hat{p}(LR_0 \,|\, \theta) = \frac{\hat{g}(LR_0 \,|\, \theta) + 1}{N + 1}\,. \tag{2.11}$$

The latter empirical p-value is thus based on the rank of LR_0 relative to its simulated counterparts.

The MMC technique involves maximizing $\hat{p}(LR_0 \,|\, \theta)$ sweeping over combinations of admissible values of θ.[14] Formally, let

$$\hat{p}_{MMC}(LR_0) = \sup_{\theta \in \Theta_0} \hat{p}(LR_0 \,|\, \theta)$$

where Θ_0 refers to the parameter space for θ conformable with the null hypothesis. The resulting MMC test is significant at level α if $\hat{p}_{MMC}(LR_0) \leq \alpha$. This can be viewed as a Monte Carlo implementation of the standard definition of the level of a test in the presence of nuisance parameters: when a test is nuisance-parameter dependent, an α level is achieved by comparing the largest p-value over all nuisance parameters consistent with the null hypothesis to α [see Lehmann (1986)]. Following the argument in Dufour (2006), this simulation-based procedure has level α, *i.e.*

$$\mathsf{P}\Big[\sup_{\theta \in \Theta_0} [\hat{p}_N(S_0 \,|\, \theta)] \leq \alpha\Big] \leq \alpha \quad \text{under } H_0.$$

The only condition needed to implement this procedure is the possibility of simulating the relevant

[14] Note that the same random draws should be used for each value of θ.

test statistic under the null hypothesis. The values of N and T (*i.e.* the number of replications and the sample size) are taken as given, and no asymptotic argument is needed.

The MMC method as described is highly related to the parametric bootstrap. Both procedures evaluate the null distribution of the test statistics under consideration by simulation. Yet the MMC and bootstrap test methods differ with respect to the treatment of nuisance parameters (here, θ). Typically, nuisance parameter point estimates are used to generate bootstrap samples; this does not guarantee level control in finite samples. In contrast, MMC p-values are simulated for all relevant nuisance parameters in order to provably control error probabilities for any sample size. So plugging $\hat{\theta}_{FCM}$ [a consistent estimator of θ under the null hypothesis] in steps (i)-(iii) above yields a parametric bootstrap p-value or, equivalently, a *local MC* (LMC) denoted $\hat{p}(LR_0 \,|\, \hat{\theta}_{FCM})$. Bootstrap procedures tend to be considerably more reliable than procedures based on asymptotic critical values. In the context of our problem however, where the asymptotic distribution may depend in a discontinuous way on nuisance parameters, it is well known that bootstrap procedures may also fail even asymptotically; see Dufour (2006) and the references therein. By contrast, the MMC procedure is immune to such failures. Of course, if $\hat{p}(LR_0 \,|\, \hat{\theta}_{FCM})$ is larger than a specific level, say 5%, there is no need to proceed with maximizing $\hat{p}(LR_0 \,|\, \theta)$ for a 5% level test. For this reason, the algorithm that maximizes the p-value function (in terms of θ) is initialized at the value used for the LMC test.

2.2. Diagnostic tests

In order to assess the underlying distributional hypotheses, it is common practice to perform diagnostic tests on estimated residuals from models such as (2.3). Because the Kalman Filter relies on the normality assumption, testing normality is typically considered as a key specification check. Using any series of residuals $\hat{u}_t$, $t = 1, ..., T$, deviations from normality are often assessed from the coefficients of skewness (Sk) and kurtosis (Ku), combined in the Jarque-Bera (JB) criterion:

$$JB = T\left[\frac{1}{6}(Sk)^2 + \frac{1}{24}(Ku-3)^2\right], \quad Sk = T^{-1}\sum_{t=1}^{T}\hat{w}_t^3, \quad Ku = T^{-1}\sum_{t=1}^{T}\hat{w}_t^4\,, \tag{2.12}$$

$$\hat{w}_t = \frac{\hat{u}_t - T^{-1}\sum_{t=1}^{T}\hat{u}_t}{\left[T^{-1}\sum_{t=1}^{T}\left(\hat{u}_t - T^{-1}\sum_{t=1}^{T}\hat{u}_t\right)^2\right]^{1/2}}\,. \tag{2.13}$$

For the model under consideration, the statistic may be computed for each series of residuals replacing, in turn, $\hat{\varepsilon}_t$, $\hat{v}_{1t}$, and $\hat{v}_{2t}$, $t = 1, \ldots, T$, for $\hat{u}_t$ in (2.12). While the $\chi^2(2)$ asymptotic null distribution is often applied in practice for JB, the fact remains that even with linear regression residuals, size distortions cannot be ruled out because of estimation uncertainty; see Dufour et al. (1998) and Dufour et al. (2003). We thus apply the MMC technique here again, to obtain exact p-values for the JB statistics. The procedure is applied as described in the previous section, replacing LR and θ by JB and ω respectively; the model tested is (2.3) for this test. The dimension of ω is larger than that of θ; maximizing the simulation-based p-value is thus more demanding for JB than for LR. Our simulation results reported in the next section show that $\chi^2(2)$ critical values can lead to very severe overrejections, which justifies resorting to simulation-based alternatives despite computational burdens.

The same method can be applied to other diagnostics, including tests for serial dependence or for contemporaneous correlation between errors (*e.g.*, to check whether v_{1t} and v_{2t} are uncorrelated). We discuss such tests as they relate to the empirical models considered below.

2.3. Reliability of standard asymptotics

It is important to remember that an asymptotic distributional theory has not been established for the test statistics described above, neither for the LR tests comparing alternative versions of TVP models nor for the diagnostic tests. In particular, usual asymptotic approximations may easily be invalid or unreliable in finite samples. In Appendix A, we present the results of a small Monte Carlo experiment which illustrates this feature, in the context of realistic designs based on parameter values obtained from our empirical study (Section 3). Among other things, test statistics for TVP models exhibit a bunch-up problem around zero, and probabilities of type I error can exceed by far the nominal levels of the tests (such as rejection rates of 99% for tests with a nominal of 5%).[15] In this paper, we simply bypass this problem by resorting to Monte Carlo test methods.

3. In-sample analysis

3.1. Data and estimated models

We study annual data on energy prices in the U.S., previously analyzed by Pindyck (1999).[16] The series for crude oil and bituminous coal extend from 1870 to 1996; for natural gas, the data cover 1919 to 1996. The nominal price series up to 1973 come from Manthy (1978) and the U.S. Bureau of the Census (1975). Pindyck (1999) updated this series through 1995 using data from the U.S. Energy Information Agency and, for 1996, the *Wall Street Journal*. The series are deflated using the U.S. wholesale price index until 1970, and the producer price index thereafter. Estimation is conducted on the logarithm of real prices. We have extended the series until 2006, following the same definitions as Pindyck.

In addition to the TVP model (2.3), we also consider the following special cases:

$$P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t \,, \quad t = 1, \ldots, T \,, \tag{3.1}$$

$$P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t \,, \quad t = 1, \ldots, T \,. \tag{3.2}$$

In the case of gas, our residual analysis backed by a historical perspective led us to introduce a two-regime variant of the TVP specifications, where the variance of ε_t is allowed to change in 1978 (while the other coefficients of the model are allowed to change according to the TVP scheme). Indeed, in the late seventies, a deregulation fundamentally altered the market for gas. Whether the

[15]Stock and Watson (1998) discussed the pile-up problem for the special case where the TVP specifications follow a unit-root.

