בס"ד

Home assignment - Jether Energy Research

Submitted by Nomi Hadar

February 2020

Part A

1. Validation of the input signals

Data size and Sampling frequency

- Size of actual data: **630,523**
- Size of forecasts data: **59,323**

The reason for the difference in data sizes might be due to the difference in sampling frequency:

- The actual values are given at **5 minutes** intervals.
- The forecasted values are given at **1 hour** interval.

And indeed, if we divide the size of actual by 12 (60/5), the result (~52,000) is closer to the size of forecasts data.

Missing values

- Actual data does not have missing values.
- There is **1** missing value in column ECM.
- There are **764** missing value in column MDA.

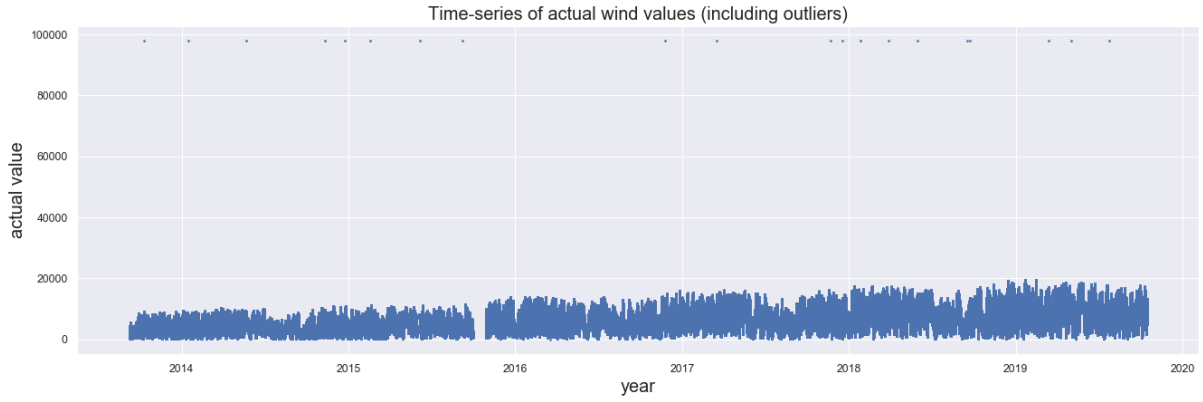
I **removed** rows where one of the values is missing.

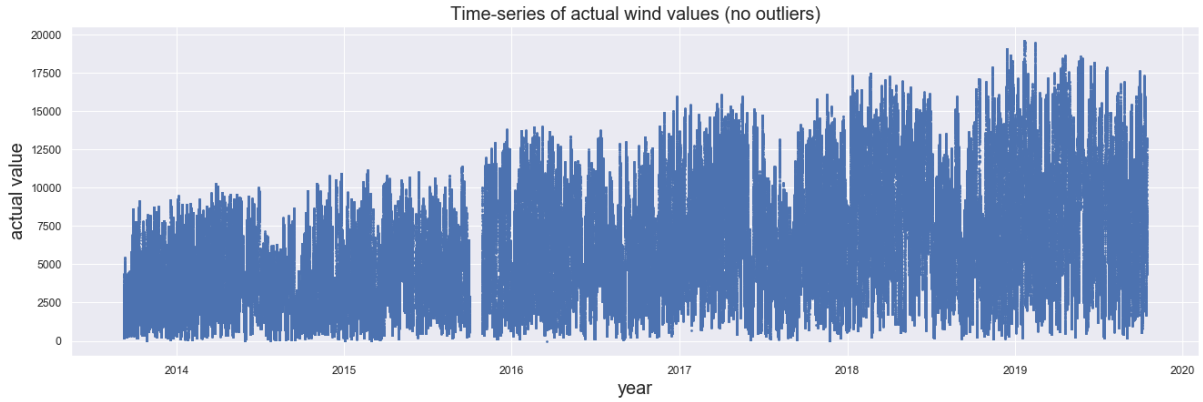
Outliers

Plotting the data against time – we can see that in the **actual** data there are several points close to the value 100,000. Those values are outliers. There are 20 such points, all with the value : **97944.2**.

To remove outliers, we can either remove these points, or replace their values by the mean/median of their neighbors (neighbors can be samples from the same hour). Since there are only 20 outliers, from different days, it will be just enough to remove them.

