Selection of Probability Distributions with a Case Study on Extreme Oder River Discharges

P.H.A.J.M. van Gelder
Delft University of Technology, Faculty of Civil Engineering, P.O. Box 5048, 2600 GA Delft, The Netherlands

J.M. van Noortwijk & M.T. Duits
HKV Consultants, P.O. Box 2120, 8203 AC Lelystad, The Netherlands

ABSTRACT: In design, one often needs to estimate extreme quantiles for use as design values. Fitting probability distributions to data is widely used for estimating these extreme quantiles. Due to little data, a difficult problem is often to select the type of distribution that can best be used. It is also seldom possible to derive the exact probability distribution theoretically. Common probability distributions in civil engineering are for instance normal, lognormal, gamma, Gumbel, Weibull, etc. Rather than choosing one particular probability distribution a priori, we propose to fit various probability distributions to the observations and to attach weights to these distributions according to how good the fits are. Existing methods are reviewed, extended and applied to the annual maximal river discharges of the Oder river in Central Europe. Finally, the methods will be compared with Monte Carlo simulation experiments.

1 INTRODUCTION

In many areas of civil engineering, the question arises which probability distribution should be used to model the load and resistance. Instead of choosing one particular probability distribution, it is also possible to consider various probability distributions and to attach weights to these distributions according to how good the fits are. The probability distribution for which the standard deviation of its predictions is large should be given less weight relative to those distributions that exhibit less scatter. Weight factors for probability distributions can be determined with different methods. In Kass and Raftery (1995) a Bayesian method is suggested to derive these weight factors. Their method was successfully applied in an economics case study by De Vos (1995) and a biometrics case study by Volinsky et al. (1996). In Tang (1980), a linear regression method was suggested to derive the weight factors. His method was successfully applied in a sea level case study by Van Gelder et al. (1996) and by Perricchi et al. (1983) in hydrology. In this paper, the two proposed methods will be reviewed briefly. Since the method of Tang is defined for comparing two probability distributions only, it appeared to have some disadvantages. In this paper, however, his method is extended to n probability distributions ($n \geq 2$). The methods are applied to estimating extreme discharges of the Oder. Finally, they are tested in a Monte Carlo simulation.

2 REVIEW OF THE METHODS

In this section, the two methods for determining the weight factors of probability distributions are described.

2.1 *Bayes factors*

Consider a data set D and two possible probability models (or hypotheses) H_1 and H_2. In the traditional approach we would determine a test statistic T and compute its p-value according to model H_1. If the test statistic of the data results in a smaller value than the p-value, then we would reject H_1. This traditional way of model testing has a lot of disadvantages. It can only be applied when two models are nested, one within the other. Furthermore it can only offer evidence *against* hypothesis H_1 under small p-values; we cannot accept H_1 under large p-values. The p-value offers only an interpretation as a long-term probability in a long repetition of the same experiment. In the Bayesian approach these disadvantages don't exist.

In the Bayesian approach, we apply Bayes theorem to the data that each of the models is supposed to predict and compute the posterior probability that a certain model is correct. There is no limit to the number of models that may be simultaneously considered, nor does any model need to be nested within any of the others.

Given prior probabilities $p(H_1)$ and $p(H_2) = 1-p(H_1)$, the data produces posterior probabilities $p(H_1|D)$ and $p(H_2|D) = 1-p(H_1|D)$. The quantity commonly used to summarise these results is the Bayes factor:

$$B=[p(H_1|D)/p(H_2|D)]/[p(H_1)/p(H_2)], \quad (1)$$

which can be reduced by Bayes theorem to:

$$B=p(D|H_1)/p(D|H_2). \quad (2)$$

So, the Bayes factor is precisely the probability of H_1 in favour of H_2 given solely by the data and its prior beliefs.

By using non-informative improper priors, $p(D|H_1) = \int L(D|H_1,\lambda)p(H_1(\lambda))d\lambda$ will not exist. Several ideas have been suggested to repair this problem:

1. Subdivide the data into two sets $D=(D_1, D_2)$ and use D_1 to generate prior information for D_2. This is an idea of Berger and Pericchi (1996) who also give suggestions how to split up D.

