## ARDL on seasonal data

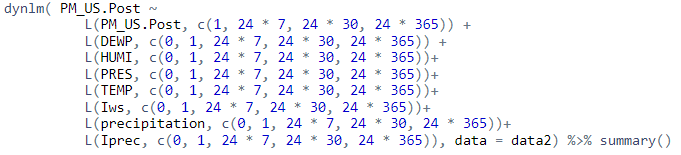
After estimating the ARIMA models, we wanted to have an alternative models to which we could compare the results. For this we choose ARDL. We initially run the ARDL only on the hourly data, but soon discovered that we could learn more about general trends in the data, by ‘smoothing’ the values.

Because of this, to estimate the models we will be using 3 data sets. The original hourly data, and 2 new sets that were created by summarizing the hourly data into daily and weekly periods. The method that we choose for this is a arithmetic mean. We investigated other options such as a truncated mean or median, but the results were not much different, so we choose the simples solution.

Since our data has both significant seasonality, and great variance in the short term, we decided to test models estimating the value od PM2.5, as well as the differences between each time period (diff(PM2.5)). We assumed that estimating values would allow us to predict general trends, and estimating differences would allow us to estimate significant changes between time periods. To test this assumption, we run both kinds of models.

## Initial models

To start the analysis, we run each model with all available variables, each in all relevant lags (daily, weekly, monthly, yearly). This created a relatively complex models that we could further optimize by removing the insignificant variables.

Fig. 1 An example of the initial ARDL model, this one estimates the hourly value of PM2.5

|  |  |  |
| --- | --- | --- |
|  | Values | Differences |
| Hourly | 0.88 | 0.02 |
| Daily | 0.45 | 0.15 |
| Weekly | 0.523 | 0.521 |

Table. 1 R^2 values for initial ARDL models

As a first measure of model performance, we choose R^2. It can be seen from table 1 that models estimating values performed much better that those estimating the differences. Only for weekly data the difference in the measure is not significant. We think that this may be because the differences between two subsequent periods are mostly random, at least in the context of available data. This is only different in weekly data, that is by its nature much more smoothed out.

## Model optimization

Since the models for estimating values performed much better, we choose them to be further optimized. For each model we removed one variable with the lowest significance (counting all lags as separate variables) and run the model again. We did this until all remaining variables were significant with p value of most 0.05.

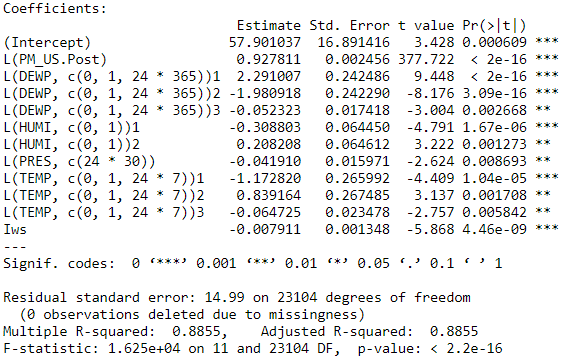
|  |  |
| --- | --- |
|  | R^2 |
| Hourly | 0.89 |
| Daily | 0.44 |
| Weekly | 0.54 |

Table 2. R^2 values for optimized ARDL models

In case of hourly and weekly data, removing variables made the model slightly better in term of predictive power. For daily data the R^2 dropped by 0.01 p.p.. This is to be expected, since this kind of optimization is done mostly to reduce the complexity of the models. In the following part, we will go through the final models, and then run tests to examine their results.

## Hourly ARDL

Fig. 2 Code for the final hourly ARDL model

Fig. 3 Results of the final hourly ARDL model

The final hourly ARDL model consists of 11 variables and was able to achieve R^2 of 0.8855. The model tells us that, for hourly PM2.5 data, the best predictors are the previous hour PM2.5, Dew Point for the current hour, previous hour, and last year, Humidity in the current and previous hour, Pressure from month ago and Temperature from current hour, previous hour and a week ago. Also significant is the current cumulated wind speed. It has to be noted, that most of this coefficients have a negative influence on the predicted value, with the exception of lagged PM2.5, current Dew Point, Humidity and Temperature. The intercept is 58, which is about equal to the median value of the PM2.5 variable.

