IAQF Student Competition 2021

Team Bond

February 2021

1 Introduction

Oil is the world's most important non-food commodity and its price impounds information from a vast range of agents, who react to economic conditions as well as anticipated changes in technology and regulation. Is this process so effective that it can tell us something about the future? In this paper, we explore this by examining whether oil prices help predict future changes in retail gasoline prices and atmospheric carbon dioxide (CO_2).

In both cases we begin with a conceptual framework of the process we attempt to model. Starting with a conceptual overview is important because we utilize non-linear modelling techniques which have the potential for over-fitting sample data. To mitigate this risk, we place a premium on parsimonious feature selection derived from our process overview. This is critical when attempting to model atmospheric CO_2 changes as this process is exceptionally complex.

2 Predicting Gasoline Prices

The conceptual framework linking oil to retail gas prices is straightforward (Figure 1). Data from California's energy commission shows more than 40 percent of retail gasoline price is the cost of crude, with 25 percent coming from taxes and other 35 percent from refinery and distribution costs.[1]. Assuming taxes change only occasionally, this framework suggests oil prices and refinery margins should be the main inputs into a model of retail gas prices.

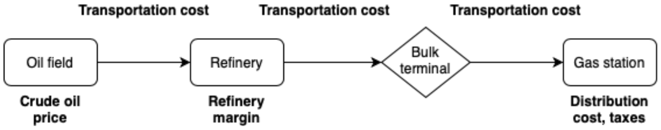

Figure 1: Simplified gasoline supply chain

2.1 Statistical Analysis of Price Series

2.1.1 Statistical Tests

Statistically oil and gas prices are very much connected. Using data from 1993-2021, we find that crude oil (WTI) futures and gasoline (RBOB) 1-month futures have a correlation coefficient of 0.95 (Table 1). This phenomenon persists if we switch from futures prices to spot retail prices - essentially adding the effect of gasoline distributors into the mix. We find that the correlation between WTI and retail gas is between 84 percent and 93 percent, decreasing in the grade of gasoline. This can be due to multiple reasons - higher taxes, lower yield, or simply the markup charged for a premium product. This effect persists whether we are in the price space or in the return space, and it remains when we lag the crude oil prices. In this paper we report results for regular grade gas only.

	Crude 1	RBOB 1	Capacity	Regular	Midgrade	Premium
Crude 1	1.00	0.95	-0.55	0.93	0.88	0.84
RBOB 1	0.95	1.00	-0.50	0.98	0.94	0.91
Capacity	-0.55	-0.50	1.00	-0.43	-0.28	-0.18
Regular	0.93	0.98	-0.43	1.00	0.98	0.96
Midgrade	0.88	0.94	-0.28	0.98	1.00	0.99
Premium	0.84	0.91	-0.18	0.96	0.99	1.00

Table 1: Correlation between crude oil, refining capacity, and gasoline prices

The relationship between oil and gas prices over time is presented in Figure 2. This visual relationship, together with our conceptual model, suggests

[1] Retrieved from https://ww2.energy.ca.gov/

the two series are cointegrated. We tested this by first performing an augmented Dickey-Fuller (ADF) test, which found both series to be nonstationary.[2] The differenced series of both commodities are stationary. We then applied Engle-Granger cointegration test to the timeseries and found both series to be significantly cointegrated.

Engle and Granger [1987] show that cointegrating time series imply long-term predictability as errors mean-revert. To assess the statistical impact of oil price changes on gasoline price changes, we implemented a vector error-correction model (VECM) in line of Chen et al. [2005]. The VECM framework allows us to distinguish the lead-lag relationship of these price series, and test for statistical significance of the parameters.

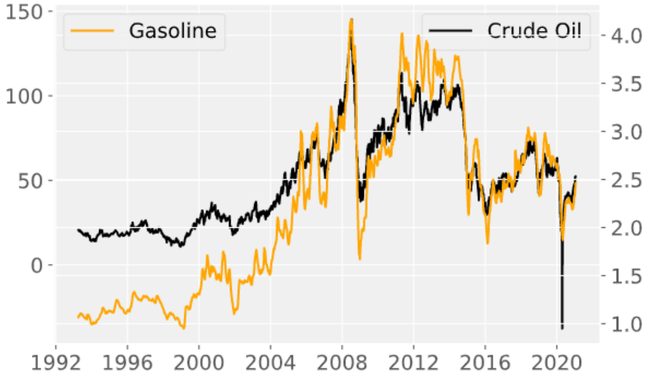

Figure 2: Price series for crude oil and gasoline

2.1.2 Analyzing Asymmetric Price Impact with Error Correction Models

In this section we use the VECM framework to test for asymmetries in the relationship between oil and gas prices. We use weekly data for all modelling and test the following two hypothesis:

- *Hypothesis 1*: People expect gasoline prices to increase when crude oil prices are rising and decrease when oil prices are falling.
- *Hypothesis 2*: The impact of crude oil on gasoline prices is same when oil prices are rising as when they are falling.

