FUNCTIONAL DATA ANALYSIS AND WAVE PROFILES DURING A STORM.

Joaquín Ortega  
CIMAT, A.C.  
Guanajuato, Gto. Mexico

Cristina Gorrostieta  
Brown University  
Providence, R.I., U.S.A.

George H. Smith  
University of Exeter  
Exeter, U.K.

ABSTRACT

Functional Data Analysis is a set of statistical tools developed to perform statistical analysis on data having a functional form. In our case we consider the one-dimensional wave profiles registered during a North-Sea storm as functional data. The waves are defined as the surface height between two consecutive downcrossings. Data is split into 20-minute periods and after registration of the waves to the interval [0,1], the mean wave is obtained along with the first two derivatives of this mean profile. We analyze the shape of these mean waves and their derivatives and show how they change as a function of the significant wave height for the corresponding time interval. We also look at the evolution of the energy, as represented by the phase diagram, as a function of significant wave height. The results show the asymmetry in vertical and horizontal scales for real data. To explore the departure of the data from the theoretical Gaussian model we simulated a storm having the same spectral evolution as the data. For each 20 minute interval the spectral density was estimated and used to simulate a Gaussian process sampled at 5 Hz. The statistical analysis was repeated and the results compared with those of real data, showing differences in the shape and the distribution of energy for the waves.

INTRODUCTION

In the last twenty years there has been considerable interest in the development of statistical tools for the analysis of data of a functional nature. The early work of Ramsay [2, 3] and Silverman [7] has been followed by a large amount of important research with applications in very different areas. Functional Data Analysis (FDA) as the field is known, has been successfully used to gain insight about data of human growth, temperature, econometric data, human movement, incidence of melanoma and speech, to name a few, (see [4, 5] and references therein). In this work we propose to investigate the shape of wind-generated random water waves using FDA techniques.

The statistical analysis of wave data is frequently based on the assumption that the sea surface elevation at a fixed point is a realization of a stationary Gaussian random process. Ocean wave data are often considered to be a realization of a stationary Gaussian random process, and much statistical analysis are based on these assumptions. However, it is well known that the Gaussian model fails in many situations and one of the reasons is that the symmetry of the Gaussian distribution implies that average Gaussian waves are both left-right and top-bottom symmetric. For a non-linear wave train this is not the case. FDA can provide tools suited to the study of the shape of wave profiles and contribute to more accurate statistical wave models that closer describe the observed waves.

In this work we use FDA techniques to examine the shape of waves during a 1999 North Sea storm that lasted over 81 hours. The data were split into 244 intervals of 20 minutes duration and for each interval we look at the mean wave profile, the standard deviation and the first two derivatives of the mean wave profile. We also look at the relation between these objects and the significant wave height, considered as a measure of the severity of the sea.

From the energy spectrum of each interval we simulated a Gaussian process of equal length. That is, we simulated a Gaussian storm having the same spectra as the real one for each 20 minute interval. The same tools were used for the analysis of both sets of data and the results show the differences existing between real (non-Gaussian) and simulated (Gaussian) waves, and how these differences vary with significant wave height.

FUNCTIONAL DATA ANALYSIS

Functional Data Analysis (FDA) as an area of Statistics has had a very strong development in last 20 years or so and is still a
very active field of research. To describe its aims we quote Ramsay and Silverman: The fundamental aims of the analysis of functional data are the same as those of more conventional statistics: to formulate the problem at hand in a way amenable to statistical thinking and analysis; to develop ways of presenting the data that highlight interesting and important features; to investigate variability as well as mean characteristics; to build models for the data observed, including those that allow for dependence of one observation or variable on another; and so on. ([4], p.1).

Very frequently the first step is the functional representation of data that has been recorded discretely. If the data for the \( i \)-th function are \( y_{i,1}, \ldots, y_{i,n} \) and were recorded at instants \( t_j \), for \( j=1, \ldots, n \) then one needs to transform the relation \( x_i(t_j) = y_{i,j} \) to a function \( x_i \) with values \( x_i(t) \). Some degree of smoothing may be adequate depending on the presence of measurement errors.

To have a common representation for all functions a basis representation is used and basis selection is a fundamental step in many cases. Popular choices are trigonometric functions and B-splines, but there are others. For computational reasons, and since there are only a finite number of data points, a finite set of functions is used, and the number of elements in the basis is related to the degree of smoothing in the representation: The smaller the basis, the smoother the functions.

