

Homogeneity aspects in statistical analysis of coastal engineering data

P.H.A.J.M. van Gelder¹ and J.K. Vrijling²

Abstract

In this paper, the problem of (in)homogeneous data in coastal engineering applications is studied. Statistical analysis of coastal engineering data such as flood data, wind data, wave height data, etc., is essentially a problem of information scarcity. Records are usually too short to ensure reliable estimates of low-exceedance probability quantiles in many practical problems. Determination of quantiles is needed for the design, construction and operation of hydraulic structures, insurance studies and protection of populated areas. To perform a frequency analysis of data from coastal engineering practice, the data first has to be tested for homogeneity. Non-homogeneous data may lead to wrong quantiles. A homogeneity test must be able to separate data sets that do not come from the same distribution. In this paper statistical and physical-based homogeneity tests will be presented.

Introduction

Statistical analysis of coastal engineering data is a problem of information scarcity. Usually records are very short. Datasets with 100 years of water levels may seem much, but when one is interested in the 10^{-4} quantile, 100 years of data is very little. Apart from the data scarcity problem, there is also a data homogeneity (or rather data inhomogeneity) problem. A basic assumption is that data is coming from one and the same process. Or in mathematical terms: the data are realisations from one and the same probability distribution function. Statistical procedures are available to check the homogeneity of a dataset. A short overview of these procedures will be given in the paper. However their weakness will appear quickly, since these procedures are not very

¹ Delft University of Technology, Faculty of Civil Engineering, P.O. Box 5048, 2600 GA Delft, The Netherlands, E-mail: P.vanGelder@ct.tudelft.nl

² Delft University of Technology, Faculty of Civil Engineering, P.O. Box 5048, 2600 GA Delft, The Netherlands, E-mail: J.Vrijling@ct.tudelft.nl

powerful to reject inhomogeneous- or accept homogeneous data. A Monte Carlo experiment will illustrate this. A second method to judge the homogeneity of a dataset is more physically based. It will be shown that this way is more powerful than the statistical procedures. A case study will be presented in order to show the physical based judgement of a coastal engineering dataset. Other physical based techniques will be reviewed in the paper. The paper will end with some conclusions and a list of references in the area of homogeneity analysis.

Homogeneity

Statistical distribution functions play an important role in coastal engineering. In the design of coastal structures they are used to determine the so-called p-quantiles. A p-quantile of a random variable is the value of that variable which is exceeded with a probability p. In coastal engineering p-quantiles in the order of 10^{-1} for small-scale defence structures to 10^{-4} for important coastal defence structures are commonly applied. In the Netherlands the seadikes for instance are designed with p-values of 10^{-4} and river dikes with p-values of $1.25 \cdot 10^{-3}$. The reason for the difference is that a possible inundation from the rivers is not so disastrous as an inundation from the sea.

The importance to estimate the correct p-quantile is quite high. A too low estimate may lead to an unsafe structure, whereas a too high estimate may lead to a conservative overdesigned structures which costs unnecessarily too much. Lots of research has been carried out for finding the best p-quantile estimation method. Well known methods are for example Maximum Likelihood (ML), Method of Moments, Least Squares (LS), Weighted Least Squares (WLS), Method of L-Moments, Bayesian methods, and many more. Methods to select the optimal distribution function are also available. Various goodness-of-fit criteria such as χ^2 , Kolmogorov-Smirnov (KS), etc. can be used for that purpose. All these p-quantile estimation techniques and distribution selection methods have one important assumption in common, and that is that the data under consideration must be homogeneous. To verify the homogeneity assumption, statistical procedures have been developed.

Homogeneity analysis has been carried out a lot in the field of flood frequency analysis (FFA). FFA tries to combine data from other sites in order to improve the accuracy of the p-quantile estimate. However, combining data from different sites may only be done when the sites can be considered homogeneous. Therefore numerous papers have appeared dealing with this problem. We mention Dalrymple's test and the L-moment X10 test (Fill et al., 1995). Homogeneity tests based on L-moment ratios have received quite some attention lately (Rao et al., 1994, Zrinji et al., 1996). Sample L-moments are less biased than traditional moment estimators.