To test the above two hypothesis, we first need to define the regions of rising and falling crude oil prices. We define these regions using an equilibrium model proposed by Chen et al. [2005]. This model utilizes the fact that gasoline and oil prices are cointegrated. The equilibrium model is defined as:

$$G_t = \zeta_0 + \zeta_1 O_t$$

where G_t denotes the gasoline price and O_t denotes the oil price. Defining regions of rising and falling oil prices using this model works better than the conventional method of using price changes because oil prices fluctuate a lot. For example, consider two scenarios, (1) The oil price rises by \$10 and then drops by \$2, (2) The oil price increases by \$7 and then increases by a further \$1. In both cases, the crude oil price increased by net \$8 and thus we would expect the gasoline price to follow this increase. The equilibrium model will treat both cases as belonging to the same regime whereas the conventional model will consider a regime shift in scenario (1). Thus, the equilibrium model is more robust to small fluctuations in the crude oil price. The falling and the rising regimes are defined with respect to what we expect from gasoline prices. Mathematically,

Rising Regime (+): $G_t < \zeta_0 + \zeta_1 O_t$
Falling Regime (-): $G_t > \zeta_0 + \zeta_1 O_t$

Based on these two regimes, we then run an error correcting model to test our hypothesis. The error correcting model is defined as:

$$\Delta G_t = \begin{cases} \gamma_+(G_{t-1} - \zeta_0 - \zeta_1 O_{t-1}) + \\ \sum_{i=1}^{p} \alpha_{+,i}\Delta O_{t-i} + \beta_{+,i}\Delta G_{t-i}, & G_{t-1} < \zeta_0 + \zeta_1 O_{t-1} \\ \\ \gamma_-(G_{t-1} - \zeta_0 - \zeta_1 O_{t-1}) + \\ \sum_{i=1}^{p} \alpha_{-,i}\Delta O_{t-i} + \beta_{-,i}\Delta G_{t-i}, & G_{t-1} > \zeta_0 + \zeta_1 O_{t-1} \end{cases}$$

where $\gamma_{+/-}$ denotes the error correcting coefficient for rising/falling regimes and $\alpha_{i,+/-}$ and $\beta_{i,+/-}$ denotes the lagged coefficients for gasoline and oil returns respectively. In this model, the lagged returns coefficients ($\alpha_{i,+/-}$ and $\beta_{i,+/-}$) govern the short-term dynamics of the gasoline prices whereas the long-term dynamics are governed by the error-correcting coefficients ($\gamma_{+/-}$).

Table 2 shows the error correcting model coefficients and their respective t-stat values for regular grade gasoline. From the table we can see that the

[2] The oil timeseries has p-value 0.15, gasoline has p-value 0.42.

Coefficients	Rising Regime (+)	Falling Regime (-)
Error Correcting (γ)	-0.010 (**-2.60**)	-0.010 (**-2.82**)
Lag1 Oil Returns (α_1)	-0.022 (-1.11)	0.008 (0.70)
Lag1 Gasoline Returns (β_1)	0.646 (**11.09**)	0.482 (**9.12**)
Lag2 Oil Returns (α_2)	-0.075 (**-3.89**)	0.011 (0.95)
Lag2 Gasoline Returns (β_2)	0.207 (**3.76**)	0.029 (0.55)

Table 2: Error correcting model estimates

gasoline prices error correct in the direction of the move in crude oil prices (γ is negative and significant) in both the falling and rising regimes. In other words, the evidence suggests that Hypothesis 1 is valid - when crude oil prices increase (decrease), gasoline prices then follow it and also increase (decrease).

We then conduct F-tests to see if oil's impact on gasoline prices is asymmetric depending on whether oil prices are rising or falling. We conduct separate tests for error correcting coefficients (long-term impact) and the lagged oil and gasoline returns coefficients (short-term impact). The null hypothesis in these tests is that the coefficients are the same in rising and falling regimes, i.e. there is no asymmetric impact. The results of the tests are shown in Table3.

Hypothesis	F-statistic (p-value)
H_0: $\gamma_+ = \gamma_-$	0.003 (0.95)
H_0: $\alpha_{+,i} = \alpha_{-,i}$ and $\beta_{+,i} = \beta_{-,i}$	5.036 (**1e**$^{-4}$)

Table 3: Testing for asymmetric behavior

The top panel shows that there is no significant evidence to conclude that the error correcting coefficients are different. However, the F-test results and p-values (in parenthesis) in the second panel are strong evidence that the lagged returns coefficients are significantly different in different regimes. What it shows is that in the short term, the autoregressive terms dominate and there is a different response to crude oil increases and decreases. The error-correction terms on the other hand are statistically the same, which means in the long term, price must converge to equilibrium at a similar speed.

