Carlos III University

Time series analysis

Final project

New Delhi weather data

Author:
Ignacio Garcia Sanchez-Migallon
Santiago Rodriguez Ferreras

March 18, 2020

Introduction

The weather is defined by the state of the atmosphere, referring to day to day temperatures and precipitations. It is driven by air pressure, temperatures and moisture differences, which can be attributed to many different reasons such as the sun's angle. It is clear how the weather impacts our daily life, as it imposes a set of constraints in ourselves such as the necessary clothing, or the water presence, etc. As temperatures or humidity varies, or even pressure, difficult situations such as hurricanes or heat waves can happen. As a result of this, weather forecasting has proven to be extremely useful.

Through the following link `https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data` we have obtained a data set with the objective of Weather forecasting in the city of New Delhi, India. It is composed of 5 different variables with daily observations from 2013 to 2017:

- Date: date of the recorded observation.
- Mean temperature in Celsius degrees.
- Humidity measured in grams of water vapor per cubic meter of air.
- Mean wind speed.
- Mean atmospheric pressure.

The weather in New Delhi is influenced heavily by the monsoon, with aspects of humid subtropical weather and semi-arid. The variations between summer and winter when it comes to temperature and precipitations are high. Summers are characterized by temperatures as high as 45^o C and monsoons from late June to mid-September. During Winter, temperatures drop to around 6^o-7^o on average.

With the collected data, the main purpose is to perform an analysis of the time series of each attribute, obtaining insights about New Delhi's weather and create a model that allows us to forecast the weather.

Contents

1 Graphical analysis

The aim of this part is obtaining some prior knowledge of the data by visual inspection through the usage of plots.

Despite the possible analysis inaccuracies derived from visual inspection, this naked eye analysis can be useful for some objectives. For example, fast pointing of seasonal character, getting an idea of which models can be fitted to the data and which should be discarded (can be useful for validating models), outlier detection, etc.

The graphical representation of the series that we are going to analyze is presented below:

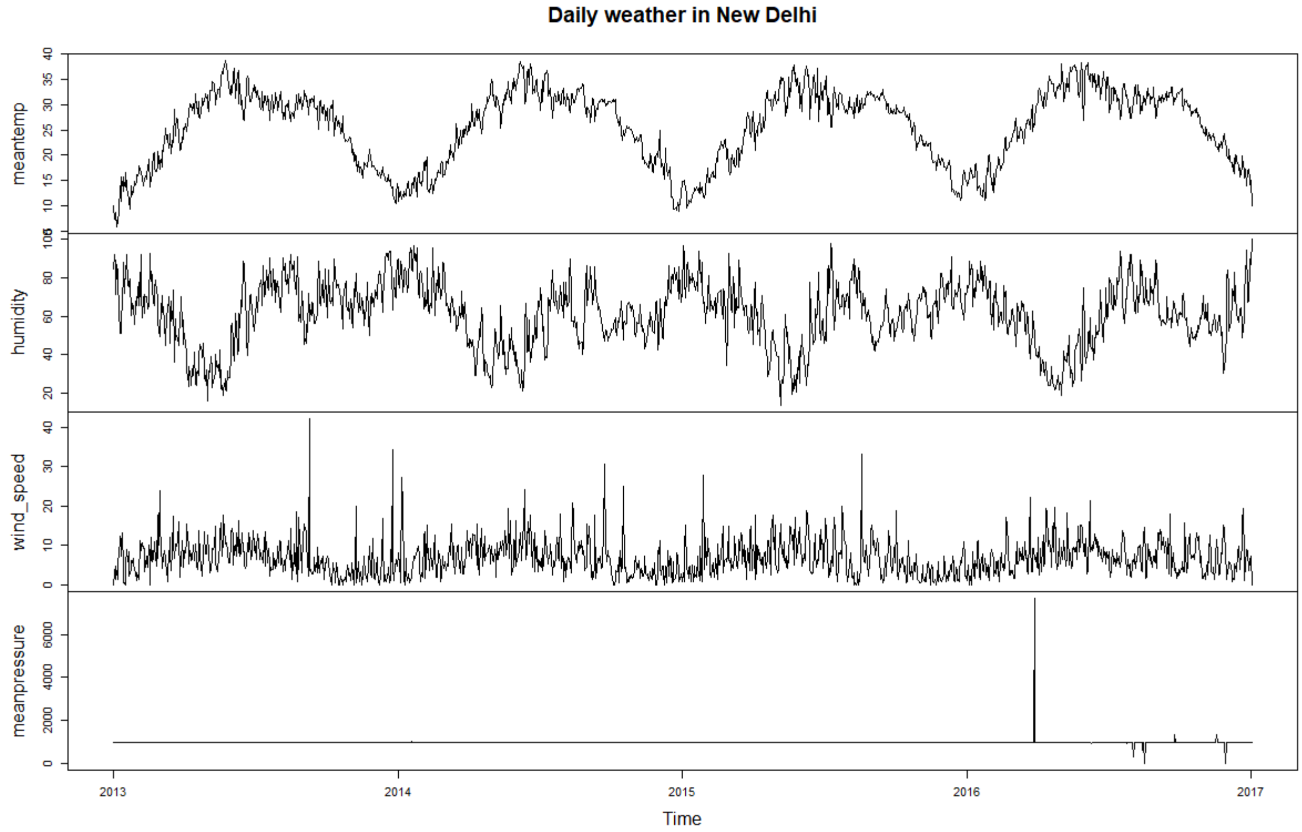

Figure 1: Graphical representation of the series

Let's start plotting the decomposed series one by one along with their ACF and PACF graphs, which will help us to achieve preliminary conclusions.