Let \( g_1, g_2, \ldots, g_k \) be the basis functions, then the \( i \)-th function on the sample is written as

\[
x_i(t) = \sum_{r=1}^{k} c_{i,r} g_r(t)
\]

where \( c_{i,r} \) for \( r=1, \ldots, k \) are the coefficients of the expansion for the \( i \)-th function in the sample. The coefficients can be obtained using a least squares approach: if the data points are \( y_{i,1}, \ldots, y_{i,n} \) and were recorded at instants \( t_j \), for \( j=1, \ldots, n \) then one seeks the coefficients that minimize the expression

\[
\sum_{j=1}^{n} \left( y_{i,j} - \sum_{r=1}^{k} c_{i,r} g_r(t_j) \right)^2.
\]

Alternatives to this procedure are weighted least squares, in which the terms of the sum above are multiplied by weights \( w_j \), and the roughness penalty approach, in which a new term is added to the sum which penalizes the irregularities in the estimated function.

Another important point is the choice of \( k \), the number of elements in the basis. If \( k \) is large, it will be possible to fit the data accurately but the influence of noise or errors in the measurements will be important. To filter out the noise one can reduce the number of elements but not too much, or one will risk losing important features of the data. Hence the choice of \( k \) entails the usual trade-off between variance and bias which commonly appears in statistical decisions. A number of algorithms have been proposed for choosing \( k \) adequately.

**DATA**

Data was recorded from the North Alwyn platform situated in the northern North Sea, about 100 miles east of the Shetland Islands (60º48.5' North and 1º44.17' East) in a water depth of approximately 130 metres. There are three Thorn EMI Infra-red wave height meters mounted on the platform and their heights are between 25 and 35 metres above the water. The data are recorded continuously and simultaneously at 5 Hz and then divided into 20 minute records for which the summary statistics of \( H_s, T_p \) and the spectral moments are calculated. For data with \( H_s > 3m \) all the surface elevation records are kept. Further details are available in Wolfram et al. [8]. Only data from the North East altimeter are used here.

Figure 1. Significant wave height calculated every 20 minutes. The line is a 9 point moving average.

The data set examined consists of 244 records, each of 20 minutes duration, sampled by the altimeter at a rate of 5 Hz., which occurred between midnight on December 23rd and 9.00 a.m. on December 26th 1999. The evolution of the significant wave during the storm is presented in figure 1. The color code for the significant wave height will be used for other figures. A similar graph for the Spectral Peak Period, calculated using the WAFO software, is given in Figure 2.

Figure 2. Peak period calculated every 20 minutes. The line is a 9 point moving average.

These measurements of the surface elevation at a fixed point are usually seen as a time series sampled at a fixed frequency and the classical approach is to analyze it as a random process.
Since our main interest here is to study the shape and variability of waves this series was segmented to obtain the individual waves that compose the series, where a wave was defined as the surface elevation between two successive downcrossings of the mean level. For this analysis we define the height to be the range between maximum and minimum values of elevation occurring between successive zero-crossing. Hence, for each interval of 20 minutes, we have a set of waves of different lengths (periods) and heights. These individual waves are the observations we consider for Functional Data Analysis.

AMPLITUDE AND PHASE VARIATION

When looking at the shape of waves one should distinguish two different types of variation, in amplitude and in phase. The first corresponds to changes in the magnitude of the main characteristics of the wave, such as positive and negative amplitudes, while the other is related to the instant when these features occur, such as the zero upcrossing and the absolute maxima and minima. In this work we focus on amplitude variation.

To explore the effect of significant wave height on phase variation we considered the three main features of a wave, aside from the start and end points, which are the upcrossing in the middle and the absolute maximum and minimum. The period of each wave was linearly transformed in time to $[0, 1]$ and the position of these three features in this scale was determined. Figure 3 shows the estimated probability density for the location of these three points for all 20 minute intervals. Each curve corresponds to an interval and is colour-coded according to the value of the significant wave height for the period, according to the colours of figure 1. It is readily seen that all curves for a given wave feature are very similar and thus $H_s$ has no effect on the location of these wave characteristics.

To study amplitude variation in the presence of variations in phase for the individual waves it is necessary to transform time so that the main features of the waves occur at approximately the same time. Otherwise, the mean waves obtained by averaging wave profiles during a time interval may not be representative of the mean behaviour of the waves. The process of transforming time for each wave to carry it to a common time interval ($[0, 1]$ in our case) is known as registration and will be explained in more detail in the next section.

REGISTRATION AND SMOOTHING.

Since data was recorded discretely, the initial step is its representation as a function using a common basis. Since there is a wide variation in the length of waves during a 20 minute interval, we started with the registration of the data. There are several possible ways of doing this ([5] Ch. 7), but in order to modify as little as possible the shape of the original waves, for each one we used a registration function $h_i(t)$ that synchronizes the three zero crossings of the wave. Thus if the $i$-th wave starts at $t = a$, has a zero upcrossing at $t = b$ and ends at $t = c$, the registration function satisfies $h_i(a) = 0$, $h_i(b) = 0.5$ and $h_i(c) = 1$. Since we also want to study the derivative of the mean waves, we also want the derivative of the registration function to be continuous. Hence, a quadratic registration function was used.

