Year-ahead forcast of electricity prices in Spain

Stephan Kirchhoff

15 May 2020

Introduction

The goal of this work is to train an algorithm that forecasts the electricity prices in Spain based on historic spot market price developments. The algorithm is trained on the electricity prices from 2014 to 2017 and tested using the 2018 prices.

The dataset is a daily time series of electricity demand, generation and prices in Spain from 2014 to 2018. It is gathered from ESIOS, a website managed by REE (Red Electrica Española) which is the Spanish TSO (Transmission System Operator). It was made available on kaggle: https://www.kaggle.com/manualrg/spanish-electricity-market-demand-gen-price/

Original energy and price data can be downloaded from : https://www.esios.ree.es/en

The idea is to forecast electricity prices for the next year, using linear regression, trend decomposition and Holt Winters seasonal method. Using these methods the electrcitiy prices could be predicted with a RMSE of below 9 EUR/MWh. The average electrcitiy price is at approximately 46 EUR/MWh.

The dataset also contains generation data from all relevant energy sources,like wind, coal, nuclear or hydro, which could serve as predictor variables. Graphical exploration and correlation analyses show that especially the wind energy generation has a strong negative correlation to the electricity price, i.e. prices are significantly lower when a lot of wind energy is generated.

Having a wind speed forecast could significantly reduce the error in the electricity price forecast. In this work the energy generation data was not considered, because no forecast was available on the considered time scale. Still, patterns in the energy generation, like wind seasons, should already be reflected in the seasonal price effects.

Methods/Analysis

Data Cleaning and Preparation

The dataset contains six variables, as seen in the following table. Data is reported in a daily frequency, in this case the Spanish electricity spot market price in Spain.

datetime	id	name	geoid	geoname	value
2014-01-01 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	25.280833
2014-01-02 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	39.924167
2014-01-03 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	4.992083
2014-01-04 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	4.091667
2014-01-05 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	13.587500
2014-01-06 23:00:00	600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	47.885417

The structure of the data can be better understood when filtering for the first reporting date. For each date various parameters are reported in a tidy format, specified in the column *name*. The Spanish spot market price is reported together with the Portuguese and French electricity spot market prices (in EUR/MWh). It also contains the expected energy demand and the energy generation plan in Spain, split by energy source (in MWh). Finally it contains data on the trading between markets, i.e. the energy that is assigned to Spain and France on the following day, and the energy export and import from and to France and Portugal (in MWh). The value for each is reported in the column *value*.

It can be seen that each parameter is assigned to an unique *id*. Only the spot market prices for Spain, France and Portgual use the same ID 600. These are differentiated via the variables *geoid* and *geoname*, which are directly linked. Both parameters are only defined for the spot market prices, to differentiate between markets.

Finally it can be seen that the spot market prices are dublicated. Once with the *name* parameter specified and once without (original data).

id	name	geoid	geoname	value
600	Precio mercado SPOT Diario ESP	3	EspaÃ±a	25.28
600	Precio mercado SPOT Diario FRA	2	Francia	28.71
600	Precio mercado SPOT Diario POR	1	Portugal	25.04
602	EnergÃa asignada en Mercado SPOT Diario EspaÃ±a	NA		566081.90
1334	EnergÃa asignada en Mercado SPOT Diario Francia	NA		171917.30
600		1	Portugal	25.04
600		2	Francia	28.71
600		3	EspaÃ±a	25.28
1119	Rentas de congestiÃ³n mecanismos implÃcitos diario Portugal importaciÃ³n	NA		10811.00
1293	Demanda real	NA		28191.60
10141	Demanda programada PBF total	NA		620107.70
10258	GeneraciÃ³n programada PBF total	NA		642771.80
10073	GeneraciÃ³n programada PBF EÃ³lica	NA		277443.90
9	GeneraciÃ³n programada PBF Ciclo combinado	NA		4497.50
10167	GeneraciÃ³n programada PBF CarbÃ³n	NA		2498.70
4	GeneraciÃ³n programada PBF Nuclear	NA		144654.60
17	GeneraciÃ³n programada PBF Gas Natural CogeneraciÃ³n	NA		75993.10
10064	GeneraciÃ³n programada PBF UGH + no UGH	NA		87564.80
14	GeneraciÃ³n programada PBF Solar fotovoltaica	NA		7027.30
3	GeneraciÃ³n programada PBF TurbinaciÃ³n bombeo	NA		15226.20

The dataset was then prepared for further processing. The *datetime* variable was changed to a datetime format (lubridate). Additional variables were added, indicating the *year*, *month*, *week*, weekday (*wday*) and *hour*. Then dublicates for the spot market prices were eliminated by removing empty rows in the column *name*. The variables in the column *name* were then renamed, using English terms (*key*). Finally the parameters in the column *key* were spread to columns to have one observation per day. Like this, correlations between spot market prices and generation and demand data could be analyzed.

```
### Spread table (to link spot prices to other variables)
energy_data_spread <- energy_data %>% spread(key=key,value=value)
```

This is an overview of the variables (*name*) and their English version (*key*)

name	id	key
Demanda programada PBF total	10141	total_demand_plan
Demanda real	1293	total_demand_actual

name	id	key
EnergÃa asignada en Mercado SPOT Diario EspaÃ±a	602	trade_ESP
EnergÃa asignada en Mercado SPOT Diario Francia	1334	trade_FRA
GeneraciÃ³n programada PBF CarbÃ³n	10167	plan_coal
GeneraciÃ³n programada PBF Ciclo combinado	9	plan_combicycle
GeneraciÃ³n programada PBF EÃ³lica	10073	plan_wind
GeneraciÃ³n programada PBF Gas Natural CogeneraciÃ³n	17	plan_gas
GeneraciÃ³n programada PBF Nuclear	4	plan_nuclear
GeneraciÃ³n programada PBF Solar fotovoltaica	14	plan_pv
GeneraciÃ³n programada PBF total	10258	total_generation_plan
GeneraciÃ³n programada PBF TurbinaciÃ³n bombeo	3	plan_hydro
GeneraciÃ³n programada PBF UGH + no UGH	10064	plan_reverse_hydro
Precio mercado SPOT Diario ESP	600	spot_price_ESP
Precio mercado SPOT Diario FRA	600	spot_price_FRA
Precio mercado SPOT Diario POR	600	spot_price_POR
Rentas de congestiÃ³n mecanismos implÃcitos diario Francia exportaciÃ³n	1118	export_FRA
Rentas de congestiÃ³n mecanismos implÃcitos diario Francia importaciÃ³n	1117	import_FRA
Rentas de congestiÃ³n mecanismos implÃcitos diario Portugal exportaciÃ³n	1120	export_POR
Rentas de congestiÃ³n mecanismos implÃcitos diario Portugal importaciÃ³n	1119	import_POR