1. Daily mean temperature:

    (a) The summary of the data seems to be plausible for the climate of this city we previously described, which is a mixture of humid subtropical and semi-arid climate ones.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
6.00   18.86   27.71   25.50   31.31   38.71
```

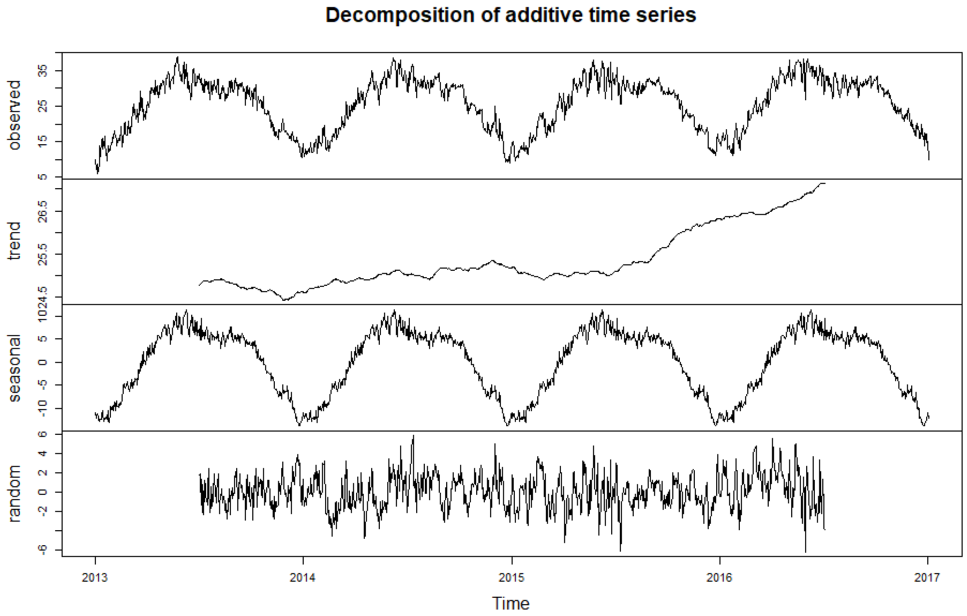

Figure 2: Decomposition of mean temperature series

(b) Regarding the decomposition of the series, we can immediately appreciate an evident seasonality.

(c) ACF and PACF are:

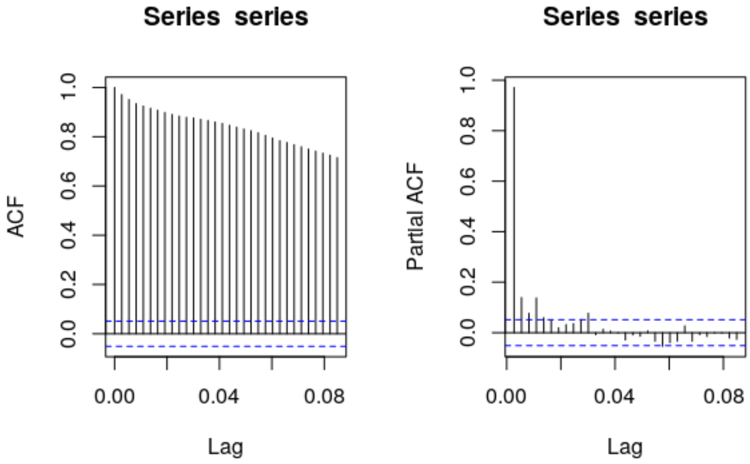

Figure 3: ACF and PACF of the mean temperature series

As we can see from the ACF plot, series are not stationary. This motivates differencing the series to achieve stationarity. The series, after that transformation, would be:

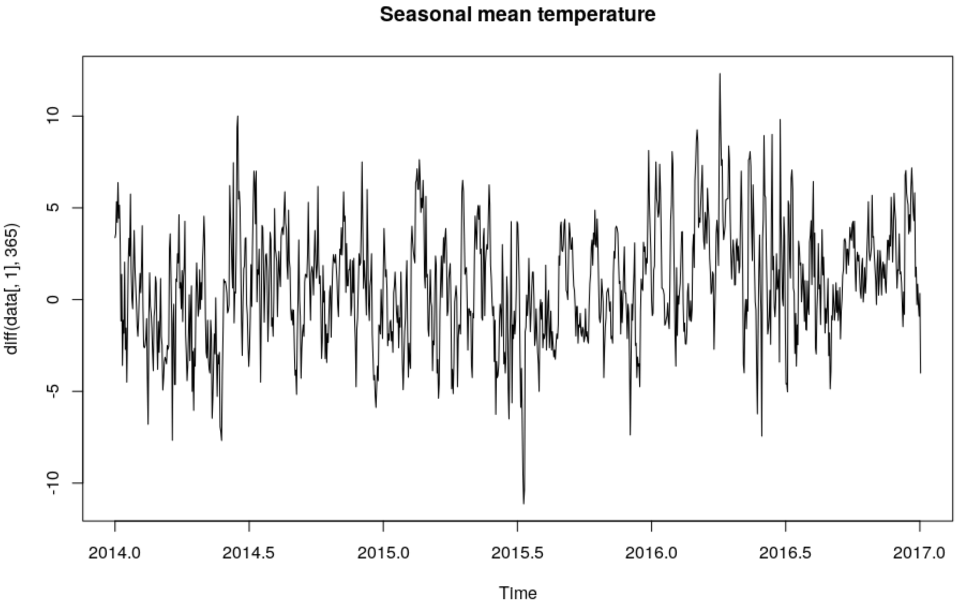

Figure 4: Plot of the differenced mean temperature series

Moreover, a surprising fact is the increasing mean temperature starting at 2015[1]. That trend is even more noticeable if we consider the heat wave occurred in 2016[2].

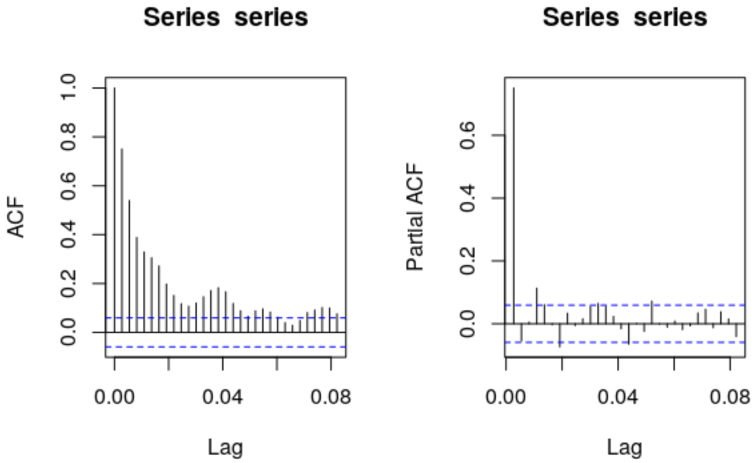

Figure 5: ACF and PACF of the differenced mean temperature series

(d) Although the ACF plot of the differenced series confirms the stationality character of the series (due to a clear exponential decay), the sinusoidal shape that is still observable makes us to have in mind a SARIMA model as a possibility when trying to fit a model.

(e) Just for checking, executing the Box-Ljung test shows a rejection of the null hypothesis indicating dependence among the different lags:

[1] Reference: wikipedia article for 2016 Indian heat wave

[2] Reference: wikipedia article for 2016 Indian heat wave

```
     Box-Pierce test
data:  series_mean_temp
X-squared = 583.66, df = 1, p-value < 2.2e-16
```

2. Humidity:

   (a) As we can see, there is high variability along the year, with days where the weather is highly dry in contrast with days of absolute humidity.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
13.43   50.38   62.62   60.77   72.22  100.00
```

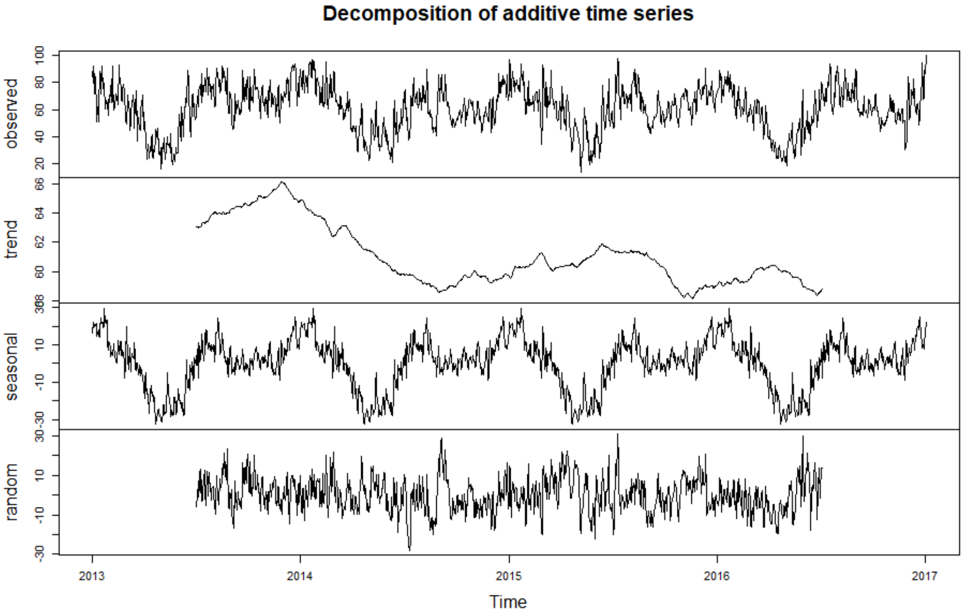

Figure 6: Decomposition of humidity series

   (b) As in the previous case, seasonality is evident. Also, we have to remark an anomaly happening every year in the middle of the summer, corresponding to the monsoon, which makes humidity grow for some days due to precipitation increase. Also, we can see a decreasing trend over the time. Despite a relief on the trend in mid 2015 (that could be caused by Gujarat cyclone[3]) the reason behind that could be the fact that India suffered a record breaking heat wave in 2015, and the region where New Delhi lies suffered another one in 2016. This could make humidity to appear smaller on average.

