Wave Height Forecasting

Alberto Guijarro Rodriguez

02/07/2021

Introduction

Wave height forecasting plays a crucial role on supporting a country coastal development. Being able to correctly define your coastal conditions allows to correctly define the design parameters to create coastal protection structures such as:

- Ports & Harbours
- Jetty walls
- Breakwaters

And even design some of the most complex and spectacular cities known to man, such as the Palm Jumeirah in Dubai.

The dataset considered for this analysis contains ocean wave information about Valencia, a coastal city in the east of Spain. The Buoy is placed on the Mediterratenan Sea, and the data can be extracted from http://www.puertos.es/en-us/oceanografia/Pages/portus.aspx.

On this project we will conduct an Exploratory Data Analysis (EDA) on the Coastal Data for the Valencia region and build a robust forecasting model on expected Significant (H_s), aiming to provide reliable forecasted values for the significant wave height.

Variables

The full description on the features and parameters can be found in the attached file "OPPE_Description.pdf", but given this is supplied by the Public Gov Organism State Ports and it is only supplied in Spanish, we will provide the following variables description,

Variable sample times

The values are aggregated on an hourly basis, however the buoy parameters they are not measured during the entire hour, for instance wind data is measured during a 10 minutes interval every hour, thus even though each hour has an average wind speed data point, this value has been calculated over a 10 minutes period.

Table below details the mean duration for each one of the observed agents:

Agent Observed	Measure Duration
Waves	30Min (Aprox)
Wind Speed	10 Min.

Agent Observed	Measure Duration
Current Speed	10 Min.
Air Temperature	Instantaneous
Air Preasure	Instantaneous
Water Salinity	Instantaneous
Water Temperature	Instantaneous

Variables description

- Date (GMT): Time Value was recorded at time zone GMT, Greenwich Mean Time
- Source of data: there are three columns and they seem to represent a change in the device being used to record the data, there is no information in the official documentation, but if we build a time series plot we can see how on the 2020-11-25 at 22:00:00 all the sources share the same value 3, and also how the number of missing values on some fields increases when compared to the previous periods.

Waves

- Scalar parameters of zero cross and espectral.
    - Significant Wave Height(m): the average wave height, from trough to crest, of the highest one-third of the waves. This is used to represent the wave height an trained observer will appreciate by simply looking where the wave is crossing (not from the coast).
    - Mean Period Tm02(s): Mean period (as an inverse of the wave length) of the measured waves, usually referred to as T_m
    - Peak Period(s): Period of the wave group with more energy, usually defined as T_p, the more regular the waves are, the closer the T_p is to the T_m, however T_p is usually larger than T_m
    - Maximum Waves Height(m): Max Wave Height measured
    - Highest Wave Period(s):
- Directional parameters
    - Waves coming-from Direc.(0=N,90=E)
    - Wave coming-from direction(0=N,90=E)
    - Mean comming-from Direc. at Waves Peak(grados): Direction of the waves with largest energy

Meteorology

Measured 3m over the water surface

- Atmospheric Pressure(hpa)
- Air Temperature(ºC)
- Wind Speed(m/s)
- Wind coming-from Direction(0=N,90=E) , mean direction of propagation

Oceanography

Measured 3m under the water surface

- Sea Temperature(ºC)
- Salinity(Practical Salinity Units, psu), salt concentration (Sodium and Chlorure) in sea water.
- Currents Mean Speed(cm/s), mean speed of the water currents
- Currents propag. Direction(0=N,90=E), mean direction of propagation