2. Apply the Schwarz Criterion:

$$-2\log B \approx -2\log\left[\frac{p(D\mid H_1,\hat{l}_1)}{p(D\mid H_2,\hat{l}_2)}\right]-(r_2-r_1)\log n, \quad (3)$$

where $\hat{l}_i$ is the maximum likelihood estimator under H_i, r_i is the number of parameters in model H_i, and n is the number of observations. In this criterion, the term $(r_2- r_1)\log(n)$ acts as a penalty term which corrects for differences in size between the models. Although the Schwarz criterion is independent of the prior density, it may be viewed as a useful approximation to $-2\log B$.

In order to test hypotheses using Bayes factors, Kass and Raftery (1995) suggested the following guidelines:

Table 1. Guidelines testing hypotheses using Bayes factors.

$2\log B$	B	Evidence against H_1
0 to 2	1 to 3	Not worth more than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
>10	>150	Decisive

Suppose $(H_1,H_2,...,H_n)$ is our collection of candidate models, and γ is our quantity of interest. Given a set of prior model probabilities $\{p(H_1),p(H_2),...,p(H_n)\}$, the posterior distribution of γ is given by

$$p(\gamma\mid D)=\sum_{i=1}^{n} p(\gamma\mid H_i,D)\,p(H_i\mid D), \quad (4)$$

where $p(\gamma|H_i, D)$ is the posterior for γ under the i-th model, and $p(H_i|D)$ is the posterior probability of this model. Averaging over models can result in a better fit than using any model individually. Despite this advantage there may arise a problem in the fact that calculation procedures can be time consuming. This problem does not exist when the number of models is small (say, n=2-5). Markov-Chain Monte-Carlo methods are developed for efficient calculation (Carlin and Louis, 1996).

Any approach that selects a single model and then makes inference conditionally on that model ignores the uncertainty involved in the model selection, which can be a big part in the overall uncertainty. This difficulty can be avoided if one adopts a Bayesian approach and calculates the posterior probabilities of all the competing models, following directly from the Bayes factors. A composite inference can then be made that takes account of model uncertainty in a simple way.

2.2 *Tang's method*

Tang (1980) proposed a linear regression model whereby discrepancy between observed data and the predicted probability model, as well as uncertainty of extrapolation from observations can be incorporated into hydrologic risk assessment. The method is based on a Bayesian linear regression analysis of the observed data (after transformation) plotted on probability paper. Given the discrepancies between data and model predictions in terms of expectations and variances of a design value, Tang's method involves combining these expectations and variances over the different probability models.

Suppose $E(Y_1)$ and $\mathrm{Var}(Y_1)$ denote the mean and variance of the design value predicted using one model, whereas $E(Y_2)$ and $\mathrm{Var}(Y_2)$ denote the mean and variance of the design value predicted using another independent model. According to Tang, an overall estimate of the expectation based on the combined information of two independent models can be determined using Bayes theorem as

$$E(X_2)=\frac{\mathrm{Var}(Y_2)E(Y_1)+\mathrm{Var}(Y_1)E(Y_2)}{\mathrm{Var}(Y_2)+\mathrm{Var}(Y_1)} \quad (5)$$

and an overall estimate of the variance as

$$\mathrm{Var}(X_2)=\frac{\mathrm{Var}(Y_2)\mathrm{Var}(Y_1)}{\mathrm{Var}(Y_2)+\mathrm{Var}(Y_1)}. \quad (6)$$

Observe that the combined estimate $E(X_2)$ will approach $E(Y_1)$ if $\mathrm{Var}(Y_1)$ is extremely small relative to $\mathrm{Var}(Y_2)$. In other words, if the first model provides an excellent fit while the second model provides some scatter, then the information of the second model could be neglected.

