A Multivariate Exploratory Data Analysis on Waves

A univariate and multivariate exploratory data analysis was performed on data collected by a wave measuring buoy off the coast of Mooloobaba, Australia, monitored by Queensland Government, Department of Environment & Science [1]. The buoy has been recording data at 30-minute interval since its deployment in February 2017; this investigation focusses on the data that were collected between January and December 2019.

The attributes analysed are defined as follows [2]:

- Wave height - the highest single wave recorded in the 30-minute window
- Period - the wave period that are producing the most energy in the 30-minute interval
- SST - surface sea temperature at the buoy, measured in Celsius

The wave measurements can be considered reasonably accurate from a trustworthy source as each buoy is calibrated yearly and the buoy has an internal memory in the case of interrupted connection with the receiver station in critical events such as cyclones [3]. However, the buoy is exposed to bodies in the ocean that can easily alter the natural motion of the instrument. Only 75 of the 17520 measurements contained the error value –99.90, these were consequently removed from further analysis. Otherwise, the data are highly interpretable and coherent - no NULL values are present.

The univariate analysis was focused on the maximum wave height. Figure 1 shows the wave height density estimates with simple measures of central tendency.

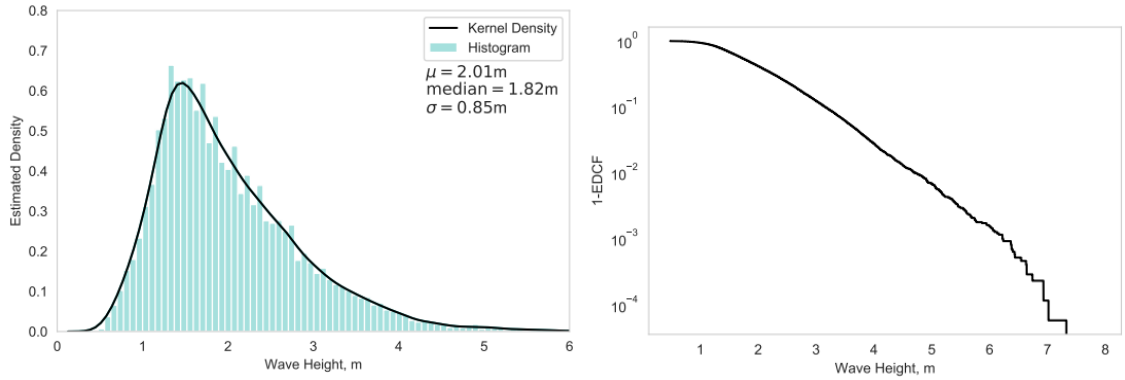

Figure 1: (Left) *Histogram and kernel density plot with a gaussian kernel of the wave height with 100 bins. (Right) The survival function, using the empirical cumulative distribution function (ECDF), of the wave height visualises the tallest waves.*

By locating the tallest histogram bar, the mode was determined to be 1.47m, which is less than the mean and median indicating that wave height is right-skewed. In other words, larger waves were recorded less often than smaller waves.

The skewness was calculated to be 1.23, indicating the variability of wave height arises larger than the mean, this can be seen in Figure 1 where the histogram bars vary between heights of 2-3m. Furthermore, the kurtosis was determined to be 2.26, close to that of a normal distribution but here less of the variance arise from the extreme values of the distribution.

Taking an interest in the wave height, period and sea temperature collectively, first the covariance matrix was determined shown in Figure 2.