בס"ד


Time-series of actual wind values (including outliers)
actual value
year


Time-series of actual wind values (no outliers)
actual value
year


Time-series of forecased wind values
ECM value
MDA value
year

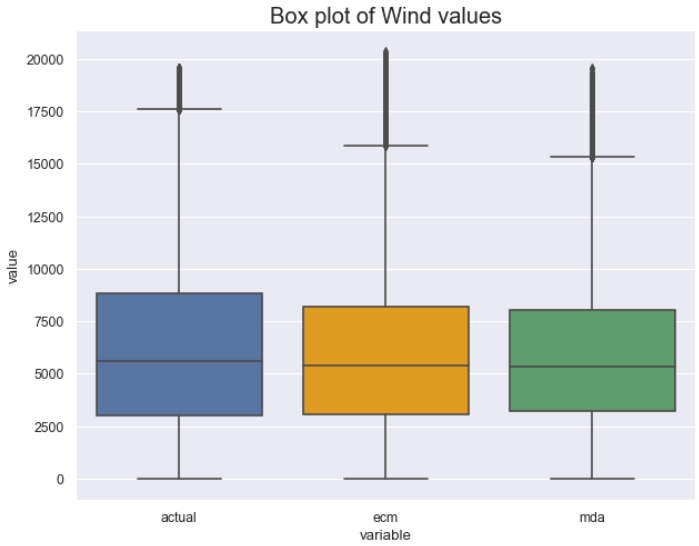
2. Data characteristics

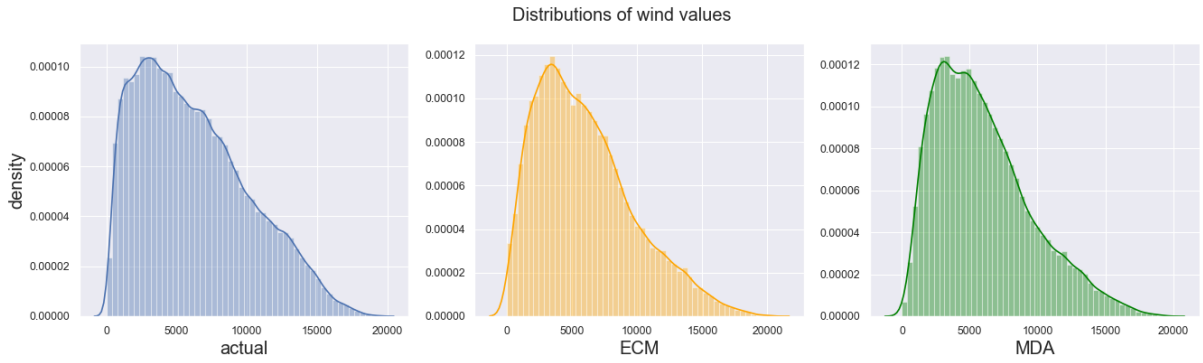
Data characteristic after removing NA values and outliers.

Description of data

- Range of actual data: 0-19,588.
- Range of ECM: 0-20,350.
- Range of MDA: 0-19,587.

	actual	*ecm*	*mda*
count	630503	58559	58559
mean	6221	5993	5992
std	3946	3754	3552
25%	2996	3077	3206
50%	5595	5375	5356
75%	8846	8204	8059





At first glance, the distributions of the three datasets are quite similar: high density in range of about (1000,8000), and then a rapid decline as values grow. However, at a closer look, it can be seen that the distributions of the forecasts are similar to each other more than to the actual, as suggested by the following points:

- **Mean:** forecasts have almost the same mean: ~5,990, while the mean of the actual is quite bigger: ~6,220.
- **Variance:** datasets differ in their variance (std^2). Actual has the highest variance, while MDA has the lowest.
- **Quantiles:** as can be seen in the box plots, the range of 50% of the data is the highest in the actual dataset.
- **Big values:** as can be seen in the distributions plots, values in range ~10,000 – 14,000, has higher density in actual distribution than in forecasts distributions.