Valencia Buoy information

- Longitude 0.20 E
- Latitude 39.51 N
- Data Sampling 60 Min
- Code 2630
- Mooring Depth 260m
- First record date: 2005-09-15
- Last record date: 2021-06-16
- Type of sensor: Directional Met-Oce
- Model: SeaWatch
- Data set: REDEXT

```
## [1] 8757   21
```

```
## tibble [8,757 x 21] (S3: tbl_df/tbl/data.frame)
## $ Date..GMT.                            : POSIXct[1:8757], format: "2020-06-15 00:00:00" "20
## $ Significant.Wave.Height.m.            : num [1:8757] 0.62 0.58 0.53 0.44 0.41 0.36 0.31 0
## $ Mean.Period.Tm02.s.                   : num [1:8757] 3.73 3.65 3.54 3.31 3.03 3 2.91 3.08
## $ Peak.Period.s.                        : num [1:8757] 4.65 4.33 4.19 4.12 3.82 3.87 4 3.87
## $ Maximum.Waves.Height.m.               : num [1:8757] 1.09 0.89 0.85 0.7 0.54 0.54 0.42 0.
## $ Highest.Wave.Period.s.                : num [1:8757] 4.45 4.34 3.99 4.27 4.27 4.7 4.83 4.
## $ Source.of.data                        : num [1:8757] 1 1 1 1 1 1 1 1 1 1 ...
## $ Waves.coming.from.Direc..0.N.90.E.    : num [1:8757] 163 164 163 170 187 194 184 158 134
## $ Wave.coming.from.direction.0.N.90.E.  : num [1:8757] 174 168 170 168 194 187 174 163 137
## $ Mean.comming.from.Direc..at.Waves.Peak.grados.: num [1:8757] 32 25 34 28 49 46 46 67 50 52 ...
## $ Source.of.data.1                      : num [1:8757] 1 1 1 1 1 1 1 1 1 1 ...
## $ Sea.Temperature.ÂºC.                  : num [1:8757] 21.7 21.6 21.6 21.6 21.5 ...
## $ Salinity.psu.                         : num [1:8757] 37.7 37.7 37.7 37.7 37.7 ...
## $ Currents.Mean.Speed.cm.s.             : num [1:8757] 28.9 28.9 30 15.2 15.6 17.5 23 17.1
## $ Currents.propag..Direction.0.N.90.E.  : num [1:8757] 182 187 208 213 233 265 271 309 343
## $ Source.of.data.2                      : num [1:8757] 2 2 2 2 2 2 2 2 2 2 ...
## $ Atmospheric.Pressure.hpa.             : num [1:8757] 1018 1018 1018 1018 1018 ...
## $ Air.Temperature.ÂºC.                  : num [1:8757] 21.6 21.4 21.2 21.6 21.4 ...
## $ Wind.Speed.m.s.                       : num [1:8757] 2.81 1.17 2.57 4.21 3.75 2.81 3.04 2
## $ Wind.coming.from.Direction.0.N.90.E.  : num [1:8757] 67 36 326 329 320 315 317 309 300 320
## $ Source.of.data.3                      : num [1:8757] 2 2 2 2 2 2 2 2 2 2 ...
```

```
##    Date..GMT.                   Significant.Wave.Height.m. Mean.Period.Tm02.s.
##  Min.   :2020-06-15 00:00:00   Min.   :0.1100             Min.   :1.860
##  1st Qu.:2020-09-14 04:45:00   1st Qu.:0.4700             1st Qu.:3.320
##  Median :2020-12-14 14:30:00   Median :0.6600             Median :3.750
##  Mean   :2020-12-14 12:01:52   Mean   :0.7668             Mean   :3.863
##  3rd Qu.:2021-03-15 19:15:00   3rd Qu.:0.9400             3rd Qu.:4.220
##  Max.   :2021-06-15 01:00:00   Max.   :4.2200             Max.   :8.130
##  NA's   :1                     NA's   :68                 NA's   :68
##  Peak.Period.s.   Maximum.Waves.Height.m. Highest.Wave.Period.s.
##  Min.   : 1.750   Min.   :0.190           Min.   : 2.270
##  1st Qu.: 4.000   1st Qu.:0.620           1st Qu.: 3.960
##  Median : 4.920   Median :0.850           Median : 4.450
##  Mean   : 5.311   Mean   :1.044           Mean   : 4.696
##  3rd Qu.: 6.250   3rd Qu.:1.280           3rd Qu.: 5.260
##  Max.   :24.800   Max.   :6.480           Max.   :20.630
```

```
##   NA's   :68        NA's   :4833          NA's   :4837
##   Source.of.data  Waves.coming.from.Direc..0.N.90.E.
##   Min.   :1.000  Min.   :   0.0
##   1st Qu.:1.000  1st Qu.: 67.0
##   Median :3.000  Median :109.0
##   Mean   :2.104  Mean   :124.1
##   3rd Qu.:3.000  3rd Qu.:158.0
##   Max.   :3.000  Max.   :360.0
##                  NA's   :68
##   Wave.coming.from.direction.0.N.90.E.
##   Min.   :   0.0
##   1st Qu.: 66.0
##   Median :106.0
##   Mean   :125.8
##   3rd Qu.:161.0
##   Max.   :359.0
##   NA's   :68
##   Mean.comming.from.Direc..at.Waves.Peak.grados.  Source.of.data.1
##   Min.   :14.00                                   Min.   :1.000
##   1st Qu.:28.00                                   1st Qu.:1.000
##   Median :35.00                                   Median :3.000
##   Mean   :36.93                                   Mean   :2.104
##   3rd Qu.:44.00                                   3rd Qu.:3.000
##   Max.   :80.00                                   Max.   :3.000
##   NA's   :4837
##   Sea.Temperature.ÂºC.  Salinity.psu.   Currents.Mean.Speed.cm.s.
##   Min.   :13.51        Min.   :37.22   Min.   :  0.00
##   1st Qu.:15.00        1st Qu.:37.81   1st Qu.:  8.60
##   Median :18.57        Median :38.04   Median :13.60
##   Mean   :19.56        Mean   :37.97   Mean   :14.94
##   3rd Qu.:24.64        3rd Qu.:38.16   3rd Qu.:19.50
##   Max.   :28.34        Max.   :38.25   Max.   :64.10
##                                        NA's   :1
##   Currents.propag..Direction.0.N.90.E.  Source.of.data.2
##   Min.   :   0.0                        Min.   :2.000
##   1st Qu.: 69.0                         1st Qu.:2.000
##   Median :153.5                         Median :3.000
##   Mean   :161.6                         Mean   :2.551
##   3rd Qu.:246.0                         3rd Qu.:3.000
##   Max.   :359.0                         Max.   :3.000
##   NA's   :1
##   Atmospheric.Pressure.hpa.  Air.Temperature.ÂºC.  Wind.Speed.m.s.
##   Min.   : 991.4            Min.   :11.09         Min.   : 0.000
##   1st Qu.:1013.6            1st Qu.:17.03         1st Qu.: 2.340
##   Median :1017.2            Median :18.75         Median : 3.980
##   Mean   :1017.2            Mean   :19.12         Mean   : 4.486
##   3rd Qu.:1021.3            3rd Qu.:21.09         3rd Qu.: 6.090
##   Max.   :1036.7            Max.   :26.56         Max.   :17.580
##   NA's   :5                NA's   :5317          NA's   :81
##   Wind.coming.from.Direction.0.N.90.E.  Source.of.data.3
##   Min.   :   0.0                        Min.   :2.000
##   1st Qu.: 64.0                         1st Qu.:2.000
##   Median :151.0                         Median :3.000
##   Mean   :161.3                         Mean   :2.551
```

```
##  3rd Qu.:250.0                          3rd Qu.:3.000
##  Max.   :357.0                          Max.   :3.000
##
```

Missing Values

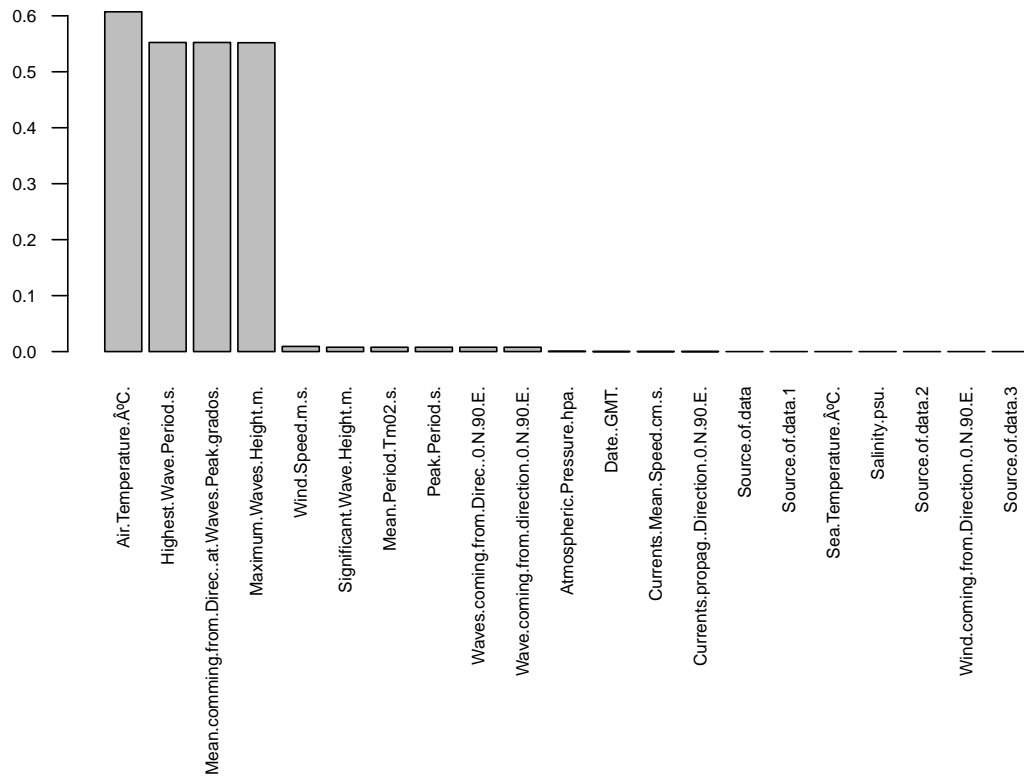Below we can see the % of null values present in our dataset

