Integration of the Political Events in the Fossil Fuels Equity Market: a PCA and Forecasting Approach

Romain A. Alfred, Hamza Chergui

Integration of the Political Events in the Fossil Fuels Equity Market: a PCA and Forecasting Approach

Romain A. Alfred[1*] and Hamza Chergui[1]

[1*]SKAIZen Lab, SKAIZen Group, 8 Avenue Ledru Rollin, Paris, 75012, Île-de-France, France.

*Corresponding author(s). E-mail(s): ralfred@skaizengroup.fr;
Contributing authors: hchergui@skaizengroup;

Abstract

In this paper, we propose a methodology to test the integration of political events from the GDELT (Global database of events, language and tone) event database in the fossil fuels equity market prices. Our methodology is based on an approach borrowed from the field of financial time series forecasting. To represent the market to be predicted, we use the PCA technique (principal components analysis) to construct an index statistically representative of our market of interest, based on an equity portfolio. Our results show that political trends calculated on the basis of the political events and geopolitical analysis are features that improve forecasting, compared with delayed mathematical transformations of the time series alone. In the calculation of political trends, we also propose a partition of the international system into geopolitical spheres. As we explain in the article, our methodology represents a first step towards a better quantification of the political risks applied to investment.

Keywords: GDELT project, financial time series forecasting, machine learning, fossil fuels market, dimension reduction, geopolitical sphere, political risk.

1 Introduction

On 24^{th} February 2022, the Russian Federation invaded Ukraine. This political event was to have a enduring impact on the dynamics of the international system, but also on the financial markets. Among these financial markets, one was particularly

affected: the fossil fuels equity market. The commercial interconnections of the two belligerent states with the international economy and their importance in the supply of raw resources, in particularly oil and natural gas, led to an energy crisis of historic proportions[1]. At its peak (on 8^{th} March 2022), crude oil[1] was +34.31% higher than at the start of the war, while natural gas[2] was +108.67% higher (on August 22^{th}). At the meantime, the MSCI world energy index[3] outperformed the MSCI world index[4] by +32.96% on June 8^{th}. Quantifying the impact of political events is therefore essential for asset managers, but also for political decision-makers.

But before looking at the quantification of the mid-term impacts of a political event in atomic terms, we need to be able to mathematically modelling the relationship between overall political information and a given market. In his 2015 article on the "determinants of fossil fuel pricing", the Dutch economist Cees van Beers demonstrated the existence of a link between different relatively static political and institutional variables and the price of fossil fuels [2]. Nonetheless, major disruptions such as the outbreak of War in Ukraine are not perceptible using this kind of data. Keeping up with political news is therefore crucial to the responsiveness of decision-makers. However, we can still ask ourselves the following question: can the mass of local and international political events that are continuously taking place around the globe be used, despite the inherent noise, to estimate fluctuations in the fossil fuels equity market? This is the issue that we will attempt to address in this paper.

To achieve this, we will implement a methodology based on a forecasting approach to the issue. We will be investigating whether a forecasting model determined by the valuation of the fossil fuels equity market and by the available political information provides a better assessment of this market than a forecasting model that does not take political information into consideration. In other words, do fluctuations in the fossil fuels equity market incorporate political events? We will use machine learning to make this forecast. The literature shows that these models often achieve better forecasts than traditional methods of financial time series analysis [3, 4].

Regarding the data to be forecast, many stock market indices already exist for our market of interest. Nonetheless, most of them weight the assets in their portfolios according to market capitalisation, leading to a representativeness bias in favour of the largest capitalisations. But if the index that is supposed to represent our market of interest is biased, then the result of the efficiency test that we are trying to implement will also be biased. We will therefore construct a statistically representative index using principal components analysis (PCA) as an information reduction technique.

The first section of the paper describes one of our contributions to the analysis of international relations, namely a proposed division of the international system into different geopolitical spheres of influence. This work will be later useful to us in the processing of dynamic political information. In the second section, we will review the state of the art in the research relating to our issue. This will enable us to take an overview of the existing solutions and to propose an innovative methodology. Next,

[1] WTI crude oil futures expressed in Dollars.
[2] Natural gas futures expressed in Dollars.
[3] Stock market index representative of the large and mid-capitalisation fossil fuels equity market in financially developed countries, weighted according to market capitalisation, expressed in Dollars.
[4] Stock market index representative of the large and mid-capitalisation equities in developed countries, weighted according to market capitalisation, expressed in Dollars.

we will mathematically modelling our problem, and then processing the market data and dynamic political information used by our machine learning models. Then, we will present our experiments and results. Finally, we will conclude and present some perspectives for future works.

2 Definition of geopolitical spheres

As suggested by the French sociologist Raymond Aron, power is "the ability to impose one's will on others" [5]. This definition, like many others in the literature, remains subjective. Despite this fact, the notion of power lies at the heart of the international relations theory. If the notion of power is subject to interpretation, the notion of great power is even more so. According to the Prussian *Generalmajor* Carl von Clausewitz, a great power is "a power capable of ensuring its own security against any other power considered in isolation" [6]. This neo-realist view of the concept emphasises the survival of the political actor as the main issue. Based on Aron's definition of power and von Clausewitz's definition of great power, we can propose a definition of great power as a political actor capable of imposing its will on small and medium powers, but also of ensuring its own security in the face of these powers, taken one by one.

At present, as it was the case in previous international systems, the organisation of powers is not set in stone or clearly specified, as it would be the case for a state governed by the rule of law. In this context, the United Nations (UN) could act as an institutionalisation of the international system. Following this hypothesis, we could therefore consider that the domination of the world's states is ensured by the permanent members of the Security council. But the prominence of these five states (United States, China, Russia, France and United Kingdom) is not sufficient to explain all the international relations that are taking place around the world. Instead, the various multilateral cooperation groups and organisations will be seen as better able to represent the hierarchies of power. Within the latter, the G7 (Group of Seven) and the BRICS+ (Brazil, Russia, India, China and South Africa+) can be seen as two clubs of great powers. The different states in each of these groups having more or less the same voice. On the basis of this hypothesis, we will classify each country as belonging to the sphere of influence of one of these great powers or as non-aligned.