The dataset is then split into a train set (*model_data*), consisting of the years 2014-2017, and the test set (*test_data*) containing year 2018 data. The goal is to train a model on the initial 4 years and test it on the last one. Any further data exploration and model training will solely use the train set.

```
### Use year 2018 as test model
model_data <- energy_data_spread %>% filter(year!=2018)
test_data <- energy_data_spread %>% filter(year==2018)
```

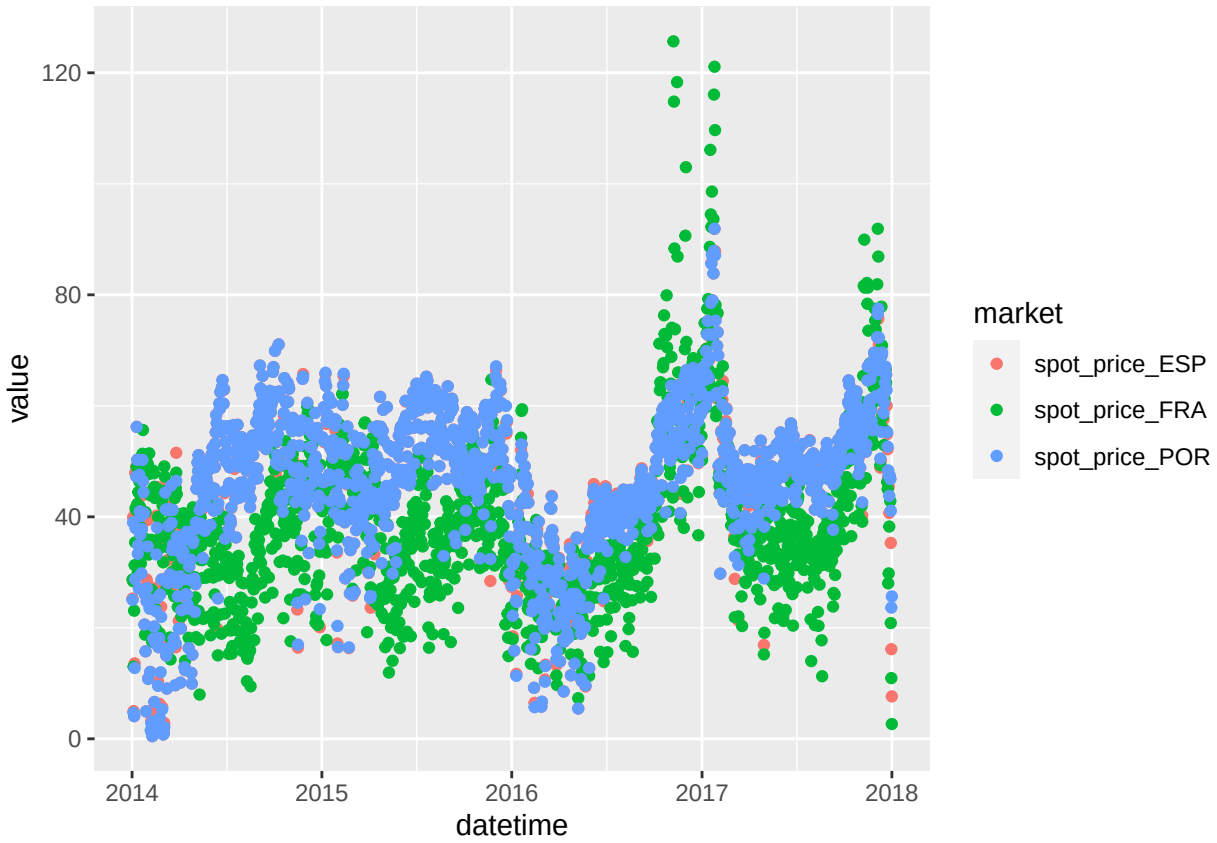
Data Analysis

Since the objective is to predict the Spanish spot market prices, the average price for years 2014-2017 was calculated in a first step.

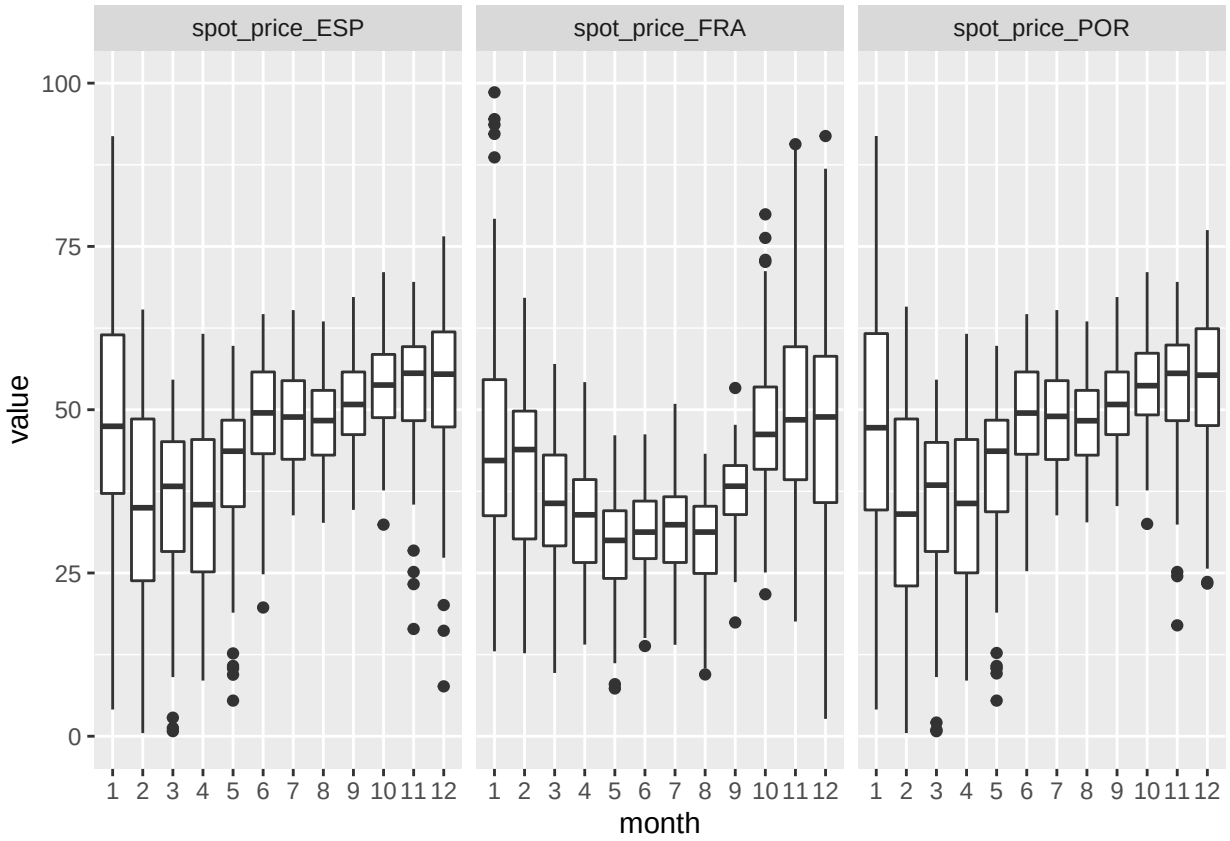
mean_price_ESP	mean_price_FRA	mean_price_POR
46.08536	38.69469	46.05715