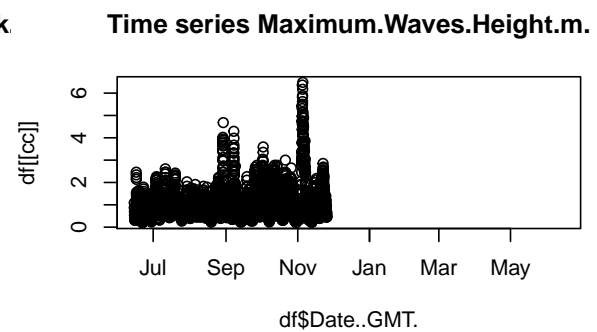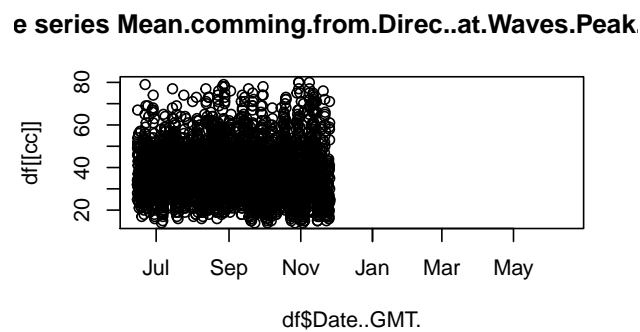- Air Temperature(ºC) has 60.02% of its values missing in the considered time period
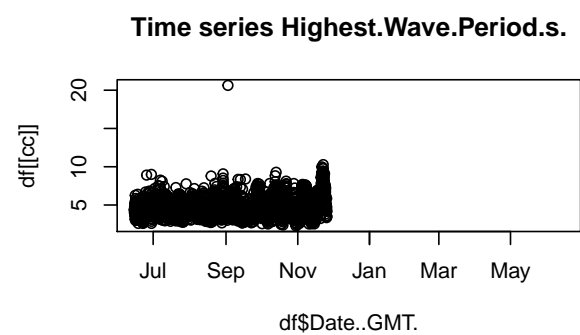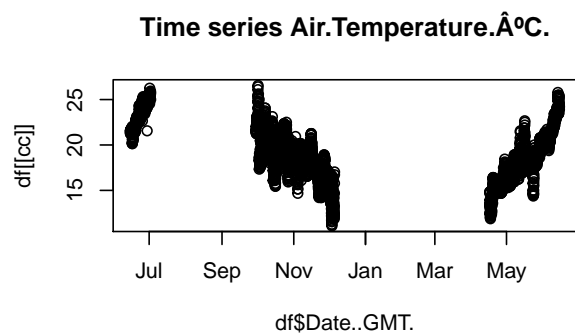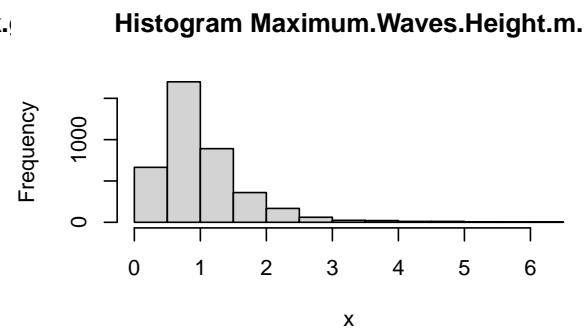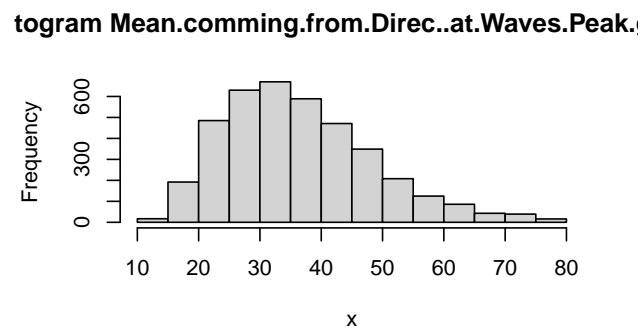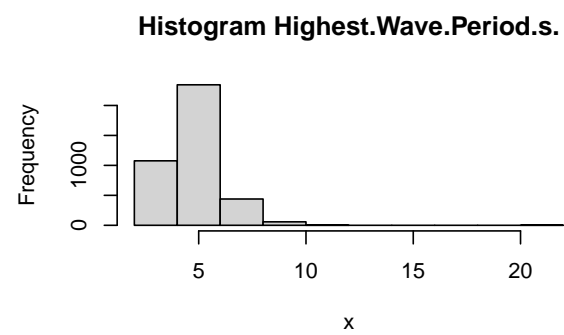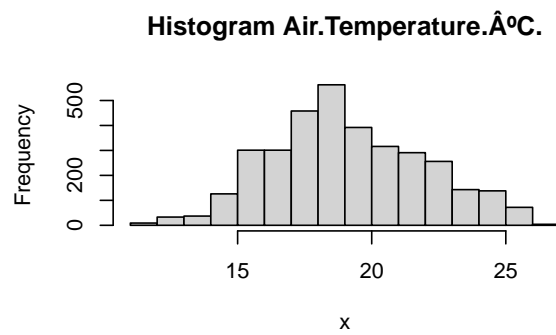
The Highest Wave records present the 55% of their records missing for the considered time period

- Highest Wave Period(s)
- Mean comming-from Direc. at Waves Peak(grados)
- Maximum Waves Height(m)