Using Bayes theorem, Eqs. (5-6) can be determined as follows. Let us consider a random sample from a normal distribution with an unknown value of the mean and a specified value of the variance. Suppose that the prior distribution of the unknown mean is a normal distribution, then the posterior distribution is also a normal distribution with parameters similar to Eqs. (5-6). As a matter of fact, the first and second model can be interpreted as the prior and observed information, respectively. For example, the posterior mean is an average of the prior mean and the sample mean, weighted inversely by the respective variances. For details, see Ang and Tang (1975, Chapter 8).

Although Tang presented the formulae for comparing design-value estimates on the basis of two probability models only, they can be easily generalised to n probability models where $n \geq 2$. If Eqs. (5-6) hold for all possible pairs of n independent probability models, then both the expectation $E(X_n)$ and the variance $\mathrm{Var}(X_n)$ can be derived using mathematical induction. This new result is proved in the appendix.

3 CASE STUDY

As a result of extreme rain during July 1997, Poland was affected by a devastating flood, the worst experienced in the past 200 years. Areas in seven voivodships in the upper and middle Oder river basin and upper Vistula river basin were flooded over 25% of their territory causing a flood damage of approximately 3 billion US dollars (Figure 1, Der Spiegel, 1997).

River floods seem to happen more often lately. A frequency analysis can be performed to determine the occurrence frequencies or return periods of extreme river floods. Estimates of the return periods of river floods are necessary in a reliability-based design of flood protection structures. Uncertainties are important in a flood frequency analysis and reliability-based design. Statistical uncertainties due to limited amounts of flood data and model uncertainties due to limited descriptive capabilities of the physical flooding process are two major uncertainties which have to be dealt with.

Figure 1. Topographical map of the Oder basin

In this case study, the approach is followed to determine the weight factors of four commonly applied probability distributions in hydrology with the two methods reviewed. The results are given in Tables 2 and 3.

Table 2. Bayes factors

Y_i	n	Weight Factors
Rayleigh	1	7%
Exponential	1	17%
Gumbel	2	66%
Lognormal	2	10%

Table 3. Tang's method (quantile 10^{-3})

Y_i	n	$E(Y_i)$	$\mathrm{Var}(Y_i)$	Weight Factors
Rayleigh	2	2599	11098	21%
Exponential	2	3673	9326	25%
Gumbel	2	3194	5111	46%
Lognormal	2	3482	32175	7%

Notice that with both methods the Gumbel distribution receives the highest weight factor.

In Figure 2, the four frequency curves (according to Tang's method) are shown together with the 51 annual maxima of the Oder river at the city of Eissenhηttenstadt (obtained from the Bundesanstalt für Gewässerkunde). From a visual inspection of the models, it is very difficult to conclude which model might be considered the best. The weight factors can therefore be very useful.

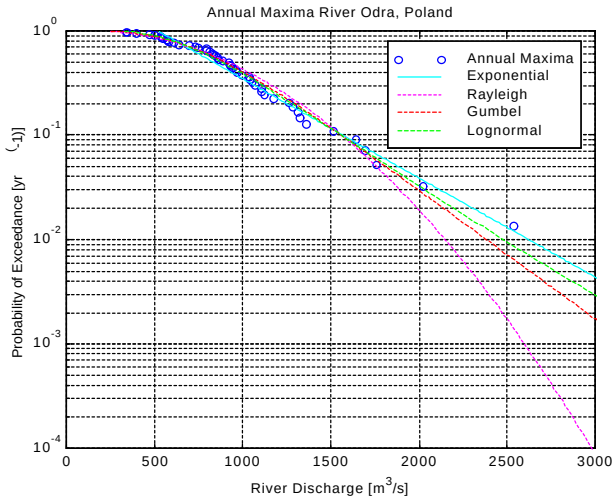

Figure 2. Frequency curves

The discharge of 1997 (2530 m^3/s) has a return period of 167 years according to the Bayes factors and 138 years according to Tang's method.

For calculation of the Bayes factors, the choice of the domain of numerical integration has been investigated. It appears not to be very sensitive. In Table 2 the results are shown for the situation in which the integration domain stretches from one third of the mean value of the parameter to three times the mean value of the parameter. However, if the integration domain is enlarged with a factor 5 around the mean value, then the following Bayes factors are obtained: 12%, 24%, 54% and 9%. And with a factor 10: 18%, 36%, 39% and 6%.

