Data Cleansing

Conrad Thiele

2019-04-21

Data preparation plays a pivotal role and in my opinion is inter-twined with data understanding as a foundation to ensuring that, any statistical analysis or algorithm, wrappers, and other assorted machine learning output accurately represent the derived sample. The output that forms the functional graph is made of up of careful transforming, cleaning and examination of data.

In table 1.1 are two samples datasets downloaded from Queensland Government open access dataset and is comprised of wave sensory information from buoys off the Queensland coast. [1][2]. If you would like to download the data used in this tut-tworiel than you can do so via:

* [Wave Data Mooloolaba](./data/mooloolaba_2018-01-01t00_00-2018-10-31t23_30.csv)
* [Wave Data Caloundra](./data/caloundra-pob_2018-01-01t00_00-2018-10-31t23_30.csv)

##Importing and viewing the dataset.

First things first, import the dataset. In this case, I will be using RStudio, however there are many software packages for performing all the task showning in the tutorial.[RStudio-Vers]

mooloolaba.waves <- read.csv(file = "./data/mooloolaba\_2018-01-01t00\_00-2018-10-31t23\_30.csv", header = T)  
caloundra.waves <- read.csv(file = "./data/caloundra-pob\_2018-01-01t00\_00-2018-10-31t23\_30.csv", header = T)

With the data loaded, it is often good to take a quick look for missing values (NA), not a number (NaN), incorrect column headings, and dates/time formats are imported correctly. This can be down with the function head() and tail() which, show a subset portion of the first and last rows.

The field names and descriptors are as follows: \* **Hs** - Significant wave height, an average of the highest third of the waves in a record (26.6 minute recording period). \* **Hmax** - The maximum wave height in the record. \* **Tz** - The zero upcrossing wave period. \* **Tp** - The peak energy wave period. \* **Dir\_Tp TRUE** - Direction (related to true north) from which the peak period. Field name Peak.Direction in CSV file. \* **SST** - Approximation of sea surface temperature

head(mooloolaba.waves)

## Date.Time Hs Hmax Tp Tz Peak.Direction SST  
## 1 1/01/2018 0:00 -99.900 -99.90 -99.900 -99.900 -99.9 -99.9  
## 2 1/01/2018 0:30 0.513 0.81 10.315 4.748 -99.9 -99.9  
## 3 1/01/2018 1:00 0.566 0.93 10.778 5.003 92.0 26.4  
## 4 1/01/2018 1:30 0.557 0.85 9.984 4.990 91.0 26.4  
## 5 1/01/2018 2:00 0.569 0.96 9.277 5.214 81.0 26.4  
## 6 1/01/2018 2:30 0.571 0.88 9.901 5.121 80.0 26.4

tail(mooloolaba.waves)

## Date.Time Hs Hmax Tp Tz Peak.Direction SST  
## 14587 31/10/2018 21:00 0.802 1.24 12.155 5.121 123 23.60  
## 14588 31/10/2018 21:30 0.771 1.34 12.111 4.856 122 23.60  
## 14589 31/10/2018 22:00 0.787 1.25 12.130 5.160 129 23.55  
## 14590 31/10/2018 22:30 0.784 1.32 11.817 5.214 120 23.55  
## 14591 31/10/2018 23:00 0.777 1.20 12.269 5.382 129 23.55  
## 14592 31/10/2018 23:30 0.816 1.18 11.968 5.261 125 23.55

*Table 1.1 Head and tail of dataset. Dir\_Tp TRUE = Peak.Direction*

##Summary with Senor Errors.

In *Table 1.1* the rows with the value -99.9 do not appear to match up with any of the **fields** and there **descriptions** associated with the sample.

There are two important reasons to remove these values, firstly, they are some kind of sensory data issue (a form of sampling error) and secondly, this will skew the average value for each column, as seen in *Table 1.2*,and will result in an inaccurate analysis.