The remaining variables presenting missing values have a ratio of under 1% meaning we could probably exclude them and still have a good enough sample size to model. For the other 4 variables presenting high levels of missing values we will study them independently and assess what is the best or more reasonable missing values strategy.

```
##  [1] "Air.Temperature.ÂºC."
##  [2] "Highest.Wave.Period.s."
##  [3] "Mean.comming.from.Direc..at.Waves.Peak.grados."
##  [4] "Maximum.Waves.Height.m."
##  [5] "Wind.Speed.m.s."
##  [6] "Significant.Wave.Height.m."
##  [7] "Mean.Period.Tm02.s."
##  [8] "Peak.Period.s."
##  [9] "Waves.coming.from.Direc..0.N.90.E."
## [10] "Wave.coming.from.direction.0.N.90.E."
## [11] "Atmospheric.Pressure.hpa."
## [12] "Date..GMT."
## [13] "Currents.Mean.Speed.cm.s."
## [14] "Currents.propag..Direction.0.N.90.E."
## [15] "Source.of.data"
## [16] "Source.of.data.1"
## [17] "Sea.Temperature.ÂºC."
## [18] "Salinity.psu."
## [19] "Source.of.data.2"
## [20] "Wind.coming.from.Direction.0.N.90.E."
## [21] "Source.of.data.3"
```

Histogram Air.Temperature.ºC.

Histogram Highest.Wave.Period.s.

togram Mean.comming.from.Direc..at.Waves.Peak.

Histogram Maximum.Waves.Height.m.

Time series Air.Temperature.ºC.

Time series Highest.Wave.Period.s.

e series Mean.comming.from.Direc..at.Waves.Peak.

Time series Maximum.Waves.Height.m.


```
## # A tibble: 4 x 5
##   name                          mean     sd   skw   kurt
##   <chr>                        <dbl>  <dbl> <dbl>  <dbl>
```

```
## 1 Air.Temperature.ÂºC.                        19.1   2.86  0.175 -0.446
## 2 Highest.Wave.Period.s.                      4.70   1.16  1.70  10.5
## 3 Mean.comming.from.Direc..at.Waves.Peak.grados. 36.9  12.1  0.733  0.367
## 4 Maximum.Waves.Height.m.                     1.04  0.694 2.52  10.1
```

After exploring the, probabilistic distributions one could argue most of them are right skewed except for the Air Temperature which is the one which presenst a closets similarity to a normal distribution (Skewness= 0.17, Kurt = -0.44). But given this is a time series, it might be possible to fit it to a trend function, so let's explore them from a time series perspective.
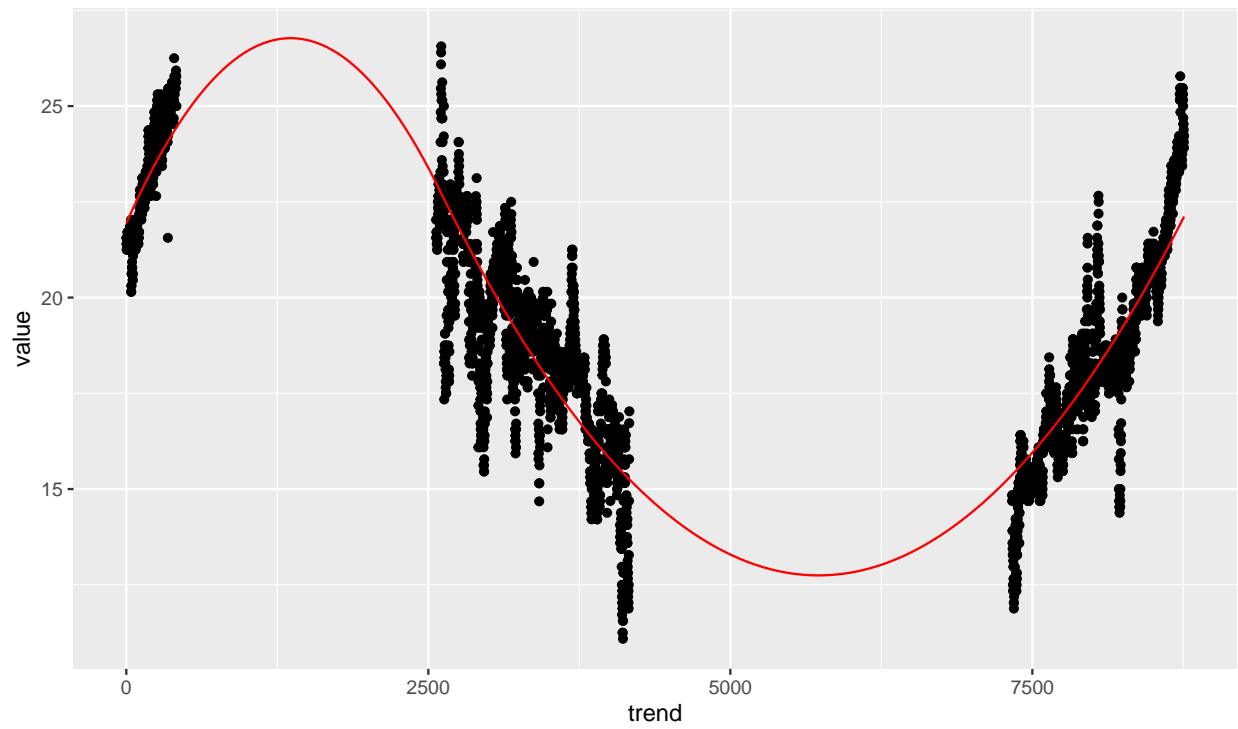
Air Temperature

