

Regularized Error Correction Framework with Multivariate Simulation for Multi Product Crack Spread Forecasting

Matheus Nogueira¹, Matheus Alves¹, Marina Dietze², Davi Valladão¹, Alexandre Street²

Pontifícia Universidade Católica do Rio de Janeiro

¹Departamento de Engenharia Industrial da PUC-Rio,

²Departamento de Engenharia Elétrica da PUC-Rio

R. Marquês de São Vicente, 225, Gávea, Rio de Janeiro, RJ, 22452-900

matnogueira@gmail.com, matheus03alves@gmail.com,

marinadietze@lampsco.com, davimv@puc-rio.br, street.puc@gmail.com

ABSTRACT

This paper proposes a probabilistic forecasting framework for the crack spread of petroleum derivatives, defined as the log-price ratio between each derivative and Brent crude oil. Motivated by a risk analysis context, the framework is designed to enable joint multivariate simulation of multiple derivatives. Inspired by the Error Correction Model (ECM) structure, the methodology captures both short- and long-term dynamics while avoiding the complexity of modeling each spot price individually. Two variants are explored: a multivariate regularized regression and a simpler univariate formulation. Both approaches are evaluated under an expanding-window scheme, using scenario-based simulations via multivariate bootstrap. A case study involving ten petroleum derivatives and two benchmark models—random walk and mean reversion—assesses forecast accuracy for horizons of up to 24 months. Results suggest that the univariate model delivers competitive or superior performance in several settings, especially for short-term forecasts.

KEYWORDS. Crack-spread. Error Corretion Models. Multivariate Simulation.

EST&MP – Statistics and Probabilistic Models; **EN&OG** – OR in Energy, Oil and Gas; **AS&DS** – Data Science and Analytics

1. Introduction

The crack spread, defined as the difference between the price of refined petroleum products and the price of crude oil, represents the gross margin earned by refineries from processing crude into finished products. As such, it is a key profitability metric in the refining sector and holds critical importance for understanding the economic viability of operations across the oil value chain. Accurate forecasting of crack spreads is essential for oil companies, investors, and policy-makers seeking to manage exposure to margin volatility, price derivatives, and support strategic decision-making. Moreover, given a risk analysis context, it is crucial to implement a framework capable of generating multivariate simulation regarding all desired derivatives.

Crude oil and refined product prices are closely related, exhibiting both strong correlations and long-run cointegration. These relationships reflect the structural ties between upstream and downstream segments of the petroleum market and imply that joint modeling of these prices can improve forecasting accuracy and risk assessment. See Asche et al. [2003]; Ederington et al. [2021, 2019] for an extended discussion regarding those relationships. Nonetheless, each product may also respond to idiosyncratic shocks, seasonal factors, and distinct supply-demand dynamics, further motivating a multivariate modeling approach.

The literature on forecasting petroleum prices has explored various model classes. For instance, Población and Serna [2016] point out that most studies assume the crack spread is stationary due to its construction from cointegrated price series. However, their empirical results show that, in many cases, this assumption does not hold. In response, they propose a two-factor model that incorporates both a non-stationary long-term component and a short-term mean-reverting factor, offering a better fit for observed futures prices. Other works include Dempster et al. [2022], who use a two-factor cointegration model for spread option valuation, and Vides et al. [2021], who apply the Fractionally Cointegrated VAR framework to capture long-memory behavior in oil and product prices. These models underscore the importance of representing both long- and short-term dynamics, but differ in their treatment of stationarity and their handling of multivariate interactions.

A particularly influential approach in this context is the Error Correction Model (ECM) framework introduced by Engle and Granger [1987], which enables forecasting based on long-run cointegration relationships and short-run adjustment mechanisms. Given the strong empirical evidence that crude oil and refined product prices are cointegrated, ECMs have become a natural modeling choice in this domain. Several works have applied ECMs to oil and product prices, including Lanza et al. [2005] and Gjolberg and Johnsen [1999], who show that deviations from equilibrium carry predictive power for short-term changes in spot prices.

This paper proposes a probabilistic methodology for forecasting the long-run dynamics of crack spreads that is inspired by the ECM structure but estimated in a single stage. Rather than modeling the spot prices of crude oil and refined products separately and computing the spread *ex post*, we model the *multiplicative crack spread* directly—defined as the difference between the logarithms of the refined product price and the crude oil price, i.e., the log-price ratio. Although the additive crack spread is more commonly used in the literature, the use of the multiplicative spread allows for modeling relative price variations, which better reflect refinery margins and guarantee non-negativity of prices, being widely adopted in financial and economic time series modeling.

