

# Inferential Analytics Project: Predicting Surf Height

Paul Englert, Lukas Fahr, Bernardo Galvão and Felix Last

## Table of Contents

1. Introduction	1
1.1 Structure	1
1.2 Data Sources	1
2. Theoretical Framework	2
2.1 Econometric Specification	2
2.2 Methodology	2
2.3 Wave Generation	3
2.4 Limitations	5
3. Preliminary Analysis	6
3.1 Variables	6
3.2 Gauss Markov Assumptions	7
4. Transformation and Modelling	13
4.1 Target Variable	13
4.2 Interpolation	13
4.3 Determination of Lag	14
4.4 Transforming Wind Direction	16
4.5 Moving Averages	20
4.6 Validation	22
4.7 Fitting the Model	24
5. Conclusion	28
A Appendix	29
A.1 IPython Notebook	29
A.2 Data and Source Code	29

# 1. Introduction

The goal of this project is to infer the surf height at Costa da Caparica, Portugal based on buoy data of the preceding days.

## 1.1 Structure

In alignment with the process of creating a powerful model to predict surf height, this paper firstly focuses on the theoretical framework (chapter 1). Specifically the econometric specification, the underlying methodology, the physics behind wave generation and the limitations of the study due to the data are discussed. Secondly this paper explains the conducted preliminary analysis, regarding the variables and the Gauss-Markov-Assumptions (chapter 3). Thirdly, the applied transformations and the final modelling process is described (chapter 4) covering the target variable, the interpolation of missing observations, the determination of the lag, the transformation of the wind direction, the moving averages and the validation of the model. Finally, a conclusion is formed (chapter 5) providing a summary and an outlook to further enhancements of the model.

## 1.2 Data Sources

Buoy data has been retrieved from the UK national weather service “Met Office”. The API<sup>1</sup> provides 24 hours of marine observations from various buoys, vessels and stations, including the K1 buoy. The buoy is approximately 550 km west of France, which is a highly relevant location for predicting surf in Portugal. As a target variable, ideally observed surf height should be used. However, this data is not easily available. Instead, surf forecasts from surfline<sup>2</sup> have been used, which are known to be highly accurate. Per day surfline publishes the wave height and other parameters for four different times through an undocumented API<sup>3</sup>. The data has been retrieved in the evening after the forecast times to ensure highest accuracy. By that time, forecasts will have been adapted in accordance with local, real-time wind data and user observations.

---

<sup>1</sup> <http://www.metoffice.gov.uk/datapoint/product/marine-observations>

<sup>2</sup> [http://www.surfline.com/surf-report/costa-da-caparica-portugal\\_44509/](http://www.surfline.com/surf-report/costa-da-caparica-portugal_44509/)

<sup>3</sup> <http://api.surfline.com/v1/forecasts/44509?resources=surf,analysis&days=6&getAllSpots=false&units=e&interpolate=false&showOptimal=false>

## 2. Theoretical Framework

The study has been conducted based on a preliminary specification of the econometric model (section 2.1), as well as a specific methodology (section 2.2). Additionally the physics behind the generation of waves has been taken into account with regards to variable selection and transformation (section 2.3). The limitations due to the used data are briefly highlighted in section 2.4 and will be further discussed in the outlook (section 5.1).

### 2.1 Econometric Specification

The goal of this project has been to infer the surf height at Costa da Caparica based on buoy data of the preceding days. The variables used have been decided based upon their significance. The model used for the prediction has the following specification:

$$surf\ height_t(BUOY) = \beta_0 + \sum_{i=1}^k [\beta_k \times MA_{24}(BUOY_{k\ t-f})] + \varepsilon_t, \text{ where}$$

$BUOY$  is the set of explanatory features,

$k$  is the number of explanatory features,

$f$  is a fixed time offset, the lag, which may be different for each variable,

and  $MA_t(x)$  is the weighted moving average of feature  $x$  over the last  $t$  periods.

### 2.2 Methodology

The analysis was conducted using the Python programming language, which is complemented by many powerful libraries for data preprocessing and analysis, such as numpy, pandas and statsmodels. Jupyter IPython notebook allowed for interactive data analysis and visualization. The appendix contains two IPython notebooks, one for querying and transforming data and another for performing analysis and modelling.

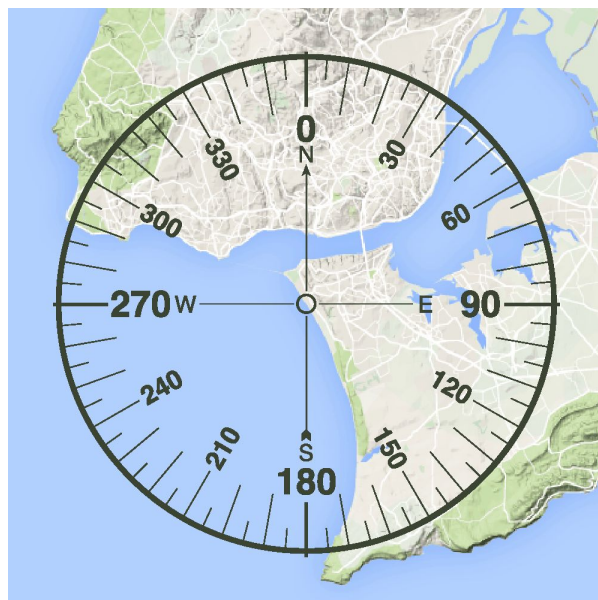
To retrieve training and test data, the above mentioned APIs were queried every 24 hours between May 18<sup>th</sup> and June 18<sup>th</sup> 2016. For this purpose, a Python script created with the first notebook was executed daily at 9pm through a Cron job on a Linux server. Retrieved data was stored in separate CSV files for each day.

## 2.3 Wave Generation

### 2.3.1 Swell

Winds and storms created by weather systems on the ocean propagate their energy onto the water generating waves that leave the boundaries of the system as *swell*. The force and the duration of the wind, as well as the distance the wave travels influence its shape and size when arriving at a coast. Consistent wind over a long period of time (implying a long distance) lets the swell gain energy, which corresponds to an increasing depth, as well as speed of the travelling water body. The swell will dissipate after leaving the systems area of influence, if it did not gather enough depth and speed or if it is not sustained by additional winds outside its origin system.

Regarding the studied coast of Costa da Caparica the dominant source of swell is a low-pressure system in the northern-atlantic. The weather system, a cyclone, moves latitudinal over the year. Due to the angle of the earth axis the cyclone is forced southwards during winter and northwards during summer. The shift of the cyclones position angles the incoming swell at the western-facing Costa da Caparica eastern during winter and southeastern during summer.



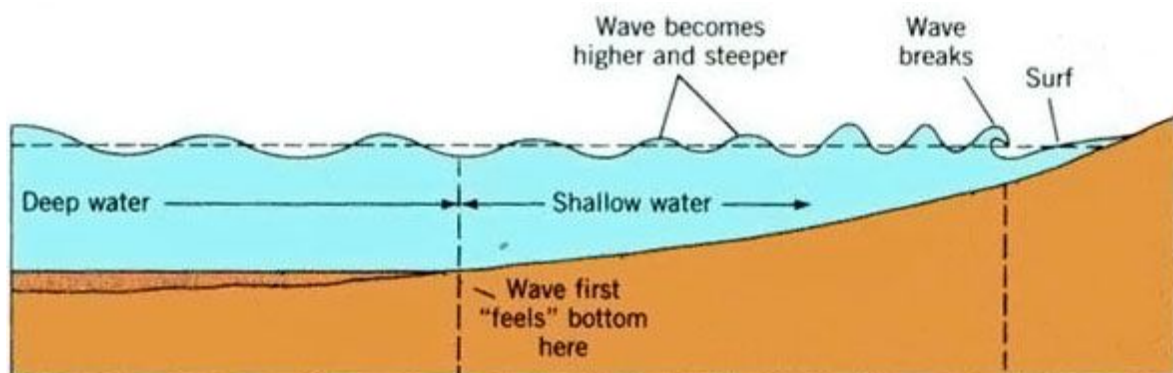
This creates strong seasonal changes in the waves reaching the coast. During winter Costa da Caparica is strongly exposed to directly incoming swell, whereas during summer the land below Cabo da Roca shields the coast from direct swell.

### 2.3.2 Period

As explained before, swell is generated and amplified by winds propagating their energy onto the water, allowing the swell to gain speed and depth. The speed of the swell is measured as the period, denoting the interval in seconds of two consecutive waves. It is therefore a proxy to the force of the swell. It can typically be observed that a larger period results in more distinct waves arriving at a coast.