Correlation to Months and Weekdays

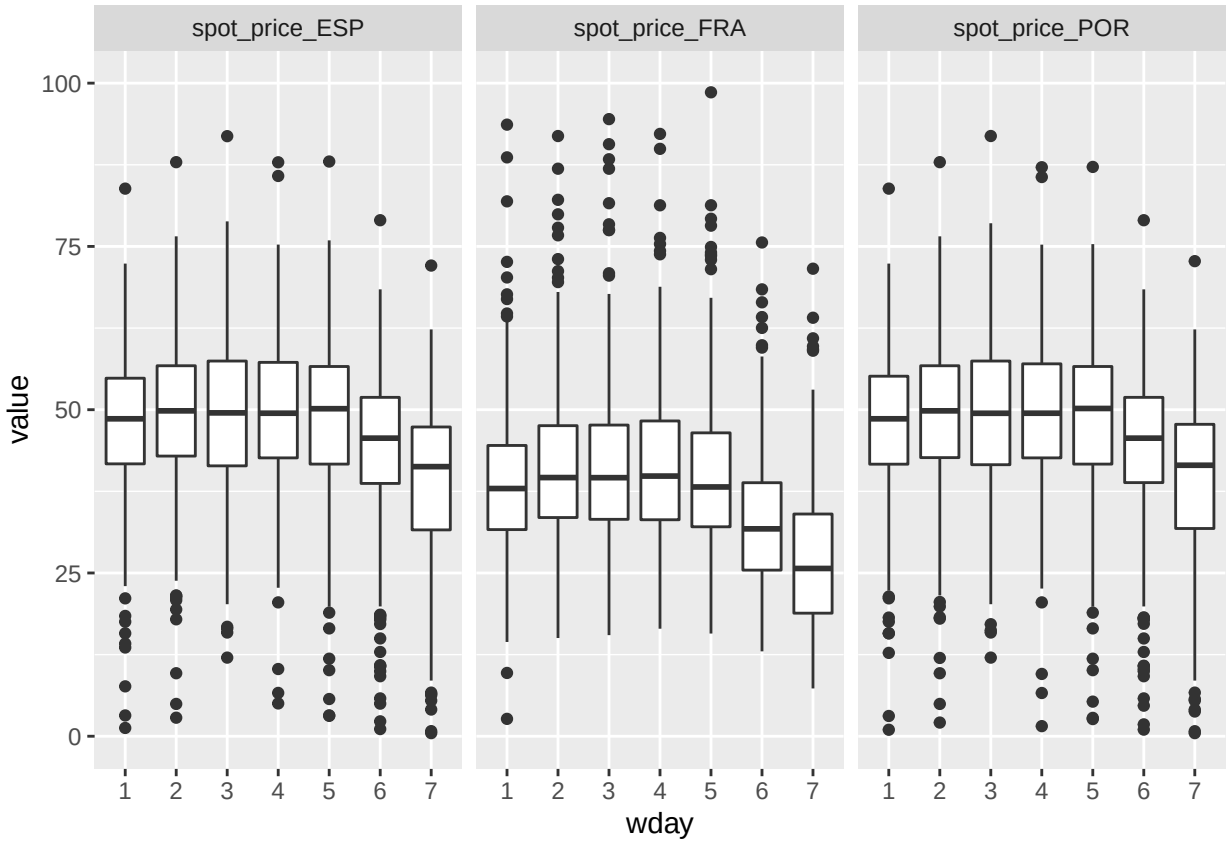
Looking at the graphical representation of the price development in years 2014-2017 you can see a seasonal pattern, with a positive trend. Also you can see that Spain and Portugal seem to have closely connected electricity markets.



To understand the seasonality better, possible patterns by month or by weekday were checked. The two box plots below show electricity prices from 2014 to 2017 by month and weekday. The price ranges are shown for Spain, Portugal and France to understand similarities and differences between markets.



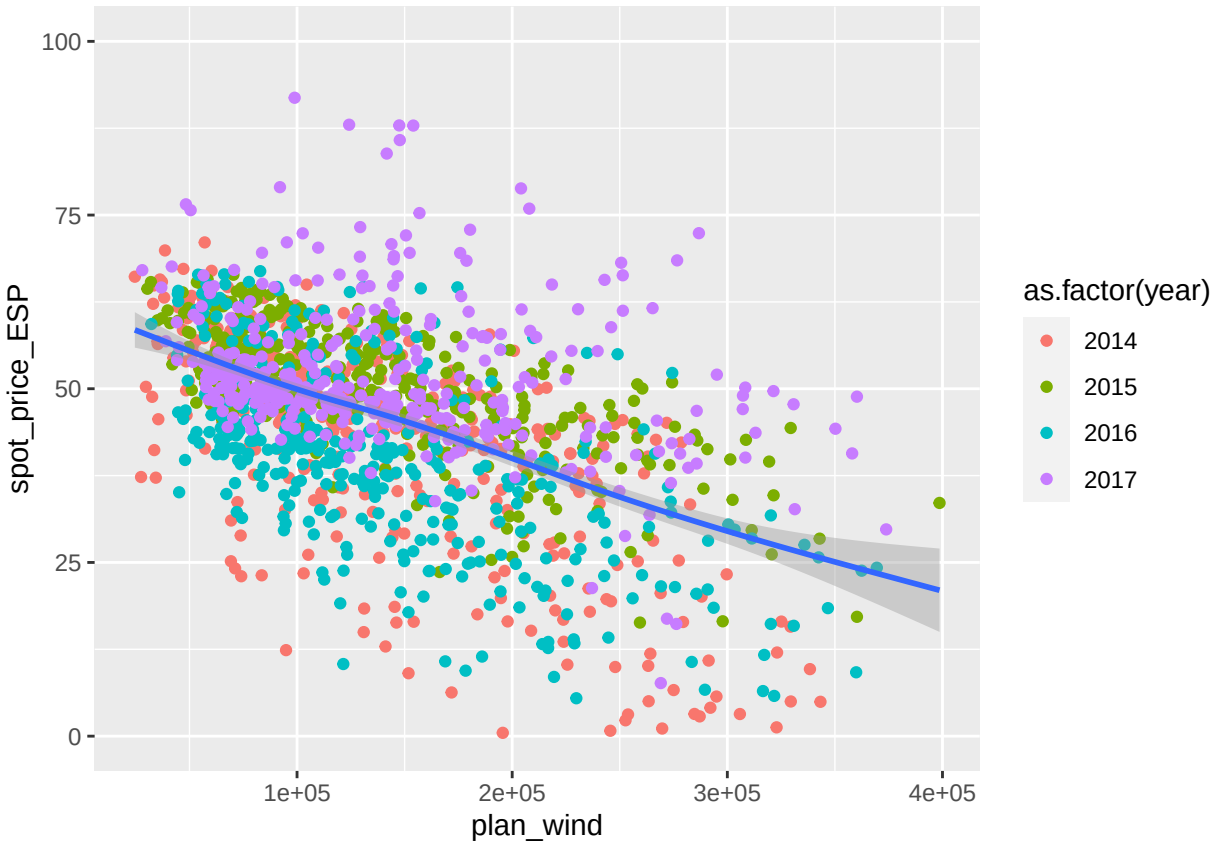
It can be seen that electricity prices in Spain and Portugal are significantly lower in spring, from February to May. In France the lowest prices can be seen in spring and summer, from March to September.