Within these great powers, despite the fact that each of them theoretically meets the definition of a great power proposed above, some are aligned with the foreign policy of another great power. This may be for security concerns in the face of another great power, as it is the case for Japan in the face of China. There can also be an alignment of great powers, based on relatively shared values and interests, as it is the case between Saudi Arabia and the United Arab Emirates. Or it could be a mix of the two, as it is the case for the great European powers relatively to the United States, in the face of the USSR yesterday and Russia today, since the War in Ukraine. The reasons may also be strategic, as it was the case for Japan during the period between the Japanese miracle and the Asian pivot in world geopolitics. During this interlude, Japan's strategy based on soft power and the economic element encouraged the country to remain within the geopolitical sphere of influence of the United States. We can thus divide the international system, which we consider to be multipolar, into

different geopolitical spheres, as shown in Figure 1. The small and medium powers of every geopolitical sphere are of course subject to the influence of several spheres and potentially of great non-aligned powers. Each state in a sphere therefore does not have the same degree of attachment to its sphere. For the sake of simplicity, however, we will assume that each state has a degree of belonging of 100% to its main geopolitical sphere. In the future, a more granular classification could be an avenue of improvement.

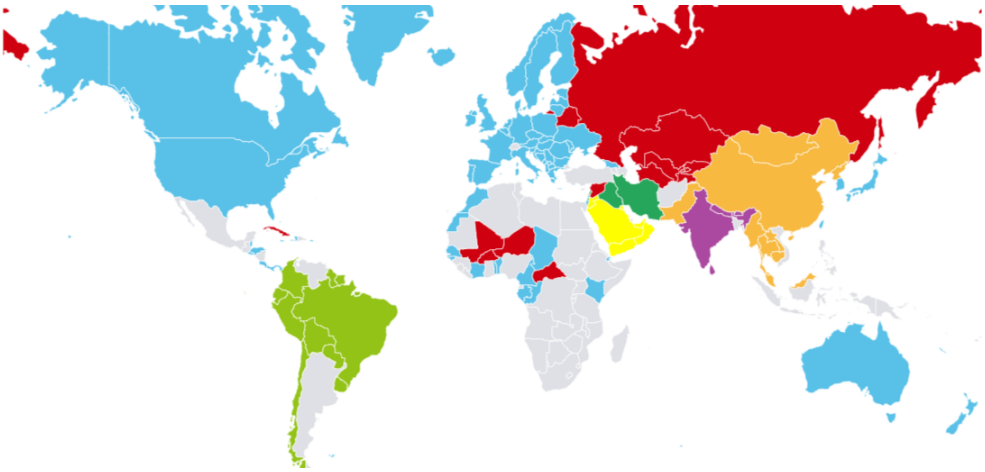


Fig. 1: Division of the international system into geopolitical spheres.
Legend: blue: western space, dark green: Shiite space, grey: non-aligned, light green: South American space, orange: Chinese space, purple: Indian espace, red: Russian space, yellow: Saudi space.

3 State of the art

In this section, we delve into the realm of forecasting financial time series using dynamic political information. Despite the growing interest in integrating such data into predictive models, existing approaches tend to rely heavily on sentiment analysis and press mentions, often overlooking broader geopolitical contexts. Our focus here is on exploring novel avenues in financial forecasting by incorporating a more comprehensive range of variables, extending beyond traditional sentiment analysis to encompass geopolitical and geographical factors.

3.1 Forecasting financial time series

A time series is a series of random variables dependant on time. In market finance, time series are the main tool for modelling financial assets. These financial time series have numerous properties, the main ones being the dependence of the random variables making them up on each other and their potential non-stationarity. A time series $X_{t_0}, \dots, X_T$ is defined as non-stationary if $\exists k \in [t_0, T]$ such that $X_{t_0}, \dots, X_k \sim \mathcal{L}_1$

and $X_{k+1}, \ldots, X_T \sim \mathcal{L}_2$, with $\mathcal{L}_1$ and $\mathcal{L}_2$ two different probability distributions. We then say that there is a market regime-switching at period k.

The main issue in time series analysis is therefore to mathematically modelling their development. Within this field, one area of study has become increasingly important as a result of technical innovations: the forecasting of time series.

The main traditional methods for time series representation are based on the autoregressive model. Introduced by the British statistician Udny Yule in 1927, this model rapidly incorporated new parameters, giving rise to the autoregressive moving average model (ARMA). In 1980, the Chinese statisticians Tong Howell and Lim Keng developed a new formulation, this time incorporating the hypothesis of non-stationarity via the threshold models. In 1988, the Norwegian statisticians John Tyssedal and Dag Tjøstheim introduced a model in which transitions are represented by Markov chains [7].

Nowadays, the forecasting of time series using machine learning models is an approach that is being explored to a growing extent by the scientific community. Most often, the models are trained using delayed data from the time series to be forecast. Among this data from the time series, we find mathematical transformations, but also technical analysis indicators. Examples include the moving average, exponential moving average, moving average convergence/divergence, rate of change, true range, relative strength index and the William's oscillator [8]. Many works also incorporate additional financial information from data providers such as Bloomberg and Reuters [9].

The most commonly used models include shallow learning models and neural networks [3]. Other methods such as evolutionary algorithms, genetic programming and agent-based models are also explored in the research literature [3].

In terms of forecast evaluation, some of the most common are the root mean square error, mean absolute error (MAE), mean absolute percentage error, mean square error and the Sharpe ratio [3].

3.2 Forecasting financial time series using dynamic political information

Within the literature, some papers are particularly interested in the forecasting of financial time series using one type of data: the political data. Using market data and structured political data from the press, as well as macroeconomic information in some cases, many authors are attempting to improve the effectiveness of forecasting models, or to demonstrate the impact of specific political variables on a given market. Nonetheless, despite the plethora of data offered by political datasets, such as the GDELT event database, the majority of approaches observed are limited to the use of variables relating to sentiment analysis and the number of press mentions of each event [10].