2.2 Prediction with Machine Learning

2.2.1 Model Selection and Hyperparameters

In the previous section we established the predictability of gasoline prices using crude oil prices and other factors. In this section we explore nonlinear machine learning (ML) models to see if the same effect persists. We explore multiple ML models - both parametric and non-parametric - to see if we not only can get better next-period predictions, but also predictability over longer periods.

Machine learning models capture non-linear relationships that can improve predictability. However, this comes at a cost. It can be difficult to explain the results as there are no fixed linear parameters to "sense check" against priors. Further there is a risk of over-fitting, leading to poor out-of-sample performance. Because it's more difficult (but not impossible) to intuitively "sense check" these models we place a premium on careful feature selection, with all our our choices motivated by the conceptual overview of the gas price process presented at the beginning of the section.

We use three different models, support vector regression (SVR), k-nearest neighbors (kNN), and feed-forward neural networks (NN). We use linear regression (LR) as the baseline for the comparison.

These are not arbitrary choices of models - SVR is an extension of linear regression where we introduce a nonlinear kernel by which we project data into a hyperspace for easier fitting. It also introduces a slack variable ε where data points that fit within the band $\hat{y} - \varepsilon < y < \hat{y} + \varepsilon$ do not count toward the error. We chose a weight of 1 in L2 error to reduce overfitting, and $\varepsilon = 10^{-11}$.

kNN is a popular nonparametric model where instead of fitting a parameter, we look toward the past to find the best period closest to the present, then average the price of gasoline in those periods to arrive at a prediction. This avoids the need to fit a model, instead requires a good choice for hyperparameter k, the number of neighbors to consider. A high k avoids overfitting, but introduces variance as it selects neighbors too far, while a low k is more precise, but are more prone to overfitting. Through backtesting we select a $k = 60$ for this case.

Finally we have neural networks, a popular and powerful framework for prediction through multiple layers of nonlinear units applied to linear transformations. The shallow network has 1 layer of 100 units, with a L2 parameter of 0.0001, and trained using Adam with momentum 0.9.

2.2.2 Model Evaluation

Since ML models do not present a simple statistical way to test performance out-of-sample, we use cross-validation to assess each model's performance. To do that, we split our data, which is weekly, into a training period (2006-2016) and a test period (2016-2020). There are no structural breaks in either training or test datasets.

Our predictors are 0-, 3-, 6-, 12-month lagged front-month crude oil and gasoline futures prices, and current refining capacities. We selected these variables since current retail gasoline prices reflect current and past refining costs and input prices. The variable of interest is t-period future retail gasoline prices for regular grade gasoline. We compare model performance using R^2 scores, which measures the variance explained by our model. The best score is 1, neutral score is 0 (if we expected a constant expected value of y), and it could be negative if our model is worse than a constant. We process the data by demeaning and scaling with an expanding window so that the variable would have mean zero and standard deviation one at each time step. This is necessary for kNN because it makes distance measures comparable for variables of different scale, such as price for gasoline and refinery capacities.

2.2.3 Results

Table 4 shows the out-of-sample results for the machine learning models and the linear regression baseline. For regular grade gasoline, the linear regression added value for up to 2 weeks. It explains about 16 percent of variance within the data. kNN performed either on-par, or better than the linear regression baseline in all t, maintaining predictive power for up to 3 weeks in the future. SVR also performed better than linear regression. While the results seem to show that machine learning models can predict gasoline prices better, we note that it is hard to make scientific conclusion given the black box nature of machine learning models. Here we attempt to shine some light on how the models operate using k-fold analysis of feature importance in figure 3.

t	LR	SVR	kNN	NN
1 week	0.16	0.16	0.16	0.14
2 weeks	0.02	0.05	0.06	0.02
3 weeks	-0.01	-0.01	0.03	-0.00
4 weeks	-0.01	-0.01	-0.02	-0.05
8 weeks	-0.01	-0.01	-0.01	-0.10
12 weeks	-0.02	-0.01	-0.04	-0.08

Table 4: Model scores, regular grade returns

Figure 3 shows feature importance scores for different models, we use it to distinguish the asymmetric impact of crude oil price in our machine learning models. The feature importance is measured by the difference in evaluation scores (in this case R^2) when the feature is normal versus when the feature is permuted. We permute the feature 10 times to generate 10 different feature importance scores. The boxplot shows median (yellow line), 25, 75 percentile (box), 1.5 times interquantile range (the vertical line), and outliers (dots).

