Coastal Engineering 77 (2013) 91–98

Contents lists available at SciVerse ScienceDirect

Coastal Engineering

journal homepage: www.elsevier.com/locate/coastaleng

Extreme value prediction via a quantile function model

Yuzhi Cai *, Dominic E. Reeve

Swansea University, United Kingdom

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 11 February 2013
Accepted 15 February 2013
Available online 21 March 2013

Keywords:
Extreme values
Quantile function models
Bayesian approach
Semi-parametric model
Parametric model
Sea-level

ABSTRACT

Methods for estimating extreme loads are used in design as well as risk assessment. Regression using maximum likelihood or least squares estimation is widely used in a univariate analysis but these methods favour solutions that fit observations in an average sense. Here we describe a new technique for estimating extremes using a quantile function model. A quantile of a distribution is most commonly termed a 'return level' in flood risk analysis. The quantile function of a random variable is the inverse function of its distribution function. Quantile function models are different from the conventional regression models, because a quantile function model estimates the quantiles of a variable conditional on some other variables, while a regression model studies the conditional mean of a variable. So quantile function models allow us to study the whole conditional distribution of a variable via its quantile function, whereas conventional regression models represent the average behaviour of a variable. Little work can be found in the literature about prediction from a quantile function model. This paper proposes a prediction method for quantile function models. We also compare different types of statistical models using sea level observations from Venice. Our study shows that quantile function models can be used to estimate directly the relationships between sea condition variables, and also to predict critical quantiles of a sea condition variable conditional on others. Our results show that the proposed quantile function model and the developed prediction method have the potential to be very useful in practice.

1. Introduction

River and coastal flooding is an acknowledged natural hazard. There are many national examples that can be quoted and that are described in the academic and wider engineering literature. Sea level rise, changes in local climate and development in flood plains have altered the hazard and generally increased the consequences of flooding, and thereby the overall risk. An improved ability to predict, quantify and manage flood probabilities is essential to protecting the public, property and infrastructure, and to maintaining a sustainable economy. More accurate predictions of extreme conditions will improve the assessment, mitigation and management of flood risk.

Flood risk may be measured by the combination of the failure probability of a system such as a sea defence system and a measure of the consequences of the resulting floods. This is commonly represented as the probability of occurrence of an extreme event (or 'hazard') multiplied by the cost of the ensuing damage and hence 'risk' is measured as a rate of expenditure. There are several commonly used approaches to studying the probability of the functional failure of flood defences. That is, the situation in which the natural conditions exceed the severity for which a structure was designed to withstand; leading to the unwanted transmission of water across the line of the defence whilst not necessarily resulting in damage to the structure itself. One of these approaches is based on the concept of a structure variable, which is a known function of flood risk variables, such as sea condition variables. In particular, the structure variable method reduces a multivariate problem to a univariate one by using a formula relating to the structural performance, (for example, the wave overtopping rate), to combine potentially correlated variables such as water level, wave height and wave direction into a single variable which may then be analysed using univariate techniques. The occurrence of failure is then defined in terms of the structure variable exceeding a specified threshold rather than the probability of occurrence of particular combinations of the primary variables. See for example, the studies of Coles and Tawn (1994) and Reeve (1998). Another approach is known as the joint probability method and provides a procedure for estimating the probability that a structure variable exceeds a critical level based on the joint analysis of all flood risk variables. Many papers and reports on the application of these methods can be found in the literature. See, for example, the studies of Owen et al. (1997), Hawkes et al. (2002, 2004) and Meadowcroft et al. (2004).

However, in using the standard methods of fitting distributions to observations, the above two approaches are mainly influenced by observations in the bulk of the distribution which may have little influence on the form of the tail, the most important part of the distribution for estimating failure, and hence may provide poor fits. Methods based on scaled error norms, (e.g., Li et al., 2008; Reeve, 1996), address this to some degree but can be difficult to 'tune' precisely. A further approach

* Corresponding author at: College of Business, Economics and Law, Swansea University, Swansea, SA2 8PP, United Kingdom. Tel.: +44 1792 606865.
E-mail address: y.cai@swansea.ac.uk (Y. Cai).

0378-3839/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.coastaleng.2013.02.003

is to estimate probability distributions by using only extreme values of the flood risk variables, see for example, the studies of Coles (2001), Coles et al. (1999), Tawn (1990, 1992) and Thompson et al. (2009). Such an approach can seem wasteful of data, particularly to those making the observations. Partly in response to this concern, modifications to traditional approaches that use more observations have been developed, (e.g., Smith, 1986).

A common feature of the above approaches is that once a probability model is established for a sea condition variable, the value of the variable corresponding to a failure probability, i.e., the quantile or the return level, needs to be calculated by inverting the estimated conditional distribution function. If the estimated conditional distribution is not a commonly used distribution, then it can be difficult to invert the distribution exactly, and hence adding another layer of inaccuracy to prediction.

Recently, there has been a surge in interest in quantile methods. These methods allow us to directly estimate the conditional quantile, i.e., the value that a sea condition variable takes with a required probability. Hence these models focus on the quantiles at a level of interest (i.e., return levels) instead of the average value of a sea condition variable, because of which, the quantile approaches to statistical modelling have been used in many areas including economics, finance and medical research. See, for example, the studies of Koenker (2005) and Gilchrist (2000).