3.2.1 Forecasting using sentiment analysis:

In her article [10], the Saudi Arabian computer scientist Rawan Alamro seeks to forecast the prices of the Tadawul all share index, which is the stock market index

representing the Saudi Arabian equity market, using political data. After filtering the events in the GDELT database over the period from 2015 to 2019 and focusing only on events with a direct link to Saudi Arabia, Alamro performs various daily aggregations on variables relating to the average tone of events and to the attention brought to them by the media. Variables derived from delayed market data are also computed. Long short term memory networks are then trained and the results are more than satisfactory: a MAE of 0,59, which nevertheless remains to be compared with an approach excluding political data.

3.2.2 Forecasting by analysing political violence:

Another interesting approach in the literature uses variables other than tonality and press attention. In his thesis [11], the American political scientist James Yonamine seeks to assess the impact of political violence involving Israeli political actors on the fluctuations of the Tel Aviv 100 index from 1991 to 2012, which includes the country's 100 largest market capitalisations. From the GDELT dataset, he extracts only events involving Israeli, Palestinian and Lebanese actors that can be classified as political violence. Yonamine defines an event of political violence as either an act of terrorism or an event of armed conflict. In terms of political information, he restricts himself to daily aggregations derived from the number of adjusted political violence events and trends calculated from the latter variable, to which he adds information about the lagged volatility of the index's logarithmic returns. On the basis of these prediction data, he uses a traditional model to forecast the time series: the generalised autoregressive conditional heteroskedasticity. He concludes that political violence does not have a statistically representative impact on the Israeli equity market.

3.2.3 Forecasting by analysing inter-state armed conflicts:

In his thesis [12], the Dutch economist van Bruggen draws on an original approach based on a range of market, political and macroeconomic data. His objective is to measure the impact of political events on the global equity market, which he represents using the logarithmic fluctuations of the MSCI all country world index (ACWI) from 1997 to 2014. For market data, he calculates various financial indicators based on the studied index. As far as political data is concerned, van Bruggen uses the GDELT database to filter out only those events that could be considered as inter-state conflicts, which are daily quoted in more than 100 articles and involve one of the 44 countries represented in the MSCI ACWI index. He then focuses his analysis on the information related to the tone, the severity and the number of mentions of each event, calculating how long they have been occurring and whether they can be associated with a religious dimension. In terms of macroeconomic variables, his analysis includes logarithmic growth in GDP, inflation and in the employment rate. The forecasting model used here is the linear regression. He concludes that inter-state armed conflicts do have a negative impact on the equity market, and even more so if they have a religious component. Moreover, the wider the press coverage of an event is, the greater its impact on the market is.

Hence, despite the availability of a large number of dynamic political variables, the vast majority of studies are limited to using the tone and the number of mentions of events in the press. The studies including other descriptive variables of an event are still too scarcely explored, and in particular variables identifying geographical and geopolitical areas wider than that of the state. This is therefore what we will be looking to explore in this research project.

4 Mathematical problem modelling

In this paper, we seek to address the issue of the integration of dynamic political information in the fossil fuels equity market. To achieve this, we will design two machine learning models: one taking into account only market data and one also including dynamic political information. We will compare the forecast error of these two models in order to determine whether dynamic political information provides information for the forecasting of the fossil fuels equity market. Regarding the data to be forecast, we will construct a statistically representative index of our market of interest using the PCA algorithm on a portfolio of mid-capitalisation, large-capitalisation and mega-capitalisation stocks from our market. For the prediction data from the market, we will use lagged data from our index as well as lagged data from mathematical transformations of the latter. For the dynamic political information, we will use data from the GDELT (Global database of events, language and tone) event database. Based on these, we will compute some different political and geopolitical variables, which in turn will allow us to compute political trends useful to our machine learning model.

Let Y be a statistically representative index of the fossil fuels equity market, Z a set of variables derived from Y and X a set of variables describing political events. Then our question amounts to pose:

$$\hat{Y_t}^Z = f^Z(Y_{t-1}, Z_{t-1}) \tag{1}$$

and

$$\hat{Y_t}^{ZX} = f^{ZX}(Y_{t-1}, Z_{t-1}, X_t) \tag{2}$$

and to verify whether

$$Error(f^{ZX}(Y_{t-1}, Z_{t-1}, X_t)) < Error(f^Z(Y_{t-1}, Z_{t-1})) \tag{3}$$

with $\hat{Y_t}^Z$ an assessment of Y at time t taking into account the time series information only, $\hat{Y_t}^{ZX}$ an assessment taking into account the time series information and the political information and $\forall k \in \{Z, ZX\}$, f^k the forecast model minimising the assessment error.

4.1 Forecasting model

As previously explained, our problem is to find two forecast models f^Z and f^{ZX} which minimise the related forecast errors.

Among the several models to forecast financial time series, we will opt for a machine learning approach. Let M^k be the prediction matrix of $\hat{Y}^k$, $\forall k \in \{Z, ZX\}$. We will

divide our dataset into two parts: a learning sample $M^k_{learn} \in \mathcal{M}_{\frac{2}{3}n,p}$ and a test sample $M^k_{test} \in \mathcal{M}_{\frac{1}{3}n,p}$, such that $(n = T - t_0 + 1)$ and p the number of prediction variables. So, to precise the definition of the problem set out at the beginning of the section, this consists in parameterising f^k as a function of M^k_{learn} and verifying whether:

$$Error(f^{ZX}(M^{ZX}_{test})) < Error(f^{Z}(M^{Z}_{test})) \tag{4}$$

$\forall k$, we are looking for the function f^k which minimises $Error(f^k(M^k_{test}))$.

