2020-06

Working paper. Economics

ISSN 2340-5031

FORECASTING GASOLINE PRICES WITH MIXED RANDOM FOREST ERROR CORRECTION MODELS.

Álvaro Escribano, and Dandan Wang.

uc3m | Universidad Carlos III de Madrid

Serie disponible en http://hdl.handle.net/10016/11

Web: http://economia.uc3m.es/

Correo electrónico: departamento.economia@eco.uc3m.es

Forecasting Gasoline Prices with Mixed Random Forest Error Correction Models

June 2, 2020

Álvaro Escribano[1] (Department of Economics, UC3M)

and

Dandan Wang (Department of Statistics, UC3M)

ABSTRACT

The use of machine learning (ML) models has been shown to have advantages over alternative and more traditional time series models in the presence of big data. One of the most successful ML forecasting procedures is the Random Forest (RF) machine learning algorithm. In this paper we propose a mixed RF approach for modeling departures from linearity, instead of starting with a completely nonlinear or nonparametric model. The methodology is applied to the weekly forecasts of gasoline prices that are cointegrated with international oil prices and exchange rates. The question of interest is whether gasoline prices react asymmetrically to increases in oil prices rather than to decreases in oil prices, the "rockets and feathers" hypothesis. In this literature most authors estimate parametric nonlinear error correction models using nonlinear least squares. Recent specifications for nonlinear error correction models include threshold autoregressive models (TAR), double threshold error correction models (ECM) or double threshold smooth transition autoregressive (STAR) models. In this paper, we describe the econometric methodology that combines linear dynamic autoregressive distributed lag (ARDL) models with cointegrated variables with added nonlinear components, or price asymmetries, estimated by the powerful tool of RF. We apply our mixed RF specification strategy to weekly prices of the Spanish gasoline market from 2010 to 2019. We show that the new mixed RF error correction model has important advantages over competing parametric and nonparametric models, in terms of the generality of model specification, estimation and forecasting.

JEL: B23, C24, C52, C53, D43, L13, L71.

Keywords: Forecasting gasoline prices, Rockets and feathers hypothesis, Cointegration, Nonlinear error correction, Machine learning, Random forest, Mixed random forest.

[1] The first author acknowledges the funding received from the Ministry of Economics of Spain (ECO2016-00105-001, MDM 2014-0431), the Community of Madrid (MadEco-CM S2015/HUM-3444) and the Agencia Estatal de Investigación (2019/00419/001) as well as the comments received in the workshop on "40 years of Cointegration" organized by FUNCAS where a preliminary version of this paper was presented.

1. Introduction

Since the seminal papers of cointegration and linear error correction models of Granger (1981) and Engle & Granger (1987), a large body of literature has emerged that aims at cointegration among economic variables. The first extension to nonlinear error correction (NEC) models was proposed by Escribano (1985, 1986) when modeling the evolution of UK money demand from the nineteenth century onwards. Since then, several authors such as Hendry & Ericsson (1991), Teräsvista and Eliasson (2001), and Escribano (2004), have used a nonlinear error correction approach to obtain stable UK money demand parameter estimates. Other economics examples using asymmetric or nonlinear error correction models include: the relationship between production, sales and inventory, (Granger & Lee, 1989); the "rockets and feathers" hypothesis of how international oil prices are transmitted to gasoline prices in most countries when firms have market power, (Borentein, Cameron, & Gilbert, 1997); asymmetries in labor markets, (Escribano & Pfann, 1998); asymmetries in gold and silver prices (Escribano & Granger, 1998) and asymmetries in the terms structure of interest rates (Enders & Granger, 1998).

All those models are dynamic and parametric models capturing departures from linearity with simple nonlinear equilibrium correction specifications affecting a single explanatory variable. The exception to this dynamic approach is based on the methodology of Teräsvirta (1994) that allows the whole dynamics of the linear model to affect the departures from linearity with smooth transition autoregressive (STAR) models. To estimate those models, a few parametric functions are used such as the exponential function or the logistic function (transition functions). To select between the those two parametric functions, several decision rules have been proposed by Teräsvirta (1994) and Escribano & Jorda (1999, 2001).

The goal of this paper is to contribute to this literature on nonlinear error correction models by proposing a more general dynamic methodology that maintains the basic idea of modeling departures from linearity, but uses the advantages of a powerful nonlinear approach suggested in the machine learning (ML) literature; the random forest (RF) approach. This RF modeling approach extends the classic regression trees approach, by using bootstrap aggregating or bagged decorrelated trees, to avoid over-fitting with highly correlated trees, in the bootstrapped training samples. The RF model is later tested in the test set. Our approach is not the usual pure RF approach, but instead we suggest a new mixed RF forest approach that combines the information learned from: a) usual time series techniques, that has

proved to forecast well when there is no abrupt changes in the economy, with b) a flexible RF approach that is able to identify the main variables related to those structural changes in the economy, and also is able to generate more robust estimates than the usual parameter estimates provided by the nonlinear least squares (NLS) approach.

The structure of the paper as follows: Section 2 briefly reviews the empirical evidence of the asymmetric reactions in the fuel market to changes in international oil prices. Section 3 discusses the cointegrating relationship that exists in the gasoline market in Spain, specifies the corresponding parametric nonlinear error correction specifications used in the literature, and explores the consumer complaint that retail gasoline prices react faster to crude oil price increases than to crude oil price decreases. Machine learning (ML) methods are briefly introduced in Section 4. Section 5 includes the main results from the empirical application of Logistic error correction models, RF error correction models, and the new mixed RF error correction approach. Section 6 discusses the forecasting comparison of those three approaches, and shows how the new mixed RF approach outperforms the rest. In Section 7, we present the main conclusion and consider further extensions.

2. Empirical evidence of asymmetric reactions in the fuel market to changes in international oil prices

Price determination in the fuel market has been a controversial issue during the last decades. In most economies, the fuel sector has been accused of having high market power and, as a result, of carrying non-competitive practices. When input costs increase, output prices increase at a faster rate; however, when input costs decrease, output prices adjust more slowly. This phenomenon is known formally as “asymmetric price transmission” and informally as “rockets and feathers”. The existence of these asymmetries is undesirable since they can be detrimental for consumers and can lead to efficiency losses.

Asymmetric price transmission has been studied for several markets across the world but it has been focused on the fuel market. In the case of the US, different studies have focused on this topic. The most common way to tackle the presence of the rockets and feathers phenomenon is by following the pioneering work of Borenstein et al. (1997). Using weekly data for different states during the period of 1986-1990, those authors created an Autoregressive Distributed Lag (ARDL) and an Error

Correction Model (ECM). The results of these empirical models show that the adjustments of spot and retail gasoline prices to changes in weekly crude oil prices are asymmetric. However, when using daily data for the same period, Bachmeier & Griffin (2003) detect no asymmetries in price transmission. Balke et al. (1998) found evidence for a persistent asymmetry by using an ECM during the period of 1987-1997. Deltas (2008) showed that retail gasoline prices respond faster to wholesale price increases than wholesale price decreases. He suggests that sticky prices and response asymmetries are the consequences of retail market power. If market power leads to higher price-cost margins, it is more likely that the price dynamics would tend to be beneficial rather than detrimental regarding profits. Furthermore, his results are consistent with different forecast methods, which show better accuracy for asymmetric models rather than symmetric models. Johnson (2002) analyzed central heating oil and gasoline price responses to changes in crude spot levels for 15 North American states. The results confirm that gasoline prices respond asymmetrically to crude oil price changes, while central heating oil reacts symmetrically. Other studies have investigated the "rockets and feathers" phenomenon in South America. Balmaceda (2008) studied the asymmetry price transmission for the case of Chile, where there is a unique refined public firm (ENAP) which belongs to the state. Instead of using time-series, a panel of weakly data regarding several service stations over time is applied. This analysis showed that the best model with which to study the asymmetry is an error correction model. Under this method, the rockets and feathers theory is met. Additionally, the study reveals that independently of the margins and geographical differentiation of service stations, the asymmetry in prices remains the same.

With respect to Europe there are also several studies. The work of Bacon (1991) is known for being one of the pioneering studies on this topic with his introduction of the term "rockets and feathers". Bettendorf et al. (2009) address the analysis for the Dutch retail gasoline market by using an Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) model. The paper shows that the volatility process is not symmetric: a negative shock to the retail price has a lower impact on the variance of the retail price than a positive shock. Further evidence of asymmetric GARCH models is provided by Torrado & Escribano (2020) for the Spanish market. Galeotti et al. (2003) examined the adjustments of the retail price of gasoline when a shock to crude oil prices occurs. The main contribution is that different countries are compared using a two-stage approach for the transmission mechanisms with the aim of resolving whether the asymmetry is at the refinery level, at the distribution level, or at both stages, by applying an ECM and also bootstrapping. The

study found symmetric pricing for Germany, Italy and the U.K., and asymmetric transmissions in the case of France and Spain using the single stage approach and the second stage of the two stage approach. The Spanish case has not been widely studied. Cotín-Pillart (2008) replicates the cumulative response functions and the ECM model developed by Borenstein et al. (1997) for the Spanish market during the periods of 1993-1998 and 1998-2005. For the first period, changes in spot gasoline prices are completely translated into retail price changes but in a symmetric way. Nevertheless, the second period shows asymmetric responses of retail prices to the spot price. More recent studies were undertaken by Jiménez & Perdiguero (2005), Perdiguero (2010), and Torrado & Escribano (2020).

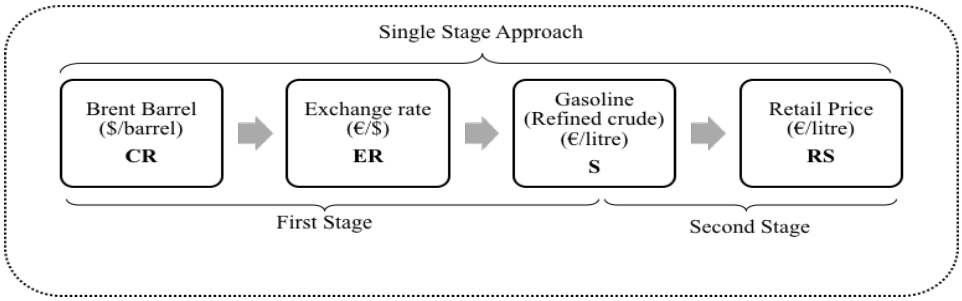
3. Cointegration and nonlinear error correction models of the Spanish fuel market

In order to estimate the potential asymmetric price transmissions in the Spanish fuel market, weekly data from January 8, 2010 until November 11, 2019. For the out-of-sample forecasting period we use the extended data from November 3, 2017 until November 11, 2019. The variables employed in this paper are the price of crude oil (**CR**) *in $/barrel of the Brent Spot Price FOB*, the *€/$ exchange rate* (**ER**), and the *pre-tax retail price of 95 octane gasoline* (**RS**), or Spanish retail price of gasoline in €. The abbreviations of the variables in capital letters refer to the series in levels and lowercase letters represent variables in *logarithms* (**cr, er, and rs**). For crude oil prices, the weekly Europe Brent Spot Price FOB ($ per barrel) from the US Energy Information Administration (EIA) is used and transformed into $/1000L. The exchange rate is obtained from the FRED. The pre-tax retail price (RS) chosen is an average of the 95 octane gasoline prices (€/liter) provided by the Spanish National Commission of Markets and Competence (CNMC is the Spanish acronym), which is also transformed[2] into €/1000L.