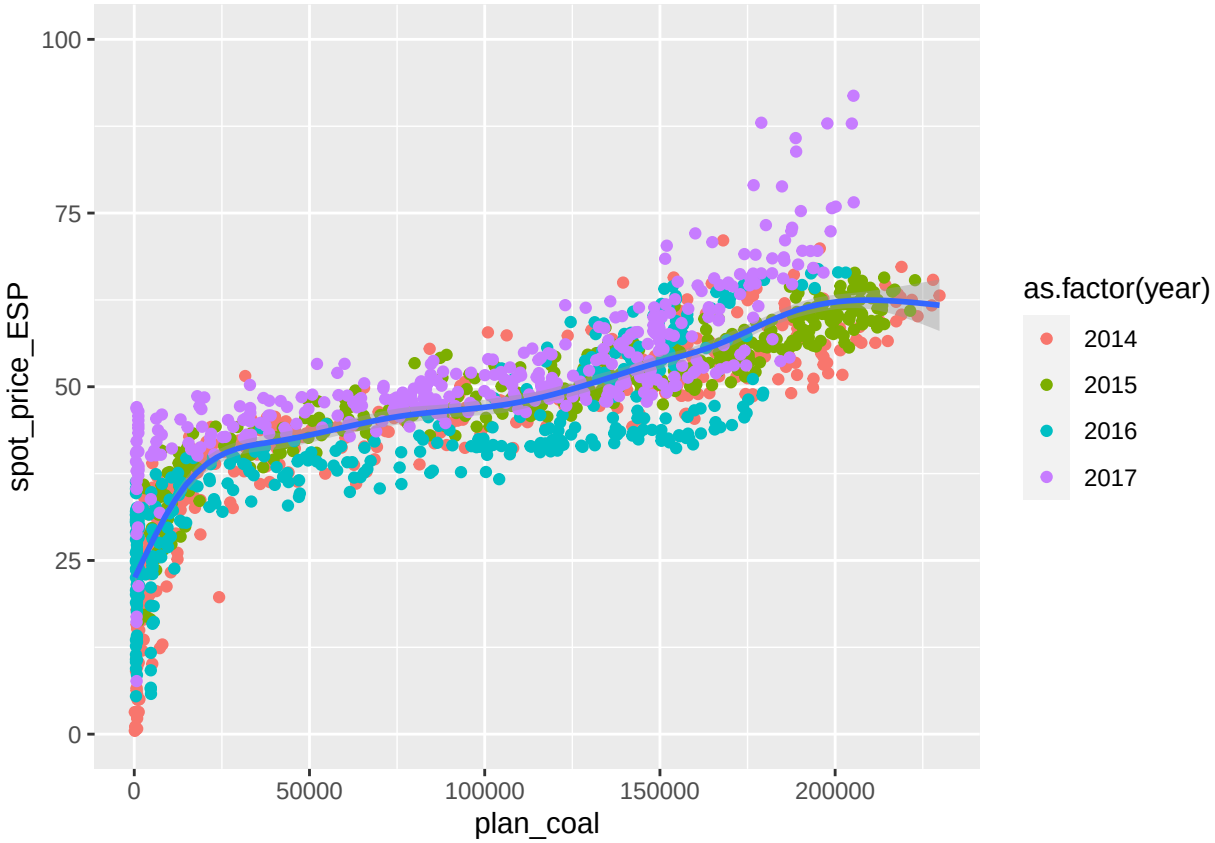
Looking at the price differences by weekday it can be seen that prices are relatively constant during the week but drop on Saturday and even more on Sunday. This trend seems to be similar for all three markets.

Correlation to Energy Sources

Not surpringly the electricity price also shows a strong correlation to the volume that different energy sources are contributing to the energy mix. The strongest correlation can be seen to coal and wind.



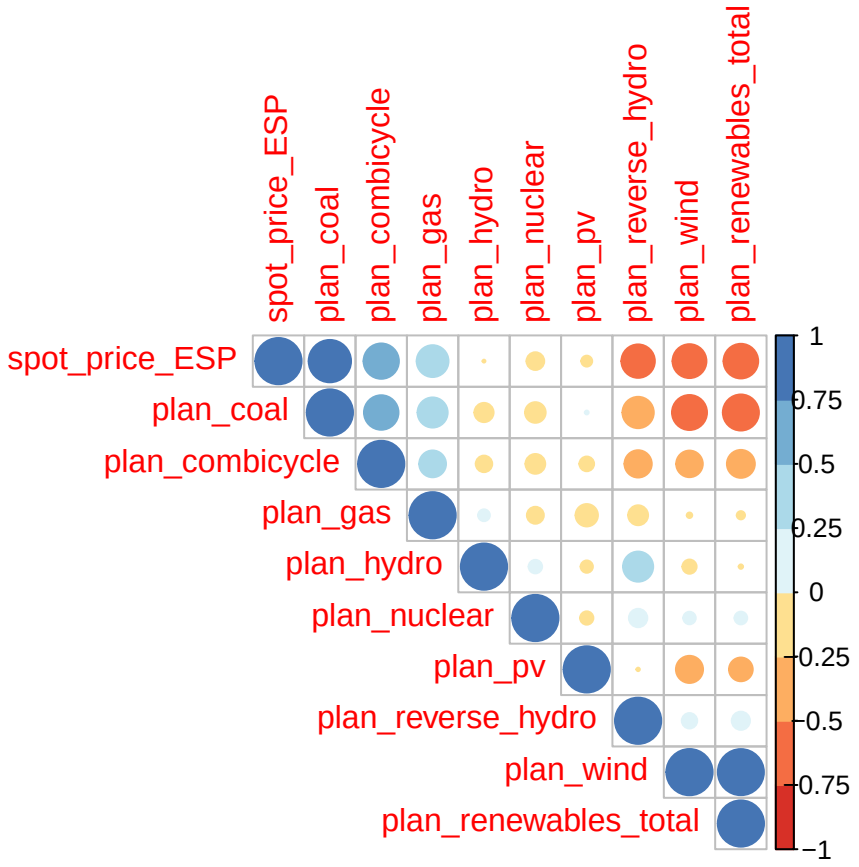
Wind energy has a negative correlation to the electricity price. The more wind energy is available, the lower the electricity price will be.



