J. Hydrol. Hydromech., *57*, 2009, 3, 162–171
DOI: 10.2478/v10098-009-0015-z

ESTIMATING HIGH LEVELS EXCEEDANCE PROBABILITIES BY POINT PROCESS APPROACH WITH APPLICATIONS TO NORTHERN MORAVIA PRECIPITATION AND DISCHARGES SERIES

DANIELA JARUŠKOVÁ

Department of Mathematics, Faculty of Civil Engineering, Czech Technical University, Thákurova 7, CZ – 166 29 Praha 6, Czech Republic; mailto: jarus@mat.fsv.cvut.cz

The paper by *Jarušková* and *Hanek* (2006) advocated application of the peaks over threshold method (POT method) for estimating the probability that a precipitation or discharges series exceeds a chosen high level. If daily precipitation amounts or average discharges are obtained at several stations one might be interested in estimating the probability that in the same time all variables of interest, e.g. precipitation amounts measured at several stations, exceed some chosen high levels. The paper explains how the method based on the point process approach may be used to get good estimates of such probabilities. Moreover, it presents some useful parametric models that were successfully applied by the author to some precipitation and discharges series of northern Moravia.

KEY WORDS: Precipitation and Discharges Series, High Level Exceedance Probabilities, Modeling Tails of Multivariate Distribution, Peaks over Threshold Method, Modeling Dependence Structure, Inverse Arguments Tail Dependence Function.

Daniela Jarušková: ODHADOVÁNÍ PRAVDĚPODOBNOSTÍ PŘEKROČENÍ VYSOKÝCH ÚROVNÍ POMOCÍ METODY BODOVÉHO PROCESU S APLIKACEMI PRO SRÁŽKOVÉ A PRŮTOKOVÉ ŘADY NA SEVERNÍ MORAVĚ. J. Hydrol. Hydromech, *57*, 2009, 3; 7 lit., 4 obr., 4 tab.

Článek navazuje na práci *Jarušková*, *Hanek* (2006), kde autoři doporučovali používání metody špiček nad prahem k odhadu pravděpodobností, s jakou srážková nebo průtoková řada překročí danou vysokou úroveň. V případě, že se denní srážková či průtoková řada měří ve více stanicích, může nás zajímat, s jakou pravděpodobností současně (to znamená ve stejný den) všechny studované řady, to je například srážkové řady měřené v několika stanicích, překročí nějaké předem stanovené vysoké úrovně. Článek vysvětluje, jak lze k odhadu takových pravděpodobností použít metodu založenou na bodovém procesu. Zároveň uvádí některé parametrické modely, které byly úspěšně použity autorkou článku pro odhady pravděpodobností překročení pro srážkové a průtokové řady na severní Moravě.

KLÍČOVÁ SLOVA: srážkové a průtokové řady, pravděpodobnosti překročení vysoké úrovně, modelování chvostů vícerozměrného rozdělení, metoda špiček nad prahem, modelování struktury závislosti, funkce závislosti chvostů daná v inverzních argumentech.

Introduction

Estimating high annual return levels of precipitation and discharges series is one of the basic problems of statistical hydrology. The problem has its parallel in estimating the probability that some given level is exceeded. The first problem consists in finding an appropriate level u_α for a given α so that $P(X > u_\alpha) = \alpha$, the second one consists in estimating $P(X > x)$ for some real x. The variable X corresponds to a quantity of interest, e. g. a daily precipitation amount or a daily average discharge. If daily measurements during several years are available, we may try to create a reasonable probabilistic model for the distribution of the studied variable. If we are interested in the probability of exceedance of some large value x, the method of peak over threshold (POT method) described in *Jarušková* and *Hanek* (2006) may be applied. The basic idea of the POT method is that the domain of possible values of the variable is split into two parts, i.e. below and above a chosen threshold. The tail above the threshold is estimated by a tail of extreme value distribution, i.e. by a generalized

Pareto distribution. The POT method belongs to a field of mathematical statistics known as "statistics of extremes". The overview of stochastic methods suggested for studying different extremal problems in hydrology has been presented by *Katz* et al. (2002).

Extreme weather conditions are often characterized not only by a very heavy rain at one site, but rather by a heavy rain on a vast area so that daily precipitation amounts at several meteorological sites across the area are large. Here we may be interested in estimating the probability

$$P\left(X_1 > x_1,...,X_k > x_k\right) = S(x_1,...x_k), \tag{1}$$

where X_i represents a daily precipitation amount at the i-th station. Similarly, supposing that a river has k tributaries, we may be interested in estimating (1), where X_i represents a daily average discharge of the i-th tributary.

In the language of mathematical statistics we suppose that our observations are realizations of independent k dimensional vectors $\{(X_{i1}, ..., X_{ik}), i = 1, ..., n\}$ with a distribution function $F(x_1, ..., x_k)$. The goal of the statistical inference is to estimate (1) for large values $x_1, ..., x_k$. We would like to recall that for any dimension k there exists a relationship between the exceedance probability (survival function) $S(x_1, ..., x_k)$ and the corresponding distribution function $F(x_1, ..., x_k)$ given by a so called union-intersection formula. For instance for a two dimensional vector (X_1, X_2) it holds:

$$\begin{aligned} &S\left(x_1,x_2\right) = P\left(X_1 > x_1, X_2 > x_2\right) = \\ &= 1 - F_1\left(x_1\right) - F_2\left(x_2\right) + F\left(x_1,x_2\right). \end{aligned} \tag{2}$$

Similarly, for a three dimensional vector (X_1, X_2, X_3) it holds:

$$\begin{aligned} &S\left(x_1,x_2,x_3\right) = P\left(X_1 > x_1, X_2 > x_2, X_3 > x_3,\right) = \\ &= 1 - F_1\left(x_1\right) - F_2\left(x_2\right) - F_3\left(x_3\right) + F_{12}\left(x_1,x_2\right) + \\ &+ F_{13}\left(x_1,x_3\right) + F_{23}\left(x_2,x_3\right) - F\left(x_1,x_2,x_3\right), \end{aligned} \tag{3}$$

where F_1, F_2, F_3, F_{12}, F_{13}, F_{23} are the distribution functions corresponding to the lower dimensions.