[16]The data were generously provided by Pindyck.

Kalman-filter can adapt to such a structural shift is an open question. For gas, we thus examine the following three specifications:

$$P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t\,, \quad t = 1, \ldots, T\,, \tag{3.3}$$

$$P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t\,, \quad t = 1, \ldots, T\,, \tag{3.4}$$

$$P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t\,, \quad t = 1, \ldots, T\,, \tag{3.5}$$

where the ε_t are independent Gaussian with standard deviation $\sigma_\varepsilon^{(1)}$ before the deregulation date and $\sigma_\varepsilon^{(2)}$thereafter. In the results reported below, 1978 is taken as the date of the variance shift.

To ensure numerical convergence in estimating TVP models, we use *simulated annealing* (a global non-gradient-based algorithm) to obtain the maximum likelihood estimators. Furthermore, we also impose, in addition to usual convergence criteria from simulated annealing, the following convergence requirements: (i) the associated LR is positive, (ii) the restricted sum of squared residuals from the pricing equation is larger than its unrestricted counterpart, and (iii) the estimated information matrix is positive definite.[17] In other words, following a "normal convergence" output, conditions (i)-(iii) are checked; if at least one does not hold which signals non-convergence, the maximization algorithm is re-initiated using the "imperfect" solution as starting value, until all conditions are met. We also imposed the stability restrictions: $0 < c_2 < 1$, $0 < c_3 < 1$, $0 < c_4 < 1$. Such restrictions are not necessary for the validity of our test procedure. Yet, we have observed that stability constraints enhance convergence and avoid corner solutions. Maximization is typically difficult to achieve in the estimation of TVP models, and the numerical burden tends to be heavy.

The number of replications used for the LMC and MMC tests is $N = 999$. Since the MMC p-value must be larger than the LMC p-value, it is not necessary to compute the MMC p-value if the former is larger than the level of the test (in this case, we use 0.05). Since the p-value function is a non-differentiable step function, we use *simulated annealing* (again) to obtain the MMC p-

[17] Our emphasis on the (2.3) form justifies requirements (i) and (ii). Invertability of the information matrix is imposed for the observed data only (since simulated data is drawn under the null hypothesis in which case this restriction does not necessarily hold).

value, using as maximization domain the relevant maximum-likelihood estimates ± 10 estimated standard errors, subject to the following restrictions: (i) the variance parameters are constrained to be nonnegative, and (ii) the parameters should satisfy the stability restrictions. In the following discussion, significance refers to a 5% test level.

Finally, for testing error normality in the case of the two-regime model for gas, $\hat{w}_t$ in (2.12) is redefined as follows:

$$\begin{aligned}\hat{w}_t &= \frac{\hat{u}_t - T_1^{-1}\sum_{t=1}^{T_1}\hat{u}_t}{\left[T_1^{-1}\sum_{t=1}^{T_1}\left(\hat{u}_t - T_1^{-1}\sum_{t=1}^{T_1}\hat{u}_t\right)^2\right]^{1/2}}, \qquad \text{before the deregulation date,} \\ &= \frac{\hat{u}_t - T_2^{-1}\sum_{t=T_1+1}^{T}\hat{u}_t}{\left[T_2^{-1}\sum_{t=T_1+1}^{T}\left(\hat{u}_t - T_2^{-1}\sum_{t=T_1+1}^{T}\hat{u}_t\right)^2\right]^{1/2}}, \qquad \text{after the deregulation date,}\end{aligned}$$

where T_1 is the sample size before the deregulation date and $T_2 = T - T_1$. In other words, we center and scale residuals using empirical means and standard deviations within each subsample. This modification introduces different pre and post-break standardizations for the residuals to account for the hypothesized break. Since our finite-sample testing approach automatically accounts for nuisance parameters, there is no need to provide an alternative distributional theory for the proposed statistic.[18]

3.2. Results

The point estimates are reported in Table 1, while the tests of parameter constancy against TVP specifications appear in Table 2. Focusing on the latter, we find evidence in favour of a TVP model (with time-varying intercept and trend) for gas and coal, but not for oil. For for all three series, the constant coefficient model cannot be rejected against a TVP specification with constant trend coefficient and time-varying intercept. Only the case of gas does a TVP specification with fixed

[18]The literature on Monte Carlo tests includes ample examples where intuitive although non-standard test statistics have been introduced and proved to outperform existing procedures whose popularity is largely due to a standard asymptotic null distribution. See for example Dufour and Khalaf (2002), Dufour et al. (2003), Dufour et al. (2004), Bernard et al. (2007), and Dufour et al. (2010), for such examples in the context of diagnostic tests.

Table 1. Parameter estimates for different energy price models

TVP-intercept-trend								
	c_1	c_3	c_2	c_5	c_4	σ_{v_1}	σ_{v_2}	σ_ϵ
Oil	.2426	.5414	.7070	.0025	.7306	.00013	.0006	.1646
	(2.25)	(.39)	(6.68)	(2.60)	(3.43)	(.0011)	(2.02)	(11.20)
Coal	.2346	.0913	.8101	.0011	.8550	.0771	.0002	1.93×10^{-5}
	(2.04))	(.82)	(8.33)	(1.41)	(10.28)	(12.81)	(2.05)	(7.62×10^{-5})
Gas	.2132	.8300	.8120	.0082	.0706	.0512	.0016	6.3×10^{-7}
	(1.22)	(4.94)	(4.74)	(1.04)	(.33)	(5.26)	(6.33)	(9.6×10^{-6})
TVP-intercept								
	c_1	c_3	c_2	c_5	c_4	σ_{v_1}	σ_{v_2}	σ_ϵ
Oil	.1898	.7372	.7372	.0026	–	.0450	–	.1693
	(.87)	(2.46)	(2.46)	(.88)	–	(.57)		(4.22)
Coal	.2320	.8173	.8173	.0009	–	.0257	–	.0734
	(1.27)	(5.62)	(5.63)	(1.09)	–	(1.75)		(7.39)
Gas	.1849	.7347	.8347	.0073	–	.0436	–	.1058
	(1.09)	(6.03)	(6.04)	(1.22)	–	(2.02)		(6.57)
TVP-trend								
	c_1	c_3	c_2	c_5	c_4	σ_{v_1}	σ_{v_2}	σ_ϵ
Oil	.2276	–	.7178	.0025	.7098	–	.0006	.1654
	(2.49)		(7.69)	(2.71)	(3.24)		(1.89)	(11.37)
Coal	.2090	–	.8306	.0010	.8456	–	.0002	.0769
	(2.17)		(10.19)	(1.47)	(10.05)		(2.21)	(12.95)
Gas	.0221	–	.9656	.0023	.3343	–	.0019	.0488
	(.31)		(31.62)	(2.43)	(2.16)		(8.60)	(4.55)
$AR(1)$ with linear trend								
	$c_1 + \phi_1$		c_2	$c_5 + \phi_2$		σ_{v_1}	σ_{v_2}	σ_ϵ
Oil	.1366		.8117	.0019		–	–	.1805
	(2.53)		(18.40)	(4.03)		–	–	
Coal	.1044		.9214	.00037		–	–	.0832
	(2.24)		(26.22)	(1.43)		–	–	
Gas	.0665		.9349	.0034		–	–	.1261
	(1.08)		(34.61)	(3.98)		–	–	

Note – The model called "TVP-intercept-trend" uses the equation: $P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$, where $\phi_{1t} = c_3 \phi_{1,t-1} + v_{1t}$, $\phi_{2t} = c_4 \phi_{2,t-1} + v_{2t}$, $t = 1, \ldots, T$. The "TVP-intercept" model is a special case of the latter where only the intercept is a random coefficient: $P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$. Similarly, in the "TVP-trend" model, only the trend coefficient is random: $P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$. The "$AR(1)$ model with linear trend" is: $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$. t-statistics are reported in parentheses.

