ISSN 2045-3345

ISSN 2045-3345

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

# PEAKS OVER THRESHOLD VS. LOGNORMAL ESTIMATES OF THE CZECH HOUSEHOLD INCOMES

Adam ČABLA
University of Economics, Prague, Czech Republic

**ABSTRACT**

Income distributions are usually long-tailed and the right tail is often important part of income inequality metrics, but it is also problematic part of income distribution to be modeled. The POT method is theoretically well established method for modeling tails of unknown underlying distribution and thus candidate to become complement of the standard estimates. The article deals with the problem of parameter estimates using deHaan and CME methods and comparing the resulting quantile estimates with the one-distributional fitting. All of these estimates are done for the net money incomes of the Czech households. The results shows, that due to the data problems deHaan method usually gives more robust estimates than CME method. The log-normal distribution usually fits the data well up to the quantile x0,995 but for the rest of the distribution, the GPD is better fitting distribution. The peaks over threshold method is then useful only for the genuine extremes and even there its estimates depends on the quality of data.

**JEL CLASSIFICATION & KEYWORDS**

■ C16 ■ D31 ■ O15 ■ PEAKS OVER THRESHOLD ■ PARAMETER ESTIMATES ■ QUANTILE ESTIMATES ■ INCOME DISTRIBUTION ■ CZECH HOUSEHOLDS

## INTRODUCTION

There are three main approaches in parametric modeling of income distributions. The first one is to model it by one of the theoretical distributions, usually of log-normal family. The second approach is to create model of finite mixture of (usually) lognormal distributions and finally the third one is to model upper and lower parts of income distribution separately, especially where there is interest in the upper part, which is usually modeled by Pareto distribution. The first two approaches in modeling Czech household´s income were for the last time used by Čabla (2011) and Malá (2010), respectively, whereas the third one appeared in the modeling of upper-median wage distribution in Bílková (2009).

In the present article the right tail distribution is of interest and generally the third approach is used and compared with the first one. The first chapters cover extreme value theory and peaks over threshold method and two parameter estimates methods used later. Then step-by-step example of the proceeding is demonstrated and finally results obtained are shown and discussed.

### Extreme Value Theory

Extreme Value Theory (EVT) is used where there is interest in the modeling of extremes of the distribution. Among its many applications belongs for example meteorology, hydrology, insurance or finance.

In modeling of extremes there are two main methods. Block maxima method considers maximums (or minimums) in random intervals, usually time periods, and the distribution of these maximums converges to the generalized extreme value distribution. Peaks over threshold (POT) method is based on the theorem, that distribution of random variables that exceeds certain, sufficiently high value called threshold, converges to the generalized Pareto distribution.

The first method can lead to the loss of information in contrast to the POT as it considers only one data point in every block, for example only one river flow every year, but usually avoids the problem of correlation in time-data series, i.e. in the given example that river flow at time t is not independent from the river flow at time t+1, which is condition of the method.

### Generalized Pareto Distribution

Values of random variable that exceed certain sufficiently high threshold u for a large class of distributions converges according to Pickands-Balkema-de Haan theorem to general Pareto distribution. As stated in Vojtěch (2011):

Let (X1, X2,…) be a sequence independent and identically distributed random variables with distribution function F. Random variables for which X > u has excess distributional function

$$F_u(y) = P(X - u \le y | X > u)$$
$$\text{for } 0 \le \omega F - u, \qquad (1)$$

where X is random variable, u is given threshold, y = x – u are excesses and ωF ≤ ∞ is right point of the underlying distribution. Then:

$$F_u \to H_{\xi,0,\beta} = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$$
$$\text{as } u \to \infty. \qquad (2)$$

Parameter ξ plays a crucial role in the behavior of the tail of distribution and general Pareto distribution can take one of the three forms: Pareto distribution if ξ > 0, exponential distribution if ξ = 0 or beta distribution if ξ < 0.

### Pareto Distribution and False Power Law

Pickands-Balkema-de Hann theorem explains why it can be convenient to use Pareto distribution in modeling high incomes distribution. Inspiring article by Perline (2005) shows that what is usually considered to be Pareto distribution is often just arbitrary truncated sample of data from another distribution. That's what he calls the false power law. He went even further and simulated finite mixture of three lognormal distributions and then truncated it. The result was that at the 90 % truncation, i.e. with using upper 10 % of the sample, the distribution mimicked the Pareto.

Truncation in these samples could be in fact just the way how the general Pareto distribution arises and with the knowledge of the extreme value theory it should be no surprise, that the truncated right tail of the distribution can take form of Pareto distribution and often does.

If the income distribution would by some hidden law followed the finite mixture of lognormal distributions as it is quite popular to model it, then use of general Pareto distribution

to model the right truncated tail is convenient as well. And if the income distribution would followed another distribution or mix of distributions, it still could be right way to model it by general Pareto distribution as well.

**Parameter Estimation in POT**

There are several estimation methods, the fist used here is de Haan method as described in Simiu and Heckert (1996).