The fuel market is a complex sector made up of several stages from the extraction of crude oil until the production, distribution, and sale of gasoline to final consumers, see Fig. 1.

[2] The unit conversion of the spot price of gasoline is derived by applying the following formula: $P_{gasoline}$ (€/1000L) = [$P_{gasoline}$ (€/t)] * [740 (kg/m^3) /1000].

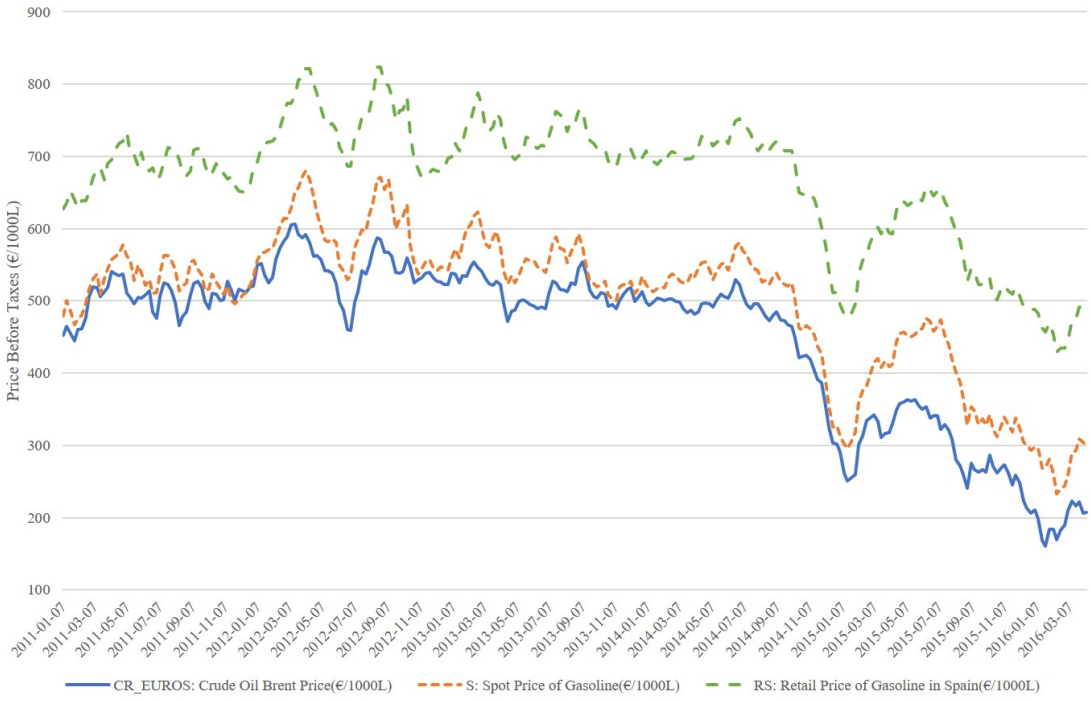
Fig. 1. Price formation along the production-distribution chain.



Note. The single stage approach establishes how final retail prices (RS) in € are related (cointegrated) with crude oil prices (CR) in $ and the exchange rate (ER in €/$).

The plots of the price variables in Fig. 2 show the relationship between crude oil prices and gasoline prices at different stages of the production-distribution chain for the 95-octane gasoline. The difference between the price series (S) and (CR_EUROS) can be a proxy of the margin in the first stage of the production-distribution chain, whereas the price difference between (RS) and (S) is a proxy for the margin in the second stage of Fig. 1.

Fig. 2. Evolution of the prices of the oil production-distribution chain.



The vast majority of the rockets and feathers investigations developed an ECM methodology due to the presence of cointegration between downstream and upstream prices.

Long-run equilibrium prices and cointegration

To model the gasoline market, we focus on a single-stage process. Escribano & Torrado (2018) also discuss the other two stages considered in Fig. 1.

Given that all the variables are I(1) or have a unit root, if a linear combination of them is stationary, I(0), then the variables are cointegrated. Recall that lowercase letters indicate that variables are in log form.

$$rs_t = rs_t^* + ecm_t = E(rs_t / x_t, \alpha) + ecm_t = \alpha_0 + \alpha_1 cr_t + \alpha_2 er_t + ecm_t \qquad (1)$$

If $ecm_t>0$ (ecm_t^+) this means that the actual price log price, rs_t , is above the expected long-run equilibrium price $rs_t^* = E(rs_t / x_t, \alpha) = \alpha_0 + \alpha_1 cr_t + \alpha_2 er_t$ and therefore future prices should decrease to reduce the disequilibrium. When $ecm_t<0$ (ecm_t^-) the actual log-price rs_t is below the expected long-run equilibrium log-price and we expect future prices to increase to correct the disequilibrium.

The single-stage cointegration equation is given by

$$rs_t = rs\alpha_0 + \alpha_1 cr_t + \alpha_2 er_t + ecm_t \ . \qquad (2)$$

In Eq. (3) we allow the asymmetric behavior to affect both the dynamics and the error correction terms (ecm_{t-1}) in a nonlinear way. Books adequately reviewing the modeling of nonlinear time series models and nonlinear error correction models are by Dufrénot & Mignon (2002) and Teräsvirta, Tjostheim, & Granger (2010).

According to the Granger Representation Theorem, the presence of a cointegrating relation implies that a valid ECM exists. However, it is not clear whether the error correction adjustment (equilibrium correction) is linear as in Engle & Granger (1987) or is nonlinear/asymmetric as in Escribano (1986, 2004), Escribano & Granger (1998), and Escribano & Pfann (1998).

Starting from the seminal work of Teräsvirta (1994), and the extensions discussed in Teräsvirta, Tjostheim, & Granger (2010), we consider a nonlinear autoregressive distributed lag (ARDL), also called smooth transition autoregressive (STAR) model, in the form of a nonlinear and time-varying error correction model.

STAR model affecting the Dynamics and the Error Correction term (STAR)

$$\Delta rs_t = a_0 + a_p(L)\Delta rs_{t-1} + a_q(L)'\Delta x_t + \\ + F(ecm_{t-1}, \Delta cr_t, \gamma, \beta)(c_0 + c_r(L)\Delta rs_{t-1} + c_m(L)'\Delta x_t) + \varepsilon_t \quad (3)$$

where $a_i(L)$ and $c_j(L)$ in Eq. (3) are finite order polynomials in the lag operator L, with all roots outside the unit circle.

The results of the study by Escribano & Torrado (2018) show that the main source of asymmetries in the gasoline prices are coming from the impact of Δcr_t on the ecm_{t-1} terms and not from the dynamic part of the autoregressive distributed lag (ARDL) variables of Eq. (3). In the empirical application described in this paper to the gasoline prices in Spain, we also find a linear behavior affecting the dynamics of the variables Δrs_t and Δx_t, but a nonlinearly behavior affecting the error correction terms from two main sources; (i) the sign of the ecm_{t-1} and (ii) whether the crude oil (cr) is increasing ($\Delta cr_t^+ = \Delta cr_t$ if $\Delta cr_t \geq 0$ and 0 otherwise) or decreasing ($\Delta cr_t^- = \Delta cr_t$ if $\Delta cr_t < 0$ and 0 otherwise).

Bivariate Nonlinear Error Correction (NEC)

$$\Delta rs_t = a_0 + a_p(L)\Delta rs_{t-1} + a_q(L)'\Delta x_t + F(ecm_{t-1}, \Delta cr_t, \gamma, \beta) + \varepsilon_t \quad (4)$$

*4th-order polynomial Model (**4th POL-ECM**), Escribano[3] (1986, 2004)*

$$F(ecm_{t-1}, \Delta cr_t, \beta) = \beta_{01} ecm_{t-1} + \beta_1 ecm_{t-1}(\Delta cr_t) + \\ + \beta_2 ecm_{t-1}(\Delta cr_t)^2 + \beta_3 ecm_{t-1}(\Delta cr_t)^3 + \beta_4 ecm_{t-1}(\Delta cr_t)^4 \quad (5)$$

The starting point of the nonlinear model specification is with a general Taylor series expansion of the function F(.,.) in Eq. (4) given by the 4th order polynomial of Eq. (5),

*4th-order polynomial Model with double threshold (**4th POL-DT-ECM**), Escribano & Torrado (2018)*

$$\Delta rs_t = a_0 + a_p(L)\Delta rs_{t-1} + a_q(L)'\Delta x_t + \beta_{01} ecm_{t-1} + \beta_1 ecm_{t-1}(\Delta cr_t) + \\ + \beta_2 ecm_{t-1}(\Delta cr_t)^2 + \beta_3 ecm_{t-1}(\Delta cr_t)^3 + \beta_4 ecm_{t-1}(\Delta cr_t)^4 + \varepsilon_t. \quad (6)$$

Following Teräsvirta (1994) the first step is to test for linearity in Eq. (6) by testing the null hypothesis that $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative of

[3] In the original nonlinear error correction of Escribano (1986, 2004) applied to the UK money demand since the 19th century, the Taylor series expansion was based on a cubic polynomial expansion of the ecm_{t-1} term instead of the variable (Δcr_t). Escribano & Torrado (2018) used this alternative specification in the Spanish fuel market but the result was statistically non-significant.

$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0$, or against the alternative hypothesis that at least some of those coefficients are different from 0. If the null hypothesis of linearity is rejected, the next step suggested in Teräsvirta (1994) follows a decision rule to select between Logistic or Exponential smooth transition specification.

Escribano & Jorda (1999, 2001) suggested the following more powerful decision rule based on Eq. (6). If the minimum p-value is obtained, rejecting the null hypothesis $H_{0L} : \beta_1 = \beta_3 = 0$, then the model is Logistic STAR. However, if the minimum p-value is obtained, rejecting the null hypothesis $H_{0E} : \beta_2 = \beta_4 = 0$, then the model is Exponential STAR.