Point process approach method

One of the methods for estimating (1) may be based on the point process approach. *Joe* et al. (1992) and *Coles* and *Tawn* (1991, 1994), who used the theoretical results by De *Haan* and *Resnick* (1977), worked out a procedure for application of this approach to real data. The method has been also explained in details by *Beirlant* et al. (2004). Its advantage consists in the fact that for a good estimates of (1) we do not need to estimate the distribution function $F(x_1, ..., x_k)$, respectively the survival function $S(x_1, ..., x_k)$, in its whole domain but it is sufficient to find a good estimator for large values of arguments only.

Estimating is made in two steps.

Step I

In the first step the one-dimensional distribution functions $F_1, ..., F_k$ are estimated. Usually we estimate the marginal distribution functions F_i, $i = 1, ..., k$, by the peak over threshold method, i.e., we choose subjectively a threshold u_i and we estimate the distribution function below u_i, i.e. $x \le u_i$, by a non-parametric estimate, e.g. by an empirical distribution function (or by its continuous version), while above u_i, i.e. for $x > u_i$, by a generalized Pareto distribution:

$$F^P\left(x\right) = \begin{cases} 1 - \left(1 + \dfrac{\xi}{\beta}\left(x-u\right)_+\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - e^{-(x-u)_+/\beta}, & \xi = 0, \end{cases} \tag{4}$$

where the parameters $\beta_i > 0$ and ξ_i are estimated by their maximum likelihood estimates. (We denote $a_+ = \max(a,0)$). The detailed description of the POT method may be found in *Jarušková* and *Hanek* (2006).

Using the above procedure we get the estimates of the one-dimensional marginal distribution functions for all coordinates $i = 1, ..., k$.

Step II

In the second step we estimate the "dependence structure". If we supposed for a while that all the one-dimensional marginal distributions would have been known we could transform the variables into the variables with any desired marginal distributions. One possibility would be to transform them into variables with standard normal marginals and to model the dependence structure by their correlation matrix. Clearly, in reality we do not know the right marginal distributions, but on the other hand, we can suppose that our estimates from the step I are so good that they differ from the right marginal distributions only negligibly. Fig. 1 shows a scatter plot of daily precipitation amounts at two chosen

stations and Fig. 2 shows a scatter plot of the transformed values

$$\left(\Phi^{-1}\left(\hat{\mathrm{F}}_1\left(\mathrm{x}_{\mathrm{i}1}\right)\right),\Phi^{-1}\left(\hat{\mathrm{F}}_2\left(\mathrm{x}_{\mathrm{i}2}\right)\right)\right),\ \mathrm{i}=1,\ldots,\mathrm{n}, \tag{5}$$

where Φ^{-1} denotes the inverse standard normal distribution function and $\hat{F}_1$ and $\hat{F}_2$ – empirical distribution functions that serve as estimates of the distribution functions F_1, F_2. We can see that the scatter plot in Fig. 2 does not look as a scatter plot of realizations of bivariate normal distributions with standard normal marginals. The transformed data exhibit stronger dependence in the upper tail than we would expect from realizations of a bivariate normal vector. Clearly, the idea to transform the variables into the variables distributed according to the normal distribution is not good.

Instead of transforming the variables into the normally distributed variables we suggest that the variables are transformed into the variables distributed according to the standard Fréchet distribution with the distribution function $G(x) = \exp(-1/x)$ for $x \geq 0$ and the inverse distribution function $G^{-1}(t) = -1/\log(t)$ for $0 < t < 1$. Theoretically, we use the transformation $Z_1 = -1/\log(F_1(X_1)), \ldots, Z_k = -1/\log(F_k(X_k))$. Practically, it means that we transform the data vectors $(x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$ into the vectors

$$\left(z_{ik},\ldots,z_{ik}\right)=\left(-\frac{1}{\log\left(\hat{F}_1\left(x_{i1}\right)\right)},\ldots,-\frac{1}{\log\left(\hat{F}_k\left(x_{ik}\right)\right)}\right). \tag{6}$$