Generally speaking, there are two types of quantile approaches: one is semi-parametric (see, for example, Koenker, 2005) and another is parametric (see for example, Gilchrist (2000)). Although some work can be found in the literature on prediction with semi-parametric quantile regression models, see, for example, those of Taylor (2005), Cai (2010c) and Cai et al. (2012) and references therein, to the authors' knowledge, little work on prediction with parametric quantile function models can be found in the literature. Therefore, the main contributions of this paper are: (a) to develop a method for extreme value prediction via a parametric quantile function model, and (b) to show, via a real data set, the differences between various statistical models commonly used in extreme value prediction. The advantages of the quantile approach are that it makes full use of all available data including both extreme and non-extreme observations, it allows us to use many non-standard distributions that may provide an improved fit to observations leading to better predictions, and it provides a natural way of dealing with multivariate problems so that predictions on a flood risk variable conditional on a set of other variables can be made. At this point it is perhaps worth emphasising that here we use the words 'model', 'estimation' to mean 'a statistical distribution used to model observations' and 'determining the best fit values of the distribution parameters' respectively.

In Section 2 we first give an outline of the two types of quantile approaches. We then introduce a special type of quantile function model and present the new forecasting method in Section 3. The application to the Venice sea-level data and comparisons with other commonly used models are presented in Section 4. Finally, Section 5 provides some further comments and conclusions.

2. An outline of the quantile approaches

2.1. Semi-parametric quantile regression model

Let Y be a continuous flood risk random variable, such as wave overtopping rate or run-up level, and $\mathbf{x} = (x_1,\ldots,x_p)$ be a vector of p covariates. Given a set of observations $\{y_i,x_{1i},\ldots,x_{pi}\}$ $(i = 1,\ldots,n)$, a semi-parametric quantile regression model (Koenker (2005)) for the τth conditional quantile of Y, denoted by $q^{\tau}_{Y|X}$ is given by

$$q^{\tau}_{Y|X} = h\left(\eta^{\tau}, x_1, \ldots, x_p\right), \tag{1}$$

where h is a known function of the covariate $\mathbf{x}$ and model parameter η^{τ} which depends on τ $(0 \le \tau \le 1)$. So, for example, if $\tau = 0.95$, model (1) gives a 95% conditional quantile of Y. Therefore, for a sequence of τ values, model (1) defines a sequence of conditional quantiles of Y. These quantiles provide an estimate of the conditional quantile function of Y. Note that model (1) does not contain an error term, hence, the whole model is semi-parametric. A simple example of (1) is the linear quantile regression model given by $q^{\tau}_{Y|X} = a^{\tau}_0 + a_{\tau}{}^{\tau}x_1 + \cdots + a^{\tau}_p x_p$ with $\eta^{\tau} = (a^{\tau}_0,\ldots,a^{\tau}_p)$.

The model parameters can be estimated by using various methods including the method based on solving the minimization problem $\min_{\eta^{\tau}} \sum_{i=1}^{n} \rho_{\tau}(u_i)$, where $\rho_{\tau}(u_i) = u_i(\tau - I_{[u_i<0]})$, in which $I_{[\cdot]}$ is an indicator function and $u_i = y_i - h(\eta^{\tau},x_{1i},\ldots,x_{pi})$, see for example, those of Koenker and D'Orey (1987, 1994) and Koenker (2005). The model parameters can also be estimated by using a Bayesian method. For example, Yu and Moyeed (2001) proposed a Bayesian approach to quantile regression with independent data; Thompson et al. (2010) proposed a Bayesian non-parametric quantile regression method with applications to sea condition variables in coastal engineering; Cai and Stander (2008) studied a Bayesian approach to a nonlinear quantile time series model. Lancaster and Jun (2010) investigated the application of Bayesian exponentially tilted empirical likelihood to inference about quantile regressions. This semi-parametric approach has been used widely. However, it can suffer from some problems. For example, the estimated conditional quantile curves may cross over, leading to a failure of the method. Different methods have been proposed to solve the crossing-over problem. See, for example, the approaches of Bondell et al. (2010) and references therein.

2.2. Parametric quantile function model

Compared with the above semi-parametric approach, much less work can be found in the literature about parametric approaches. Gilchrist (2000) gave an excellent introduction to this parametric approach. A general parametric quantile function model is given by

$$Q_Y(\tau|\xi,\mathbf{x}) = h_1\left(\eta_1, x_1, \ldots, x_p\right) + h_2\left(\eta_2, x_1, \ldots, x_p\right)Q(\tau,\gamma), \tag{2}$$

where $\xi = (\eta_1,\eta_2,\gamma)$ is the model parameter vector; h_i $(i = 1, 2)$ are known functions of $\mathbf{x}$ and η_i, $h_2(\eta_2,x_1,\ldots,x_p) > 0$ and $Q(\tau,\gamma)$ are the quantile functions of the error term with explicit mathematical expression. A special case of model (2) is the linear quantile function model given by $Q_Y(\tau|\xi,\mathbf{x}) = a_0 + a_1x_1 + \cdots + a_px_p + Q(\tau,\gamma)$ with $h_1(\eta_1,x_1,\ldots,x_p) = a_0 + a_1x_1 + \cdots + a_px_p$, $h_2(\eta_2,x_1,\ldots,x_p) = 1$ and $\eta_1 = (a_0,\ldots,a_p)$.

It is seen that in the parametric approach, the number of parameters has been significantly decreased as the model parameters do not depend on τ, and the monotonicity of the conditional quantile function of Y can be guaranteed due to the fact that $Q_Y(\tau|\xi,\mathbf{x})$ is a well defined conditional quantile function. Gilchrist (2000) discussed various methods for estimating ξ. Cai (2009, 2010a, 2010b) proposed Bayesian approaches to estimating parameters of different types of model (2), including polynomial, multivariate and time series quantile function models.

In this paper we focus on a special type of quantile function models with details given in the next section.

3. The prediction method

3.1. Polynomial quantile function model

As mentioned above, different choices of h_1, h_2 and Q in model (2) lead to different quantile function models. If there is only one covariate x, then the dependence between Y and x may be modelled by a