Clearly for each model RBOB (gasoline futures price) is the most important feature for all models, while capacity and refinery margin are not important in most models, except kNN, where refinery margin is the second most important. This can partly explain the outperformance of kNN compared to other models, since it is utilizing more information from data.

Figure 3 also shows evidence of the asymmetry detected in the VECM model from the previous section. The LR, kNN, and neural networks models see a significantly greater feature importance for the oil price when it is above the equilibrium predicted by crude oil price (+Crude 1, similarly defined as the previous subsection), than when it is below the equilibrium price. This is potential evidence of tactical price adjustments by retailers, adjusting gas prices higher when oil rises but not lowering commensurately when oil falls.

In summary, oil prices have some ability to predict changes in gasoline prices over the next week, and several machine learning methods (SVR, kNN) can possibly extending this a few weeks further into the future. The relationship is strongest when oil prices are rising.

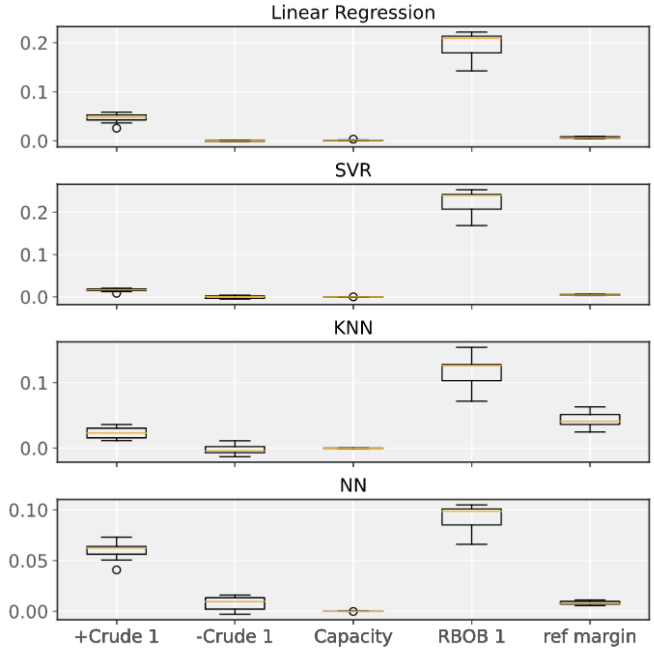

Figure 3: Feature importance across models, regular gasoline

3 Predicting CO_2 Concentration

In this section we focus on the second question - predicting CO_2 concentration levels in the atmosphere. In contrast to predicting gasoline prices using crude oil prices - a case where there is a clear economic link between the two - predicting CO_2 concentration is fundamentally different. CO_2 levels change due to a variety of different causes, and we expect a model which largely relies on only one aspect (e.g crude oil prices) to have a lower ”signal-to-noise ratio”, and less predictive power, than we saw in section 2. We therefore begin this section by describing our conceptual framework, which provides an overview of the key drivers of changes in CO_2 concentration. We next provide the problem formulation, followed by the features and prediction model, and lastly, a discussion about the implications of our forecasting framework.

3.1 A Conceptual Framework

There are two main goals for starting with a ”conceptual framework” to model the changes in CO_2 levels in the atmosphere. First, we get a clearer picture of what models, or what features, can assist us in creating a CO_2 level forecast. Second, we get a better sense of what we *cannot* model - and how to properly analyze the forecast performance given the nature of the problem.

We break down changes in CO_2 into changes in consumption and emissions. Those changes can have either a negative or positive effect on CO_2 levels. The diagram below depicts our model. The top section of the diagram illustrates that CO_2 changes are driven by fossil fuel emissions, renewable energy uses and organic sources like plants and animals. The second panel shows the key sources of fossil fuel emissions. In our modelling we will build and test proxies for each of the three drivers of CO_2 changes, with our fossil fuel proxies informed by the industrial composition of energy usage.

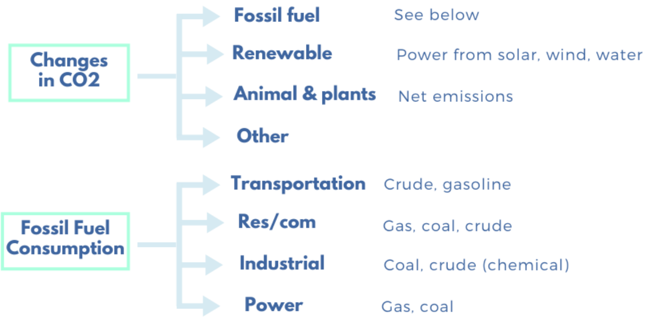

Figure 4: Model for changes in CO_2 levels

3.2 Formulating the Prediction Problem

As a first step to define the forecast goal, we begin by choosing the *frequency* of our model. CO_2 levels are observed daily. However, we choose to down-sample this series to a monthly frequency. The reason for that is two-fold. First, some of the features we have are available only on a monthly basis. Second, our framework is not well-suited for high-frequency (say daily or weekly) changes, and it is less likely to derive a meaningful prediction in those cases. Taking a longer horizon (say yearly), on the other hand, would result in a small sample, creating possible ”overfitting” issues in a supervised machine learning framework.