   (c) ACF and PACF are:

[3] Reference: wikipedia article for 2015 Gujarat cyclone

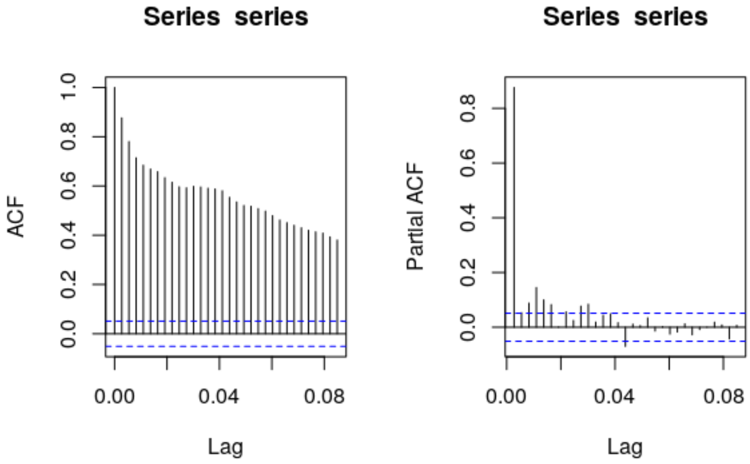

Figure 7: ACF and PACF of the humidity series

Similar to the case before, the differenced series are needed in order to achieve stationality. The series, after that transformation, would be:

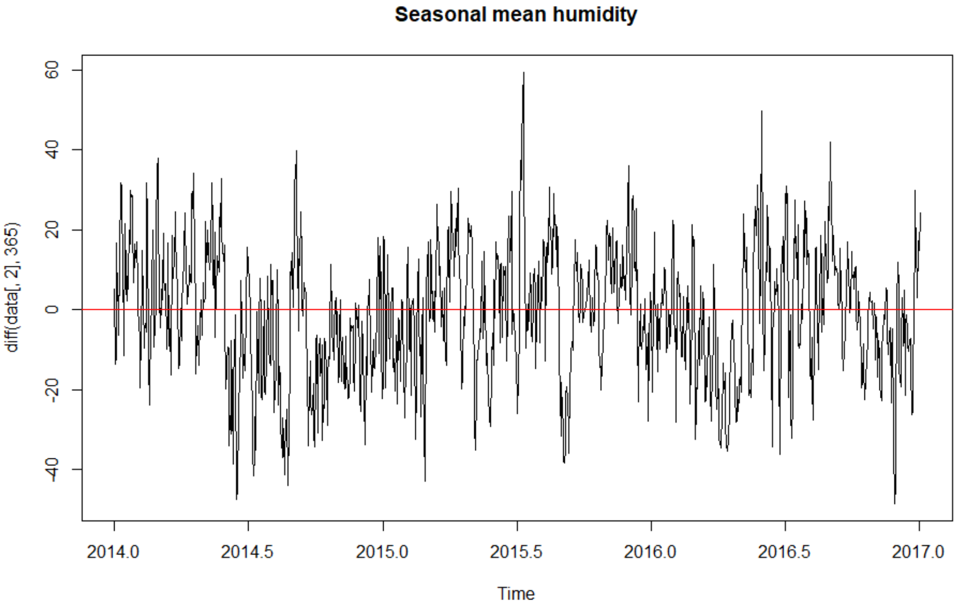

Figure 8: Plot of the differenced humidity series

It is worth mention the outliers that appear at mid year, which will be analized later; also, its noticeable a variation in the trend from 2016 onwards.

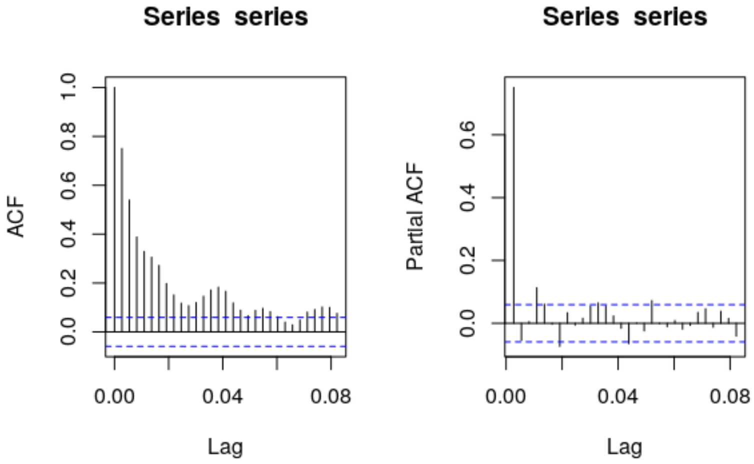

Figure 9: ACF and PACF of the differenced humidity series

(d) As in the previous example, the ACF plot of the differenced series confirms the stationality character of the series, due to a clear exponential decay. However, seasonality its still noticeable due to cycle that can be apreciated, requiring considering an SARIMA model as a good option

(e) For the PACF, the positive and negative variation observed makes us to think that the MA part is really neede, and has high importance.

(f) As done before, executing the Box-Ljung test show a rejection of the null hypothesis, implying correlation on different lags of the series:

```
Box-Pierce test
data:  series
X-squared = 618.04, df = 1, p-value < 2.2e-16
```