Coal energy has a positive correlation to the electricity price. The more coal energy is produced the higher the electricity price will be. It can be assumed that coal energy is produced whenever less cheap wind energy is available, which will be offered at very low prices if needed, since to fuel is needed.

This assumption can be confirmed by looking at the correlation values between the electricity spot price and the different energy sources. As expected, there is a strong positive correlation between the spot price and the coal energy production (*plan_coal*), and a negative correlation to wind energy (*plan_wind*) and reverse hydro energy (*plan_reverse_hydro*), i.e. energy from storage. Also there is a negative correlation between coal on one hand and wind and reverse hydro on the other. Whenever a lot of wind energy and reverse hydro energy is available, little coal energy is planned.

```
### Calculate correlation betwwen energy sources and to spot prices
model_data <- model_data %>% mutate(plan_renewables_total=plan_hydro+plan_pv+plan_wind)

impact_generation <- model_data %>% select(spot_price_ESP,
                                           plan_coal,plan_combicycle,
                                           plan_gas,plan_hydro,
                                           plan_nuclear,plan_pv,
                                           plan_reverse_hydro,plan_wind,
                                           plan_renewables_total) %>%
  cor(use="complete.obs")

corrplot(impact_generation, type="upper", order="original",
         col=brewer.pal(n=8, name="RdYlBu"))
```



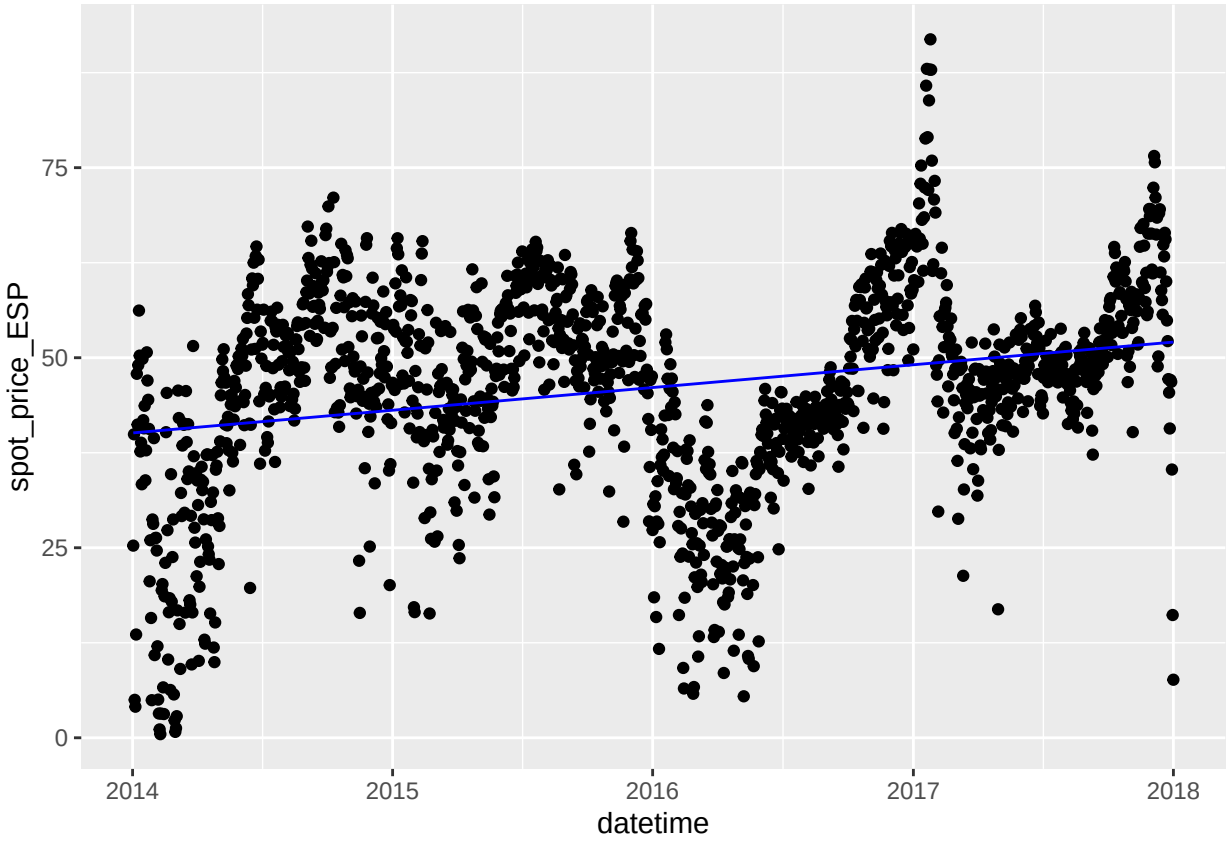
Training of Prediction Model

Considering our initial findings a prediction algorithm is trained on the train set.

Regression Line

In a first step a simple regression model is trained on the train set. Representing a the regression line together with the electricity prices in a scatter plot, shows that a general positive trend is reflected but seasonal effects lead to strong errors in the prediction.