Next, we note that carbon-dioxide concentration values exhibit two strong behaviors: trend and seasonality. It is well known that many time series in finance, physics and other areas can be decomposed into several components: given a time series y_t, we may often write:

$$y_t = T_t + S_t + R_t$$

where T_t is a *trend* component (which is not necessarily linear), S_t is a *seasonal* component (a sine-like wave function with a constant period which fluctuates around 0), and R_t - defined as the rest - is called the *residual* component.

Below is the time-series decomposition for the CO_2 concentration values. We observe a strong seasonal component with one-year period.

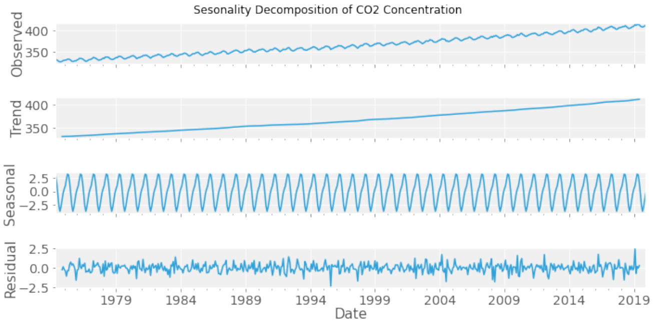

Figure 5: Trend (T_t), seasonal (S_t) and residual (R_t) components

There is a well-known scientific reason for this phenomenon called the ”Seasonal CO_2 Cycle”. The CO_2 concentration is affected by plants (one of the components of our framework) in a different way throughout the year, leading to higher CO_2 levels during the colder months[3]. Since the seasonal component is predictable, we focus on the *de-seasonalized* time series $T_t + R_t$. We note that the decomposition is done by using one-sided moving averages, and therefore does *not* introduce any ”forward-looking” bias. Therefore, our model will provide a monthly prediction for the changes in the de-seasonalized CO_2 level: $\Delta(T_t + R_t)$, using a set of features, which we discuss below.

3.3 Creating Features

Our features closely follow our framework, and include a variety of different sources and ideas. We note that crude oil prices are used in several places, both directly and indirectly.

Our approach to this problem is to carefully ”hand-craft” our features, deriving a small number of them to use in our model. Given the scientific nature of the problem, a pure ”data-mining” approach can lead to the omissions of important variables while introducing an ”overfitting” issue that we wish to avoid. Similar to Section 2, we validate the significance of each individual feature using a regression benchmark - and although the final model is far from linear, doing so gives an idea of the applicability of the defined features.

3.3.1 Emitted CO_2 Dynamics in the Atmosphere

The most immediate features to use is the total energy emissions values. Doing so correctly, however, is not trivial, and some careful ”feature-engineering” is required. It turns out that CO_2 emitted into the atmosphere decays exponentially, much like radioactive substances decay over time. The exact formula is given by the Bern Carbon Model (see, for example, [Kharecha and Hansen, 2008]),

$$w^B(t) = 0.18 + 0.14e^{-t/420} + 0.18e^{-t/70} + \\ 0.24e^{-t/21} + 0.26e^{-t/3.4}$$

Here, $w^B(t)$ signifies the proportion of CO_2 remaining in the atmosphere after t years. The figure describes the weights $w^B(t)$ for 50 years after emission.

The changing weights over time mean that, when analyzing the impact of emissions on CO_2 levels, in order to correctly account for the decay we need to apply convolution over a rolling window with $w^B(t)$. Therefore, in our model, the monthly level of CO_2 emissions and the change in CO_2 emissions are re-

[3] Retrieved from https://svs.gsfc.nasa.gov/4565

placed by the Bern-Decay adjusted emission[4]:

$$e^B(t) := \Delta^k \sum_{s=0}^{t} e(s)w^B(s) \text{ for } k = 1,2$$

where $e(s)$ is the emission at time s, and w^B are the decay weights defined above. These features seem to have a significant impact on emissions, giving t-statistics of 4.5 and 6.7 for $k = 1$ and $k = 2$, respectively.