Table 2. Tests of constant parameter model versus Pindyck's specifications

	TVP-intercept-trend		TVP-intercept		TVP-trend	
	LR	p-value	LR	p-value	LR	p-value
Oil	2.60	.424	0.14	.886	2.46	.236
		($> 5\%$)		($> 5\%$)		($> 5\%$)
Coal	20.04	.001	2.22	.252	3.45	.247
		(.001)		($> 5\%$)		($> 5\%$)
Gas	36.95	.001	3.88	.202	29.73	.001
		(.001)		($> 5\%$)		(.001)

Note – For the definitions of the models, see the note to Table 1. In each case, the time invariant model counterpart is tested against the TVP specification. The LR statistic is defined in (2.8); the LMC p-value is reported with the MMC counterpart in parentheses.

Table 3. Residual-based normality tests for models with significant TVP specifications

Model		Test on ε_t		Test on v_{1t}		Test on v_{2t}	
		JB	p-value	JB	p-value	JB	p-value
TVP-intercept-trend	Coal	114.78	.010	25717.60	.210	208.90	.63
			(.010)		($> 5\%$)		($> 5\%$)
	Gas	79.47	.010	32.18	.430	30.61	.86
			(.010)		($> 5\%$)		($> 5\%$)
TVP-trend	Gas	39.35	.039	–	–	9.41	.006
			(.135)		–		(.994)

Note – For the definitions of the models, see the note to Table 1. The JB statistic is defined in (2.12); the LMC p-value is reported with the MMC counterpart in parentheses.

intercept and time-varying trend coefficient turns out to be statistically significant, which suggests that TVP effects are mainly present in the trend coefficient in this case. For coal, TVP effects on both the intercept and the trend are significant when both are included, whereas individually they are not. This finding suggests a multicollinearity (or an identification) problem, as in linear regressions when an F-test on two coefficients on two coefficients is significant while the associated *t*-tests are not). Nevertheless, our residual analysis reported below calls for caution in interpreting this conclusion.

Residual plots from the pricing error equation for the TVP-intercept-trend models (and the two-regime TVP model for natural gas) are provided in Appendix B. The application of usual serial dependence tests revealed no significant departures from the *i.i.d.* hypothesis.

From Table 3 and relying on the MMC test, we see that despite a significant LR test, normality of the pricing error terms is rejected within the TVP-intercept-trend models for both coal and gas. In contrast, for gas, normality is not rejected for the TVP-trend model. In this case, the LMC p-values are less than 5%, so the LMC and MMC p-values yield conflicting decisions. Given the extent of size distortions revealed by our Monte Carlo experiment, we prefer to rely on the MMC p-values to avoid spurious rejections. In this regard, it is worth noting that using the asymptotic critical values would have led to rejecting normality for all models analyzed.

We also explored the possibility of allowing for contemporaneous correlation between v_{1t} and v_{2t} in the TVP model. Although this could be an attractive extension of Pindyck's original model, it appears difficult to implement in practice. Indeed, for all series considered, our attempts to estimate such an extension of the TVP model led to numerical difficulties and convergence typically failed. This may be associated with identification issues or the fact that sample sizes are too small to allow reliable estimation. A fortiori, this makes testing the absence of correlation difficult. For this reason, and though we believe that TVP-filtered diagnostic tests deserve further theoretical work beyond the scope of this paper (our small-scale Monte Carlo study points clearly in this direction), we shall mainly focus on deviations from normality to assess the considered TVP specifications.

Upon inspection of the residual plot for coal (Appendix B), we suspect that departures form normality may be driven by the end-of-sample residuals. Our forecasting exercise (reported in the next section) supports this observation. In the case of gas, visual inspection of the residual plot confirms our decision to consider for a two-regime model for the variance. One must however guard against hasty conclusions based on graphs, since the residual plot of the two-regime TVP specification suggests a similar pattern.

From Table 5, we see that TVP effects are not significant against a two-regime time invariant specification.[19] Here the LMC p-values are less than 5% (the associated χ^2 critical value also signals rejection at this level), so the LMC and MMC p-values yield conflicting decisions. In line with our Monte Carlo experiment conformable with this design, we again prefer to rely on the MMC p-value. Low power cannot be ruled out, since a two-regime model may be over-parameterized.

Overall, our in-sample LR testing exercise rejects the fixed-coefficient model in favour of Pindyck's model for coal and gas, though not for oil. However, our independence and normality tests suggest that Pindyck's models is too restrictive to conclusively fit the data. For these reasons, these findings should be interpreted in conjunction with our forecasting results.

4. Out-of-sample forecasts

In line with Pindyck's original objective, we complement our in-sample analysis with a forecasting exercise. In this case, we consider two further benchmark models commonly considered to analyze the long-run dynamics of energy prices, namely the random walk model with or without drift (the latter leading to the *no-change forecast*), and a fixed coefficient quadratic trend model. The competing models are defined in the notes to tables 6 and 7. In addition to a long-run forecasting analysis similar to Pindyck's, we also consider a real-time approach. In other words, we recursively estimate all models under consideration (fixed-coefficient as well as time-varying parameter models) and forecast in real-time: we derive one-step-ahead out-of-sample forecasts, where parameter estimates are updated at every step of the procedure. For both long-run and real-time exercises,

[19] Although we only report normality tests for significant TVP specifications, it is worth noting that the LMC test for TVP two-regime residuals did not detect deviations form normality.