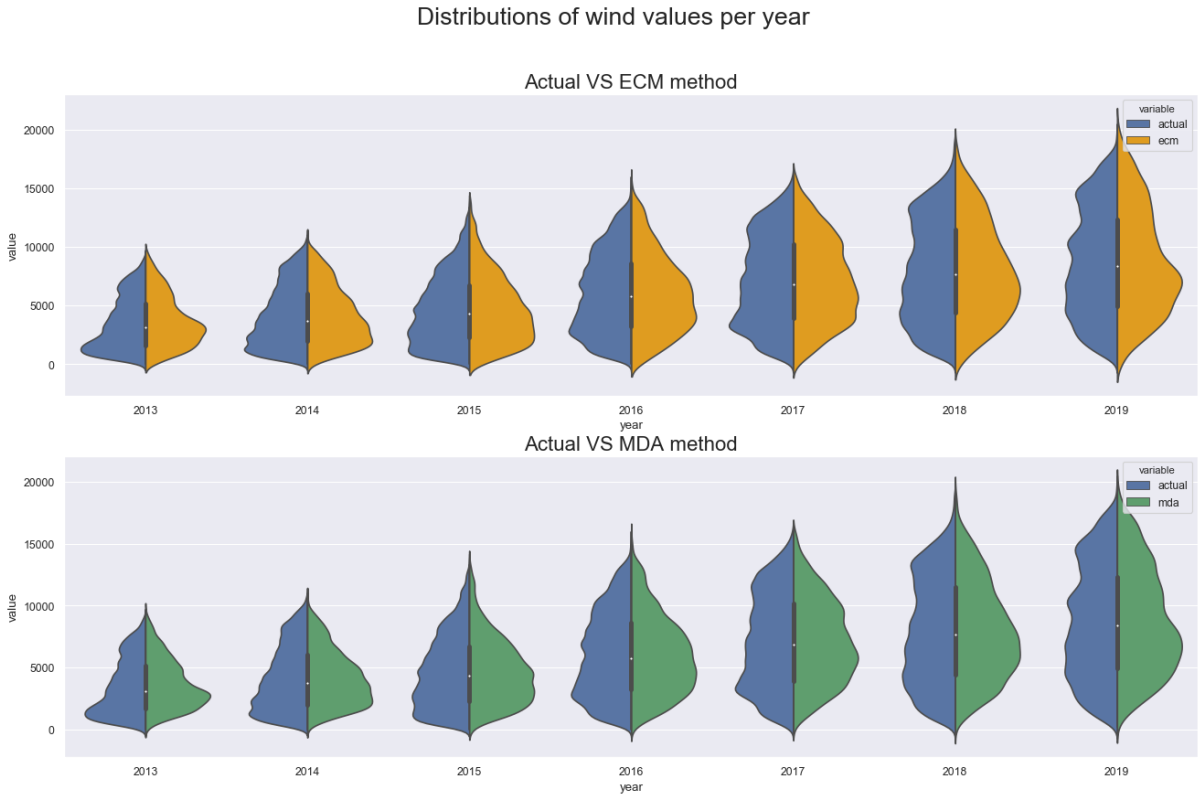
More points to note:

- **Total range:** Total range of ECM forecast is bigger than of actual. While maximal value of actual is 19,588, the maximal value of ECM is 20,350, a difference of about ~760. In contrast, the maximal value of MDA is similar to actual.

Distributions per year

The following figure shows violin plots of actual vs foracast, per year. First thing to notice, is that range of data grows as the years progress. The reason for this fact is not clear to me, and is probably has to do with sampling.

Second, for all years, the distibutions of actual and foracast are not substantially different, for both forecasts. However, as mentioned before, the forecasts distributoins are more similar to each other (orange VS green) than to actual. In general, it can be seen that the forecast distributions are smoother comapre to actual.



Statistical hypothesis testing

Kolmogorov-Smirnov test for 2-samples. This is a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.

I conducted the following tests:

- Actual and ECM are from the same distribution.
- Actual and MDA are from the same distribution.
- ECM and MDA are from the same distribution.

For *pval=0.05*, all three hypotheses were rejected. This means that the forecasts are not identically distributed, and that none of them comes from the same distribution of actual.

Kolmogorov-Smirnov test for goodness of fit. This performs a test of the distribution G(x) of an observed random variable against a given distribution F(x). Under the null hypothesis the two distributions are identical, G(x)=F(x).

I conducted the following tests:

- Actual comes from the gamma distribution.
- Actual comes from the log-normal distribution.

For *pval=0.05*, hypotheses were rejected. This means actual data does not come from the gamma or the log-normal distributions. To find the distribution of the actual data, more tests should be conducted, but in most cases "real" data doesn't fit to well-known distribution.

3. Evaluate error

Error is defined as: **actual - forecast**.

To evaluate error, we first have to match the sampling frequency of actual and forecasts. To this end, we **average** the actual wind values **over an hour**. Second, we merge on time (i.e., we select samples which share the same time-stamps in actual and forecasts).

We are left with **51,823** samples.

