Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

HYDROLOGICAL AND METEOROLOGICAL DATA AND MODIFICATIONS OF STATISTICAL DISTRIBUTIONS - HEURISTIC

BUDÍK Ladislav (CZ)

Abstract. The presented paper is dedicated to new approaches of fitting theoretical curves (transformed survival function) to daily average discharges, daily average air temperatures and air humidity and daily precipitation, using generalized log-normal distribution using 5 parameters and modified normal distribution.

Keywords: LN5 distribution, probability optimization method, discharge, daily air temperature, air humidity, precipitation

Mathematics Subject Classification: 62P12

1 Introduction

It is necessary in hydrology to use methods of mathematical statistics and numerical mathematics. They can effectively help even if data does not strictly satisfy all conditions, which are necessary for using such methods. One problem being solved using such mathematical methods is fitting theoretical curves with cumulative frequency curves from measured data.

Cumulative frequency function is an inverse function to the survival function.

By fitting theoretical curve one can obtain several estimated parameters for a given profile. It is then possible to easily analyze data and deduce values that cannot be directly measured – extrapolate beyond measured data (occurrence of rare phenomena such as floods). Deduced values can be used for deriving cumulative frequency curves in unobserved profiles using geographical and physical conditions in profiles (for example using regression equations for parameters of distribution).

2.1 Theoretical frequency curves and methods of parameters optimization on theoretical frequency curves of discharges

It is now necessary to describe the transformation of normal distribution into the so-called LN5 distribution. It is a generalized LN3 transform. LN5 distribution is a logarithmic-exponential transformation of normal distribution. In LN5 transformation is in contrast to

LN3 transformation used similarity of curves using parameter a. This transformation is given by the formula $Y = a \cdot e^{sign(\sigma X+\mu)\cdot|\sigma X+\mu|^b} + y_0$, where *a*, *b* and y_0 are parameters, σ is variance of the normal distribution, μ is mean shift of this distribution and *X* is standardized normal distribution [1], [2]. If one sets the parameters *a* and *b* to be equal to 1, then one obtains the LN3 distribution.

Various different metrics that optimize quality of fit are used for fitting theoretical curves. The quality of the different metrics can be compared with the help of simulations or their application using large data sets. In this way one can assess the quality of results using different metrics.

The only metric used in hydrology is the least square method derived for symmetrical distribution or weighted least square metrics used for LN2, LN3 or Pearson distribution. In climatology, it is used for other distributions, e. g. the Weibull distribution.

There is one problem with the weighted least square method. When values close to zero are used, one divides differences between measured values and theoretical values by value close to zero and then the smallest values (though they have usually lower accuracy) have greater weight than the other values. In an ideal case, the weight given to the highest and smallest values would decrease slightly. All the known metrics try to optimize the sum (usually to minimum) of metric using some relation difference between values of theoretical curve and measured one in the same probabilities for being equal or exceeding their values. This method is used for optimization of any function with measured values that characterize it.

An experiment has been done for creating a method not based on minimization of some function of differences between measured and theoretical data but which uses minimization of sum of differences between probability of measured frequency and the same theoretical value. Such a method could work generally with different frequency curves. It could be universally applied and allow deriving theoretical function quality. Such method could be usable even for survival functions because it is nearly the same. It would be possible to obtain more parameters of theoretical distribution, which could subsequently be analysed further.

When using method of probability optimization one does not minimize expression $\sum_i |F(p_i) - y_i|^d$, where $F(p_i)$ is a theoretical frequency curve, p_i is a probability for being equal or exceeding their values, y_i are aligned to measured values and *d* is an exponent (usually second power).In case of even numbers one does not need absolute value.

Using the method of probability optimization the parameters of frequency curve by heuristic iteration process can be obtained. It is necessary to choose convenient starting parameters that largely depend on user's experience.

Description of probability optimization method: suppose we have $r \geq 1$ groups of the same values y_i in a sample of *n* measured values. They have different empiric probabilities for being equal or exceeding their values. Such situation is caused mostly by inaccuracy of measurements. In an ideal case same values occur with zero probability. In *i* group we have n_i equal values y_i. Their empirical probabilities for being equal or exceeding probability values are $p_{i1}, \ldots, p_{in_i}$. Theoretical probability for equal or exceeding value y_i is marked p_i. Then $F(p_i) = y_i$. We choose convenient starting parameters for LN5 distribution and by iteration process

find parameter estimation of such distribution so that the expression $\sum_{i=1}^{r}\sum_{j=1}^{n}\left|p_{ij}-p_i\right|^d$ is minimal.