Table 4. Two-regime model for natural gas: parameter estimates

	c_1	c_3	c_2	c_5	c_4	σ_{v_1}	σ_{v_2}	$\sigma_\varepsilon^{(1)}$	$\sigma_\varepsilon^{(2)}$
TVP-intercept -trend	.0393	.3378	.6552	.0198	.8616	.0423	.0010	$8.59.10^{-7}$	.1413
	(.25)	(1.66)	(5.75)	(2.75)	(74.85)	(5.31)	(3.89)	$(3.1.10^{-5})$	(4.95)
TVP-intercept	.2955	.8616	.7174	.0123	–	.0616	–	.0092	.1540
	(1.56)	(7.59)	(3.76)	(1.25)	–	(4.43)	–	(.18)	(5.62)
TVP-trend	.0539	–	.6861	.0179	.8572	–	.0010	.0399	.1463
	(.40)	–	(7.50)	(3.26)	(82.72)	–	(4.86)	(5.94)	(5.42)

	$c_1+\phi_1$	c_2	$c_5+\phi_2$	σ_{v_1}	σ_{v_2}	$\sigma_\varepsilon^{(1)}$	$\sigma_\varepsilon^{(2)}$
$AR(1)$ with linear trend	.0036	.9642	.0030	–	–	.0762	.1765
	(.06)	(36.97)	(3.52)	–	–		

Note – The model denoted *TVP-intercept-trend*, corresponds to $P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$, where ε_t are independent Gaussian with standard deviation $\sigma_\varepsilon^{(1)}$ before the deregulation date and $\sigma_\varepsilon^{(2)}$thereafter. The special cases denoted *TVP-trend* and *TVP-intercept*, correspond to $P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$, and $P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$, respectively; in the above, $\phi_{1t} = c_3 \phi_{1,t-1} + v_{1t}$, $\phi_{2t} = c_4 \phi_{2,t-1} + v_{2t}$, $t = 1, \ldots, T$. The model denoted $AR(1)$ with linear trend corresponds to $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$. Numbers in parentheses are t-statistics.

Table 5. Two-regime model for natural gas: TVP tests

TVP-intercept-trend		TVP-intercept		TVP-trend	
LR	58.41	LR	50.76	LR	55.71
p-values	.015	p-values	.023	p-values	.023
	(.999)		(.988)		(.999)

Note – For the definitions of the models denoted *TVP-intercept-trend*, and *TVP-trend*, see notes to Table 4. In each case, the time invariant model counterpart is tested against the TVP specification. The LR statistic is defined in (2.8); the LMC p-value is reported with the MMC counterpart in parentheses. *Empirical rejections* refers to the proportion of simulated statistics that exceed the χ^2 critical values for a test with level 5%. Statistics are simulated under the time invariant counterpart of each TVP model (the null hypothesis) with parameters as estimated (under the null hypothesis) from the gas price series.

we also analyze forecasts obtained through model averaging, using unweighted forecast means and medians, as well as weighted means based on information criteria (AIC and BIC). We report mean squared forecast errors, and we rely mainly on the latter to analyze the results.

To the best of our knowledge, tests of forecasting performance are available for either nested or non-nested model comparisons. Given the serious size problems with in-sample tests and the nonregular features of the models studied here, we prefer to avoid an inferential out-of-sample analysis.[20] As in Pindyck's study, a forecast horizon starting in 1976 is considered for oil and coal. The long-run forecasts span the 1976-2006 horizon and do not use post-1976 data. In the real-time exercise, we first forecast 1976 prices using pre-1976 data; after 1976, we add a year of data to the estimation sample one year at a time, reestimate the model using each new sample, and forecast the subsequent year using only the information in the estimation sample. In the case of gas, we follow the same procedure except that, for numerical stability, we use a forecasting horizon starting in 1996. Recall that the series on gas prices starts in 1919, whereas for oil and coal, our data go back to 1870.

For the oil series, out-of-sample results favour time-invariant models for long-run and real-time forecasts. This concurs with our in-sample findings. Interestingly, the random-walk model emerges as the best forecasting tool: the random walk with drift outperforms all other models for long-run forecasts and the real-time exercise selects the no-change forecast. For related work, see Alquist and Kilian (2009) and the survey of Hamilton (2009).

In contrast, the long-run and real-time exercises lead to conflicting results regarding TVP effects for coal: the no-change forecast outperforms all other forecasts in real-time whereas the long-run exercise selects the TVP-intercept-trend specification. Here again, our in-sample analysis does not contradict this finding. Recall that we found evidence in favour of the latter model through LR tests, yet we also detected significant departures from normality. Residual plots however indicate that such departures appear to be associated with observations near the end of our sample. So most of

[20]Developing forecast comparison tests that account for boundary nesting is a worthy research objective beyond the scope of the present paper.

Table 6. Mean squared forecast errors (1976-2006)

	Oil		Coal		Gas	
	Long run	Real time	Long run	Real time	Long run	Real time
	TVP models					
TVP-intercept-trend	.3394	.1110	**.1590**	.0040	.1950	.0647
TVP-intercept	.3421	.1055	.1651	.0068	.6936	.1058
TVP-trend	.3394	.1084	.1602	.0055	.2265	**.0586**
	Fixed coefficient models					
$AR\,(1)$ with second-order trend	.2441	.1049	.4845	.0101	1.8051	.0794
$AR\,(1)$ with linear trend	.3394	.1147	.2044	.0113	**.1278**	.0676
$AR\,(1)$	.6447	.1367	.2438	.0057	.3157	.0850
Random walk with drift	**.1550**	.1162	.4809	.0084	.2995	.0904
No-change forecast	.1907	**.0568**	.3438	**.0026**	.6837	.0655
	Model averaging					
AIC-weighted	.3312	.1071	.2640	.0075	.2044	.0650
BIC-weighted	.3972	.1142	.3650	.0080	.2724	.0724
Unweighted average	.2819	.0958	.3029	.0058	.4499	.0645
Median	.2921	.1044	.2866	.0059	.3200	.0715

Note – The model called "TVP-intercept-trend" uses the equation: $P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$, where $\phi_{1t} = c_3 \phi_{1,t-1} + v_{1t}$, $\phi_{2t} = c_4 \phi_{2,t-1} + v_{2t}$, $t = 1, \ldots, T$. The "TVP-intercept" model is a special case of the latter where only the intercept is a random coefficient: $P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$. In the "TVP-trend" model, only the trend coefficient is random: $P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$. The $AR(1)$ model with second-order trend is: $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_6 t^2 + c_2 P_{t-1} + \varepsilon_t$. The "$AR(1)$ model with linear trend", the "$AR\,(1)$ model" and the "Random walk with drift" are restricted forms of the latter given by: $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$ $P_t = (c_1 + \phi_1) + c_2 P_{t-1} + \varepsilon_t$ and $P_t = (c_1 + \phi_1) + P_{t-1} + \varepsilon_t$, respectively. The no-change forecast is produced by a random walk without drift: $P_t = P_{t-1} + \varepsilon_t$. Numbers in bold indicate the smallest mean squared forecast error.

