Statistics for Spatio-Temporal Data: Germany's Solar power generation analysis

Santoro Arnaldo mat:822274

A.A. 2018/2019

1 Introduction

1.1 Background

The constant menace of global warming pushes towards green power sources exploitation, such as wind and solar power. This can be seen also in EU spending into funds for green energy production, which now makes from 2% to 4% of the total energy expenditure of the EU, Germany and Italy being two leading countries in European solar power business.

Since the technologies to exploit these resources are relatively new, they are quite uneffective for various reasons: - the source of energy is not continuous because it depends on the environment - storing the power in excess is very expensive - solar panel production is not carbon neutral, and panels take from 6 months to 4 years to become carbon neutral, depending on the panels' quality and durability.

While battery technology improvements and technology advancement are making the private solar generation ever closer to be carbon neutral, the main source of uncertainty in energy production depends on the weather and panels quality.

This makes the kind of data analysed here of public interest, and appropriate analysis may be useful to European states, to power companies, and (maybe) to the private citizens.

1.2 Data

Data source https://data.open-power-system-data.org/time_series/latest/

Documentation https://nbviewer.jupyter.org/github/Open-Power-System-Data/datapackage_timeseries/blob/2019-05-15/main.ipynb

The OPEN POWER SYSTEM DATA site provides data on solar and wind power in European states. The main data sources are the various European Transmission System Operators (TSOs), the ENTSO-E Power Statistics and the ENTSO-E Transparency Platform (more info in the links). The datasets contain data from 2005 until 2019 on solar and wind power generation, consumption, capacity, and other variables, but many entries are missing in different measure for the various contries . There are three types of datasets available: * A dataset with observations every 15 minutes. * A dataset with observations every 30 minutes. * A dataset with observations every hour.

The analysed dataset will be Germany's solar power generation data from Jan 2015 to Jan 2019); this choice is due to the relatively low number of missing observations. In order to simplify the analysis the data has been aggregated to monthly means of daily TW/h generated.

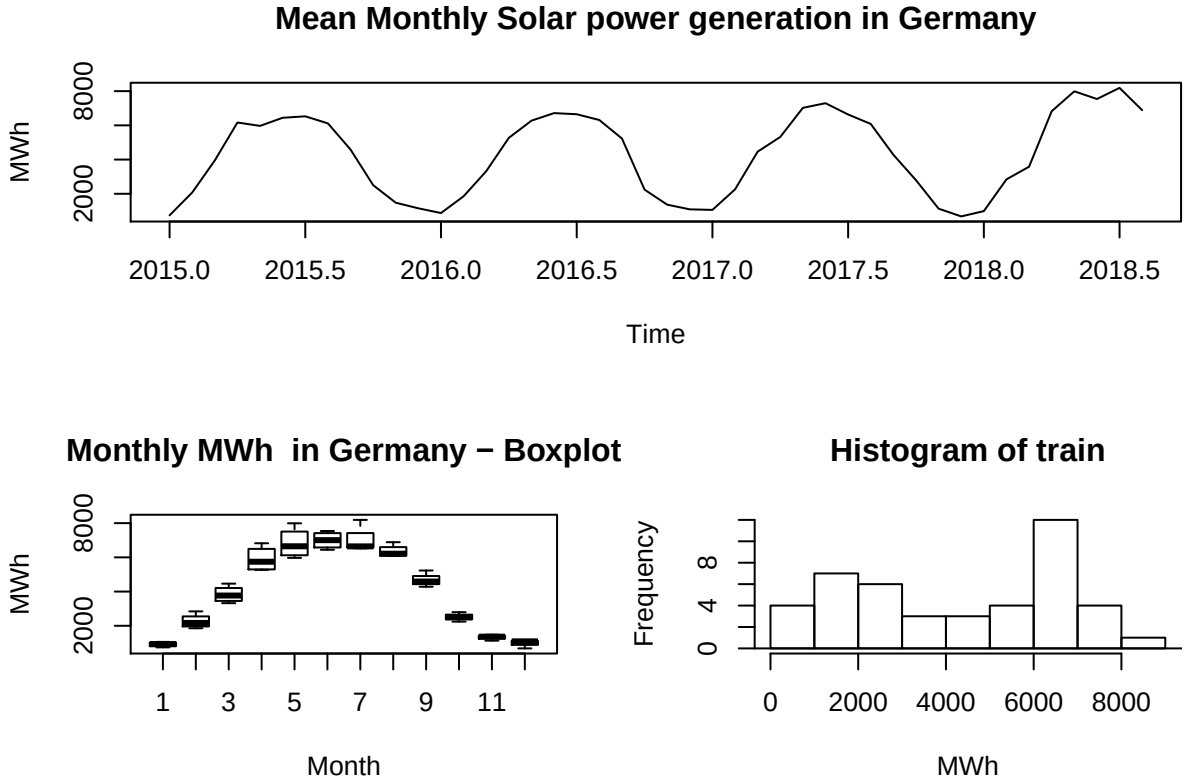
Thus, due to the relatively few number of observations obtained (49) I choose to split the dataset into train and test subsets, conserving only the last 4 observation for testing.