5 Market data

In this section, we focus on constructing a statistically representative index for the fossil fuels equity market, crucial for evaluating its efficiency regarding political information. Utilizing PCA, we aim to design an index that overcomes the biases of traditional global indices like MSCI World Energy or MSCI ACWI Energy. By narrowing our analysis to companies predominantly deriving sales from oil and natural gas industries, we exclude coal due to its distinct market dynamics. Through PCA, we seek to obtain a matrix representing the fluctuations of our representative index, enabling us to construct a robust measure of market performance. Additionally, we explore feature engineering of the financial time series, incorporating elements such as moving average, volatility, and moment of market regime-switching to enhance the accuracy of our forecasting models. This synthesis ensures a comprehensive evaluation of the fossil fuels equity market's efficiency in response to dynamic political information, facilitating a deeper understanding of the interplay between financial market dynamics and geopolitical factors.

5.1 Construction of a statistically representative index through PCA

In our study, we seek to verify the efficiency of the fossil fuels equity market regarding political information. To achieve this, we need to statistically represent the fossil fuels equity market. We represent this as a time series. As mentioned earlier in the 1 Introduction, there are already a multitude of stock market indices representing specific financial markets. Nonetheless, the market capitalisation weighting used to build the majority of these indices induces significant bias. We will therefore design an index that is more statistically representative of our market of interest than a traditional global index such as the MSCI world energy or the MSCI ACWI energy.

There are numerous dimension reduction methods in the literature that are applicable to our issue. The most widely used and oldest method is the PCA. The latter consists to reduce the number of dimensions of a dataset while trying to explain the initial variance as much as possible. There are several alternatives to the PCA, such as the linear discriminant analysis (LDA) and the locality preserving projections (LPP). More robust forms have also been developed by the scientific community. One example is the robust PCA (RPCA), which is effective in cases where the data are highly correlated or decorrelated from one another. The RPCA, but also the expectation RPCA (ERPCA) and the non-iterative proper orthogonal decomposition for singular value

decomposition (NIPOD-SVD), are also more efficient if there are many outliers. The generalized power sparse PCA (GPSPCA), on the other hand, offers a better dimension reduction when we faced with a very large number of variables. Other techniques exist, such as the project pursuit (PP) and the singular value decomposition (SVD), but they present problems in achieving a stable result for the former and in facing outliers for the latter. In this article, we will concentrate on the use of PCA to construct our statistically representative index. Nonetheless, it might be fruitful to explore the potential benefits of these other techniques [13].

Consider a portfolio φ of p fossil fuels equities. In our study, we will restrict our scope to companies that derive the majority of their sales from oil and natural gas industries, two primary sources of fossil fuels that account for the equivalent of 66,26% to 68,01% of global fossil fuels consumption between 2015 and 2023 [14]. We have therefore excluded coal from the analysis because of the distinct dynamics of this market. Indeed, because of environmental considerations, coal is a primary source of energy whose production, export and consumption follow downward trends once a country reaches a certain stage of economic development.

Let P_t be the closing prices and D_t the dividends on the portfolio's assets at time t. We define $P_t^{'}$ as the price adjusted for dividends such that $P_t^{'} = (P_t + \sum_{i=t_0}^{t} D_i)$. We then obtain $r_t = (\frac{P_t^{'}}{P_{t-1}^{'}} - 1)$ the periodic returns on the portfolio's assets.

PCA is a statistical method which uses a reduced centred matrix $x \in \mathcal{M}_{np}$ to obtain a matrix $x^{'} \in \mathcal{M}_{np^{'}}$ such that $n \geq p \geq p^{'}$. Each factor $x_{ij}^{'}$, $\forall i$, is associated with an eigenvector K_j, an eigenvalue λ_j, an explained variance $V_j = \lambda_j / (\sum_{m=1}^{p^{'}} \lambda_m)$ and C_{kj}^2 the quality of representation of the variable k at the factor j such that $C^2 = (K\sqrt{\lambda})^2$.

To construct our index, we will therefore center, reduce and augment r and obtain Δ the fluctuations of our representative index Y using the linear combination $\Delta_t = \sum_{j=1}^{p} \omega_j r_{tj}$ with $\forall k$, $\omega_j = \underset{V}{argmax}\ C_{kj}^2$, the percentage of the fluctuation explained by each stock. This gives us our statistically representative index $Y_t = \prod_{i=t_0}^{t} (1 + \Delta_i)$.

5.2 Feature engineering of the financial time series

Once we have obtained our target variable Y, we will calculate various variables which we will use as prediction data. First of all, we look at the forecast variables derived from the financial time series to be forecast. For each observation Y_t, we associate a set of forecast variables $Z_{t-1,j}$, $\forall j \in [1, 10]$.

We have:

$\mu_t = \frac{1}{t-t_0} \sum_{i=t_0}^{t} \Delta_i$ the moving average (over the period $[t_0, t]$),

$\sigma_t = \sqrt{\frac{1}{t-t_0-1} \sum_{i=t_0}^{t} (\Delta_i - \mu_t)^2}$ the moving volatility,

$s_t = \frac{1}{(t-t_0-1)\sigma_t^3} \sum_{i=t_0}^{t} (\Delta_i - \mu_t)^3$ the estimation of the moving skewness,

$\kappa_t = \frac{1}{(t-t_0-1)\sigma_t^4} \sum_{i=t_0}^{t} (\Delta_i - \mu_t)^4$ the estimation of the moving kurtosis,

$\hat{k}_t$ the estimation of the moving moment of market regime-switching,

$\mu_t^{(k)} = \frac{1}{t-\hat{k}_t} \sum_{i=\hat{k}_t}^{t} \Delta_i$ the moving average over the period $[\hat{k}_t, t]$,

$\sigma_t^{(k)} = \sqrt{\frac{1}{t-\hat{k}_t-1}\sum_{i=\hat{k}_t}^{t}(\Delta_i - \mu_t)^2}$ the moving volatility over the period $[\hat{k}_t, t]$,
$s_t^{(k)} = \frac{1}{(t-\hat{k}_t-1)\sigma_t^3}\sum_{i=\hat{k}_t}^{t}(\Delta_i - \mu_t)^3$ the estimation of the moving skewness over the period $[\hat{k}_t, t]$,
$\kappa_t^{(k)} = \frac{1}{(t-\hat{k}_t-1)\sigma_t^4}\sum_{i=\hat{k}_t}^{t}(\Delta_i - \mu_t)^4$ the estimation of the moving kurtosis over the period $[\hat{k}_t, t]$.