Six different linear and nonlinear parameterizations of error correction models were estimated in Escribano & Torrado (2018) with linear and nonlinear equilibrium correction terms and the class of selected models was the following Logistic-ECM.

*Smooth Transition Error Correction Models (**LOGISTIC-ECM**)*

$$F(ecm_{t-1}, \Delta cr_t, \gamma, \beta) = \beta ecm_{t-1} \left(\frac{1}{1 + \exp(-\gamma \Delta cr_t)} \right). \tag{7}$$

The nonlinear specification of Eqs. (3) and (7) are discussed in Teräsvirta & Eliasson (2001) with a different dataset and was also considered in Escribano & Torrado (2018). However, they found that model (7) or the bivariate double threshold extension of model (7) provided the best nonlinear models for weekly gasoline prices in Spain.

*Double Threshold Logistic Error Correction Models (**DT-LOGISTIC-ECM**)*

$$\begin{aligned} F(ecm_{t-1}, \Delta cr_t, \gamma, \beta) = \beta_1^+ ecm_{t-1}^+ \left(\frac{1}{1 + \exp(-\gamma \Delta cr_t)} \right) + \\ + \beta_2^- ecm_{t-1}^- \left(\frac{1}{1 + \exp(-\gamma \Delta cr_t)} \right). \end{aligned} \tag{8}$$

To do that, they started with the following Taylor series approximation of Eq. (4) and Eq. (8) given in Eq. (9),

$$\Delta rs_t = a_0 + a_p(L)\Delta rs_{t-1} + a_q(L)'\Delta x_t + \beta_{01} ecm_{t-1} +$$
$$+ \beta_{1p} ecm^{+}_{t-1}(\Delta cr_t) + \beta_{2p} ecm^{+}_{t-1}(\Delta cr_t)^2 + \beta_{3p} ecm^{+}_{t-1}(\Delta cr_t)^3 + \beta_{4p} ecm^{+}_{t-1}(\Delta cr_t)^4 + \quad (9)$$
$$+ \beta_{1n} ecm^{-}_{t-1}(\Delta cr_t) + \beta_{2n} ecm^{-}_{t-1}(\Delta cr_t)^2 + \beta_{3n} ecm^{-}_{t-1}(\Delta cr_t)^3 + \beta_{4n} ecm^{-}_{t-1}(\Delta cr_t)^4 + \varepsilon_t.$$

The decision rule of Escribano & Jorda (1999, 2001) applied to Eq. (9) is the following: if the minimum p-value is obtained rejecting the null hypothesis $H_{0L}: \beta_{1j} = \beta_{3j} = 0$, for j=p or n, then the models are Logistic STAR. However, if the minimum p-value is obtained, rejecting the null hypothesis $H_{0E}: \beta_{2j} = \beta_{4j} = 0$ for j=p or n, then the model is Exponential STAR. Based on this decision rule they selected the Logistic STAR model for gasoline prices in Spain.

4. Random forest methods

Machine learning (ML) methods have cast light on the data analysis. Medeiros et al. (2019) claim that the machine learning gains in mean squared errors reach up to 30% in the US CPI inflation in the two years out-of-sample forecasting compared to the traditional random walk, autoregressive, and unobserved components of stochastic volatility models. In particular the random forest (RF) (Breiman, 2001), outperforms compared to other machine learning methods, i.e., deep neural networks, boosted trees, and a polynomial model estimated either by LASSO or adaLASSO.

Compared to the decision tree algorithm, in the RF the processes of finding the root node and splitting the feature nodes run randomly. The RF model randomly searches the models that fit the subset of the dataset instead of searching the best model that fits the whole dataset, hence the variance in models of trees are reduced. It leads to a reduction in the possibility of overfitting and the building of models that are better trained for future predictions. These good properties received much attention in the field of economics by such as Scornet et al. (2015) and Wagner & Athey (2018).

As an ensemble method, the RF using the tree bagging process tends to yield the high accuracies, by combining the predictions from multiple machine learning algorithms. Here, we present a brief picture about how random forest works, taking 600 trees as the example. The RF uses the bagging process rather than the

boosting process, which means that the trees in random forests are run in parallel. There is no interaction between these trees while building the trees.

Fig. 3 represents the basic structure of a random forest (RF) where:

(a) the RF randomly selects “k” features from total “m” features where k<<m, for example, in section 5.2, $\Delta \mathrm{rs}_t = \mathrm{RF}(\Delta \mathrm{cr}_t, \Delta \mathrm{cr}_{t-1}, \Delta \mathrm{er}_t, \Delta \mathrm{er}_{t-1}, \Delta \mathrm{rs}_{t-1}, \Delta \mathrm{rs}_{t-2}, \mathrm{ecm}_{t-1}) = \mathrm{RF}(X_t)$, m=7 features in the RF,

(b) then among these “k” features, the node (Tree 1 the first node) is obtained by using the best split (the splits are determined to minimize the sum of squared errors),

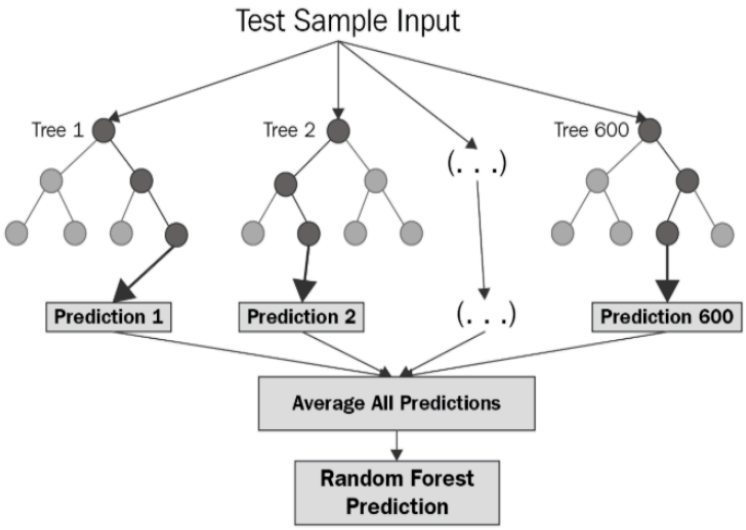
(c) splitting the node into daughter nodes using best split,

(d) following steps (a) to (c), until reaching “l” number if nodes,

(e) by repeating steps (a) to (d) “n” (600 trees) number times (bagging procedure), then the “n” number of trees are built,

(f) the final prediction would be the average of predictions in all these “n” (600) trees.

Fig. 3. Random Forest Structure.



Biau & Scornet (2016) summarized the RF algorithm in Breiman L(2001). There are 4 important parameters in the RF mechanism in the RF setting in the software:

(1) $a_N \in \{1,\ldots,N\}$: the number of sampled data points in each tree, N is total number of observations in the training dataset;

(2) mtry ∈ {1, . . . , m}: the number of possible directions for splitting at each node of each tree, m is the total available features;

(3) nodesize ∈ {1, . . . , a_n }: the number of examples in each cell below which the node is not split.

(4) n: number of trees.

These parameters are selected according to the minimum mean squared errors.

The RF builds robust models (trees), improving the accuracies. However, it is computationally very expensive, since many trees need to be made during the training process (Singh, 2018). The process (Mixed-RF) we wish to build would be able to reduce the tree size.

In economics, for decades the majority of the literature has a focus on the functionals of its distribution in parametric and non-parametric approaches. However, a mis-specified parametric model might lead to misleading results and well-known over-fitting problems, while non-parametric modeling suffers from less precision and the curse of dimensionality. Robinson (1988) offers an interesting semi-parametric modeling approach which provides consistent estimators when modeling $\beta' X_t + \theta(Z_t)$, where $\theta(Z_t)$ is the unknown non-linear function and X_t is neither completely dependent on Z_t nor necessarily independent on it. An alternative approach, with flexible correction terms including power series similar in spirit to our approach that also uses power series, was suggested in Newey (2009).

In most multivariate time series models, the dynamic behavior of economic variables is explained by a combination of linear dynamic parametric forms with nonlinear forms measuring temporal departures from linearity, as in Eq. (3). See Teräsvirta, Tjostheim, & Granger (2010) for a general discussion of this approach. Based on this mixed approach we suggest a new semi-parametric modeling.

In this paper, we propose a new mixed strategy to specify dynamic RF models. James & Wineland (2010) show that the mixed dynamic linear model, using OLS+RF, performs much better than adaLASSO+RF and RF in modeling and out-of-sample forecasting when the defined linear part is independent of the unknown pattern function $\theta(Z_t)$. However, in our approach, we allow for some dependency between the specified linear part of the models and the non-linear parts. We show that the linear part converges after the first iteration. For interpretation purposes and to open the black-box of the RF, we study the features of importance, the individual condition expectations, and the interaction between features and dependent variables.

5. Empirical application to the gasoline market

In order to compare the estimation and forecasting ability of different procedures, we focus on the logistic-ECM, the random forest, and the new procedure to estimate a mixed random forest based on the single-stage co-integration relation of gasoline prices provided by Eq. (2).

The weekly data set starts on January 8, 2010 and ends on November 11, 2019. The first 405 observations are used as the *training dataset* (from January 8, 2010 to October 27, 2017), and the *test dataset* (Forecasting period) runs from November 3, 2017 to November 11, 2019 with 102 forecasting periods.

5.1 LOGISTIC-ECM model (Non-linear least squares)

In this section, we look at the parametric estimation of the logistic-ECM model of Eq. (7) using the non-linear least squares (NLS) estimation, with the initial values set up according to Escribano & Torrado (2018). The results of estimating model (10) by NLS are included in Table 1 and Fig. 5.

$$\Delta rs_t = \hat{a}_0 + \hat{a}_q(L)'\Delta x_t + \hat{a}_p(L)\Delta rs_{t-1} + \hat{\beta} ecm_{t-1}\left(\frac{1}{1+\exp(-\hat{\gamma}\Delta cr_t)}\right) + \hat{\varepsilon}_t . \qquad (10)$$

Table 1 Nonlinear error correction model of the rate of growth of gasoline prices (Δrs_t)

Explanatory variables	Coefficients	Std. errors	t-ratio	P-values
Δcr_t	0.336	0.019	17.85	< 2e-16 ***
Δcr_{t-1}	0.057	0.026	2.23	0.02612 *
Δer_t	0.404	0.066	6.08	2.76e-09 ***
Δer_{t-1}	0.143	0.070	2.05	0.04103 *
Δrs_{t-1}	0.142	0.048	2.96	0.00323 **
Δrs_{t-2}	0.078	0.036	2.16	0.03135 *
Logistic	β = -0.158	0.032	-4.97	9.86e-07 ***
β**ecm$_{t-1}$*[1/(1+exp(-$\gamma\Delta cr_t$))]**	γ = 48.75	39.374	1.24	0.21637
Significance: 0 '***', 0.001 '**', 0.01 '*'				

Escribano & Torrado (2018) show that in this parametric logistic-ECM model the ecm at t-1 interacts (nonlinear equilibrium correction) with the rate of change of international oil prices (Δcr_t) thus affecting the evolution of gasoline prices, see Fig. 5.1. The gasoline price reaction to ecm_{t-1} is significant but small when the price of crude oil is decreasing but when the crude oil price is increasing, the gasoline price adjustment to the ecm_{t-1} term is significant and much faster, see Fig. 5.2. Fig. 5.3 shows that the gasoline price equilibrium reaction (ecm_{t-1}) is non-linear and changes with the level of (Δcr_t).