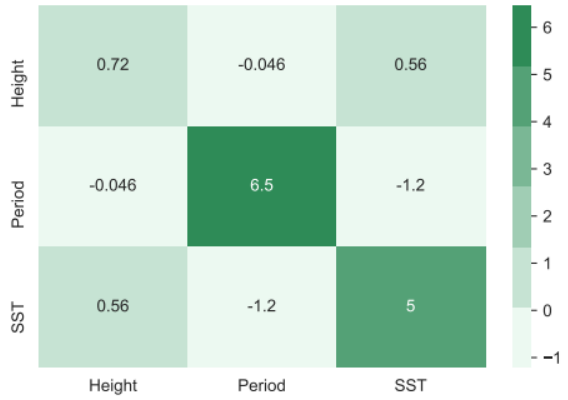

Figure 2: *Covariance matrix of the maximum wave height, time period recorded and the surface sea temperature at the time the wave passed.*

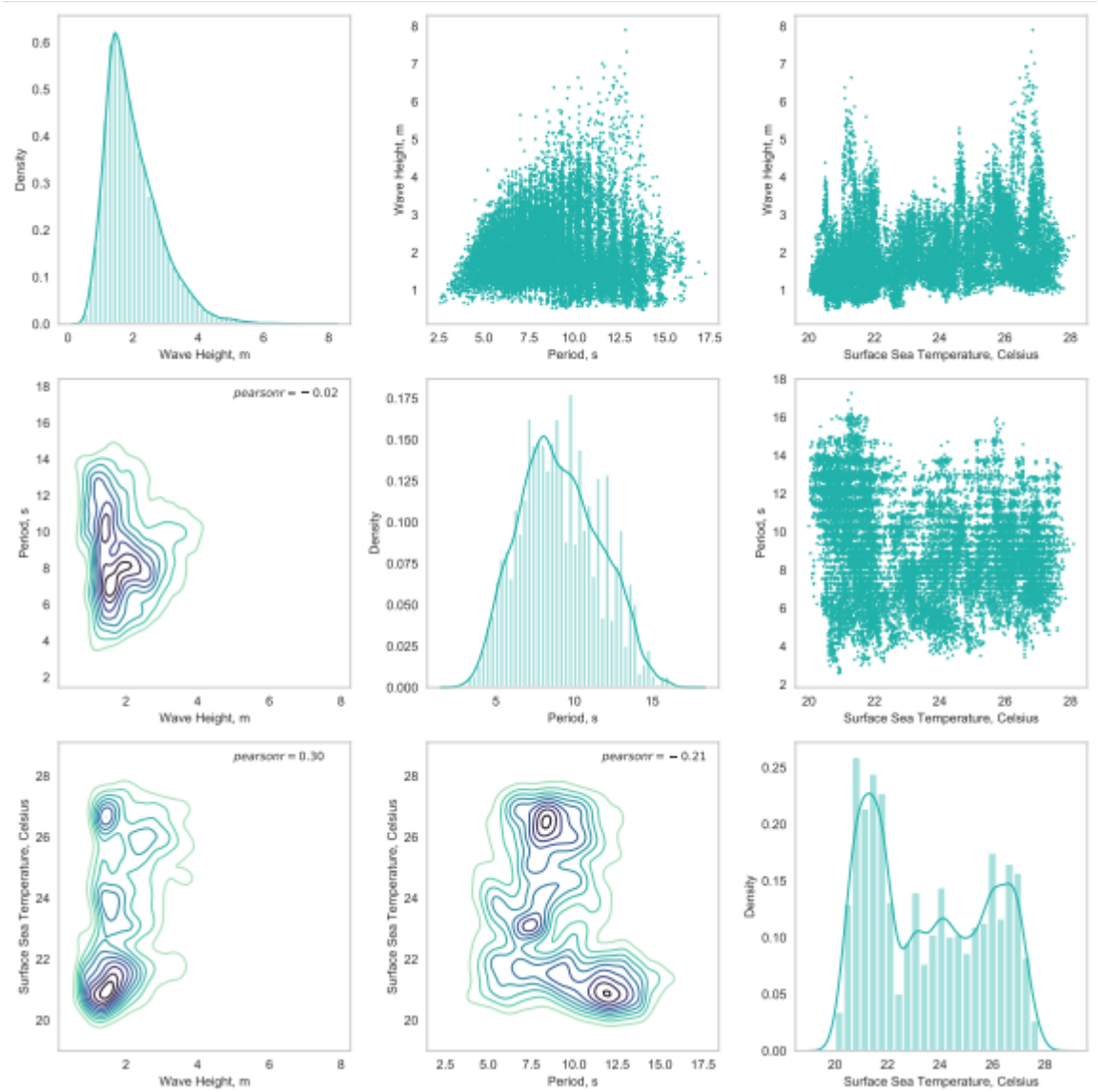

Figure 3: *Multivariate analysis of the maximum wave height, time period recorded and the surface sea temperature at the time the wave passed.*

The matrix plot showing the scatter and densities plots outline that there is very little correlation between each of the three variables, verified by the Pearson correlation coefficients close to 0. However, in the multivariate kernel density plots the modes are revealed – in the mid-left plot the waves with the mode height determined earlier can have time period across the range of periods measured but typically are observed at around 7s or 10s.

The bottom-middle plot shows the most common regions of wave periods against sea temperature, the 3 peaks may relate to the 3 common states of waves across the year. For example, the peak at 12s and 21ºC may be waves measured in winter where regular storms bring swell and hence raises the time period of the waves.

References

[1] Queensland Government, 2019, *Wave data –2019.* Available at: https://www.data.qld.gov.au/dataset/coastal-data-system-waves-mooloolaba/resource/d80243cd-5bca-41a0-b0f8-4d4fac168d94?truncate=30&inner_span=True (accessed at 13/11/2020)

[2] Queensland Government, 2018, *Mooloolaba WR Metadata 2018, Available at:* https://www.data.qld.gov.au/dataset/aafb52d6-7bb8-4601-966b-5153bd35d4f8/resource/d80243cd-5bca-41a0-b0f8-4d4fac168d94/download/mooloolaba_verifieddata.csv (accessed at 13/11/2020)