1.3 Methods

After some exploratory analysis of the data, some tests about the stationarity of the series will be performed. After that, the series will be decomposed, this has allowing trend studying. In order to do some forecasting, a Seasonal ARIMA model has been built.

2 Methodology

2.1 Data exploration and manipolation



In figure 1 you can find the monthly means of daily MWh generation, followed by its distribution and variance. As you can see, the series has a clear seasonal pattern, and is slightly trended in the last part. The histogram shows a bimodal distribution; this is due to the effect of seasonality.

Looking at the month boxplot, we can see that solar generation is high during summer, and drops near zero during winter; the highest variance occurs in spring and, in a lesser extent, during autumn when the weather is more unstable. Three kinds of tests have been performed: * the Ljung-Box test examines whether there is significant evidence for non-zero correlations at given lags, with the null hypothesis of independence; * another test we can conduct is the Augmented Dickey–Fuller (ADF) t-statistic test to find if the series has a unit root (a series with a trend line will have a unit root and result in a large p-value); * Lastly, we can test if the time series is level or trend stationary using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. In this test, the null hypothesis is that the data are stationary, and we look for evidence that the null hypothesis is false, therefore a low p-value will indicate a signal that is not trend stationary.

We can apply a seasonal differentiation at lag 12 in order to remove a possible seasonal trend, or a simple differentiation, or both. The justification for this choice lies in the previous test results. Both the single differentiation and the raw series fail all tests, since the Ljung-Box Q statistic has a p-value less than 2.2e-16 (for h = 2*12 = 24), and in the ADF test, given an alernative hypothesis of stationarity, a p-value of 0.7792 for the raw series, and of 0.4244 for the differentiated series (for k=12) is returned. The KPSS test reports a p-value greater than 0.1. These tests leads us to apply a differencing operator at lag 12 in order to remove a possible seasonal trend, as suggested in the previous section. After this Ljung-Box reports a p-value of 0.9827 , in the ADF test a 0.5268 p-value is returned, while KPSS has worsened. These results are show more evidence for stationarity, and confirm the presence of a seasonal trend, requiring a seasonal differenciation (D = 1) of the series, but are not conclusive. If both differencing operators are applied we obtain equally