3. Mean wind speed:

(a) Mean wind speed is also a variable with high variability, having days with low winds and days with moderate wind speed. In any case, some peaks can be observed which may indicate the existence of potential outliers.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.000   3.475   6.222   6.802   9.238  42.220 
```

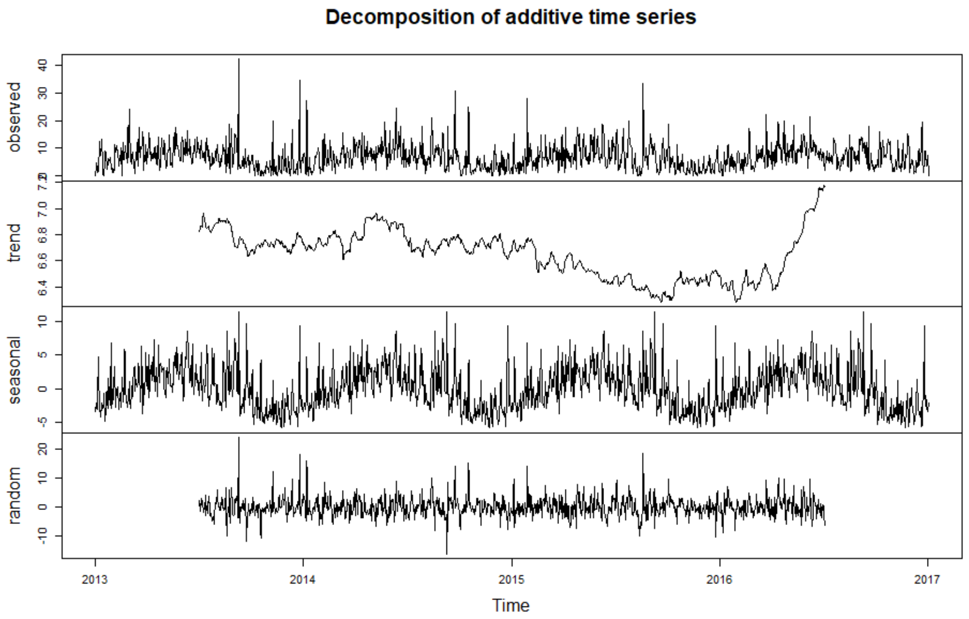

Figure 10: Decomposition of wind series

(b) One again, seasonality is evident. The monsoon is notable here, due to increasingly high values at summer, coinciding with humidity. Later in this project, we will later identify them as innovations. A strange increasing trend is also noticeable from 2016 onwards.

(c) ACF and PACF are:

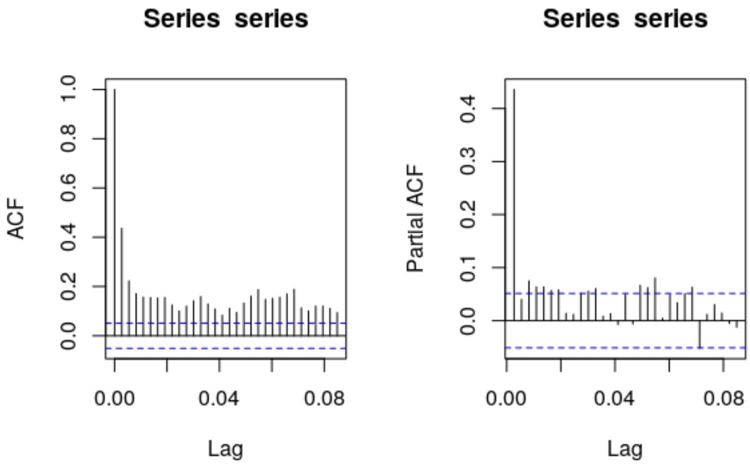

Figure 11: ACF and PACF of the wind series

Despite having a strong decay for the first coefficients (ACF plot), the correlation is still present. This motivates differencing the series:

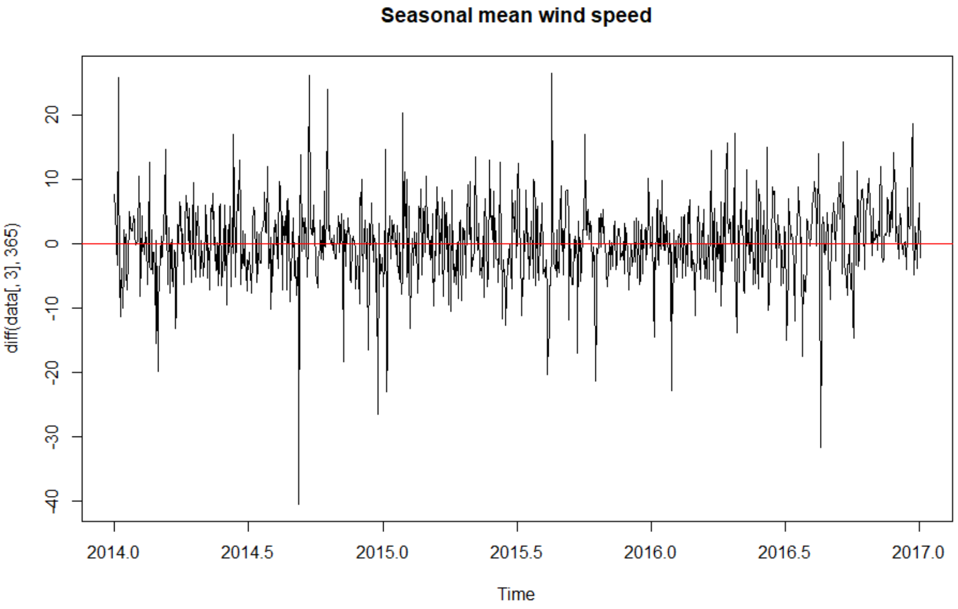

Figure 12: Plot of the differenced wind series

After looking that plot, trend starting at mid 2016 is clear. Also, the regular outliers corresponding to the monsoon events are detectable.

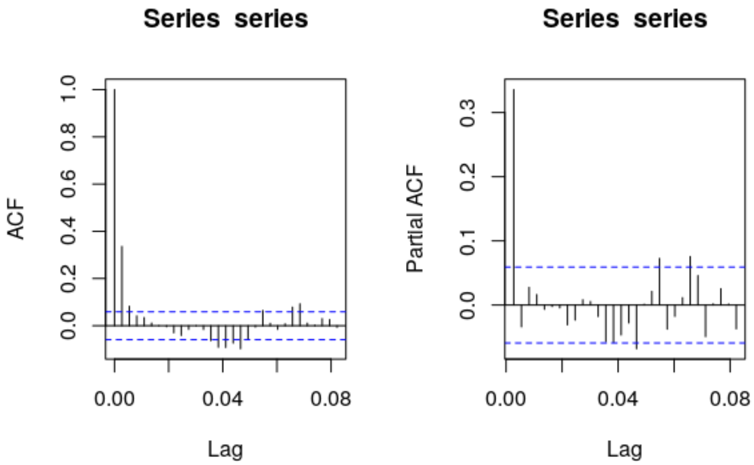

Figure 13: ACF and PACF of the differenced wind series

(d) The ACF and PACF shape proves there is still some kind of seasonality, due to their sinusoid shape, despite being not much strong. For startings, we can asure that the MA part is quite important, due to the non zero coefficients at the ACF part. However, modeling this series might require more complex model (due to the fact that there is still some seasonality remaining).

(e) As done before, executing the Box-Ljung test shows again a rejection of the null hypothesis indicating correlation among different lags:

```
Box-Pierce test
data:  series
X-squared = 123.67, df = 1, p-value < 2.2e-16
```

4. Mean atmospheric pressure:

   (a) A quick summary of the data for this series shows ridiculous values. The first and evident outlier we detect is the minimum atmospheric pressure, since it is impossible to be negative. Furthermore, a maximum value as large as 7000 is also impossible. They seems to be caused by a wrong measurements or malfunctions of the measuring instruments. Therefore, this series will need a special treatment regarding outlier detection.