Fig. 5.1 Single-stage logistic adjustment function.

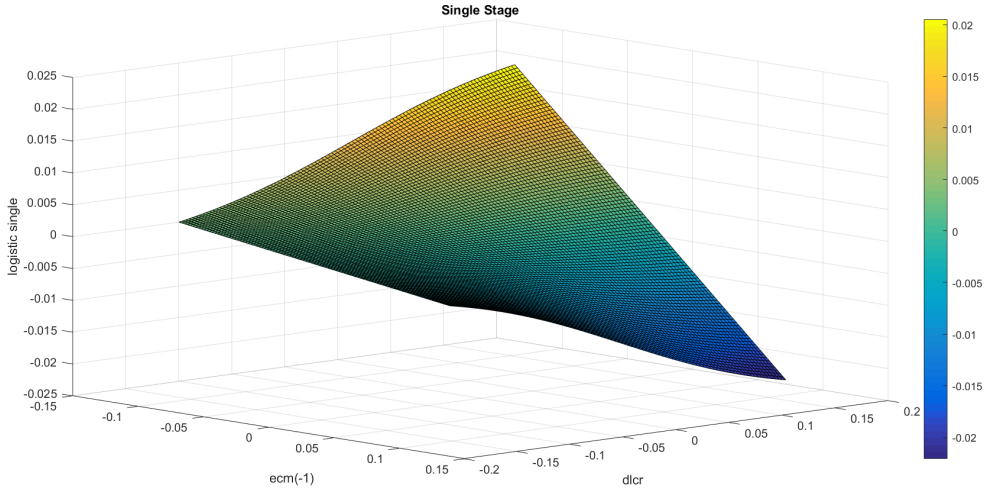


Fig. 5.2 Single-stage logistic adjustment function maintaining ecm_{t-1} fixed.

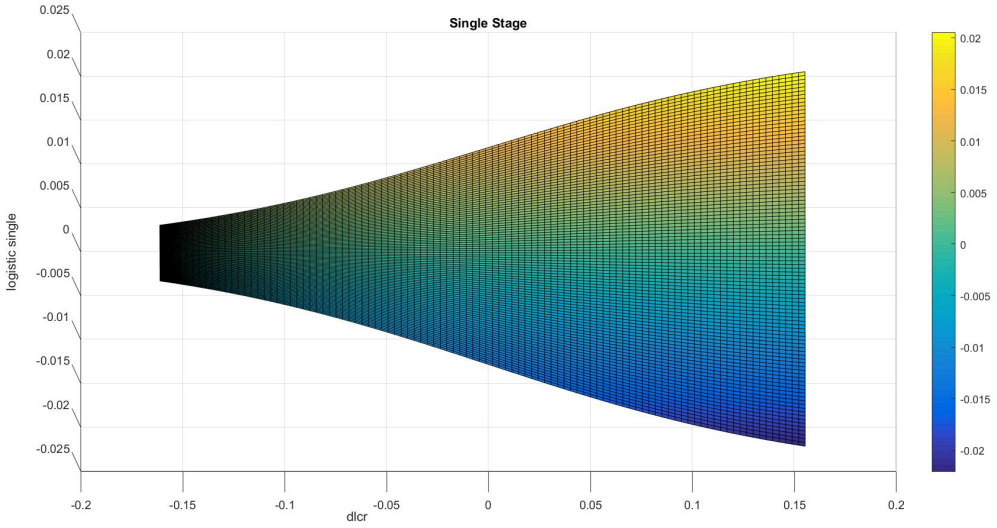
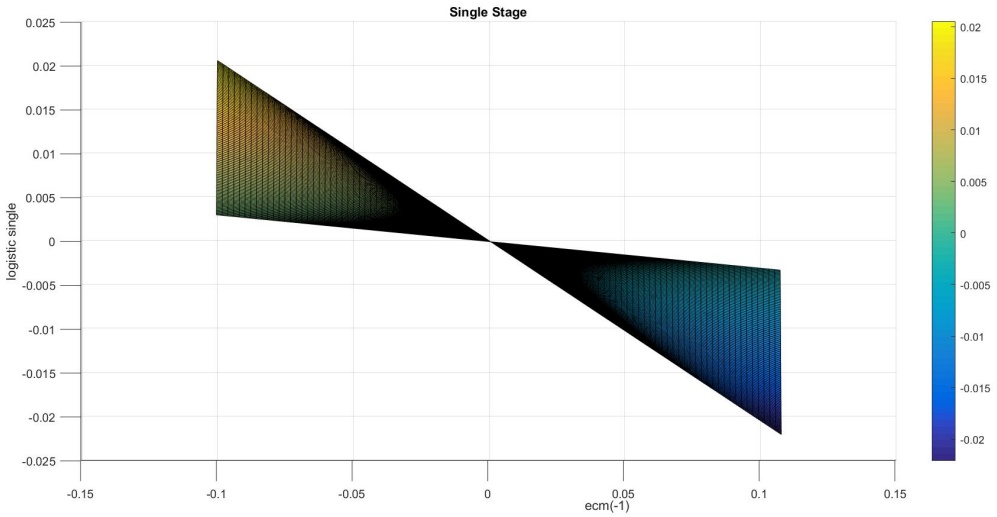


Fig. 5.3. Single-stage logistic adjustment function maintaining Δcr_t fixed.



5.2 Random forest

We consider the features that appear in Section 5.1, Table 1, hence:

$$\Delta \mathrm{rs}_t = \mathrm{RF}(\Delta \mathrm{cr}_t, \Delta \mathrm{cr}_{t-1}, \Delta \mathrm{er}_t, \Delta \mathrm{er}_{t-1}, \Delta \mathrm{rs}_{t-1}, \Delta \mathrm{rs}_{t-2}, \mathrm{ecm}_{t-1}) = \mathrm{RF}(X_t) \quad (11)$$

Based on the least root mean squared errors and highest R-squared, the best parameters for pure Random Forest (RF) are: the number of variables available for splitting at each tree node(mtry) is 7, the minimum node size is 5, the maximum node size is 49, and the number of trees is 2000.

The importance of each feature is presented in Table 2. The ecm_{t-1} and Δcr_t are the two most important features in the RF.

Table 2 Random Forest variable importance

Variable	Overall
Δcr_t	151.50
ecm_{t-1}	37.87
Δcr_{t-1}	36.09
Δer_t	34.45
Δrs_{t-1}	23.90
Δer_{t-1}	14.89
Δrs_{t-2}	8.31

We then extract the best model in the RF to visualize and interpret random forest models. The order of the features is sorted in the reading direction by importance of variables. In Fig. 6, the y-axis has the cross validated (leave-one-out cross validation, in each process the optimized parameters are performed on n-1 of n dataset pairs, and then the performance of the tuned algorithm is tested on the pair that have been left out, then repeating this process n times) contributions of each x-feature, i.e., the change of the predicted probabilities for different values of each x-feature. The plotting illustrates the main effects, as contributions by each feature were plotted against their respective feature values. In each sub-graph in Fig. 6, the x-axis represents the feature values of each regressor in vector X_t, see Eq. (11), and the black line represents the partial functions of each regressor on the Δcr_t. By employing the leave-one-out k-nearest neighbor gaussian kernel estimation, the goodness of fit was obtained (R-squared) for each regressor to the dependent variable (Δcr_t), which evaluates how

well each feature contribution can be explained as a main effect. From Fig. 6, we see clearly that Δcr_t is the main effect when explaining gasoline prices, which is with R-squared=0.99.

The main effect of Δcr_t, indicated by a non-linear pattern (S shape) of the first sub-graph in Fig. 6, has a Logistic shape function of the contribution of Δcr_t to the rate of change in gasoline prices, Δrs_t. The second graph (in the reading direction) in Fig. 6 shows the error correction contribution (ecm_{t-1}) contribution to the rate of change in gasoline prices and as expected it is decreasing, as was also obtained in Fig. 5.3, but the decreasing is slow. The rest of the regressors present the positive effect on the gasoline prices and the partial effects of each regressor seem more linear than Δcr_t when the value of the regressor changes. A color gradient along the most influential feature, Δcr_t (cr_diff1 on the graph), was applied to search for interactions, for example in the first graph in Fig. 6. The observations in red color help discover the interactions with the other independent variables also presenting in red color in graphs 2 to 7 of Fig. 6 (in the reading direction). We can observe that there is latent interaction with ecm_{t-1}.

The interaction between Δcr_t and ecm_{t-1} and the non-linear pattern (S-shape) of Δcr_t, confirms the empirical results of Escribano & Torrado (2018) that, in the parametric modeling, the logistic-ECM fits the data better compared to the rest of the models, where the nonlinear error correction term is given by $\beta ecm_{t-1}[1/(1+\exp(-\gamma \Delta cr_t))]$ in Eq. (10).