Tang's method is concerned with the weight factors with respect to a certain quantile. If instead of the 10^{-3} quantile the 10^{-2} quantile is studied, then the following differences can be noted:

Table 4. Tang's method (quantile 10^{-2})

Y_i	n	$E(Y_i)$	Var(Y_i)	Weight Factors
Rayleigh	2	2118	9904	18%
Exponential	2	2621	6752	26%
Gumbel	2	2383	4058	44%
Lognormal	2	2478	15131	12%

Since Tang's method is based on a linear regression analysis, it can only be applied to probability distributions with two parameters. For the Rayleigh, exponential and Gumbel distribution, both the location and scale parameter are fitted; for the lognormal distribution the mu and the sigma. When one- or three-parameter distributions are studied, Tang's method must be adapted.

Notice that the number of parameters of the probability distribution has a large influence on the variance of the *p*-quantile and therefore also on the weight factor. Some kind of correction factor could be introduced to take account of this influence. Such a penalty term was also incorporated in Kass and Raftery's Bayes factors. However, it will not be a subject of attention in this paper.

In the above case study, differences have been found between both methods. In order to test the performance of both methods, Monte Carlo simulations have been performed. The results are presented in the next section.

4 PERFORMANCE OF THE METHODS

Monte Carlo simulations have been performed to determine the posterior model probabilities as a function of the number of samples, and the distribution type from which the simulations were generated (Rayleigh, exponential, Gumbel and lognormal). In all cases, diffuse prior model probabilities have been used, $\theta_1=\theta_2=\theta_3=\theta_4=1/4$, and uniform parameter priors have been used. The following integrals were calculated numerically:

$$K_1=\int L_{exp}(\lambda|D)p(\lambda)d\lambda,$$

$$K_2=\iint L_{gumb}(\delta,\lambda|D)p(\delta,\lambda)d\delta d\lambda,$$

$$K_3=\int L_{ray}(\delta|D)p(\delta)d\delta,$$

$$K_4=\iint L_{lognorm}(\delta,\lambda|D)p(\delta,\lambda)d\delta d\lambda,$$

with L_{exp}, L_{gumb}, L_{ray}, $L_{lognorm}$ the exponential, Gumbel, Rayleigh and lognormal likelihoods, respectively. According to Kass and Raftery (1995), the weight factors are defined by:

$$\theta_i=K_i/(K_1+K_2+K_3+K_4), \qquad i=1,...,4.$$

The weight factors according to Tang's method have also been determined. In order to exclude statistical variability, the determination of the weight factors has been performed 1000 times and the mean values are presented in Tables 5a-d. In each cell the value on the l.h.s. gives the weight factor according to the Bayes factors; the value on the r.h.s. according to Tang's method.

Tables 5a-d. Simulation results (l.h.s. Bayes; r.h.s. Tang)

Simulation from exponential distr. (c.o.v. 100%; fixed value)

	n=10		*n*=20		*n*=50	
Rayleigh	0	0.35	0	0.30	0	0.22
Exponential	0.99	0.23	0.99	0.32	0.99	0.44
Gumbel	0.01	0.24	0.01	0.26	0.01	0.25
Lognormal	0	0.17	0	0.13	0	0.07

Simulation from Rayleigh distribution (c.o.v. 48%; fixed value)

	n=10		*n*=20		*n*=50	
Rayleigh	0.43	0.50	0.72	0.53	0.88	0.56
Exponential	0.48	0.18	0.23	0.14	0	0.10
Gumbel	0.09	0.29	0.05	0.32	0.12	0.33
Lognormal	0	0.02	0	0.01	0	0.01

Simulation from Gumbel distribution (c.o.v. 20%)

	n=10		*n*=20		*n*=50	
Rayleigh	0	0.40	0	0.38	0	0.36
Exponential	0.95	0.17	0.83	0.17	0	0.15
Gumbel	0.04	0.25	0.15	0.29	0.93	0.34
Lognormal	0.01	0.17	0.02	0.15	0.07	0.15