```
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.042 1001.580 1008.563 1011.105 1014.945 7679.333
```

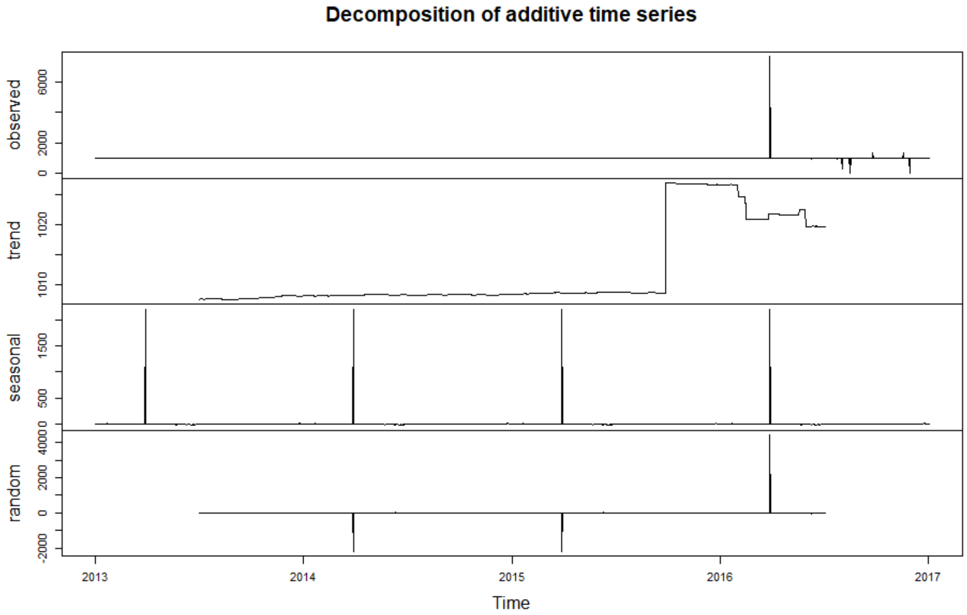

Figure 14: Decomposition of pressure series

   (b) As we pointed out before, since the outlier values are much different from the rest of the series, we are getting distorted results just by decomposing the series. Therefore, any further analysis to this series will be biased or will not make sense. As a result, the analysis of this series will be reserved until the outliers are being accounted for.

Summary

Beside the pressure series, which require an immediate outlier removal prior any analysis, the data seems to be pretty consistent: no big outliers were detected, and remaining seasonality makes us prone to estimate SARIMA models. Another thing to consider is a change in the trend from 2016 onward, that can be observed in all series. The reasons causing this change in trend could be tracked in some of them (temperature, humidity), but the actual reasons before that could be more complex; for example, the changes in trend could be associated to a global climate change which has accentuated since 2015[4].

[4] Reference: World Meteorological Organization

2 Outlier analysis and model fitting

This section consist on the fitting of a suitable model for each of our time series. The finding a suitable model is subject to many constraints, and different issues that are inherent to data can obscure the finding of a suitable model: this is the case of the outliers. In order to find a proper model, we need to determine if our data has anomalies that can potentially affect the results of the modelization of the data.

The outliers analysis will be performed via one-dimensional kmeans clustering. Applying one-dimensional kmeans algorithm to each of our series allows to explore graphically and see the different groups and infer insights. Exploring graphically these clusters together with domain-knowledge can help to easily spot outliers in the data, like extreme values or incorrect imputations in the data. Based on this insight obtained, the model of each series will be fitted on the premise of existing outliers or not. One of the main reasons behind performing this previous analysis instead of fitting each model through the R function tso that helps identify outliers, is the computational costs. Given the existent data, the computation time of fitting a model with outliers is in the order of hours. Furthermore, in the case of extreme meteorological events such as heavy storms,fitting a model on the premise of existing outliers might lead to the incorrect identification of the values registered in those occasions as outliers.

However, what is fitting a model? Fitting a model consist mainly on using algorithms to learn the relationship between the predictors and the outcome. In this case, the predictors can be a mixture of past observations and innovations. Given the current data, as it is analyzed in a univariate fashion, we are going to model the variables through ARIMA models. ARIMA stands for Auto-regressive Integrated Moving Average model and they are defined by the parameters p, d and q. The first term, p, defines the order of the auto-regression, while d is the number of differences needed to achieve stationarity in the series and q is the order of the moving average part. These models, assume that the series can be seen as:

$$X_t = \mu_t + a_t$$

where μ_t is the conditional mean of the series and a_t is white noise.

Furthermore, when the series has an stationary component as some of our variables they can be modelled through SARIMA(p,d,q)(P,D,Q) models. A series follows a SARIMA process if the following equation holds:

$$\Phi(B)\Phi_S(B^S)\nabla^d\nabla_S^D X_t = (1+\theta_1 B)(1+\theta_S B^S)a_t$$

.

The parameters p,d,q will be obtained via simulation with R through the usage of auto.arima function (or tso under the assumption of the existence of outliers) together with the insights obtained throughout the exploratory analysis.

Mean of temperatures

Based on the series observed throughout the graphical analysis, there does not seem to be outliers in the temperature. After performing clustering in its own values, we have obtained the two groups that can be observed in figure 15

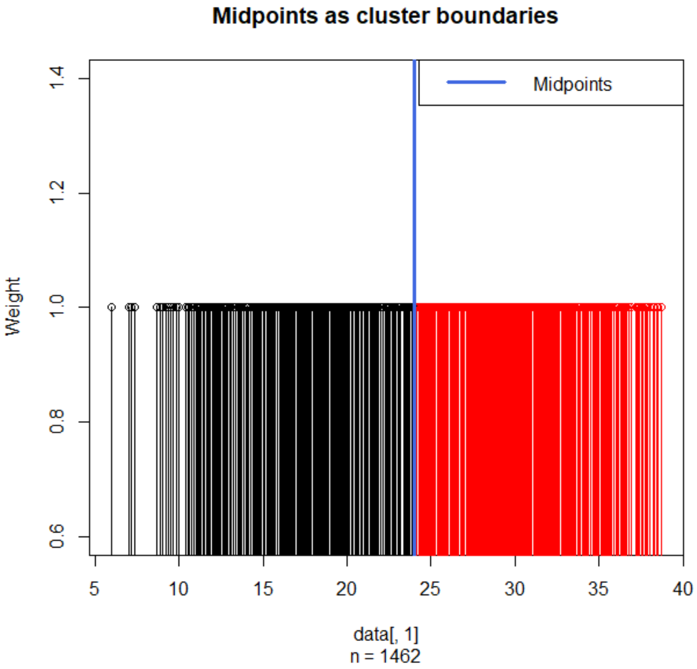

Figure 15: Clusters of mean temperature.

We can see two different groups in the data, all of them among normal values. It is assumed that the groups correspond to spring-summer and fall-winter. Based on this and the results of the exploratory analysis of the series, we can conclude that this series can be estimated through an ARIMA model.

The obtained ARIMA model corresponds to ARIMA(3,1,1)(0,1,0)[365] with the following coefficients:

- ar1 = 0.7188
- ar2 = 0.0102
- ar3 = -0.0375
- ma1 = -0.9883
- σ^2 estimated as 4.591

