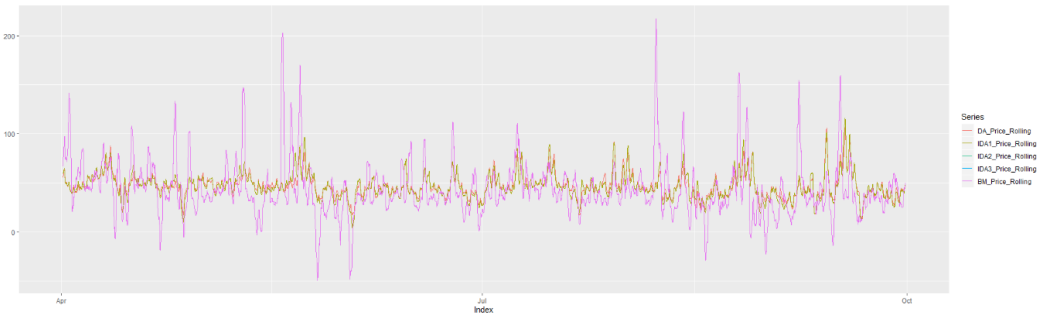

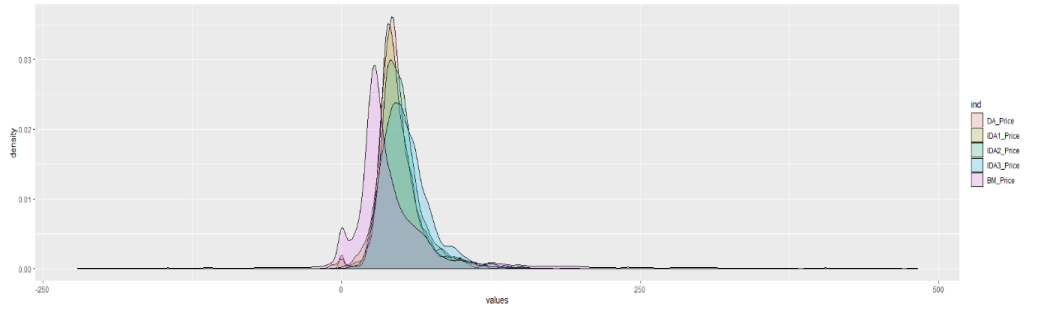

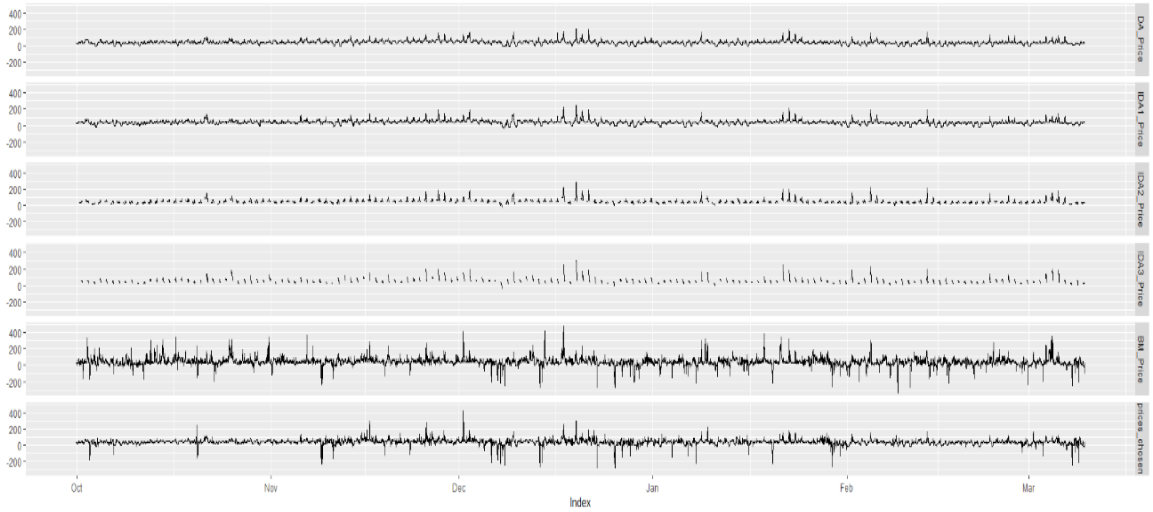

Energy Price Prediction

PREDICTIVE MODELLING AND TIME SERIES ANALYSIS

Rory Cox | 20/03/2020

Project Outline

An energy provider purchases energy at auction each half-hour of every day. Energy is most often purchased as a futures contract, where a price is agreed upon in advance of the half-hour for which the energy is provided. Energy futures can be purchased on four different markets, or purchased directly on the BM (Balancing Market) where volumes not traded in other markets (imbalance volumes) remain to be traded or settled. In chronological order of time of auction these are: DAM (Day-Ahead Market), IDA1 (Intra-Day 1), IDA2 (Intra-Day 2), IDA3 (Intra-Day 3) and BM (Balancing Market). The market auction periods run in such a way that some markets are not available for each half-hour period, prices are published shortly after auction bidding closes.

Energy Period	Markets Available
23:00 – 11:00 Exclusive	DA, IDA1, BM
11:00 – 17:00 Exclusive	DA, IDA1, IDA2, BM
17:00 – 23:00 Exclusive	IDA, IDA1, IDA2, IDA3, BM

Table 1 – Market Availability

The provider has set two tasks, each associated with a separate dataset:

1. For the period 31 March 2019 to 30 September 2019, create predictive models which attempt to optimize an energy purchase strategy. Describe the factors that drive the strategy.
2. Assuming a market rule change on 1 October 2019, which has made forecasts for demand and weather unavailable, again create predictive models which attempt to optimize a purchase strategy on the new data from 30 Sep 2019 to 09 March 2020.

Section 1:
Exploratory Data Analysis (EDA)

The purposes of EDA in relation to the task can be summarized as the following:

- To Inspect the data to confirm that it is in the expected format.