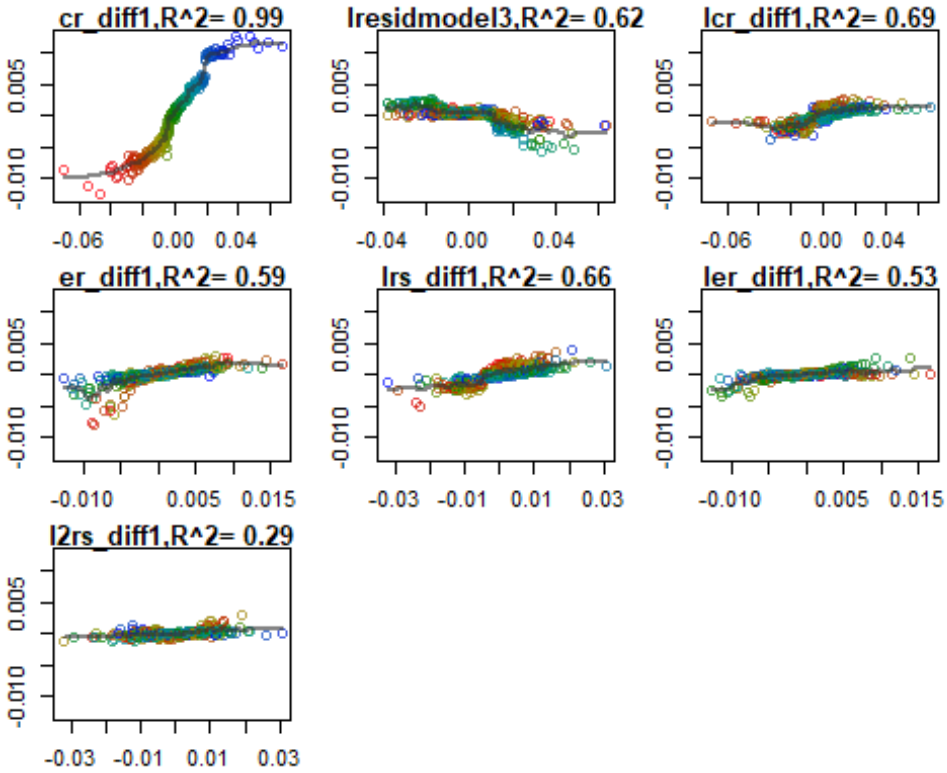
In Fig. 6 the color gradient suggests, that Δcr_t interacted with ecm_{t-1} due to the vertical color gradient in the plot of ecm_{t-1}. In Fig. 7 their combined feature contributions were plotted in the context of both features, Δcr_t and ecm_{t-1}. In this 3D plot it is observed, that the 2D rule of color gradients of interacting features was a basic consequence of the 2D projections from this 3D graph. There is no large deviation of feature contributions from the fitted gray color plot. Thus, it is evident that any structure of S_t,

$$S_t = (\Delta rs_t, \text{and } [\Delta cr_t, \Delta cr_{t-1}, \Delta er_t, \Delta er_{t-1}, \Delta rs_{t-1}, \Delta rs_{t-2}, ecm_{t-1}]) \quad (12)$$

related to Δcr_t and ecm_{t-1} is well explained by the joint nonlinear effect of both features Δcr_t and ecm_{t-1}. The goodness of this fit was 0.97. Therefore, this second order effect plot was an appropriate representation of Δcr_t and ecm_{t-1} contributing to the target Δrs_t.

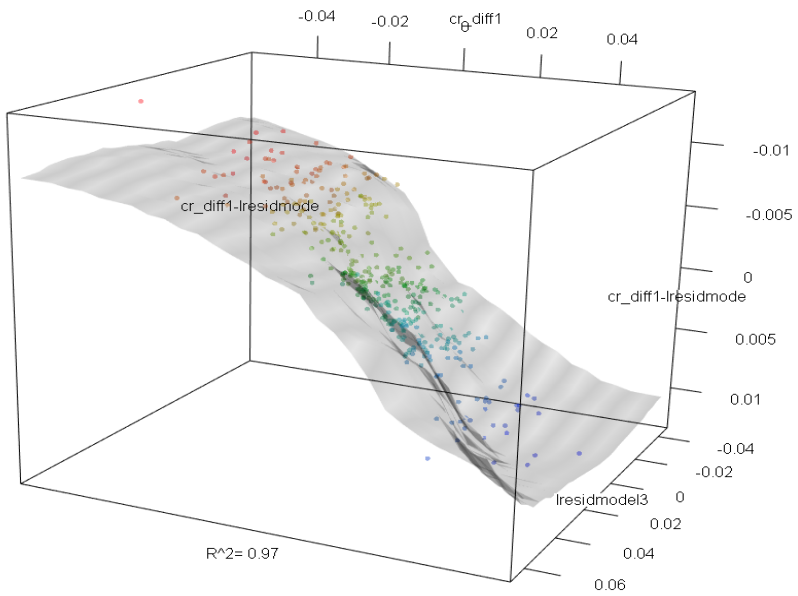
The depicted saddle-point structure of Fig. 7 was expected, as the product of Δcr_t and ecm_{t-1} contributed additively to the target Δrs_t.

Fig 6.



Note: cr_diff1 denotes Δcr_t, lresidmodel3 denotes ecm_{t-1}, lcr_diff1 denotes Δcr_{t-1}, er_diff1 denotes Δer_t, lrs_diff1 denotes Δrs_{t-1}, ler_diff1 denotes Δer_{t-1} l2rs_diff1 denotes Δrs_{t-2}. R^2 denotes R-squared.

Fig. 7. The 3D interaction plot of Δcr_t and ecm_{t-1} on Δrs_t.



Note: cr_diff1 denotes Δcr_t , lresidmodel3 denotes ecm_{t-1}, cr_diff1-lresidmode(Z-axis) denotes the interacted effect of Δcr_t and ecm_{t-1} on Δrs_t . R^2 denotes R-squared.

5.3 Random forest nonlinear error correction models

The starting point in the model specification of the new mixed random forest (Mixed RF) approach, is to use the best parametric model of Escribano & Torrado (2018) in Eq. (10), estimated by nonlinear least squares (NLS) in Table 1.

From Eq. (10) we subtract the estimated linear dynamic part of the model and look for the best RF specification for the rest using the two most important features; ecm_{t-1} and Δcr_t .

$$\Delta r\tilde{s} \equiv \Delta rs_t - \hat{a}_0 + \hat{a}_p(L)\Delta rs_{t-1} + \hat{a}_q(L)'\Delta x_t = R\hat{F}(ecm_{t-1}, \Delta cr_t) \tag{11}$$

Next, we subtract the estimated nonlinear RF terms in Eq. (11) from the dependent variable Δrs_t,

$$\Delta r\breve{s} \equiv \Delta rs_t - R\hat{F}(ecm_{t-1}, \Delta cr_t) \tag{12}$$

and use the new dependent variable from Eq. (12) to estimate the new parameters of the linear dynamic terms based on Eq. (13).

$$\Delta r\breve{s} = \Delta rs_t - \hat{a}_0 + \hat{a}_p(L)\Delta rs_{t-1} + \hat{a}_q(L)'\Delta x_t + \hat{\varepsilon}_t \tag{13}$$

The iterations between the estimated models in Eq. (12) and Eq. (13) are running until convergence. This new semi-parametric modeling approach is called *Mixed RF*.

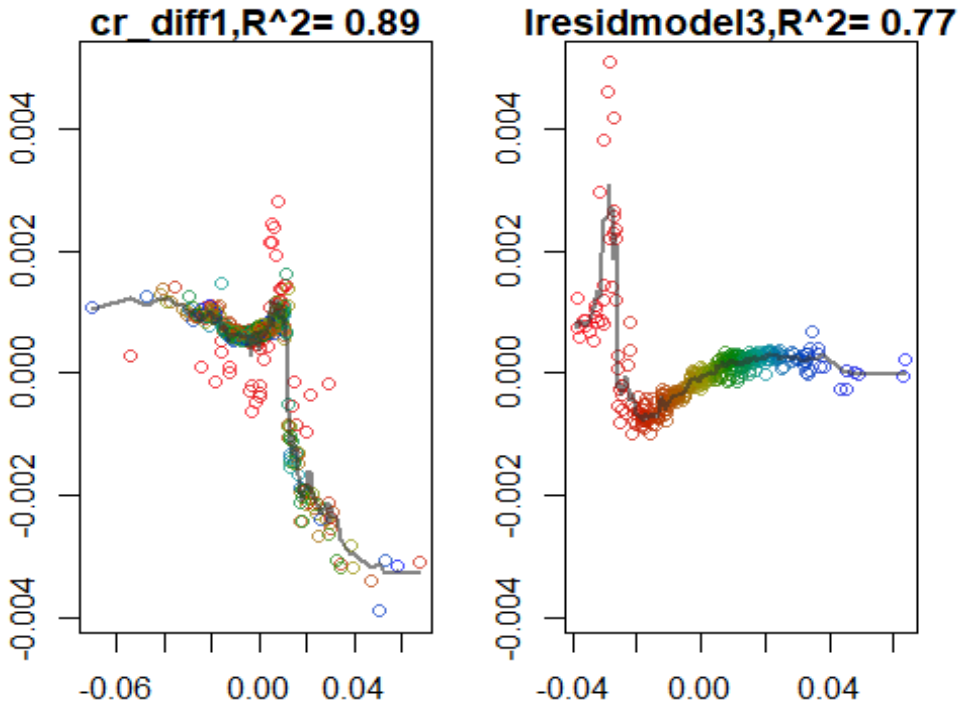
Based on the least root mean squared errors and highest R-squared, the best parameters for RF is: number of variables available for splitting at each tree node is 1, the minimum node size is 1, the maximum node size is 6, and the number of trees is 259. Certainly the computing cost in the Mixed RF is much smaller than in the RF. The importance of each feature is:

Table 3. Random forest variable importance in Mixed-RF

Variable	Overall
ecm_{t-1}	18.15
Δcr_t	5.48

The effects in the plot of Δcr_t and ecm_{t-1} are clearly of a non-linear pattern, representing the underlying additive non-linear reaction of Δrs_t to changes in Δcr_t and ecm_{t-1}. In Fig. 8, the y-axis has the cross-validation (CV) contributions of each x-feature, i.e., the change of the predicted probabilities for different values of each x-feature. The plotting illustrates the main effects, as feature contributions by each feature were plotted against their respective feature values. In each sub-graph in Fig. 8, the x-axis represents the feature values of each regressor in Eq. (11), and the black color line presents the partial functions. By employing the leave-one-out k-nearest neighbor gaussian kernel estimation, the goodness of fit was obtained (R-squared) for each regressor, which evaluates how well each feature contribution can be explained as a main effect. From Fig. 8, clearly Δcr_t is the main effect for explaining gasoline prices, which is with R-squared=0.89. The first sub-graph of Fig. 8 shows clearly that the equilibrium adjustment is more active when increasing the crude oil prices but the reaction is really small when there are negative changes on the crude oil prices. Fig. 8 shows that the range of positive ecm_{t-1} is greater than the range of negative ecm_{t-1} and more plots are located into the positive range. A color gradient along the most influential feature, Δcr_t, is applied to search for interactions. We can observe that there is a latent interaction with ecm_{t-1}.