These results are not surprising as it was expected to have a seasonal component as this variable is based on the changing temperature during the years. Also, the estimation have a small variance, which is reasonable as temperatures changes are seldom very extreme from day to day. Basically, temperatures seem to follow a model based on the past three days temperature, with higher influence of the immediate past day and innovations with moderate, decreasing effect.

Mean humidity

Initial analysis of humidity indicated the absence of outliers in the data. The results obtained through clusterization in figure 16 back up this initial hypothesis. We can observe two different clusters, however, the second cluster is composed by only one observation with a value of 100. As this value is between the limits of what could be logical, we are not going to consider it an outlier.

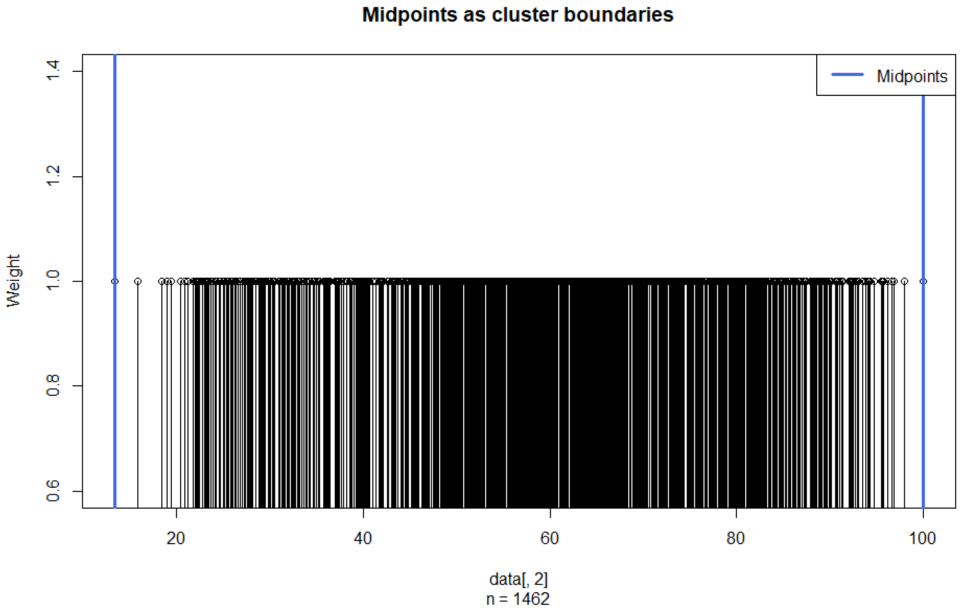

Figure 16: Clusters of mean humidity.

The estimated model for this data is an ARIMA(2,0,0)(0,1,0)[365]. Again we see that it is a seasonal arima model which was also suspected initially, as humidity in New Delhi follows a seasonal pattern heavily marked by the drier seasons and the monsoon season. The coefficients estimated are the following:

- ar1 = 0.7949
- ar2 = -0.0549
- $\sigma^2 = 115.8$

The variance in this case is quite high, however as humidity can vary a lot between days due to rains we assume this a an acceptable value. From this model we can infer that humidity is heavily influenced mainly by the previous day with a very small influence of the two-days past readings. With a high variance explained by meteorological events such as rains.

Mean wind speed

When it comes to wind speed, the variability along the seasons is smaller as seen in the exploratory analysis. However, it has certain spikes that may indicate wrong imputations in our data. Clustering yields several groups as seen in figure 17

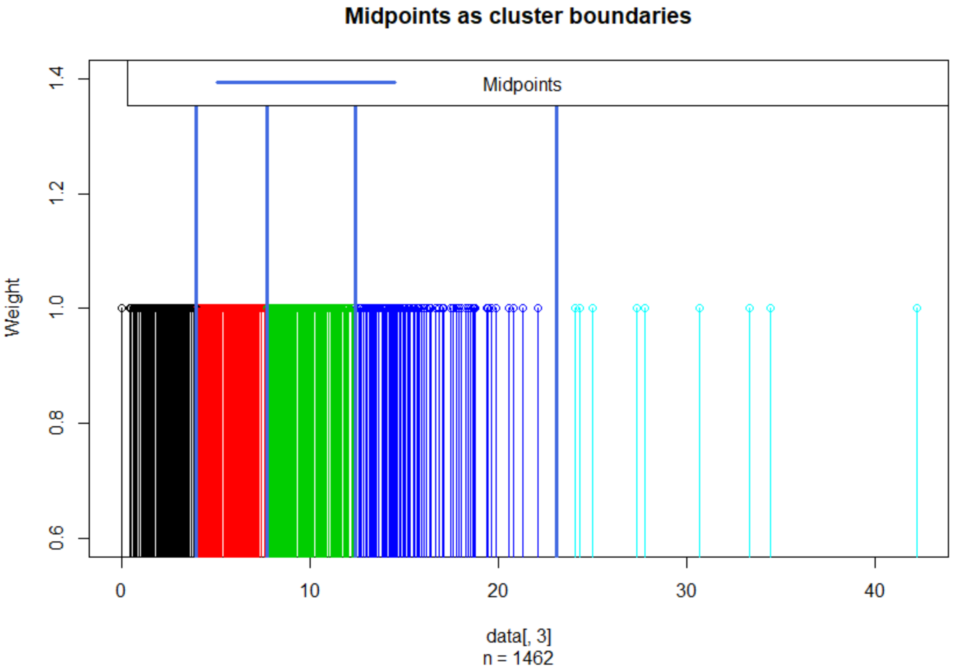

Figure 17: Clusters of mean wind speed.

The clusterization results show the existence of 5 different groups based on the wind speed. However, the fifth blue, is composed by a handful of observations with high speed values. Consulting weather information about New Delhi, and observing the general shape of the series, shows that high speed winds are not common at all in New Delhi, as a result, in order to avoid the contamination of our model by these high values, even though they might be real values, we are going to treat them as outliers and fit a model that identify them, in the case they are actually identified as such.

After estimating the model and its outliers, the results obtained showed that there are no outliers in the data, even after explicitly looking for them. As a result, the estimated model is an ARIMA(2,0,1) with non-zero mean. We can see that unlike humidity and temperature, the model does not have a seasonal component and its stationary by itself. The estimated coefficients are the following:

- ar1 = 1.3176
- ar2 = -0.3269
- ma = -0.9481
- mean = 6.7612
- σ^2 = 16.24

As we can see, our model is mainly influenced by the wind speed of the previous day, with big a noticeable effect of the past innovation too. The influence of the two-days-past observation is smaller but still existent. It is interesting to notice that on average there is always wind in New-Delhi.

Mean pressure

When it comes to atmospheric pressure, the values are generally centered around 1013mb. As a result, it is easy to spot outliers as values highly deviation from 1013 are imputation errors. Observing the clustering results of figure 18 and the series itself, it easy to notice the existence of unreal values.

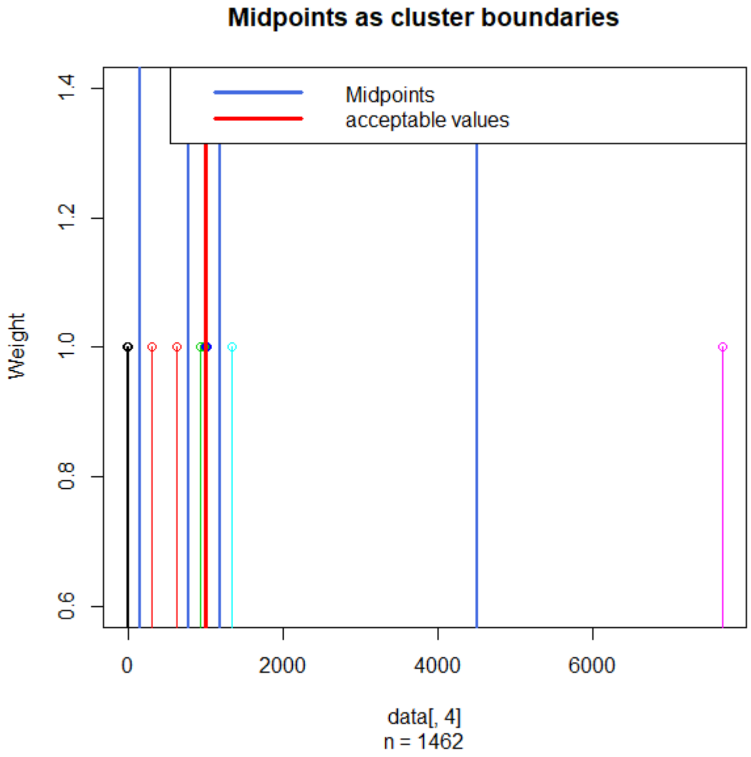

Figure 18: Clusters of mean pressure.

In the figure, an extra red line has been added signaling 1013mb. As a result, only the values located in the cluster in which the red line is located, are real values. All the other observations are clearly imputing mistakes as those values are not achievable in the Earth. As a result, pressure's model will be estimated taking into account the existence of outliers. Also, the data will be log-transformed as a way to make its computation faster.

The estimated model is an ARIMA(2,1,0)(0,1,0)[365]. Again, we see a seasonal component to the model which is expected, specially due to the monsoon period. It has detected 10 Additive outliers, which are isolated spikes that corresponds to an external error or exogenous change of the time series at a particular time point. Inspecting the series we see that the model has successfully identified all the spikes seen in the original data. The estimated AR(p) coefficients of the series are the following:

- ar1 = 0.7173
- ar2 = 0.1890

As we can see the series is driven by the previous observation and influenced by the observations from 2-days-past plus its seasonal component.

3 Weather forecasting

Weather forecasting has always been a trending topic, as knowing the weather of future days gives great advantage as it allows planing accordingly to the meteorological conditions. Nevertheless, forecasts are never absolutely accurate, as a result, our objective will be to generate predictions for future observations for each of our variables with a measure of uncertainty: a 95% confidence interval.