- To identify outliers that might be symptomatic of corrupted data and could negatively influence any models.
- To gain an understanding of the distributions of variables.
- To Identify relationships between variables and any trends in the data.

Format

After loading the data and joining into a data frame indexed by time period, it was firstly confirmed that data was in a tidy format suitable for model building. Market availability for each half hour period was then confirmed to correspond with Table 1.

Outliers

Boxplots of each variable were plotted to identify any outliers. The boxplot for the "Actual Demand" variable revealed that energy demand was 0 at 05:30 on 11 April 2019. This value was found to be an error in the data, and the data point was accordingly replaced with the mean demand of the two adjacent energy half-hour periods.

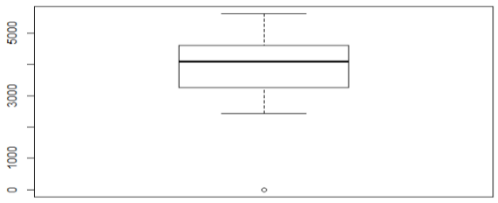

Figure 1 – Demand Boxplot

Variable Distributions

The dependent variables (i.e. the prices of each market) were the focus of EDA of the variable distributions. Boxplots were first used to see the dispersion of data. Boxplots revealed that the median prices for each market all fell between 0 and 100. The lowest median price was the BM price, followed by DA price, IDA1 price, IDA2 price and finally IDA3 price. BM price was also significantly more variable than its counterparts, and even contained negative values which corresponded with the contextual understanding of the data, as BM is the market for unsettled energy, where producers may even pay companies to buy their energy to avoid energy overload. Density plots were also useful to see the overlap between variables and understand the proportion of extreme values. Histograms faceted by market showed the counts of each market and their distributions, revealing as expected that IDA2 and IDA3 had fewer values for price as the IDA2 and IDA3 markets are only available for certain energy half-hour periods each day.