By modeling the spread directly, we avoid the challenges of modeling pairwise correlations between product prices, thus reducing model complexity. That being said, we estimate one model for each derivative crack spread via regularized regressions and, in line with the multivariate nature of the problem, we generate scenario-based simulations by resampling model residuals, preserving the dependence structure across derivatives under common market shocks.

We evaluate the out of sample predictive performance of the proposed methodology in

a case study involving the monthly crack spread of ten petroleum derivatives with an expanding window approach and 24 months ahead probabilistic forecast through a multivariate bootstrap algorithm. The remaining paper is organized as follows: Section 2 presents the ECM framework and the proposed model, Section 3 presents the case study implemented to evaluate the proposed methodology and Section 4 concludes the work.

2. Methodology

This work aims to compute probabilistic forecasts for the crack spread of petroleum derivatives by modeling the multiplicative crack-spread of each derivative. The approach is conditional on the Brent crude oil log-spot price and is inspired by the Error Correction Model (ECM) framework, which captures both long-term equilibrium and short-term adjustments. This structure allows us to incorporate long-term relationships between the products with each other and the Brent spot price while preserving the multivariate nature of the system.

It is well established that petroleum derivatives' prices are cointegrated, which justifies the use of an ECM framework to model and forecast those time series (see Asche et al. [2003], Lanza et al. [2005] and Ederington et al. [2021, 2019]). However, instead of a two-step methodology, being the first step modeling the log-spot prices and, secondly, computing the crack-spread, we have chosen to model the spread directly. There are two reasons to choose the direct methodology. First, following the argument from Población and Serna [2016] and Dempster et al. [2022], modeling the spread directly evades the difficulty of modeling the correlation between derivative prices. Secondly, although not necessarily, the spreads can be stationary, while the prices themselves are always non-stationary. See Población and Serna [2016] and Dempster et al. [2022] for different conclusions regarding the stationarity of the additive spread.

That being said, one could define two common spreads between the prices of a petroleum derivative and the crude oil price. First, there is the (additive) crack-spread, defined as the difference between the prices in level (not log transformed):

$$y_{t,i} = d_{t,i} - B_t \quad (1)$$

where $d_{t,i}$ is the spot price of a petroleum derivative (Gasoline, Nafta or QAV, for instance) and B_t is the Brent crude oil, which is commonly used as the crude oil benchmark price, as stated in Ederington et al. [2019].

Moreover, one could define a multiplicative crack spread, as the logarithm of the ratio between the derivative price and the crude price.

$$y_{t,i} = \ln(d_{t,i}) - \ln(B_t) = \ln\left(\frac{d_{t,i}}{B_t}\right) \quad (2)$$

Equation 2 is the definition used in this work and the variable we aim to model and forecast. Its use allows for modeling relative price variations, which are more representative of refinery margins and better reflect the percentage changes that are critical in the decision-making process of refining operations. Moreover, the logarithmic transformation of prices is widely adopted in financial and economic modeling due to its ability to stabilize variance, handle skewed distributions, and guarantee the non-negativity of prices, making it particularly suitable for time series forecasting in energy markets.

2.1. Error Correction Models

To model the short run dynamic of a derivative spot price, this project methodology takes inspiration from the Error Correction Model (ECM) framework developed by Engle and Granger [1987]. In this framework, one can represent a long term equilibrium between two cointegrated time

series, which means that while each time series may be non-stationary, their linear combination is $I(0)$, that is, stationary. The framework's next step is to define a short-term dynamic which is repeatedly corrected by the deviation of the long-run equilibrium, expressed explicitly in the short-term equation through the error correction term. The ECM will then estimate the speed at which the dependent variable adjusts to restore equilibrium following a shock.

First, one could define a long-run equilibrium between two time series y_t and x_t as

$$y_t = \alpha^E + \beta^E x_t. \quad (3)$$

where α^E and β^E are the long-run equilibrium parameters to be estimated. It is important to notice that this equilibrium is expressed in contemporary time, since y_t is explained by x_t . Those time series are commonly two price series but, in our case, they represent the multiplicative crack spread and the Brent crude price.

From this expression, it is possible to define the long-term equilibrium deviation as the error of the long-term equation 3.

$$\varepsilon_t^E = y_t - \alpha^E - \beta^E x_t \quad (4)$$

Once the long-term equilibrium is specified, one could argue that the observed time series y_t follows a different dynamic than the one from equilibrium. As stated in Asche et al. [2003], Equation 3 describes the situation when prices adjust immediately. However, often there will be a dynamic adjustment pattern accounted by introducing lags of the two prices. Equation 5 introduces those terms, as follows.

$$y_t = c + \delta_1 x_t + \delta_2 x_{t-1} + \mu y_{t-1} + \nu_t \quad (5)$$

where ν_t is a random error $\nu_t \sim \mathcal{N}(0, \sigma^2)$.