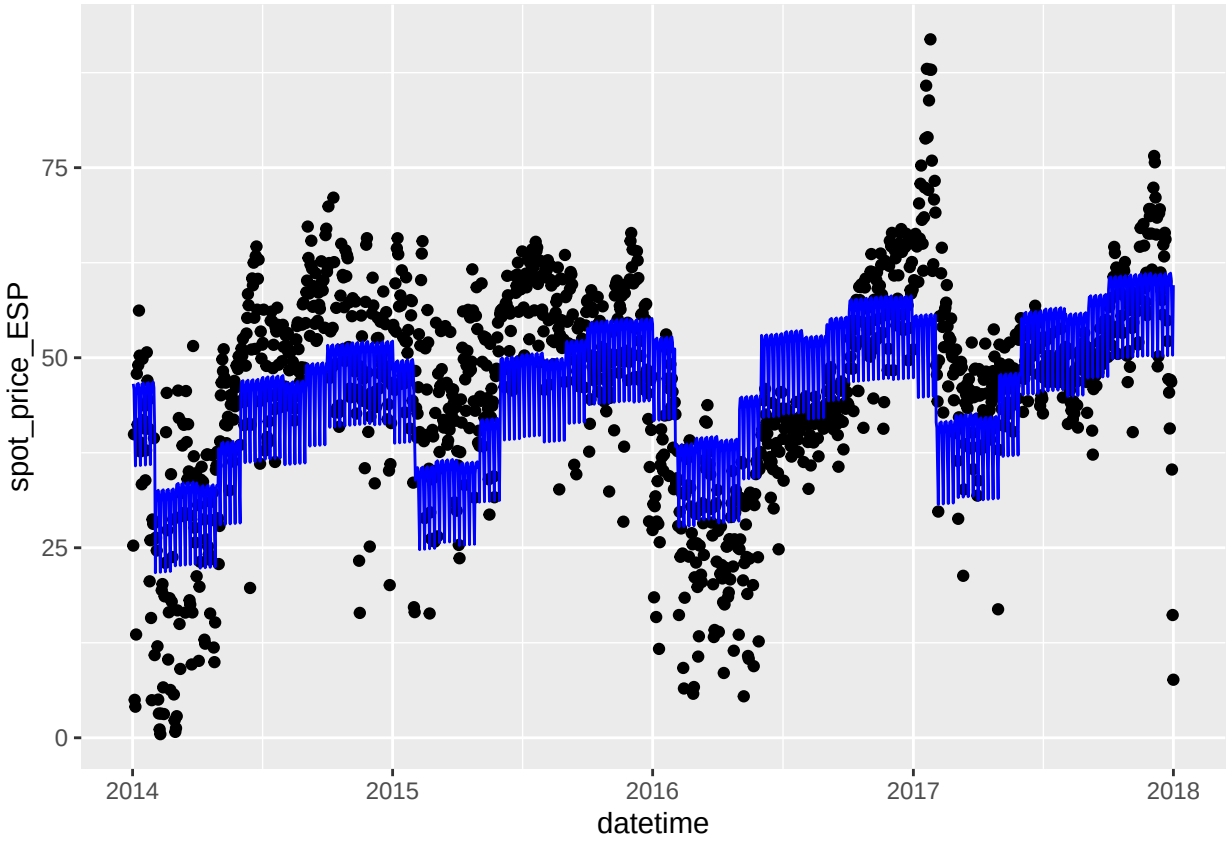
```
### Calculate regression line through spot prices
regline <- model_data %>% lm(spot_price_ESP~datetime, data=.)

model_data <- model_data %>% mutate(price_regline=predict(regline,newdata=.))
```



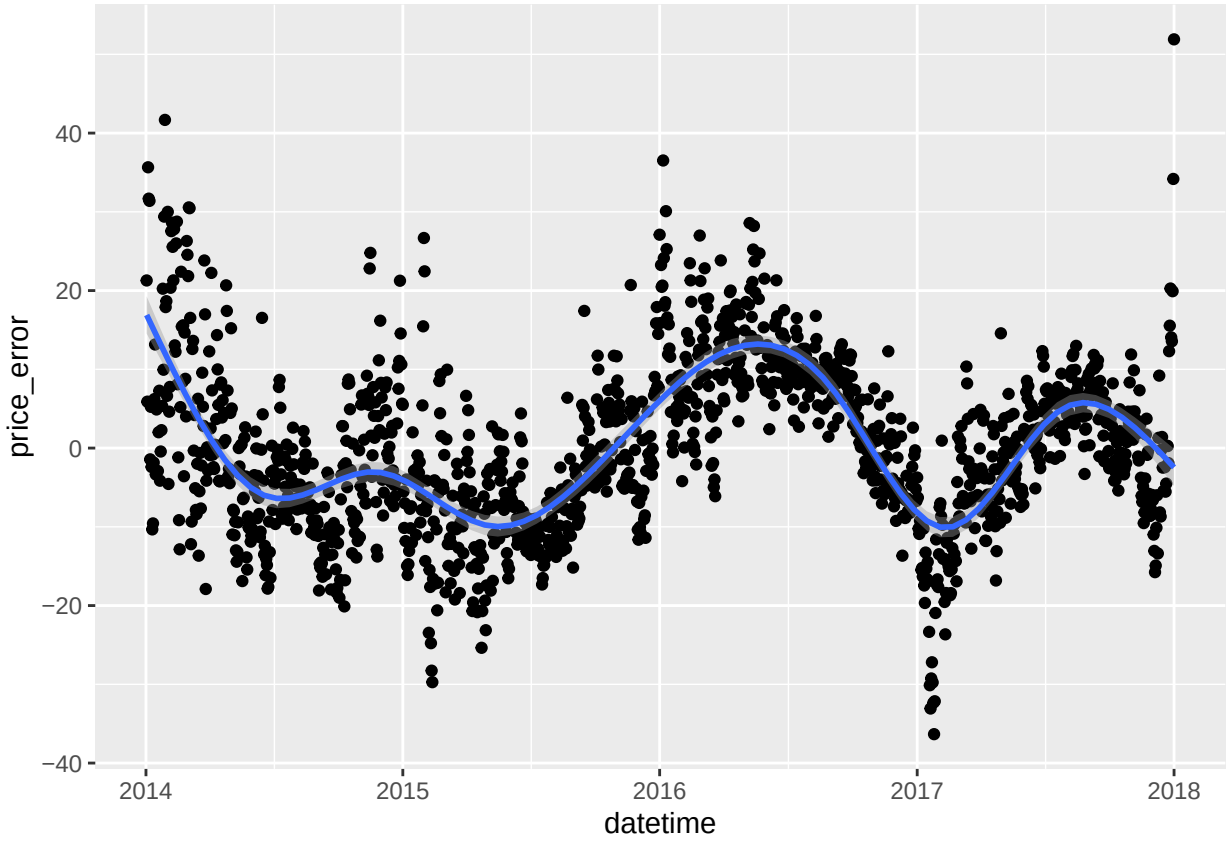
Trend Decomposition

In a second step a seasonal pattern is added to the regression line. The average error between the regression line and the actual electricity price is calculated by month (*bias_m*). This monthly "bias" is added to the prediction. The same was done for calendar weeks (*bias_w*) and weekdays (*bias_d*). The graph above shows the predicted prices for the regression line in combination with the monthly bias and the bias by weekday. It follows the actual electricity prices already much closer than the regression line alone.

```
### Calculate monthly bias (delta to regression line)
bias_month <- model_data %>% group_by(month) %>%
  summarise(bias_m=mean(spot_price_ESP-price_regline,na.rm=TRUE))

model_data <- model_data %>% left_join(bias_month, by="month") %>%
  mutate(price_month=price_regline+bias_m)
```



To understand the remaining error below graph shows the error for the years 2014-2017. There seem to be some general trends that are not reflected yet. Since the upward and downward trends have different durations, it might be difficult to include these in a seasonal pattern, though.