The variable seems to be following a cyclic pattern, which would make sense, the air temperature would be expected to be higher during the summer periods and lower during the winter, plot below shows how we have the trend points for the series, we have the initial increment from June 2020, then around October it starts to show a downward trend and after May it starts to pick up again.
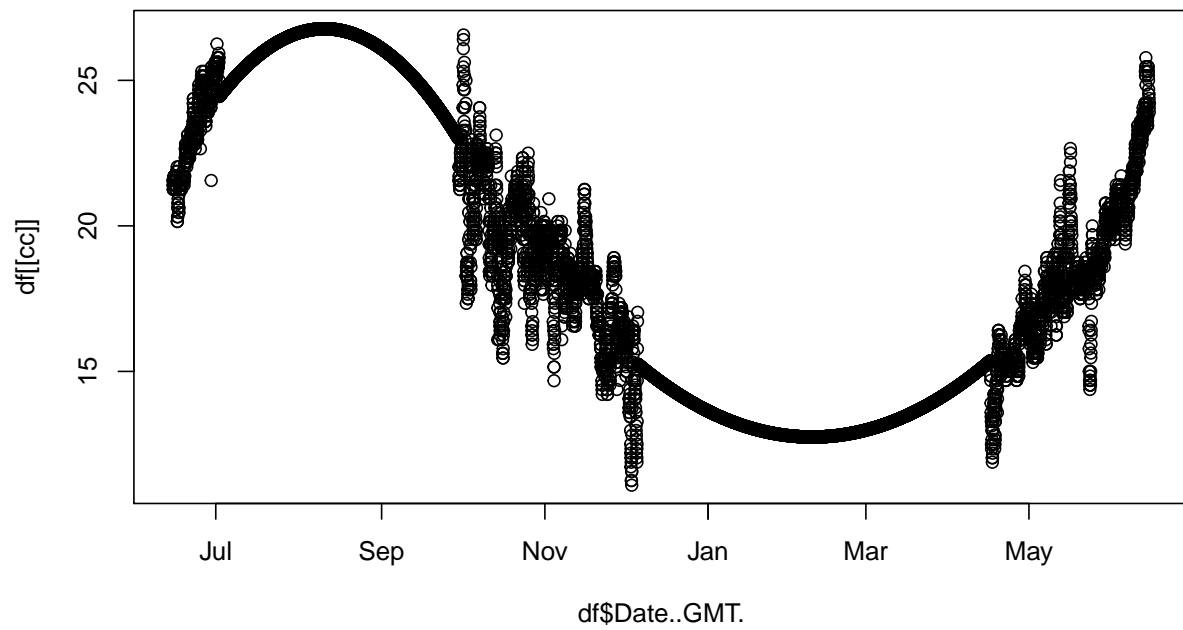
Time series Air.Temperature.ÂºC.

df$Date..GMT.

8
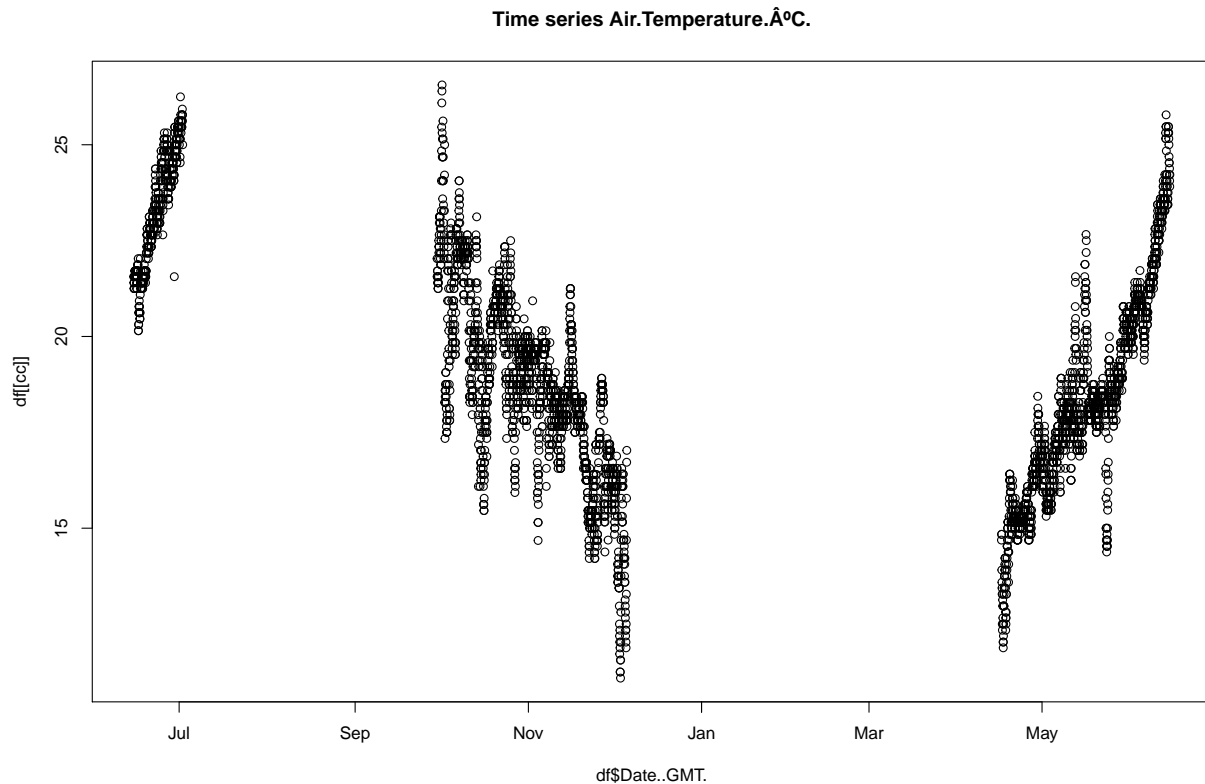
Adjusted Time series plot

## Time series for Air Temperature after missing values filled

Other three columns

The variables Highest Wave Period(s), Mean comming-from Direc. at Waves Peak(grados) and Maximum Waves Height(m) do not present any obvious overarching trend, therefore we will use the mean value to fill the missing values.

Remaining missing values

Given they represent less than 1% of the total values, we will use the mean value for all of them and drop all the records for the missing values of **"Significant Wave Height(m)"**, given this is our target value.
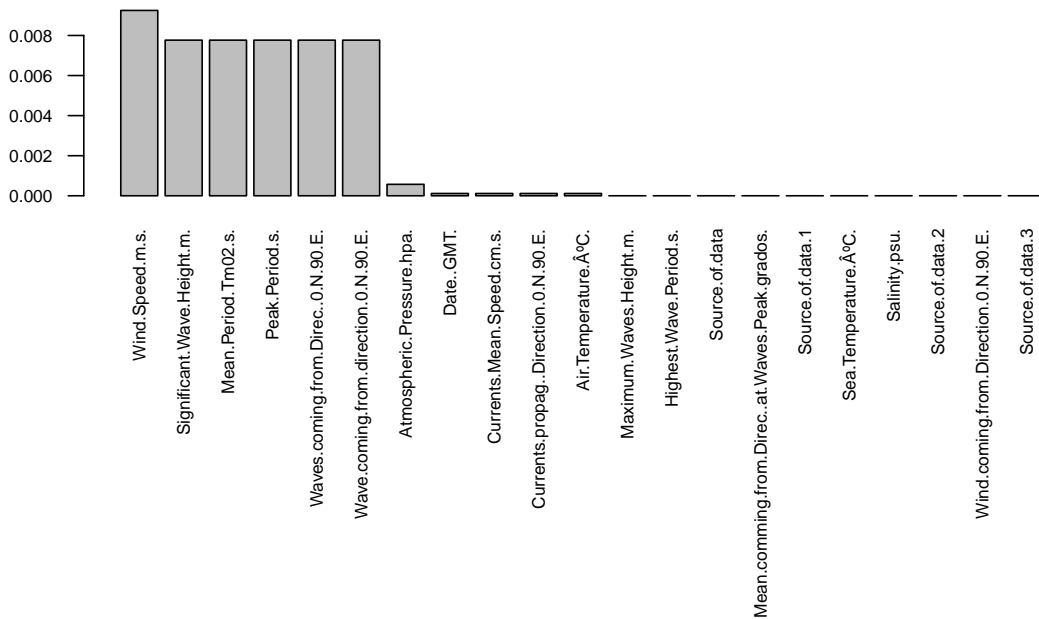