## Daily ARDL

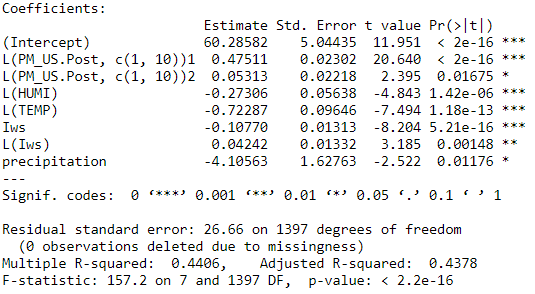
 Fig. 4 Code for the final daily ARDL model

Fig. 5 Results of the final daily ARDL model

The final daily ARDL model has a R^2 value of 0.44, and is made of 6 variables, all of which are significant. To estimate daily PM2.5 we start with the intercept of 60, then apply positive influence of the PM2.5 levels from a day and 10 days before, then apply negative influence of yesterday’s Humidity and Temperature, as well as todays Wind Speed and Precipitation. The only other positive influence is the previous day Wind Speed.

## Daily ARDL

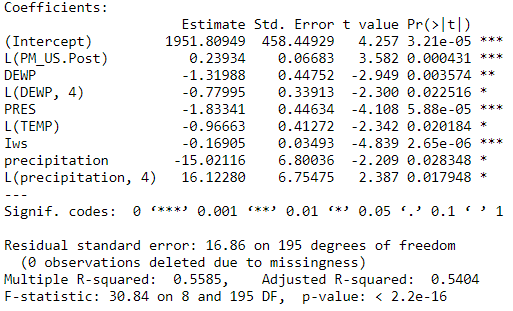
 Fig. 6 Code for the final weekly ARDL model

Fig. 7 Results of the final weekly ARDL model

The final model is the weekly ARDL. It has a R^2 of 0.54, and has 8 variables, more than the daily model, but less than hourly. For this model the most important influence is the intercept. Positive coefficients are the previous week PM2.5 values and previous month Precipitation. All other coefficients are negative. They are: this weeks and previous months Dew Point, this week Pressure, last week Temperature, and this week Wind Speed and Precipitation.

## Similarities between models

Although the models were run on different periodicity data, their results are similar. All of them include a lagged PM2.5 variable, as well as Temperature and Wind Speed. Two of three include Dew Point, Humidity, Pressure and Precipitation, in some form. There is not a single variable that appears in only a single model.

The R^2 values of daily and weekly models are fair, but not perfect. At 0.44 and 0.54 there is a significant variability that is not explained by the models. The hourly ARDL with its R^2 of 0.89 is surprisingly good, especially when compared to other models.

## Plots of the models

## Hourly

Fig. 8 Hourly model plots

*Residuals vs Fitted* – the plot is mostly linear but there are some outliers.

*Normal Q-Q* - follows the normal distribution, except for the most extreme values that are significantly different.

*Scale-Location* - values lie mostly in the left-hand side corner.

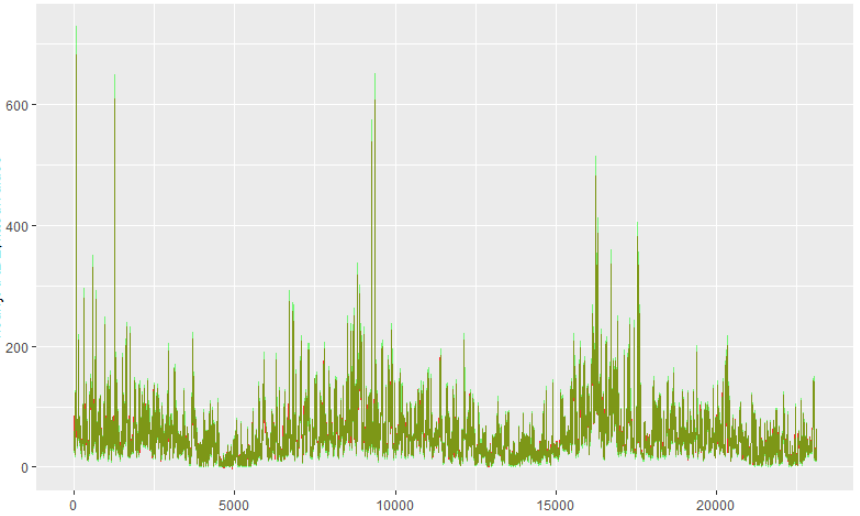
*Residuals vs Leverage* - there is one observation outside the Cook's distance.

Fig. 9 Differences between real and predicted values. The plots overlap in nearly all places, producing a green color.