Description of errors

	ECM error	*MDA error*
count	51823	51823
mean	-10	-6
std	1552	1495
min	-8433	-7400
25%	-883	-907
49%	-17	-67
50%	15	-33
51%	45	1
75%	916	920
max	8779	8457

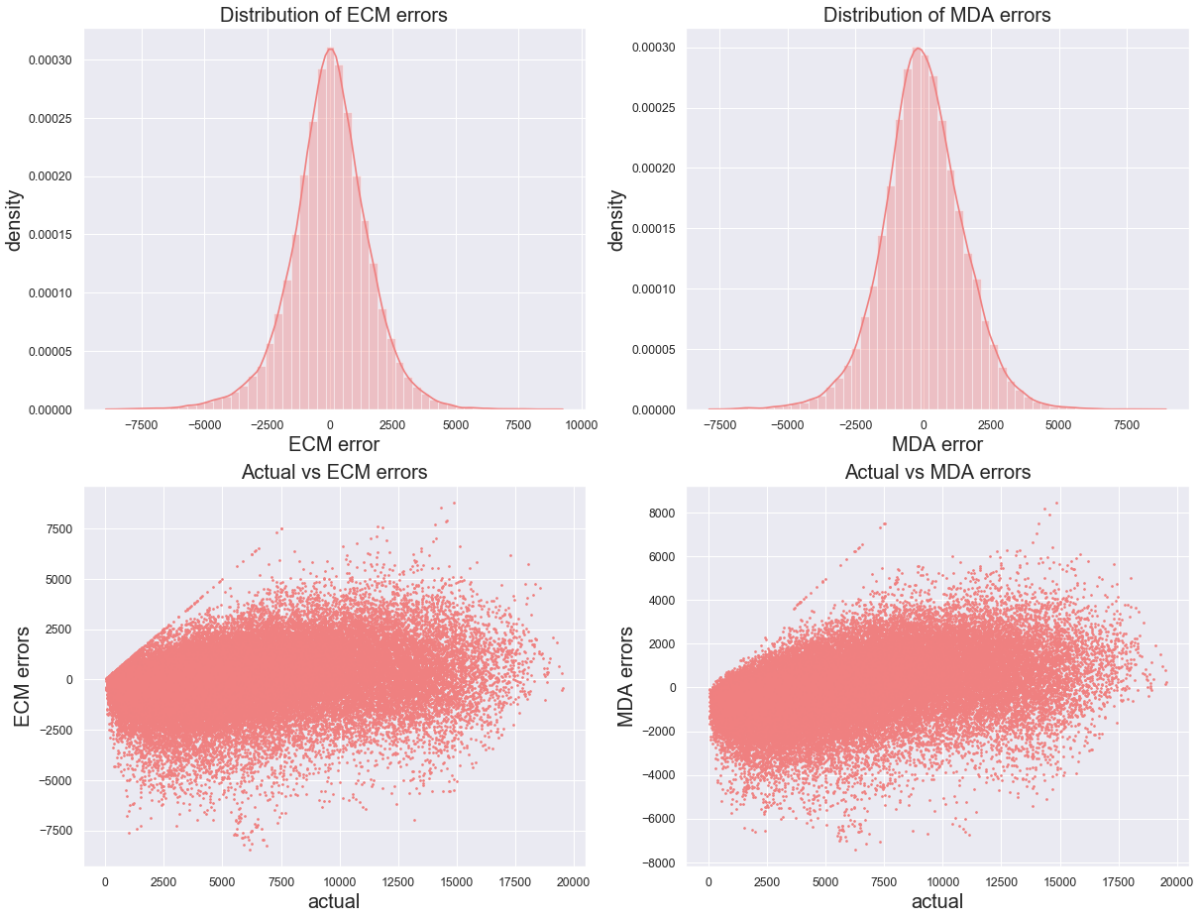
- **Mean:** Mean error of ECM is slightly bigger than of MDA. Mean error of both forecasts are negative and close to 0. This means that the sums of the positive and the negative errors are similar, with a little tendency to overestimation, i.e., in average, forecasted values are bigger than actual, rather than smaller.
- **Variance:** variance of ECM errors is slightly bigger than of MDA errors.

- **Range:** range of values of ECM is bigger than of MDA, in both directions, but not symmetrically: In the negative side, range of values is bigger in about 1,000 (-8,433 vs -7400), while in the positive side it is bigger in about 300 (8779 vs 8457).
- **Quantiles:** From the fact that signs are reversed in the ~50% quantile, we can deduce that number of positive values equals numbers of negative values.

Plot errors

The figure below shows the distributions of errors, as well as scatter plots of actual vs errors. The two distributions of ECM and MDA errors look similar. It can be seen that most errors are in range (-2500, 2500). The distributions are quite symmetrical around zero– equal number of positive and negative values.