Simulation from lognormal distr. (c.o.v.20%)

	n=10		*n*=20		*n*=50	
Rayleigh	0	0.44	0	0.44	0	0.09
Exponential	0.57	0.12	0.80	0.12	0	0.43
Gumbel	0.34	0.26	0.18	0.29	0.77	0.31
Lognormal	0.07	0.17	0.02	0.16	0.23	0.16

The weight factors according to both methods can distinguish competing models, even when the length of record is quite short. However, the Bayes factors perform better than Tang's method. As the sample size increases upto *n*=50, the true model comes out with a model probability close to 100% in the case of Bayes factors, whereas Tang's method has probabilities around 50%. For the lognormal distribution both methods perform quite bad. One has to have sample sizes of the order 100-200 before the lognormal distribution is "recognised".

Also notice that for Gumbel generated data, the incorrect exponential model had higher posterior probabilities for low sample sizes. This is a bit surprising, since the Gumbel distribution even has one parameter more than the exponential distribution. In Figure 3, this behaviour has been analysed in more detail with Bayes factors for small sample sizes. Note the high standard deviation of the weight factors.

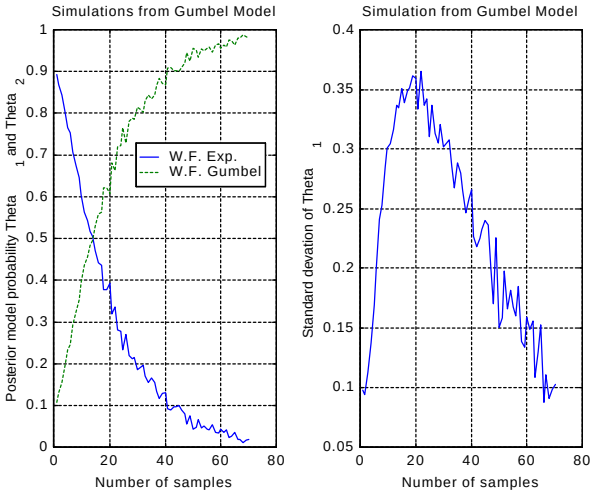

Figure 3. The Exponential and Gumbel models.

5 CONCLUSIONS

Probability exceedance curves are used frequently in many areas of safety and reliability. In this paper, the question how to select a certain probability distribution has been tackled by using Bayes factors and Tang's method. Both methods are reviewed, extended, investigated with Monte Carlo experiments and applied to a case study. Bayes factors appear to perform better than Tang's method and are therefore suggested to be used in distribution selection.

REFERENCES

Ang, A.H.S., and Tang, W.H., 1975, *Probability concepts in engineering planning and design, Volume* 1, Wiley.

Berger, J.O., and Perichi, L.R., 1996, The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association,* 91, 109-122.

Bundesanstalt für Gewässerkunde, 1997, *Discharge dataset of the Oder at Eisenhüttenstadt*, Koblenz, Germany.

Carlin, B.P., and Louis, T.A., 1996, *Bayesian and Empirical Bayes Methods of Data Analysis*, Chapman and Hall.

De Vos, A.F., 1996, "Fair and predictive Bayes factors for comparison of regression models", Technical Report, Vrije Universiteit Amsterdam.

Der Spiegel, 1997, *Die Schlacht an der Oder,* Germany, July 1997.

Kass, R.E., and Raftery, A.E., 1995, Bayes Factors, *Journal of the American Statistical Association*, June 1995, Vol. 90, No. 430, 773-795.

Pericchi, L.R., and Rodriguez-Iturbe, I., "On some problems in Bayesian model choice in hydrology", *The Statistician*, Vol.32, 1983.

Tang, W.H., 1980, Bayesian frequency analysis, *Journal of the Hydraulics Division,*Vol. 106, No. HY7, pp. 1203-1218.

van Gelder, P.H.A.J.M., Vrijling, J.K., and Slijkhuis, K.A.H., 1997, Coping with uncertainty in the economical optimization of a dike design, *Proceedings 27th IAHR Congress, Water for a Changing Global Community*, San Fransisco, p. 554-559.