We assume that our index Y is non-stationary, i.e. that it follows several market regimes. In our study, we will assume the existence of only two market regimes over the analysis period. The same study could be carried out with a higher number of market regimes. To determine the moving moment of market regime-switching, we assume that the fluctuation Δ of our index follows a normal distribution. We then estimate the moment of market regime-switching with

$$\hat{k}_t = \underset{k_t}{argmax}\ (-\frac{k_t}{2} ln\ \hat{\sigma}_{t,1}^2 - \frac{t-k_t}{2} ln\ \hat{\sigma}_{t,2}^2) \tag{5}$$

with
$\hat{\sigma}_{t,1}$ the estimation of the standard deviation of Δ over the period $[t_0, k_t]$,
$\hat{\sigma}_{t,2}$ the estimation of the standard deviation of Δ over the period $[k_t + 1, t]$.

6 Dynamic political information

Henceforth, we will be focusing on dynamic political data. As stated in Subsection 4.1, our aim is to design a machine learning model that takes political information into account. To achieve this, we will use the information provided by the GDELT event database. In this section, we will therefore describe the different selected variables for our analysis. Then, we will present the different political as well as geopolitical variables calculated on the basis of these initial variables. As we discussed in Section 3, the regrouping of information in larger geographical and geopolitical ensembles are still under-explored. Finally, we present the different political trends extracted, which we will use to train our forecasting model.

6.1 GDELT event database

To represent the available dynamic political information, we need a sufficiently representative database. The GDELT event database allows us to address this issue. From press articles from around the world, political events have been extracted using natural language processing. This database contains a range of information about each event, including the actors involved, its location, its tone, the number of associated references and its nature [15]. Regarding the actors, we find a set of variables specifying at different levels of granularity their location, their typology, their ethnicity and their religion.

For each day, we have available a number ($k \geq 0$) of political events. For an analysis period extending from t_0 to T, we therefore have ($\sum_{t=t_0}^{T} k_t \geq 0$) political events associated with ($T - t_0 + 1$) prices to be forecast. We therefore proceed to aggregate

the political information in order to match the $(T - t_0 + 1)$ political observations in the GDELT event database with the $(T - t_0 + 1)$ prices.

To perform these aggregations, we use the following GDELT variables:

Time variable:

- *Day*: the date of the political event.

Variables related to typology:

- *GoldteinScale*: the classification of the political event according to the Goldstein scale, defined over a range $[-10, 10]$,
- *EventRootCode*: the level 1 of granularity of the type of political event according to the CAMEO (Conflict and mediation event observations) classification, example: 20 = unconventional mass violence,
- *EventBaseCode*: the level 2 of granularity of the type of political event according to the CAMEO classification, example: 204 = use of weapons of mass destruction,
- *EventCode*: the level 3 of granularity of the type of political event according to the CAMEO classification, example: 2042 = detonate of nuclear weapons.

Location variable:

- *ActionGeo_CountryCode*: the FIPS10-4 code of the country in which the political event took place.

Variables relating to the actors:

- *Actor1Type1Code* and *Actor2Type1Code*: the level 1 of granularity of the types of actors of the political event according to the CAMEO classification,
- *Actor1Geo_CountryCode* and *Actor2Geo_CountryCode*: the FIPS10-4 code of the countries of origin of the actors of the political event.

Variables relating to information diffusion modalities:

- *AvgTone*: the average tone of the documents mentioning the political event, defined over a range $[-100, 100]$,
- *NumSources*: the number of information sources mentioning the political event,
- *NumMentions*: the number of references to the political event in the source documents.

6.2 Computation of political and geopolitical variables

Among all the political events taking place around the world, some have a greater impact on the price of crude oil and natural gas, and therefore, indirectly, on the stock prices of these two sectors. We therefore hypothesise that events corresponding

to economic sanctions, armed conflicts, institutional changes, internal troubles and commercial agreements are more likely to have an impact on our market of interest. Especially if they involve OPEC+ (Organisation of Petroleum Exporting Countries+) member states with regard to the oil sector. We therefore compute from the variables *EventRootCode*, *EventBaseCode* and *EventCode* the variables *EconomicalSanctions*, *ArmedConflict*, *InstitutionalChange*, *InternalTroubles* and *CommercialAgreement*.

The GDELT database provides information about the geographical location of political events, as well as the countries of origin of the actors involved. For our study, it might be interesting to extract trends from coherent geographical areas that are larger than a single country. We are therefore going to operate two types of regrouping. For the location of the event, we will extract trends within different regions of the world. To do this, we will use the classification by sub-regions of the UN geographical scheme[5].

If we now adopt the point of view of the actors, it is the relationship between different geopolitical ensembles that we are interested in, and not just the geographical factor. Among the geopolitical ensembles with a potentially greater impact on the fossil fuels market are the OPEC+ countries, the political actors from countries with the majority of the world's natural gas reserves, and whether or not a country is a great power. Finally, we will also be grouping them by geopolitical sphere, as defined in Section 2. To capture potential disruptions, it might be interesting to observe whether the action takes place between two actors from different geopolitical spheres. From *Actor1Geo_CountryCode* and *Actor2Geo_CountryCode*, we define the variables *ActionGeo_UNGeoSubregion*, *Actor1Geo_OPEP+*, *Actor2Geo_OPEP+*, *Actor1Geo_MainNaturalGasReserve*, *Actor2Geo_MainNaturalGasReserve*[6] [16], *Actor1Geo_GreatPower*, *Actor2Geo_GreatPower*, *Actor1Geo_GeopoliticalSphere* and *Actor2Geo_GeopoliticalSphere* as follows.