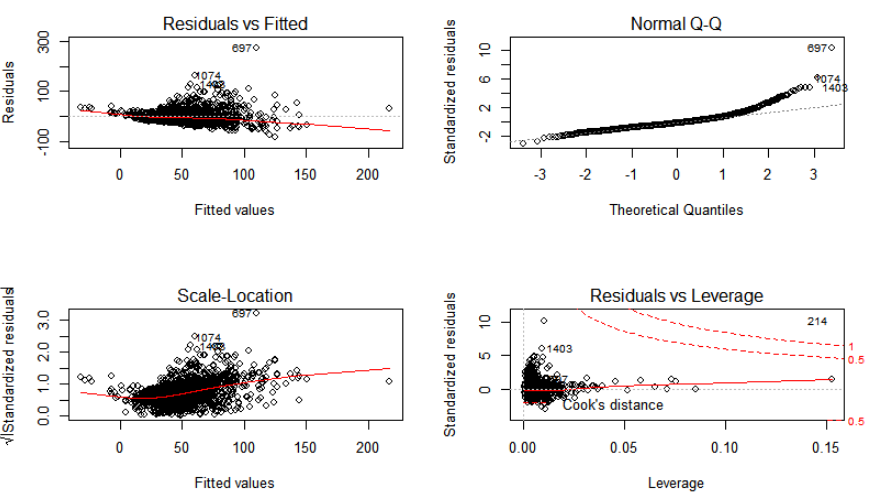
Daily

Fig. 10 Daily model plots

*Residuals vs Fitted* - no extra patterns, the residuals form a group in the middle.

*Normal Q-Q* - some deviation from the distribution in the extreme positive side.

*Scale-Location* - the points are not spread out evenly, but sit mostly in the middle part of the plot.

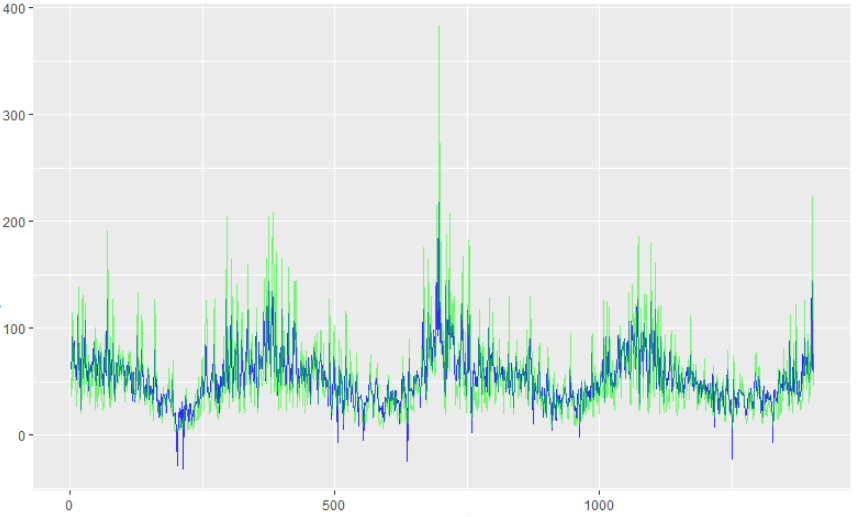
*Residuals vs Leverage* - there is one observation outside the Cook's distance (different than the one in hourly).

Fig. 11 Differences between real and predicted values. The real values (green) have much stronger outliers than the predicted ones (blue).

## Weekly

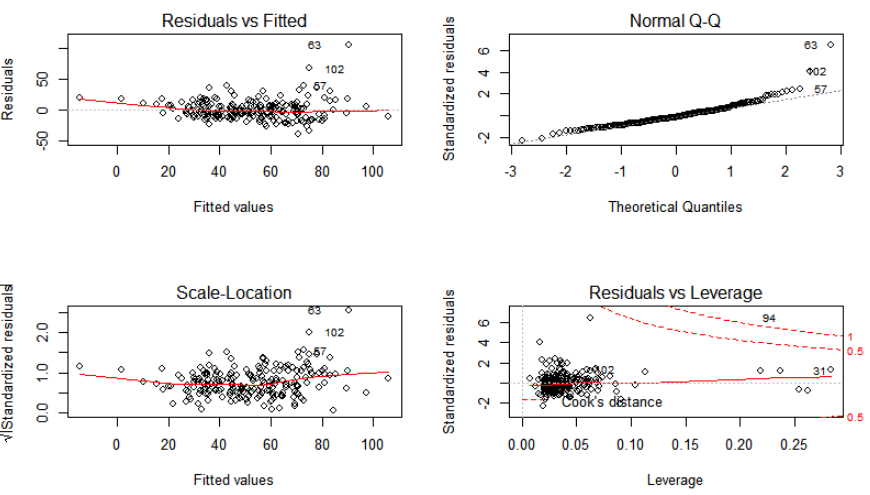


Fig. 12 Weekly model plots

*Residuals vs Fitted* - mostly linear pattern, but there are some outliers.