We have a lot of experience with working with hydrological data and we know that it is necessary to norm such data to average of 1. This is necessary to do if one wants to compare parameters of different curves. This is necessary to do in order to have the possibility to obtain relationships of the physical-geographical conditions of catchments. The value of normed average 1 follows from transforming of average of normal distribution, which is 0. After an exponential transformation it is e^0 =1.

2.2 Characteristics of probability optimization method

Above described method of parameters estimation seems to possess several interesting characteristics, but they are not accurately mathematically proven.

a) When using simulated data it was shown that the parameters estimation is very close to the real parameters of distribution regardless of symmetry or asymmetry of distribution.
b) The results seem to better match the course of densities similarly to the method of maximal likelyhood. However, the probability optimization method is known to work even with additive and multiplicative coefficients in expressions of distribution (it is important even in LN3 and especially LN5 distribution).
c) Method of maximal reliability has tendencies to fail under parameter estimation for data sets where groups of same values exist. Probability optimization method solves this problem very well.
d) Except for one case, in all so far tested discharge data sets (about 100 data sets) the estimation of parameters was independent of the size of d power of the difference between theoretical and empirical frequency probabilities in the range of 0.1 to 4. When using the commonly used methods for estimation based on differences between measured and theoretical values it depends on power used.
e) A simulation experiment was performed using nine different parameters. Then relative differences between simulated and theoretical values of extremely high and low values were compared. The data range in each simulation of LN5 distribution was 824 values. Frequency probability for extremely high values was chosen to be $1.214\text{x}10^{-7}$ and for extremely low values 1 - $1.214\text{x}10^{-7}$. Parameter estimations of the frequency curve were obtained by both the method of optimization of probability and the weighted least square method. Results are shown in Table 1.

Method	Average relative difference for extremely low discharges	Average relative difference for extremely high discharges
Optimization probability	0.253550	0.149132
Weighted least square	0.295576	0.172721

Tab. 1. Average relative differences - simulated and theoretical values.

As can be seen, the method of probability optimization provides better results in both cases.

f) Further analyzed were removed 10 % and then 20 % values at both tails and their effects. Frequency probabilities of the other values were retained. Data represented flows on the river Balinka in Baliny in the years 1997-2016. Estimations made by the relative least square method and method of probability optimization was depicted in Tab. 3. First row of the table gives estimations for all values, in the 2nd row 10 % values are omitted at the start and end of frequency curve, in the 3rd row 20 % is omitted as well. Last column gives extrapolated value of frequency probability 2.7397×10^{-5}. Course of frequency curves is shown on Fig. 2 and 4.

Probability optimization method						
	Est. a	Est. b	Est. μ	Est. σ	Est. y_0	Extrapolation [$m^3.s^{-1}$]
All data	2.1952	0.9123	-1.2300	1.0003	-0.0541	20,78
20 % omitted	2.0364	0.9606	-1.2337	1.0334	0.0012	23.79
40 % omitted	2.0475	0.9968	-1.2737	1.0410	0.0260	25.76
Weighted least square method						
	Est. a	Est. b	Est. μ	Est. σ	Est. y_0	Extrapolation [$m^3.s^{-1}$]
All data	11.7674	1.1309	-2.5912	0.8008	-0.0135	15.81
20 % omitted	14.4738	0.6889	-5.1105	2.1082	-0.0667	100.25
40 % omitted	14.2035	0.6825	-5.1623	2.2018	-0.0592	111.63

Est. = estimation.

Tab. 2. Parameters LN5 distribution estimations obtained by two methods under releasing of values on the tails.

Using the table one can see that at this point on the river the least square method gives the least stable estimates for the parameters of theoretical frequency curve compared to the relative least square method. Same goes for the extrapolated values.

Similar tests were done for the other six profiles – with one exception the relative least square method has always been the least stable one.