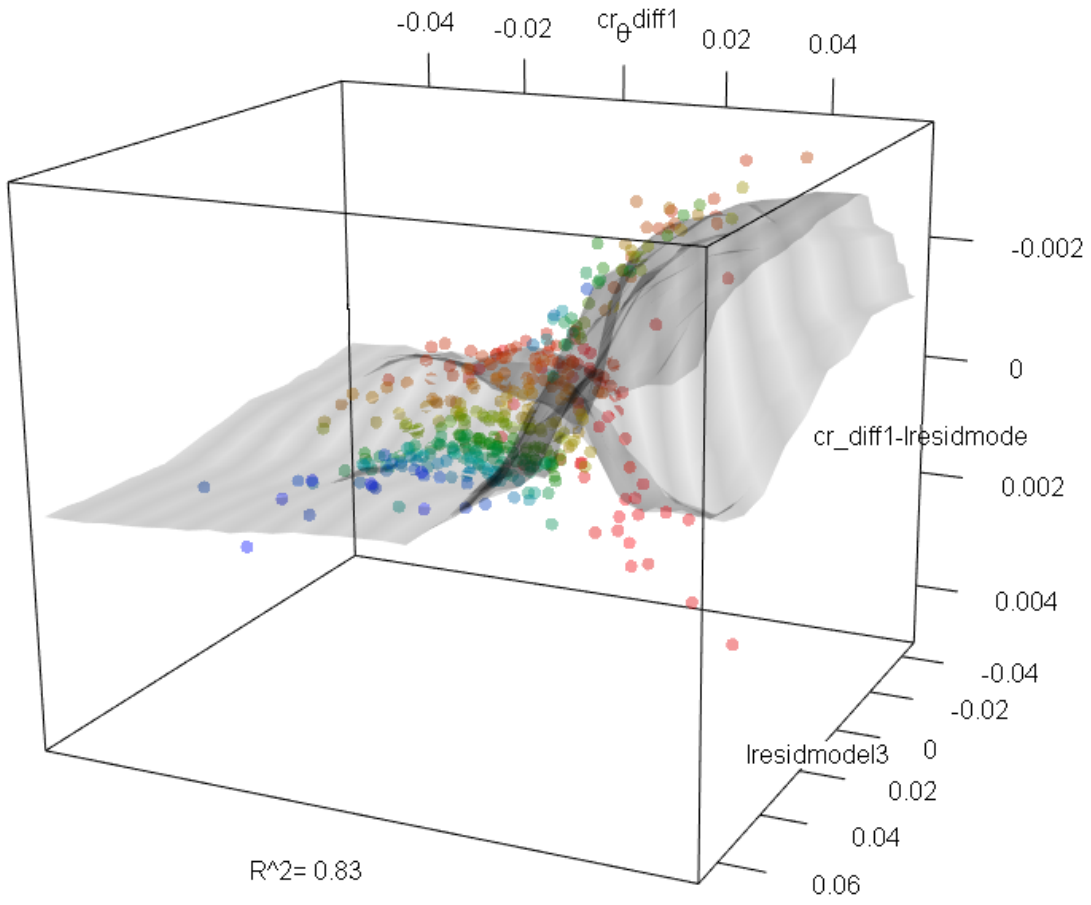
Fig. 8. Nonlinear reaction.



Note: cr_diff1 denotes Δcr_t, lresidmodel3 denotes ecm_{t-1}, R^2 denotes R-squared.

In Fig. 8 the color gradient suggests that Δcr_t interacted with ecm_{t-1} due to the vertical color gradient in the plot of ecm_{t-1}. In Fig. 9 their combined feature contributions are plotted in the context of both features, Δcr_t and ecm_{t-1}. The goodness of this fit is 0.83. Therefore, this second order effect plot is an appropriate representation of Δcr_t and ecm_{t-1} and contributes to the adjustment of Δrs_t. Fig. 9 also provides evidence of the empirical finding of Escribano & Torrado (2018); the main and largest equilibrium corrections occur when the international crude oil prices are increasing.

Fig. 9. Combined feature contributions.



Note: cr_diff1 denotes Δcr_t , lresidmodel3 denotes ecm_{t-1}, cr_diff1- lresidmode(Z-axis) denotes the interacted effect of Δcr_t and ecm_{t-1} on Δrs_t . R^2 denotes R-squared.

By subtracting the estimated RF(ecm_{t-1}, Δcr_t) from Δrs_t, see Eq. (12), and estimating by OLS the parameters of Eq. (13), we are able to check the fast convergence of the parameter estimates of the dynamic linear part of the model as well as the nonlinear

part. By comparing the final estimation results of NLS, Eq. (10) and Table 1, with the Mixed Random Forest of Table 3, we observe that they are very similar, although they are not identical. In the following section we compare the forecasting ability of each nonlinear model.

Table 3 Linear parameter estimates with Mixed RF

	Coefficients	Std. Errors	t-ratio	P-values
Δcr_t	0.330	0.019	17.16	< 2e-16 ***
Δcr_{t-1}	0.085	0.025	3.31	0.0010 **
Δer_t	0.383	0.067	5.67	2.71e-08 ***
Δer_{t-1}	0.210	0.071	2.98	0.0031 **
Δrs_{t-1}	0.133	0.049	2.69	0.0074 **
Δrs_{t-2}	0.084	0.037	2.28	0.0231
Significance levels: 0 '***', 0.001 '**', 0.01 '*'				

6. Forecasting comparison of three methods with training and test datasets

Table 4 presents the model performance based on both training and test datasets. Clearly, the mixed random forest (Mixed RF) model has the lowest root mean square error (RMSE) and mean absolute error (MAE) in both training and testing datasets while the random forest (RF) has the worse RMSE and MAE.

Table 4. Comparing the model performance in each method

	Training dataset			Testing dataset		
	Logistic-ECM	RF	Mixed-RF	Logistic-ECM	RF	Mixed-RF
RMSE	0.0055918	0.005971	0.0052444	0.0042073	0.0042271	0.0041819
MAE	0.0041978	0.004498	0.0039386	0.0030761	0.0030323	0.0030551

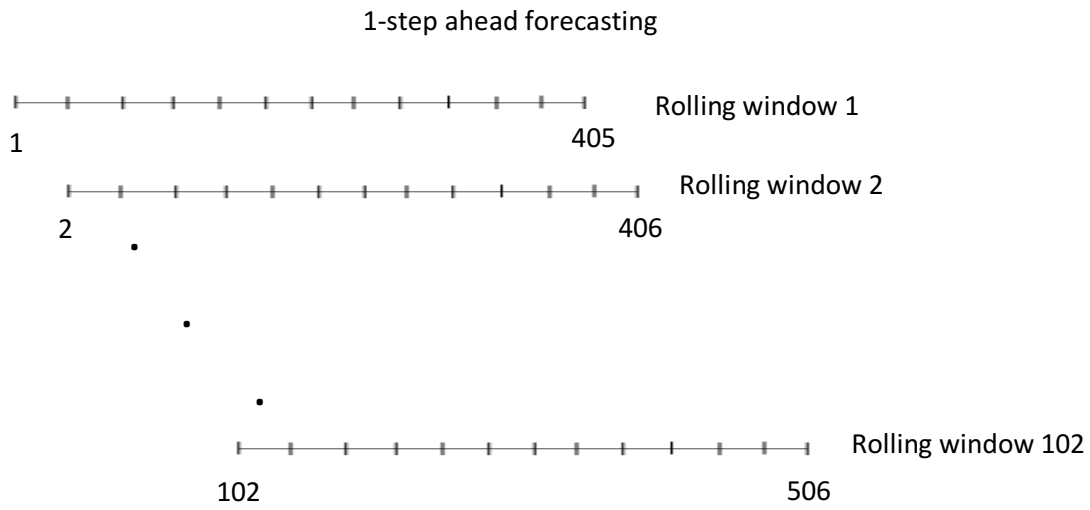
In all three models, the RMSE and MAE in the training dataset are bigger than in the testing dataset, hence the over-fitting might not be the issue here. Overall, the proposed Mixed RF over-performed compared to the other two models in both the training and testing datasets.

6.1 Forecasting fuel prices with competing between these three models

In the previous section, we showed that the proposed Mixed RF outperformed the other two models in terms of model fit. Our goal now is to evaluate the out-of-sample forecasting performance of these models based on the use of rolling windows in each model where the window size is 405 periods of weekly data (sub-sample). The forecasting period goes from November 3, 2017 to November 11, 2019 (102 periods). We estimate each model in each sub-sample (405 periods in our case), then we predict the h-step (1,2…12 in our case) ahead forecasts.

Example: 1-step ahead forecasting:

1. Estimate the first rolling window which is period 1 to period 405, then we forecast period 406 which is the first forecast period.
2. Estimate the second rolling window which is period 2 to period 406, then we forecast period 407 which is the second forecast period.
3. Keeping the procedure until the last rolling window period 102 to period 506, then we forecast the period 507 which is the last forecast period.



The graphs of the out-of-sample forecasting of 1-week, 4-week, and 12-week ahead are given in Figs. 10.1, 10.2, and 10.3, respectively. The 1-step ahead forecast (Fig.

10.1) in Mixed RF and RF catch the true value and much closer to it than the Logistic ECM model. In both 4-step and 12-step ahead forecast, the Mixed RF and RF capture the change of gasoline price more than the Logistic ECM.