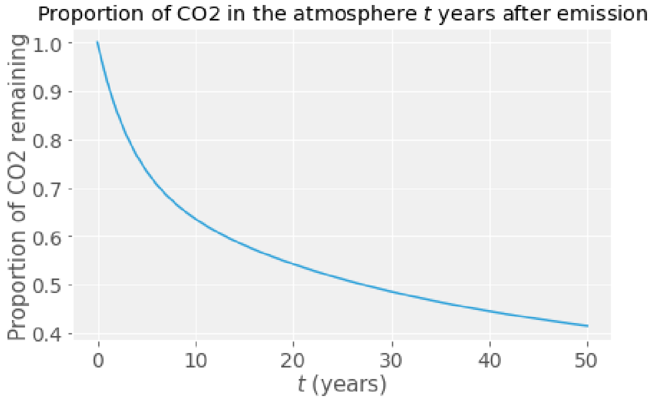

Figure 6: Decay of Emitted CO_2 in the Atmosphere, Bern Carbon Model

3.3.2 Economic Activity

Another metric which is closely related to CO_2 emissions is economic activity.

In the simplest terms, an increase in economic growth is linked to an increase in energy consumption, and therefore emissions. An important exception, which makes this connection less trivial, is that - as mentioned in our framework - renewable energy consumption could also increase, and some studies (see Lahiani et al. [2017]) suggest that this is indeed the case, at least in some countries and for longer time horizons. Therefore, in order to account for the part of the economy related to non-renewable energy, we directly use the prices of crude oil, natural gas and coal. Among these three, oil is known to have the largest effect on emissions[5] and the data suggests that changes in oil price give the best proxy for fossil-fuel emitting economic activity, giving a t-statistic of around 2.2. To be as parsimonious as possible, we choose to include only the changes in crude oil price, as both coal and natural gas are highly correlated to it (around 60% correlation) and do *not* seem to provide any additional predictive power to our model (for example, trying the first principle component of the three performed worse than directly using oil prices).

3.3.3 Animals

Going back to the diagram outlined above, recall that animals are another source of emissions. Studies show (Rotz [2018]) that different stages in the production cycle of dairy, beef and similar products create CO_2 emissions. Much like when giving a proxy for economic activity, we choose to approximate this part using a related financial asset data. As a proxy for the demand of dairy products, we obtain a series of milk prices. Using those prices directly, however, would not be very meaningful to model emissions, as external shocks in milk demand, as well as inflation, are known to effect the price. To mitigate the aforementioned effects we introduce a - somewhat creative - new feature, defined as the ratio between milk prices and eggs prices, or more specifically, the ratio between one gallon of fortified milk and dozen grade A large eggs. The logic is that both products are food staples but the production of milk involves significant CO_2 emissions while egg production does not. An increase in the price of milk relative to eggs thus helps us proxy for the relative change in CO_2 emissions from food production. Taking *changes* in this ratio gives a significant t-statistic of 2.2 (while milk price gives a lower, non-significant t-statistic of 0.4, confirming our assumption).

3.3.4 Energy Companies Valuation

Another way to derive information on CO_2 levels from financial markets is to look at the valuations of energy companies. We take the S&P Energy Index, which consists of the biggest US energy companies, and regress its cyclically adjusted P/E ratio on that of the S&P 500 (known as the Shiller CAPE index), and on changes in oil and gas prices. The rolling regression provides a simplistic model for an energy

[4] Since the convolution acts like an integral and represents cumulative values, we take the first derivative, which is analogous to the monthly emissions, and the second derivative, analogous to change in monthly emissions.

[5] Retrieved from https://www.eia.gov

company's valuation:

$$\Delta\text{Energy CAPE} = \beta_1 \Delta\text{Shiller Index} + \beta_2 \Delta\text{Oil} + \beta_3 \Delta\text{Gas}$$

We then extract the *unexplained part*, i.e the regression residual, and use it as a feature in our CO_2 prediction model. The logic behind it is as follows: apart from general market multiples and oil and gas prices, the unexplained part of the energy sector CAPE relates to future changes in CO_2 levels: changes in the trend may impose regulations, policies and ultimately a shift to renewable alternatives - all hurting the value of those companies. The data confirm our assumption, and we indeed notice a *negative* correlation to CO_2 levels, with a t-statistic of around 1.8.

It is worth mentioning that the R^2 of the rolling regression is decreasing over time, suggesting that those other considerations (like regulations and policies) are becoming a larger part of the valuation models of energy companies in recent years.

