Time Series Analysis of PM2.5 in Shanghai

Brief comparison of Predictive Power and Efficiency between

ARDL and ARIMA models

Korneliusz Krysiak, Michalina Cieślak, Michał Szałański

May 2019

Abstract

Models Overview

In order to investigate the current level of PM2.5, and predict the future levels of this particle in the city of Shanghai, a case study was conducted. We choose this data, because Shanghai is one of the most polluted Cities in the World, hence they gather a lot of data about air pollutants, and its derivatives. We will analyze only air pollution with respect to all Shanghai, combining outputs from different district to one. A statistical models, Autoregressive Integrated Moving Average (ARIMA), and Autoregressive Distributed-lagged (ARDL) model were used for this purpose. Data contained 31876 non-NULL observations, with 13 independent variables, from the date 2011.31.12 to 2015.31.12 . ARIMA results suggested that the best results can be obtained using SARIMA, while ARDL showed that for this type of data (multiple seasonalities) it is rather hard to perform a significant prediction. Following this study, you can find useful framework for future purposes of exploring this topic even more, with basic explanatory data analysis done, and models trained.

Keywords: Time Series, ARIMA, ARDL, PM2.5, SARIMA, Predictive Models, Econometrics, EDA

Introduction

The aim and foundation of the analysis.

Nowadays, air pollutants play significant role in our climate change debate. World organizations like WHO, Greenpeace etc., tries to persuade governments to lead a change in green energy, and reusable production, to diminish the impact of our non-ecofriendly behavior. Trivial is fact that, to make a change, there is a need to research, both in terms of gathering and analyzing data of any kind, which can help us understand where lies the problem, and how to effectively build prevention methods. Hence, as data about air pollutants gathered is of enormous size, we will focus only on one of their family, observed in a city of Shanghai. We will analyze, a family of Atmospheric aerosol particles - precisely: Particulate Matter 2.5, which stands for „fine particles with a diameter of 2.5 μm or less". We will analyze it occurrence from the year 2011 to 2016, with underlying aim to build predictive model which can help us see in which direction the pollution is going. As a good manner we will test two models for this Time Series Data Set, which will be ARDL and ARIMA, we will provide show explanatory data analysis, then build a model and compare the results. At the end we will describe our findings, and focus on the future usage of this analysis for different types of pollutants.

A. Useful definitions (taken entirely from internet, as we are not experts in this matter):

Particulate Matter (PM) is the generic term used for the mixing of solid and liquid particles in the air. Based on its origin, suspended particulates in the atmosphere are classified as primary and secondary. The primary particles are emitted directly by the source, such as those

caused by the re-suspension of dust and the combustion in diesel vehicles, and the secondary particles are formed by chemical reactions in the atmosphere, such as sulphates and nitrates, which are formed from the combustion gases that react with water vapour[1].

Knowledge of particle diameter is important because it is an indication of its source and the place of its deposition in the respiratory tract (Marques, 2000). Atmospheric particles have diameters typically within the 0.001 to 500 μm size range. The larger particles are less harmful to health, since they deposit in the soil very quickly due to the effect of gravity. The same does not happen with the smaller particles (below 10 μm), which are small enough to remain in suspension in the atmosphere and there remain for hours or even days. Therefore, they can travel considerable distances from the emission source (Seinfeld and Pandis, 1998).

Owing their more pathogenic nature, inhalable particles are defined by the US-EPA (United States Environmental Protection Agency) in a number of fractions depending on their mean size (EPA, 2003). Of special importance to human health, the fraction defined as PM10 is sub-divided in two sub-fractions:

- PM2.5: Particles with mean aerodynamic diameter of 2.5 μm or less (fine particles), which can bypass the respiratory defences and invade the lung parenchyma, causing inflammatory responses of the body;
- PM10–2.5: Particles with mean aerodynamic diameter of 10 μm or less (coarse particles) from which the PM2.5 fraction is removed. As such, despite being in the respirable diameter range, they are generally retained in the upper airways.

[1] https://www.researchgate.net/publication/297929542_Time_series_analysis_of_PM25_and_PM10-25_mass_concentration_in_the_city_of_Sao_Carlos_Brazil

In addition to health problems, the PM harms the environment, altering nutrients and chemical balances in water bodies, and can cause erosion and stains in structures and monuments. According to Cappiello (2002), PM is the main cause of visibility reduction in many parts of the USA, reducing levels of solar radiation that reach the ground. As a result of this reduction in solar radiation, the particles change the soil temperature and influence negatively the growth of plants[2].

B. Used Models Overview (general description based on internet sources):

ARIMA: these models are, the most general class of models for forecasting a time series, which we can made to be "stationary" by differencing or in conjunction with nonlinear transformations such as logging or deflating. A random variable that is a time series is stationary if its statistical properties are all constant over time. i.e., its short-term random time patterns always look the same in a statistical sense. The latter condition means that its *autocorrelations* (correlations with its own prior deviations from the mean) remain constant over time, or equivalently, that its *power spectrum* remains constant over time.
ARIMA is also combinations of AR I, and MA processes, which stands for as follows, Autoregressive, Integrated and Moving Average processes which gives us Autoregressive Integrated Moving Average Process.

[2] up

ARIMA(p,d,q) process is often given by:

A $$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

nonseasonal **ARIMA(p,d,q)** model consists of:

- **p** which is the number of autoregressive terms,
- **d** which is the number of nonseasonal differences needed for stationarity, and
- **q** which is the number of lagged forecast errors in the prediction equation.

Full reference to deep insight about ARIMA models can be found here: https://people.duke.edu/~rnau/411arim.htm

ARDL: Autoregressive distributed lag model is used to model relationship between variables in a single-equation time-series setup. Its often given by:

$$y_t = c_0 + c_1 t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=0}^{q} \beta_i' \mathbf{x}_{t-i} + u_t,$$

In distributed lag models, both **y** and **x** are typically random. That is, we do not know their values prior to sampling. To accommodate for this stochastic process, we assume that the **x's** are random and that the error term **e** is independent of all **x's in the sample - past, current, and future**. This assumption, in conjunction with the other multiple regression assumptions, is sufficient for the least squares estimator to be unbiased.

The assumptions of the distributed lag model are

1. $y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_q x_{t-q} + e_t$
2. y and x are stationary random variables, and e_t is independent of current, past, and future values of x.
3. $E(e_t) = 0$
4. $\text{var}(e_t) = \sigma^2$
5. $\text{cov}(e_t, e_s) = 0 \qquad t \neq s$
6. $e_t \sim N(0, \sigma^2)$

Full reference to deep insight about ARDL models can be found here: http://rstudio-pubs-static.s3.amazonaws.com/243419_0a7cebeb66b049fea263f8ce72590afd.html

Literature Overview

Brief introduction to the topic based on the literature used.

Along this paper, our aim was to predict future levels of PM2.5 in the city of Shanghai, basing on the data from 2012 - 2016. However, first we had to clarify our thesis, aims and statements, finding possibly similar research along the paper already published. We were using them as a basis to our research, and in some cases reflecting to solutions, proposed by the authors.

First of all, we found it crucial to use either ARDL or ARIMA model, when it comes to analysis of PM2.5. Most of papers stick to this pattern, and show great results. However, usage of ARDL is not very popular among researchers due to the fact, that the data on which they build their models, contains only level of PM2.5 and Date. Our Data however contains multiple different variables like: temperature, air pressure etc, per date, which were used as independent variable in our analysis, to check wether exogenous variables like this can reflect on higher level of PM2.5 in the future.

In each paper there was a comparison between at least two models performance, in general the models presented were either: ARIMA, SARIMA, ARDL or Neural Network Models. Most of this articles revealed that either ARIMA, SARIMA or ARDL can be used for similar studies elsewhere. However, first paper pointed that „if the best possible representation is pursued, the more complex SARIMA approach is probably the preferred choice". Papers also showed that statistical analysis of the parameters and the residuals analysis should not be discarded, because even with a lower value of AIC, the model can still present insignificant parameters or correlation over time. They noted also the fact that time series models have a limited period of time for forecasting (up to 12 future observations is a reliable number of forecasts).

Data Overview

Our Data Set has been taken from: https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities

Data gathers 52854 observations for each Chinese biggest City from 2010 to 2016. It has 14 variables with explanation which goes as follows:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- season: season of data in this row
- PM: PM2.5 concentration (ug/m^3)
- DEWP: Dew Point (Celsius Degree)
- TEMP: Temperature (Celsius Degree)
- HUMI: Humidity (%)
- PRES: Pressure (hPa)
- cbwd: Combined wind direction
- Iws: Cumulated wind speed (m/s)
- precipitation: hourly precipitation (mm)
- Iprec: Cumulated precipitation (mm)

Missing data are denoted by NA's. As for the clear analysis, we've omitted those dates where NA's exist, so in final gave us data set of 31 876 observations from dec.2011 to dec.2015.

The original dataset, retrieved from UCI Machine Learning Repository, was collected for 5 Chinese cities and consists of hourly observations between January 1st, 2010 to December 31st, 2015. It contains information about PM2.5 concentration, as well as meteorological details like dew point, temperature, humidity, pressure, combined wind direction, cumulated wind speed, hourly precipitation and cumulated precipitation. Our analysis focuses on Shanghai pollution problem with the complete observations gathered for the period of Dec 28th, 2011 to Dec 31st, 2015. The dataset lacks values for 24 days - between Jul 1st, 2014 and Jul 25th, 2014.