### 2.3.4 Bottom contours

As mentioned above the depth of a waves body increases over distance. Once swell reaches a coast, the decreasing depth of the ocean forces the moving water body of the wave to grow above the water level. Additionally the movement over the coasts bottom creates friction slowing the lower part of the wave down. The wave will grow and shift forward until its surface becomes too steep to sustain itself - the wave will collapse and distribute its energy into the shallow water (see figure<sup>4</sup> below for graphical explanation).



### 2.3.5 Local Winds

Although winds are generating swell on the ocean, local winds at the coastal regions can radically influence the waves shape and size. According to their direction, local winds can be classified into three types: offshore, sideshore and onshore. Sideshore winds, travelling along the coast line do not have significant influences on the waves size - thus will not be further considered. Offshore winds travel from land towards the open ocean, in opposite direction of the incoming swell. Sufficiently strong offshore winds are able to sustain the waves growth above the limits explained in the above section, creating higher and more powerful waves.

<sup>4</sup> Source: <http://magicseaweed.com/help/near-shore-effects/breaking-waves>

Contrarily, onshore winds (travelling land inwards) force the waves to collapse in an earlier stage. Additionally, onshore winds create local wind swells that merge with the incoming ground swell, typically creating chaotic and disrupted waves.

## **2.4 Limitations**

The K1 buoy has been positioned to specifically measure the northern-atlantic swells, while the portuguese coast is influenced by multiple swells from different directions at a time. This implies that, ideally several buoys aligned spherically around Costa da Caparica would yield the highest accuracy. Using only the K1 buoys data renders the model possibly inaccurate.

Additionally this study is restricted with respect to detail, since local winds at Costa da Caparica are not being included in the model, ignoring possible negative, as well as positive effects on the surf height.

### 3. Preliminary Analysis

The explanatory analysis performed in this work was mostly done by means of iteratively transforming data and observing the effect on correlations and the final model, as elucidated in [chapter 4](#). However, some preliminary exploration was done to understand the data set and the assumptions that could be imposed on a linear model.

#### 3.1 Variables

The collected training data contains 25 columns, most of which are of type floating point number. The index, i.e. the first column, contains hourly timestamps for each observation. The time zone UTC+00:00 is the same for Portugal and the buoy.

##### 3.1.1 Dependent Variable

Most variables were discarded at the first step of analysis, because they belong to the target data set. This set contains a number of features about observed surf at the shore. Dependent features were not used as independent variables, because the project goal was to predict the wave height when the target data is not available. Thus the respective variables would not be available when using the model to make an inference.

Only maximum and minimum surf height were retained in order to compute the target variable, average surf height. Both variables are stated in feet, which has a fixed scalar conversion to meters.

##### 3.1.2 Independent Variables

Wind direction is the only variable of the explanatory set which is stated as string. It is measured as cardinal direction. Wind directions are conventionally stated as the direction from which the wind originates. This means, a value of ‘N’ for this variable indicates wind blowing from North to South. The precision of measurement goes up to the secondary-intercardinal direction, i.e. North West (NW, intercardinal) and West North West (WNW, secondary-intercardinal) are both valid values for this variable. As per convention, naval vessels measure wind speed in knots, corresponding to one nautical mile per hour.

Wave height is recorded in meters, which differs from the unit of the target variable. Wave height can be measured in different ways depending on whether the largest or smallest waves

are considered. It is assumed that significant wave height (average of highest  $\frac{1}{3}$  of waves) is provided, but this could not be verified.

The average period of waves is measured, indicating the frequency of waves passing the buoy. The unit of measure is seconds and values are strictly positive.

Lastly, the buoy also measures three temperatures (air, sea and dew point) which are all stated in degrees Celsius. Air pressure is measured in hectopascal, corresponding to one millibar.

## **3.2 Gauss Markov Assumptions**

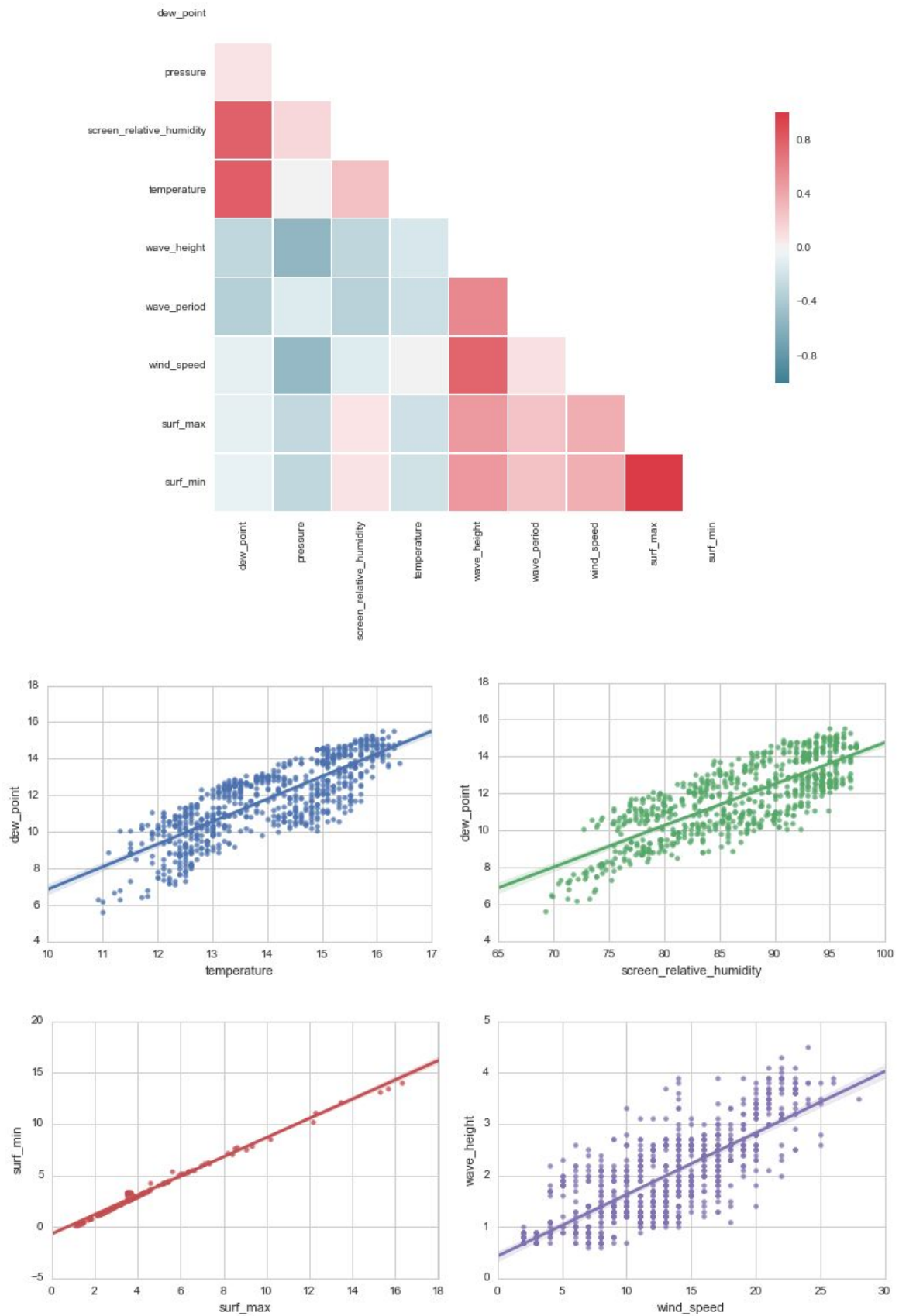
### **3.2.1 Linearity in Parameters**

The specification as outlined in [chapter 2](#) follows linearity in parameters, i.e. the target variable is a linear combination of parameters. Thus, the model does not violate this assumption.

### **3.2.2 No Perfect Collinearity**

By definition of the variables, there should be no variable which is a perfect linear combination of another. To identify strong linear correlations which may hint towards perfect collinearity, a correlation matrix was visualized as a heatmap. Examining the relationships among the highly correlated variables closer, it was verified that there is no perfect collinearity. The highest correlation is between maximum surf height and minimum surf height, but these are not explanatory variables. This assumption is therefore regarded as satisfied.





### 3.2.3 Zero Conditional Mean

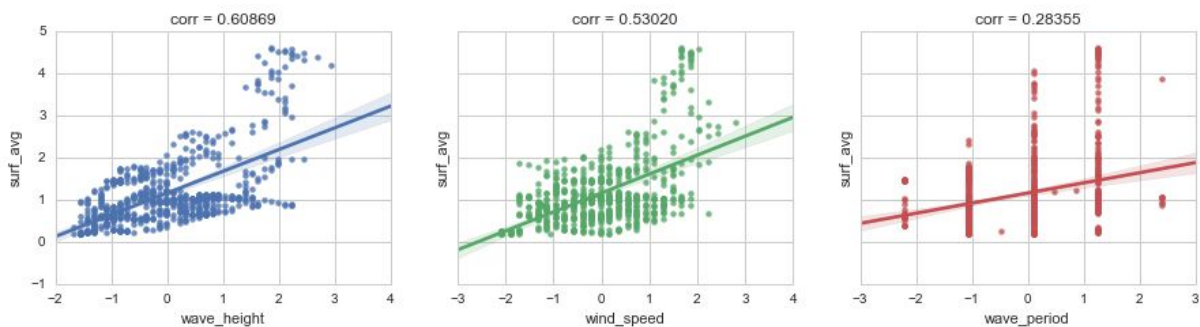
This assumption is violated if there are unobserved variables in the error term correlated with one of the independent variables. The violation leads to biased estimates of the model coefficients.