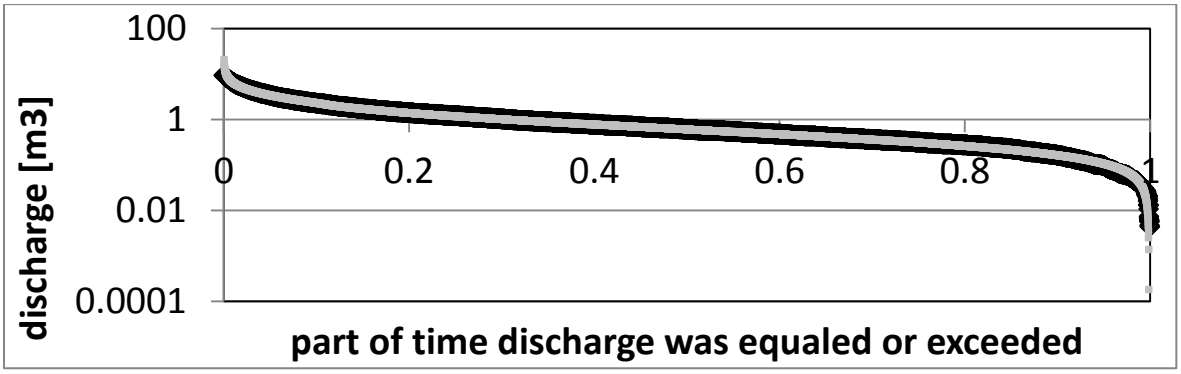

Fig. 1. Fitting theoretical distribution LN5 on daily discharges of river Balinka in Baliny between 1997-2016, probability optimization method.

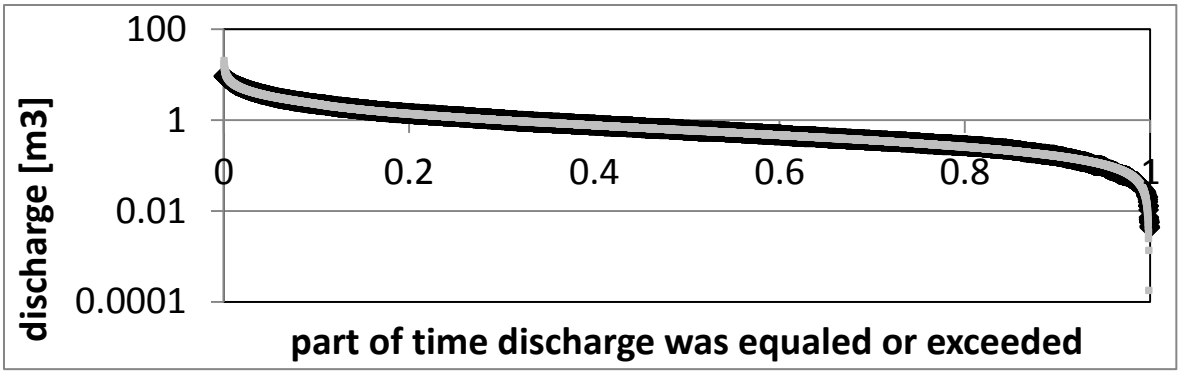

Fig. 2. Example of fit when omitting 20 % data at the tails, weighted least square method/

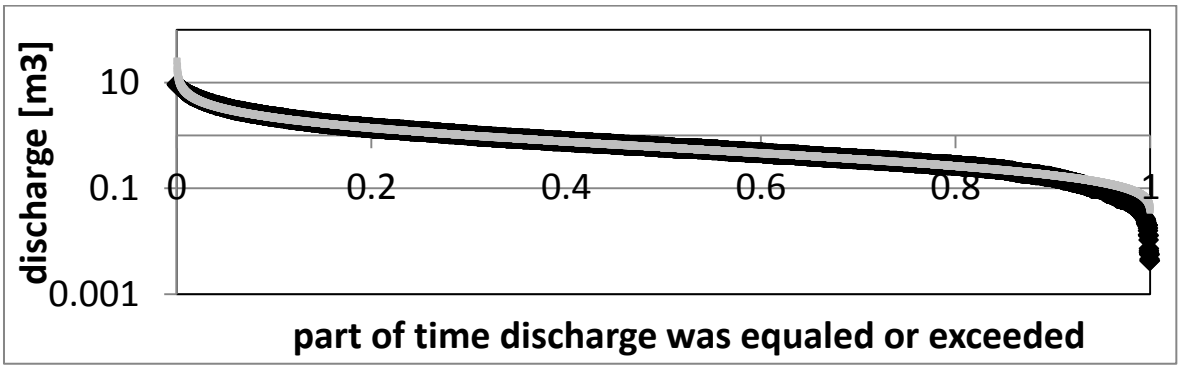

Fig. 3. Example of fit when omitting 20 % data at the tails, probability optimization method.