If y_t and x_t are $I(1)$, equation 5 may generate spurious coefficients and may lack economic interpretation. With that in mind, it would be useful to write this dynamic in first difference. To achieve that one could subtract y_{t-1} in both sides of the equation.

$$y_t - y_{t-1} = c + \delta_1 x_t + \delta_2 x_{t-1} - (1 - \mu)y_{t-1} + \nu_t \quad (6)$$

$$\Delta y_t = c + \delta_1 x_t + \delta_2 x_{t-1} - (1 - \mu)y_{t-1} + \nu_t \quad (7)$$

Equation 7 is expressed in first difference so the dependent variable is stationary. Now, it would be useful to transform x_t , which can be $I(1)$, into a first difference notation. In order to do so, we can add and subtract $\delta_1 x_{t-1}$ in the right hand side, as follows.

$$\Delta y_t = c + \delta_1 x_t - \delta_1 x_{t-1} + \delta_1 x_{t-1} + \delta_2 x_{t-1} - (1 - \mu)y_{t-1} + \nu_t \quad (8)$$

Organizing the terms in the right hand side, we find the short term dynamic which directly depends on the long-run equilibrium lagged error, multiplied by a parameter λ which captures the speed at which the dependent variable adjusts to restore equilibrium following a shock. Note that the term in parentheses, multiplied by λ is exactly the lagged equilibrium error from Equation 4.

$$\Delta y_t = c_\lambda + \delta_1 \Delta x_t - \lambda (y_{t-1} - \alpha - \beta x_{t-1}) + \nu_t \quad (9)$$

It is important to notice that, in equation 9, both series are expressed in first difference and, therefore, are stationary. In this equation, the term multiplied by λ is exactly the deviation from the long-run equilibrium from Equation 4, as α and β are the coefficients from this equilibrium in Equation 3. That being said, λ expresses the speed at which the short-run dynamics adapt to long-run deviations.

In Engle and Granger [1987], the authors describe in detail a two-step estimation for an ECM, in which the first step is to estimate the long-run equilibrium and, with the lagged error of this model, estimate the short-term dynamic. Our methodology deviates from this approach, since we propose a model specification capable of being estimated in a single step.

2.2. Proposed Model

Inspired by the ECM framework and taking advantage of the multivariate nature of the problem, we have proposed the following specification for each derivative multiplicative crack as a function of both the Brent log-spot prices and other derivatives lagged multiplicative cracks:

$$\mathbf{y}_t = \mathbf{a} + \mathbf{b}B_t + \mathbf{c}B_{t-1} + \boldsymbol{\phi}^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (10)$$

where $\mathcal{T}$ is the set of time indexes, $\mathbf{y}_t$ is the vector with the multiplicative crack of all derivatives in time t , B_t is the Brent log-spot price in time t , $\mathbf{y}_{t-1}$ is the multiplicative crack of all derivatives in time $t-1$, $\mathbf{S}_t$ is a seasonal dummy matrix, and $\boldsymbol{\varepsilon}_t$ is a random error vector term ($\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$).

In order to take advantage of the regularization during the estimation of the model and, also, follow the steps of the ECM framework, we have reparametrized equation 10 as:

$$\mathbf{y}_t = \mathbf{a} + \mathbf{b}B_t + \mathbf{c}B_{t-1} + (\mathbf{I} - \mathbf{D})^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (11)$$

where the parameter vector $\boldsymbol{\phi}$ is reparametrized as $(\mathbf{I} - \mathbf{D})$, $\mathbf{I}$ being the identity matrix and $\mathbf{D}$ being the new parameter matrix for the lagged values $\mathbf{y}_{t-1}$. The advantage of this formulation is the fact that the l_1 and l_2 penalization will attempt to shrink $\mathbf{D}$ to zero and, in consequence, $\boldsymbol{\phi}$ will go to one, indicating the existence of a unit root in the model. This is a useful approach since $\mathbf{y}_t$, the multiplicative crack spread, may not be stationary as previously stated.

Moreover, we can follow two more steps in order to obtain the ECM framework equation in which the model is specified with respect to the first difference of the time series.

$$\mathbf{y}_t = \mathbf{a} + \mathbf{b}B_t + \mathbf{c}B_{t-1} + \mathbf{I}^\top \mathbf{y}_{t-1} - \mathbf{D}^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (12)$$

$$\mathbf{y}_t - \mathbf{y}_{t-1} = \mathbf{a} + \mathbf{b}B_t + \mathbf{c}B_{t-1} - \mathbf{D}^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (13)$$

$$\Delta \mathbf{y}_t = \mathbf{a} + \mathbf{b}B_t + \mathbf{c}B_{t-1} - \mathbf{D}^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (14)$$

Equation 14 is the model final specification.