Not only homogeneity tests have been developed in the field of FFA. Also literature is available from the behavioural sciences such as sociology and psychology. Well known statistical tests are for example the Mann-Whitney test, and the Wald Wolfowitz test (Harnett, 1970). The Mann-Whitney test is a non-parametric homogeneity test which can test the nul-hypothesis that two independent datasets are

coming from the same distribution. The performance of this test can be studied with help of Monte Carlo simulations. A dataset 1 with a given size can be simulated from a given distribution function. Also a dataset 2 can be generated and the two datasets can be compared with each other from a homogeneity point of view. Under the null-hypothesis that the datasets come from the same distribution, the test statistic should be normally distributed with mean 0 and standard deviation 1. If the Mann-Whitney test is analyzed in this way, it will appear that the test is not so powerful for small sample sizes. Only for large sample sizes in the order of 100 or more, the test may reject or accept the null-hypothesis with high reliability. Also the recently developed L-Moment techniques have difficulties to judge the homogeneity of a dataset. A Monte Carlo experiment was designed for that purpose. A sample of size 40 is generated from a certain extreme value distribution (Gumbel¹⁰). From this sample the ordinary moments and the L-moments are calculated. Another sample of size 40 is generated from a quite different extreme value distribution (Normal¹⁰) with the same mean and standard deviation however. Its ordinary and L-moments are calculated. The experiment is repeated 100 times and the moments and L-moments values of each sample are depicted in figure 1. Note that as well as the ordinary moments graphic as the L-moments graphic show a non-distinguishable behaviour between the Gumbel¹⁰ and Normal¹⁰ distribution. In other words: it is impossible to separate the two distribution functions with the traditional and newly developed L-moment techniques.

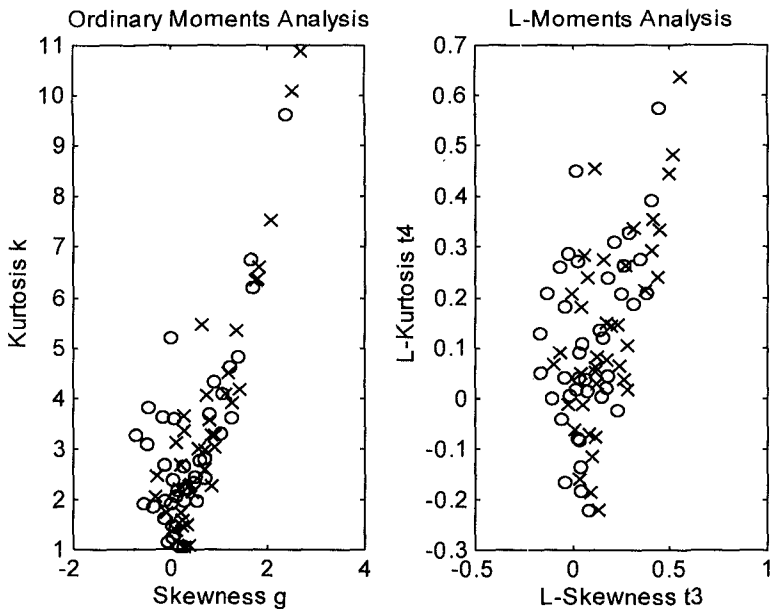


Figure 1. Homogeneity analysis with L-moments techniques.
Data from Gumbel¹⁰ are indicated with x; data from Normal¹⁰ with o.

The before mentioned homogeneity tests do not include any physical arguments to judge the data. In the following part we will propose procedures in order to examine the homogeneity of a data set on basis of physical arguments.

Physical based homogeneity analysis

Rather than a statistical analysis of the data, the data is examined on basis of its physical origin.

Physical arguments to judge the homogeneity of a data set are for instance:

1. Type of spectrum (swell, wind, single or double peaked wave height spectrum);
2. Season, calendar period of the data set (sea level data in the winter or summer);
3. Physical characteristics of the phenomena (breaking, non-breaking waves).

Combining statistical tests with physical arguments leads to more homogeneous data than only applying the statistical tests. This will be shown in a case of wave height and peak period data measured in the Bay of Bengal near the city of Madras.

Case study

In this case study registrations of the wave rider campaigns in the Bay of Bengal near the city of Madras during the south-west monsoon from mid-April to mid-August 1993 were used for analysis. A set of peaks over threshold values is available with significant wave heights. The set consists of 144 values and the threshold is given by 39cm. A statistical data analysis gives the following figure 2.

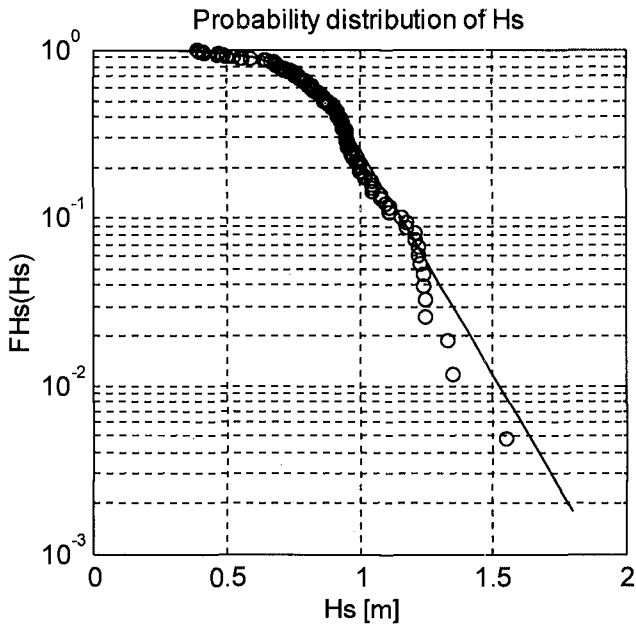


Figure 2. Wave height data with optimal fit (Normal).