2.3 Fitting cumulative frequency curve for daily precipitation

To pass fit theoretical cumulative frequency curve through daily precipitation is relatively easy. Using the described methods the fit can be made including dry days. Days with precipitation are fitted by part of LN5 where values of transformed normal distribution are greater than zero. Negative values, which after transformation give very negligible values near zero, represent dry part of the curve. LN5 distribution seems to be two - parametric distribution for precipitation. Only parameters a and μ vary. The other three seem to be, at least for the region of Czech Republic, constant.

Using all five parameters gives insignificantly better result, but the question is whether the difference is only result of errors in measurements.

The values of "constant" parameters are $b = 0.18708$, $\sigma = 277769.0352$, $y_0 = 0$.

Fig. 4 shows the cumulative frequency curve of the daily precipitation at Babice nad Svitavou station between 1997 – 2016. Data are in logarithmic scale. Parameter estimation $a = 0.00008899$ and $\mu = -81874.33072$ were obtained by the probability optimization method.

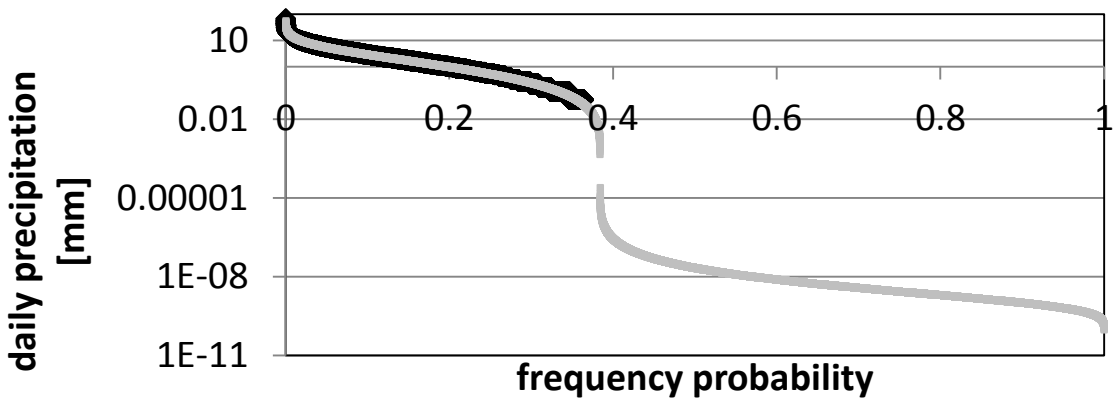

Fig. 4. Example of theoretical frequency curve of daily sum of precipitations.

2.4 Fitting cumulative frequency curve for air humidity

In order to fit air humidity data in % it was necessary to perform easy data adjustment.

When calculating ratio between saturation deficit in % and air humidity in %, the resulting ratio lies between 0 and ∞. It is then possible to estimate parameters of LN5 distribution and then divide values on theoretical frequency curve by saturation deficit and obtain theoretical frequency curve for air humidity data. Resulting theoretical curve can only have values between 0 % and 100 %. An example of such curve is depicted on Fig. 5. Data comes from the station in Pec pod Sněžkou. Parameters estimation for ratio between saturation deficit and air humidity are: a = 0.001376, b = 3.0635, μ = 6.7442, σ =2.60119, y_0 = 0.01505.