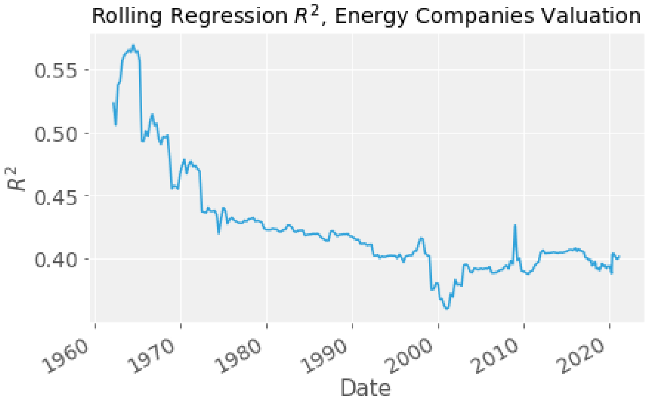

Figure 7: R-Squared of Rolling Regression, Simple Energy Industry Multiple Model

3.3.5 Renewable Energy

Lastly, we note that a missing piece of the puzzle is changes in renewable energy. In contrast to the other elements we described above, the influence of renewable energy on CO_2 concentration is somewhat trickier: an increase in renewable energy consumption can mean an overall increase in economic activity and, consequently, an *increased* level of CO_2 emissions. There are, however, two possible indirect ways to incorporate this aspect.

- One is a similar approach to our energy sector valuation approach: an increase in the sector-specific valuation multiples can be indicative of expected increases in CO_2 levels, and the need to utilize zero emission energy sources. To try this idea, we took the P/B multiples of leading renewable energy companies, regressed on that of the S&P, and used the residual in the model.
- A second approach to give is to analyze carbon credit data. An increase in the price of the right to emit CO_2 can accelerate a shift to renewable sources of energy.

These two implementations gave some interesting results: both gave a non-negligible t-statistics of 1.8 and 1.7 respectively, but their impact on CO2 emissions is *positive* - meaning that they likely relate to increased economic activity rather than an increased usage of renewable sources of energy. Combined with the fact that neither gave an improvement in our non-linear model, we decided not to include these features. We note, however, that those ideas might play an increasingly significant role in the future.

3.4 The Model

In this part we outline our forecast model. As explained in the problem formulation, we predict the one month difference in de-seasonalized CO_2 level, using the features described above. For concreteness, we list below the features we use:

- Bern decay adjusted CO_2 emissions, and its change.
- Change in crude oil price.
- Change in quotient of milk price over eggs price.
- Energy companies valuation regression residual.

Standard feature processing techniques (normalization, filling missing values) are then applied to derive a matrix of features.

3.4.1 Choosing A Model

Our intuition and understanding of the features and their effect on the problem indicate a non-linear behavior. We have several features which are related to emissions, and their cross impact is not easily described by a linear model. Similarly to the first part, we try different models, and compare their performance - MAE on the hold-out period. kNN provides a simple non parametric alternative; SVR is another popular method, and a shallow NN introduces non-linear activation functions to add complexity to the model. In addition to these, we try a Random Forest model, which has the advantage of having a highly non-linear decision boundary, and can be suited to this model, in which our features are different in nature and their interaction is potentially complex.

3.4.2 Evaluation and Results

We evaluate the forecast in several steps. First, we split our data in time approximately 70% and 30% Train/Test, and get the following results for the MAE of the test set (note that here, *lower* values indicate *better* results):

Model	LR	SVR	kNN	NN	RF
MAE	0.94	0.92	0.76	0.73	0.71

Table 5: Model MAE, Test Set

Next, we choose the best performing model - the Random Forest regression - and perform both rolling window (of size 50 months) and expanding window prediction schemes. Both have the advantage of taking more recent observations into account, while the first one can be more adaptive to changes in the dynamics of the problem.

Training Scheme	Static	Rolling	Expanding
MAE	0.71	0.55	0.53

Table 6: Random Forest Model MAE, Test Set, Different Training Schemes

We see that the expanding window model performs better, hinting that the dynamics of the problem change over time, but that even older observations are valuable in predicting recent values. Since we predicted the change in CO_2 values, taking cumulative sum gives the actual levels, and we get that our model gives the following prediction (for the test set):

Overall, we get a meaningful forecast which significantly outperforms the linear regression benchmark (see figure 8).