It is almost certain that there are other relevant variables which could not be observed. One example would be another buoy's measurements. That buoy's measurement would differ from the observations used in this study, but could potentially be highly correlated due to being affected by the same storms and cyclones.

The assumption is thereby violated, which will lead to an overestimation of the effects of the observed variables. However, the goal of this project was to predict surf height, and not to accurately estimate the true effects of each explanatory variable.

### 3.2.4 Homoscedasticity

This assumption cannot be immediately tested, because the relationship between dependent and independent variables is assumed to contain a lag. Thus, there may or may not be heteroscedasticity in unlagged data, while after applying a lag the opposite could be true. The following figures show the relationship of some variables to the target after the lag estimated in [chapter 4](#) has been applied.



Judging visually, the variance of errors does not appear constant. For wind speed, wave height and wave period, increasing values also lead to an increased variance of error terms. As a consequence, the homoscedasticity assumption is violated.

#### 3.2.4.1 Breusch-Pagan tests

Different specifications were attempted in order to find homoskedastic errors. For these specifications, only the variables wind direction, wave height and wind speed were selected

based on problem knowledge. Since the goal of this test was to determine the presence of heteroskedasticity, it is considered sufficient to include this subset of variables. The different specifications tested are:

#### Linear

$$surf\ height_t = \beta_0 + \sum_{i=1}^k [\beta_k \times BUOY_{k\ t-f}] + \varepsilon_t$$

#### Linear - Log

$$surf\ height_t = \beta_0 + \sum_{i=1}^k [\beta_k \times \log(BUOY_{k\ t-f})] + \varepsilon_t$$

#### Log - Linear

$$\log(surf\ height_t) = \beta_0 + \sum_{i=1}^k [\beta_k \times BUOY_{k\ t-f}] + \varepsilon_t$$

#### Log - Log

$$\log(surf\ height_t) = \beta_0 + \sum_{i=1}^k [\beta_k \times \log(BUOY_{k\ t-f})] + \varepsilon_t$$

where f was chosen in accordance with the lags estimated in [section 4.3](#).

The results are outlined in the following table:

	Linear	Linear - Log	Log - Linear	Log - Log
<b>BP</b>	241.15	62.018	243.24	98.978
<b>df</b>	2	2	3	3
<b>p-value</b> ≤	2.2e-16	3.412e-14	2.2e-16	2.2e-16

It can be concluded from these results that none of the specifications has homoskedastic errors. Thus, t- and F-statistics are invalid when estimated using OLS. Wald test and GLS could instead be used in order to perform significance tests in the presence of heteroscedasticity.

### 3.2.5 No Correlation of Errors

#### 3.2.5.1 t-Test for Autocorrelation of Order 1

This section outlines the results of testing for serial correlation of order 1. For this purpose,

$u_t = \rho u_{t-1} + v_t$  is estimated. The coefficient  $\rho$  was estimated to be 0.9183 with a p value of  $2e^{-16}$ . Thus, there is a high autocorrelation for lags of order 1.

#### 3.2.5.2 Breusch-Godfrey test for autocorrelation of order higher than 1

In order to find out the true order of the autocorrelation present in the model, the

Breusch-Godfrey test was chosen over the Durbin-Watson test. In a preliminary phase, lagged errors until lag of 10 were included in the right-hand side of the equation. However, from the fourth lagged error onwards, no lagged error was significant. For this reason it is presented here the Breusch-Godfrey test for three lagged errors, meaning that the following unrestricted equation was regressed:

$$u_t = \alpha_0 + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \alpha_3 u_{t-3} + \text{wave height}_t + \text{wind direction}_t + v_t$$

The results of this regression are outlined in the following table:

Coefficient of	Estimate	p value
(Intercept)	-0.002067	
u1	0.676475	< 2e-16
u2	0.176052	0.000138
u3	0.097630	0.009738
wind direction	0.044521	2.04e-08
wave height	-0.027084	0.000659

The test statistic, which asymptotically follows the chi-square distribution with  $p$  degrees of freedom,  $p$  being the number of lags on the error term and  $n$  being the number of observations after lagging, is computed:

$$nR^2 = \text{chi-square}(p) = 604.9285$$

As chi-square(3), for a significance level of 5%, is 12.838, the null hypothesis of no serial correlation is rejected.

### **3.2.6 Normality**

Since errors are heteroscedastic, they are not normally distributed with constant variance. This assumption is thereby violated.

### **3.2.7 Conclusion**

Except for the assumptions of linearity in parameters and no perfect collinearity, all Gauss Markov assumptions are violated. This should be regarded as an indication that a linear model cannot be used for the problem at hand. Rather, it points out the limitations of the specification and the interpretation of results.

The suspected violation of the linearity in parameters condition hints toward the fact that inferential accuracy of the linear model will likely be low since the model cannot capture the full complexity of the process of wave generation. However, a linear model also has limited risk of overfitting compared to polynomial functions and might provide less extreme results when extrapolating.

Coefficient estimation is biased, and the effects of the observed variables will be overstated. From this it can be derived that the coefficients may not be interpreted as the true effect of each variable. Instead it can be expected that the effects of wind speed, for example, measured at the buoy will be much less in reality than estimated by the model.

Finally, the heteroscedastic nature of the data implies that OLS will have higher sampling variance than other estimators. This discovery allows for the use of models which can deal with the inconsistent dispersion.

## 4. Transformation and Modelling

This chapter is concerned with preparing the data to be in the right format for being used in the model. Moreover, variables need to be transformed so that their ability to predict the target can be maximized. Lastly, the model is fitted to the data.

### 4.1 Target Variable

The target variable to be predicted is the average surf height in meters. The data set contains the maximum and minimum surf height in feet. Dividing the sum of maximum and minimum by two yields the surf average. The unit conversion from feet to meters was performed by multiplying the result by 0.3048.

### 4.2 Interpolation

The target variable was recorded four times per day, starting at 1 am and repeating every six hours. Since the explanatory data is observed hourly, the target variable was upsampled to fit this frequency. This was deemed necessary to ensure the availability of enough data to fit a model. Furthermore, the values could be interpolated, which can generally be considered safe. It is not clear which function describes the changing of the wave height, but linear interpolation should not create significant deviation from the “true” values and was thus used as upsampling method.

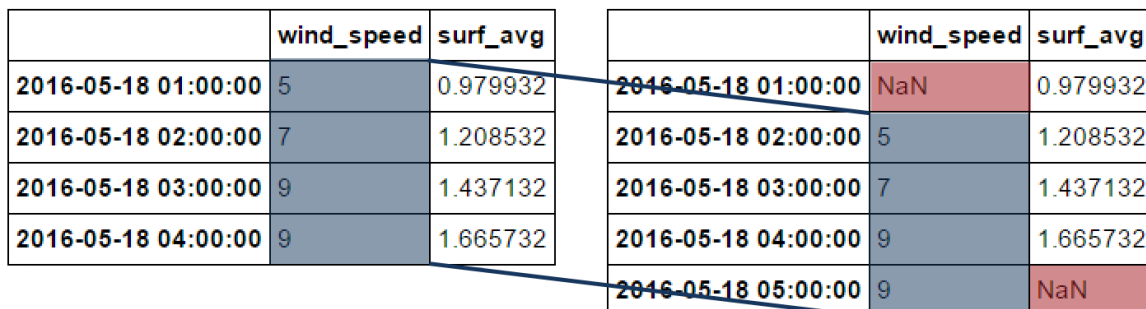
In addition, the hourly buoy data had few (less than ten) gaps which were filled by linear interpolation to simplify analysis and avoid missing values between observations. Buoy and Portugal lay in the same time zone, so there was no conversion necessary to join explanatory to dependent observations.

For wind direction, filling of missing values was performed in a different manner. After the conversion to degrees as discussed in [section 4.4](#), interpolation could have lead to undesired results. For example, if the wind shifted from NNW (340°) to NNE (20°), the linearly interpolated value between these observations would be S (180°), while N (360°) would be more appropriate. Since the number of missing values was very low and for matter of simplicity, missing values in wind direction were forward filled, meaning that they were replaced with the value of the last preceding observation. In the example discussed above, a missing value between the two observations would therefore be filled with NNW (340°).