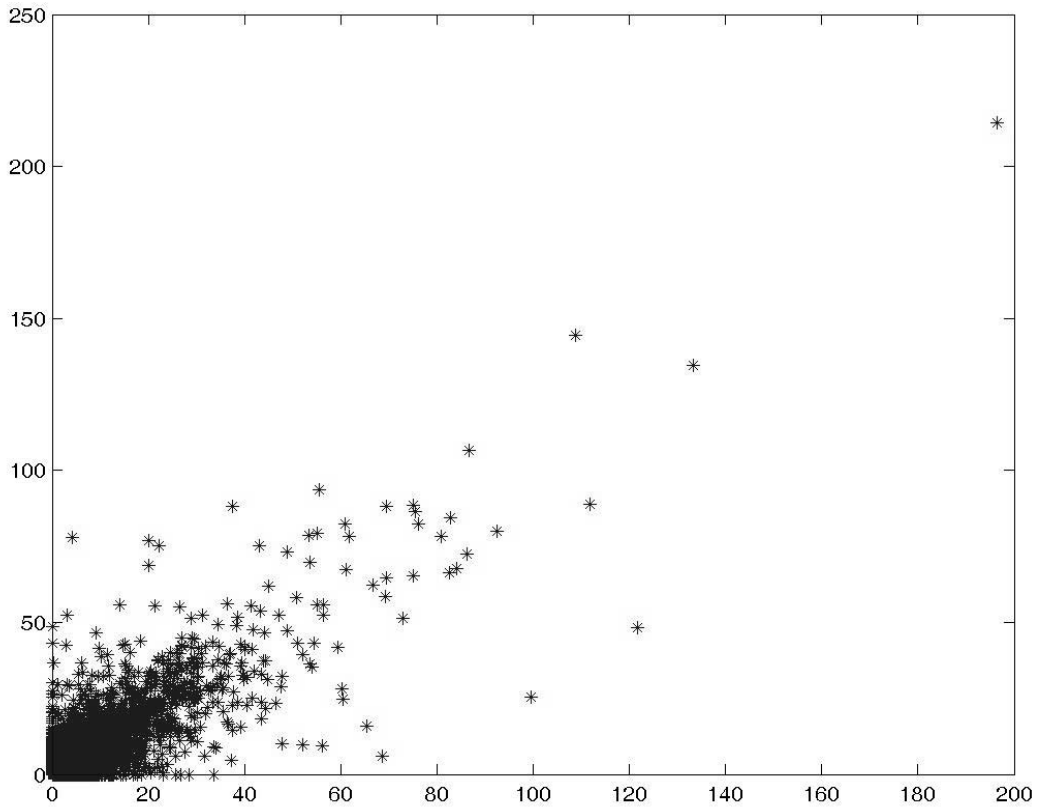

Fig. 1. Scatter plot of daily precipitation amounts measured in two stations.
Obr. 1. Rozptylový graf denních srážkových úhrnů měřených ve dvou stanicích.

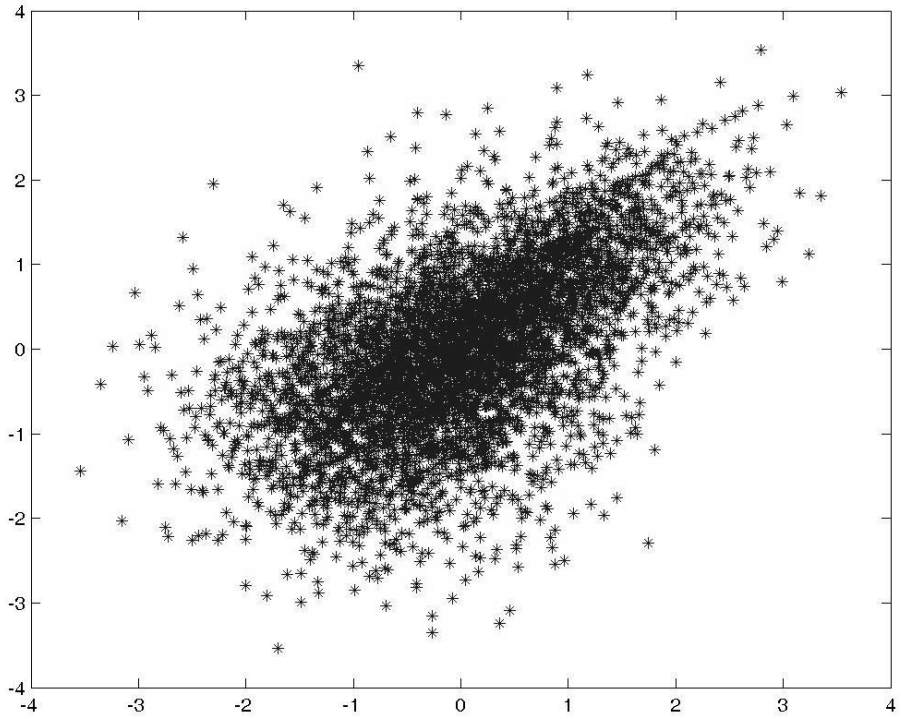

Fig. 2. Scatter plot of daily precipitation amounts measured at two station after transforming the data into standard normal variates using (5).
Obr. 2. Rozptylový graf denních srážkových úhrnů měřených ve dvou stanicích po transformaci na normálně rozdělené veličiny za použití (5).

The idea to approximate the upper tail of a multivariate vector $(Z_1, ..., Z_k)$ by a tail of extreme value distribution comes from *De Haan* and *Resnick* (1977). They proved that a subset of the transformed data with all coordinates exceeding high thresholds form a point process that can be approximated by a Poisson process defined on R^k with a nonhomogeneous intensity measure Λ. The connection between the measure Λ and the distribution function of the transformed variables $Z_1, ..., Z_k$ for $(z_1, ..., z_k)$ large can be expressed with the help of a so-called inverse arguments tail dependence function A as follows:

$$P\left(Z_1 \le z_1,...,Z_k \le z_k\right) \approx e^{-A\left(z_1,...,z_k\right)},$$
$$A\left(z_1,...,z_k\right) = \Lambda\left(\left(\left(0,z_1\right)\times...\times\left(0,z_k\right)\right)^c\right), \qquad (7)$$

where $((0, z_1) \times ... \times (0, z_k))^c$ is a complement of the multidimensional rectangle $((0, z_1) \times ... \times (0, z_k))$. For more information see *Beirlant* et al. (2004).