However, it could be useful to add one more step in order to better fit this model into the ECM framework, by explicitly writing the lagged long run error into the short term dynamic. Equation 15 shows this step.

$$\Delta \mathbf{y}_t = \mathbf{a} + \mathbf{b}B_t - \lambda(\mathbf{D}_\lambda^\top \mathbf{y}_{t-1} - \mathbf{c}_\lambda B_{t-1}) + \boldsymbol{\beta}^\top \mathbf{S}_t + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathcal{T} \quad (15)$$

where $\mathbf{D} = \lambda \mathbf{D}_\lambda$ and $\mathbf{c} = \lambda \mathbf{c}_\lambda$ and the expression $\mathbf{D}_\lambda^\top \mathbf{y}_{t-1} - \mathbf{c}_\lambda B_{t-1}$ is the lagged error from a long term model directly written into the short term model equation.

By writing this model in a single equation as 14, it is possible to estimate it directly as a penalized regression using the ElasticNet algorithm with AdaLasso penalization, from Zou and Hastie [2005]. Moreover, each derivative equation is independent from all others, thus allowing for independent estimations. The model can be written as the following mathematical programming problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}_i^M, \boldsymbol{\varepsilon}_{t,i}} \quad & \sum_{t \in \mathcal{T}} \varepsilon_{t,i}^2 + \lambda \left[\alpha \|\boldsymbol{\theta}_i\|_1 + (1 - \alpha) \|\boldsymbol{\theta}_i\|_2^2 \right] \\ \text{s.t.} \quad & \Delta y_{t,i} = a_i + b_i B_t + c_i B_{t-1} + \mathbf{D}_i^\top \mathbf{y}_{t-1} + \boldsymbol{\beta}_i^\top \mathbf{S}_t + \varepsilon_{t,i}, \quad \forall t \in \mathcal{T} \end{aligned} \quad (16)$$

where $\theta_i^M = [a_i \ b_i \ c_i \ D_i \ \beta_i]^\top$, $\lambda > 0$, $\alpha \in [0, 1]$, the subscript i indicates each derivative from the set of derivatives $\mathcal{D}$ and the index M indicates a multivariate model, which means that all derivative lags are regressors in the right hand side.

Even though a regularized estimation could be sufficient to select all of the derivatives that are useful to predict a single derivative spread, we have also tested an univariate model. This decision has two main arguments in its favour. First, even though it is well known that the derivatives prices are highly correlated, it is not safe to conclude in priori that the spreads are also highly correlated (Figure 1 in Section 3 will discuss this). Moreover, from a forecasting lens, a more parsimonious model can have a better generalization capacity for an out-of-sample simulation, which is this work's main objective.

That being said, Equation 17 exhibits the second proposed model, being an univariate version of equation 14. This means that, in the right hand side, only the given derivative will be used as a lagged regressor, instead of all available derivatives.

$$\Delta y_{t,i} = a_i + b_i B_t + c_i B_{t-1} + \phi_i y_{t-1,i} + \beta^\top \mathbf{S}_t + \varepsilon_{t,i} \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma_i) \quad (17)$$

Moreover, the univariate model can also be written in a similar mathematical programming manner:

$$\begin{aligned} \min_{\theta_i^U, \varepsilon_{t,i}} \quad & \sum_{t \in \mathcal{T}} \varepsilon_{t,i}^2 + \lambda \left[\alpha \|\theta_i\|_1 + (1 - \alpha) \|\theta_i\|_2^2 \right] \\ \text{s.t.} \quad & \Delta y_{t,i} = a_i + b_i B_t + c_i B_{t-1} + \phi_i y_{t-1,i} + \beta^\top \mathbf{S}_t + \varepsilon_{t,i}, \quad \forall t \in \mathcal{T} \end{aligned} \quad (18)$$

where $\theta_i^U = [a_i \ b_i \ c_i \ \phi_i \ \beta_i]^\top$, $\lambda > 0$, $\alpha \in [0, 1]$, the subscript i indicates each derivative from the set of derivatives $\mathcal{D}$ and U indicates the univariate model.

2.3. Multivariate Simulation

Once all derivatives models are fitted, we have implemented a multivariate simulation through a bootstrap of each model's residuals with common sampled indexes. This procedure is described in Algorithm 1. A multivariate scenario simulation is crucial in this context given the correlation between all derivatives.