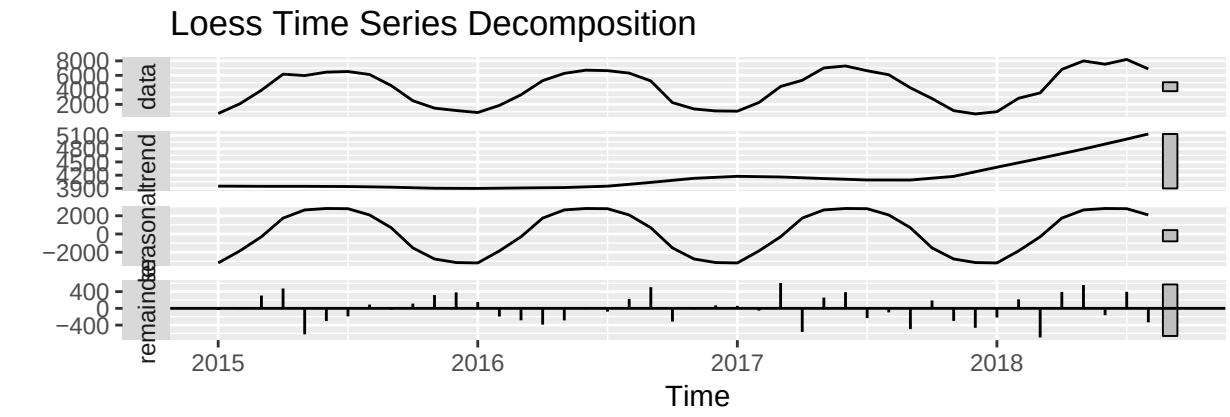
satisfactory results: an acceptable p-value in the Box-test equal to 0.5922, a still high ADF test p-value of 0.6599, and a KPSS t-statistic slightly higher than 0.1, making it a riskier but valid option, expecially useful to generate alternative models.