The advantageous property of Λ is the following. If we transform the Cartesian coordinates $(z_1, ..., z_k)$ into the coordinates $(r,\omega_1, ..., \omega_k)$ (that resemble the spherical coordinates) by the transformation $r = z_1 + ... + z_k$, $\omega_1 = z_1/r$, ... , $\omega_k = z_k/r$ then the intensity measure Λ factorizes:

$$\Lambda\left(dr, d\omega\right) = R\left(dr\right)H\left(d\omega\right) \qquad (8)$$

with the measure R having a density function $g(r) = 1/r^2$ for $r > 0$. The measure H, usually called the spectral measure, is given on the set $S_k = \{(\omega_1, ..., \omega_k), \omega_i \ge 0, i = 1,... , k, \omega_1 + ...+ \omega_k = 1\}$, e.g. for $k = 2$ the measure H is given on the vertex $\{ (0, 1), (1, 0) \}$; for $k = 3$ the measure H is given on a triangle $\{(1,0,0), (0,1,0), (0,0,1) \}$. The factorization (8) means that for any $\Omega \subset S_k$

$$\frac{\Lambda\left\{\left(r,\omega\right); r > r_1 \cap \omega \in \Omega\right\}}{\left\{\Lambda\left(r,\omega\right); r > r_1\right\}} = \frac{\Lambda\left\{\left(r,\omega\right); r > r_2 \cap \omega \in \Omega\right\}}{\left\{\Lambda\left(r,\omega\right); r > r_2\right\}}. \qquad (9)$$

The goal of the statistical inference in the step 2 is to estimate the spectral measure H. We know that a Poisson process with an intensity measure Λ is a good approximation for the subset of data that are

far away from the origin. In practice it means that we choose subjectively some threshold *ro*, transform the values $\{(z_{i1}, ..., z_{ik}), i = 1, ..., n \}$ into $\{(r_i, \omega_{i1}, ..., \omega_{ik}), i = 1, ..., n\}$ by the above introduced transformation and deal with the subset of data $\Omega_{ro} = \{(r_i,\omega_{i1}, ..., \omega_{ik}), r>ro \}$ only. Factorization (8) of Λ enables to find an estimate of the spectral measure H or its density h if it exists. Instead of dealing with $h(\omega_1, ..., \omega_k)$ on the set S_k we often estimate $h_s(\omega_1, ..., \omega_{k-1}) = h(\omega_1, ..., 1 - \omega_1 - ... - \omega_{k-1})$, i.e. for $k = 2$ we estimate $h_s(\omega) = h(\omega,1- \omega)$, for $k = 3$ we estimate $h_s(\omega_1, \omega_2) = h(\omega_1, \omega_2, 1- \omega_1 - \omega_2)$ etc. Most frequently we model the function h_s by a known mathematical function with unknown parameters $(\theta_1, ..., \theta_p)$ and estimate these parameters by their maximum likelihood estimators. More precisely we search such values of parameters that maximize

$$\sum_{\{i,(r_i,\omega_{i1},...,\omega_{ik})\in\Omega_{ro}\}} \log h_s\left(\omega_{i1},...,\omega_{i,k-1};\theta_1,...,\theta_p\right). \tag{10}$$

After having estimated the spectral density we may proceed by estimating the inverse arguments tail dependence function A by replacing the true density function h by its estimate $\hat{h}$ in the expression (11). For some models there exists an explicit formula for the integral (11), for some others the integral has to be calculated numerically.

$$A\left(z_1,...,z_k\right) = \int_{S_k} \max\left(\frac{\omega_1}{z_1},...,\frac{\omega_k}{z_k}\right)\times$$
$$\times h\left(\omega_1,...,\omega_k\right) d\omega_1...,d\omega_k. \tag{11}$$

Replacing the tail dependence function A in (7) by its estimate $\hat{A}$ the multivariate distribution function may be approximated for large values of arguments by

$$\hat{F}\left(x_1,...,x_k\right) \approx$$
$$\approx \exp\left\{-\hat{A}\left(\frac{-1}{\log\left(\hat{F}_1\left(x_1\right)\right)},...,\frac{-1}{\log\left(\hat{F}_k\left(x_k\right)\right)}\right)\right\}. \tag{12}$$

Finally, the exceedance probability may be calculated using the union-intersection formula.

Models

We present here several models that belong to a family of so-called logistic distributions. It is convenient that the inverse arguments tail dependence function has an explicit expression so that no numerical integration (11) is needed.

Bivariate logistic distribution

The spectral density function h has the form:

$$h\left(\omega_1,\omega_2\right) =$$
$$= \left(\varphi-1\right)\left(\omega_1,\omega_2\right)^{-1-\varphi}\left(\left(\frac{1}{\omega_1}\right)^{\varphi}+\left(\frac{1}{\omega_2}\right)^{\varphi}\right)^{1/\varphi-2}, \tag{13}$$

while for the function A it holds:

$$A\left(z_1,z_2\right) = \left(\left(\frac{1}{z_1}\right)^{\varphi}+\left(\frac{1}{z_2}\right)^{\varphi}\right)^{1/\varphi}. \tag{14}$$

The model has one parameter $\varphi > 1$ that expresses the dependence between the variables. The larger the value of φ the stronger the dependence. Sometimes instead of the parameter φ the parameter $\alpha = 1/\varphi$ is used.