Even if the spreads are not as correlated as the spot prices, they all take part in the same market, which means they are exposed to the external shocks affecting the market itself, such as crisis scenarios. By bootstrapping the same residuals indexes for all derivatives, we keep the multivariate nature of the problem.

Although the algorithm describes the simulation for the multivariate model, it is perfectly suitable for the univariate version as well. The only difference is the X_t definition, which, for the univariate model, should be $X_{t,i} \leftarrow [1, B_t^{(s)}, B_T, y_{T,i}]^\top$ if $t = 1$ or $X_{t,i} \leftarrow [1, B_t^{(s)}, B_{t-1}^{(s)}, y_{t-1,i}]^\top$ if $t \neq 1$.

2.4. Expanding window evaluation

To account for potential structural changes and evolving dynamics in the time series, we adopt an expanding window approach for model evaluation. This method allows us to simulate a realistic forecasting environment by re-estimating model parameters as new data becomes available, capturing time-varying relationships and avoiding look-ahead bias. By assessing model performance across multiple periods, the expanding window provides a more robust and adaptive framework for evaluating out-of-sample predictive accuracy.

Algorithm 1 Multivariate Bootstrap Simulation Algorithm

```

1: Input: Number of observed periods  $T$ , simulation horizon  $H$ , number of scenarios  $S$ 
2: Output: Simulated values for all derivatives  $\hat{y}$ 

3: Set random seed
4: for each scenario  $s = 1$  to  $N_S$  do
5:   for each time step  $t = T + 1$  to  $T + H$  do
6:     if  $t = 1$  then
7:        $X_t \leftarrow [1, B_t^{(s)}, B_T, \mathbf{y}_T]^\top$ 
8:     else
9:        $X_t \leftarrow [1, B_t^{(s)}, B_{t-1}^{(s)}, \hat{\mathbf{y}}_{t-1}^{(s)}]^\top$ 
10:    end if
11:    Randomly sample the bootstrap index  $t_\varepsilon \in \{1, \dots, T\}$ 
12:    for each derivative  $i \in \mathcal{D}$  do
13:       $\beta_i \leftarrow$  model coefficients for derivative  $i$ 
14:       $\varepsilon_i \leftarrow$  residuals of the model for derivative  $i$ 
15:       $\Delta \hat{y}_{t,i}^{(s)} \leftarrow X_{t_\varepsilon}^\top \beta_i + \varepsilon_{t_\varepsilon,i}$ 
16:      if  $t = 1$  then
17:         $\hat{y}_{t,i}^{(s)} \leftarrow y_{t,i} + \Delta \hat{y}_{t,i}^{(s)}$ 
18:      else
19:         $\hat{y}_{t,i}^{(s)} \leftarrow \hat{y}_{t-1,i}^{(s)} + \Delta \hat{y}_{t,i}^{(s)}$ 
20:      end if
21:    end for
22:  end for
23: end for

```

For each product and step-ahead, we have chosen 2 accuracy metrics: the Root Mean Squared Error (RMSE) to evaluate point forecasts, and the Conditional Rank Probability Score (CRPS) to evaluate the out-of-sample simulations. Both metrics's definitions are given below.

$$\text{RMSE}_i = \sqrt{\frac{1}{T} \sum_{t=1}^T (\hat{y}_{t,i} - y_{t,i})^2}; \quad \text{CRPS}_i(F, y_i) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}\{z \geq y_i\})^2 dz$$

3. Case Study: Long-run Probabilistic Forecasting the Crack-spread of 10 Derivatives

In order to evaluate the proposed methodology, we have developed a case study with 10 oil derivatives monthly time series. The products are: two prices for Gasoline (N and U), two prices for Diesel (N and U), Jet Fuel, two prices for Fuel Oil (10000 and 30000), Nafta, Butane and Propane. Furthermore, the Dated Brent oil price was used as the crude oil reference spot price.

All prices are nominal, which means that no inflation rate was discounted in their values and they are expressed in dollars per barrel. All series were gathered from S&P Global Platts, covering the period from October 2007 to December 2024, totaling 207 monthly observations.

3.1. Case study parameters

This case study was performed under an expanding window approach, with 30 windows. The first window has a length of 125 months, starting in 2007-10 and ending in 2018-02 while the last has 183 months, starting in 2007-10 and ending in 2022-12. There is a constant value of

2 months included in each consecutive window. The simulations' horizon is always of length 24 months with 1000 scenarios in each window.