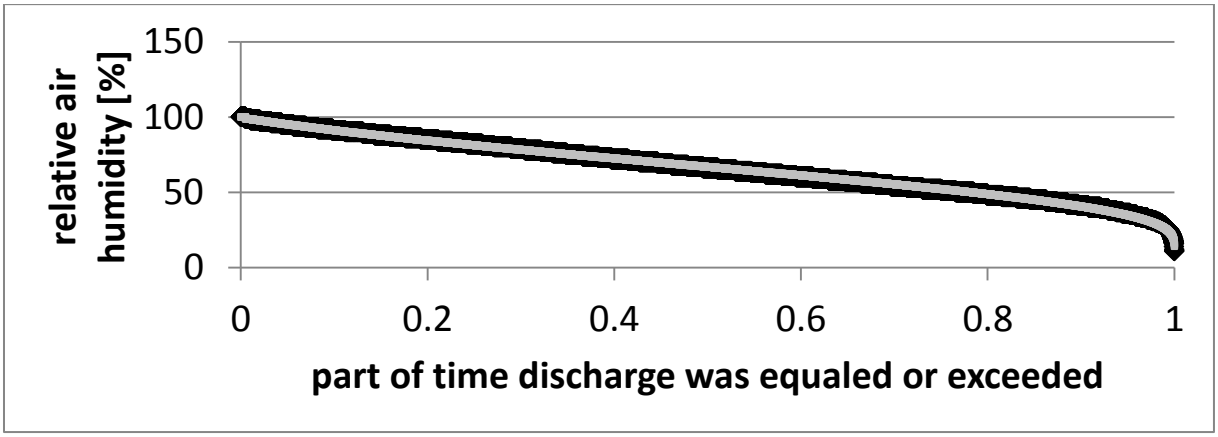

Fig. 5. Course of air humidity theoretical frequency curve.

2.5 Fitting theoretical frequency curve for daily air temperatures

Air temperatures measured in °C are usually fitted by normal distribution though they are a little asymmetric. They have extended end towards the lower values. It is possible to make them symmetrical using the Stephan – Boltzmann law: warm energy emitted from an absolutely black body with absolute temperature T_1 to the space with temperature T_2 is

proportional to the difference of fourth powers of absolute temperatures of the body and the space (in Kelvin). For easier work with data the result, it is convenient to divide it by 100 000 000. Frequency curve of such data is symmetric (Fig. 6). The particular values are from station in the city Jihlava. If it was to be fitted with a similar quality as described above it is necessary to use power transformation of normal distribution with the power marked as b (it is assumed that $b \neq 1$).

$$Y = sign\left[(X+1)\sigma + \mu\right] \cdot a \cdot \left|(X+1)\sigma + \mu\right|^{b} + y_0 - 1 \text{ for } X \geq 0$$

$$Y = sign\left[(X-1)\sigma + \mu\right] \cdot a \cdot \left|(X-1)\sigma + \mu\right|^{b} + y_0 + 1 \text{ for } X < 0$$

Subtraction or addition of 1 ensures that positive and negative part of theoretical frequency curve will be continuous. After parameter estimation of transformed data it is possible to make reverse transformation from values proportional to the emitted energy – Fig. 6 to the temperature in °C – Fig. 7. It turned out that similarly to daily precipitation it is not necessary to use all parameters but apparently at least in the Czech Rep., only one parameter is necessary, and it is the parameter a. Using all five parameters has the same consequences as precipitation.

Values of "constant" parameters are approximately $b = 0.33967$, $\mu = 0$, $\sigma = 0.0001$, $y_0 = 1$.

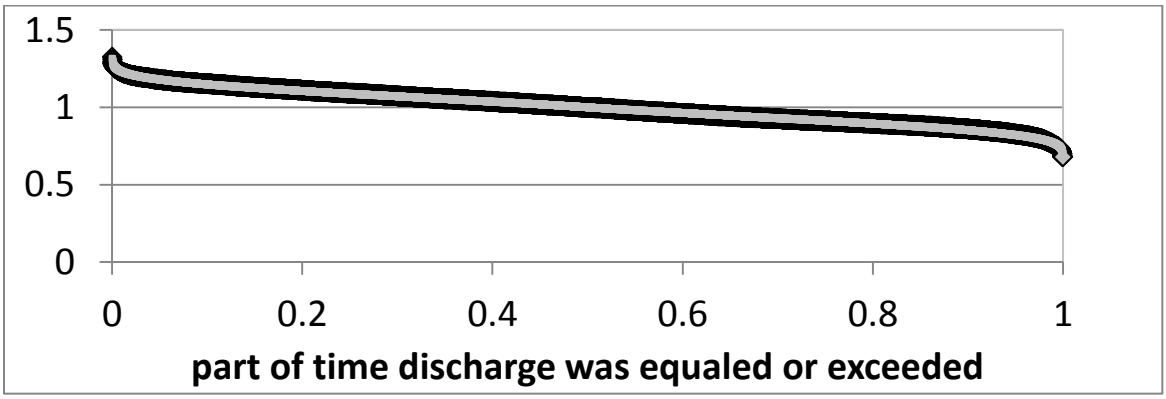

Fig. 6. Theoretical curve of exceeding values relative to emitted energy $a = 10.45342$.