### 4.3 Determination of Lag

As detailed in [chapter 1](#), waves are - amongst others - generated by wind propagating its energy into the water. This makes clear that when wind is measured at the buoy it takes time until the waves it generated reach a given spot at the coast. The exact time depends on the speed of the wave, which is influenced by many more factors than just the measured wind speed. Computing the exact time between each observation and its effect at the coast would be unfeasible with the scope of the given data set. Instead, in order to detect how much each explanatory variable lags, the linear correlation could be maximized.

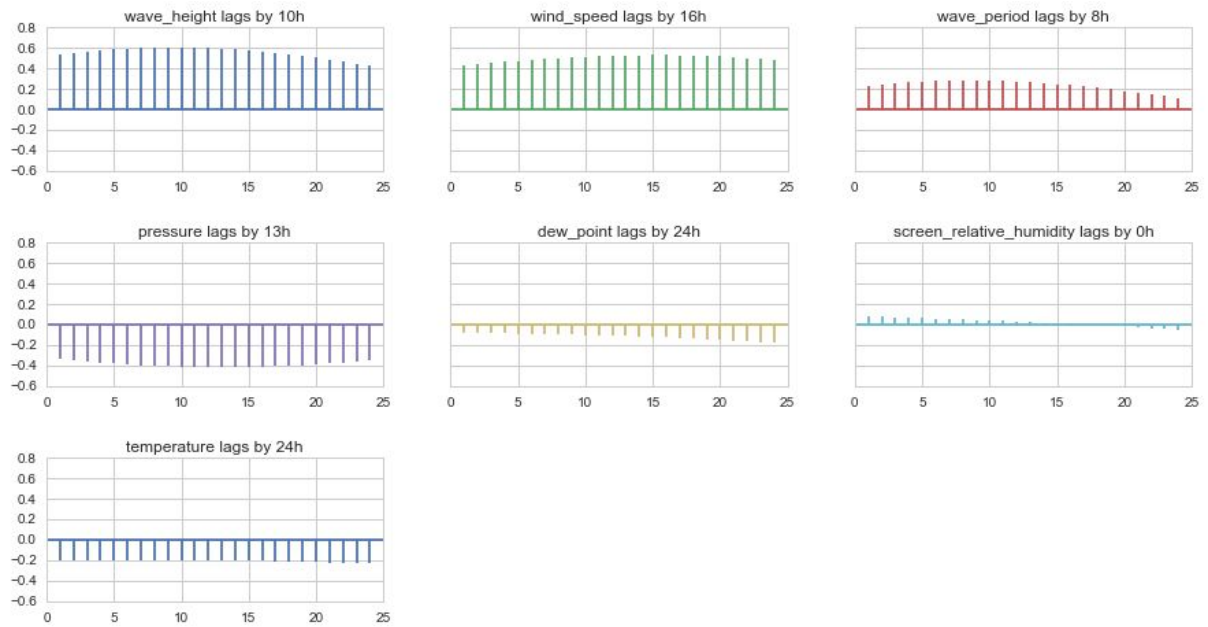
It is assumed that every observed variable does not take instant effect (i.e. lags), due to the mere distance between surf spot and buoy. Another assumption is that this lag is under 24 hours, which could be verified by calculating wave speeds of some observations. A lag can be implemented by shifting a column (variable) downwards while holding index, target variable and all other variables in place. The process creates missing variables above the lagged variable and below the other variables.



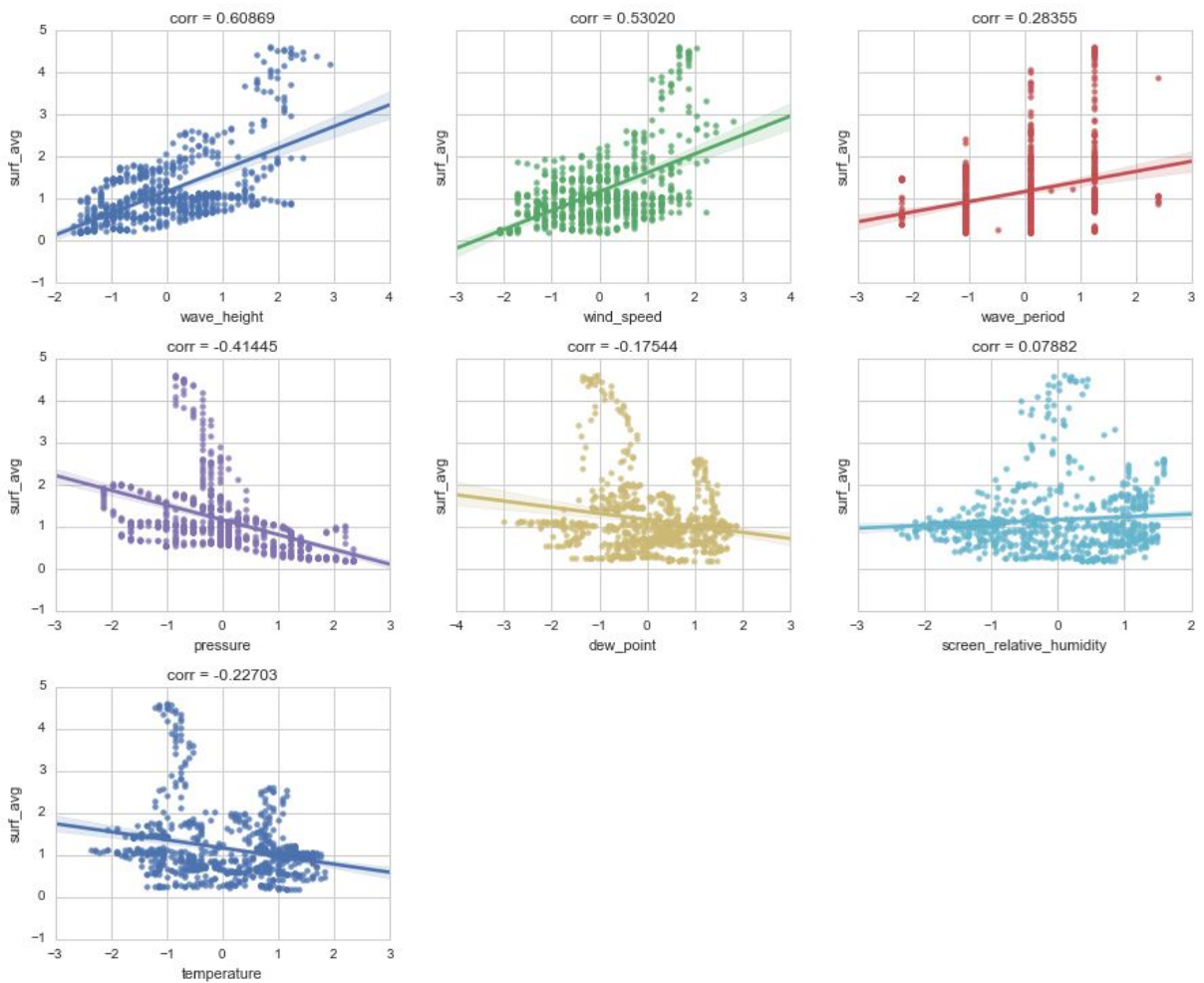
	wind_speed	surf_avg
2016-05-18 01:00:00	5	0.979932
2016-05-18 02:00:00	7	1.208532
2016-05-18 03:00:00	9	1.437132
2016-05-18 04:00:00	9	1.665732

	wind_speed	surf_avg
2016-05-18 01:00:00	NaN	0.979932
2016-05-18 02:00:00	5	1.208532
2016-05-18 03:00:00	7	1.437132
2016-05-18 04:00:00	9	1.665732
2016-05-18 05:00:00	9	NaN

To estimate the true lags, each variable was lagged by 1, 2, ..., and 24 hours and the Pearson correlation to the target was computed. The following figure shows one plot for each variable; the horizontal axis displays the applied lag, while the vertical axis corresponds to the correlation coefficient. In addition, the lag with the highest correlation is printed above each plot.



The distribution of each variable after the determined lags were applied is shown in the following figure.



