**Correlation analysis of EUAs**

Stefania Albarosa

**Chapter 1,2,3**

Note: I did this for the EUA market, but with available data such analysis could be done on any carbon market.

**The variables included:**

* 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

Notes: I have a chapter in the jupyter notebook dedicated to picking a value from the Spark and Dark spread, but I could have also used another variable, I had limited time and therefore information on the variables. The settings were set to represent coal and gas spark and dark spreads. My logic was that fuel switch values integrate the coal and gas prices. It must be noted that within all the calculations the spark and dark spread the carbon price is also accounted for which is a limitation since it causes multicollinearity and makes the analysis less clear. Regardless I added it because I thought it could be interesting, but this might be something to work around.

All data is 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:**These are some handy plots that immediately give us some idea about the correlation between all variables (scatter plots) as well as the distribution of each variable (bar graphs).

I will only copy the first row of values from each pariplot (one for each dataframe) to show the relationship between the EUAs and the 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 work better with normally distributed data, and to check weather a data is normally distributed we can use bar charts (we have those in the pairplot), but we can also use qq plots which are considered a better for this purpose. The more dots on the red line, the more normally distributed the data is.

The qq plots seem fairly similar within all datframes, indicating that the data is fairly normally distributed. I will add the qq plots of the logged values here:  
A graph of a graph of a graph

Description automatically generated with medium confidence

**Correlation matrix**This will give us an initial picture of what variables are strongly correlated to 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**, electricity and gas are also strongly correlated, we will have to watch out for this.
* Negative overall correlation with stocks
* Very weak correlation with coal

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

**Models:**

These models give us a good idea about the impact of different variables. I only used it on the variables I chose, but many other things could be done. Adding weather data, economic indicators, policy changes. I could have also spent more time on avoiding multilinearity because it really impacted the model. I could also look into applying the seasonal decomposition of time series for seasonal effects. Many things could be done, but sometimes a simple correlation matrix, especially about a market I might not know much about is already very helpful.

The idea behind doing these models, is that based on which model perform bests, which we can see from indicators (R\_squared ect..), if complex enough they could also be used for prediction. In this paper I simply would like to show what is the process of building a model. Even if it can’t be used for prediction (because that is way more complicated to do).

**Linear regression:** It will look at each independent variable and how individually they correlate to the dependent variable. I firs ran the model with all independent variables, but because of multi-correlation some variables appeared as not significant in their contribution to EUA price (crude oil and gas) because their impact was accounted by the model in the electricity variable. For now I made the choice to make the model stronger, to take these variables out based on their low p value (see jupyter notebook), but if I had more time I would work on avoiding the multicollinearity.

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.

Because there is strong multicollinearity, the next step would be to do a Principal Component Analysis, which would reduce our variables by combining the ones that correlate with each other. I didn’t have time to do that in this research, but it would be very recommended.

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

**Disclaimer:** Because of the high multicollinearity, these results are not reliable but I just wanted to show the methodology for calculating the contribution of different independent variables to a dependent variable.

I used ANOVA analysis for this, this is just to calculate the significance of each variable in determining EUA price fluctuations. In this model;

Electricity prices explain 92.22% of the variations observed in EUA prices in the model.

Coal prices explain only 5.22% of the variations in EUA prices. This indicates a weaker relationship compared to electricity.

Stoxx explain only 0.188% of the variations in EUA prices, suggesting a very weak relationship.

FS explain only 1.10% of the variations in EUA prices.

**VAR analysis:**

This model can analyse correlation with a lag in time. I did 1 day lag for the variables but seems like all the correlations are not statistically significant (prob < 0.05) for EUAs when adding a lag.

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**

These are more advanced models using machine learning, that would be used for prediction, and they work better with non-linear relationships. I am not so familiar with these types of models so these were just made out of curiosity.

**They are** better at capturing complex, non-linear relationships in the data than linear regressions.

Results:

RMSE of 1.375 (Random Forest) is less than the average standard deviation, indicating relatively good performance.

RMSE of 1.497 (Gradient Boosting) is slightly higher but still within the range of the average standard deviation, suggesting acceptable performance.