Table 1. Summary of the data.

	mean	sd	median	mad	min	max	range	skew	kurtosis	se
PM 2.5 conc.	53.41	43.07	42.0	28.17	1.0	730.0	729.0	2.59	13.88	0.24
dew point	11.02	9.69	12.0	11.86	-17.0	28.0	45.0	-0.27	-0.97	0.05
humidity	69.43	17.78	72.0	18.12	13.1	100.0	86.9	-0.61	-0.29	0.10
pressure	1016.14	8.99	1016.0	10.38	990.0	1040.0	50.0	0.04	-0.93	0.05
temperature	17.23	9.23	18.0	11.86	-4.0	41.0	45.0	-0.05	-0.96	0.05
cumul. wind speed	50.74	72.43	21.0	26.69	0.0	691.0	691.0	2.76	10.59	0.41
precipitation	0.15	1.08	0.0	0.00	0.0	61.6	61.6	21.43	735.72	0.01
cumul. precip.	1.01	7.44	0.0	0.00	0.0	226.4	226.4	17.86	431.56	0.04

All of the variables are rather highly dispersed. In Shanghai, during the observed period, the concentration of fine particles with a diameter of 2.5um ranged from almost non-existent (1 ug/m^3) to very large numbers (730 ug/m^3). The mean PM2.5 concentration for each year exceeds 50 ug/m^3 (with the value of 60 ug/m^3 in 2013), which is more than twice as much as standard threshold adopted in other countries[3]. Furthermore, the values recorded in 2014 and 2015 are also higher than the annual averages for China[4].

[3] European Commission (http://ec.europa.eu/environment/air/quality/standards.htm), United States Environmental Protection Agency (https://www3.epa.gov/region1/airquality/pm-aq-standards.html)

[4] Lei Jiang et al., "Comparison of Ground-Based PM2.5 and PM10 Concentrations in China, India, and the U.S.", *International Journal of Environmental Research and Public Health 15*, no.7 (2018)

Skewness results indicate that 3 variables (dew point, pressure and temperature) are normally distributed. The other ones, except humidity, are highly negatively skewed.

According to Figure 1, PM2.5 concentration intensifies in winter and follows a pattern with relatively high peaks at the beginning of each year (one or two first months) and significant drop near July, which implies seasonality. Such behavior is consistent with already conducted studies[5]. Starting from 2013, from year to year, the highest values seem to be decreasing, which is confirmed in the means. At the critical point PM2.5 concentration reaches the value of over 600 ug/m^3. The linear part visible after the 6th month in 2014 is associated with formerly mentioned lack of data.

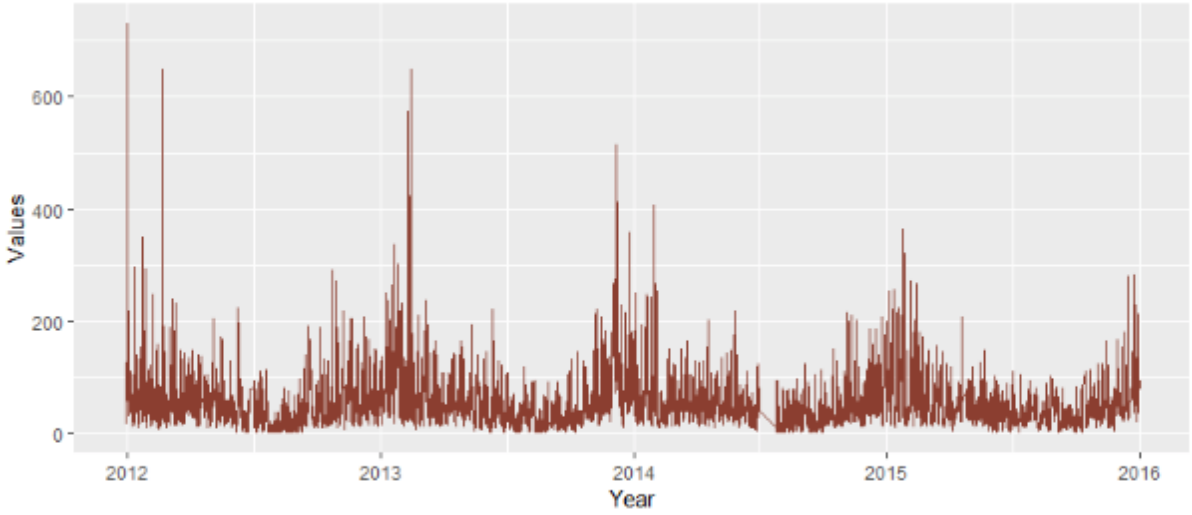


Figure 1. Plot of PM 2.5 concentration.

Figure 2 shows, among others, the graphs for dew point and temperature. It can be noticed that they follow a similar pattern - minimum values obtained at the beginning and high peaks, more or less, in the middle of each year. This dependency can be easily explained by looking at the Magnus formula for dew

[5] Wei Meng et al., "Seasonal and diurnal variations of ambient PM2.5 concentration in urban and rural environments in Beijing", *Atmospheric Environment 43*, No.18 (2009), Yingjun Chen et al., Characteristics of organic and elemental carbon in PM2.5 samples in Shanghai, China", *Atmospheric Research 92*, No.4 (2009)

point calculation[6], which considers two components: actual air temperature and relative humidity. Having that in mind, it becomes obvious that the dew point might be highly correlated with both variables. It can be further verified using the Pearson correlation. In the case of dew point and temperature, it is 88%, which confirms the presumption of correlation. For humidity, value is lower and equals to, around, 42%. Pressure is another variable highly dependent on temperature, thus, showing also the relation with dew point. Here, we can examine negative correlation - peaks in the middle of each year, visible on the pressure graph, correspond to downs for two other variables.

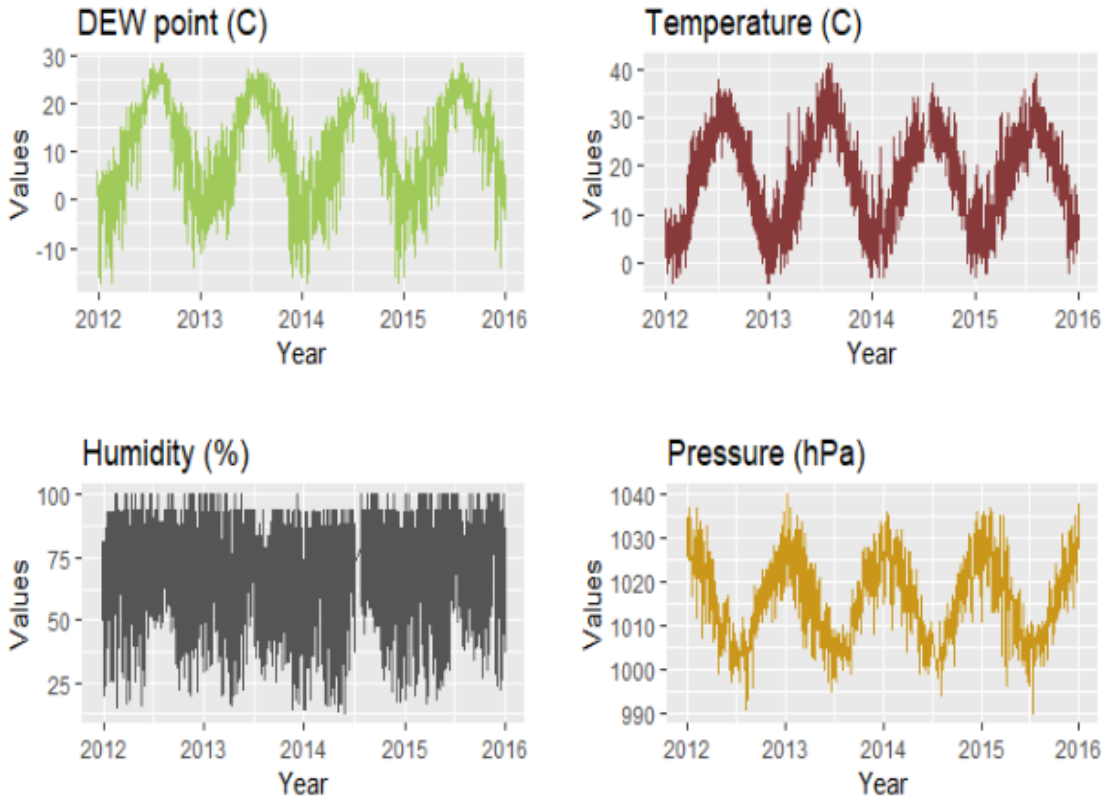


Figure 2. Plots of dew point, temperature, humidity and pressure

Cumulated and hourly precipitation are, not surprisingly, following a similar pattern (Figure 3). The correlation between those two variables is equal to 98%, which means that 98% of the time they move in the same direction. The maximum value recorded for this meteorological variable was equal to 61.6

[6] Wikipedia (https://en.wikipedia.org/wiki/Dew_point#Calculating_the_dew_point)

mm (for the cumulated precipitation 226.4 mm). The highest fallouts were denoted during spring/summer period with relatively small values occurring in winter. Such behavior is consistent with other studies investigating the relationship between seasons and meteorological conditions in Shanghai. They emphasize that precipitation is one of the main determinants of PM2.5 concentration at this time of the year[7].