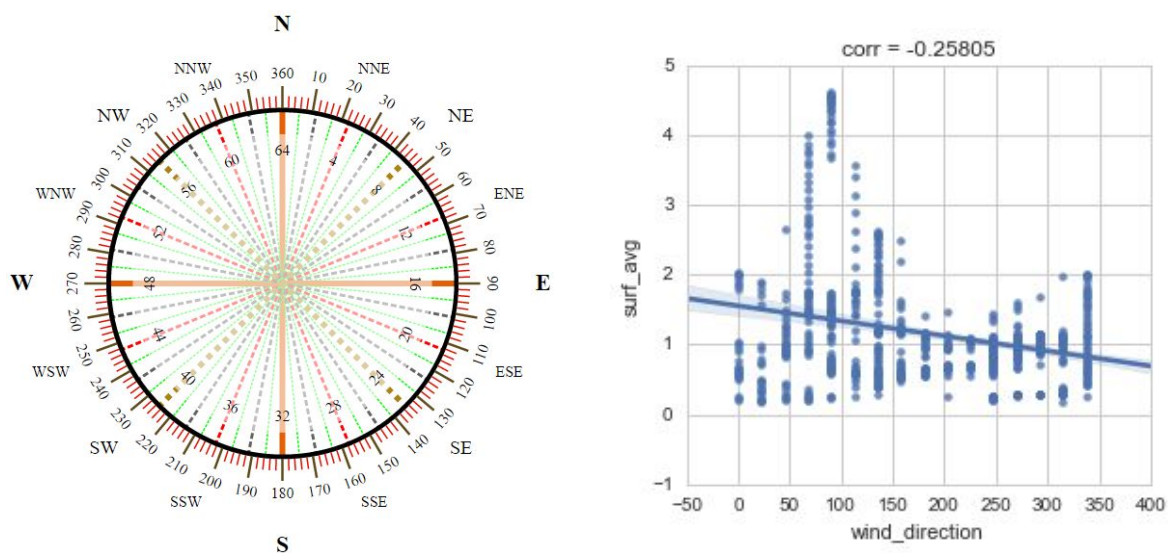


The variable wind direction was omitted in the lag analysis because it is assumed that its lag is tied to the lag of wind speed. It was therefore decided that wind direction is shifted the same number of periods as wind speed.

## 4.4 Transforming Wind Direction

Wind direction is the only nominal variable in the data set. As discussed in [chapter 3](#), it is measured in cardinal directions indicating the direction the wind comes from. Since the linear model cannot work with nominal data, this variable has to be transformed into a numeric type.

A natural first step is the mapping of cardinal directions to degrees. As can be seen in the left figure, each wind direction spans across an arc of a circle. The center of that arc is mapped to the textual wind direction to generate a degree encoded wind direction column. For North,  $0^\circ$  or  $360^\circ$  can be used; it was arbitrarily decided that  $0^\circ$  be used. The resulting column contains floating point numbers between  $0^\circ$  and  $340^\circ$  (corresponding to NNW). The resulting distribution can be seen in the right figure.



One issue identified with this encoding is that the distance between  $360^\circ$  and  $0^\circ$  should be equal to zero, but cannot be captured by a linear model. Two different approaches were tested to overcome this problem, which are elucidated in the following sections.

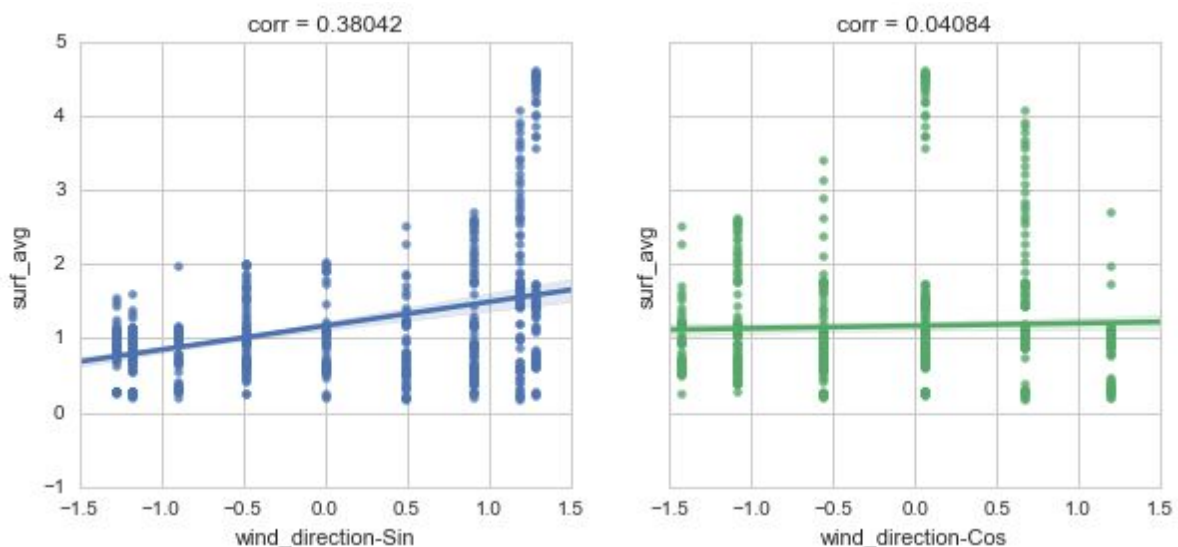
### 4.4.1 Sine Cosine Transformation

Degrees are a measure of an angle in a circle. In the so-called unit circle, sine and cosine can be used to describe an angle. Thus, degrees between 0 and 360 can be transformed into sine

and cosine without loss. As the following example shows, in this encoding there is zero difference between  $1^\circ$  and  $361^\circ$ .

	degrees		degrees-Sin	degrees-Cos
0	361	0	0.017452	0.999848
1	1	1	0.017452	0.999848

After transforming the wind direction with this approach, a scatter plot with a line of best fit shows the correlation. The Pearson correlation is quite low for both variables and they do not appear to be linearly correlated with the target. This is because a high value for sine or cosine does not have intuitive or known effects on surf height.



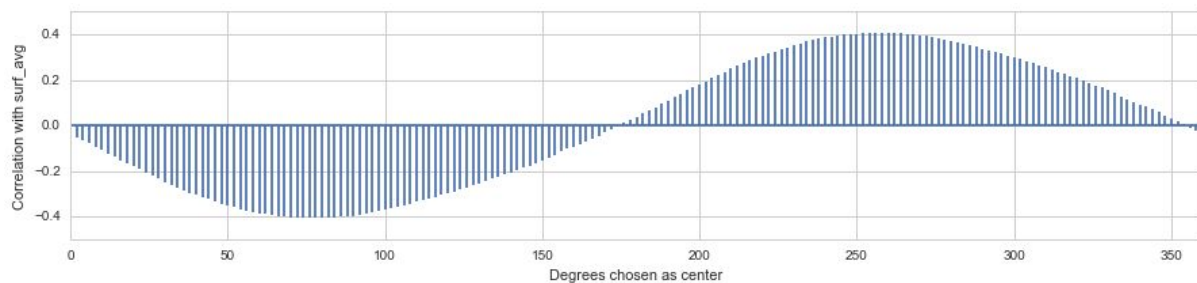
#### 4.4.2 Degree Deviation Transformation

For a linear model to be able to capture the relationship between wind direction and wave height, the effect needs to be linear, or at least monotonically increasing. As shown in the figure, the target location for prediction faces approximately South West (approximately  $230^\circ$ ). The importance of a spot's orientation is detailed in [chapter 1](#). According to the theoretical background of wave generation, the wind direction for maximizing surf height should be perpendicular to the surf spot, blowing toward it. In this way, the generated waves would travel in the direction of the surf spot.

Judging from the direction of Costa da Caparica, it is expected that the highest correlation with the target is achieved by creating a variable which contains the absolute distance of the wind direction to a center around  $230^\circ$ . Thus, wind coming roughly from the South West

would correspond to the minimum of  $0^\circ$  and the maximum distance should be  $180^\circ$  for wind coming from North East. In which direction the wind varies should not matter - the farther from the center, the smaller the expected wave height. This explains why an absolute distance was used.

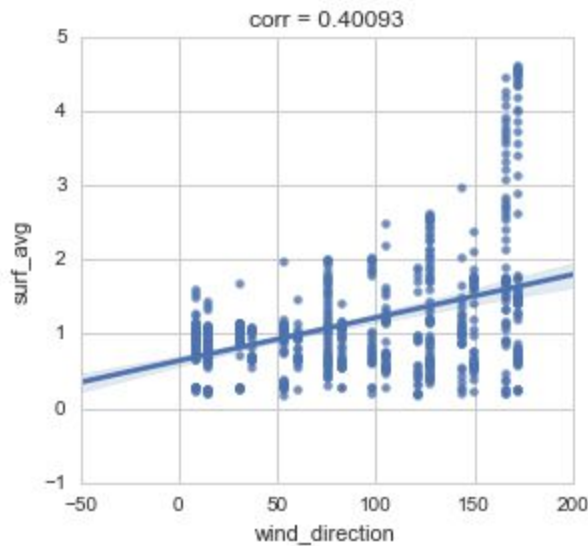
Rather than relying purely on problem knowledge, the goal of maximizing linear correlation was pursued. For this purpose, the deviation of the wind direction from every angle (in fact, angles  $0^\circ, 2^\circ, 4^\circ, \dots, 358^\circ$ ) was computed into a new variable. For each of these newly created variables, the correlation to the target was calculated. The result can be seen in the following figure.