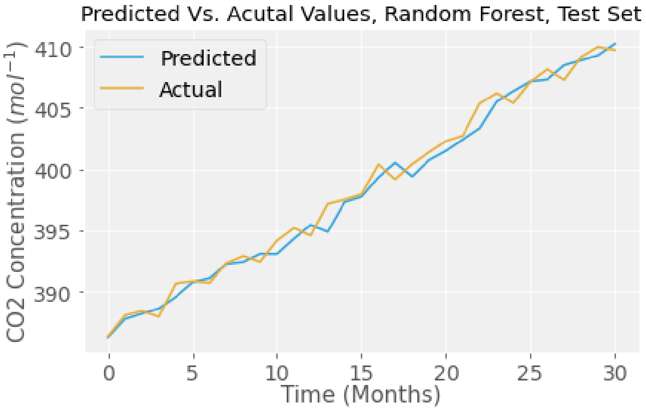

Figure 8: Model Forecast

3.4.3 Feature Importance

To better understand the impact oil prices, as well as other features, have on our prediction, we find the *Shapley values* of our different features. While there are several approaches to understand the individual contribution of each feature to a supervised prediction model, this concept - taken from game theory - is both robust, meaningful, and illustrative. The idea behind it is to find the average marginal contribution of a feature to some random subset of the other features - an idea reminiscent of the Shapley value of a player in a cooperative game. We get the following values for our model:

In figure 9, E1 denotes the bern-decay adjusted emissions, E2 the changes in E1, O the oil price, ME the milk price over eggs, and CR the regression residual for the energy sector valuation. Each subplot describes the distribution of the Shapley values, with higher values meaning a more significant impact on the model. We see that crude oil has a significant predictive power in our model, but its overall contribution is rather marginal. This is not surprising, as the two energy-emissions variables - having the largest importance in our model - likely subsume some of the information in oil price changes.

4 Implications and Discussion

Going back to the seminal paper of Roll [1984], financial markets seem to contain information way

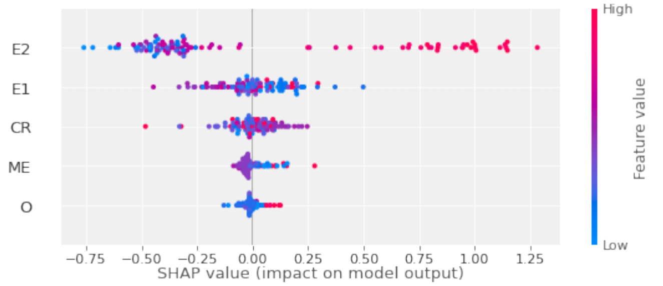

Figure 9: Random Forest Model, feature importance (Shapley values)

beyond their ”financial scope”. The impact external factors have on markets mean that, in some instances, asset prices can provide explanatory power over variables like weather forecasts and atmospheric CO_2 levels that extend beyond market themselves.

Since the publication of the original paper in 1984, there has been a tremendous progress both in developing and leveraging non-linear machine learning models. We saw that using machine learning models improves the forecast significantly, and exposes the non-linear nature of the problem. A careful analysis of the features importance shows that financial variables have predictive power and - when used correctly - can help in modeling subsequent changes in both consumer grade gasoline prices and CO_2 levels.

Given the direct economic and statistical relationship between crude oil and gasoline, it is inevitable that we find strong predictive power of crude oil prices on gasoline prices. We exploit the cointegration relationship to show the asymmetric price impact of crude oil price changes to retail gasoline price changes. By computing feature importance scores for machine learning models, we show that the asymmetric effect persists when we move into nonlinear models, and features like refinery margin add significant predictive power.

In the future, efforts to combat climate change mean that emission levels are going to decrease and - consequently - atmospheric CO_2 is going to change its trend and decrease as well. In that new regime, we suspect that crude oil prices will not be as meaningful a feature as it has in the past, but other financial instruments related to renewable energy sources might. Some of these financial instruments might even not exist yet or do not play an important role in current financial markets. Indeed, since financial markets exist to facilitate resource allocation for real-world needs there is a very strong case to be made for creating securities with payoffs directly linked to atmospheric CO_2. The evidence from our work supports Roll’s original idea that such securities would provide society with an important signal of our progress in solving the problem of climate change.

References

L.-H. Chen, M. Finney, and K. S. Lai. A threshold cointegration analysis of asymmetric price transmission from crude oil to gasoline prices. *Economics Letters*, 89(2):233–239, 2005.

R. F. Engle and C. W. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

P. A. Kharecha and J. E. Hansen. Implications of “peak oil” for atmospheric co2 and climate. *Global Biogeochemical Cycles*, 22(3), 2008.

A. Lahiani, A. Sinha, and M. Shahbaz. Renewable energy consumption, income, co emissions, and oil prices in g7 countries: The importance of asymmetries. *The Journal of Energy and Development*, 43(1/2):157–191, 2017.

R. Roll. Orange juice and weather. *The American Economic Review*, 74(5):861–880, 1984. ISSN 00028282. URL http://www.jstor.org/stable/549.

C. A. Rotz. Modeling greenhouse gas emissions from dairy farms. *Journal of Dairy Science*, 101 (7):6675–6690, 2018. ISSN 0022-0302. doi: https://doi.org/10.3168/jds.2017-13272. URL https://www.sciencedirect.com/science/article/pii/S002203021731069X.