6.3 Computation of political trends

Using the GDELT variables mentioned above and the computed variables, we will perform various aggregations. The difficulty lies in the loss of information that aggregation can represent. But it is also an opportunity because the process allows trends to be identified and observations to be hierarchised.

We introduce a factor ω_{tj} which will allow us to weight political events according to their importance in the press opinion (variable *NumSources*) and their veracity (*NumMentions*). We then have:

$\forall t, i \in [1, n_t], j$, with n_t the number of political events at the period t,

$$\omega_{tij} = \omega'_{tij} / (\sum_{i=1}^{n_t} \omega'_{tij}) \tag{6}$$

and

[5] Methodology: standard country or area codes for statistical use (M49). United Nations Statistics Division. https://unstats.un.org/unsd/methodology/m49/

[6] Only countries whose natural gas reserves represent more than 5% of the world's reserves are included.

$$\omega'_{tij} = X_{t,i,NumMentions} \frac{X_{t,i,NumSources}}{max(X_{t,i,NumSources})}. \quad (7)$$

We then carry out various simple aggregations, as well as combinatorial aggregations, including two or three variables (not counting the weighting by the ω factor, which amounts to combining two additional variables). By period, we extract the observable information in appendix A.

7 Experiments

To address our issue, we extract stock prices and dividends from the fossil fuels market. As explained in Section 1, our focus in the fossil fuels market will be on the oil and gas sectors. To better represent our market of interest, we will restrict ourselves to the mega-capitalisation, large-capitalisation and mid-capitalisation stocks, i.e., those with a market capitalisation higher than $2bn. This removes a large number of stocks that are too illiquid and therefore susceptible to adding noise to the market index that we are attempting to build.

For our analysis, we will focus on the period over which the political events in the GDELT version 2.0 dataset are available, i.e., from February 2015[7]. We will stop at the end of December 2023. This corresponds to 4,2 million political events, i.e., 1 292 events per week.

We obtained 943 stocks via the Yahoo Finance Equity screener[8], then 936 stocks using Yahoo Finance's Python library over our period of interest, which we reduce by PCA into a statistically representative index (see Figure 2), using the method presented in Subsection 5.1.

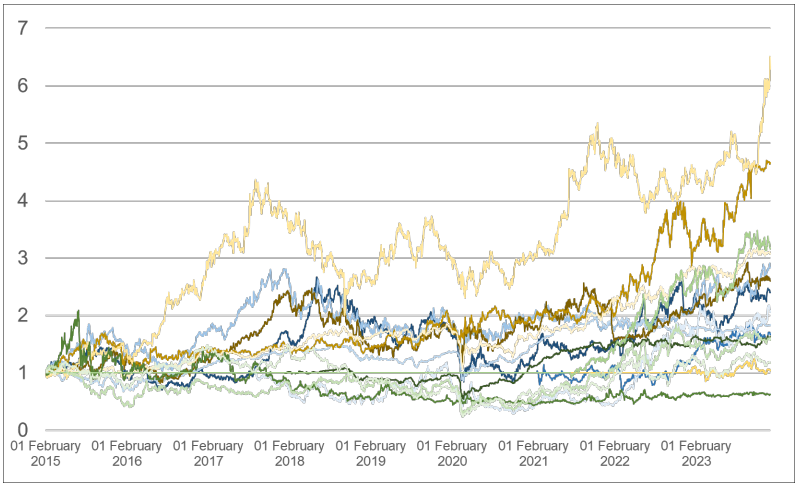
Over the period, 21% of the variations in the fossil fuels equity market were explained by the Halliburton Company, 5% by the Murphy Oil Corporation, 2,5% by the Marathon Petroleum Corporation, 2,0% by the Sinopec Shandong Taishan Petroleum Corporation, 1,8% by the Pembina Pipeline Corporation, 1,6% by TotalEnergies and 67% by the 511 other oil and/or gas companies.

After calculating the political and geopolitical variables, and then calculating the weekly political trends, we train the different models mentioned in Subsection 4.1 and measure their forecast errors.

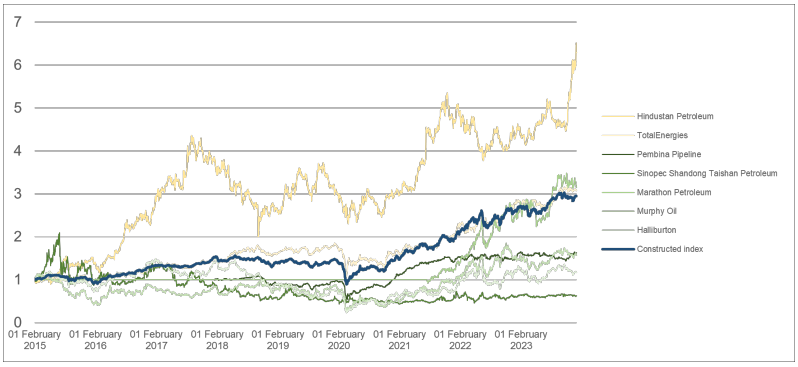
7.1 Forecast error

Regarding the assessment of the error, we favour indicators used in market finance which are efficient to measure the deviation of a portfolio from a benchmark index. Indeed, the forecast time series $\hat{Y}$ can be compared to an asset portfolio seeking to replicate a benchmark index, which in our case corresponds to our time series to be forecast Y. The commonly used indicators which can be adapted to our issue are the beta, the deviation of returns at the end of the analysis period (or error) (11) and the tracking error (TE).

[7] GDELT. https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/
[8] Yahoo Finance. https://fr.finance.yahoo.com/screener/new?lang=fr-FR&region=FR



(a) Main contributors to fossil fuels equity market variance.



(b) Constructed index representative of the fossil fuels equity market variance.

Fig. 2: Construction of a statistically representative index of the fossil fuels equity market from February 2015 to December 2023, using PCA.