From the scatter plots, we can learn the following: For actual values in range (0,75000), most errors are negative, indicating that low values tend more to overestimation. Other than that, distribution of points is quite uniform, with a slight tendency to grow as actual values grows.



Statistical hypothesis testing

Normal test: tests the null hypothesis that a sample comes from a normal distribution. I conducted the following tests:

- ECM errors come from the normal distribution.
- MDA errors come from the normal distribution.

For *pval=0.05*, hypotheses were rejected. This means that although the distribution form of ECM and MDA errors resembles the normal distribution, neither of them are normally distributed.

4. A quantitative analysis of the performance of the forecasts

For the quantitative analysis of the forecasts performance, I used three evaluation metrics:

- Mean absolute error (MAE).
- Root mean square error (RMSE).
- R squared (R^2).

The difference between MAE and RMSE is that RMSE gives more weight to points further away from the mean.

Results:

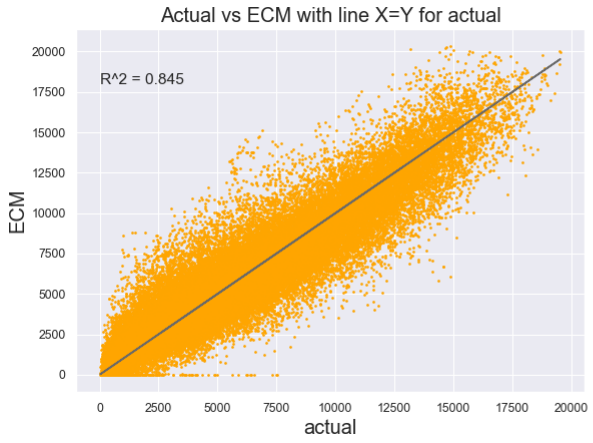
	ECM	***MDA***
MAE	1166	1147
RMSE	1552	1494
R^2	0.85	0.86

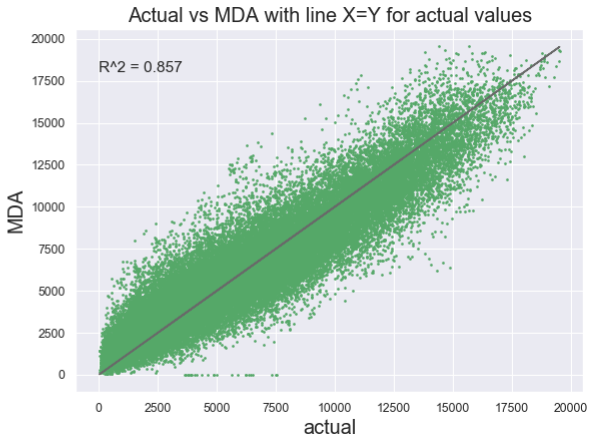
We can see that metrics values are similar between ECM and MDA, with a slight advantage to the MDA forecast.

Result of MAE means that the average deviation of the forecasted value from the true value, is about 1,150. Under the assumption wind values are in range 0 to 20,000, this amount is about 6% of values range $(\frac{1150}{20000} * 100)$.

Plots of actual vs forecasts

Following plots show actual vs forecasts, with a line X=Y for the actual values, with R^2 values.