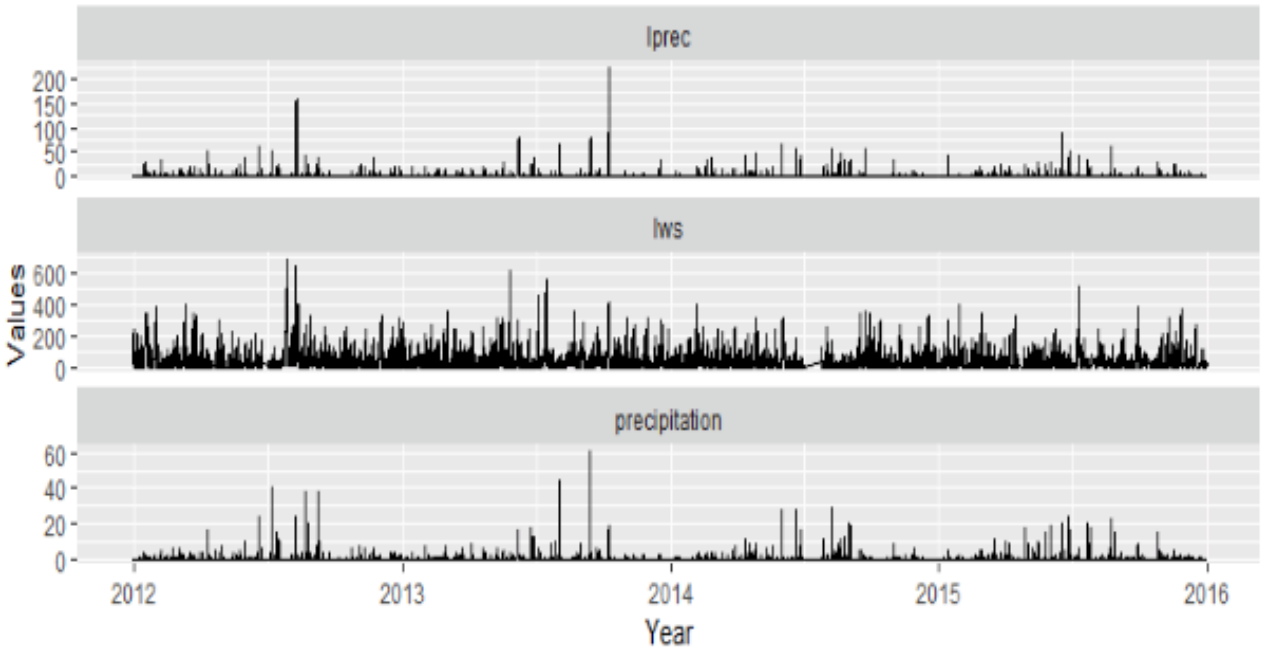


Figure 3. Plots of cumulated precipitation, cumulated wind speed and hourly precipitation.

The accumulated wind speed accrued and declined, more or less, randomly with a few outstanding peaks in the 2nd half of 2012, 2013 and 2015. The mean value for the whole analyzed period was equal to 50.74 m/s. Based on the graph analysis there is no certainty about the seasonality.

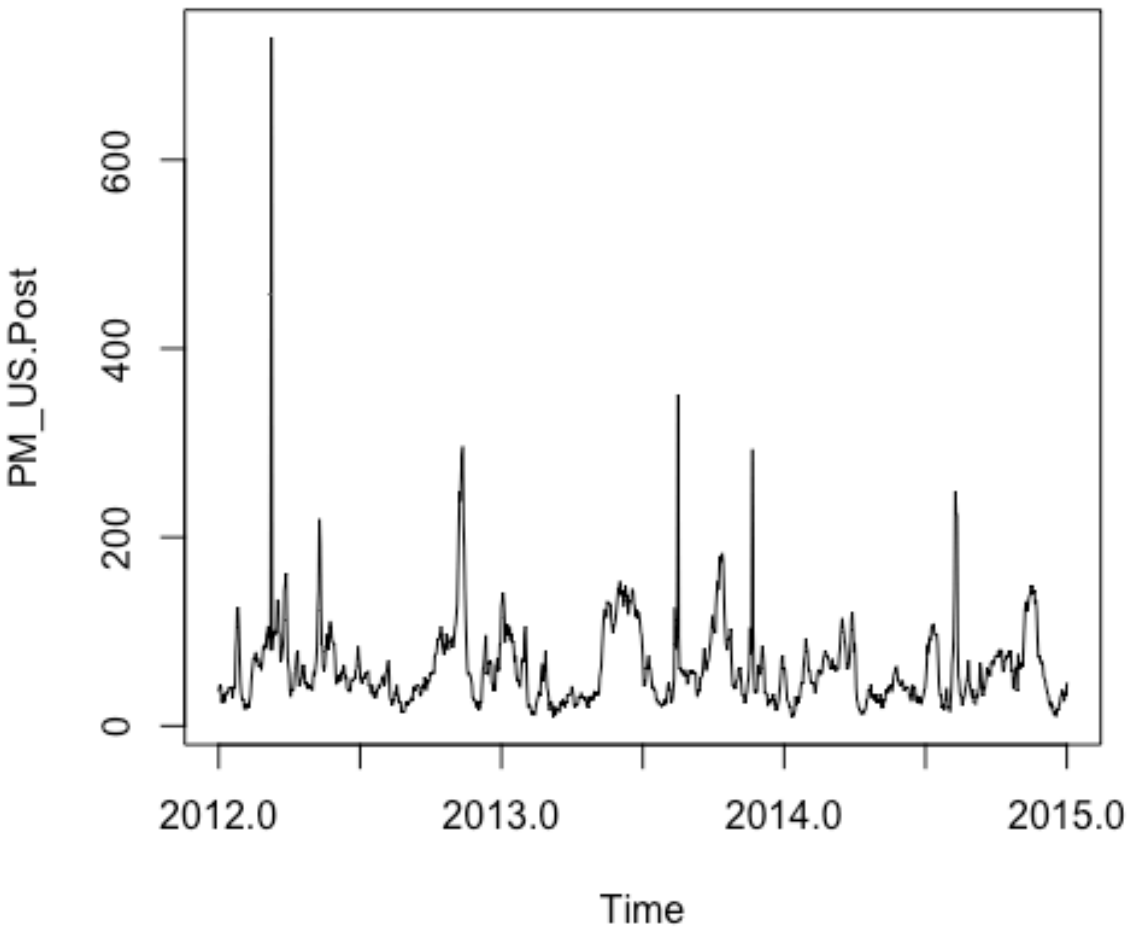
[7] Danlu Chen et al., "Understanding meteorological influences on PM2.5 concentrations across China: a temporal and spatial perspective", *Atmospheric Chemistry and Physics 16*, No.8 (2018)

ARIMA/SARIMA Models

Seasonal Data and R approach

During our ARIMA model estimation, we were forced to use SARIMA model, because of multi-seasonality in our data. We performed HEGY test (Hylleberg, Engle, Granger and Yoo Test for Seasonal Unit Roots) to make sure about this multi seasonality. We've found out that in our data there are at least 3 different seasonalities: daily, weekly, monthly and even Yearly. Unfortunately, as data set is quite big, and it took a lot of time to perform these seasonality tests, to avoid enormous time complexity, we had to adjust our data and make it daily not hourly. Distribution of our PM2.5 variable is shown on the plot 1.

Plot 1. Distribution of PM2.5 over time

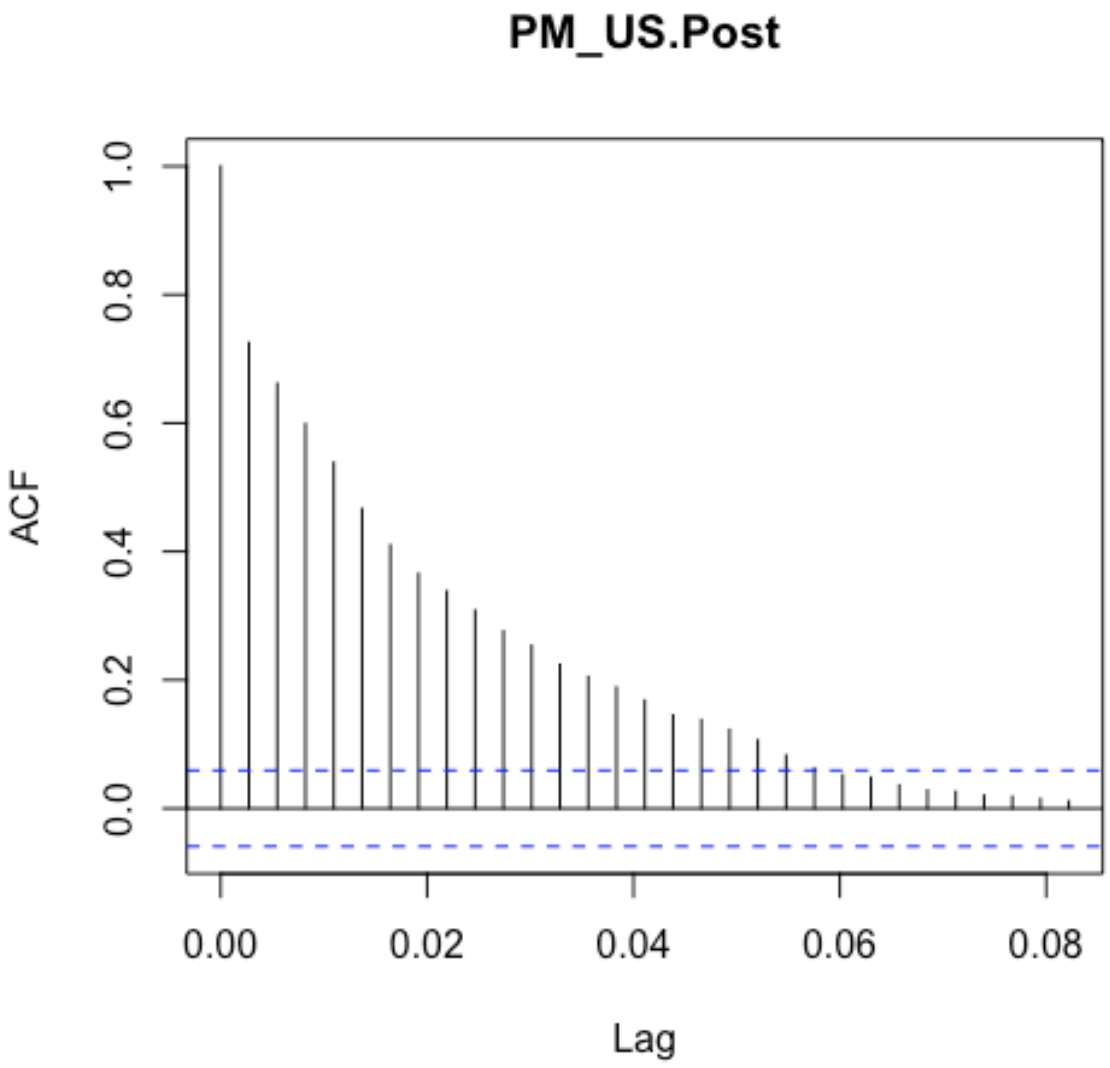


HEGY test on the daily set gave us results about overall seasonality that is statistically significant. It appeared that only daily seasonality is still significant. We've also performed tests for stationarity, which indicated that our process is stationary and doesn't need differentiation.

