**Correlation analysis of EUAs**

**Notes**

Stefania Albarosa

**Chapter 1,2,3**

This analysis focuses on the EUA market, although similar methods could be applied to any carbon market with available data. The variables included in the analysis are:

* Crude oil benchmark = oil
* TTF gas = gas
* German electricity = electricity
* Rotterdam coal = coal
* EuroStoxx50 = stoxx
* Fuel switch (data from Spark and Dark spread graph) = fs

Selection of fuel switch values:

A chapter in the Jupyter notebook is dedicated to selecting a value from the Spark and Dark spread section in Bloomberg. The settings were configured to represent coal and gas spark and dark spreads. I selected the fuel switch. The logic was that fuel switch values integrate coal and gas prices. It should be noted that in all calculations, the carbon price is accounted for in the spark and dark spread, introducing multicollinearity and reducing the clarity of the analysis. Despite this limitation, I wanted to further explore it the relationship between fuel switch and EUA prices so I included it.

All data were sourced from Bloomberg.

**The dataframes:**

I made three dataframes,

* Original values
* Daily percentage changes
* Logged values (most commonly used for analysis)

The format of all dataframes:

A screenshot of a graph

Description automatically generated

**Chapter 4: Data exploration**

**Pariplots:**Pairplots provide an initial understanding of the correlation between variables (through scatter plots) and the distribution of each variable (via bar graphs). Here, only the first row of values from each pairplot is shown to illustrate the relationship between EUAs and other variables.

original data

A graph of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

percentage

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Logged

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**Data exploration:**Models often perform better with normally distributed data. To assess normality, bar charts (included in pairplots) and qq plots were used. The qq plots across all dataframes suggest that the data is fairly normally distributed. Below are the qq plots for the logged values.

A graph of a graph of a graph

Description automatically generated with medium confidence

**Correlation matrix**

The correlation matrix provides an initial picture of the variables' relationship with EUAs.

Original:

A screenshot of a graph

Description automatically generated

Percentage:

A screenshot of a graph

Description automatically generated

Logged:

A screenshot of a graph

Description automatically generated

Results:

* Strong correlation with gas, electricity and fuel switch. **BUT**, it should be noted, that electricity and gas are also strongly correlated.
* Negative overall correlation with stocks
* Very weak correlation with coal

**Chapter 5: Linear Models: Linear regression and VAR analysis**

**Models:**

These models provide insights into the impact of different variables. The analysis was limited to the selected variables, but could be expanded to include weather data, economic indicators, and policy changes. Addressing multicollinearity is crucial as it significantly affects the models. Seasonal decomposition of time series could also be applied for seasonal effects.

The goal was to demonstrate the model-building process, even if the models aren't predictive (a more complex task).

**Linear regression:** Linear regression examines the correlation between each independent variable and the dependent variable individually. Initially, all variables were included, but multicollinearity caused some variables (crude oil and gas) to appear insignificant due to their impact being accounted for by electricity. The model was improved by removing these variables based on their low p-value.

This is the improved version without the crude oil and gas variables. But this model is still not perfect because the Cond.No. value which represents multicollinearity is still high.

A screenshot of a computer code

Description automatically generated

**Multicollinearity: VIF**The model still had a high Cond. No. even after I have experimented with removing a lot of the variables, which is why I did this analysis to check how strong is the multicollinearity.

A white background with black text

Description automatically generated

**Interpretation of multicollinearity**

VIF = 1: No multicollinearity.

1 < VIF < 5: Moderate multicollinearity.

VIF > 5: High multicollinearity that might require correction.

VIF > 10: Very high multicollinearity, typically a sign that the variable should be removed or combined with other variables.

Due to strong multicollinearity, Principal Component Analysis (PCA) is recommended for reducing variables by combining correlated ones, though it was not conducted in this study due to time constraints.

**Impact of each variable on EUA prices:**

**Disclaimer:** High multicollinearity makes these results unreliable; the intention is to demonstrate the methodology for calculating the contribution of independent variables to a dependent variable.

Using ANOVA analysis, the contribution of each variable to EUA price fluctuations was calculated:

**Electricity prices** explain 92.22% of EUA price variations.

**Coal prices explain** 5.22%.

**Stoxx explains** 0.188%.

**Fuel Switch (fs) explains** 1.10%.

**VAR analysis:**

VAR (Vector Autoregression) models, designed for time series data, analyze correlations with a time lag. A 1-day lag was used, but no significant correlations (prob < 0.05) were found, indicating same-day correlations only.

A screenshot of a black and white document

Description automatically generatedA screenshot of a black and white document

Description automatically generated

**Chapter 6: Non-Linear models: Random forest and Gradient Boosting model**

**Advanced Models**

These machine learning models are better suited for capturing non-linear relationships and are typically used for prediction. They were explored out of curiosity despite limited familiarity.

**Results**

Random Forest: RMSE of 1.375, less than the average stand deviation of the dataframe, indicating relatively good performance.

Gradient Boosting: RMSE of 1.497, slightly higher but still acceptable within the range of the average standard deviation.