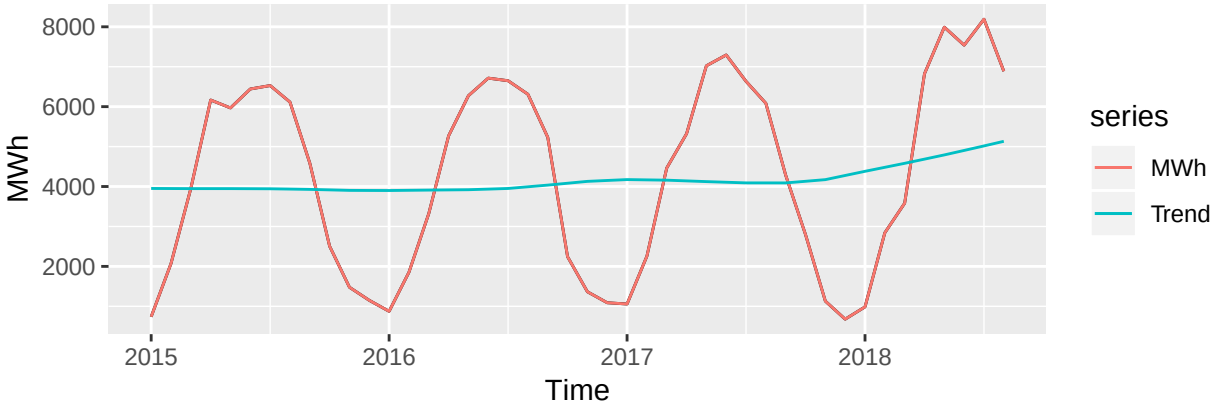
2.2 Series decomposition

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```
##
## Attaching package: 'gridExtra'
```

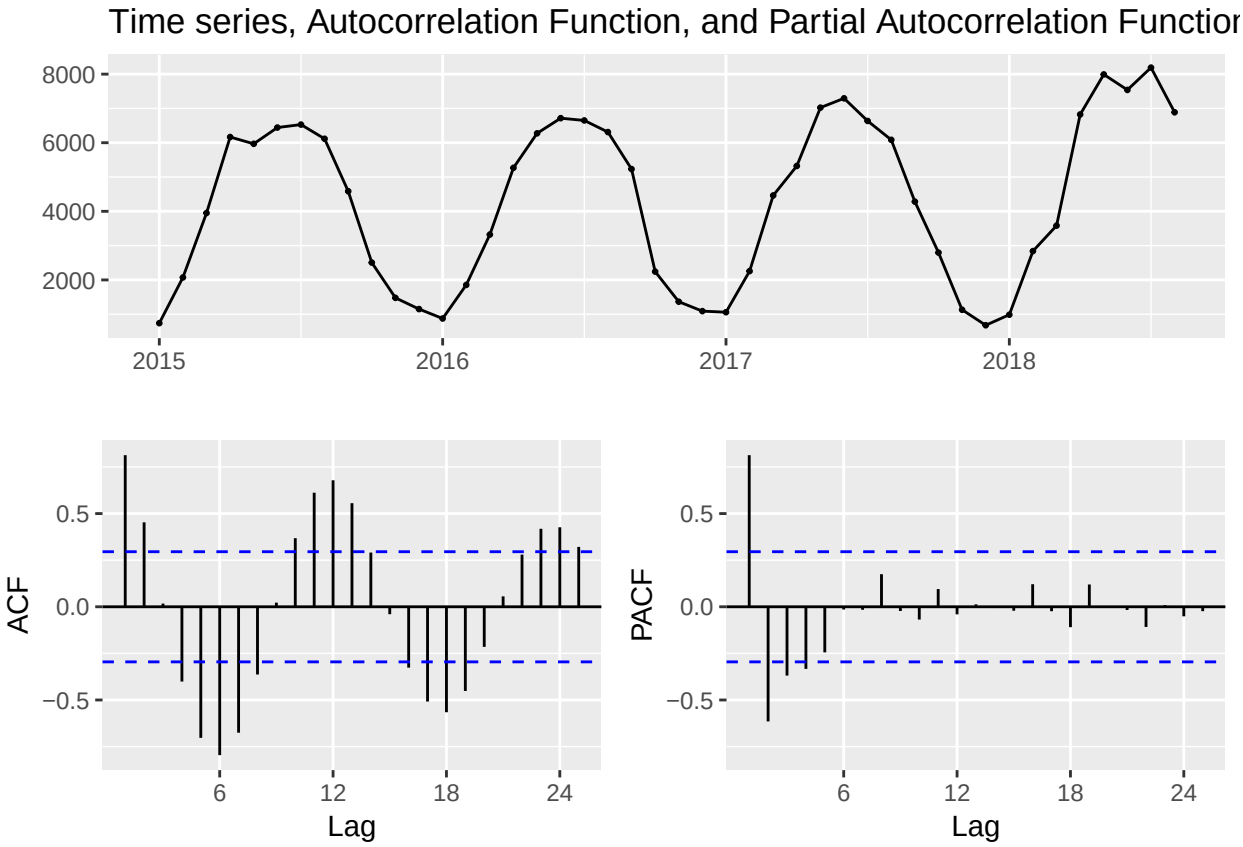
```
## The following object is masked from 'package:dplyr':
##
##     combine
```





Here can find the seasonal decomposition of the series in its additive form. Loess seasonal decomposition has been used with a yearly seasonality. As described in section 2.1 the series is characterised by a trend that is increasing in the rightmost part. During the first two years the trend was quite stationary, then in 2017 we have a slight deviation from the mean, but in the last year the trend has increased visibly, probably due to an exceptionally sunny year. About the seasonal component: we can observe a significant amplitude of approximately 4000 MWh, with slightly faster increase during spring, and a slower decrease during winter, with a relatively stable highs during summer.

2.4 Model Fitting



As you can see from the figure above, the autocorrelation function shows the pattern of seasonaly trended series, as we saw in the previous analysis. It should be noted that only the last few years showed an increasing trend and its autocorrelation pattern is decreasing: this implies that undifferenced models should not be totally discarded. We have talked only about seasonal patterns, but the PACF seems to suggest the presence of some kind of AR process.

In view of the use of an automated model selection I take the liberty to use the riskier option and start with a model differenciated once both in the seasonal and nonseasonal component.

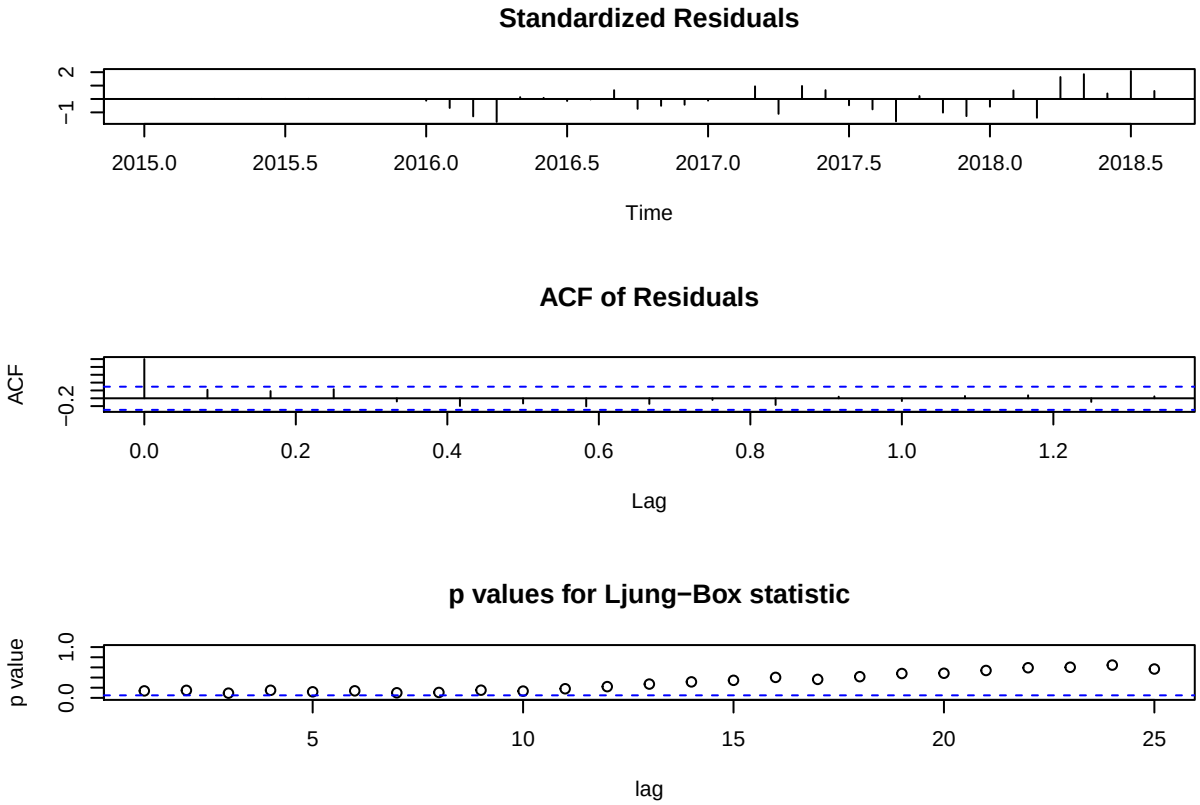
This model (D=1) doesn't passes the Ljung-Box test, so I proceeded to add parameters, and then remove all parameters whose absolute value was less than twice the standard error. Along this procedure I adopted an automatic model selection (the criterion was AICc) to select the third model. The results are three models, the first two manually selected, the third selected by the algorithm: - the first is an ARIMA(1,1,0)(0,1,0)12 (hereafter called model 1); - the second is an ARIMA(0,1,1)(1,1,0)12 (hereafter called model 2); - the third is an ARIMA(0,0,0)(0,1,1)12 (hereafter called model 3).

Model 1 is much simpler than model 2, and model 3 is the simplest but also includes a drift; moreover all parameters of model 3 doesn't satisfy the requirement for which the ratio between the estimated coefficients and their standard error is greater than two.