*Normal Q-Q* - very nice distribution, except for some outliers in the positive side.

*Scale-Location* – no significant patterns can be observed.

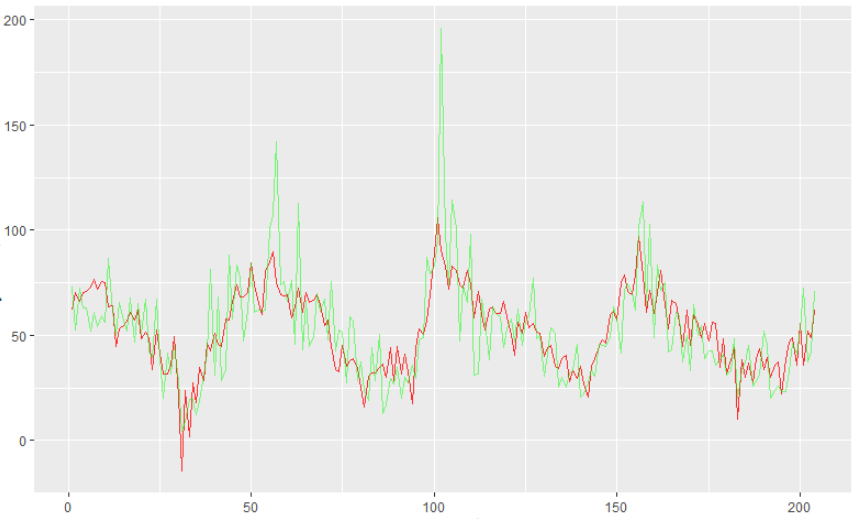
*Residuals vs Leverage* - One significant outlier, also different than previously.

Fig. 13 Differences between real and predicted values. Like previously, the real values (green) have much stronger outliers and variability than the predicted values (red). This time, the models did not create its own outliers.

## Testing the models

Breusch-Godfrey test for serial correlation of the residuals

The Breusch-Godfrey test is used to find serial correlation. We will use it to examine the residuals of the models, up to order 5.

For the hourly data the p values are all smaller than 0.001, which indicates a strong serial correlation of the residuals. This may in part explain the high R^2.

For the daily and weekly models all p-values are bigger than 0.1, which indicates that there is no serial correlation between the residuals.

Jarque - Bera Normality Test of the residuals

For all periods the p values are smaller than 0.001, which indicates that the residuals are not normally distributed. This may be in part due to some significant outliers in the residuals.

Breusch-Pagan test for homoscedasticity

For hourly and daily data, the p values are smaller than 0.001, which means that the results are heteroscedastic. This may be due to the strong seasonal character of the data.

For the weekly data the p value is equal to 0.0521, which is close to the threshold of 0.05. Because of this, it is not possible to accurately judge the homoscedasticity.

Summary

We choose to run 6 different models, for 3 types of periodicity. Based on these, we selected 3, and optimized them. The resulting models had R^2 values of 0.89, 0.44 and 0.54. They were able to accurately predict the general trend of the PM2.5 variable, but failed to explain the significant period to period variability that resulted in strong outliers in the real data.

Based on the models and the plots (fig. 8 to 13), it seems that the data has a very strong yearly seasonality, as well as strong period to period variability, but no significant trend. Because of this findings, we want to explore the dataset by trying to decompose the seasonal effects.

The Breusch-Godfrey, Jarque – Bera and Breusch-Pagan tests showed significant shortcomings of the resulting models. These results may be in part due to the seasonal and highly variable nature of the PM2.5 variable, but may also indicate some underlining problems in the estimations. Despite this, the models were able to provide good enough predictions for the data (see fig. 9, 11 and 13).

Based on this results, we can say that in a business environment this models could be used to predict the general trend of the PM2.5 variable, but will not be able to predict significant outliers.