Multivariate symmetric logistic distribution

The model is a generalization of the preceding bivariate logistic model for $k \geq 2$. It has one parameter $r > 1$ that expresses the over-all dependence. The larger the value of r the stronger the dependence. The assumption that the dependence between any couple of variables X_i, X_j, $i \neq j$, $i, j = 1, ..., k$ is the same, seems to be too restrictive but the model gives very often reasonable results and is easy to deal with. The spectral density has a following form:

$$h\left(\omega_1,...,\omega_k\right) =$$
$$= \prod_{j=1}^{k-1}\left(jr-1\right)\left(\prod_{j=1}^{k}\omega_j\right)^{-(r+1)}\left(\sum_{j=1}^{k}\frac{1}{\omega_j^r}\right)^{1/r-k}. \tag{15}$$

The function A may be expressed as follows:

$$A\left(z_1,...,z_k\right) = \left(z_1^{-r}+...+z_k^{-r}\right)^{1/r}. \tag{16}$$

Trivariate asymmetric logistic distribution

To capture the dependence between any pair of variables is not an easy task. It is possible to do it for the three dimensional case by the following

model. We applied this model to estimate the exceedance probabilities for the precipitation series measured at three meteorological stations. The spectral density has a following rather complicated form, see Eq. (17):

$$h(\omega_1,\omega_2,\omega_3) = \\ = \frac{1}{2}a\left(\frac{1}{\omega_1^{\theta}},\frac{1}{2\omega_2^{\theta}},\frac{1}{\omega_3^{\theta}}\right)\frac{1}{\omega_1^{\theta+1}}\frac{1}{\omega_2^{\theta+1}}\frac{1}{\omega_3^{\theta+1}}, \tag{17}$$

where

$$a(y_1,y_2,y_3) = (\theta-1)(2\theta-1)\left(\left(y_2^{\delta_1}+y_1^{\delta_1}\right)^{\frac{1}{\delta_1}}+\left(y_2^{\delta_2}+y_3^{\delta_2}\right)^{\frac{1}{\delta_2}}\right)^{\frac{1}{\theta}-3}$$

$$\left(\left(y_2^{\delta_1}+y_1^{\delta_1}\right)^{\frac{2}{\delta_1}-2}\left(y_2^{\delta_2}+y_3^{\delta_2}\right)^{\frac{1}{\delta_2}-1}y_1^{\delta_1-1}y_2^{\delta_1-1}y_3^{\delta_2-1}+\left(y_2^{\delta_1}+y_1^{\delta_1}\right)^{\frac{1}{\delta_1}-1}\left(y_2^{\delta 2}+y_3^{\delta_2}\right)^{\frac{2}{\delta_2}-2}y_1^{\delta_1-1}y_2^{\delta_2-1}y_3^{\delta_2-1}\right)+$$

$$+(\theta-1)\theta.\left(\left(y_1^{\delta_1}+y_2^{\delta_1}\right)^{\frac{1}{\delta_1}}+\left(y_2^{\delta_2}+y_3^{\delta_2}\right)^{\frac{1}{\delta_2}}\right)^{\frac{1}{\theta}-2}$$

$$\left((\delta_2-1)\left(y_2^{\delta_1}+y_1^{\delta_1}\right)^{\frac{1}{\delta_1}-1}\left(y_2^{\delta_2}+y_3^{\delta_2}\right)^{\frac{1}{\delta_2}-2}y_1^{\delta_1-1}y_2^{\delta_2-1}y_3^{\delta_2-1}+\right. \tag{18}$$

$$\left.+(\delta_1-1)\left(y_2^{\delta_2}+y_3^{\delta_2}\right)^{\frac{1}{\delta_2}-1}\left(y_2^{\delta_1}+y_1^{\delta_1}\right)^{\frac{1}{\delta_1}-2}y_1^{\delta_1-1}y_2^{\delta_1-1}y_3^{\delta_2-1}\right).$$

The function A may be expressed as follows:

$$A(z_1,z_2,z_3) = \\ = \left(\left(\frac{1}{z_1^{\theta\delta_1}}+\frac{1}{\left(2z_2^{\theta}\right)^{\delta_1}}\right)^{\frac{1}{\delta_1}}+\left(\frac{1}{z_3^{\theta\delta_2}}+\frac{1}{\left(2z_2^{\theta}\right)^{\delta_2}}\right)^{\frac{1}{\delta_2}}\right)^{\frac{1}{\theta}}. \tag{19}$$

The model has three parameters $\theta > 1$, $\delta_1 > 1$, $\delta_2 > 1$. The parameter θ expresses the baseline dependence between the variables Z_1, Z_3, while the parameters δ_1 and δ_2 add some dependence to the respective pairs Z_1, Z_2 and Z_3, Z_2.

Applications

The studied data are daily measurements, i.e. the daily precipitation amounts or the average discharges. We are interested in the probability that in the same day the measurements in all stations exceed certain given levels. Of course, we are especially interested in high levels that are on the border of the domain where the values were observed, or even beyond it, it means in such levels where it is unreasonable to use relative frequencies as estimators.

There are two aspects that should be considered when studying daily measurements. The first one is the dependence between the neighboring observations and the second one is the seasonality. It was shown by *Jarušková* and *Hanek* (2006) that if these aspects are not taken into account then exceedance probabilities are usually slightly overestimated. The problem of seasonality may be solved by splitting the series into more homogeneous parts corresponding to different seasons. The problem of dependence is more difficult to solve. If we are interested in the probability that a daily measurements exceed some given levels we can use a declustering technique to get a good estimate. However, the probability that during a year the measurements in all stations will exceed in the same day the given levels may be affected by this dependence. Despite

suggestions of different authors a simple way how to incorporate the dependence into the model does not exist .