- model 1 AIC=494.39 AICc=495.28 BIC=498.69 RMSE=574.5916 MASE=0.8055076
- model 2 AIC=491.66 AICc=492.55 BIC=495.96 RMSE=488.6486 MASE=0.705224
- model 3 AIC=504.87 AICc=505.73 BIC=509.27 RMSE=465.3085 MASE=0.6423455

Above we can find the information criteria that characterise these models, along with some accuracy measure. The first two models are quite similar while the third is very simple. By observing the information criteria alone

Model 2 overcomes model 1, but the latter is simpler, and for this reason it will be kept and tested in the next section. It must be noted that, while the third model seems to perform the worst, it possess a different differencing order so the data on which the likelihood is computed differs, thus making these information criteria unreliable.



Above we can see model 3's residual diagnosis. As we can see, in the first year there are no residuals, as an effect of the seasonal differentiation, and in the last year the residuals increase in frequency and size, probably as effect of the trend, but not so much to be significant. In fact the ACF shows that the residuals are quite uncorrelated. The residuals of model 1 and 2 are very similar in behaviour, but perform much better in the Ljung-Box test (omitted for the sake of brevity) and can be accepted. The Ljung-Box p-values for model 3 are almost too small 10 lags showing that the residuals from the model are not weakly stationary. For this reason the model is risky to work with, if not unreliable. Even so, the third model is the one with both the smallest root mean square error and mean absolute scaled error and for this reason will be validated

3 Results

In the next section the models will be validated. First of all the forecast produced will be compared with the test set as described in section 1.2. After that, the cross-validation of the model will be performed.

3.1 Forecasting

In the figure above the forecasting of the last three months of 2018 and january 2019 (h=4) is drawn, with its real data underlayed. All models' means do not properly catch the real data; nevertheless, the prediction interval covers it, even if in the second model the coverage is barely caught. As ARIMA is used here to forecast just a few steps it catches properly the data, which is not guaranteed for longer forecast intervals using ARIMA models. As said in the last section, from the RMSE and MASE (and all other accuracy measures) it appears that the simplicity of the third model (and the more stationary premises) beats the complexity of the first two models.

3.2 Cross-Validation