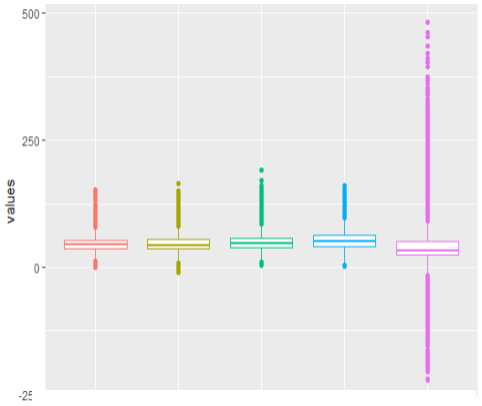

Figure 2 – Market Boxplots

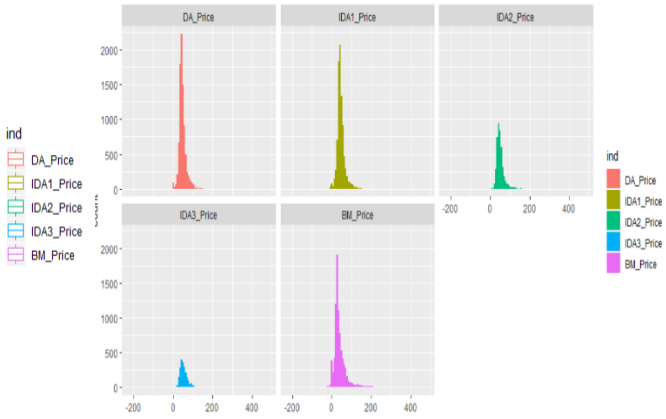

Figure 3 – Histograms faceted by Market

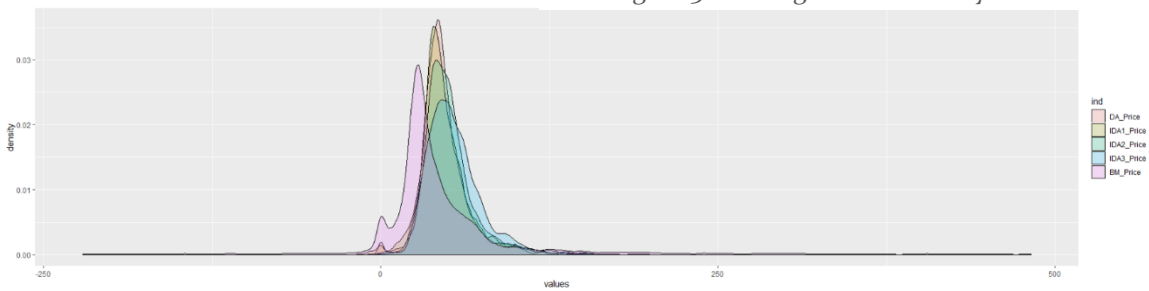

Figure 4 – Market Prices Density Plots

Variables Relationships and Trends

Creating an xts object allowed for plotting of the data as a time series, and creating variables for rolling averages of 30 half-hour periods made trends much easier to identify compared to plotting the raw prices.

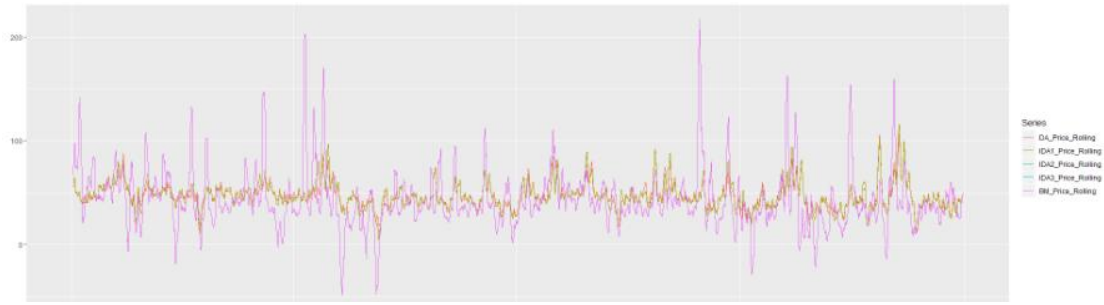

Figure 5 – Market prices over period

The prices tended to follow each other approximately, so one might expect prices to have been influenced by the same variables. However, BM prices particularly seemed to be much more volatile.