Example 1

The data describes daily average discharges [$m^3 s^{-1}$] of Opava and Opavice measured at Krnov in the period 1. 11. 1963 – 31. 10. 2003, i.e. the both series consist of n = 16 071 observations. We denote by X_1 a daily average of Opava while by X_2 a daily average of Opavice. Suppose that we are interested in $P(X_1 > x_1, X_2 > x_2)$ for (x_1, x_2) = (40, 20), (45, 25), (55, 30), (100, 50).

We proceed in two steps. In the first step we estimate the marginal distributions of X_1 and X_2 by the POT method. The thresholds are chosen to be equal to the 95% quantiles of the observations.

T a b l e 1. The chosen thresholds and the estimates of the parameters of the generalized Pareto distributions for estimating the marginal distribution functions of daily discharges of Opava and Opavice.
T a b u l k a 1. Vybrané hodnoty prahů a odhady parametrů Paretova rozdělení pro odhady marginálních distribučních funkcí denních průtoků Opavy a Opavice.

River	u_i	$\hat{\beta}_i$	$\hat{\xi}_i$
Opava	11.50	4.411	0.404
Opavice	4.58	2.430	0.425

In the second step we transform the observed values and maximize (10) with the spectral density function h given by (13). The max-likelihood estimate of the parameter $\hat{\varphi}$ of the bivariate logistic distribution is 2.627. The histogram of the angular components $\{\omega_{1i}, i = 1, ..., n\}$ together with the spectral density function $h_s(\omega_1) = h(\omega_1, 1 - \omega_1)$ given by (13) is shown in Fig. 3. We see that the fit is not bad.

Tab. 2 presents the estimated exceedance probabilities using the bivariate logistic model (column 2). Column 3 shows the estimates of the same probabilities by simple relative frequencies. It seems that the estimates based on the stochastic model agree well with the relative frequencies. However, for larger values of arguments they slightly overestimate the probabilities of interest.

T a b l e 2. The exceedance probabilities $P(X_1 > x_1, X_2 > x_2)$ estimated by the suggested method and by the relative frequencies.
T a b u l k a 2. Pravděpodobnosti překročení $P(X_1 > x_1, X_2 > x_2)$ odhadnuté navrhovanou metodou a relativními četnostmi.

(x_1, x_2)	Estimates of probabilities	Relative frequencies
(40, 20)	0.00153	0.00161
(45, 25)	0.00103	0.00118
(55, 30)	0.00065	0.00062
(100, 50)	0.00017	0.00012

Example 2

To assess the probability of extreme wet weather conditions we have chosen three station in northern Moravia with different precipitation characteristics located not extremaly close to each other: Heřmanovice (HE), Albrechtice – Žáry (ZY), Lichnov (LI). The data set consists of n = 15131 daily precipitation amounts [mm] measured at each of these stations from the period 1/1/1960 – 6/2/2005 (some data are missing).

In the first step we estimate the marginal distribution functions using the POT method with the thresholds equal to the 95% quantiles of all observations. Tab. 3 presents the threshold values and the parameters of the Pareto distribution.

T a b l e 3. The chosen thresholds and the estimates of the parameters of the generalized Pareto distribution for estimating the marginal distribution functions of daily precipitation amounts at Heřmanovice, Albrechtice – Žáry and Lichnov.
T a b u l k a 3. Vybrané hodnoty prahů a odhady parametrů Paretova rozdělení pro odhady marginálních distribučních funkcí denních srážkových úhrnů ve stanicích Heřmanovice, Albrechtice – Žáry a Lichnov.

Station	u_i	$\hat{\beta}_i$	$\hat{\xi}_i$
HE	12.4	9.17	0.24
ZY	11.0	7.80	0.08
LI	9.5	7.50	0.04

In the second step we transform the data and model the dependence structure by a trivariate asymmetric logistic distribution. Fig. 4 presents a scatter plot of two first angular components calculated from the studied data.

The estimates of the parameters of the asymmetric logistic distribution obtained by the maximum likelihood method, i.e. by maximizing (10) with h defined by (17) are equal to $\hat{\theta}$ = 1.773, $\hat{\delta}_1$ = = 1.235, $\hat{\delta}_2$ = 1.221. For comparison we also model the dependence structure by a trivariate symmetric logistic distribution with the spectral density (15). The maximum likelihood estimate $\hat{r}$ = 1.81.

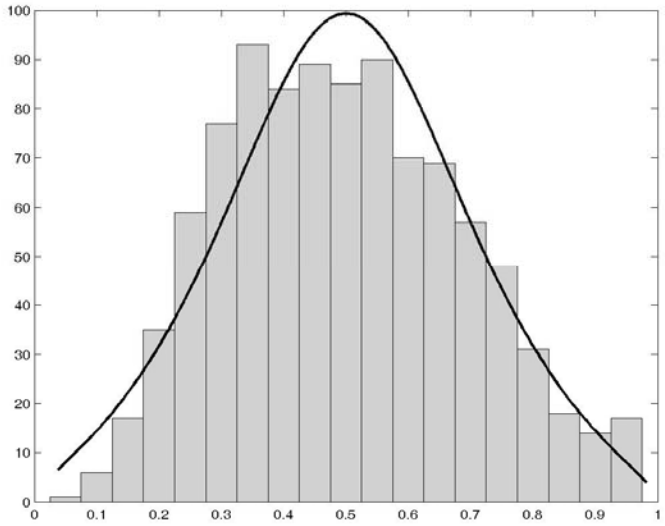

Fig. 3. Histogram of the angular components corresponding to daily average discharges of Opava and Opavice and the estimated spectral density of bivariate logistic model.
Obr. 3. Histogram úhlové složky spočtené z denních průměrných průtoků Opavy a Opavice a odhadnutá spektrální hustota logistického rozdělení pro dvě proměnné.