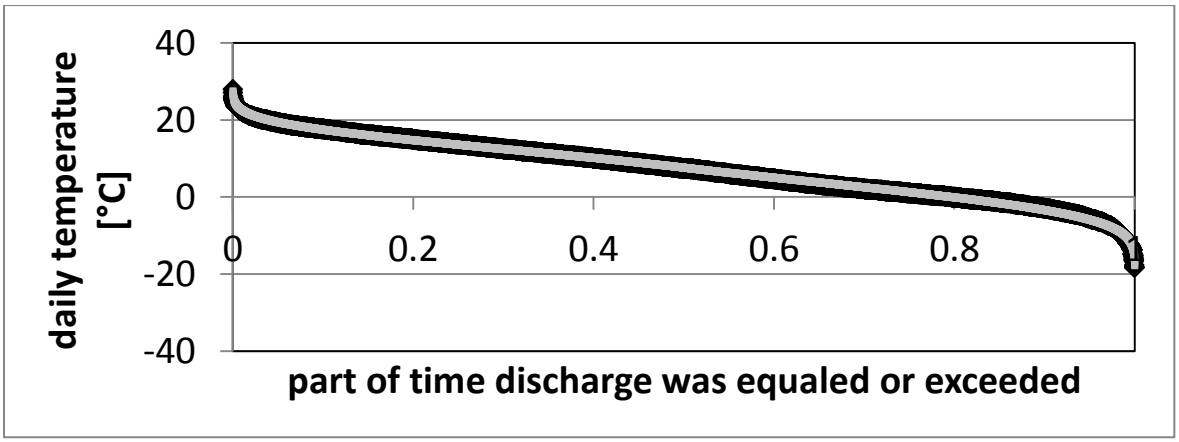

Fig. 7. Theoretical curve of exceeding temperature in °C.

3 Conclusion

It can be said that theoretical frequency curves LN5 and power transformation of the normal distribution are more accurate than the theoretical frequency curves currently used in hydrology and climatology. If probability optimization method is used, estimations of parameters are fairly stable against omitting data at the tails of theoretical frequency curve from optimization process. Their role in theoretical frequency curves must be preserved. This also means that probabilities of the measured values of the frequency curves must be preserved as well. Further advantage is that potential extrapolations of the theoretical frequency curve do not reach nonsense values even when using probabilities in the order of the age of the Earth. This is a consequence of the parameter b being < 1. If this way extrapolated data are accurate remains unknown. In addition, the Earth's climate is constantly changing. Above stated conclusion would only be correct if current climate was the same in the past 5.109 years. In the meantime, it is possible to say that these results have physical basis. Except for air temperatures where this is clearly given by type of transformation of input data (physical basis of the modification of the standard normal distribution remains questionable), this is not completely clear. More questions remains than there are answers.

Acknowledgement

I would like to thank my wife Marie for help with the mathematical part of text and revision of the whole text, thank Gejza Wimmer for a math consultation and Jáchym Brzezina for revision text translated to English language.

References

[1] BUDÍK, LADISLAV, Skládání křivek překročení časových řad průtoků (Superposition of cumulative frequency curves of flows) Bratislava, Aplimat 2015, 134-141

[2] BUDÍK, LADISLAV, KUKLA, PAVEL, ŠERCL, PETR, Odhady parametrů křivek překročení průtoků v nepozorovaných povodích a jejich optimalizace – výsledky, (Parametres estimation of frequency curves in unmeasured catchments and their optimization), Aplimat 2013

Current address

Budík Ladislav
Czech Hydrometeorological Institute, Praha-Komořany
Branch Brno
Kroftova 43, 616 67 Brno, Czech Republic
E-mail: budik@chmi.cz