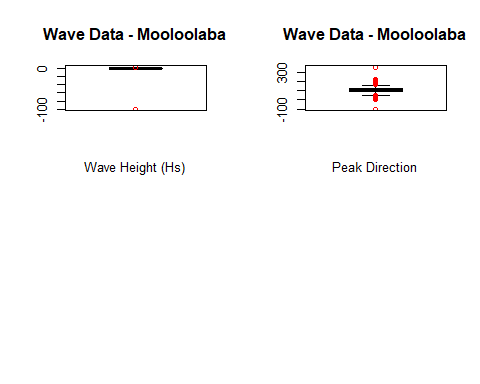
summary(mooloolaba.waves)

## Date.Time Hs Hmax Tp   
## 1/01/2018 0:00 : 1 Min. :-99.9000 Min. :-99.900 Min. :-99.900   
## 1/01/2018 0:30 : 1 1st Qu.: 0.8037 1st Qu.: 1.350 1st Qu.: 7.526   
## 1/01/2018 1:00 : 1 Median : 1.1120 Median : 1.870 Median : 9.132   
## 1/01/2018 1:30 : 1 Mean : 1.0747 Mean : 1.929 Mean : 9.072   
## 1/01/2018 10:00: 1 3rd Qu.: 1.6080 3rd Qu.: 2.700 3rd Qu.: 10.901   
## 1/01/2018 10:30: 1 Max. : 4.2570 Max. : 7.262 Max. : 21.121   
## (Other) :14586   
## Tz Peak.Direction SST   
## Min. :-99.900 Min. :-99.9 Min. :-99.90   
## 1st Qu.: 5.033 1st Qu.: 91.0 1st Qu.: 21.00   
## Median : 5.567 Median :105.0 Median : 22.95   
## Mean : 5.513 Mean :102.5 Mean : 23.09   
## 3rd Qu.: 6.256 3rd Qu.:119.0 3rd Qu.: 25.95   
## Max. : 10.146 Max. :358.0 Max. : 28.65   
##

*Table 1.2 Adapted Five number summary with mean values*

Another way to determine outliers is via boxplot as it does a true five number summary. As each of columns range of values (min - max) are completely different, it is best to do a boxplot for each variable (column). For this tutorial, the columns Wave Height(Hs) and Peak Direction were chosen. (see*Figure 1.1*)

par(mfrow=c(2,2))  
boxplot(mooloolaba.waves$Hs,  
 data=mooloolaba.waves,  
 outcol="red",  
 xlab="Wave Height (Hs)",  
 main="Wave Data - Mooloolaba")   
  
boxplot(mooloolaba.waves$Peak.Direction,  
 data=mooloolaba.waves,  
 outcol="red",  
 xlab="Peak Direction",  
 main="Wave Data - Mooloolaba")

 *Figure 1.1 Wave height and Peak Direction Boxplot*

The box portion of the plot, in *Figure 1.1*, which, represents the range between the first and the third quartile is difficult to interpret. This often happens when there is an extreme distance between minimim and maximum values as produced by the sensory error (-99.9)

As such, it is difficult to determine anything of real value from this plot other than, it definitely has outliers and secondly, they possibly need to be removed. However, before removing an outliers you should always ask yourself these questions:

* What is abnormal and therefore, what does the team/individual consider to be normal?
* What should we do with the outliers?
  + Is this in fact a new opportunity.
  + A beginning of a new trend.
  + Is this something this person knows and is it possible other will learn from this outlier behaviour and start a new trend.
  + Or is this a sign of a massive changes in circumstance.[3]

An outlier(s) are “data point(s) that are an abnormal distance from other values in a dataset.”, and as such, abnormal will need to be defined. [3] Outliers are often found using a five number summary or can be shown using a boxplot.