5. Performance in different periods and in different ranges of the wind signal

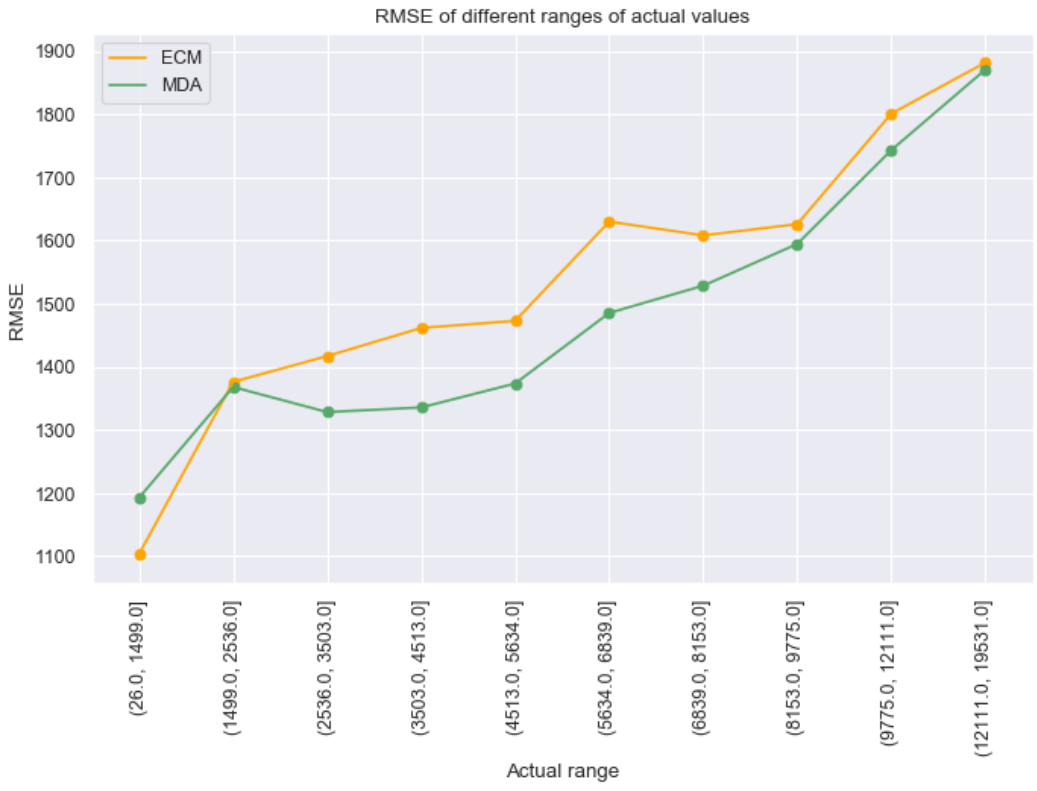
Different ranges of wind signal

To evaluate performances of forecasts in different ranges of values, actual data was divided into 10 quantiles. RMSEs were then calculated for each quantile, and were plotted against the quantiles.

Following figure shows the same trend for both forecasts: the higher the values range, the higher the RMSE. In both forecasts there is a sharp increase in RMSE in edge quantiles (1 and 9-10).

MDA forecast performs better for all quantiles, except of the first quantile. The difference between the forecasts' performances is notable mainly in quantiles 3-7, which contain values in range [2500, 8200].

Statistical hypothesis testing – I used ANOVA test to test the hypothesis that the RMSEs of ECM and MDA have the same population mean. For *pval=0.05*, hypothesis was not rejected.



Different periods

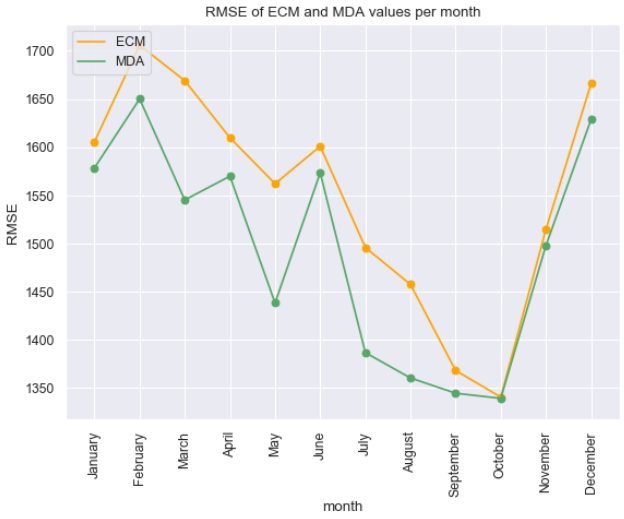
To evaluate performances of forecasts in different time periods, actual data was divided into months and years (separately). RMSEs were then calculated for each period, and were plotted against periods. Right figure: means of actual values per month.

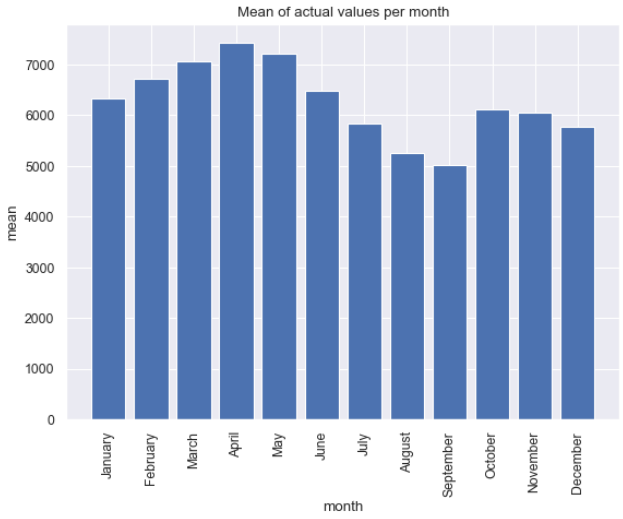
Months: Top figure shows that forecasts behave quite similar along the months.