After registration the waves were represented in functional form with the following procedure. First, each wave was individually described in a functional way by using a b-spline basis so that the resulting function interpolated the original data. After all the waves had their own functional profile, they were evaluated at a sequence of 500 equally spaced points in $[0, 1]$ with the objective of creating a common basis for all the waves. This sequence was used to create a b-spline basis with knots at the grid points, a total of 502 basis functions. This basis was used to represent all the waves.

MEAN AND STANDARD DEVIATION

The sample mean and standard deviation are obtained using the following formulae:

$$\bar{x}(t) = \frac{1}{N} \sum_i x_i(t); \quad s(t) = \left(\frac{N - 1}{N}\right) \sum_i \left(x_i(t) - \bar{x}(t)\right)^2$$

Figure 3. Mean wave for 244 intervals of 20 minutes as a function of significant wave height.

Figure 4. Standard deviation for 244 intervals of 20 minutes as a function of significant wave height.

For each 20 minute interval the mean and standard deviation was obtained and the resulting functions are shown in figures 3
and 4, with a colour coding corresponding to the significant wave height of the time interval introduced on figure 1. It is readily seen that the oscillation of the mean waves increases with \( Hs \) and that there is an asymmetry between the two halves of the waves, which shows more clearly in the standard deviation. There is more variability in the second part of the wave, which corresponds to the crest.

**DERIVATIVES OF THE MEAN WAVES**

The first two derivatives of the mean wave for each interval were calculated (figures 5 and 6). They show more clearly the difference with pure sinusoidal waves, since the derivatives of a sine are the same sine function with a phase shift of \( \pi/2 \). As can be seen, the first derivative's shape also changes with the significant wave height of the period, positive and negative values become more pronounced as the significant wave height increases, while the zero crossings, which correspond to the extreme values of the wave profile, remain approximately fixed.

![Figure 5. First derivatives for the mean wave for 244 periods of 20 minutes as a function of significant wave height.](image)

![Figure 6. Second derivatives for the mean wave for 244 periods of 20 minutes as a function of significant wave height.](image)

The absolute value of the first derivative is a maximum at the zero crossings of the wave. Also there is an apparent symmetry with respect to a vertical line passing through the midpoint of the wave. This symmetry resembles that of a cosine function, although the actual values differ. The second derivative is also dependent on the significant wave height (figure 6). The acceleration is null at the endpoints and the midpoint, i.e., at zero crossings. As for the first derivative, the shape is similar for all intervals but the values increase with significant wave height. In this case there is an apparent symmetry with respect to the midpoint. The shape of this derivative is clearly different from a sine function. Maxima and minima for the acceleration do not occur at the minima and maxima of the wave but are closer to the midpoint. The mean position for the maximum is 0.376 with 95% of the values falling in the interval (0.343, 0.397). As for the minimum, the mean position is 0.636 and the corresponding 95% interval is (0.613, 0.667).

**PHASE-PLANE PLOTS**

A very useful tool to investigate the evolution of energy during a wave cycle is the phase-plane plot, which is simply a plot of first (x-axis) versus second (y-axis) derivatives, i.e., acceleration vs. velocity. A sinusoidal wave \( x(t) = \sin(2\pi t) \) corresponds to an oscillating pendulum or a spring with a weight at the end. Using appropriate scales in both axes, the phase-plane plot in this case will be a circle. The graph describes the exchange of potential energy, proportional to acceleration and kinetic energy, which is proportional to the square of velocity. The derivatives are \( x'(t) = 2\pi \cos(2\pi t) \) and \( x''(t) = -(2\pi)^2 \sin(2\pi t) \). At points \( \pi, 3\pi, \ldots \) the first derivative is 0 while the second has maximal absolute value, which correspond to a situation where the spring is at one of its endpoints and all the energy is potential. On the other hand, at points \( 0, 2\pi, \ldots \) the situation is reversed: there is 0 acceleration while the velocity has maximal absolute value and the spring is at its equilibrium position. The evolution of the graph shows the exchange of potential and kinetic energy. Moreover, the amount of energy in the system is related to the size of the graph.

![Figure 7. Phase-plane diagram for the mean waves.](image)
energy as the wave evolves. Moreover, the distance of any given point on the graph to the centre is proportional to the total energy present in the wave at that moment.

Figure 7 presents the phase-plane plot for the mean wave for all 20 minute intervals. The graph evolves clockwise starting in section A, which corresponds to the wave going from the mean sea level to its minimum; in B the wave starts to ascend, going from its minimum to its midpoint at the mean sea level. Hence the upper half of the graph (sections A and B) corresponds to the first half of the wave, when it is below the mean sea level, while sections C and D correspond to the second half of the wave, C corresponds to the wave going from the midpoint to its maximum and D from the maximum to the endpoint.