[3] Queensland Government, 2018, *About Wave Monitoring,* Available at: https://www.qld.gov.au/environment/coasts-waterways/beach/monitoring/waves (accessed at 13/11/2020)

```
# Load the libraries

# Basic numerics
import numpy as np
import scipy as sp
import scipy.stats as st

# Data handling
import pandas as pd

# Graphics
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid", {'axes.grid' : False})

# Statistical tools
from statsmodels.distributions.empirical_distribution import ECDF
import statsmodels.api as sm

# Create a directory for figures
import os
if not os.path.exists('Figures'):
    os.makedirs('Figures')
```

```
# Take a peak at the data
wave_df = pd.read_csv('./mooloolaba_verifieddata.csv')
print(wave_df.shape)
wave_df.head()

wave_df.describe()

# Removed the readings containing the error value -99.9
for column in wave_df.columns[1:]:
    wave_df = wave_df[wave_df[column] > -99.90]
wave_df.describe()
```

```
# Focus on wave height column
height = wave_df.Hmax.values

# First draw a histogram
ax = sns.distplot( height, bins=100, color = "lightseagreen", label="Histogram" )

# Add a kernel density estimate
sns.kdeplot( height, color="black", label="Kernel Density")

# Find mean, median and standard deviation
mu = np.mean(height)
med = np.median(height)
sd = np.sqrt(var)

textstr = '\n'.join((
    r'$\mu=%.2f$m' % (mu, ),
    r'$\mathrm{median}=%.2f$m' % (med, ),
    r'$\sigma=%.2f$m' % (sd, )))

# Add labels
plt.xlabel('Wave Height, m')
plt.ylabel('Estimated Density')

# Add tick marks abd set the limits of the axes
plt.xticks((0,1,2,3,4,5,6))
plt.xlim([0,6])
plt.ylim([0,0.8])
plt.tight_layout()
```

```
# draw a text box
props = dict(boxstyle='round', facecolor='white', alpha=0.5)

# place box in upper left in axes coords
ax.text(0.72, 0.83, textstr, transform=ax.transAxes, fontsize=12,
        verticalalignment='top', bbox=props)

# Save a PDF version, then display the result here too
plt.savefig('Figures/2019waves_distplot.pdf')
plt.show()
```

```
# Use a KDE to estimate the mode
mykde = sns.kdeplot(height, color='k')
xm, ym = mykde.get_lines()[0].get_data()
mo = xm[np.argmax(ym)]

#Compute variance, sd, skewness and kurtosis
var = np.var(height, ddof=1)
sd = np.sqrt(var)
skew = st.moment(height,3)/(sd**3)
kurt = (st.moment(height,4)/(sd**4))-3

print("The mode is", mo,
      "\nThe variance is", var,
      "\nThe standard deviation is", sd,
      "\nThe skewness is", skew,
      "\nThe kurtosis is", kurt)
```

```
#Import the ECDF with
from statsmodels.distributions.empirical_distribution import ECDF

# Plot the ECDF
ecdf= ECDF(height)

# Plot the survival function
plt.step(ecdf.x, 1-ecdf.y, color='k')
plt.yscale('log')
plt.xlabel("Wave Height, m")
plt.ylabel("1-EDCF")

plt.savefig('Figures/2019waves_survplot.pdf')
```

```
# Focus on wave height, period and sea temp
X = wave_df[["Hmax", "Tp", "SST"]]

# Find covariance matrix
S = np.cov(X, rowvar=False)

# Plot a heat map
sns.heatmap(S, xticklabels=["Height", "Period", "SST"], yticklabels=["Height", "Period", "SST"],
            annot=True, cmap=sns.light_palette("seagreen"))

plt.savefig('Figures/2019waves_covplot.pdf')
```

```
names=["Wave Height, m", "Period, s", "Surface Sea Temperature, Celsius"]
X = wave_df[["Hmax", "Tp", "SST"]]

# Form plot matrix
plt.figure(figsize=(12,12))
for i in range(0,3):
    for j in range(0,3):
        plt.subplot(3,3,1+i+(3*j))
        if i==j:
            # Draw histograms and KDEs on the diagonal using
            # whichever version of Seaborn command is appropriate
            seabornVersionStr = sns.__version__
            versionStrParts = seabornVersionStr.split('.')
            if( int(versionStrParts[1]) < 11 ):
                # Use the older, now-deprectaed form
                sns.distplot(X.iloc[:,i],color="lightseagreen")
            else:
                # Use the more recet form
                sns.kdeplot( X.iloc[:,i], color="black", label="Kernel Density")
                sns.histplot( X.iloc[:,i], stat="density", color = "lightseagreen" )

            # Add labels
            plt.xlabel(names[i])
            plt.ylabel('Density')
        else:
            if i<j:
                # Plot two-dimensional KDEs below the diagonal
                kd = sns.kdeplot(np.ravel(X.iloc[:,i]), np.ravel(X.iloc[:,j]),cmap="mako_r")

                # Calculate and write into text box the correlation coefficent
                (r, p) = st.pearsonr(X.iloc[:,i],X.iloc[:,j])

                textstr = r'$pearsonr=%.2f$' % (r, )

                props = dict(boxstyle='round', facecolor='white')

                # place a text box in upper left in axes coords
                kd.text(0.60, 0.97, textstr, fontsize=9,
                        verticalalignment='top', transform=kd.transAxes, bbox=props)
            else:
                # Put scatterplots above the diagonal
                sns.scatterplot(X.iloc[:,i],X.iloc[:,j],facecolor="lightseagreen",marker=".", linewidth=0, s=20)


            # Add labels
            plt.ylabel(names[j])
            plt.xlabel(names[i])

plt.tight_layout()
plt.savefig('Figures/2019waves3.pdf',format='pdf')
```