The beta (8) measures the covariations between the forecast and the series to be forecast, adjusted by the volatility of the latter. A beta higher than 1 means that the forecast amplifies the variations of our market index. A positive beta of below 1 means that the forecast is more conservative than the index to be forecast. And a negative beta means that the forecast varies in the opposite direction to the index.

We then define a measure that we will name the target beta's absolute error (TBAE). The target beta being 1. The aim is then to minimise the TBAE (9).

The TE (12) is used to measure the volatility of the deviations between the forecast and the series to be forecast.

Model	Prediction data	β	$\lvert 1-\beta \rvert$	Annualised MAE	Error	TE
KNN regression	**Political events + Feature engineering**	**0.87**	**0.13**	**0.10%**	**0.53%**	**0.94%**
KNN regression	Political events	0.85	0.15	2.88%	-3.41%	1.22%
Huber regression	Political events	0.81	0.19	1.48%	7.31%	3.50%
KNN regression	Feature engineering	0.84	0.16	4.31%	-6.33%	1.23%
ARD regression	Feature engineering	0.42	0.58	-1.81%	3.42%	1.90%

Table 1: Forecasting models evaluation by cross-validation.

Finally, we will also measure the MAE (10), so that our results can be compared with other works in the literature. This measurement will be annualised to avoid overestimating the error due to the length of the analysis period.

$$\beta = \frac{Cov(\hat{\Delta}_{test}, \Delta_{test})}{\sigma_{test}}, \tag{8}$$

$$TBAE = |1-\beta|, \tag{9}$$

$$Annualised\ MAE : (1 + \frac{1}{\frac{1}{3}n} \sum_{t \in I_{test}} |\Delta_{test,t} - \hat{\Delta}_{test,t}|)^{\frac{52}{\frac{1}{3}n}} - 1 \tag{10}$$

with I_{test} the periods of the test sample,

$$Error : \hat{Y}_{test,t_{max}} - Y_{test,t_{max}} \tag{11}$$

such that $t_{max} = max(I_{test})$,

$$TE = \frac{1}{\frac{1}{3}n - 1} \sqrt{\sum_{t \in I_{test}} (\hat{\Delta}_{test,t} - \Delta_{test,t})^2}. \tag{12}$$

7.2 Results

This produces the following results: presented in Figure 3 and Table 1. We can see that the KNN regression taking into account the political information and the market data from the index (KNN-PE+FE) obtains better results for all the selected evaluation measures. This model has a TBAE of 0.13, compared with a TBAE of 0.15 for the KNN regression taking account only of political information (KNN-PE) and a TBAE of 0.16 for the KNN regression only including market data (KNN-FE). Our KNN-PE+FE model also achieves an annualised MAE of 0.1%, i.e. it overestimates fluctuations in the fossil fuel equity market by an average of +0.1% per year. A TE of 0.94%, i.e. less than 1, puts it in line with an index fund seeking to replicate our build up market index. So we can conclude that the fossil fuels equity market does integrate political events.

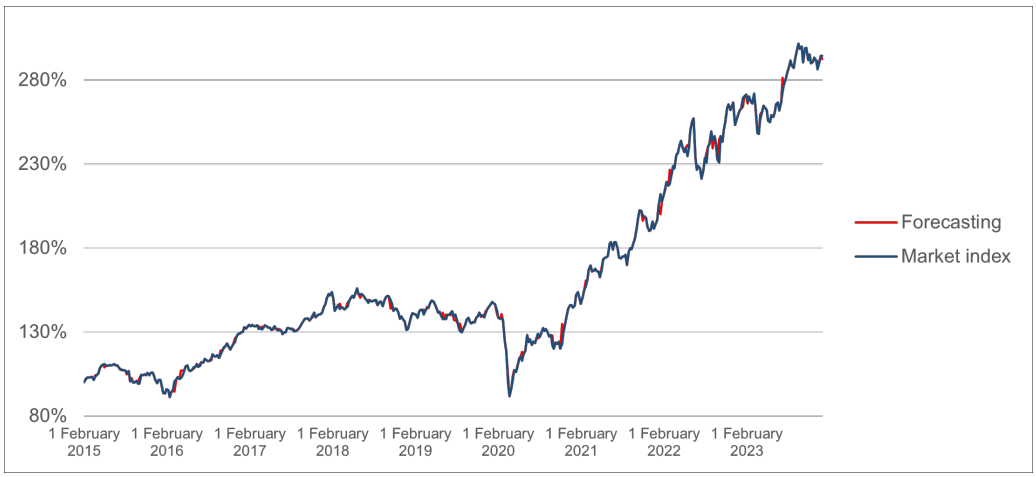


Fig. 3: KNN-PE+FE model forecasting compared to the built index (cross-validation step 3).

8 Conclusion

The link between the politics and the fossil fuels market has always been well known. The link between relatively static political data and this market has already been mathematically demonstrated. But the issue of measuring the impact of the mass of political events that take place every week on this market, despite their noise, remains scarcely addressed.

In this article, we have therefore proposed a method to test the integration of dynamic political information into fossil fuels equity market prices. A comparative approach of forecasting errors of two machine learning models: one integrating this information and one excluding it ; has been implemented. To address this issue, the optimal forecasting model, as well as the appropriate prediction variables and error evaluation formula, were sought for each case. A statistically representative index of our market of interest was also built by the PCA from over 900 mid-capitalisation, large-capitalisation and mega-capitalisation oil and gas stocks.

Over the period from February 2015 to the end of December 2023, our approach has shown that political events are well integrated into the fossil fuels equity market. The political trends – calculated from the GDELT event database – mainly based on the importance of events in the press opinion, their veracity, the geopolitics of their actors, their geographical location, their tone and their typology enable us to improve the pricing of our market of interest.