Table 7. Mean squared forecast errors, two-regime model for natural gas

	Long Run	Real time
TVP-intercept-trend	.2860	**.0634**
TVP-intercept	.5302	.0661
TVP-trend	.2589	.0644
$AR\,(1)$ with second-order trend	1.40	.0794
$AR\,(1)$ with linear trend	**.0917**	.0676
$AR\,(1)$	.2182	.0850
Random walk with drift	.2822	.0904
No-change forecast	.6837	.0655
Model averaging, AIC-weighted	.2858	.0664
Model averaging, BIC-weighted	.2853	.0830
Model averaging, unweighted	.3980	.0658
Model averaging, median	.2856	.0718

Note – The models compared can be summarized as follows. (1) TVP-intercept-trend: $P_t = c_1 + \phi_{1t} + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$, where $\phi_{1t} = c_3 \phi_{1,t-1} + v_{1t}$, $\phi_{2t} = c_4 \phi_{2,t-1} + v_{2t}$, $t = 1, \ldots, T$.; (2) TVP-intercept: $P_t = c_1 + \phi_{1t} + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$; (3) TVP-trend model: $P_t = (c_1 + \phi_1) + c_5 t + \phi_{2t} t + c_2 P_{t-1} + \varepsilon_t$; (4) $AR(1)$ with second-order trend: $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_6 t^2 + c_2 P_{t-1} + \varepsilon_t$; (5) $AR(1)$ with linear trend: $P_t = (c_1 + \phi_1) + (c_5 + \phi_2)\, t + c_2 P_{t-1} + \varepsilon_t$; (6) $AR(1)$: $P_t = (c_1 + \phi_1) + c_2 P_{t-1} + \varepsilon_t$; (7) random walk with drift: $P_t = (c_1 + \phi_1) + P_{t-1} + \varepsilon_t$. The no-change forecast corresponds to the random walk with no drift: $P_t = P_{t-1} + \varepsilon_t$. The pricing errors ε_t are independent Gaussian with variance $\sigma_\varepsilon^{(1)}$ before the deregulation date, and $\sigma_\varepsilon^{(2)}$ thereafter. Numbers in bold indicate the smallest mean square forecast error.

these forecasts are not affected by these outlying observations.

In the case of gas, when we restrict focus to the single regime case, the time invariant model with a linear deterministic trend is preferred as a long-run forecasting tool; in contrast, the TVP-in-trend specification outperforms all other specifications for real-time forecasting. This model was indeed not rejected using in-sample LR and normality tests. Accounting for the two-variance case improves the performance of the AR(1) model with linear trend; the two-variance version of this model is preferred to the single variance case, and also emerges as the best overall specification for long-run forecasting. Here again, this is the model favoured by our in-sample analysis. The real-time exercise allowing for two variances supports the TVP-intercept-trend specification, but this model remains marginally dominated by the single regime TVP-trend special case.

Overall, our out-of-sample exercise selects a random-walk specification for long-run and real-time forecasting of the oil price, and Pindyck's model for both real-time forecasting in the case of gas and long-run forecasting in the case of coal. We do not claim generality, because results depend (as usual) on the forecasting horizon considered. At any rate, we find that our in-sample test results do not conflict with our forecasting study.

The fact that real-time and long-run exercises suggest different decisions may be due to several factors. The TVP models may themselves be structurally unstable, so allowing for recursive estimation could in a way circumvent this problem. From the structural stability perspective, even linear models including the unit-root case are given a fair chance since the drift parameter estimates can adjust to additional observations. From a different perspective, the nonlinear specifications may suffer from non-negligible estimation uncertainty, so allowing point estimates to recursively adapt with incoming data may alleviate this difficulty. Continuous updating in real-time has practical advantages, when the economic question at hand does not call for a long-run forecast. The small samples under consideration may also drive some of the observed discrepancies between the in-sample and the out-of sample results. For further discussion, see Inoue and Kilian (2004, 2006) and Hansen (2009*b*).

Finally, we find that averaging does not improve forecasts relative to the best performing model, neither from a long-run nor from a real-time perspective. Recent econometric studies suggest that combining forecasts from several models can improve accuracy, in the presence of structural instabilities; see Clark and McCracken (2009) or Hansen (2009*a*). In this regard, our exercise is telling, since TVP models have not been formally considered in this emerging literature so far. Of course, our findings depend on the competing models, the data as well as on the forecasting horizons. One issue is however worth raising. On comparing the unweighted to the weighted averages, one may suspect that weights based on information-criteria are not adapted to TVP models. Our results thus suggest that further research in this direction is a worthy objective.

5. Conclusion

This paper tests the statistical significance of Pindyck's (1999) suggested class of econometric models for the behavior of long-run real energy prices. These postulate mean-reverting prices with continuous and random changes in their level and trend, using Kalman filtering for the estimation. In such contexts, the distributions of the test statistics are typically non-standard and depend on nuisance parameters. We conduct a small-scale simulation study based on empirically relevant designs to illustrate the serious implications of these problems for applied work. Exploiting simulation-based procedures to address this issue, we report results for both LR tests for TVP and for residual-based normality tests.

A further contribution of this paper is to examine in-sample and out-of-sample fit for the specific class of models at hand. We thus complement our in-sample analysis with a long-run as well as a real-time forecasting exercise. We also produce and analyze model-averaged forecasts, to assess the effectiveness of combining various models.

Our in-sample LR tests select Pindyck's model for coal and gas, though not for oil. However, our independence and normality tests suggest that the model is too restrictive to conclusively fit the data. Allowing for non-Gaussian disturbances in TVP models and for non-zero correlations between different disturbances, stand out as potentially useful extensions. Our out-of-sample exercise selects the random-walk specification for oil, and Pindyck's model for real-time forecasting in the case of gas as well as for long-run forecasting in the case of coal. From the methodological perspective, our results suggest that developing further diagnostic in- and out-of-sample methods for TVP models is a worthy research objective.

References

Ahrens, W. A. and Sharma, V. R. (1997), 'Trends in natural resource commodity prices: Deterministic or stochastic?', *Journal of Environmental Economics and Management* **33**, 59–74.

Alquist, R. and Kilian, L. (2009), 'What do we learn from the price of crude oil futures?', *Journal of Applied Econometrics* **forthcoming**.

Andrews, D. W. K. (2000), 'Inconsistency of the bootstrap when a parameter is on the boundary of the maintained hypothesis', *Econometrica* **68**, 399–405.

Andrews, D. W. K. (2001), 'Testing when a parameter is on the boundary of the maintained hypothesis', *Econometrica* **69**, 683–733.

Barsky, R. B. and Kilian, L. (2004), 'Oil and the macroeconomy since the 1970s', *Journal of Economic Perspectives* **18**, 115–134.

Berck, P. and Roberts, M. (1996), 'Natural resource prices: Will they ever turn up', *Journal of Environmental Economics and Management* **9**, 122–137.

Bernard, J.-T., Idoudi, N., Khalaf, L. and Yelou, C. (2007), 'Finite sample multivariate structural change tests with application to energy demand models', *Journal of Econometrics* **141**, 1219–1244.

Clark, T. E. and McCracken, M. W. (2009), Averaging forecasts from vars with uncertain instabilities, Discussion paper, FEDS Working Paper No. 2007-42.

Cortazar, G. and Naranjo, L. (2006), 'An N-factor gaussian model for oil futures prices', *The Journal of Future Markets* **26**, 243–268.

Cortazar, G. and Schwartz, E. S. (2003), 'Implementing a stochastic model for oil future prices', *Energy Economics* **25**, 215–238.

Dufour, J.-M. (1997), 'Some impossibility theorems in econometrics, with applications to structural and dynamic models', *Econometrica* **65**, 1365–1389.

Dufour, J.-M. (2003), 'Identification, weak instruments and statistical inference in econometrics', *Canadian Journal of Economics* **36**(4), 767–808.

Dufour, J.-M. (2006), 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics', *Journal of Econometrics* **133**, 443–478.

Dufour, J.-M., Farhat, A., Gardiol, L. and Khalaf, L. (1998), 'Simulation-based finite sample normality tests in linear regressions', *The Econometrics Journal* **1**, 154–173.