As was to be expected, the graph is symmetrical, since observations were mapped from a  $360^\circ$  range to a  $180^\circ$  range.

The highest absolute correlation is at  $262^\circ$  (and  $82^\circ$ ). In other terms, wind blowing from  $262^\circ$  mapped to  $0^\circ$  resulted in the highest positive correlation. The positive correlation indicates that surf height is expected to be low when wind comes from that direction and high when it blows in that direction. This is almost the opposite of the expected result;  $230^\circ$  were expected to be the ideal source of wind, not direction.

The distribution of transformed wind direction set to the degree deviation from this  $262^\circ$  can be seen in the following plot.

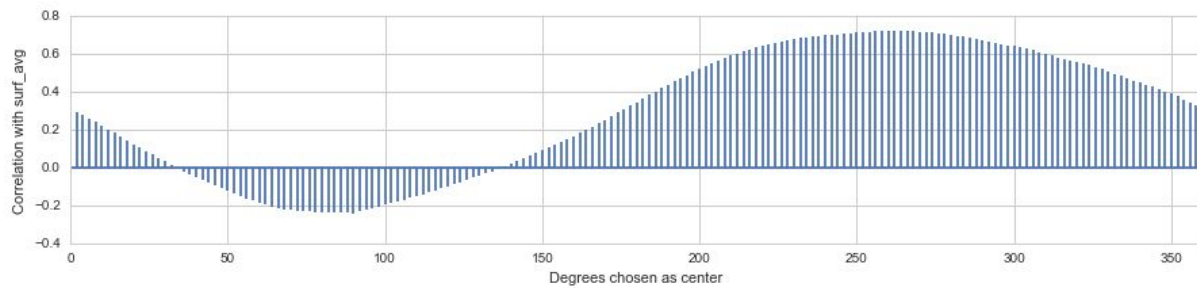


The correlation coefficient is still rather low with approximately 0.40. Low values of the target variable can be observed regardless of the wind direction. The linear correlation appears to be caused mostly by very high waves of over two meters which could be considered as outliers. A possible reason for the lacking relationship of wind direction and the surf height might be that waves are not strongly affected by wind when they pass the buoy; their main source is in cyclones further offshore.

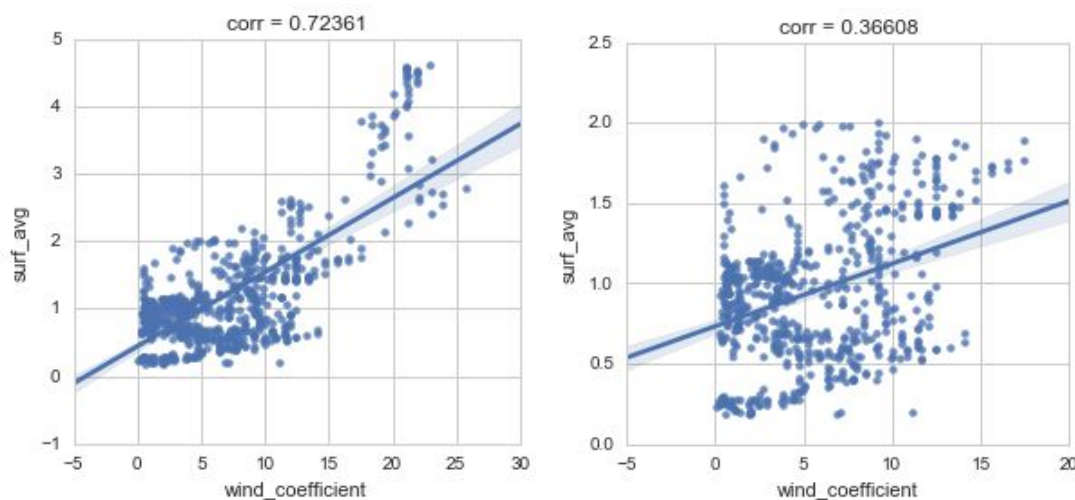
#### 4.4.3 Wind Coefficient

In another attempt to gather predictive power from measured wind data, wind speed and wind direction were combined into a single variable. This idea follows the rationale that strong wind blowing in the wrong direction doesn't increase wave height. Similarly, if the wind direction is perfectly aligned with the wave direction, but the wind speed is low, it will have no effect on wave height. For this purpose, a wind coefficient was introduced, computed by the following equation:

$wind\ coefficient_t = (wind\ direction\ deviation_t \div 180) \times wind\ speed_t$ , where the wind direction deviation is  $180^\circ$  for winds blowing from  $262^\circ$  and 0 for winds coming from  $82^\circ$ . This was derived by maximizing Pearson correlation as shown in the following graph:



The resulting correlation appeared to be quite high at first (left figure), but was weakened significantly when target values above 2m (outliers) were excluded (right figure).



## 4.5 Moving Averages

While predictions based on a model fitted to the previously transformed data showed reasonable results, one issue identified was that the predictions fluctuated much stronger than the true observations. While they were mostly centered around the actual observations the variance was much higher than the variance of the target values, which follow a clear line.

In an attempt to overcome this, fluctuation in the explanatory variables was reduced by computing rolling means. In this process, a window from one observation to a given number of preceding observations and calculates the average of values in that window. This is done for each observation of each variable. For the first observations, the window size must be decreased so that valid values can be computed.

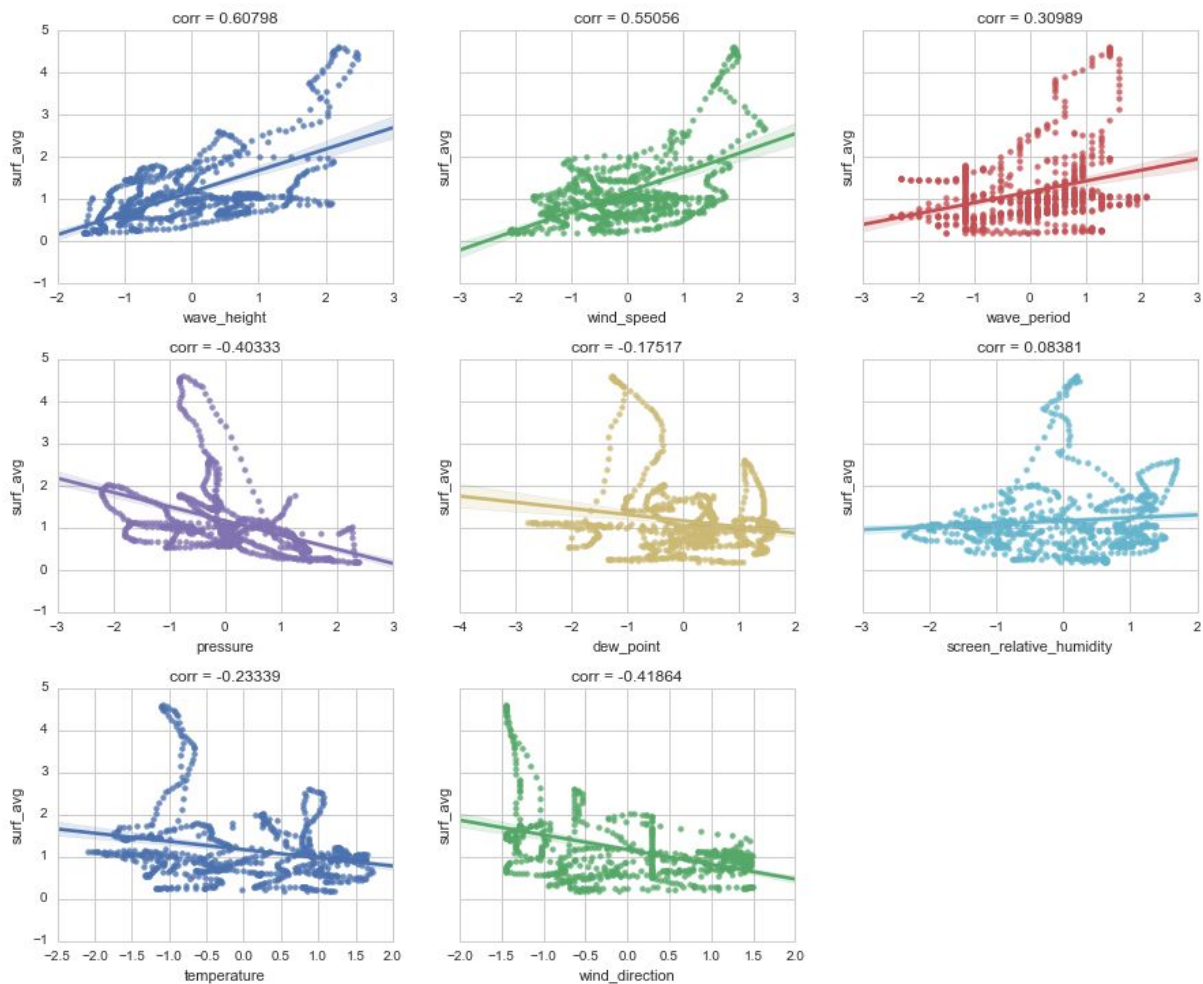
In order to give more importance to the most recent observation of the window, the moving average can be weighted. A common version of weighted rolling means is the exponentially weighted moving average, in which the weight of the observations in the window decreases

exponentially with inverse recency. In this way, observations at the recent end of the window are highly weighted, while observations at the other end are weighted close to zero.