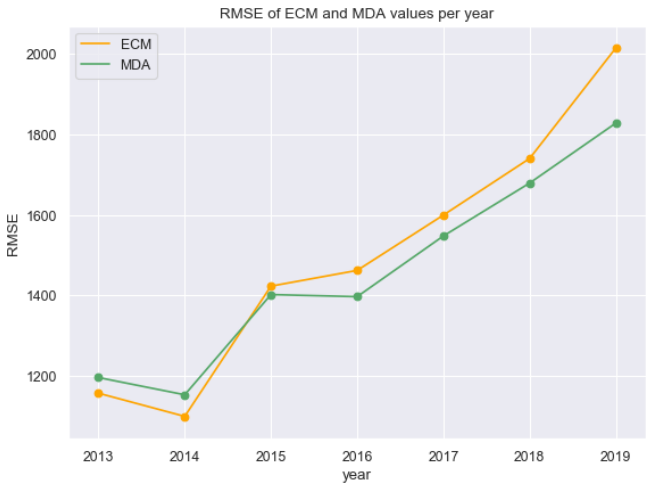
MDA forecast performs better for all months. The difference between the forecasts' performances is notable mainly in months March, May, July and August. There does not seem to be a relationship between these performances and the means shown on the right plot.

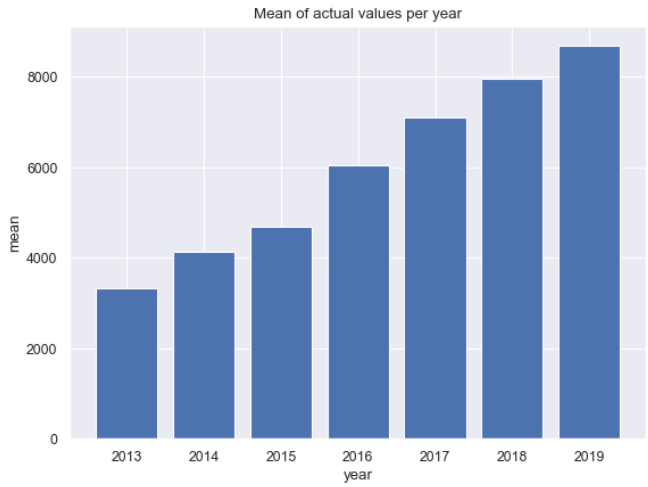
Years: Top figure shows that forecasts behave very similar along the years. **MDA** forecast performs better for all years, except of 2013-2014. The difference between the forecasts' performances is notable mainly in 2019. Here we can see a relationship between the increasing trend performances and the means shown on the right plot: the larger the mean, the larger the RMSE (except in 2014). This is consistent with previous results of different ranges of wind signal.

Statistical hypothesis testing – I used ANOVA test to test the hypotheses that the RMSEs of ECM and MDA have the same population mean for months and for years. For *pval=0.05*, hypotheses were not rejected.









6. Write an analysis tool

Instruction of this section were not detailed enough. I made the following assumption: input arguments are actual and forecast datasets (which share same timestamps), and evaluation metric can be MAE or RMSE.

Such utility can be useful to visualize trends of errors along time.

We can use it in the following situation: we have a forecast method that we use continuously, and meanwhile we are working on improving accuracy (for example, by adding features, or by testing different hyper-parameters). We can use the utility to examine the performances of our model – whether it gets better with time (decreasing trend) or not (static or increasing trend).

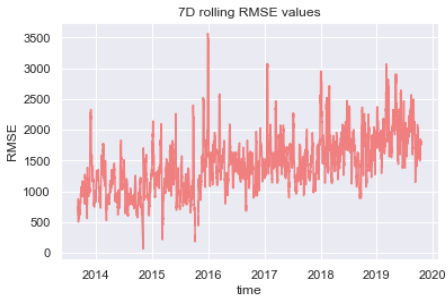
In addition, this utility can be used to compare performances of different forecasts methods over time.