The probabilistic principle behind point forecast is the conditional mean such that:

$$\widehat{X}_{n+k/1:n} = E[X_{n+k}|x_1, ..., x_n]$$

On the other hand, the probabilistic principle behind the prediction confidence intervals is as follows:

$$\sigma^2_{X_{n+k}-\widehat{X}_{n+k}} = Var(X_{n+k} - \widehat{X}_{n+k}|x_1, ..., x_n)$$

With this in mind, and approximation of a 95% confidence interval for our forecast will be calculated through:

$$\widehat{X}_{n+k/1:n} \pm 1.96\sigma_{X_{n+k}-\widehat{X}_{n+k}}$$

In this specific case, our data contains a test set composed of new hundred observations. As a result, we are going to perform a 100-step-ahead prediction and compare with our test set.

When it comes to temperatures, we estimated an ARIMA(3,1,1)(0,1,0)[365] mainly affected by the past innovation and the past day values as per its coefficients value. It is easy to see in figure 19 that while most of the the true values fall inside our 95% confidence interval, our predicted temperatures and the real one are not exactly the same. Nonetheless, we could consider it an acceptable forecast as throughout the span of 100 days, we are able to forecast adequately all the temperature values with a confidence of 95%.

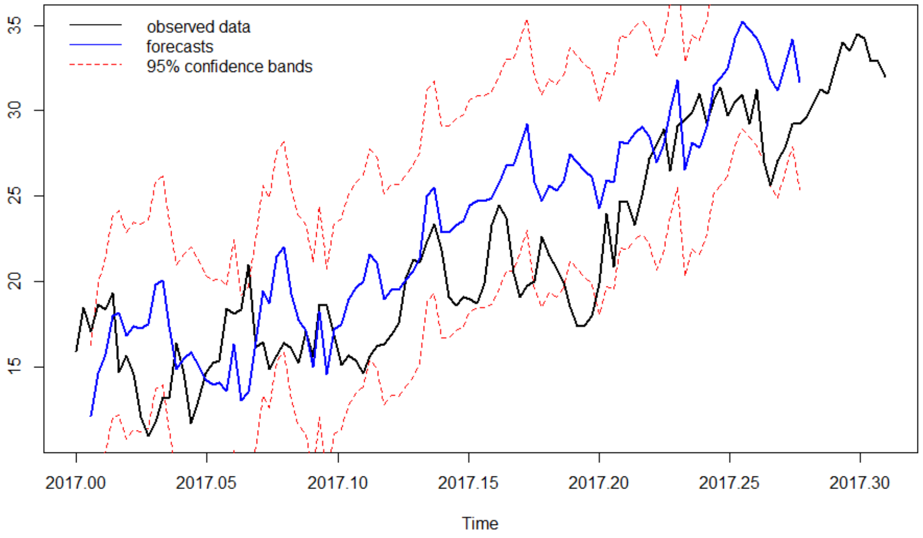

Figure 19: Forecast of the next hundred temperature values

In the case of humidity, the predicted values can be seen in figure 20. Figure 20 show that our predictions are quite accurate, as they follow closely the real values and every point lies inside our confidence interval. As a result, we can conclude that the humidity is accurately predicted by our estimated model.

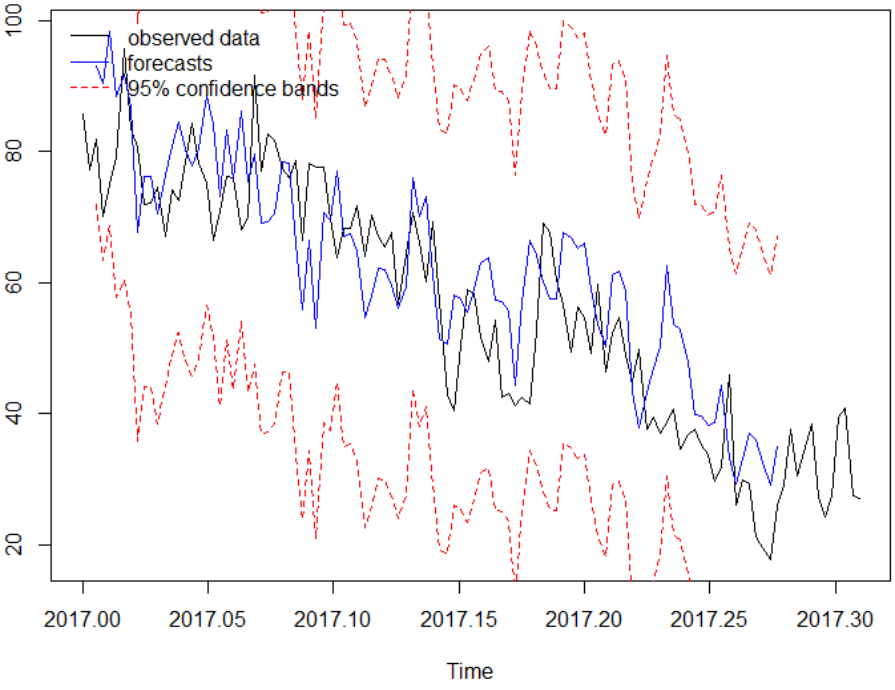

Figure 20: Forecast of the next hundred humidity values

On the other hand, the wind predictions are not as accurate as the temperature and humidity ones. As we can see in figure 21, the wind predictions are mainly based in the mean value of the series, 6.76. This is due to the lack of a seasonal component in our estimated model. However, the real values are located mainly inside the 95% confidence interval except for peak winds in certain occasions. However, as the wind series is stationary, we could consider an acceptable forecast estimating that every day the wind speed is the mean speed plus or minus the confidence regions.

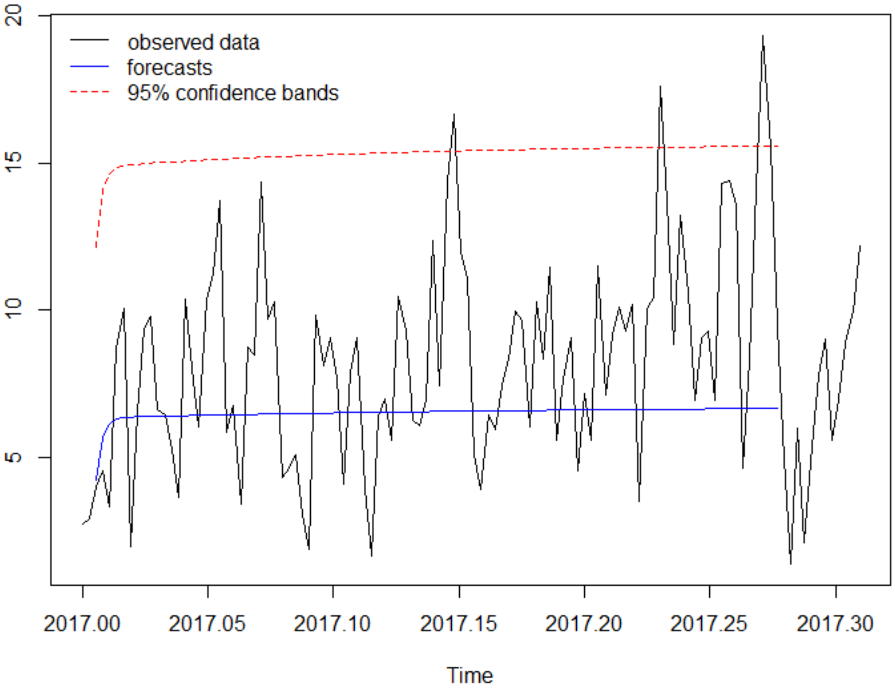

Figure 21: Forecast of the next hundred wind speed values

At last, the forecasting results of the atmospheric pressure are shown in figure 22. As we can see, the test data also has an outlier as it has an initial observation with value 0. However, during the rest of the predictions, our data accurately predicts the atmospheric pressure values. It is interesting to notice that in this case the size of the 95% confidence interval make it slightly useless as it expands to values that are impossible to achieve on a physical level. Nonetheless, our predicted values match closely the real values throughout the 100 days span.

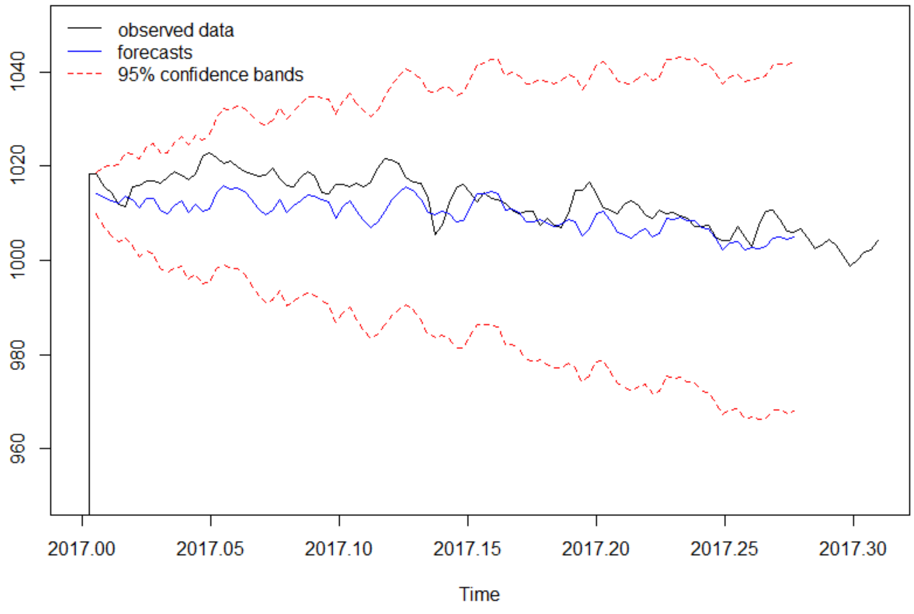

Figure 22: Forecast of the next hundred atmospheric pressure values

4 Multivariate model

One last idea we would like to cover is the fitting of a multidimensional model. In this part, we will not expose any rigorous theory, on the contrary, we will try to make use of already programmed R functions for fitting multivariate time series models.

The model we chose to fit is a VARMA: a multivariate ARMA model. A SVARMA model would perform a better fit, however, we opted for the former since similar results can be achieved by differencing the series. Moreover, R commands for fitting a VARMA model are quite easy to use (in the sense of automation), and they are computationally efficient for R standards, making calculations impressively fast.

First of all, we will briefly describe what is a VARMA model. Lets supose X_t is a k dimensional time series. Then X_t is a vector autoregressive moving average process, shortened VARMA(p,q), if

$$\Phi(B)X_t = \Phi_0 + \theta(B)a_t,$$

where Φ_0 is a constant vector, $\Phi(B)$ and $\theta(B)$ are two matrix polynomials, and a_t is a sequence of i.i.d distributed random vectors with mean zero and positive definite covariance matrix. Further conditions are required, but we will not list them here.

As we can see, a VARMA(p,q) model its just the multidimensional intuitive extension of an ARMA (p,q) model.

For fitting VARMA models we have to study their structural specification, which means find the structure of a multivariate linear time series so that a well defined VARMA model can be identified. That techniques are related to dimensionality reduction and variable selection.

The approach chosen is the Kronecker index approach. It consists on specifying the maximum order of the AR and MA polynomials for each component. Specifically, the Kronecker index method specifies an index for each component. Jointly, these indexes specify a VARMA model for the series. The following code is used for finding those indices through the usage of MTS library:

```
library(MTS)

series = ts(DailyDelhiClimateTrain, start= c (2013,01,01), frequency = 365)

series [,2] = diff(series [,2], lag=12)
series [,3] = diff(series [,3], lag=12)
series [,4] = diff(series [,4], lag=12)
series [,5] = diff(series [,5], lag=12)

>  index = Kronid(series[,-1], 4)$index
```

The former code yields the next results:

```
h =  0
Component =  1
square of the smallest can. corr. =  0.9516413
    test,   df, &  p-value:
[1] 4401.295   16.000    0.000
Component =  2
square of the smallest can. corr. =  0.7599014
    test,   df, &  p-value:
[1] 2072.29   15.00    0.00
Component =  3
square of the smallest can. corr. =  0.1533115
    test,   df, &  p-value:
[1] 241.645  14.000   0.000
Component =  4
square of the smallest can. corr. =  0.01234027
    test,   df, &  p-value:
[1] 18.023 13.000  0.157
A Kronecker index found
=============
h =  1
Component =  1
Square of the smallest can. corr. =  0.007954852
    test,     df, p-value & d-hat:
[1] 11.473 12.000  0.489  1.010
A Kronecker found
Component =  2
Square of the smallest can. corr. =  0.009223945
    test,     df, p-value & d-hat:
[1] 13.314 12.000  0.347  1.010
```

```
A Kronecker found
Component =  3
Square of the smallest can. corr. =  0.007903928
    test,     df, p-value & d-hat:
[1] 11.525 12.000  0.485  0.999
A Kronecker found
============

Kronecker indexes identified:
[1] 1 1 1 0
```

Therefore, the Kronecker indexes for each component is $(1, 1, 1, 0)$. Using this data, we could know the maximum order of the AR and MA parts.

We use that information for estimating the parameters of the model. The next command makes all the operations needed for that.

```
> Kronfit(series[,-1],index)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    0    0    0    2    2    2    0
[2,]    0    1    0    0    2    2    2    0
[3,]    0    0    1    0    2    2    2    0
[4,]    2    2    2    1    0    0    0    0
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    0    0    0    2    2    2    2
[2,]    0    1    0    0    2    2    2    2
[3,]    0    0    1    0    2    2    2    2
[4,]    2    2    2    1    0    0    0    0
Number of parameters:  28

Coefficient(s):
        Estimate  Std. Error  t value Pr(>|t|)
 [1,]  2.096e-02   3.488e-01    0.060  0.95209
 [2,]  9.814e-01   6.795e-03  144.435  < 2e-16 ***
 [3,]  3.726e-03   3.313e-03    1.125  0.26068
 [4,]  3.359e-02   2.734e-02    1.229  0.21917
 [5,] -1.747e-01   3.611e-02   -4.838 1.31e-06 ***
 [6,]  5.873e-03   7.585e-03    0.774  0.43873
 [7,] -5.547e-02   3.206e-02   -1.730  0.08367 .
 [8,]  2.484e-04   2.311e-04    1.075  0.28246
 [9,]  7.933e+00   2.027e+00    3.914 9.08e-05 ***
[10,] -1.368e-01   3.734e-02   -3.662  0.00025 ***
[11,]  8.916e-01   1.926e-02   46.300  < 2e-16 ***
[12,]  3.154e-01   1.444e-01    2.185  0.02887 *
[13,]  1.038e+00   1.619e-01    6.413 1.43e-10 ***
[14,]  3.566e-02   3.989e-02    0.894  0.37138
```

```
[15,] -2.694e-01   1.627e-01   -1.656  0.09782 .
[16,] -1.372e-03   1.122e-03   -1.223  0.22145
[17,]  1.900e+00   9.929e-01    1.913  0.05573 .
[18,]  9.216e-02   1.854e-02    4.971 6.68e-07 ***
[19,] -6.650e-03   9.303e-03   -0.715  0.47468
[20,]  4.365e-01   7.539e-02    5.789 7.06e-09 ***
[21,] -1.807e-01   8.055e-02   -2.243  0.02490 *
[22,] -8.673e-02   1.865e-02   -4.649 3.33e-06 ***
[23,] -1.230e-01   9.049e-02   -1.360  0.17391
[24,]  6.142e-05   5.446e-04    0.113  0.91020
[25,]  1.085e+03   4.650e+01   23.328  < 2e-16 ***
[26,]  1.364e+00   8.799e-01    1.551  0.12102
[27,]  4.962e-01   4.383e-01    1.132  0.25761
[28,]  1.271e+00   3.037e+00    0.419  0.67557
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
---
Estimates in matrix form:
Constant term:
Estimates:  0.021 7.933 1.9 1084.792
AR and MA lag-0 coefficient matrix
      [,1]  [,2]  [,3] [,4]
[1,] 1.000 0.000 0.000    0
[2,] 0.000 1.000 0.000    0
[3,] 0.000 0.000 1.000    0
[4,] 1.364 0.496 1.271    1
AR coefficient matrix
AR( 1 )-matrix
       [,1]   [,2]  [,3] [,4]
[1,]  0.981  0.004 0.034    0
[2,] -0.137  0.892 0.315    0
[3,]  0.092 -0.007 0.436    0
[4,]  0.000  0.000 0.000    0
MA coefficient matrix
MA( 1 )-matrix
       [,1]   [,2]  [,3]  [,4]
[1,]  0.175 -0.006 0.055 0.000
[2,] -1.038 -0.036 0.269 0.001
[3,]  0.181  0.087 0.123 0.000
[4,]  0.000  0.000 0.000 0.000

Residuals cov-matrix:
           [,1]      [,2]       [,3]         [,4]
[1,]  2.6489583 -8.397233  0.3722017    -1.525279
[2,] -8.3972327 62.046108 -7.5303095    -6.898324
[3,]  0.3722017 -7.530310 15.8916163    -9.393283
[4,] -1.5252788 -6.898324 -9.3932835 32421.124479
```

```
----
aic=  17.654
bic=  17.75527
```

Notice the low AIC and BIC values achieved, indicatting a good fit of the model. Also, we have to note the short computational time spent on that calculations, proving the worth of this approach.

The goal of this part was to show briefly what a multivariate model is, and how it could be fitted (since identification is not straightforward, more steps are needed rather than just the parameter estimation), we are not going to perform more computations regarding the fitted model, but the model has been accurately determined with the coefficients we have presented before.

5 Conclusions

The prior study that was carried out over the weather observations showed the general need of differencing for achieving stationarity. Furthermore, initial conclusions about outliers (or the inexistence of them), type of model to fit, and remaining seasonality could be carefully drawn.

However, a more informative analysis yielded decisive information about each variable. In this part, complex algorithms were needed for perform such analysis, involving auto-fitting methods as well as one dimensional clustering. Forecasting were realized for 100 days ahead.

1. Mean temperature: methodological analysis showed no outliers. SARIMA model was fitted, since changes from day to day were smooth. Forecasting was straightforward, getting an accurate prediction.

2. Humidity: similar behaviour to temperature, in the sense of model fitting but forecasting as well. Seasoning corresponding to monsoons and summer/winter variation.

3. Mean wind speed: despite analysis proved the existence of outliers, they finally did not affect the model fitting. However, the absence of seasonality provoked this series to be fitted with an ARMA model. This model fitting causes the forecasting to be awful, being the best prediction similar to the unconditional mean, which is not realistic.

4. Mean atmospheric pressure: this series required much more preprocessing than the others. Analysis output showed 10 additive outliers, caused, presumably, by instrumental error measuring. Finally, an SARIMA model was fitted. Forecasting confidence bounds was unrealistic, by the limitation of the variables. Despite that, prediction was pretty accurate too.

Finally, a multivariate model was fitted, in order to carry out an attempt for gathering all variables in one model, including the relations between them. Despite being a relative simple model (the multivariate equivalent for an ARMA model), posterior analysis showed that the model is quite informative with respect to the data.

Therefore, we can conclude that meteorological variables, despite having a characteristic randomness, can be studied (individually) applying Time Series methods, achieving, in general, good results, not necessarily using complex methods in the process.