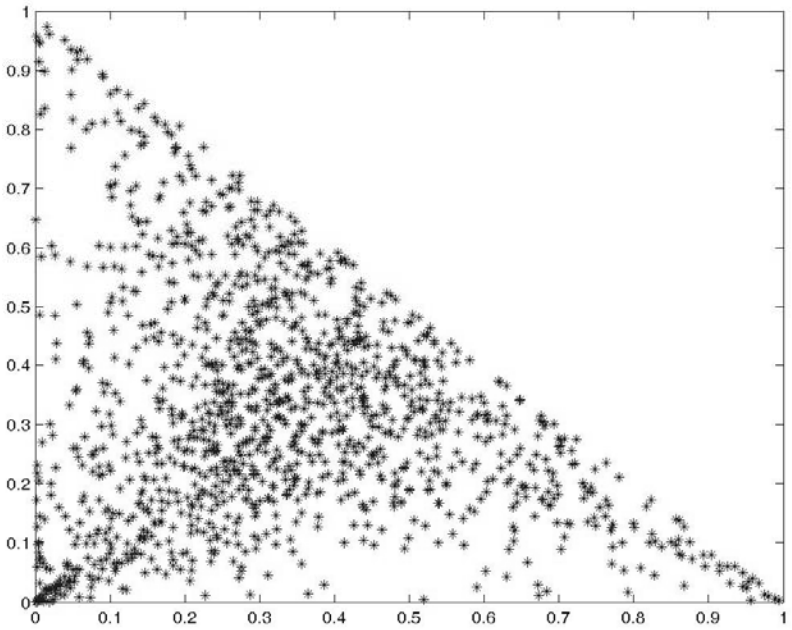

Fig. 4. Scatter plot of the first two angular components corresponding to the daily precipiation amounts measure at the stations HE, RE, LI.
Obr. 4. Rozptylový graf prvních dvou úhlových složek počítaných z denních srážkových úhrnů měřených ve stanicích HE, RE, LI.

Tab. 4 shows the estimated exceedance probabilities for several triples of levels. The real exceedance frequency was equal to 0 for all considered triples.

T a b l e 4. The estimated exceedance probabilities $P(X_1 > x_1, X_2 > x_2, X_3 > x_3)$ when the dependence structure was modeled by the asymmetric logistic distribution (column 4) and by the multivariate symmetric distribution (column 5).
T a b u l k a 4. Odhadnuté pravděpodobnosti překročení $P(X_1 > x_1, X_2 > x_2, X_3 > x_3)$, jestliže byla závislost mezi proměnnými modelována pomocí asymetrického logistického rozdělení (sloupec 4) nebo pomocí vícerozměrného symetrického rozdělení (sloupec 5).

x_1	x_2	x_3	Estimated probability	Estimated probability
75.0	83.3	34.5	$56.9\ 10^{-6}$	$30.0\ 10^{-6}$
108.9	50.9	51.1	$162.6\ 10^{-6}$	$183.1\ 10^{-6}$
196.5	125.0	51.9	$3.9\ 10^{-6}$	$2.3\ 10^{-6}$
133.3	57.0	20.0	$117.3\ 10^{-6}$	$102.6\ 10^{-6}$

Conclusion

We have described a method for estimating the probability that several series exceed at the same time, e.g. in the same day, some chosen levels. The method is a generalization of the peak over threshold method for a multivariate case. The method is based on the idea that a tail of a multivariate distribution may be approximated by a tail of a multivariate extreme value distribution. Unlike as in a one-dimendional case, in the multivariate case the tail of an extreme value distribution cannot be described by a parametric family of distributions but there exists a relationship between a multivariate extreme value distribution and an intensity measure of a point process. The goal of the statistical inference is to estimate an angular component of this measure. For estimating the spectral density we recommend a parametric family of distributions that is called a multivariate logistic distribution family.

Our experience with application of multivariate logistic distributions for modelling a dependence structure between discharge series as well as precipitation series is good. However, we have to admit that exceedance probabilities for very high levels were usually slightly overestimated. Moreover, while the method is not difficult to apply when we deal with two or three variables, modelling the dependence structure for a vector with more than three components is a rather complicated problem. In such a case we have to simplify the situation, e.g. to suppose that the dependence between any two variables is the same and to use a multivariate symmetric logistic distribution.

The hydrologists like to estimate the probability given in years rather than in days. The quality of estimates is affected by seasonality and dependence between observations in subsequent days. The effect of these two factors in one-dimensional case was discussed by *Jarušková* and *Hanek* (2006). In the multivariate case the situation is similar. According to our experience, despite the mentioned problems the method yields reasonable results.

Acknowledgemen. The study presented was partly carried out within the framework of the project MSM6840770002, GAČR 201/09/0775.

REFERENCES

BEIRLANT J., GOEGEBEUR Y., TEUGELS J., SEGERS J., DE WAAL D., FERRO CH., 2004: Statistics of extremes – theory and applications. Wiley, 490 pages.

COLES S., TAWN J., 1991: Modelling extreme multivariate events. J. R. Statist. Soc. B, *53*, 377–392.

COLES S., TAWN J., 1994: Statistical methods for multivariate extremes: an application to structural design. Appl. Statist., *43*, 1–48.

DE HAAN L., RESNICK S., 1977: Limit theory for multivariate sample extremes. Z. Wahrscheinlichkeitstheorie verw. Gebiete, *40*, 317–337.