Analysing target variable

When plotting the variable it is quite obvious it is a time series which presents several levels of seasonality, the usual levels of seasonality experience by the sea state can be found in several nautical bibliography books, but as a rule of thumb waves in the ocean present at the very least the following levels of seasonality:

- Every 12hrs
- Daily (Mon to Sun)
- Moon Cycles (every 28 days)
- YearSeasons, which we will model with the per month indicator

If we had longer tailed data spanning over several years we might be able to identify further seasonal information periods, but given our dataset spans for the lenght of 1 year, we will focus on the most frequently identifiable seasonal events a wave time series might experience over the course of a year.

Which means we will add those levels of information to our dataset to highlight those events

Lagged Time series plot
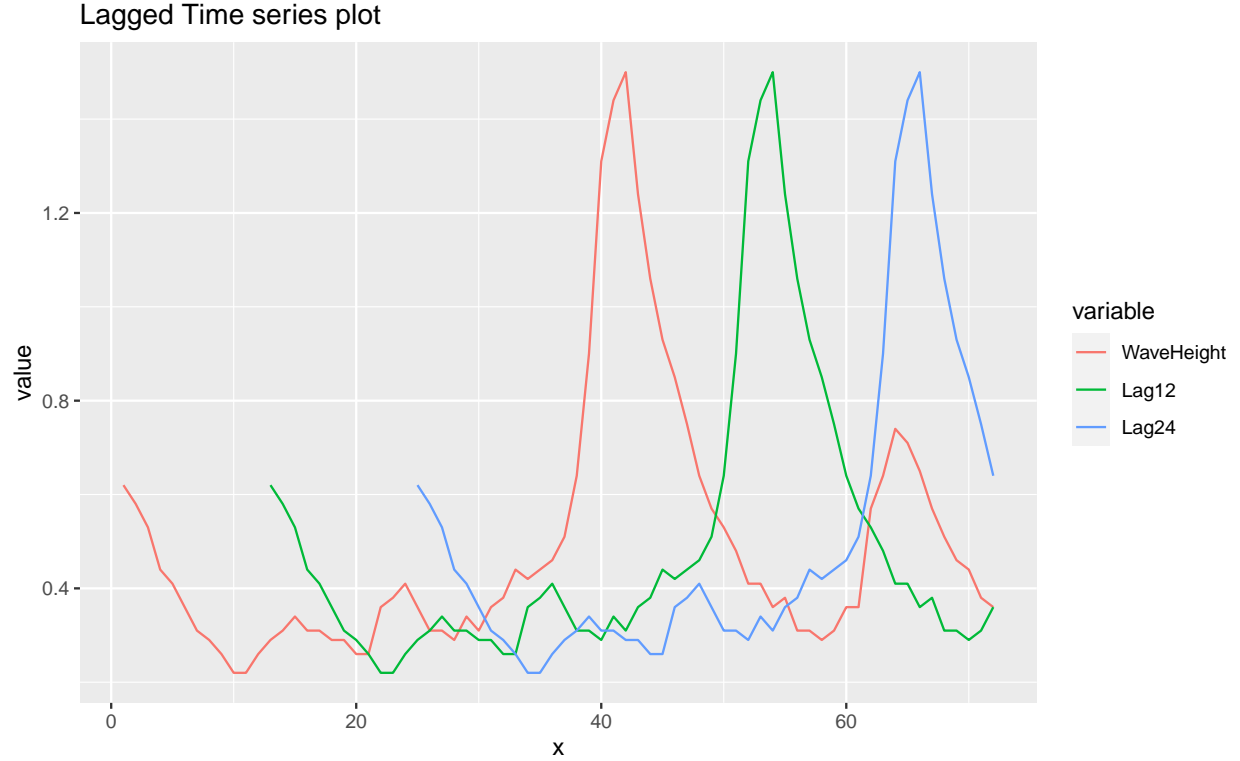
Model selection and cross validation

From a purest perspective, this is a time series forecasting model, whereby we aim to predict the wave height based on some seasonality and trend, however, what we've also got are some measurements of the sea state: Water Temperature, Salinity, Wind speed, Wave direction, which we could use to help improve our model and move away from a time series forecasting model towards a multivariate regression model.

Given we've got no baseline for our model we will build a baseline model which will consist mainly of the average:

$$\hat{y} \approx \bar{y}$$

And use it as the baseline to beat. As output measures to consider model performance we will use two metrics:

- Root mean square error

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

- Mean absolute error

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Results

Our top performer model seems to be the random forest regressor which outperforms all the other considered models on the cross validation test we used, and improves our baseline average model by reducing the RMSE by 70% and the MAE in 71% in comparison.

```
##              Model      RMSE        MAE
## 7      RandomForest 0.1375911 0.09811789
## 2 LinearRegression 0.2145928 0.15589945
## 3            Ridge 0.2371095 0.16673819
## 1          Baseline 0.4577411 0.33479927
```

Results Critique

However how good is our model, should we pursue alternatives? A good way to do this, would be by analysing the residuals out of our fitted model, in this case the top performer RandomForestRegressor, the summary performance parameters out of our cross validated samples are:

- RMSE: 0.1370972 m
- MAE: 0.09823503 m

Which can be interpreted as the expected model standard deviation against unseen data is of 0.137 m (or 14cm).