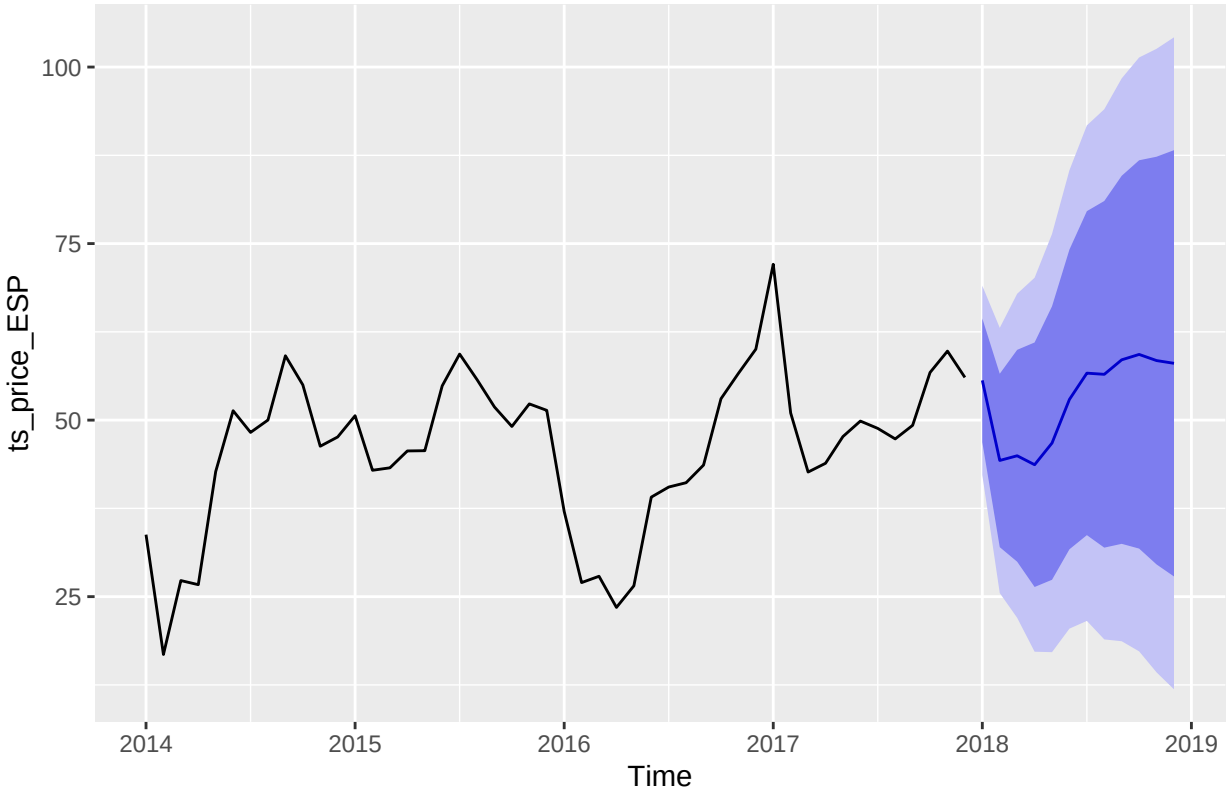


Holt Winters Method

A second method was applied that automatically identifies patterns, to check if a higher prediction accuracy could be achieved. The Holt Winters method decomposes the data into the general level, a (linear) trend and a seasonal pattern. It uses exponential smoothing, thus putting higher weighing on recent data. To use Holt Winters' method the electricity prices needed to be converted into a time series object first, with a frequency of 12 (monthly). The calculation is calculating the weighted average and already generates the prediction for 12 months ahead, i.e. the year 2018.

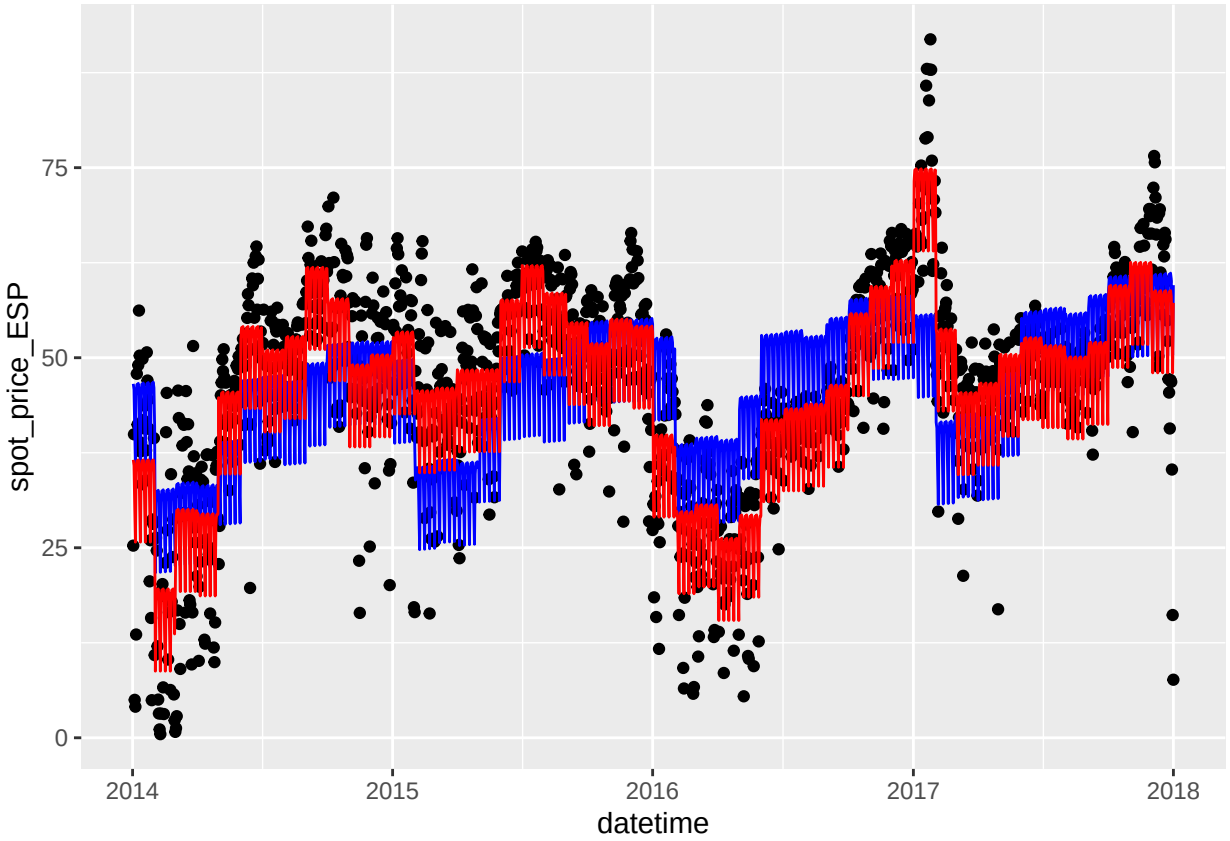
Following graph shows the smoothing curve through years 2014 to 2017 and the prediction for 2018. The blue area shows the confidence intervall for the prediction.

```
# Calculate average spot price by month
spot_price_ESP <- model_data %>% group_by(year,month) %>%
  summarize(spot_price_ESP=mean(spot_price_ESP))

# Convert spot prices into time series
ts_price_ESP <- ts(spot_price_ESP[,3], frequency=12, start=c(2014,1))

# Calculate Holt Winters model, including forecast for 2018
ts_price_hw <- hw(ts_price_ESP, h=12, seasonal="additive")
```



The time series is then converted back into a data frame. To improve the accuracy the bias by weekday (*bias_d*) is added to it. The Holt Winters method can only consider a maximum frequency of 24.