The pair plot of wind, demand, time and mean price across all markets revealed relationships that were correlated but clearly non-linear. Demand was unsurprisingly positively correlated with price, while wind was negatively correlated with price, as wind is essentially free energy. Time was negatively correlated with price; it appeared energy became non-linearly less expensive over the period. The dataset provided many forecast variables for demand and wind, each provided at different times in the auction cycle. The pair plot for the forecast variables was nearly identical to that of the actual variables, with negligibly different correlations. This indicated that the forecasts were good predictors for the actual values for demand and wind.

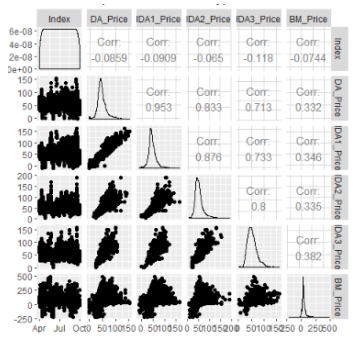

Figure 6 – All Prices

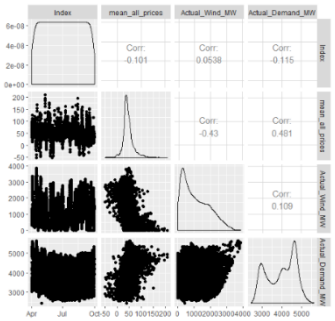

Figure 7 - Price, Time and Actual Demand and Wind

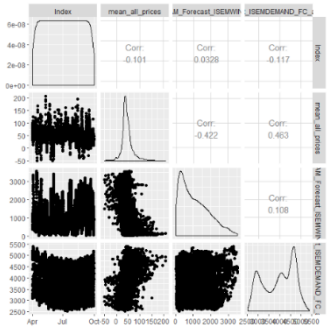

Figure 8 - Price, Time and Forecast Demand and Wind

Modelling Prices with Generalised Additive Models (GAMs)

Generalised Additive Models (GAMs) were used to model prices, which capture non-linear relationships by fitting non-linear smooth functions. The main advantages of GAMs in this case were their ability to capture nonlinear relationships, their ability to use large numbers of explanatory variables, and their ability to distinguish explanatory variables which are significant from those which are not. While it would have been a simple exercise to fit a GAM for each market based on the variables provided, predict prices and choose the market for each half-hour period with the lowest predicted price, there are 2 improvements that were made to increase the success of the pricing strategy.

Feature Engineering

The first improvement was to add further significant features to the GAMs. In addition to the forecasts for wind and demand, and the time variable, the effect of sentiment and speculation could be captured in the GAMs by including lag variables for the prices. The lagged variables engineered were:

1. The mean of all prices 24 hours prior to the energy half-hour period.
2. The rolling mean of all prices over week prior to the energy half-hour period (lagged by 24 hours so as to only use available prices, so from 8 days prior to 1 day prior).

Fitting a GAM based on the variables provided alone resulted in a fairly poor fit, with an R-Squared value of 62.2%, but this rose to 67.3% after including the new lagged variables, and all variables were statistically significant. The fit was still suboptimal, with non-normally distributed residuals, but there was a marked improvement.