In the future, it will therefore be necessary to precisely assess the impact of each calculated variable on the performed forecast. Among other things, to assess the impact of the proposed division of the international system into different geopolitical spheres, and therefore the accuracy of the analysis carried out. This would be a contribution to the analysis of contemporary international relations thanks quantitative ways. Another avenue for further exploration lies in the use of more complex types of neural networks,

as well as large language models (LLM). The continuous improvement of forecasting models and their explanatory power will make it possible to quantify the links between cause and effect and thus better measure the political risk applied to investment.

Appendix A Computed political trends

- The weighted average classification according to the Goldstein scale and its development,
- The number of events of each type (granularity 1 of the CAMEO classification) and its development,
- The number of events taking place in each sub-region of the world (according to the classification of the UN geographical scheme) and its development,
- The number of actors of each type (granularity 1 of the CAMEO classification) and its development,
- The number of events involving two actors from OPEC+ member countries and its development,
- The number of events involving actors from different geopolitical spheres and its development,
- The number of events involving actors from OPEC+ countries and from different geopolitical spheres, and its development,
- The number of events involving actors from countries with major natural gas reserves and from different geopolitical spheres, and its development,
- The number of events involving actors from great powers and different geopolitical spheres, and its development,
- The number of events involving actors from OPEC+ countries, great powers and different geopolitical spheres, and its development,
- The number of events involving actors from countries with major natural gas reserves, from great powers and from different geopolitical spheres, and its development,
- The weighted average tone and its development,
- The weighted average Goldstein classification in each sub-region of the world (according to the classification of the UN geographical scheme) and its development,
- The weighted average Goldstein classification of events involving two actors from OPEC+ member countries and its development,
- The weighted average Goldstein classification of events involving actors from different geopolitical spheres and its development,
- The weighted average Goldstein classification of events involving actors from OPEC+ countries and from different geopolitical spheres, and its development,
- The weighted average Goldstein classification of events involving actors from countries with major natural gas reserves and from different geopolitical spheres, and its development,
- The weighted average tone of events corresponding to economic sanctions decreed by or against an actor from an OPEC+ member country, and its development,
- The weighted average tone of events corresponding to armed conflicts involving an actor from an OPEC+ member country and its development,

- The weighted average tone of events corresponding to institutional changes within OPEC+ member countries and its development,
- The weighted average tone of events corresponding to internal troubles within OPEC+ member countries and its development,
- The weighted average tone of events corresponding to commercial agreements between two OPEC+ member countries and its development,
- The weighted average tone in each sub-region of the world (according to the classification of the UN geographical scheme) and its development,
- The weighted average tone of events involving two actors from OPEC+ member countries and its development,
- The weighted average tone of events involving actors from different geopolitical spheres and its development,
- The weighted average tone of events involving actors from OPEC+ member countries and from different geopolitical spheres, and its development.

References

[1] Alfred, R.: Impact de la guerre en ukraine sur le marché de l'energie fossile. Master's thesis, Université Paris 1 Panthéon-Sorbonne, Paris, France (July 2022)

[2] Van Beers, C., Strand, J.: Political determinants of fossil fuel pricing. Political Economy and Instruments of Environmental Politics, 71 (2015)

[3] Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M.: Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. Applied Soft Computing **90**, 106181 (2020) https://doi.org/10.48550/arXiv.1911.13288

[4] Bartram, S.M., Branke, J., Motahari, M.: Artificial Intelligence in Asset Management. CFA Institute Research Foundation, Charlottesville (2020)

[5] Aron, R.: Paix et Guerre Entre les Nations. Calmann-Lévy, Paris, France (2004)

[6] Allain, J.-C., Frank, R.: Chapitre 6: La hiérarchie des puissances. In: Frank, R. (ed.) Pour l'Histoire des Relations Internationales, pp. 169–186. Presses Universitaires de France, Paris, France (2012)

[7] Zakoïan, J.-M.: Modèle autoregressif à un seuil. In: Annales de l'ISUP, vol. 37, pp. 85–114 (1993)

[8] Krollner, B., Vanstone, B., Finnie, G.: Financial Time Series Forecasting With Machine Learning Techniques: a Survey. In: European Symposium on Artificial Neural Networks: Computational Intelligence and Machine Learning, Bruges, Belgium, pp. 25–30 (2010)

[9] Ding, X., Zhang, Y., Liu, T., Duan, J.: Deep learning for event-driven stock prediction. Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)

[10] Alamro, R., McCarren, A., Al-Rasheed, A.: Predicting Saudi Stock Market Index by Incorporating GDELT Using Multivariate Time Series Modelling. In: Alfaries, A., Mengash, H., Yasar, A., Shakshuki, E. (eds.) Advances in Data Science, Cyber Security and IT Applications. ICC 2019. Communications in Computer and Information Science, vol. 1097, pp. 317–328. Springer, Riyadh, Saudi Arabia (2019). https://doi.org/10.1007/978-3-030-36365-9_26

[11] Yonamine, J.E.: A nuanced study of political conflict using the global datasets of events location and tone (gdelt) dataset. PhD thesis, Pennsylvania State University, State College, United States (March 2013)

[12] Van Bruggen, B.O.: International conflict's impact on the stock market: a gdelt project-based event study. Master's thesis, Erasmus University Rotterdam, Rotterdam, Netherlands (September 2018)

[13] Nanga, S., Bawah, A.T., Acquaye, B.A., Billa, M.-I., Baeta, F.D., Odai, N.A., Obeng, S.K., Nsiah, A.D.: Review of dimension reduction methods. Journal of Data Analysis and Information Processing **9**(3), 189–231 (2021) https://doi.org/10.4236/jdaip.2021.93013

[14] Ritchie, H., Rosado, P.: Fossil fuels. Our World in Data (2017). https://ourworldindata.org/fossil-fuels

[15] Schrodt, P.A.: Cameo: Conflict and mediation event observations event and actor codebook. Technical report, Pennsylvania State University (2012)

[16] Energy Institute – Statistical Review of World Energy — with major processing by Our World in Data: "Gas proved reserves" [dataset]. Energy Institute, "Statistical Review of World Energy" [original data]. https://ourworldindata.org/grapher/natural-gas-proved-reserves (2023)