Dufour, J.-M. and Khalaf, L. (2002), 'Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions', *Journal of Econometrics* **106**(1), 143–170.

Dufour, J.-M., Khalaf, L. and Beaulieu, M.-C. (2003), 'Exact skewness-kurtosis tests for multivariate normality and goodness-of-fit in multivariate regressions with application to asset pricing models', *Oxford Bulletin of Economics and Statistics* **65**, 891–906.

Dufour, J.-M., Khalaf, L. and Beaulieu, M.-C. (2010), 'Multivariate residual-based finite-sample tests for serial dependence and GARCH with applications to asset pricing models', *Journal of Applied Econometrics* **forthcoming**.

Dufour, J.-M., Khalaf, L., Bernard, J.-T. and Genest, I. (2004), 'Simulation-based finite-sample tests for heteroskedasticity and ARCH effects', *Journal of Econometrics* **122**(2), 317–347.

Gibson, R. and Schwartz, E. S. (1990), 'Stochastic convenience yield and the pricing of oil contingent claims', *Journal of Finance* **45**, 959–76.

Godfrey, L. G. (1996), 'Some results on the Glejser and Koenker tests of heteroscedasticity', *Journal of Econometrics* **72**, 275–299.

Hamilton, J. D. (2003), 'What is an oil shock', *Journal of Econometrics* **113**, 363–398.

Hamilton, J. D. (2009), 'Understanding crude oil prices', *The Energy Journal* **30**, 179–206.

Hamilton, J. D. and Herrera, A. M. (2004), 'Oil shocks and aggregate macroeconomic behavior: The role of monetary policy', *Journal of Money Credit and Banking* **36**, 265–286.

Hansen, B. E. (2007), 'Least squares model averaging', *Econometrica* **75**, 1175–1189.

Hansen, B. E. (2008), 'Least squares forecast averaging', *Journal of Econometrics* **146**, 342–350.

Hansen, B. E. (2009*a*), 'Averaging estimators for regressions with a possible structural break', *Econometric Theory* **35**, 1498–1514.

Hansen, P. R. (2009*b*), In-sample fit and out-of-sample fit: Their joint distribution and its implications for model selection, Discussion paper, Stanford Unibversity.

Inoue, A. and Kilian, L. (2004), 'In-sample or out-of-sample tests of predictability: Which one should we use?', *Econometric Reviews* **23**(4), 371–402.

Inoue, A. and Kilian, L. (2006), 'On the selection of forecasting models', *Journal of Econometrics* **130**(2), 273–306.

Khalaf, L. and Kichian, M. (2005), 'Exact test for breaks in covariance in multivariate regressions', *Economics Letters* **95**, 241–246.

Kilian, L. (2008*a*), 'A comparison of the effects of exogenous oil supply shocks on output and inflation in the G7 countries', *Journal of the European Economic Association* **6**, 78–121.

Kilian, L. (2008*b*), 'The economic effects of energy price shocks', *The Journal of Economic Literature* **46**, 871–909.

Kilian, L. (2008*c*), 'Exogenous oil supply shocks: How big are they and how much do they matter for the U.S. economy?', *Review of Economics and Statistics* **90**, 216–240.

Kilian, L. (2009), 'Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market', *American Economic Review* **99**, 1053–1069.

Kilian, L. and Park, C. (2009), 'The impact of oil price shocks on the U.S. stock market', *International Economic Review* **50**, 1267–1287.

Kilian, L., Rebucci, A. and Spatafora, N. (2009), 'Oil shocks and external balances', *Journal of International Economics* **77**, 181–194.

Kim, C.-J. and Nelson, C. R. (1999), *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, The MIT Press, Cambridge, Massachusetts.

Lee, J., List, J. A. and Strazicich, M. C. (2006), 'Non-renewable resource prices: Deterministic or stochastic trends?', *Journal of Environmental Economics and Management* **51**, 354–370.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses, 2nd edition*, John Wiley & Sons, New York.

Livernois, J. (2009), 'On the empirical signifiance of the hotelling rule', *Review of Environmental Economics and Policy* **3**, 22–41.

Manthy, R. S. (1978), *Natural Resource Commodities: A Century of Statistics*, John Hopkins Press, Baltimore.

Moshiri, S. and Foroutan, F. (2006), 'Forecasting nonlinear crude oil future prices', *The Energy Journal* **27**, 81–85.

Pindyck, R. S. (1999), 'The long run evolution of energy prices', *The Energy Journal* **20**, 1–27.

Pindyck, R. S. (2001), 'The dynamics of commodity spot and futures markets: a primer', *The Energy Journal* **22**, 1–29.

Postali, F. A. S. and Picchetti, P. (2006), 'Geometric brownian motion and structural breaks in oil prices: A quantitative analysis', *Energy Economics* **28**, 506–522.

Regnier, E. (2007), 'Oil and energy price volatility', *Energy Economics* **29**, 405–427.

Sadorsky, P. (2006), 'Modeling and forecasting petroleum futures volatility', *Energy Economics* **28**, 467–488.

Schwartz, E. S. (1997), 'The stochastic behavior of commodity prices : Implications for valuation and hedging', *The Journal of Finance* **52**, 923–973.

Schwartz, E. S. and Smith, J. E. (2000), 'Short-term variations and long-term dynamics in commodity prices', *Management Science* **46**, 893–911.

Slade, M. E. (1982), 'Trends in natural resource commodity prices: An analysis of the time domain', *Journal of Environmental Economics and Management* **9**, 132–137.

Slade, M. E. (1988), 'Grade selection under uncertainty: Least cost last and other anomalies', *Journal of Environmental Economics and Management* **15**, 189–205.

Stock, J. H. and Watson, M. W. (1998), 'Median unbiased estimation of coefficient variance in a time-varying parameter model', *Journal of the American Statistical Association* **441**, 349–358.

Stock, J. H., Wright, J. H. and Yogo, M. (2002), 'A survey of weak instruments and weak identification in generalized method of moments', *Journal of Business and Economic Statistics* **20**(4), 518–529.

Tabak, B. M. and Cajueiro, D. O. (2007), 'Are the crude oil markets becoming weakly efficient over time? a test for time-varying long-range dependence in prices and volatility', *Energy Economics* **29**, 28–36.

Appendix

A. Simulation study

To provide further motivation for the methodology employed in this paper, we report here the results of a small Monte Carlo experiment on the properties of asymptotic procedures in the context of the models studied in this paper. We consider in turn the behavior of tests for restricted TVP or constant parameter models against Pindyck's specification, and residual-based normality tests.

Indeed, the algorithm [described in steps (i)-(iii) of Section 2.1] which underlies the definition of the Monte Carlo p-value function may be exploited to assess the size of the asymptotic test based on the statistics under consideration, namely LR and JB. For presentation ease, let us first focus on the LR test problem. In this case, the N simulated test statistics $LR_i(\theta)$, $i = 1, \ldots, N$, correspond to a data generating process (DGP) that satisfies the null hypothesis (with a given choice of θ). So, if we compare each $LR_i(\theta)$ to the asymptotic critical value from the χ^2 distribution (with degree of freedom equal to the number of coefficients fixed by the null hypothesis) for a test with (nominal) level α, and count rejections over the N replications, we obtain an estimate of the empirical probability of type I error (or level) for the design considered (and a specific value of θ).