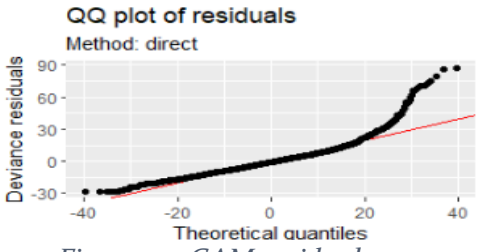

Figure 9 – GAM residuals- no lagged variables

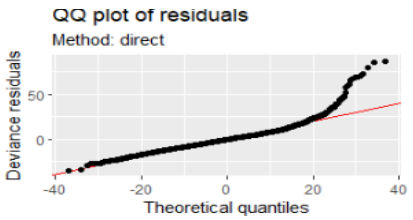

Figure 10 – GAM residuals– incl. lagged variables

Price Interdependency – Iterative Approach

The second improvement was to recognise that energy prices were strongly interdependent; energy prices revealed in the auction cycle significantly impacted future prices. This interaction could be factored into the strategy by adopting an iterative modelling approach whereby models were recalibrated using previous prices as soon as they became available. This iterative method is described as follows:

Stage 1: No prices information is available.
All applicable market prices are predicted using the DAM forecasts and lagged variables.
If the DAM is predicted to be the cheapest price, the actual DAM price is bought. Otherwise move to stage 2.
Stage 2: DAM price is available.
All applicable market prices are predicted using models that now include the DAM price and most recent forecasts.
If the IDA1 is predicted to be the cheapest price, the IDA1 price is bought. Otherwise move to stage 3.
Stage 3: DAM price and IDA1 price is available…
… (process continues in this way until the final stage) …
Stage 5: BM price remains as no previous markets have been predicted to have the lowest price at any stage.

Section 1 Results

Implementing a function to prepare the data with the relevant variables, splitting the data into a train and test set and building a total of 15 GAMs to account for each stage of the price selection process, and finally using a further function to implement the pricing strategy, yielded the following results.

```
"DA Price Mean: 47.039"
"IDA1 Price Mean: 46.663"
"IDA2 Price Mean: 50.861"
"IDA3 Price Mean: 54.082"
"BM Price Mean: 42.788"
"Chosen Prices Mean: 40.336"
```

Table 2 - Modelling Results

The average price per half hour after implementing the strategy was 40.336, which was lower than what would be achieved from consistently selecting one market. The chosen prices can be seen plotted over all the testing data in figure 12.

The distribution of the mean value of chosen prices over 100 randomly selected half-hour periods could be approximated using bootstrap aggregation. This is useful if the provider also placed value in the variance of the mean price, and would like to estimate the probability of spending outside predetermined limits over a set number of half-hour periods. The distribution was approximately normal, with a mean of 40.71, a standard deviation of 3.22, and was not significantly skewed.

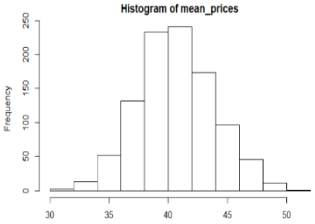

Figure 11 - Bootstrap Resampling - Mean Distribution

The results show that the strategy was successful in predicting market prices with enough accuracy to achieve a relatively low mean price for each half-hour period.