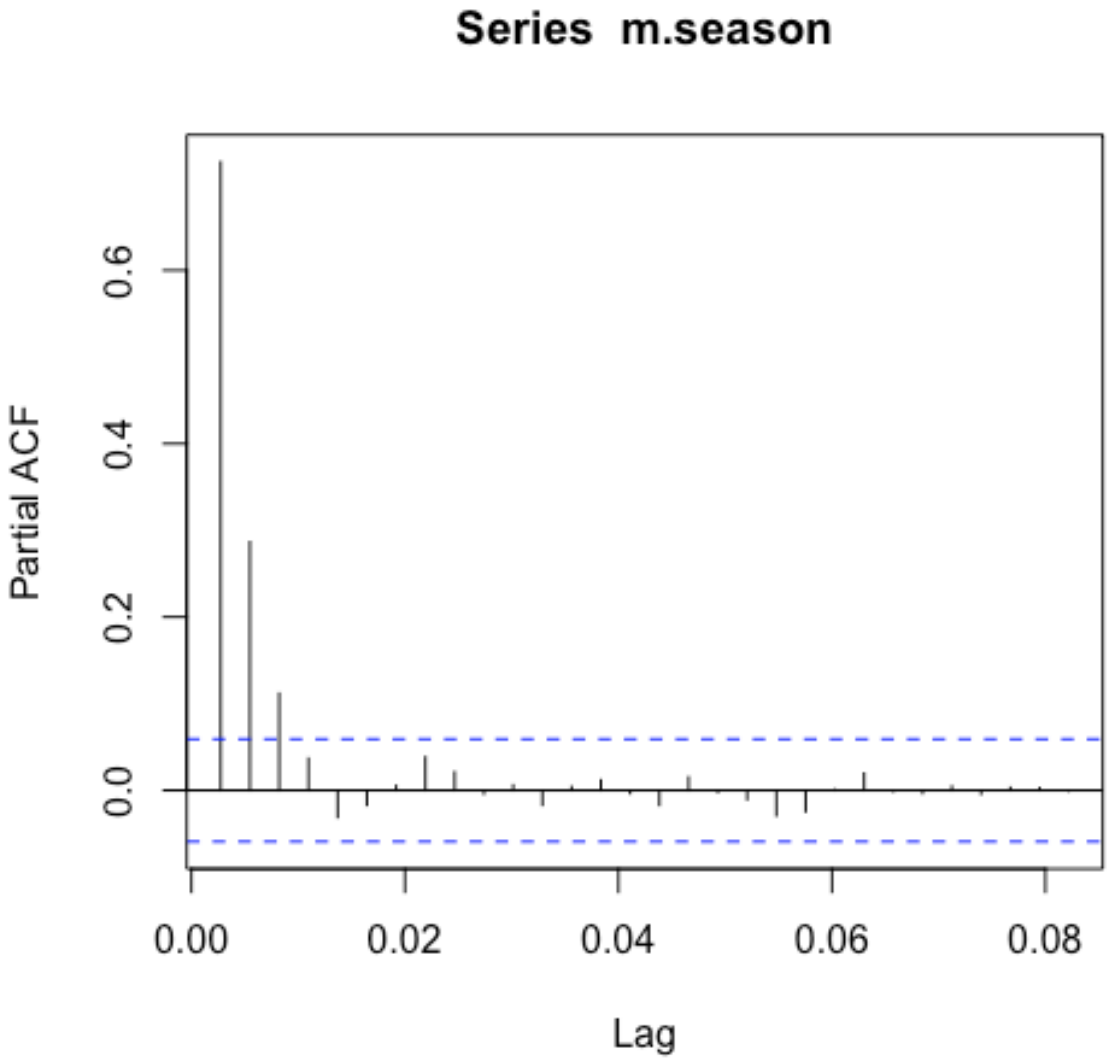
Yet, when it comes to seasonality, as we use SARIMA model, we won't get rid of it by our own - step by step. Instead we will use R-cran TBATS and auto.arima functions to handle it automatically.

Our PACF and ACF graphs of this process is presented on Plot 2. And Plot 3.

Plot 2. ACF

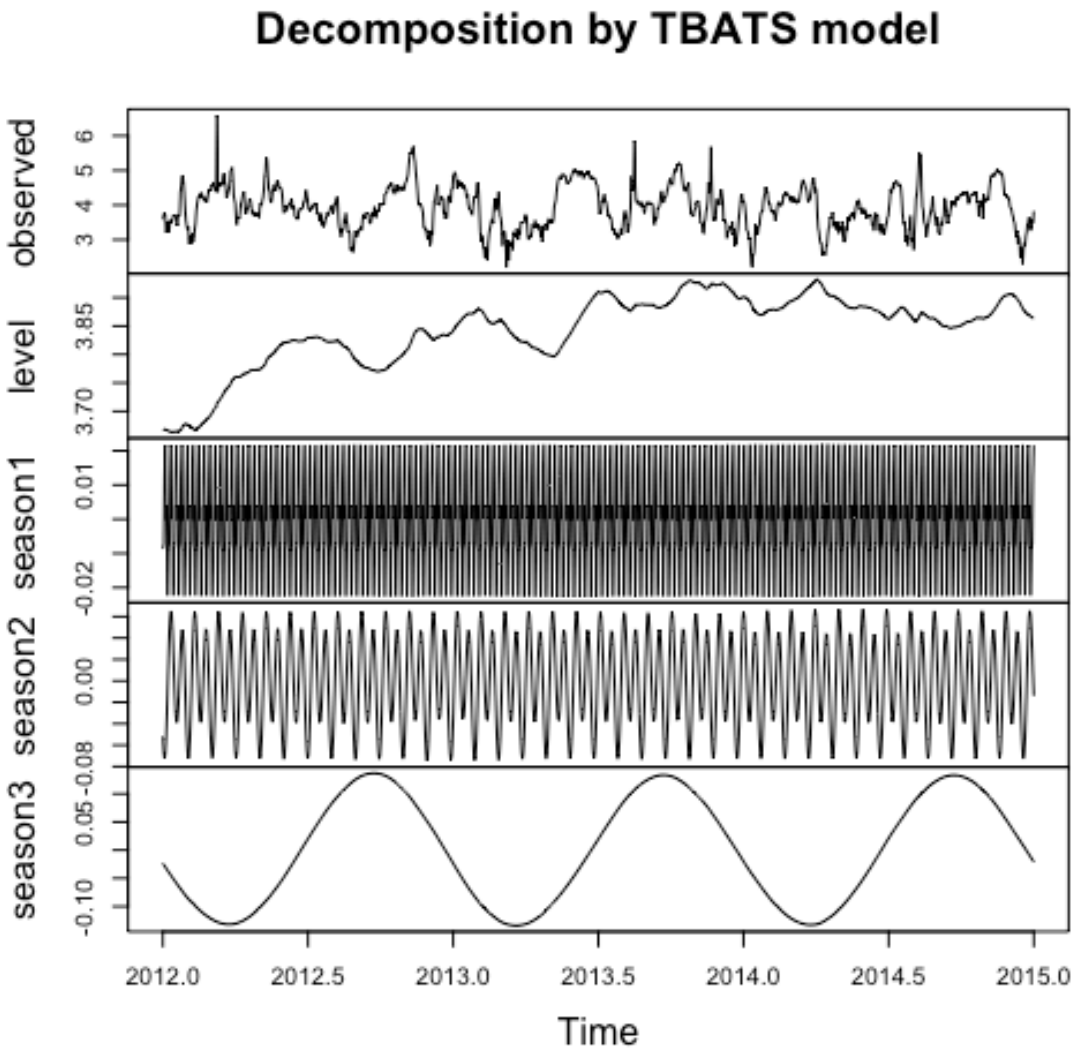


Plot 4.



In Plot 4. We presented how TBATS decomposition of our data look like when it comes to seasonality. We can see clearly that it appears to be 3 seasonalities, Weekly, Monthly and Yearly. As we are limited by computing power, and our PC's efficiency, we get rid of hourly data to make it easier to analyze. Next we will try to make a one year in future prediction basing on our fitted SARIMA model, and check how the plot for this data looks like. We will use combination of SARIMA and LinearModels to make our predictions more efficient and robust to drifts and seasonality.

Plot 5.