The wave starts on section A with maximal negative velocity and zero acceleration. As the wave goes down, the graph moves right and up, indicating increasing velocity and acceleration. At the wave's minimum the speed is zero, but the acceleration is still increasing. Some time later the acceleration begins to decrease and reaches the value zero as the speed is maximal at the midpoint of the wave, i.e. the zero upcrossing. The acceleration continues to decrease, as does the speed until the latter reaches zero at the maximum of the wave. The speed continues to decrease while the acceleration increases, and at the endpoint of the wave, there is zero acceleration and maximal negative speed.

Figure 8. Integrated phase diagram as a function of significant wave height.

The phase diagram is clearly left-right asymmetric, but there is also an up-down asymmetry that is not so obvious. To explore this and the relation between significant wave height and this diagram we calculated the integral for the positive half of the phase-plane plot (that is, the area between the x-axis and the diagram) and the same for the negative half of the plot. According to the previous interpretation in terms of energy, these integrals represent the total energy present during each half of the wave. Figure 8 shows the values of the integrals for each half of the graphs as a function of significant wave height. Both are quadratically increasing with $H_s$ but the energy in the upper half, which corresponds to the first half of the wave, is almost always smaller than the energy in the second half. A regression of the total energy (sum of both integrals) on the square root of significant wave height gives an excellent fit ($R^2 \approx 0.983$), as would be expected.

**SIMULATED GAUSSIAN STORM**

It is well-known that the Gaussian model for waves does not work well during storm conditions, when non-linear effects are present and play an important role. To explore the departure of the data from the theoretical Gaussian model we simulated a storm having the same spectral evolution as the data. For each 20 minute interval the spectral density was estimated and used to simulate a Gaussian process sampled at 5 Hz. This process was done using the WAFO software [1].

The previous statistical analysis was repeated for the simulated waves. The mean waves for the simulated Gaussian storm show more symmetry than real waves, and their variability, as measured by the standard deviation is also more symmetric, as figure 9 shows.

![Standard Deviation](image)

Figure 9. Standard deviation for simulated waves.

To quantify the differences between the mean wave profile determined from the actual data and the Gaussian simulation for each 20 minute time interval we calculated both the maximum ($L_\infty$) distance between the two curves and the integrated ($L_1$) distance between the curves for each interval. Figure 10 shows the results along with a local regression (loess) curve. Both graphs show that the distance increases with $H_s$, as would be expected.

![Distance between Mean Waves](image)

Figure 10. (left) Maximum distance between mean waves. (right) Integrated distance between mean waves.
There are also important differences in the first two derivatives. Figure 11 show the phase-plane plots for all periods for the simulated storm. The curves corresponding to simulated data show more regularity and symmetry than those of real waves. Measurements show that there is almost no difference between the total energy present in sections A and C. For section B simulated Gaussian waves have consistently a higher proportion of the total energy f the wave in this quarter, while for section D the opposite is true. These differences are stable with respect to changes in $H_s$.

Figure 10. Phase-plane diagram for simulated waves.

CONCLUSIONS

Using FDA techniques we have been able to analyze the shape of mean waves and their principal characteristics as well as their variability. The results clearly show the vertical (amplitude) and horizontal (time) asymmetry arising from the non-linear properties of the higher water waves. This type of analysis allows for a better understanding of the effect of significant wave height on the shape of mean waves. The consideration of the phase diagram for mean waves gives insight on the distribution of energy during the evolution of the waves and its dependence on significant wave height.

On the other hand these tools permit the comparison of simulated waves using the Gaussian model with real waves, and the analysis of their differences. The results show not only that the Gaussian model is not valid under the conditions considered, which is well known, but also how the shape of mean waves, their variability, and their derivatives differ and how the distribution of the energy as the wave evolves under this model is different from what is observed in real data. These comparisons may help in future works to select nonlinear models for storm waves that produce simulations with a better fit to what is actually observed under real conditions.

ACKNOWLEDGMENTS

The authors would like to thank Total E&P UK for the wave data from the Alwyn North platform. The software WAFO [1] developed by the Wafo group at Lund University of Technology, Sweden, available at [http://www.maths.lth.se/matstat/wafo](http://www.maths.lth.se/matstat/wafo), was used for the calculation of all Fourier spectra and associated spectral characteristics, as well as for the simulation of Gaussian processes. The FDA software used was developed by J.O. Ramsay, G. Hooker and S. Graves [6] and was used for the analysis of functional data. This work was partially supported by CONACYT, Mexico, Proyecto Análisis Estadístico de Olas Marinas. George Smith would like to gratefully acknowledge the support offered by Scottish and Southern Energy PLC. Part of these result appear in Cristina Gorrostietas's M.Sc. thesis at CIMAT.

REFERENCES