Following graph compares the regression line combined with monthly and daily bias (blue) to the Holt Winters smoothing curve combined with daily bias (red). It can be seen that the smoothing curve follows the actual electricity prices closer in the train set. The question is which method will predict the prices better in the test set.



Results

The following table shows the RMSE values both in the train set and the test set.

method	accuracy_train	accuracy_test
naive	13.490757	15.623775
regline	13.042198	10.945090
regline+monthly bias	11.259274	8.985962
regline+bias by calendar week	10.890647	9.339583
regline+monthly and daily bias	10.595086	8.851033
hw by month	7.857276	9.193364
hw by month+daily bias	6.901872	9.065285

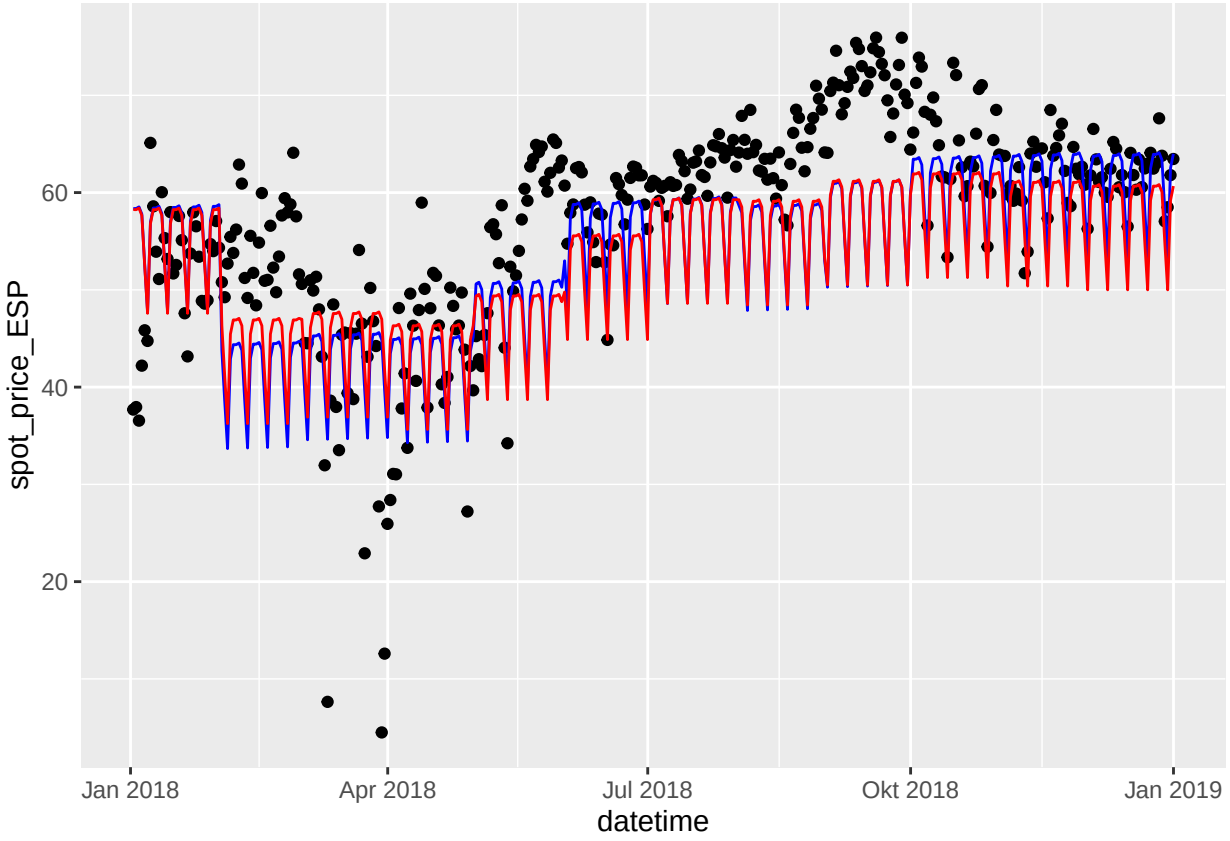
Using the naive method for prediction, i.e. the average electricity price, delivers a RMSE of 13.49 in the train set and a RMSE of 15.62 in the test set. Based on a simple linear regression line the prediction can be slightly improved in the train set, and provides an even better result in the test set, down to 10.95.

Adding the monthly and weekly bias improves the prediction both in the train and the test set. Using the monthly bias provides a slightly higher RMSE in the train set (11.26) than the weekly bias (10.89). In the test set the monthly bias provides a slightly better accuracy (8.99). Adding the daily bias further improves the RMSE down to (10.6) in the train set and (8.85) in the test set. This is the lowest RMSE of the algorithms that were considered.

In comparison the Holt Winters method predicts the electricity with high accuracy in the train set (RMSE

of 7.86), but shows a higher RMSE of 9.19 in the test set. This is slightly higher than the simple linear regression line in combination with the monthly bias. Adding the daily bias brings down the RMSE down to 9.07 in the test set.

Following graph shows the predicted prices based on a regression line combined with monthly and daily bias (blue) and the Holt Winters prediction with daily bias (red) in the scatter plot of 2018 electricity prices.



Conclusion

Based on a time series of electricity prices in Spain from 2014 to 2018 an algorithm was trained to predict electricity prices for one year ahead. The electricity prices from 2014 to 2017 were used to train the algorithm, and the data from 2018 was used as test set. By identifying seasonal effects by month, calendar week and weekday, an effective algorithm could be trained. The accuracy of the algorithm was measured using the RMSE.

The naive method delivered a RMSE of 15.62 in the test set. Combining a linear regression line with the average bias by month and the bias by weekday could reduce the RMSE to 8.85 in the test set. A similar result was achieved using the Holt Winters method with a frequency of 12, in combination with the bias by weekday, providing a RMSE of 9.07.

While this was already a good improvement to the naive method, further improvement should be possible in future works. A clear correlation between the electricity price and the energy mix could be shown. Especially coal and wind energy have a strong correlation to the electricity price. This should help to achieve a higher accuracy with additional information on weather forecasts.