Volinsky, C.T., Madigan, D., Raftery, A.E., and Kronmal, R.A., 1996, “Bayesian Model Averaging in Proportional Hazard Models: Assessing Stroke Risk”, Technical Report no.302, Dept. of Statistics, University of Washington.

APPENDIX

THEOREM:

Let

$$E(X_1) = E(Y_1),\ \mathrm{Var}(X_1) = \mathrm{Var}(Y_1),$$

as well as

$$E(X_n) = \frac{\mathrm{Var}(Y_n)E(X_{n-1}) + \mathrm{Var}(X_{n-1})E(Y_n)}{\mathrm{Var}(Y_n) + \mathrm{Var}(X_{n-1})},$$

$$\mathrm{Var}(X_n) = \frac{\mathrm{Var}(Y_n)\mathrm{Var}(X_{n-1})}{\mathrm{Var}(Y_n) + \mathrm{Var}(X_{n-1})},$$

$n = 2,3,\ldots$, then

$$E(X_n) = \frac{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)E(Y_i)}{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)},$$

$$\mathrm{Var}(X_n) = \frac{\prod_{j=1}^{n}\mathrm{Var}(Y_j)}{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)},$$

$n = 2,3,\ldots$

PROOF:

The proof follows by mathematical induction. As the basis of induction,

$$E(X_2) = \frac{\mathrm{Var}(Y_2)E(X_1) + \mathrm{Var}(X_1)E(Y_2)}{\mathrm{Var}(Y_2) + \mathrm{Var}(X_1)} = \frac{\mathrm{Var}(Y_2)E(Y_1) + \mathrm{Var}(Y_1)E(Y_2)}{\mathrm{Var}(Y_2) + \mathrm{Var}(Y_1)},$$

and

$$\mathrm{Var}(X_2) = \frac{\mathrm{Var}(Y_2)\mathrm{Var}(X_1)}{\mathrm{Var}(Y_2) + \mathrm{Var}(X_1)} = \frac{\mathrm{Var}(Y_1)\mathrm{Var}(Y_2)}{\mathrm{Var}(Y_2) + \mathrm{Var}(Y_1)}.$$

Using the induction hypothesis, it follows that

$$E(X_n) = \frac{\mathrm{Var}(Y_n)E(X_{n-1}) + \mathrm{Var}(X_{n-1})E(Y_n)}{\mathrm{Var}(Y_n) + \mathrm{Var}(X_{n-1})}$$

$$= \frac{\mathrm{Var}(Y_n)\sum_{i=1}^{n-1}\left(\prod_{\substack{j=1\\ j\neq i}}^{n-1}\mathrm{Var}(Y_j)\right)E(Y_i) + E(Y_n)\prod_{j=1}^{n-1}\mathrm{Var}(Y_j)}{\mathrm{Var}(Y_n)\sum_{i=1}^{n-1}\left(\prod_{\substack{j=1\\ j\neq i}}^{n-1}\mathrm{Var}(Y_j)\right) + \prod_{j=1}^{n-1}\mathrm{Var}(Y_j)}$$

$$= \frac{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)E(Y_i)}{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)}$$

and

$$\mathrm{Var}(X_n) = \frac{\mathrm{Var}(Y_n)\mathrm{Var}(X_{n-1})}{\mathrm{Var}(Y_n) + \mathrm{Var}(X_{n-1})}$$

$$= \frac{\mathrm{Var}(Y_n)\prod_{j=1}^{n-1}\mathrm{Var}(Y_j)}{\mathrm{Var}(Y_n)\sum_{i=1}^{n-1}\left(\prod_{\substack{j=1\\ j\neq i}}^{n-1}\mathrm{Var}(Y_j)\right) + \prod_{j=1}^{n-1}\mathrm{Var}(Y_j)}$$

$$= \frac{\prod_{j=1}^{n}\mathrm{Var}(Y_j)}{\sum_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(Y_j)\right)}$$

for $n = 2,3,\ldots$