To get empirically relevant designs, the parameter vector tested θ is set to the value $\hat{\theta}_{FCM}$ obtained from our data model under the relevant null hypothesis [as given in Table 1 of Section 3]. We report the empirical rejections corresponding to $\hat{\theta}_{FCM}$, for each energy series considered. Since the same observation holds in the case of $JB_i(\omega)$, $i = 1, \ldots, N$, we also report the empirical rejections obtained on setting $\omega = \hat{\omega}_{TVP}$, where $\hat{\omega}_{TVP}$ is the unrestricted ML estimate of the TVP model. The values of θ and ω so chosen remain fixed for the purpose of the simulation.

In the context of TVP tests, usual statistics tend to bunch-up around zero, which casts doubts on the appropriateness of χ^2 critical values. Our simulation design allows us to assess this problem for the LR tests. Specifically, in addition to the empirical size, we report the number of simulated $LR_i(\theta)$ which are close to zero (formally, between 0 and 0.01). This exercise aims to be illustrative.

Table 8. LR tests of constant parameter model versus Pindyck's specification: Monte Carlo study

Model	TVP-intercept-trend		TVP-intercept		TVP-trend	
	Empirical rejections	Mass around zero	Empirical rejections	Mass around zero	Empirical rejections	Mass around zero
Oil	21.4	56.6	2.5	1.4	6.0	23.2
Coal	3.9	2.0	3.6	5.5	10.5	20.0
Gas	8.6	3	8.2	2.8	16.8	10.3

Note – *Empirical rejections* refer to the proportion of simulated statistics that exceed the χ^2 critical values for tests with level 5%. Statistics are simulated under the null hypothesis, namely model (2.6) with $\hat{\theta}_{FCM}$ as calculated from each price series. *Mass around zero* refers to the proportion of simulated statistics that are close to zero (formally, that lie between 0 and 0.01); the latter aims to illustrate the bunching-up-at-zero problem. For the definitions of the models denoted *TVP-intercept-trend*, *TVP-intercept* and *TVP-trend*, see notes to Table 1.

Table 9. TVP tests against a two-regime model for natural gas: Monte Carlo study

	TVP-intercept-trend	TVP-intercept	TVP-trend
Empirical Size	99.1	99.2	99.3

The MMC method is not affected by this problem.

Results for LR tests of parameter constancy restrictions against TVP models are reported in tables 8 and 9. When the alternative is a model with one TVP regime, we see that the probability of type I error can substantially exceed 5% (the nominal significance level), with rejection rates exceeding 20%. Against the two-regime model, we observe empirical levels close to 100% (rather than the nominal level of 5%). Over-parameterization may be driving the latter result. Tests against the two-regime model suffer from an additional identification problem arising from the possibility of a non-existent break, in which case estimating two different values for the pricing error variance may lead to serious estimator instability for all model parameters.

Our results also reveal a noticeable pile-up problem at zero. This problem may be expected for Student-*t* statistics, as discussed by Stock and Watson (1998) for the special case where the TVP specifications follow a unit-root. For this reason, one must avoid over-interpreting the close-to-zero

Table 10. Residual-based normality tests: Monte Carlo study

Model		Test on ε_t	Test on v_{1t}	Test on v_{2t}
TVP-intercept-trend	Oil	4.9	35.5	90.0
	Coal	4.2	100	100
	Gas	0.0	100	100
TVP-trend	Gas	63.8	–	2.3

Note – *Empirical rejections* refer to the proportion of simulated statistics that exceed the $\chi^2(2)$ critical value for a test with level 5%. For the definitions of the models for the definitions of the models denoted *TVP-intercept-trend*, and *TVP-trend*, see notes to Table 1. Statistics are simulated for given these models, with $\hat{\omega}_{TVP}$ as calculated from the indicated price series.

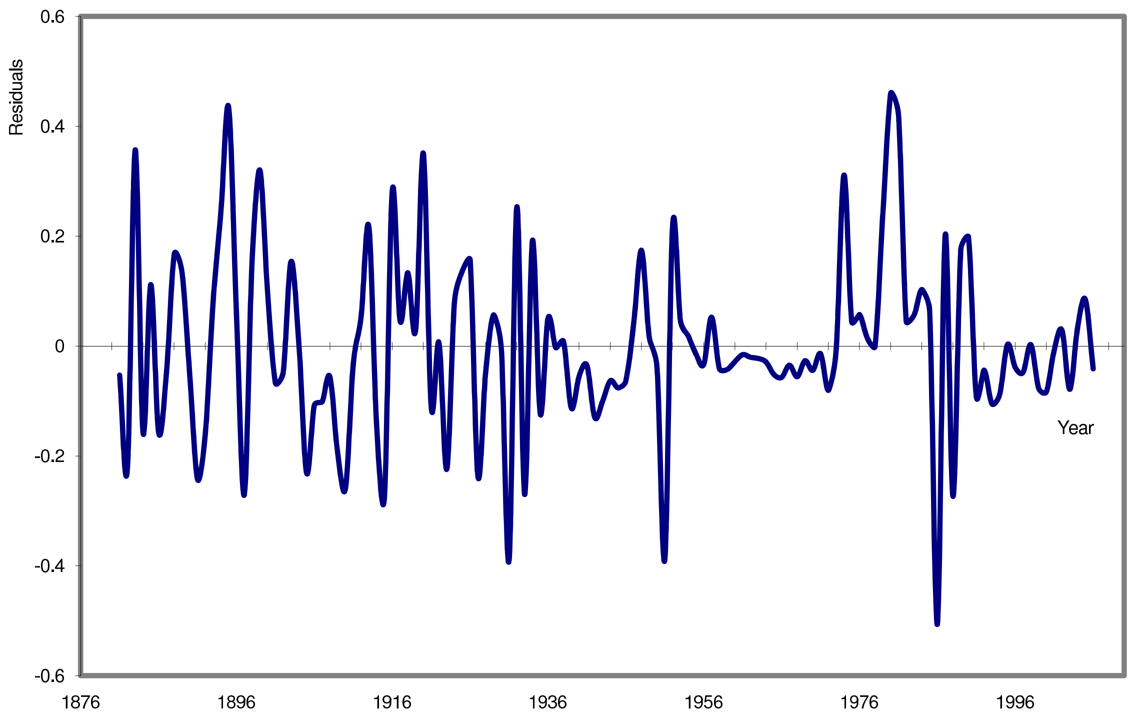
t-statistics in Tables 1 and 4. Here we find that LR statistics to test TVP may also have a mass point at zero. Overall, the Monte Carlo study of the LR test suggests that: (i) usual critical values can lead to spurious detections of TVP effects, and (ii) considering simulation-based alternatives is particularly worthwhile when the sample size relative to the number of parameters is tight.