Given the few observations kept for testing the data, this section is more important as a demonstration of the technique, than actual usefulness in the analysis, but is important also to verify the ability of the models to predicting data. In order to check the accuracy of the model without the need of long term forecasting, cross validation has been performed, in this case, due to the dependency of the data, the procedure must be ran sequentially. Given the few observations kept for testing the data, this section is more important as a demonstration of the technique. The coefficients of previous models are fixed and considering the full dataset (test and training sets) a subset is taken as training data,and the rest is used as validation set. After that, the model will predict the following month's power generation, the square error of such prediction with respect to the real observation is computed and the latter is moved from validation set to the training set. This process is iterated for the entire dataset. At the end of the precedure the rooted mean of previous errors is computed.

These results confirm what was said in the previous sections: the automatically selected model beats both in accuracy and simplicity both models, as model 3 RMSE from Cross Validation is 613.733, while model 2 has a RMSE of 640.3908, and model 1 has an even higher RMSE.

4 Discussion and further perspectives

The models provided can catch the monthly behaviour of Germany's solar power generation, especially model 3.

Decomposing the time series we revealed the presence of an increasing trend. Some explaination comes from three facts: the exceptionally high number of sunny days in 2018, the continuous increase in power capacity (can be seen also in the original dataset), the increase in EU investiments in solar power.

Since the finality of this paper was to test the tools rather than reach optimal results in the analysis I took the liberty to choose some models mainly to add variety, therefore a more result-focused research could lead to more insights more efficiently.

More accurate results could be obtained by analysing more data, as well as using a more fine-grained dataset such the hourly dataset or even finer.

Real model should take into account the numerous other variables, such as capacity, raw generated power, and many others provided by the same source as this data. Finally the dataset used for the analysis is just a small subset of the enormous data on solar and wind power in the whole Europe, which is likely to become increasingly important in the next years.