In order to estimate the model with the ElasticNet algorithm, we have used the GLMNet.jl package, which is a wrapper of the original *glmnet* package written in R from Friedman et al. [2010]. The information criteria used to select the best λ from the optimization model 18 was the AIC and the parameter that defined the l_1, l_2 combination was set to $\alpha = 0.5$.

In order to access the quality of our proposed methodology, we have implemented two benchmarks commonly used for spot and crack spread modeling: a random walk model and a mean reversion model. We have chosen to implement two benchmarks because of the ongoing discussion about the stationarity of the crack-spread (see Población and Serna [2016] and Dempster et al. [2022]). With that in mind, the Random Walk is the most common benchmark and the appropriate model for a non-stationary series. On the other hand, a Mean Reversion model is more appropriate for a stationary series.

Moreover, in order to provide a fair comparison with the proposed methodology, both benchmarks are conditional on the Brent oil value, since they are fitted directly in the multiplicative crack spread. Their definitions are as follows:

$$\textbf{Random Walk: } y_{t,i} = a_i + y_{t-1,i} + \varepsilon_{t,i} \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma_i) \quad (19)$$

$$\textbf{Mean Reversion: } y_{t,i} = a_i + \phi_i y_{t-1,i} + \varepsilon_{t,i} \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma_i) \quad (20)$$

where $y_{t,i} = \ln \left(\frac{d_{t,i}}{B_t} \right)$ as in equation 2 with $d_{t,i}$ as a single derivative spot price and B_t as the Brent oil spot price.

It is interesting to notice that both benchmarks are particular cases of the univariate model from equation 17. They were fitted in the same manner our proposed models were estimated: independent models for each derivative multiplicative crack with a multivariate bootstrap out of sample simulation.

3.2. Empirical Results

In this section, we present the proposed case study empirical results. First, in order to empirically visualize the multivariate nature of the problem, Figure 1 exhibits the correlation coefficient between all derivatives and the Brent, both in log-spot prices first difference and in multiplicative crack. As expected, the prices themselves are highly correlated. The multiplicative crack, on the other hand, even though exhibits some degree of correlation between all derivatives, has a weaker correlation when compared with the prices themselves. This is important because it can justify the evaluation of a univariate model, such as proposed in Section 2.2.

Since the main goal is to evaluate the proposed models' out of sample forecasting accuracy for up to 24 months in comparison with two benchmarks, the case study results will be presented as tables with the point forecasting and simulation errors. They are all mean values regarding all 30 expanding windows. It is a crucial reminder that all models' simulations are conditional on the real Brent price, since no model was implemented for the crude oil price itself, and its value is necessary to compute the forecasted crack spread.

In all tables, we will refer to the Random Walk model as RW, the Mean Reversion as MR, the multivariate proposed model from Equation 14 as Multi and the univariate version from Equation 17 as Uni.

First, Tables 1 and 2 exhibit the mean RMSE and CRPS metrics from all windows for 3 different horizons of 1, 12 and 24 months ahead so that it is possible to evaluate short, medium and long-term forecasting accuracy.

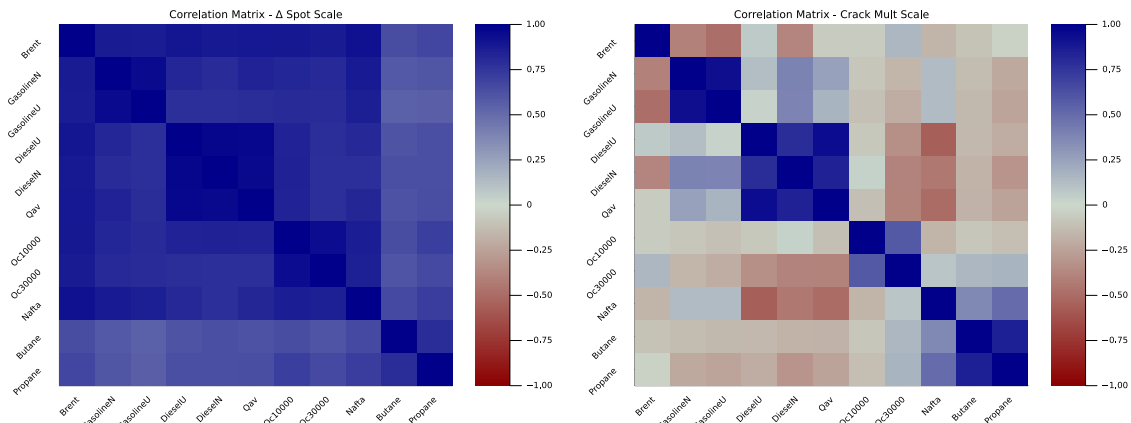


Figure 1: Correlation matrix for all derivatives and Brent in first difference of log-spot price and in multiplicative crack.