JARUŠKOVÁ D., HANEK M., 2006: Peaks over threshold method in comparison with block-maxima method for estimating return levels of several northern Moravia precipitation and discharges series. J. Hydrol. Hydromech., *54*, 309––319.

JOE H., SMITH R., WEISSMAN I., 1992: Bivariate threshold methods for extremes. J. R. Statist. Soc. B, *54*, 171–183.

KATZ R., PARLANGE M., NAVEAU P., 2002: Statistics of extremes in hydrology. Advances in water resources, *25*, 1287–1304.

Received 12. December 2008
Scientific paper accepted 9. June 2009

ODHADOVÁNÍ PRAVDĚPODOBNOSTÍ PŘEKROČENÍ VYSOKÝCH ÚROVNÍ POMOCÍ METODY BODOVÉHO PROCESU S APLIKACEMI PRO SRÁŽKOVÉ A PRŮTOKOVÉ ŘADY NA SEVERNÍ MORAVĚ

Daniela Jarušková

Metoda popsaná v článku umožňuje odhad pravděpodobnosti, s jakou několik časových řad ve stejnou dobu překročí stanovené vysoké hodnoty. Jedná se o statistickou metodu, která na základě naměřených dat a vhodně zvoleného modelu odhaduje pravděpodobnosti překročení velmi vysokých úrovní, to znamená například i takových úrovní, které během doby měření nikdy nebyly překročeny.

Z matematické teorie extrémů plyne, že vhodným stochastickým modelem pro modelování distribuční funkce vícerozměrného rozdělení pro velké hodnoty argumentů, tj. pro modelování takzvaných „chvostů rozdělení“, jsou „chvosty“ vícerozměrných extremálních rozdělení. Vícerozměrná extremální rozdělení však netvoří jednu parametrickou třídu, pouze mají určité charakteristické vlastnosti, které musí mít i hledaný model. Vhodný model je třeba vybrat na základě zkušenosti s podobnými typy problémů. Při volbě modelu je třeba vzít v úvahu jak shodu modelu s daty, tak i výpočetní složitost úlohy.

Metoda bodového procesu je zobecněním metody špiček nad prahem (POT metody) pro vícerozměrný případ. Odhad distribuční funkce náhodného vektoru pro velké hodnoty argumentů probíhá ve dvou krocích. V prvním kroku se odhadují distribuční funkce jednotlivých složek vektoru pomocí POT metody a v druhém se odhaduje závislost mezi složkami pro velké hodnoty argumentů. Jedním z možných modelů pro modelování závislosti je třída logistických rozdělení. Naše zkušenosti ukazují, že se dá tento model úspěšně použít pro vektor se dvěma i třemi souřadnicemi.

Z praktického hlediska je modelování všech možných závislostí pro vektor s více než třemi složkami velmi obtížné. Navrženou metodu jsme použili pro odhad pravděpodobnosti překročení vysokých úrovní pro srážkové a průtokové řady z jedné oblasti severní Moravy. Jedná se o řady, které byly podrobně studovány v článku *Jaruškové* a *Haneka* (2006). Odhadovali jsme pravděpodobnost, že dvě, případně tři řady měřené v různých stanicích překročí ve stejnou dobu určité stanovené hodnoty. Pro hodnoty, které jsou nižší než dosažená maxima jsme porovnávali odhady získané pomocí popsané metody s použitím logistických modelů s relativními četnostmi. Zdá se, že shoda mezi oběma odhady je dobrá. Cílem statistické inference je však především odhadnout pravděpodobnosti překročení úrovní, které leží mimo oblast naměřených dat. Kvalitu takových odhadů však mohou prověřit jen budoucí měření.

Popsaná metoda vychází z asymptotických teoretických výsledků pro nezávislé stejně rozdělené náhodné vektory. Je zřejmé, že tyto předpoklady nejsou pro denní srážkové a průtokové řady splněny. Zkušenost ukazuje, že v takovém případě jsou odhadnuté pravděpodobnosti obvykle trochu vyšší než skutečné hodnoty.

Na závěr bych ráda zdůraznila, že navržená metota, podobně jako POT metoda v jednorozměrném případě, by měla sloužit jen jako metoda pomocná.

Seznam symbolů

$S(x_1, ..., x_k)$ – pravděpodobnost překročení (funkce přežití),
$F(x_1, ..., x_k)$ – sdružená distribuční funkce,
$F^P(x)$ – distribuční funkce zobecněného Paretova rozdělení,
u – práh pro zobecněné Paretovo rozdělení,
ξ, β – parametry zobecněného Paretova rozdělení,
$\hat{\xi}, \hat{\beta}$ – odhady parametrů zobecněného Paretova rozdělení,
Φ^{-1} – inverze distribuční funkce standardního normálního rozdělení,
G – distribuční funkce standardního Fréchetova rozdělení,
$Z_1, ..., Z_k$ – transformované veličiny se standardním Fréchetovým rozdělením,
$A(z_1, ..., z_k)$ – funkce závislosti chvostů,
Λ – intenzita nehomogenního Poissonova procesu,
H – spektrální míra,
h – spektrální hustota,
φ – parametr spektrální hustoty dvojrozměrného logistického rozdělení,
r – parametr spektrální hustoty vícerozměrného symetrického logistického rozdělení,
$\theta, \delta_1, \delta_2$ – parametry spektrální hustoty třírozměrného asymetrického logistického rozdělení.
$F_i(x_i)$, $i = 1, ..., k$ – marginální distribuční funkce,
$\hat{F}_i$, $i = 1, ..., k$ – empirické distribuční funkce.