After testing values between 1h and 15h, the window size was empirically set to 12 hours. The size of this window only minimally impacted Pearson correlation, but had strong effect on behavior of model curve. As a result, the function  $MA_t(x)$  from the model specification is restated as follows<sup>5</sup>:

$$MA_t(x) = \alpha \times x_t + (1 - \alpha) \times MA_{t-1}(x) \text{ for } t > \text{window size},$$

$$\text{where } \alpha = \frac{2}{\text{window size} + 1} \text{ and } \text{window size} = 12$$



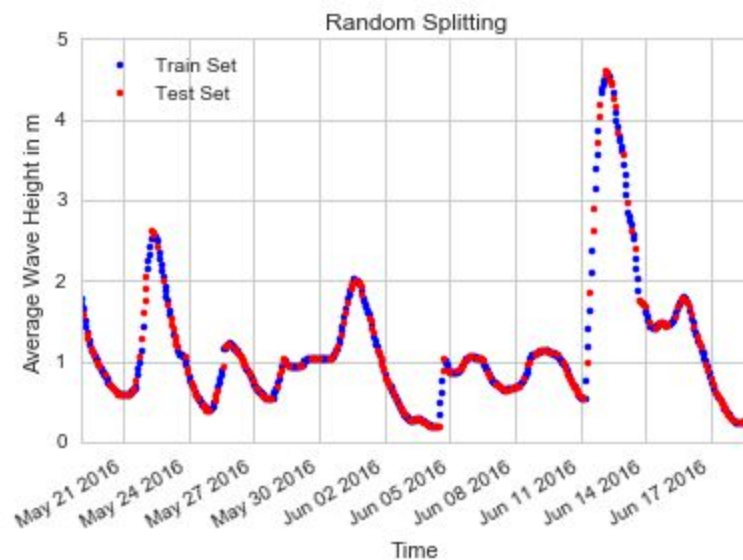
Unshifted by 4h and moving averages over 8h

<sup>5</sup> Specifying exponentially weighted moving average with span (window):  
<http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.ewma.html>



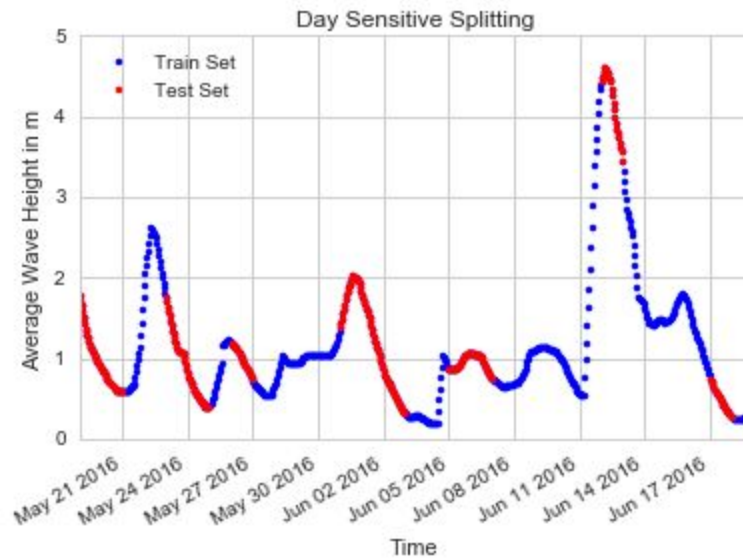
## 4.6 Validation

To ensure a sufficiently good prediction of the surf height the obtained dataset from the K1 buoy has been split into a training and a validation set. Typically these splits are performed by randomly sampling 60% of the dataset into the training set and 40% into the validation set respectively. Applying this technique to the time series at hand resulted in the following distribution of observation:

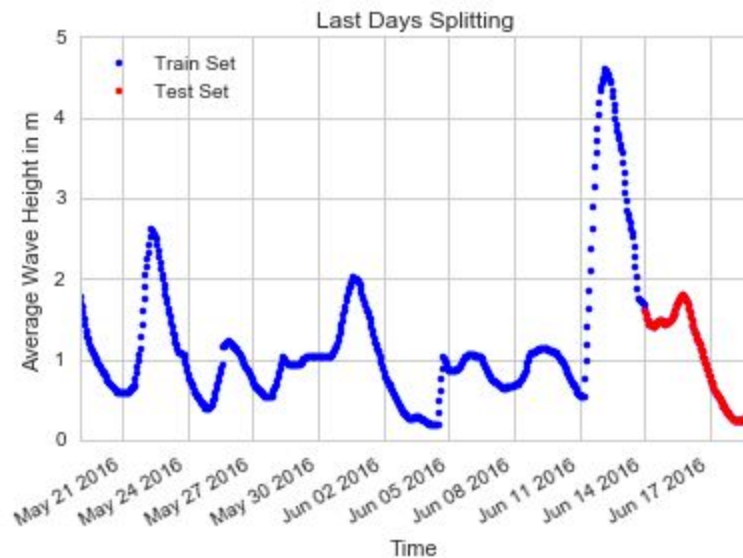


It becomes apparent that this kind of splitting will allow the model to simply interpolate the training observations, which will result in high predictive power regarding the validation set. Unfortunately the predictive capability obtained by the validation set might not reflect the true error of prediction. This does not imply a necessarily bad predictive power, but makes the model less comparable.

It could be possible to overcome this problem by combining the random splitting with the exclusion of complete days (24 observations, see figure below) and thus prevent the model from being trained too close to the validation observations.



When working with time series, it is common practice to retain a continuous set of training observation and using another continuous set for validation (see figure below). Consequently, the results of prediction on the validation set have been worse than sampling the training and validation set randomly without regards to timely continuity. Nevertheless the error is likely to reflect the true prediction error when applying the model to truly new observations.



The final split of the data set into training and validation set has been made with a ratio 87:13. This corresponds to 648 observations (27 days) in the training set and 96 observations (4 days) in the validation set (out of 31 total days, corresponding to 744 observations).

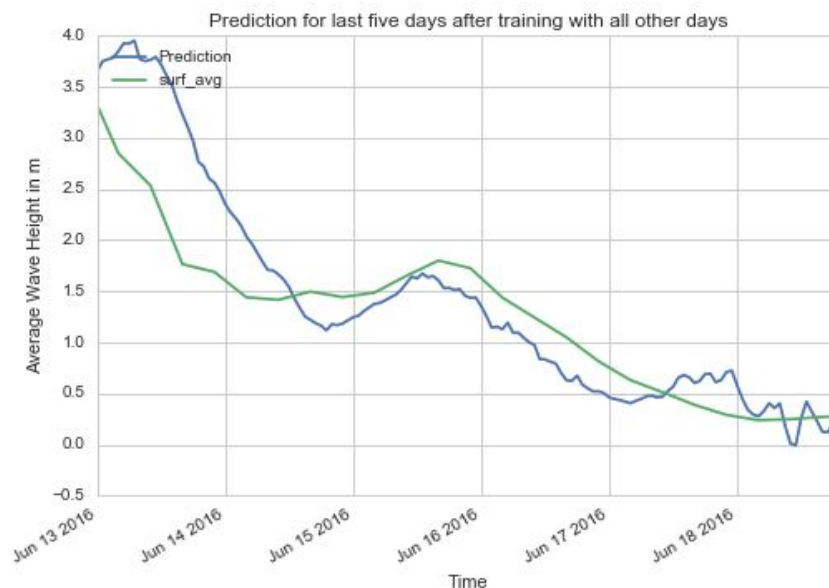


## 4.7 Fitting the Model

This section is concerned with fitting linear models using OLS based on different training data and analyzes the results.

### 4.7.1 Entire Feature Set

In this run, all features from the explanatory set are used to fit a model. The training data is split into a training set consisting of 26 days and a test set with the last 5 days. The resulting prediction can be seen in the following figure. All preprocessing methods mentioned in this chapter have been used to prepare the training and test data; wind direction was transformed with the degree deviation method and the wind coefficient was added to the feature set.