However how good is the achieved standard deviation? If we consider the mean expected wave is of 0.7667m we are looking at a coeficient of variation of $cv = std / mean = 0.18$, which means our standard deviation has the size of 18% our mean value, which is not despicable, and makes our confidence interval for the expected value of the significant wave height to oscillate $\hat{y} +/- 1.96\text{RMSE}$

```
##          Model      RMSE        MAE
## 7 RandomForest 0.1375911 0.09811789
```

At this point we've got several options to consider:

- Optimizing our selected ML algorithm, we could use grid search for hyperparameter tunning of the model in order to make it more robust against unseen data.
- Changing the probelm solving approach, move from a regressor to classifier.
- Conduct feature selection and dimenstionality reduction
- Consider variables interaction

However, there is a clear factor on this particular data we are analysing which we've clearly disregarded and it is the fact this is a time series, and time series usually present the following property **autocorrelation**, which could be loosely translated as *is the future value of my series influenced by the value which had yesterday?*, for a casual description of the concept we could refer to the wikipedia definition https://en.wikipedia.org/wiki/Autocorrelation, however there is plenty of literature in the topic.

If we were to plot the autocorrelation plot for our residuals, we could see the chart below, which shows a strong autocorrelation between the current value and its past values, which would make sense, one might argue the current sea state could be in part defined by the previous state and some additional external variables.

Series residuals_swh
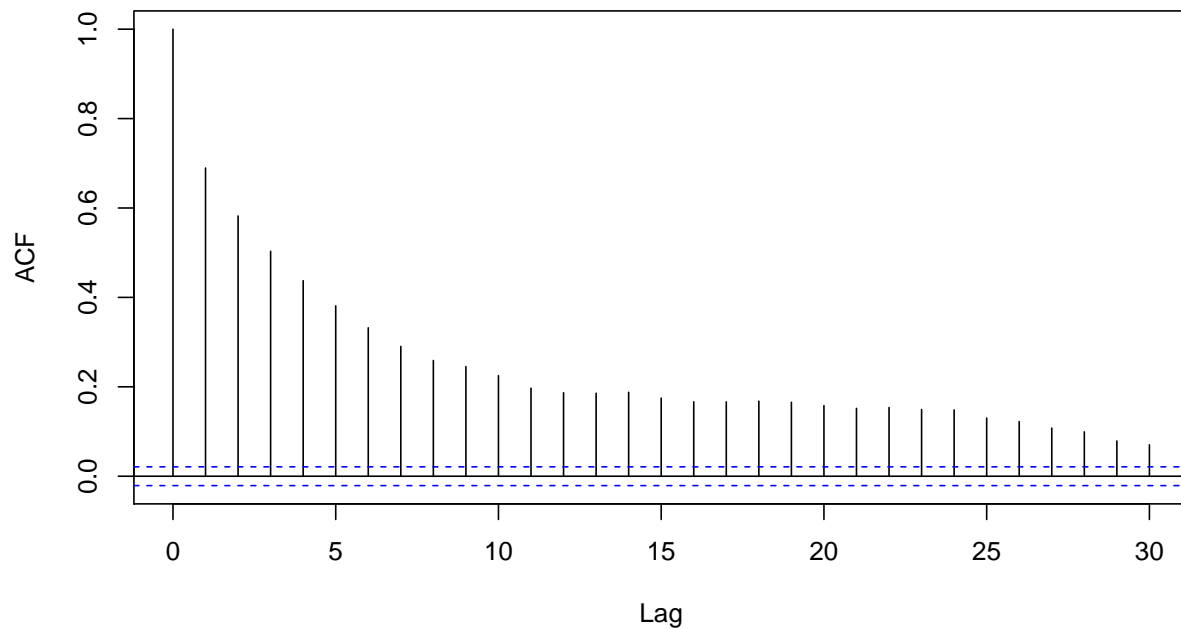
However if we were going to shift our approach from regression to time series forecasting, we would need to change several of the approached taken in our model, and given this has not been addressed as part of the course modules we will provide a general overview of the problem and how we could get a glance of the effect of autocorrelation in our current model.

If we were to build an autoregressive model, and there is plenty of literature in the topic, where we could use simple naive, naive seasonal methods, down to data smoothers such as Exponential or Smoothing splines, and for much more advanced complex problem ARMA, ARIMA and SARIMAX models, which have their own fine tuning requriements.

And from a scoring perspective, which is what will affect the most to our current approach, we will need to change the cross validation approach and use time series cross validation, which means we will train our model with the first n sequential data points to predict the from n + 1 to n + h next data points, where h is our forecast horizon, **sequential** being the key here, given we will try to predict the future using the past data points, if we were to use the future to predict the some of its future values we will get an overestimation of the model's performance.

For our example on what autocorrelation could be applied to our model, we will assume h = 1, meaning our target is to predict the next period t + 1 with the information up to t, plus the sea state information at t + 1 (which are all the columns we know the values for at t + 1).

The simplest way to do this, would be to shift our Wave height column 1 time period and use the past value to predict the next, below we've added the the column SWH_tp as the previous value for the Significant wave height.

Lagged Time series plot
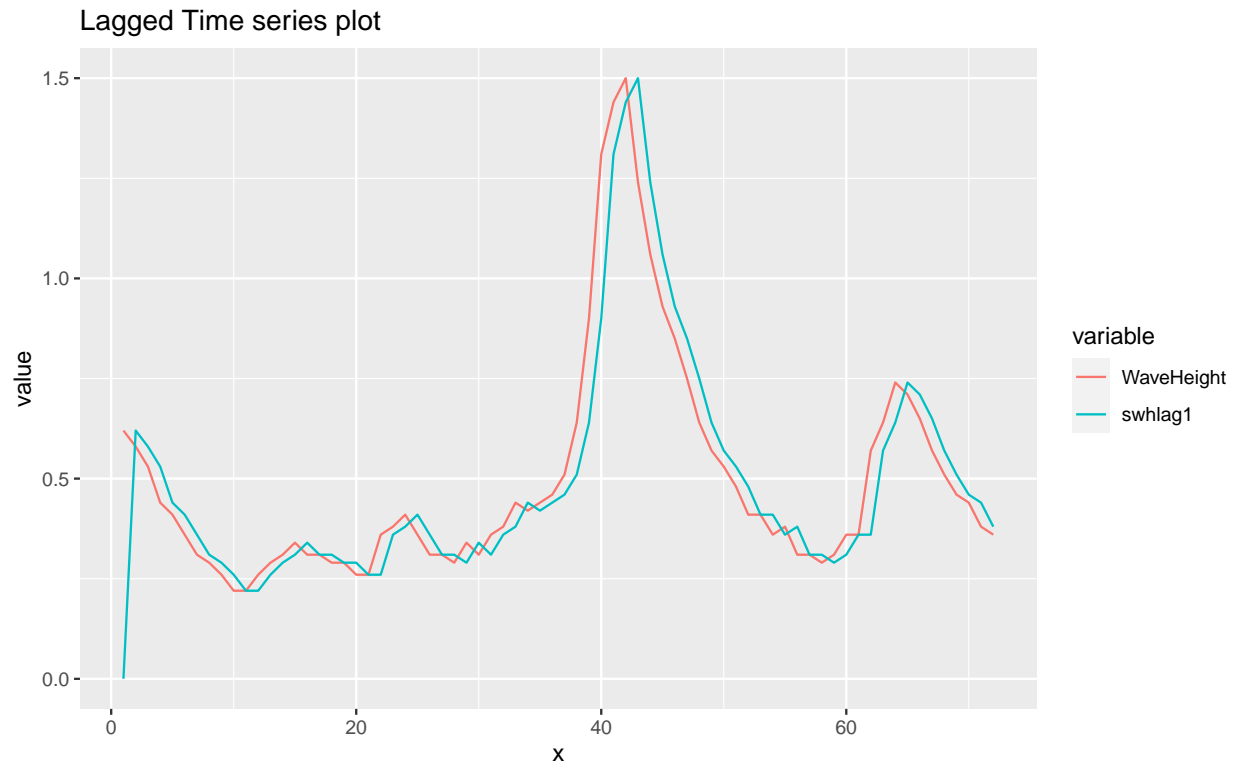
And if now we retrain the model and plot the autocorrelation function once more, we can see how it has dissapeared in its entirety meaning, we've modelled the effect of autocorrelation in our model.