Horizon	RMSE											
	1				12				24			
Models	RW	MR	Multi	Uni	RW	MR	Multi	Uni	RW	MR	Multi	Uni
GasolineU	0.065	0.06	0.071	0.061	0.083	0.089	0.105	0.105	0.095	0.09	0.11	0.105
GasolineN	0.063	0.059	0.057	0.05	0.082	0.091	0.101	0.095	0.095	0.095	0.096	0.094
DieselU	0.075	0.067	0.061	0.057	0.171	0.128	0.117	0.118	0.204	0.129	0.129	0.124
DieselN	0.089	0.087	0.071	0.063	0.182	0.168	0.145	0.145	0.204	0.151	0.177	0.159
JetFuel	0.066	0.064	0.059	0.057	0.172	0.132	0.125	0.127	0.208	0.133	0.142	0.133
Fo10000	0.063	0.058	0.077	0.061	0.126	0.122	0.128	0.129	0.171	0.122	0.182	0.135
Fo30000	0.065	0.063	0.083	0.064	0.187	0.126	0.177	0.145	0.18	0.096	0.162	0.109
Nafta	0.052	0.043	0.043	0.041	0.138	0.111	0.1	0.103	0.172	0.107	0.101	0.099
Propane	0.111	0.106	0.115	0.108	0.422	0.339	0.385	0.331	0.489	0.295	0.416	0.261
Butane	0.099	0.096	0.099	0.087	0.364	0.302	0.299	0.293	0.426	0.257	0.291	0.236

Table 1: Mean RMSE from all windows and 1, 12 and 24 months ahead.

Horizon	CRPS											
	1				12				24			
Models	RW	MR	Multi	Uni	RW	MR	Multi	Uni	RW	MR	Multi	Uni
GasolineU	0.038	0.035	0.042	0.036	0.058	0.051	0.065	0.065	0.077	0.051	0.063	0.065
GasolineN	0.036	0.034	0.033	0.028	0.057	0.052	0.066	0.059	0.074	0.054	0.056	0.058
DieselU	0.039	0.035	0.036	0.033	0.097	0.078	0.068	0.07	0.113	0.081	0.084	0.077
DieselN	0.039	0.037	0.038	0.033	0.098	0.09	0.086	0.085	0.11	0.081	0.114	0.099
JetFuel	0.036	0.033	0.033	0.03	0.1	0.08	0.071	0.075	0.12	0.083	0.093	0.083
Fo10000	0.032	0.031	0.036	0.03	0.073	0.067	0.07	0.07	0.099	0.066	0.099	0.069
Fo30000	0.036	0.034	0.046	0.036	0.107	0.07	0.102	0.079	0.104	0.051	0.084	0.058
Nafta	0.03	0.025	0.024	0.023	0.078	0.067	0.057	0.062	0.102	0.062	0.058	0.057
Propane	0.061	0.058	0.061	0.057	0.259	0.218	0.247	0.211	0.284	0.173	0.227	0.146
Butane	0.057	0.056	0.058	0.052	0.217	0.189	0.184	0.184	0.242	0.148	0.164	0.132

Table 2: Mean CRPS from all windows and 1, 12 and 24 months ahead.

First, there is not a clear conclusion regarding the best model, since the results differ with the horizon. When evaluating the point forecasting with RMSE in Table 1, the univariate model exhibits better results in 6 of the 10 derivatives in the short run. However, in the medium horizon, all models perform better in a similar number of cases, while in the long run, the univariate model is tied with the MR benchmark. When evaluating the simulations with CRPS in Table 2 the results are

similar. In the short term, the univariate model outperforms the other 3 in 8 of the 10 derivatives. This is not the case for the longer horizons. For $h = 12$, the MR benchmark is the best model for 4 derivatives and, for $h = 24$, for 5 of the 10 chosen products. In general, evaluating just those 3 horizons, the univariate model seems to be the best performer. This result is interesting since the univariate model outperforms the multivariate one in almost all derivatives and horizons.

In order to access the forecasting accuracy on all windows, Tables 3 and 4 exhibit the mean RMSE and CRPS regarding all windows and all 24 months from the out of sample simulation. In a point forecast comparison, no model consistently outperforms the other, even though the univariate model has the best accuracy for 6 products. When evaluating the whole simulation, the univariate model is also dominant in 5 out of 10 cases, while the multivariate and MR show better results for 3 products each. One important result is that the RW model consistently exhibits the worst accuracy for both point and probabilistic forecast, while the stationary benchmark results are comparable with the proposed methodology, even outperforming it in some cases. This fact may be an argument for the stationarity of the multiplicative crack-spread, but, in order to confirm this hypothesis, more investigation should be conducted.