Fig. 10.1

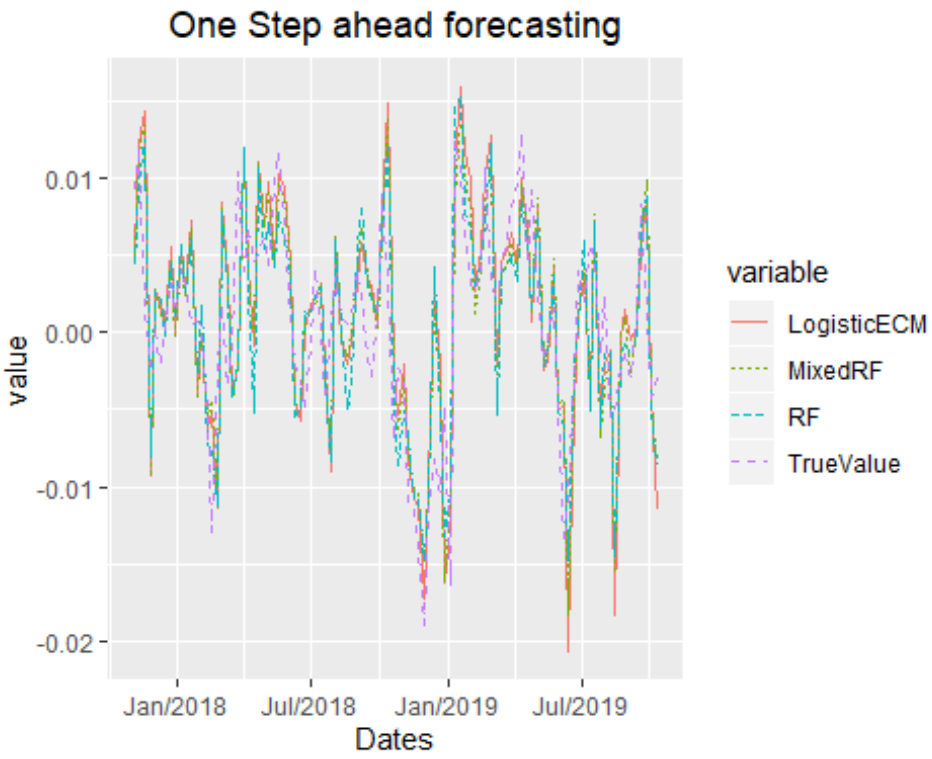


Figure 10.2

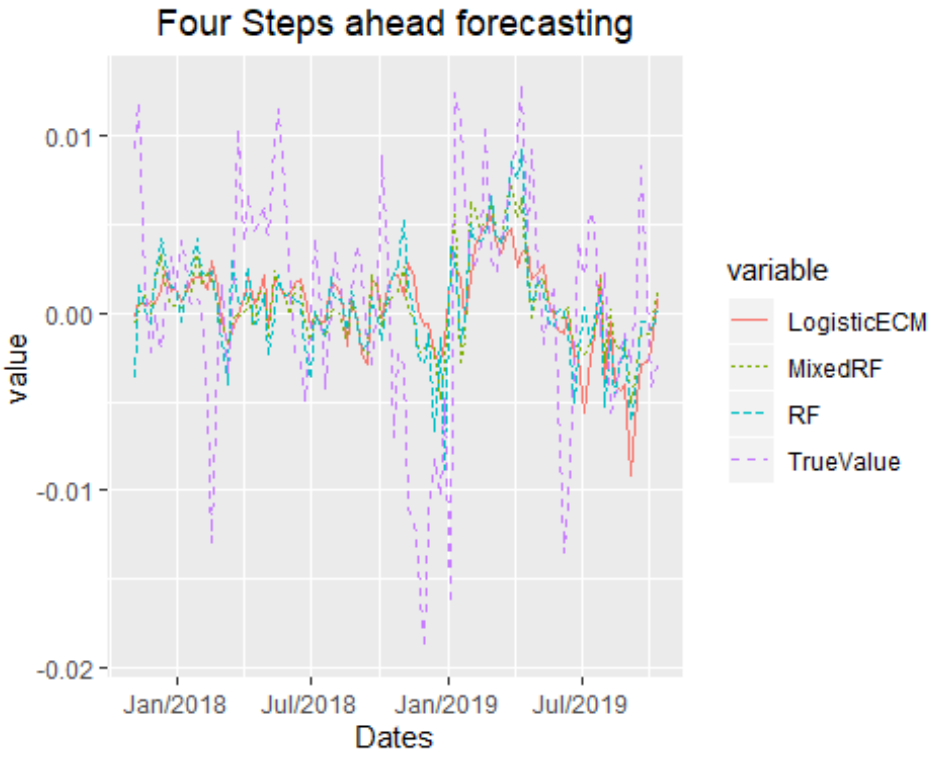
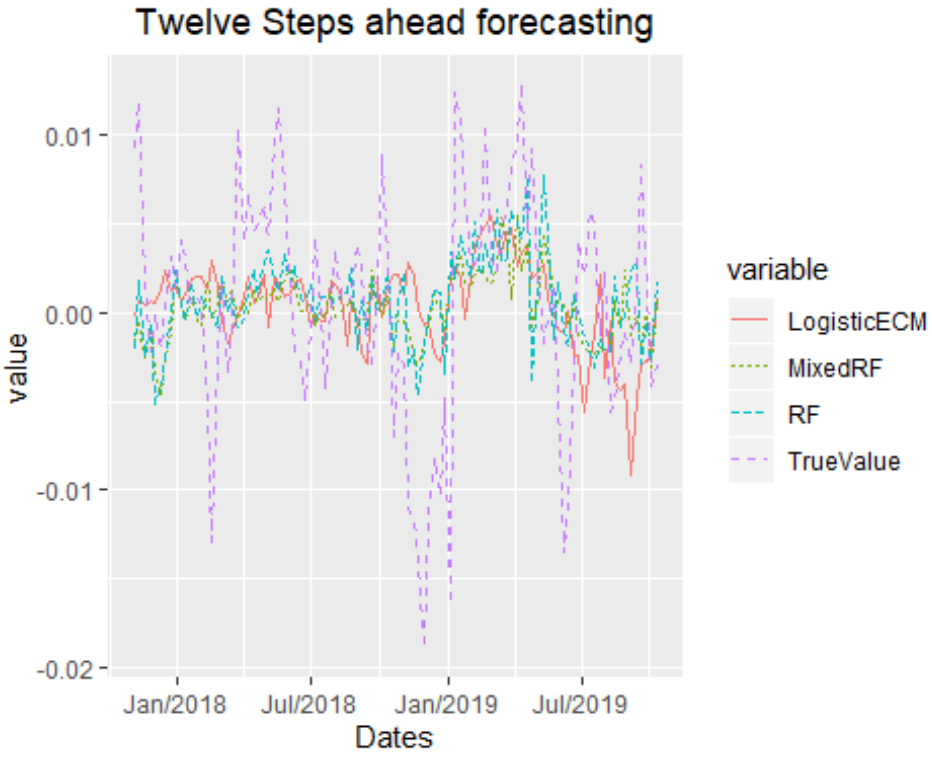

Four Steps ahead forecasting
0.01
0.00
-0.01
-0.02
value
Jan/2018
Jul/2018
Jan/2019
Jul/2019
Dates
variable
LogisticECM
MixedRF
RF
TrueValue

Figure 10.3


Twelve Steps ahead forecasting
0.01
0.00
-0.01
-0.02
value
Jan/2018
Jul/2018
Jan/2019
Jul/2019
Dates
variable
LogisticECM
MixedRF
RF
TrueValue

Then, we evaluate each model's performance using the Model Confidence Set (MCS) proposed by Hansen et al. (2011). The MCS procedure aims to determine the Superior Set Models (SSM). Any model from the SSM has statistically equal forecasting capabilities. MCS presents the test between every pair of models. The null hypothesis is the equal predictive ability (EPA), which means that statistically all the models have the same predicting ability. The alternative is that models have different predicting abilities. If the null hypothesis is accepted, then we have obtained the SSM, otherwise, the procedure eliminates the worst performing model and the procedure is repeated. At 95% confidence level, all the models are not rejected for all the forecast results of the steps. However, at 80% confidence level, the non-linear parametric model is rejected in the 1-step ahead forecast.

Hence, the Mixed RF and the RF present an equal predictive ability in our dataset from the MCS test. We then evaluate the RMSEs and MAEs of these models to compare the accuracy of these models. The graphs of RMSEs and the MAEs of the out-of-sample forecast based on the 1-step to 12-step ahead forecasts are given in Figs. 11.1 and 11.2. We observe that the best forecasting performance is given by the Mixed RF since it has the lowest RMSE and MAE for all of the two forecasting horizons considered; short-term and middle-term on the average. However, the parametric logistic ECM model now gives the worst out-of-sample forecasting performance out of the three competing nonlinear modeling specification procedures considered. The RF performs slightly better in the long-term forecasting, with lower RMSEs and MAEs from 8-step ahead to 11-step ahead.

Fig. 11.1

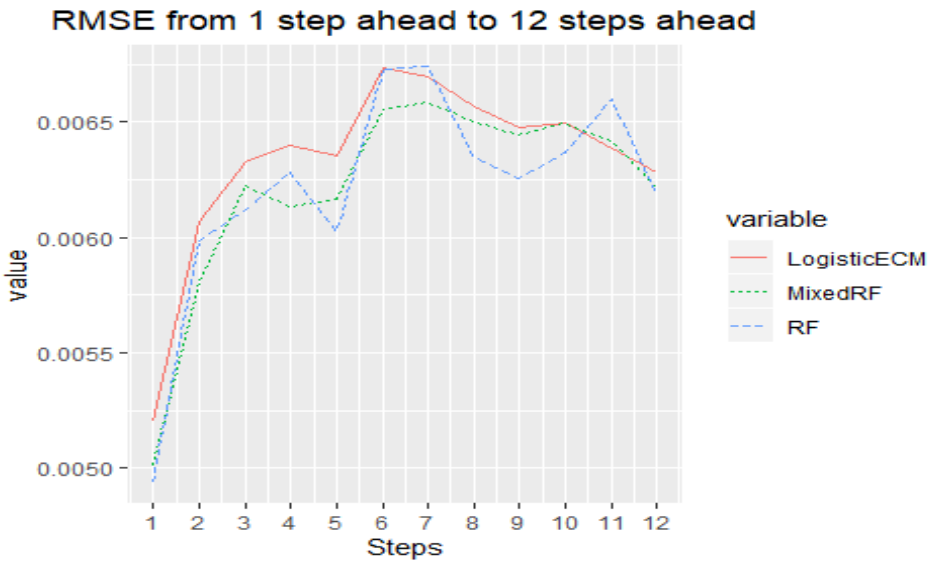
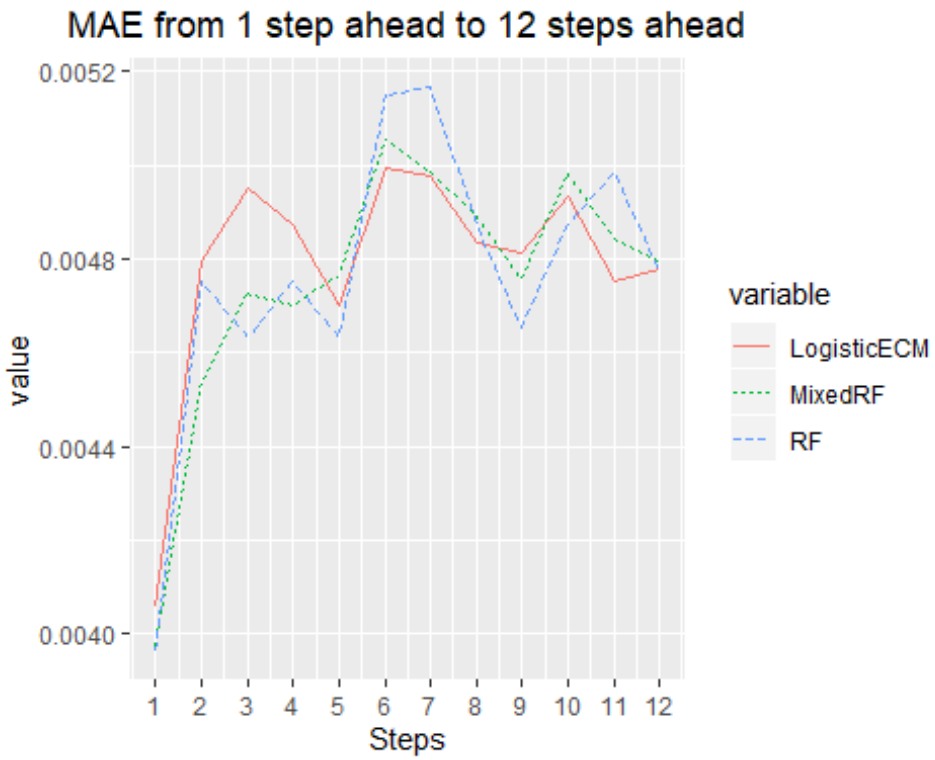