A statistical homogeneity check on the data set with 144 values as described in the previous section does not lead to a rejection of the data set. Also a visual inspection of figure 2 does not suspect inhomogeneous data.

An investigation on the origins of wave generation in the Bay of Bengal however leads to:

1. Locally generated waves; these waves are generated in the Bay of Bengal either by the north-east or south-west monsoon or by hurricanes.
2. Swell; these waves are generated in the Roaring Forties, south of Cape Town. They reach the southern point of India from the south-west in approximately three days.

From the given data set with 144 significant wave heights, also the corresponding peak periods were available. A visual inspection of the data set with the peak periods immediately leads to the conclusion of inhomogeneity (figure 3). The waves can be separated into two classes. The first class contains waves with periods of about 5 sec. The second class contains waves of about 12 sec.

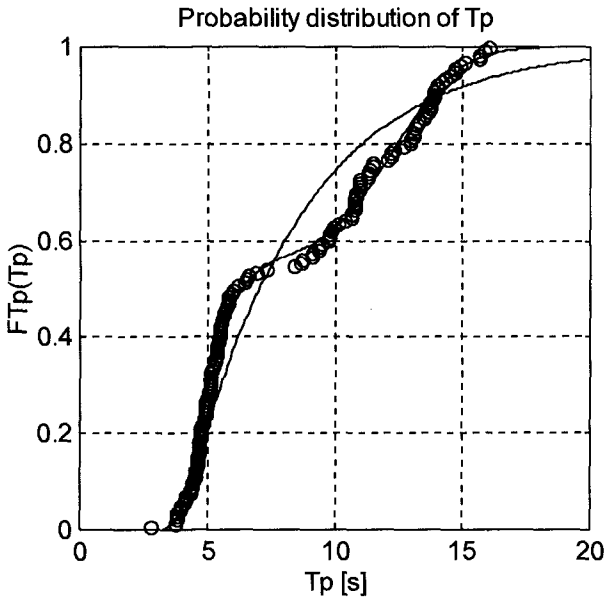


Figure 3. Peak periods data with Exponential inhomogeneous fit and Combined homogeneous fit.

Trying to model the probability distribution of the peak periods by one single distribution function leads to unacceptable deviations, as can be seen from figure 3. Therefore the inhomogeneous data set had to be split up into two sub sets. Each sub set is modeled by its own distribution and the probability model of the total data set is obtained by combining both sub models (figure 3).

Returning to the original goal: a probability distribution for the significant wave height, the following result is obtained (figure 4). The significant wave height with exceedance probability of 10^{-3} is 1.55m instead of 1.90m; a difference of about 30%. This might have led to an oversized coastal structure. However, the multivariate analysis of the data (including wave period in the analysis together with wave heights) had shown the inhomogeneous behaviour of the second tuple.

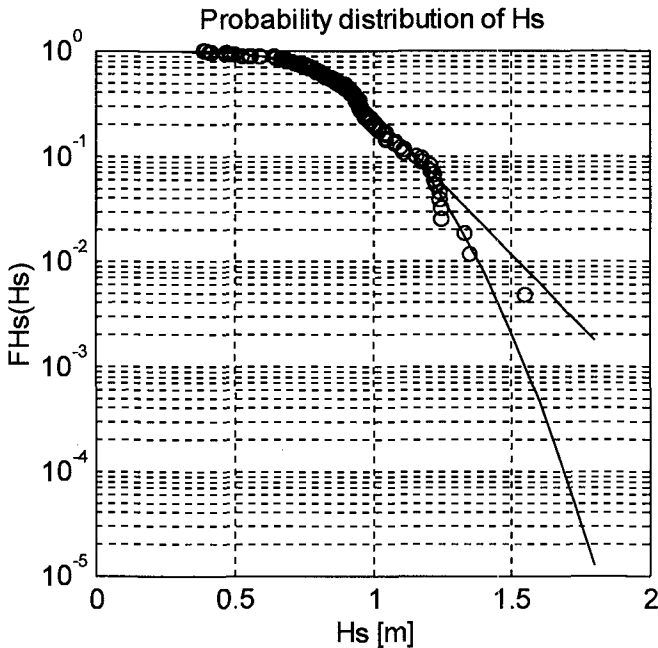


Figure 4. Comparison of in-/homogeneous fit.