Models	RMSE			
	RW	MR	Multi	Uni
GasolineU	0.096	0.087	0.101	0.099
GasolineN	0.112	0.094	0.093	0.088
DieselU	0.155	0.115	0.106	0.107
DieselN	0.167	0.147	0.139	0.13
JetFuel	0.158	0.123	0.118	0.118
Fo10000	0.124	0.104	0.133	0.114
Fo30000	0.163	0.112	0.158	0.126
Nafta	0.134	0.095	0.087	0.087
Propane	0.377	0.287	0.349	0.278
Butane	0.331	0.26	0.262	0.242

Table 3: Mean RMSE values from all windows and full horizon.

Models	CRPS			
	RM	MR	Multi	Uni
GasolineU	0.064	0.049	0.061	0.06
GasolineN	0.07	0.054	0.058	0.053
DieselU	0.087	0.071	0.064	0.066
DieselN	0.087	0.079	0.085	0.078
JetFuel	0.089	0.073	0.069	0.071
Fo10000	0.071	0.058	0.073	0.061
Fo30000	0.093	0.061	0.088	0.069
Nafta	0.076	0.055	0.05	0.051
Propane	0.221	0.173	0.205	0.164
Butane	0.192	0.157	0.155	0.144

Table 4: Mean CRPS values from all windows and full horizon.

4. Conclusion

This work proposed a probabilistic modeling framework for forecasting the crack-spread of petroleum derivatives based on their multiplicative spread relative to Brent crude oil. By directly modeling the spread rather than the spot prices themselves, the approach simplifies the estimation of cointegrated relationships while preserving the multivariate dynamics between derivatives. The methodology relies on an error correction structure adapted to a penalized regression setting, allowing flexible yet regularized inference of long- and short-term dynamics.

Empirical results from an expanding-window simulation using 10 petroleum derivatives and two benchmark models suggest that the univariate ECM-inspired model achieves competitive or superior performance, particularly in short and medium-term forecasts. The multivariate model did not consistently outperform the univariate formulation, reinforcing the idea that spreads may be weakly correlated compared to prices.

A limitation of the current formulation is that forecasting remains conditional on the true future values of the Brent price, which restricts its applicability in fully out-of-sample or real-time forecasting environments. As a next step, integrating a model for Brent spot prices, thereby removing this conditionality, would enhance the methodology's realism and autonomy. This extension would allow practitioners to generate joint distributions for the Brent and derivative prices, leading to a truly end-to-end probabilistic forecast of crack spreads.

In addition, further research should investigate the stationarity properties of different spread formulations, as well as explore non-linear or regime-switching structures to account for structural breaks in oil markets.

Acknowledgments

The authors would like to thank Petrobras, specially the Market Risk and Stochastic Premises sector, for the financial and institutional support that made this work possible.

References

- Asche, F., Gjølberg, O., and Völker, T. (2003). Price relationships in the petroleum market: an analysis of crude oil and refined product prices. *Energy economics*, 25(3):289–301.
- Dempster, M., Medova, E., and Tang, K. (2022). Long-term spread option valuation and hedging. In *Commodities*, p. 125–146. Chapman and Hall/CRC.
- Ederington, L. H., Fernando, C. S., Hoelscher, S. A., Lee, T. K., and Linn, S. C. (2019). A review of the evidence on the relation between crude oil prices and petroleum product prices. *Journal of Commodity Markets*, 13:1–15.
- Ederington, L. H., Fernando, C. S., Lee, T. K., Linn, S. C., and Zhang, H. (2021). The relation between petroleum product prices and crude oil prices. *Energy Economics*, 94:105079.
- Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, p. 251–276.
- Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22.
- Gjølberg, O. and Johnsen, T. (1999). Risk management in the oil industry: can information on long-run equilibrium prices be utilized? *Energy Economics*, 21(6):517–527.

- Lanza, A., Manera, M., and Giovannini, M. (2005). Modeling and forecasting cointegrated relationships among heavy oil and product prices. *Energy Economics*, 27(6):831–848.
- Población, J. and Serna, G. (2016). Is the refining margin stationary? *International Review of Economics & Finance*, 44:169–186.
- Vides, J. C., Carmona, M., Fera, J., and Golpe, A. A. (2021). Modeling the united states crack spread: market efficiency, persistence and the verleger hypothesis. *Energy Sources, Part B: Economics, Planning, and Policy*, 16(10):951–970.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.