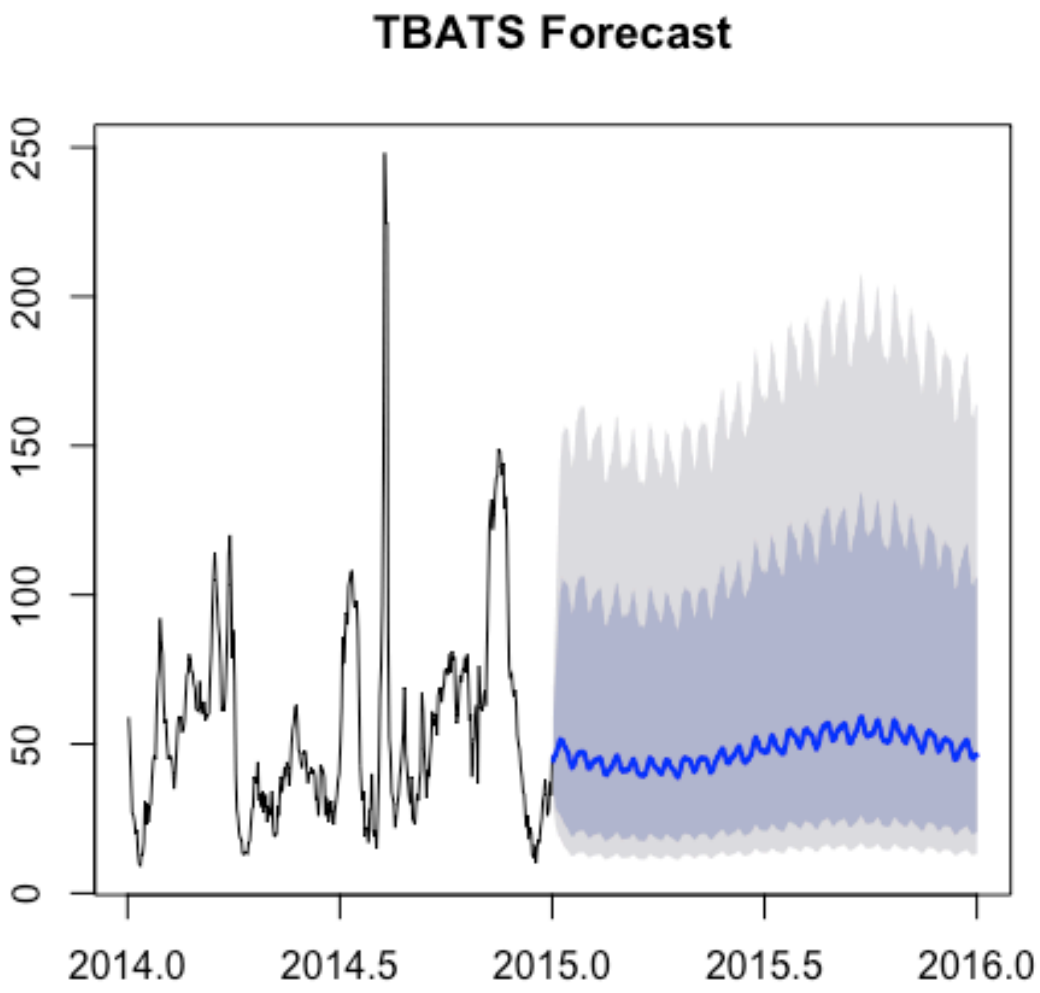


As far as we performed Auto-Arima on our seasonal data set, we had to set parameter seasonal to TRUE. The output of this function resulted in best model which appeared to be ARIMA(1,0,1).

```
Series: m.season
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1      ma1     mean
      0.9027  -0.4091  61.6392
s.e.  0.0162   0.0335   5.3719

sigma^2 estimated as 872.8:  log likelihood=-5265
AIC=10538   AICc=10538.04   BIC=10558

Training set error measures:
                      ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -0.005956944 29.50209 12.09796 -11.47822 22.30402 0.2692528 -0.01123176
```

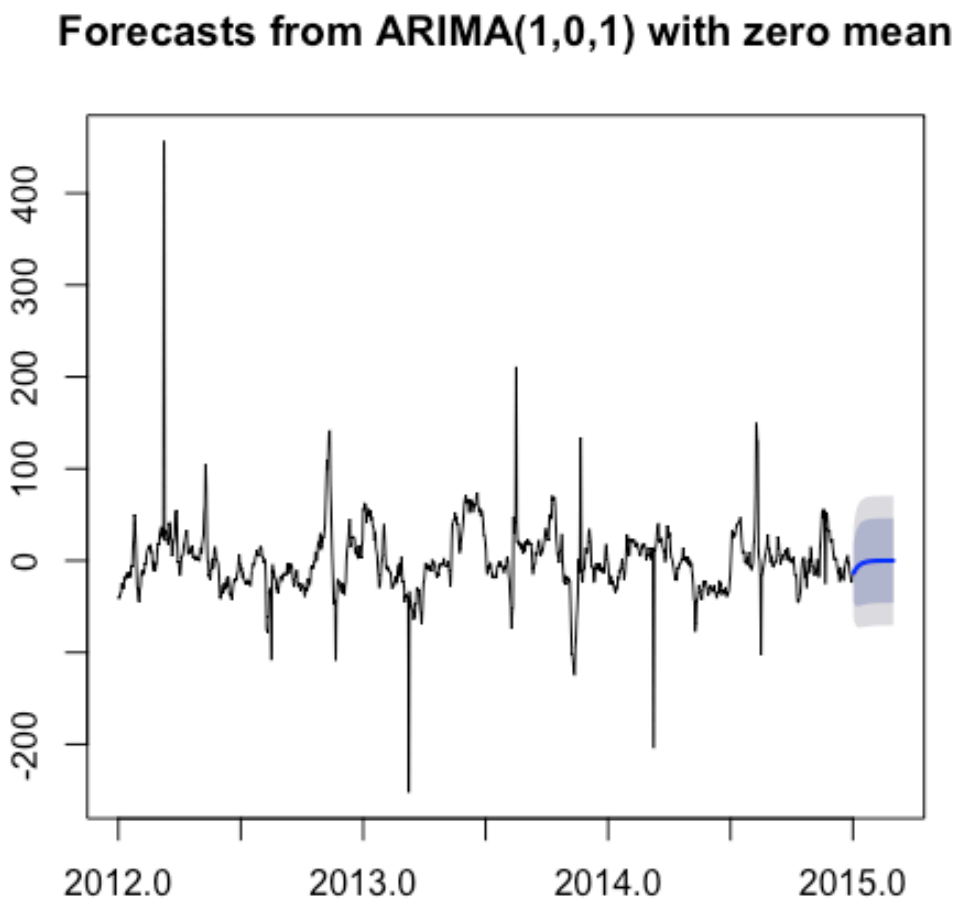
Having in mind our SARIMA results, first we made independently automatic TBATS forecast for next year. The results is shown in Plot 6. As we can conclude it oscillate around the mean of this time series, and shows us a 1% and 5% confidence interval. We can see though that process is indeed stationary.

Plot 6.

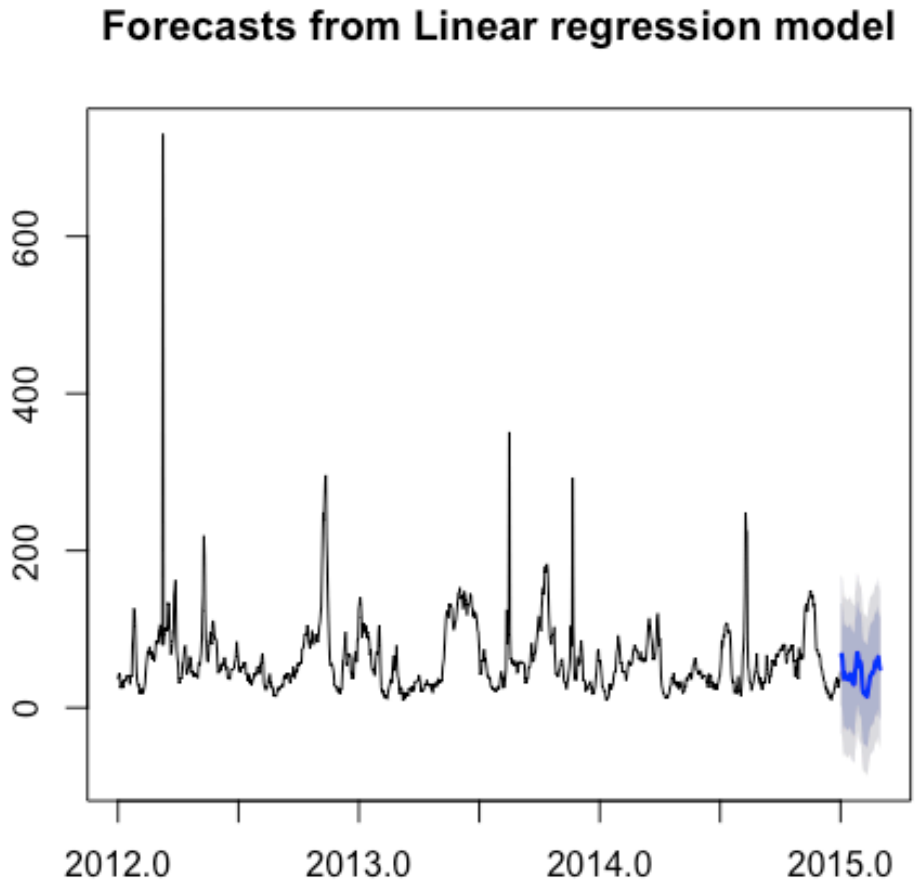


After, we've fit our time series to time series linear model, using daily data. We will use it residuals to strengthen our SARIMA model in terms of prediction power. Yet, this time we will make prediction only for next 60 days to be more precise. Plots 7 and 8 respectively shows forecast using ARIMA and LinearModels.

Plot 7.