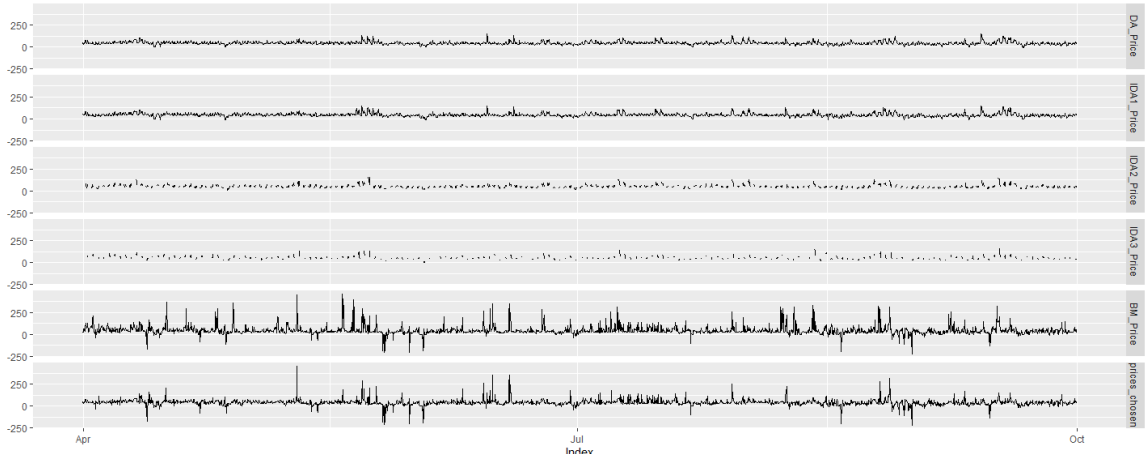

Figure 12 – All market prices and chosen prices (chosen prices in final row)

Section 2:

In the second part of the task, the forecast variables were no longer provided, and a strategy to choose the lowest possible energy prices for each half-hour was still required. The variance explained by the forecast variables therefore needed to somehow be otherwise explained. One option was to create a forecast variable for the difference between Demand and Wind (Net Demand) using an ARIMA (Autoregressive Integrated Moving Average) model. It will be demonstrated that an ARIMA model was not suitable for modelling the Net Demand. Instead, the addition of lagged variables for the Net Demand was used, just as price was in section 1. These were incorporated into GAMs to predict prices using the same iterative method.

Attempt to Fit Arima Model

Autoregressive Moving Average (ARMA) models attempt to fit predict future values of a variable by using its previous values, taking advantage of any repetitive trends in a time series. ARIMA models are a generalisation of a standard ARMA model, which include an initial differencing step which transforms a non-stationary time series into a stationary time series. An ARIMA model would be expected to perform well in predicting the Net Demand variable if there are predictable trends in Net Demand over time.

After creating a variable for the Net Demand, an ACF (Auto-Correlation Function) plot reveals that the Net Dmeand variable becomes stationary after the second difference. The size of the ACF values shows the level of correlation between a value and the nth value preceding it (after differencing). By the 7th lag value, there were no significant coefficients. After overfitting, an MA value of 3 was determined to produce the lowest AIC, and therefore the final model was an ARIMA model of order 3, 2, 7.

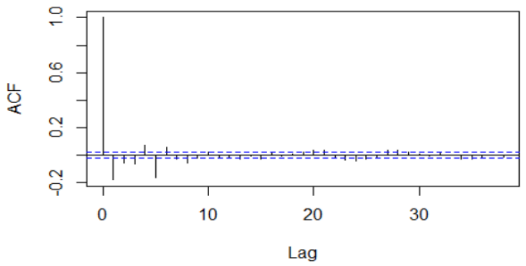

Figure 13: ACF Plot

However, after splitting the data into train and test such that test data chronologically followed the training data in order to test the accuracy of future Net Demand on the testing data, the accuracy was far too low for the variable to be used. The RMSE between predicted and actual values was 1648.