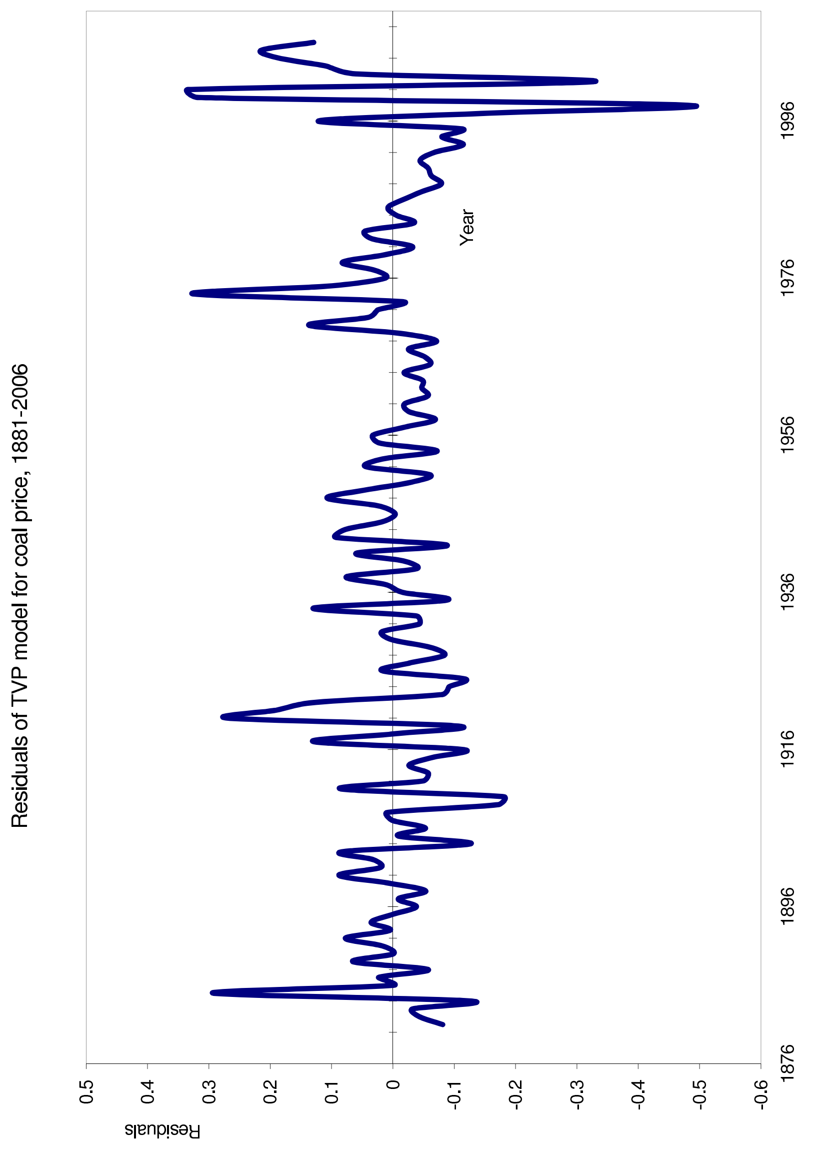
Monte Carlo results for the normality tests are reported in Table 10. Here again, we see that the rejection rates for the χ^2-based test fluctuate from 0 to 100%, even with the single-regime designs; this implies that the usual Jarque-Bera χ^2 test is inappropriate for TVP-filtered residuals. This (alarming) fact does not seem to be known in this literature. Given the popularity of such tests in practice, our results - that are particularly revealing because we have modeled our simulation design based on empirical data - suggest that bootstrap-based alternatives are clearly called for. Consequently, for the diagnostic tests applied (and discussed below), we consider MMC p-values when commonly used asymptotic procedures are significant, to guard against spurious rejections.

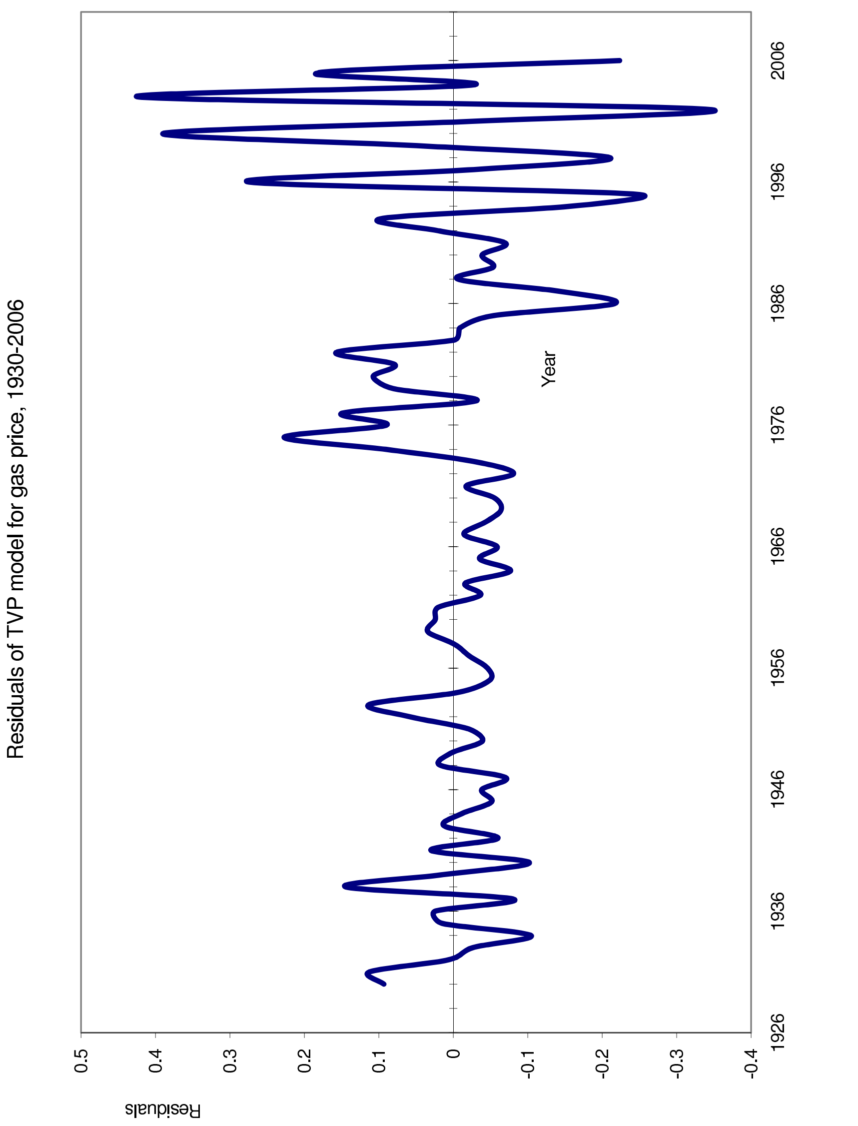
B. Residuals of TVP models

The graphs provided in this section give the residuals of the four TVP models whose estimated coefficients appear in tables 1 and 4.

Residuals of TVP model for oil price, 1881-2006




Residuals of TVP model for coal price, 1881-2006
Residuals
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
1876
1896
1916
1936
1956
1976
1996
Year



Residuals of TVP model for gas price, 1930-2006
Residuals
Year
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
1926
1936
1946
1956
1966
1976
1986
1996
2006

Residuals of TVP model for gas price; two variance case, 1930-2006