Series residuals_swh
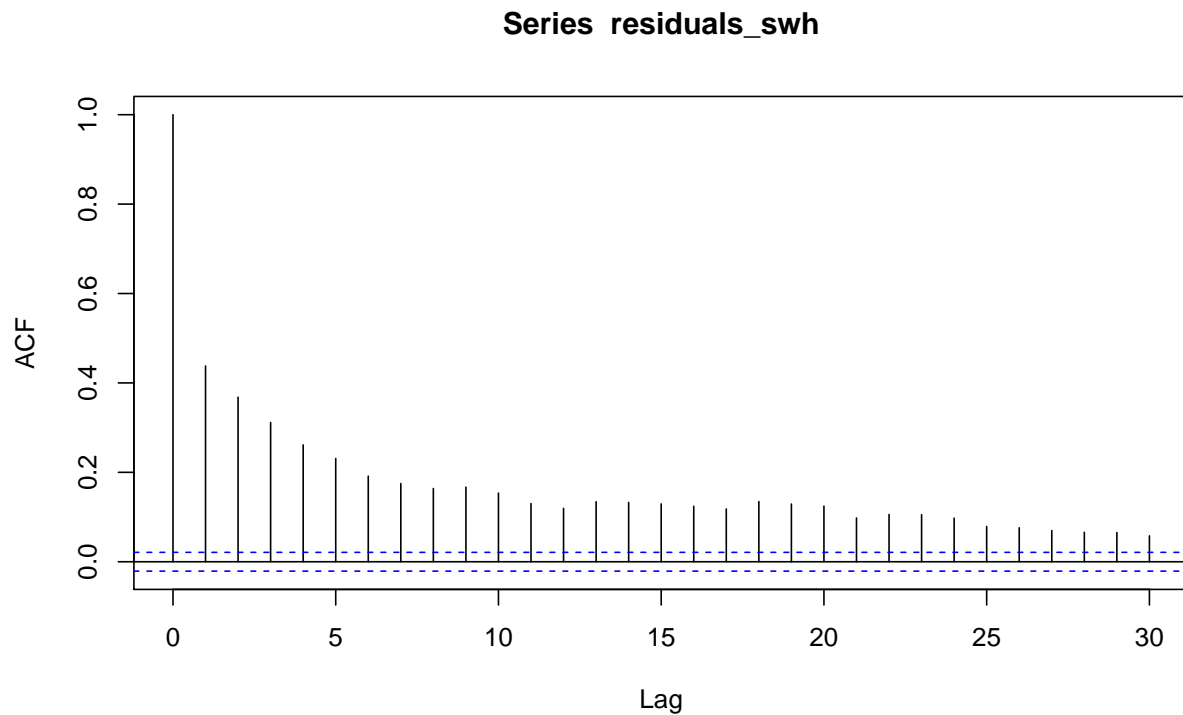
And if we were to measure the RMSE against unseen data we would obtain the following output measures:

- RMSE: 0.09722259m (previous 0.1370972 m)
- MAE: 0.06794571m (previous 0.09823503 m)

We've reduced our expected standard deviation against unseen data from 0.137m down to 0.097m (29% drop), and now the coefficient of variation is of 0.127, which means our standard deviation is about 13% of our variable mean, showing how we could improve our regressor and probably try also more time series oriented models when comparing performance.

```
##   mtry       RMSE  Rsquared        MAE      RMSESD   RsquaredSD        MAESD
## 1   50 0.09802685 0.9649882 0.06771029 0.004761088 0.003230721 0.001731437
```

Explaning considerations on what needs to be considered to build a robust time series model goes beyond the scope of this paper, but all the general rules for time series forecasting apply such as, guaranteeing time series continuity, identifying outlier, defining missing values and outliers strategy, normalizing our series and making it stationary, etc...

Conclusion

We've developed a regression model to predict the Significant Wave Height (m) using a set of input variables which measure the sea state at any given time period, and entertained the idea of swicthing to a time series model which could describe the dynamics of our data much better.

It is clear that in order to decide what to do next we should consider our intent behind making those predictions, is this data that might be used to make safety recommendations to the population?, are we after defyning the extreme sea state for water infrastructure development? or will this be used by our nautical fleet when deciding to go fishing or sailing?

There is a wide range or applications we could be using this data for, and depending on the customer needs we might need to build one kind of model or another, and probably adjust accordingly the success measure, but one very crucial factor to define would be our prediction time horizon, this is, how many periods h into the future our wave will need to be predicted for.

One of the biggest weaknesses of our developed model are the predictors used in our input matrix, given these are measured sea conditions **at the time of prediction**, this means we will probably not have this data available at the time of forecasting, meaning some of our features will have an associated degree of uncertainty at the time of forecasting if we are going to be using them, which builds the case even further towards switching into a time series forecasting model.

We've written this paper as a demonstration on how we could use some of the learned techniques on the DataScience course over EDX could be used for, but as previously mentioned if our aim was to build a robust prediction tool we will need to take a complete different approach by which we follow a twofold approach:

- Listen to the voice of the customer (aspect which can't be performed here given this is a technical paper)
- Build a tool which utilizes the most reliable and useful data at reach at the point of prediction.