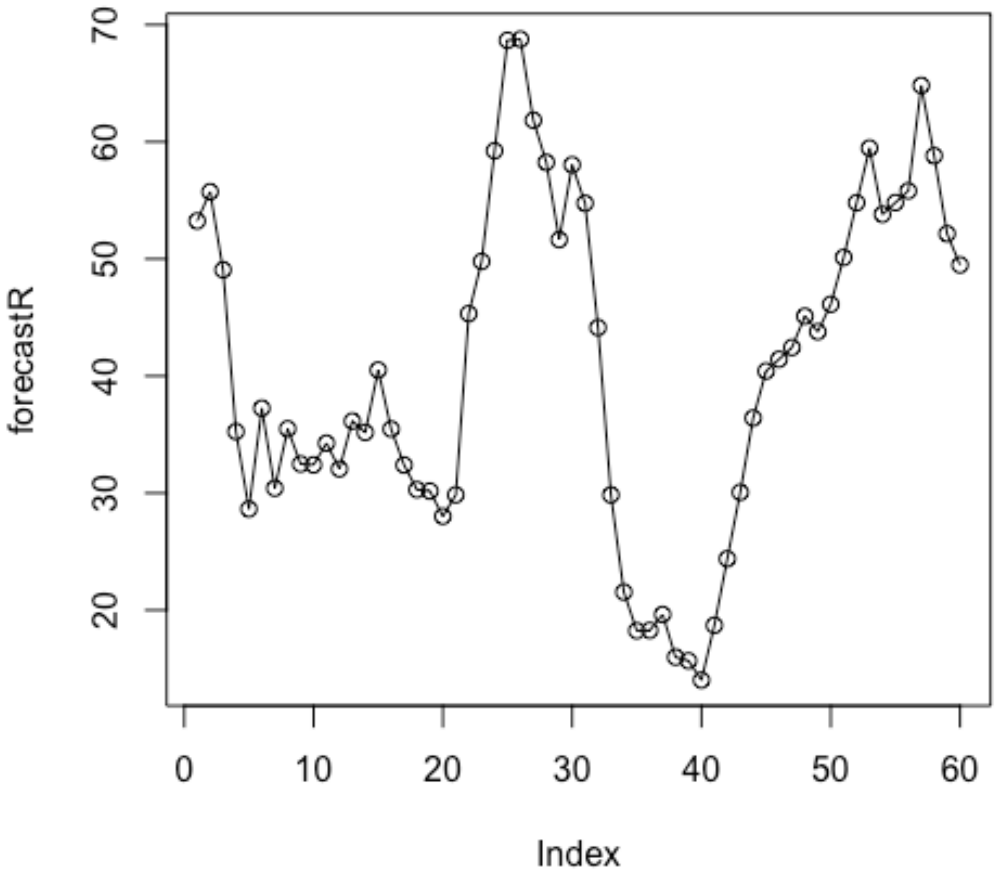

Forecasts from ARIMA(1,0,1) with zero mean
400
300
200
100
0
-200
2012.0
2013.0
2014.0
2015.0

Plot 8.


Forecasts from Linear regression model
600
400
200
0
2012.0
2013.0
2014.0
2015.0

Next we will use combined residuals to make graph for the next 60 days prediction. This is shown on the Plot 9. We can easily say that there is a significant pattern with ups and downs, and it resembles natural deviations from mean which is in our case - 40.

Plot 9.



In summary, we can say that it's possible to model seasonal data in R, but it's quite hard task. It's recommended in here to make use of TBATS from „uroot" library in R, and also of „auto.arima" to make the process fast and efficient. However, dealing with hourly data is time and memory consuming, thus for this kind of analysis we skipped this part, as execution of some loops and methods took approximately few hours…

ARDL Models

Seasonal Data and R approach

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.

```
dynlm( PM_US.Post ~
          L(PM_US.Post, c(1, 24 * 7, 24 * 30, 24 * 365)) +
          L(DEWP, c(0, 1, 24 * 7, 24 * 30, 24 * 365)) +
          L(HUMI, c(0, 1, 24 * 7, 24 * 30, 24 * 365))+
          L(PRES, c(0, 1, 24 * 7, 24 * 30, 24 * 365))+
          L(TEMP, c(0, 1, 24 * 7, 24 * 30, 24 * 365))+
          L(Iws, c(0, 1, 24 * 7, 24 * 30, 24 * 365))+
          L(precipitation, c(0, 1, 24 * 7, 24 * 30, 24 * 365))+
          L(Iprec, c(0, 1, 24 * 7, 24 * 30, 24 * 365)), data = data2) %>% summary()
```

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

```
hourlyARDL <- dynlm( PM_US.Post ~
          L(PM_US.Post) +
          L(DEWP, c(0, 1, 24 * 365)) +
          L(HUMI, c(0, 1))+
          L(PRES, c(24 * 30))+
          L(TEMP, c(0, 1, 24 * 7))+
          Iws, data = data2)
hourlyARDL %>% summary()
```

Fig. 2 Code for the final hourly ARDL model

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    57.901037  16.891416   3.428 0.000609 ***
L(PM_US.Post)                   0.927811   0.002456 377.722  < 2e-16 ***
L(DEWP, c(0, 1, 24 * 365))1     2.291007   0.242486   9.448  < 2e-16 ***
L(DEWP, c(0, 1, 24 * 365))2    -1.980918   0.242290  -8.176 3.09e-16 ***
L(DEWP, c(0, 1, 24 * 365))3    -0.052323   0.017418  -3.004 0.002668 **
L(HUMI, c(0, 1))1              -0.308803   0.064450  -4.791 1.67e-06 ***
L(HUMI, c(0, 1))2               0.208208   0.064612   3.222 0.001273 **
L(PRES, c(24 * 30))            -0.041910   0.015971  -2.624 0.008693 **
L(TEMP, c(0, 1, 24 * 7))1      -1.172820   0.265992  -4.409 1.04e-05 ***
L(TEMP, c(0, 1, 24 * 7))2       0.839164   0.267485   3.137 0.001708 **
L(TEMP, c(0, 1, 24 * 7))3      -0.064725   0.023478  -2.757 0.005842 **
Iws                            -0.007911   0.001348  -5.868 4.46e-09 ***
```

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

```
weeklyARDL <- dynlm( PM_US.Post~
          L(PM_US.Post) +
          DEWP +
          L(DEWP, 4) +
          PRES +
          L(TEMP)+
          Iws +
          precipitation +
          L(precipitation, 4), data = weeklyData)
weeklyARDL %>% summary()
```

Fig. 6 Code for the final weekly ARDL model

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1951.80949  458.44929   4.257 3.21e-05 ***
L(PM_US.Post)         0.23934    0.06683   3.582 0.000431 ***
DEWP                 -1.31988    0.44752  -2.949 0.003574 **
L(DEWP, 4)           -0.77995    0.33913  -2.300 0.022516 *
PRES                 -1.83341    0.44634  -4.108 5.88e-05 ***
L(TEMP)              -0.96663    0.41272  -2.342 0.020184 *
Iws                  -0.16905    0.03493  -4.839 2.65e-06 ***
precipitation       -15.02116    6.80036  -2.209 0.028348 *
L(precipitation, 4)  16.12280    6.75475   2.387 0.017948 *
```

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.

Hourly

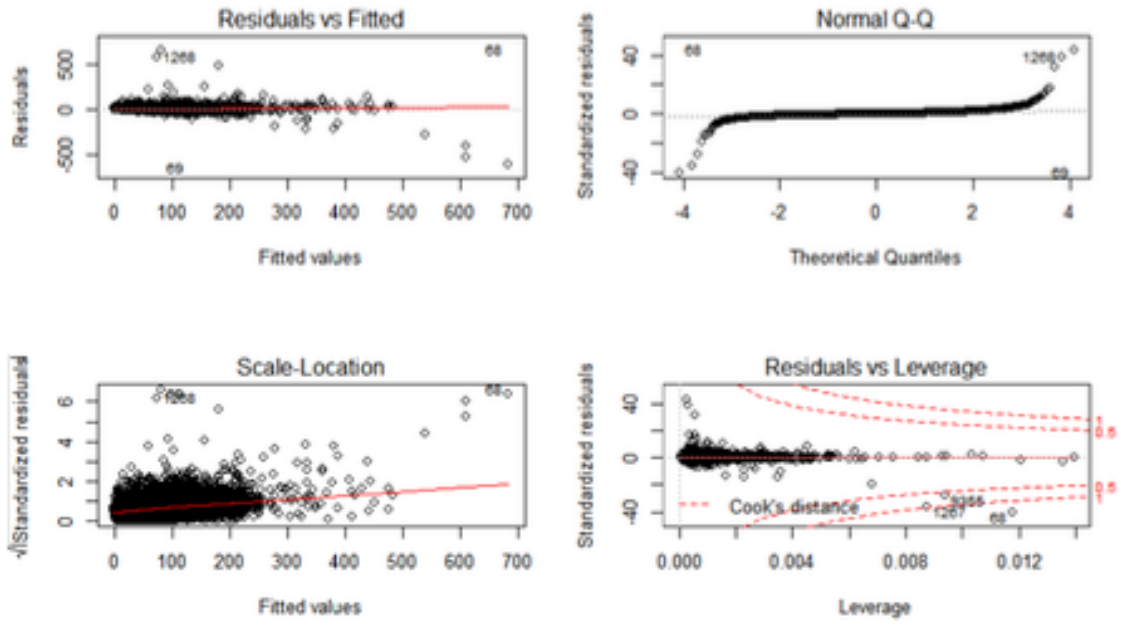


Fig. 8 Hourly model plots

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.

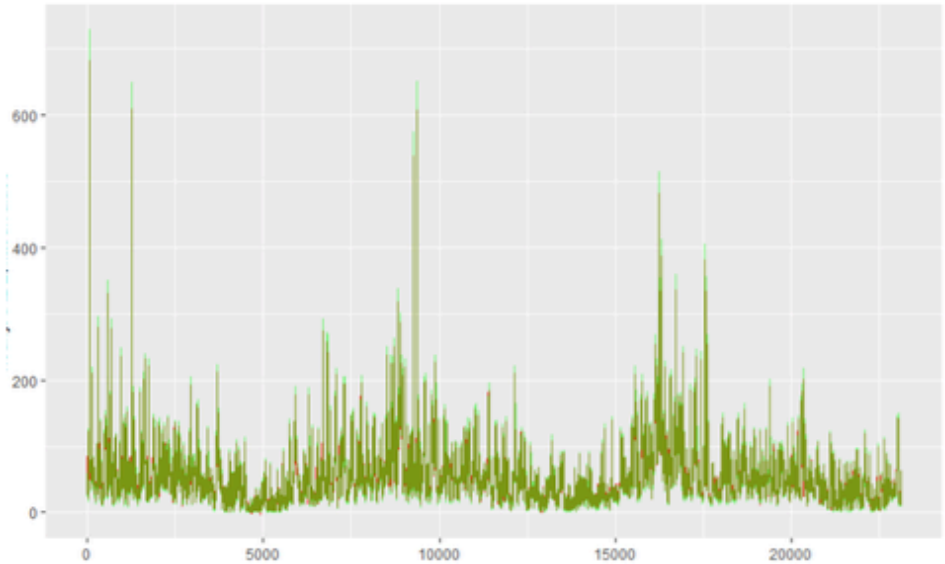


Fig. 9 Differences between real and predicted values. The plots overlap in nearly all places, producing a green color.