North-European climate

Coastal engineering data from Northern European seas differ from the data presented above. Monsoons and hurricanes are not present in Northern Europe. However, strong winds can occur during the winter months November, December and January. These winds may yield extreme wave heights and water levels along the coasts. The analysis of such datasets also requires homogeneity studies. One of the first extensive homogeneity studies of the water levels along the Dutch coast was performed in 1960 by the Delta Committee. They homogenize the dataset of water levels by looking at the trajectory of the depression that caused the high water level. Extreme water levels can only occur when the depression follows a trajectory through a certain area. From the meteorological archive the area could be confined to:

- At 10° W between 51° N and 62° N;
- at 0° W between 52° N and 61° N;
- at 7° E between 52° N and 61° N.

The cause of an extreme water level is found in the trajectory of the depression combined with the behaviour of the body of water in the North Sea basin. Therefore

the statistics of the water level data is confined to those water levels that were caused by the dangerous trajectories. Such a physically based homogeneity selection method can also be successfully applied at other locations.

Multivariate analysis

What was seen in the case study was that the homogeneity of a univariate dataset can be judged by extending the univariate set to a multivariate dataset and analyzing the other tuples of the dataset. This principle can be applied to all kind of studies and appears to be very powerful. Consider for example the total yearly precipitation in the Netherlands (figure 5).

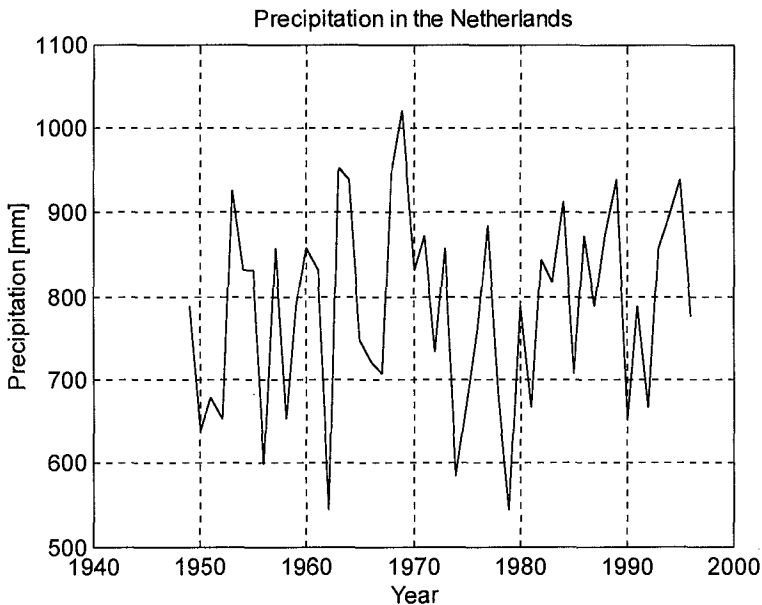


Figure 5. Total yearly precipitation with high variation coefficient.

Although the total yearly precipitation might be expected to be a stable quantity, a variation coefficient of 15% is observed. The univariate dataset of precipitation at each year can be extended to a multivariate dataset with (precipitation, dominant wind direction). Due to the geographic location of the Netherlands, winds from the west are very humid and bring a lot of rain; winds from the east however are very dry. Therefore two different processes can be distinguished and the precipitation dataset should be splitted into at least two subsets.

Conclusions

Homogeneity aspects in the statistical analysis of coastal engineering data have been discussed in this paper. Apart from statistical tests to judge the homogeneity of a data set, also physical arguments have to be included in the judgement. A case study of significant wave height data has been presented to illustrate the differences between the tails of distributions when the homogeneity aspects are left out of consideration.

References

Delta Committee, (1960) Considerations on storm surges and tidal waves (in Dutch), *Rapport Deltacommissie*, Deel 3, Staatsdrukkerij, Den Haag.

Fill, H.D., Stedinger, J.R., (1995) Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple's test, *Journal of Hydrology*. v 166 p 81-105.

Harnett, Donald L., (1970) *Introduction to Statistical Methods*, Addison-Wesley, London.

Mutua, Francis M., (1994) The use of the Akaike Information Criterion in the identification of an optimum flood frequency model, *Hydrological Sciences Journal*, 39, 3, June 1994.

Rao, A Ramachandra. Hamed, Khaled H., (1994) Frequency analysis of Upper Cauvery flood data by L-moments, *Water Resources Management*. v 8 n 3 1994. p 183-201.

Zrinji, Zolt. Burn, Donald H., (1996) Regional flood frequency with hierarchical region of influence, *Journal of Water Resources Planning & Management-ASCE*. v 122 n 4 Jul-Aug 1996. p 245-252.