##Removing Sensor Errors

In this case, The data point, -99.9, is most likely a sensor error and will be removed. (see *Table 1.3.*)

mooloo.RM.outlier <- mooloolaba.waves[!(apply(mooloolaba.waves[,2:6], 1,  
 function(y) any(abs(y + 99.9) < 1e-9))),]

summary(mooloo.RM.outlier)

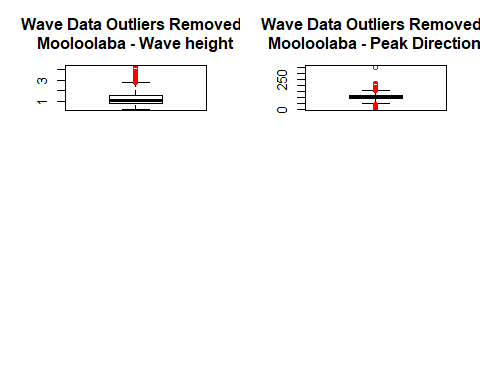
## Date.Time Hs Hmax Tp   
## 1/01/2018 1:00 : 1 Min. :0.294 Min. :0.510 Min. : 2.835   
## 1/01/2018 1:30 : 1 1st Qu.:0.805 1st Qu.:1.350 1st Qu.: 7.529   
## 1/01/2018 10:00: 1 Median :1.112 Median :1.870 Median : 9.147   
## 1/01/2018 10:30: 1 Mean :1.248 Mean :2.103 Mean : 9.260   
## 1/01/2018 11:00: 1 3rd Qu.:1.609 3rd Qu.:2.700 3rd Qu.:10.903   
## 1/01/2018 11:30: 1 Max. :4.257 Max. :7.262 Max. :21.121   
## (Other) :14541   
## Tz Peak.Direction SST   
## Min. : 3.378 Min. : 5 Min. :19.80   
## 1st Qu.: 5.036 1st Qu.: 91 1st Qu.:21.00   
## Median : 5.568 Median :105 Median :23.00   
## Mean : 5.693 Mean :103 Mean :23.43   
## 3rd Qu.: 6.257 3rd Qu.:119 3rd Qu.:26.00   
## Max. :10.146 Max. :358 Max. :28.65   
##

*Table 1.3 Adapted Five number summary with mean values outliers removed.*

As a result of this one crucial step in removing the sensory error produces a positive mean shift from *Table 1.2 - Table 1.3*. The data appears to have normalised, i.e. the mean and median values are almost identical, so it would be reasonal to say that it is close to a normal distribution and has gone from being non parametric to parametric.

##Box Plot - Sensor Error removed.

par(mfrow=c(2,2))  
boxplot(mooloo.RM.outlier$Hs,  
 data=mooloo.RM.outlier,  
 outcol="red",  
 main="Wave Data Outliers Removed:\nMooloolaba - Wave height")   
  
boxplot(mooloo.RM.outlier$Peak.Direction,  
 data=mooloo.RM.outlier,  
 outcol="red",  
 main="Wave Data Outliers Removed:\nMooloolaba - Peak Direction")

 *Figure 1.2 Wave height and Peak direction for Mooloolaba outliers removed*

In examining the first plot on the left, Wave height, the average wave height for Mooloolaba is 1 metre, still during storms/cyclones, one would expect the wave height could reach between the ranges of 3-4 metres.

In turn, looking at the Peak Direction plot, second graph on the right, the average direction is ~100 degree, yet, as there are 360 degrees in a circle, then the outliers ~350-0 degrees are completely feasible and should be kept.

In both instances, the abnormal values are crucial to explaining the whole picture and as such, it would be detrimental for the analyst to exclude them and would produce what is know as an **error due to bias**.

To sum up, missing values are not always labled as Not Availabe (NA) as is the case with the sensor data reading -99.9. Removing incorrect values, such as these, can make a dramatic improvement in understanding the basics of the data at hand and hopefully now the reader will have a better understand of the need to examine outlier value and question whether or not to remove them.

### References

[1] “Coastal Data System - Waves (Mooloolaba) - Datasets | Data | Queensland Government.” [Online]. Available: <https://data.qld.gov.au/dataset/coastal-data-system-waves-mooloolaba>. [Accessed: 12-Dec-2018]

[2] “Coastal Data System - Waves (Caloundra) - Datasets | Data | Queensland Government.” [Online]. Available: <https://data.qld.gov.au/dataset/coastal-data-system-waves-caloundra>. [Accessed: 12-Dec-2018]

[3] E. Davila, “Outliers.” [Online]. Available: <https://www.lynda.com/Business-Skills-tutorials/Outliers/427473/531429-4.html>. [Accessed: 29-Apr-2019]