Daily

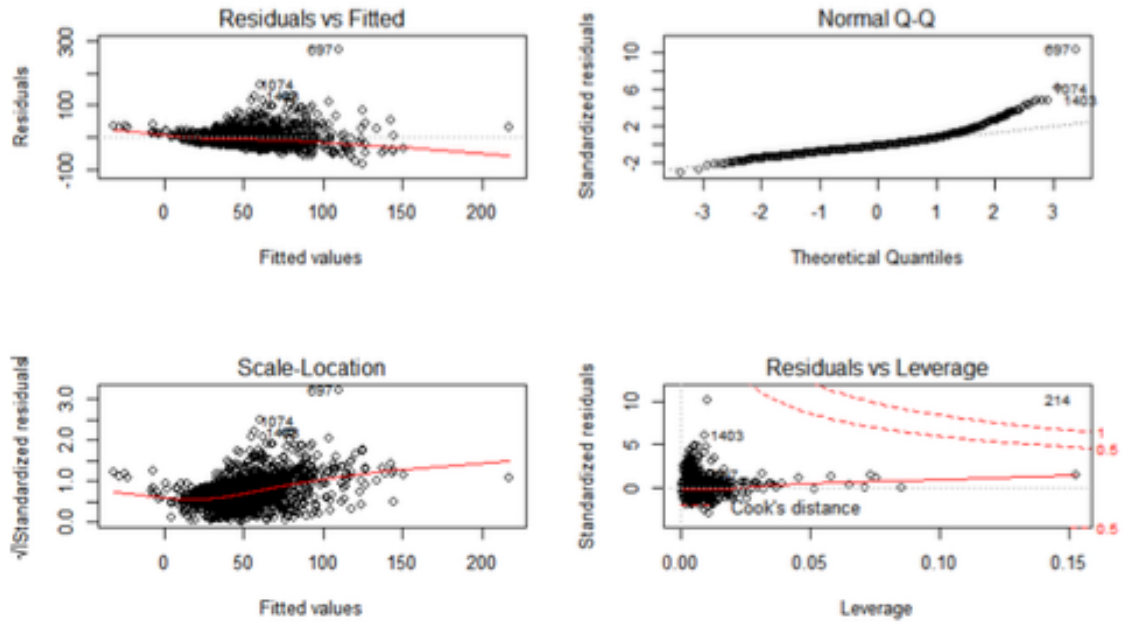


Fig. 10 Daily model plots

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).

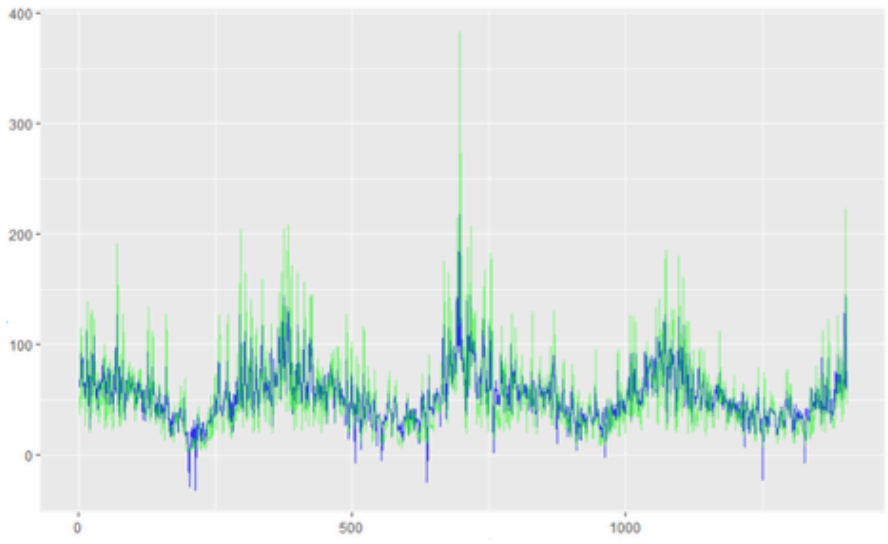


Fig. 11 Differences between real and predicted values. The real values (green) have much stronger outliers than the predicted ones (blue).

Weekly

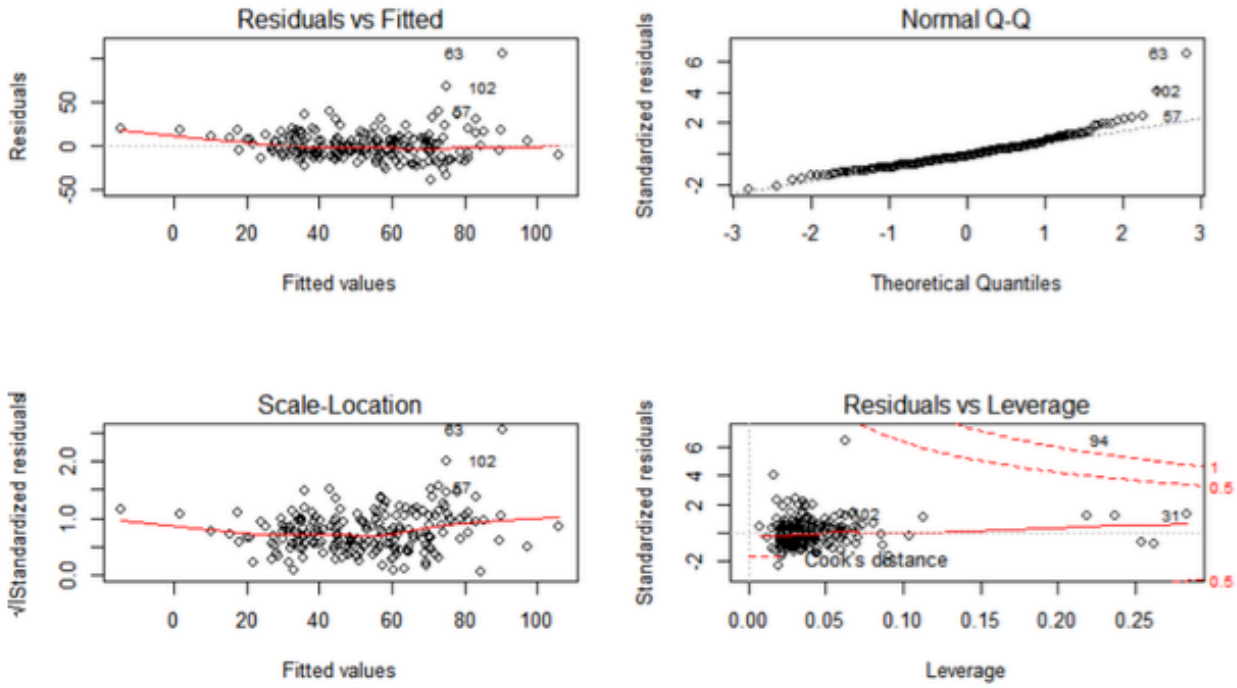


Fig. 12 Weekly model plots

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.

First ARDL Modeling Findings

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.

Yet, in the next part we will try to repeat the ARDL approach with deseasonated daily data in JDemetra.+

ARDL Models

Deseasonated Data: JDemetra+ approach

In the case of seasonally adjusted data in JDemetra+, the variables do not correlate higher than 50% and, besides cumulated precipitation, are symmetrical or moderately skewed (Appendix 1).

Table 2. Summary of ADF test with p-values for Breusch-Godfrey test for non-seasonal data.

	PM	HUMI	DEWP	TEMP	PRES	lprec	lws	prec
lags	2	1	1	0	1	2	2	2
pv bgtest (1)	0.58	0.53	0.69	0.73	0.47	0.64	0.57	0.75
pv bgtest (2)	0.24	0.60	0.79	0.28	0.24	0.73	0.77	0.68
pv bgtest (3)	0.16	0.11	0.29	0.05	0.11	0.76	0.18	0.76
pv bgtest (4)	0.08	0.19	0.19	0.10	0.18	0.69	0.15	0.70
pv bgtest (5)	0.07	0.28	0.13	0.11	0.28	0.80	0.24	0.70
t stats	-7.62	-7.02	-7.16	-10.32	-8.21	-8.23	-7.36	-8.38

According to the ADF test, calculated on differences with manually adjusted lags, all of the variables are stationary - the t-statistic is lower than the 5% critical value (-1.95), which indicates that the null hypothesis can be rejected. Lags were adjusted using the Breusch-Godfrey test for residuals serial correlation (of order from 1 up to 5) - until all p-values were bigger than 0.05 significance level.

The model selection was based on the general-to-specific approach - different combinations of lags were tested and the formula with the lowest BIC and AIC value was chosen. First, we analyzed the behavior of models with all of the variables having the same number of lags. We selected the one with the smallest values for previously mentioned criterions. Then, we tested several options by adding and subtracting the number of lags. Such deduction led to a model with 1 lag for PM2.5 concentration, dew point, cumulated precipitation

and temperature, 1 and 2 lags for cumulated wind speed and pressure and 1,2 and 3 lags for humidity and hourly precipitation. There were 7 statistically significant variables (Appendix 2). The next tested model took into account only those significant ones and eliminated another variable – the lagged dew point with a p-value larger than 10% significance level (Appendix 3).