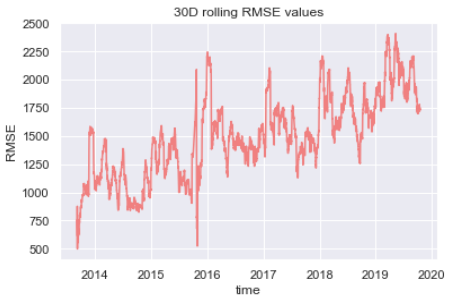
Results on provided forecasts:

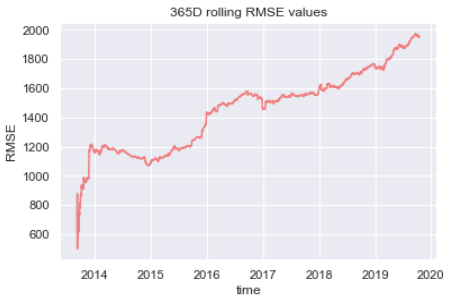
For both forecasts, there is an increasing RMSE trend over time, which means decreasing in performances. This result is of course consistent with the previous plot (RMSEs per year).

Note that the longer the period, the smaller the RMSE range of values (due to rolling). For all periods, maximal values of **MDA** RMSEs are smaller than of **ECM** RMSEs.

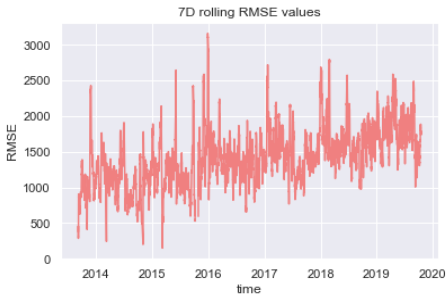
Plots for **ECM** forecast, RMSE metric, and the periods: week, month, year

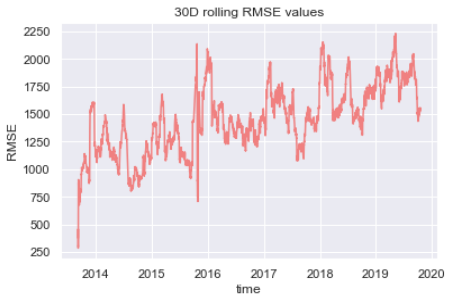


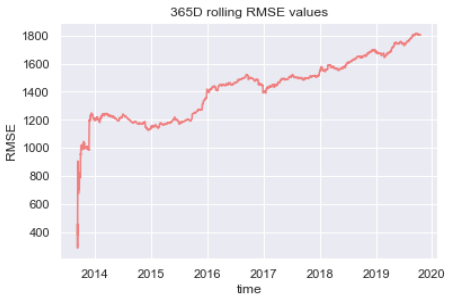




Plots for **MDA** forecast, RMSE metric, and the periods: week, month, year







7. Conclusion

Given the above analysis, I would recommend using the **MDA method**.

Following is a summary of the results supporting this conclusion:

- MAE / RMSE values of MDA are slightly better than of ECM.
- Performances of MDA in different ranges and periods exceed ECM performances. Exceptions: low ranges (0-2000), where ECM does better, and except of years 2013-2014, which are characterized by low mean of values. Nonetheless, the difference in performances in these exceptions is small.
- Maximal rolled RMSE value of MDA is less than of MDA (for all tested periods).

Another conclusion is related to the evaluation metric:

For some unknown reason, we saw that range of wind values increases with years (from 10,000 in 2014 to 20,000 in 2019, see violin plot). We showed that the larger the average wind values – the larger the mean error. Now, if forecasting the value 2,000 with an absolute error of 300 is much worse than forecasting the value 19,000 with an absolute error of 300, then we might consider use **relative error** as our evaluation metric. In this way, size of errors will not be dependent on wind ranges.

Part B

The United States Wind Turbine Database (USWTDB) provides the locations of land-based and offshore wind turbines in the United States, corresponding wind project information, and turbine technical specifications. The **Trent Mesa** wind project is one of the project included in this database.

I download a CSV file (here) of data, end extracted the rows of Trent Mesa farm. File contains a row for each turbine in the farm. Main attributes:

- **t_state:** State where turbine is located.
- **t_county:** County or county equivalent where turbine is located.
- **p_name:** Name of the wind power project that the turbine is a part of.
- **p_year:** Year the wind power project became operational and began providing power.
- **p_tnum**: Number of turbines in the wind power project.
- **p_cap:** Cumulative capacity of all turbines in the wind power project, in megawatts (MW
- **t_cap:** Turbine rated capacity in kilowatt (kW). The manufacturer's stated output power at rated wind speed.
- **t_ttlh:** Turbine total height.
- **Xlong, ylong:** Latitude and longitude.

In the Trent Mesa project there are 83 turbines.

Wind turbine rated capacity is the amount of power a wind turbine can produce at its rated wind speed. If we know to forecast wind speed/power in the area where the Trent Mesa farm is located, then we can get forecast the amount of power each turbine will produce. If we sum these amounts, we will get the forecasted wind generation of the Trent Mesa project.