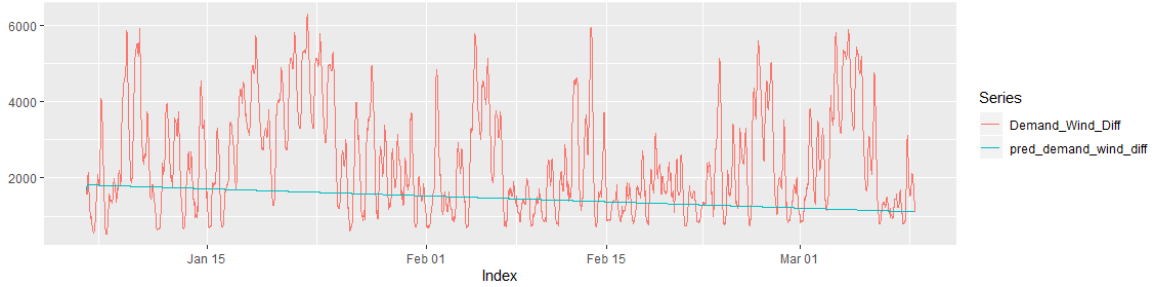

Figure 14 – Actual vs Predicted values for Net Demand

Alternative to ARIMA: Feature Engineering

As ARIMA did not provide an adequate replacement for the demand-wind difference variable, the same approach as in section 1 was used, which was to engineer lagged features. The features used were:

1. The Net Demand 24 hours prior to the energy half-hour period.
2. The rolling mean Net Demand over week prior to the energy half-hour period (lagged by 24 hours so as to only use available information, so from 8 days prior to 1 day prior).

Part 2 Results

Using the same iterative approach as in section 1, the results on the testing data were as follows. The mean price was 38.447 using the proposed strategy which was again lower than consistently choosing any 1 market.

Bootstrapping was also used to approximate the distribution of the mean price over 100 random half-hour periods, which was approximately normal with a mean of 40.33, and a standard deviation of 3.2 and no significant skew. The chosen prices can be seen plotted over the testing data in figure 16.

```
"DA Price mean: 42.893"
"IDA1 Price mean: 42.405"
"IDA2 Price mean: 50.601"
"IDA3 Price mean: 56.999"
"BM Price mean: 42.132"
"Chosen Price mean: 38.447"
```

Table 3 – Modelling Results

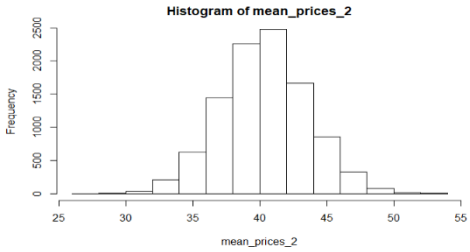

Figure 15 - Bootstrap Resampling - Mean Distribution

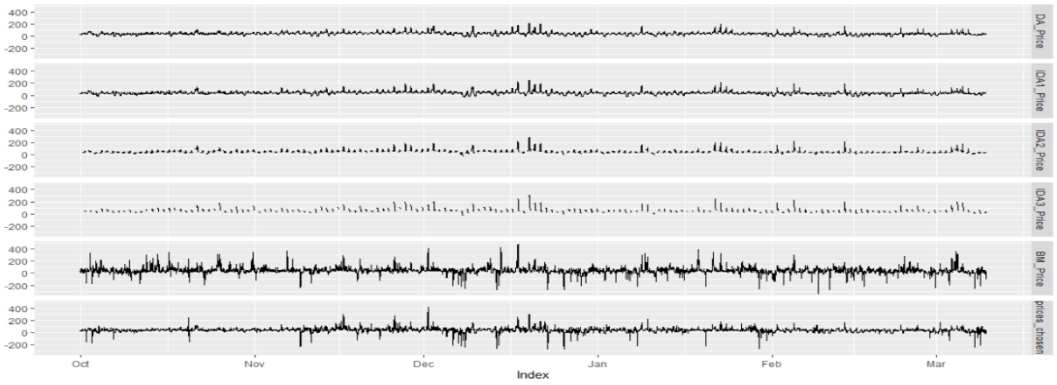

Figure 16 – All market prices and chosen prices (chosen prices in final row)

Conclusions

In both tasks, a successful strategy was implemented to choose low average energy prices. In each task, non-linear relationships between variables demanded the use of non-linear predictive models which included engineered lagged variables, and since prices were also significantly interdependent, an iterative approach to modelling proved successful. In the second task, an ARIMA model was found to be an unsuitable replacement for the removed forecast variables, though the introduction of lagged Net Demand variables contributed enough to the models to achieve a strategy that was successful in selecting low average energy prices.