The final model uses 6 regressors to analyze the ΔPM2.5 concentration (Table 3). The model is statistically significant, given the p-value close to 0, and 67% of the variation of the PM2.5 concentration can be explained by the input variables. According to the Breusch-Pagan test, the homoscedasticity is present (p-value = 0.1187). Furthermore, residuals are normally distributed (null for Jarque-Bera test on residuals is not rejected) and there is no autocorrelation - p-values for the Breusch-Godfrey test (up to order 5) are higher than 0.01 (Appendix 4).

Table 3. Summary of the final ARDL model for non-seasonal data

	coefficients	t-value	Pr(>|t|)
L(d(tsPM))	-0.45834	-4.865	2.03.e-05***
L(d(gtsHUMI),3)	0.44841	2.839	0.00723**
d(tslws)	-0.26884	-5.301	5.16e-06***
L(d(tsprec),2)	60.94906	5.542	2.42e-0.6***
L(d(tsPRES), c(1:2))1	0.24573	5.082	1.03e-0.5***
L(d(tsPRES), c(1:2))2	0.12627	2.796	0.00807**

Significance level: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1

According to Table 3, in the ARDL model for non-seasonal data, all of the variables are highly significant (p-values lower than at least 5% level). With a one unit increase in the differentiated cumulated wind speed (tslws), we can observe a 0.27 decrease of the dependent variable. In the case of pressure, the 2nd lag influences ΔPM2.5 concentration as well. It can be seen, that with an additional lag added, this variable has a smaller effect on the target variable.

What is interesting, the previous period value of the dependent variable has an influence on ΔPM2.5 concentration in time t_0. One unit increase in the lag of the differentiated variable PM2.5 causes a 0.46 unit decrease of this variable in examined time. Moreover, the results show that the effect of precipitation, in comparison to all other regressors, is the highest. One unit increase in precipitation increases the value of ΔPM2.5 concentration by 61. This may be due to the fact that the majority of observations for this variable takes values close to 0, more than 75% of them do not exceed 0.2, meaning that the differences are even smaller. In general, meteorological variables, that were found to be statistically significant in the ARDL model were listed as one of the most influential in various researches[8].

The model's predictability was tested using predict() function in R. It was performed on the test data, which consist of the last 15 months of the seasonally adjusted dataset. The real and predicted differences of PM2.5 concentration significantly vary from each other - for some months the predicted values are wrong about more than 20 units (Appendix 5). The R^2, which verifies the accuracy of the prediction, exceeds the accepted interval of 0 and 1, meaning that there are possible mistakes in the calculation of the prediction. This means that we are not able to draw conclusions whether using ARDL model on seasonally adjusted data is a proper prediction method.

[8] Limini Jiao et al., "Influences of wind and precipitation on different-sized particulate matter concentrations (PM2.5, PM10, PM2.5-10)", *Meteorology and Atmospheric Physics*, (2017)

Danlu Chen et al., "Understanding meteorological influences on PM2.5 concentrations across China: a temporal and spatial perspective", *Atmospheric Chemistry and Physics 16*, No.8 (2018)

Findings and Summary

The PM2.5 concentration in Shanghai during the period between 2012 to 2015 followed a significant pattern with the highest values occurring in the winter season. Moreover, the majority of the meteorological variables also showed some dependencies on a particular part of the year. Because of the strong yearly, weekly and daily seasonality we performed different ARDL models. First test and results seemed optimistic. For hourly, daily and weekly models, the predicted and actual PM2.5 concentration values were at many points the same or followed a similar pattern. However, other statistics required, so that the outcomes may be considered as significant, were not fulfilled. In the case of seasonally adjusted data (in JDemetra+), the situation was inverted. All of the ARDL assumptions, for example, homoscedasticity, normality of residuals or no autocorrelation, were met, however, the values of the prediction did not allow to draw any significant conclusions.

The comparison of Linear regression, ARIMA and SARIMA showed that the latter one was the best predictor. The results of this model were the most accurate among others.

Having that in mind, it can be said that ARDL might not be the best possible solution for seasonal data. Mostly because choosing the right amount of lags and the best version of the model requires a lot of work and, compared to other models, yields rather poor results. We think that the best choice for this kind of input is SARIMA, which automatically takes into account seasonality. It is rather not recommended to use hourly data, because it it yields high computational problems. All of the drawbacks of the ARDL might be the reason, why in most of the reviewed articles the authors use ARIMA or neural networks.

Bibliography

An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China: https://www.researchgate.net/publication/224810617_An_application_of_ARIMA_model_to_predict_submicron_particle_concentrations_from_meteorological_factors_at_a_busy_roadside_in_Hangzhou_China

Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations: http://home.deib.polimi.it/guariso/inglese/arima-vs-nn.pdf

Time series analysis of PM2.5 and PM10-2.5 mass concentration in the city of Sao Carlos, Brazil: https://www.researchgate.net/publication/297929542_Time_series_analysis_of_PM25_and_PM10-25_mass_concentration_in_the_city_of_Sao_Carlos_Brazil

Modeling and forecasting demand for electricity in New Zeland: A comparison of two approaches: https://www.mssanz.org.au/MODSIM01/Vol%203/Fatai.pdf

Forecasting Ozone Concentrations Using Box-Jenkins ARIMA Modeling in Malaysia: https://thescipub.com/pdf/10.3844/ajessp.2018.118.128

Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling: https://mpra.ub.uni-muenchen.de/39277/1/MPRA_paper_39277.pdf

Modelling Seasonality and Trends in Daily Rainfall Data: http://papers.nips.cc/paper/1429-modelling-seasonality-and-trends-in-daily-rainfall-data.pdf

Modelling Considerations in the Seasonal Adjustment of Economic Time Series: https://www.census.gov/ts/papers/Conference1983/HillmerBellTiao1983.pdf

A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000WR900012

Model-data synthesis of diurnal and seasonal CO2 fluxes at Niwot Ridge, Colorado: https://pdfs.semanticscholar.org/075c/e4a3311f62c6556cd0e44eaa80b13aed5a87.pdf

Robust Seasonal Adjustment by Bayesian Modelling: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/36299/b1833042.0001.001.pdf?sequence=2&isAllowed=y

Appendix

APPENDIX 1 Summary of seasonally adjusted dataset

	mean	sd	median	mad	min	max	range	skew	kurtosis	se
tsPM	52.55	9.48	52.45	8.53	28.90	80.32	51.43	0.71	1.26	1.35
tsDEWP	11.48	1.17	11.63	0.99	8.71	14.13	5.41	-0.18	-0.11	0.17
tsHUMI	69.67	5.98	69.56	4.48	56.82	80.64	23.82	-0.11	-0.61	0.85
tslprec	0.96	1.16	0.65	0.47	0.00	6.88	6.88	3.11	11.89	0.17
tslws	50.89	16.90	47.31	15.42	26.68	102.31	75.64	1.01	0.41	2.41
tsTEMP	17.61	0.82	17.52	0.68	16.05	20.25	4.20	0.96	2.23	0.12
tsprecipitation	0.15	0.07	0.14	0.06	0.03	0.41	0.38	1.10	2.79	0.01

APPENDIX 2 1st ARDL model

	coefficients	t-value	Pr(>|t|)
L(d(tsPM))	-0.29424	-2.086	0.048274*
d(tsDEWP)	-1.65714	-1.013	0.321709
L(d(tsDEWP))	-2.43440	-1.791	0.086529.
d(tsHUMI)	0.22140	0.807	0.428196
L(d(tsHUMI),c(1:3),1)	0.27420	1.094	0.285283
L(d(tsHUMI),c(1:3),2)	-0.26614	-1.434	0.165156
L(d(tsHUMI),c(1:3),3)	0.47688	2.329	0.029022*
d(tslws)	-0.26079	-3.809	0.000903***
L(d(tslws),c(1:2),1)	0.08873	1.204	0.240820
L(d(tslws),c(1:2),2)	0.08268	1.145	0.263886
d(tsprec)	25.60625	0.875	0.390727
L(d(tsprec),c(1:3),1)	38.24913	1.256	0.221765
L(d(tsprec),c(1:3),2)	65.98071	3.335	0.002878**
L(d(tsprec),c(1:3),3)	-9.93379	-0.502	0.620171
d(tslprec)	-1.46873	-0.914	0.370053
L(d(tslprec))	-0.53620	-0.365	0.718541
d(tsTEMP)	1.02034	-0.430	0.671240
L(d(tsTEMP))	-0.16326	-0.079	0.937382
d(tsPRES)	0.03746	0.655	0.519182
L(d(tsPRES), c(1:2))1	0.23981	4.467	0.000176***
L(d(tsPRES), c(1:2))2	0.12156	2.281	0.032141*

Significance level: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1

APPENDIX 3 The model with one insignificant variable (dew point)

	coefficients	t-value	Pr(>|t|)
L(d(tsPM))	-0.50215	-4.993	1.44e-05***
L(d(tsDEWP))	-1.00415	-1.198	0.23834
L(d(tsHUMI),3)	0.42262	2.666	0.01131*
d(tslws)	-0.28961	-5.432	3.70e-0.6***
L(d(tsprec),2)	59.57216	5.418	3.85e-0.6***
L(d(tsPRES), c(1:2))1	0.24589	5.114	9.90e-06***
L(d(tsPRES), c(1:2))2	0.13255	2.932	0.00575**

Significance level: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1

APPENDIX 4 P-values for Breusch-Godfrey test for the final model

	p-value
pv bgtest (1)	0.4887
pv bgtest (2)	0.1131
pv bgtest (3)	0.06179
pv bgtest (4)	0.1187
pv bgtest (5)	0.03619

APPENDIX 5 Comparison of real and predicted results

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
real	-7.63	-4.01	13.64	-5.94	8.89	-5.42	-9.60	-3.13	9.51	-0.38.	5.84	-9.19
predicted	-5.56	-9.19	28.27	-19.77	1.30	18.37	-20.64	5.11	-12.18	18.66	-2.38	-15.28