Figure 11.2



7. Conclusions

In this paper we described our testing of the consumer complaint that retail gasoline prices react faster to crude oil price increases than to crude oil price decreases. The relevance of fuel prices in the daily lives of people have attracted the attention of economists and competition authorities and subsequently the fuel market has been studied in detail for several countries. To date, no consensus has been reached for the Spanish sector since most of the empirical analyses previously undertaken, apart from those of Escribano & Torrado (2018) and Torrado & Escribano (2020), did not find strong empirical evidence of asymmetric price behavior in gasoline markets. To tackle the "rockets and feathers" phenomenon of oil price changes those authors introduced an interaction term with previous long-run equilibrium error (ecm_{t-1}), from the cointegrating relationship, as the transmission channel of changes in oil prices to changes in gasoline prices. These nonlinear models gain flexibility by introducing double threshold error correction terms; the error correction terms not only depend on whether the equilibrium error is positive or negative but also on whether the crude oil prices are rising or falling.

In this paper, for training the mixed RF modeling we took an updated weekly series of variables for the period of January 2010 to October 2017. For the out-of-sample forecast (test set) of gasoline prices we used the period of November 2017 to November 2019. The RF modeling in Table 2 indicates that the two main features of overall importance are the increases or decreases of oil prices (151.5) and the previous equilibrium errors (37.87). When applying the mixed RF modeling, the overall feature importance is reversed, with the previous equilibrium error having an importance of (18.15) and the increases or decreases of oil prices having a variable importance of (5.48). However, in both modeling approaches those are always the two main variables entering nonlinearly. Furthermore, from Fig. 6 to Fig. 9, both RF approaches identify that the Logistic approximation is the closest parametric functional form, as was previously established by Escribano & Torrado (2018) using a nonlinear error correction, the Logistic ECM.

The asymmetric results found in Spain in the oil market, show that sophisticated bivariate short-run nonlinearities are present in the gasoline market prices. Those gasoline price reactions depend on two main aspects; (a) whether the oil price increases or decreases and (b) whether the price levels of gasoline are above or below their long-run expected price levels. This long-run co-integrating equation relates the price of gasoline (€) with the international price of oil ($) and the corresponding exchange rate (€/$).

The mixed RF model corroborates the previous asymmetric results of Escribano & Torrado (2018). The disequilibrium in the long-run prices of gasoline is adjusted very slowly (slow equilibrium correction) when the international oil prices are decreasing but this equilibrium correction is very fast (fast equilibrium correction) when oil prices are increasing.

In terms of model fit, in both the training dataset and the test dataset, the best model is the new mixed RF. Similar results are also obtained when doing out-of-sample forecasts. Using a rolling window of 102 periods of weekly data, Figs. 10.1 to 10.3, the graph of the out-of-sample for 1-week, 4-week and 12-week ahead forecasting. In both 4-step and 12-step ahead forecasting, the Mixed RF and the RF capture the change of gasoline price better than the Logistic ECM. From the graphs of RMSEs and the MAEs of the out-of-sample forecast, based on the 1-step ahead to the 12-step ahead forecasts, Figs. 11.1 and 11.2, we conclude that the best forecasting performance is given by the Mixed RF, for short-term and medium-term on average. The parametric logistic ECM model now gives the worst out-of-sample forecasting performance.

References

Bachmeier, L. J., & Griffin, J. M., 2003, New evidence of asymmetric gasoline price responses, *Review of Economics and Statistics*, *85*(3), 772-776.

Bacon, R. W., 1991. Rockets and feathers: the asymmetric speed of adjustment of UK retail gasoline prices to cost changes. *Energy economics*, *13*(3), 211-218.

Balke, N. S., Brown, S. P., & Yucel, M. K., 1998. Crude oil and gasoline prices: an asymmetric relationship?. *Economic Review-Federal Reserve Bank of Dallas*, 2.

Balmaceda, F., & Soruco, P., 2008. Asymmetric dynamic pricing in a local gasoline retail market. *The Journal of Industrial Economics*, *56*(3), 629-653.

Biau, G., & Scornet, E., 2016. A random forest guided tour. TEST 25, 197–227.

Borenstein, S., Cameron, C. A., & Gilbert, R., 1997. Do gasoline prices respond asymmetrically to crude oil price changes?, *Quarterly Journal of Economics*, *112*, 305-339.

Breiman L., 2001. Random forests. *Mach Learn* 45:5–32

Cotín-Pilart, I., Correljé, A. F., & Navarro, M. B. P., 2008. (A) Simetrías de precios y evolución de márgenes comerciales en el mercado español del gasóleo de automoción. *Hacienda Pública Española*, (185), 9-37.

Deltas, G., 2008. Retail gasoline price dynamics and local market power. *The Journal of Industrial Economics*, 56(3), 613-628.

Dufrénot G., & V. Mignon., 2002. Recent Developments on Nonlinear Cointegration with Applications to Macroeconomics and Finance. *Kluwer Academic Publishers.*

Enders W., & Granger., C. W. J. 1998. “Unit-root test and asymmetric adjustment with an example using the term structure of interest rates”. *Journal of Business and Economic Statistics*, 16, 304-311.

Engle, R., & Granger., C. 1987. Cointegration and error correction: Representation, estimation and testing. *Econometrica* 55: 251-276.

Escribano A. 1985. Non-linear Error Correction: The cases of Money Demand in United Kingdom 1878-1970. *Mimeo, Department of Economics*, University of California San Diego

Escribano A. 1986. Identification and modelling of economic relationships in a growing economy. *Ph.D Thesis, Department of Economics*. *University of California San Diego*.

Escribano A. 2004. Nonlinear error correction: The case of the money demand in UK (1878-2000). *Macroeconomics Dynamics*, *8*, 76-116.

Escribano, A., & Granger., C. W. J. 1998. Investigating the Relationship Between Gold and Silver Prices. *Journal of Forecasting.* Vol. *17*, 81-107, 1998.

Escribano A., & Pfann., G. 1998. Nonlinear error correction, asymmetric adjustment and cointegration. *Economic Modelling 15*, 197-216.

Escribano A., & Jordá., O. 1999. Improved testing and specification of smooth transitions regression models, in P. Rothmand (ed.), Nonlinear Time Series Analysis of Economic and Financial Data. *Kluwer Academic Publishers*, 289-319.

Escribano A. & Jordá., O. 2001. Testing Nonlinearity: Decision Rules for Selecting between Logistic and Exponential STAR Models. *Spanish Economic Review*. 3, 193-209.

Galeotti, M., Lanza, A., & Manera, M., 2003. Rockets and feathers revisited: an international comparison on European gasoline markets. *Energy economics*, 25(2), 175-190.

Granger C. W. J, & Lee., T. H. 1989. Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. *Journal of Applied Econometrics* 4, 145-159.

Hansen P. R., Lunde, A., & Nason., J. M. 2011. The Model Confidence Set. *Econometrica*, vol. 79(2), pages 453-497.

Hendry D. F., & Ericsson., N. R. 1991. An econometric analysis of U.K. money demand in *Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna Schwartz. *American Economic Review* 81, 8-38.

Jiménez, J. L., & Perdiguero, J., 2005. Medición de la colusión a través del parámetro de conducta: el caso de los hidrocarburos en Canarias. Mimeo.

Johnson, R. N., 2002. Search costs, lags and prices at the pump. *Review of Industrial Organization*, 20(1), 33-50.

Learning Methods. *Journal of Business & Economic Statistics*, DOI: 10.1080/07350015.2019.1637745.

Marcelo C., Medeiros, & Gabriel, F. R. Vasconcelos, Álvaro Veiga & Eduardo Zilberman., 2019. Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine.

Newey W. K. 2009. Two-step series estimation of sample selection models. *Econometrics Journal*, Vol 12, 217-229.

Perdiguero, J. 2010. Dynamic pricing in the Spanish gasoline market: a tacit collusion equilibrium. *Energy Policy*, Vol. 38(4), pp. 1931-1937.

Robinson P. M. 1988. Root-N-Consistent Semiparametric Regression. *Econometrica*, Vol 56, 4, 931-954.

Schimert J., & Wineland A. 2010. "Coupling a Dynamic Linear Model with Random Forest Regression to Estimate Engine Wear", *Annual Conference of the Prognostics and Health Management Society, North America*.

Scornet, E., Biau, G., & Vert, J.-P., 2015. Consistency of Random Forests. *The Annals of Statistics*, 43, 1716–1741.

Singh A., 2018. "How to interpret a Random Forest Model (Machining Learning with Python)", https://www.analyticsvidhya.com/blog/2018/10/interpret-random-forest-model-machine-learning-programmers/.

Teräsvirta T. 1994. Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89, 208-18.

Teräsvirta T., & Eliasson, A. C. 2001. Nonlinear error correction and the UK demand for broad money, 1878-1993". *Journal of Applied Econometrics*, 16, 277-88.

Teräsvirta, T., Tjostheim, D., & Granger, C.W.J. 2010. Modelling Nonlinear Economic Time Series. *Oxford University Press*.

Torrado, M., & Escribano, A. 2020. European Gasoline Markets: Price Transmission Asymmetries in Mean and Variance. *Applied Economics*. https://doi.org/10.1080/00036846.2020.1739224.

Wagner, I., & Athey, S. 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association,* 113, 1228–1242.