The model has an  $R^2$  value of 0.458 and a slightly lower adjusted  $R^2$  of 0.449. Its coefficients are the following:

<b>wave_height</b>	1.9710
<b>wind_speed</b>	-1.6388
<b>wave_period</b>	-0.6432
<b>pressure</b>	0.1647
<b>dew_point</b>	-0.0705
<b>screen_relative_humidity</b>	-0.0337
<b>temperature</b>	0.6120
<b>wind_direction</b>	1.4844
<b>wind_coefficient</b>	2.0071

The variables pressure, dew point and screen relative humidity have p-values above 0.05 and should be excluded by t-test. However, the t-statistics are invalid since homoscedasticity does not hold. The curve shows high fluctuations and varies by up to 1m off the true target.

#### 4.7.2 No Moving Averages

As outlined in [section 4.5](#), a model trained without using moving averages revolves around the correct curve, but shows strong fluctuation. This can be seen in the following prediction of a model using the entire feature set, and all discussed transformations.

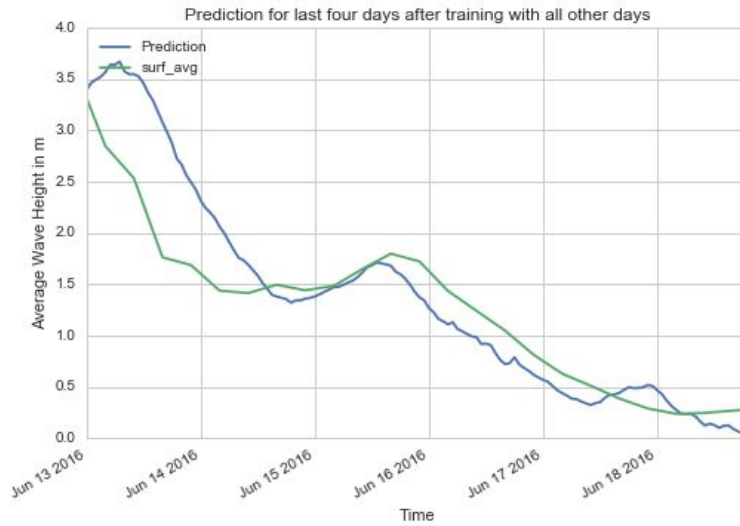


The adjusted  $R^2$  of the model is 0.374. The coefficients of the model are:

<b>wave_height</b>	1.0127
<b>wind_speed</b>	-0.6323
<b>wave_period</b>	-0.0523
<b>pressure</b>	0.0544
<b>dew_point</b>	-0.1076
<b>screen_relative_humidity</b>	0.0115
<b>temperature</b>	0.4631
<b>wind_direction</b>	0.6329
<b>wind_coefficient</b>	1.0423

### 4.7.3 Features based on Problem Knowledge

For training the model, the same split of training and test data is used as in the previous case. The same transformations are applied. However in this case, only variables that are expected to have a true impact on the target variable are included. Wave period is dropped due to its low correlation with the target. In addition, wind speed and wind direction are excluded since they are included in the function of the wind coefficient.



The model has an adjusted  $R^2$  value of 0.293 and both coefficients have p-values of 0. The specification is:

$$surf\ height_t = 0.6746\ EWMA_{12}(wave\ height_{t-11}) + 0.3149\ EWMA_{12}(wind\ coefficient_{t-17})$$

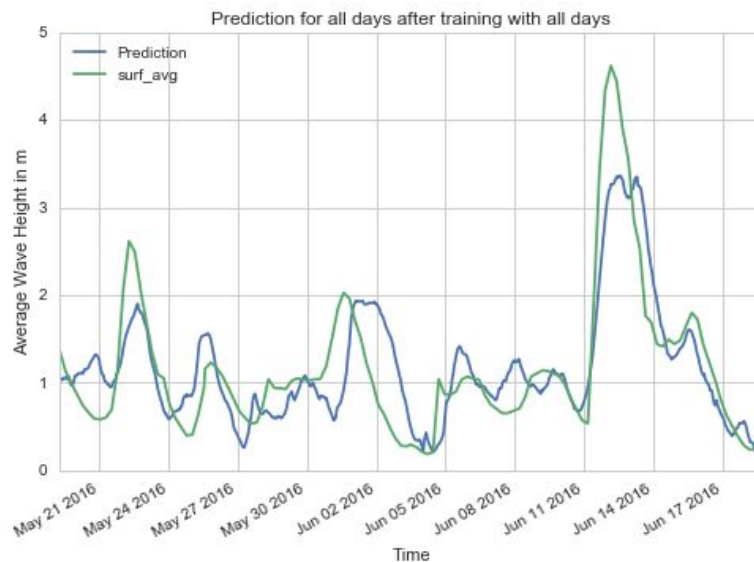
The predicted curve is much more stable than any of the other fitted models. Also, it barely has prediction errors larger than 0.5m, often even less. Another advantage of this model is its simplicity. It only uses two variables (three from the original feature set), but still reflects the movements of the true target curve.

### 4.7.4 Final Model

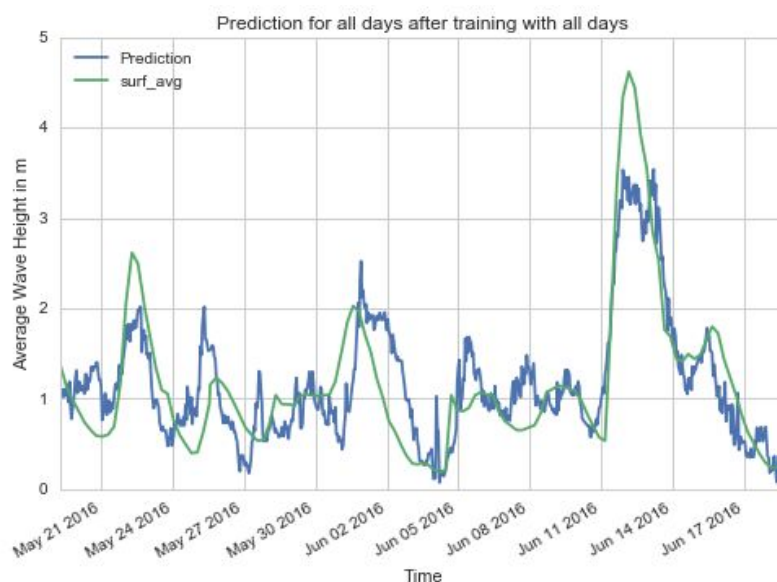
The final model is built on the two features selected based on problem knowledge, but is trained with all available training data. The resulting adjusted  $R^2$  is 0.233 and the specification is:

$$surf\ height_t = 0.3693\ EWMA_{12}(wave\ height_{t-11}) + 0.5475\ EWMA_{12}(wind\ coefficient_{t-17})$$

Predicting the entire data set, there are strong discrepancies visible. However, the general shape of the curve is resembled by the prediction. Strikingly, the prediction is lagging behind the true curve in some points, like between May 30<sup>th</sup> and June 2<sup>nd</sup>, while at other times it increases and drops at the same time as the true data.



For the purpose of further illustrating the purpose of moving averages, the prediction made by a corresponding model which does not rely on this transformation is shown in the following figure.



## 5. Conclusion

The model's prediction closely resembles the shape of the true process. Nevertheless, the predicted line is much less stable. Moving averages, however, greatly improve the smoothness of the curve. This steadiness is an important property since the observed surf height rarely changes quickly.

Another factor which affects the stability is the surf height of the period preceding the predicted period. This factor should be included by incorporating autoregressive properties into the model.

While the estimated lag seems to be close to the true travel time of waves to the shore, this isn't true for every period of the time series. This observation can likely be attributed to the fact that the time between observed buoy conditions and their effect at the targeted location depends on the speed of the waves - a factor, which is not incorporated in the model.

Additionally, and as mentioned before in [section 2.4](#), the surf height at Costa da Caparica is influenced by multiple swells from different directions. Since this study only considers the K1 buoy data, the precision could potentially be improved by adding other data sources.

A fundamental problem with the model applied to the problem is, that the complexity of the wave generation in Costa da Caparica is higher than a linear model can represent. Thus the model is not accurately depicting the relationship between surf height and swell or wind data.

# **A Appendix**

## **A.1 IPython Notebook**

Please find the IPython notebooks containing the source code online:

### **A.1.1 Analysis and Modelling Notebook**

This notebook contains the core of the analysis.

<http://nbviewer.jupyter.org/github/felix-last/wave-forecast/blob/master/wave-forecast.ipynb>

### **A.1.2 Data Acquisition Notebook**

This notebook was used to explore the APIs and to derive the data pull script later used to automatically gather data daily.

<http://nbviewer.jupyter.org/github/felix-last/wave-forecast/blob/master/grab-data.ipynb>

## **A.2 Data and Source Code**

Data and other source code can be found online as well:

<https://github.com/felix-last/wave-forecast>