Let k be the number of observations above threshold u. We have λ = k/n where "n" is the length of the record. The highest, the second highest,.. k-th highest, (k+1)th highest variates are denoted $X_{n,n}$ $X_{n-1,n}$ …, $X_{n-(k+1),n}$ respectively. Compute quantities:

$$M_n^{(r)} = \frac{1}{k}\sum_{i=0}^{k-1}(\log(X_{n.-1,n}) - \log(X_{n-k,n}))^r$$

for r = 1,2. (3)

The estimators of ξ and β are then:

$$\hat{\xi} = M_n^{(1)} + 1 - \frac{1}{2\left\{1 - (M_n^{(1)})^2 / (M_n^{(2)})\right\}}$$

$$\hat{\beta} = u M_n^{(1)} / \rho_1$$

ρ1 = 1 for ξ ≥ 0 otherwise ρ1 = 1/(1-ξ). (4)

The second used method is CME method as described by Gross, Heckert, Lechner and Simiu (1995):

The CME (conditional mean exceedance) is the expectation of the amount by which a value exceeds a threshold u, conditional on that threshold being attained. If the exceedance data are fitted by the GPD model and ξ < 1 and β+uξ > 0, then the CME vs. u plot should follow a line with intercept β/(1-ξ) and slope ξ/(1-ξ). The linearity of the plot is an indicator of the appropriateness of the GPD model. Estimates of ξ and β are thus obtained from the slope and intercept of the straight line fit to the CME vs. u plot.

This fit is done by least maximum square estimates.

**Threshold Determination**

The theory does not propose any objective method for threshold determination, there are mainly graphical ad hoc approaches on which good summarizing article was provided by Tanaka and Takara (2002).

The approach used in this paper is to contrast estimates of shape parameter ξ and number of observations above threshold. The less the observations above threshold the higher the variance of gamma is. On the other hand higher threshold means better GPD approximation of the tail, therefore with rising number of observations above threshold comes higher bias of the estimate. It means that over intervals where the bias is small the plot should be horizontal.

Another possible graphical approach can be based on the CME vs. u plot. Where there is a straight line, there should be GPD model appropriate, so the highest possible threshold should be set at the point of the beginning of this line.

**Three-parameter lognormal distribution**

Three-parameter lognormal distribution is often used to model income distributions and is usually considered to be good at fitting central part of the distribution of interest, but the fitting of the tails is often problematic. In this article this distribution was chosen to be comparative distribution to show the possible positive outcomes of the using of POT method. The estimates of the parameters are done by the maximum likelihood method.

**Data**

Data used in this work are net money incomes of the Czech households and come from the Czech Statistical Office´s (CZSO) surveys in the years 1992, 1996, 2002 and 2005 through 2009. Years 1992, 1996 and 2002 were covered by mikrocensus surveys while the others were covered by EU-SILC surveys. Data from the year 2010 are not available yet.

Example: The year 1992

In this chapter the concrete proceeding is shown for the net money income of the Czech households in the year 1992.

The threshold determination as described above is shown in Figure 1 for de Haan estimation method and in Figure 2 for CME estimation method.

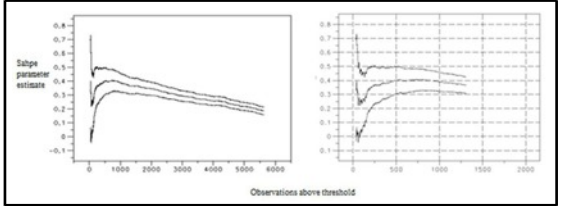
Upper and lower lines show 95% confidence interval and middle line shows the estimate itself. High variance produces large jumps in estimate at the beginning especially where there are less than 500 observations.

With de Haan method as soon as at 1 000 observations above threshold the estimate begins lowering which could mean that bias is taking place. From the closer look is seen that the similar estimate of shape parameter is given with approximately 500 – 900 observations above threshold which gives threshold between 176 847 and 202 992. With lesser threshold and more observations above it there is narrower confidence interval, so with this approach the threshold is determined at value 176 847. As in this year there were 16 234 households in the survey, there are approximately 5.54 % of them above threshold and so subject to modeling.

Parameter estimates are thus ξ = 0.3982 and β = 47 820.

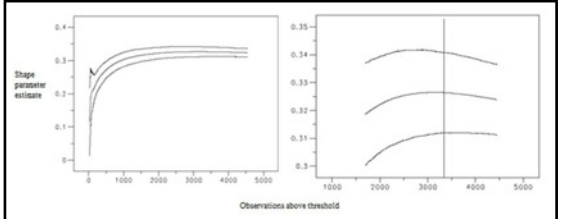
With CME method estimate seems to be quiet stable around 3 000 observations above threshold and closer look reveals that from approximately 3 300 observations above threshold the estimate begins to lower which is about 20.33 % of the households. The threshold is then 123 504 and parameter estimates are ξ = 0.3263 and β = 32 839.

Figure 1: Threshold determination for the year 1992 – de Haan




Source: CZSO, own calculations

Figure 2: Threshold determination for the year 1992 – CME



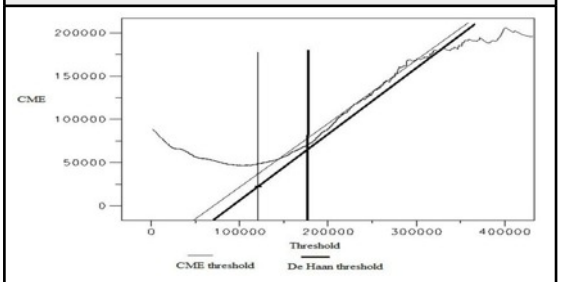

Source: CZSO, own calculations

Figure 3 shows CME vs. u plot, the second method of threshold determination described previously. The plot suggests that the threshold is actually underestimated and should be put somewhere around the threshold obtained by de Haan method, but GPD fits to the data doesn´t seem to favor any of the two thresholds considerably.

Having parameters estimated obtaining high quantiles estimates is quite simple. The three quantiles to be estimated are $x_{0.95}$, $x_{0.99}$ and $x_{0.999}$ which mean the estimated income of the 950th, 990th 999th highest earning households out of 1000 randomly chosen households.

As for example de Haan method deals with the 5.54 % of the highest incomes, the 95th highest income in the whole dataset is quantile $y_{0.0974}$ of the GPD with given parameters.

Figure 3: CME vs. u plot with highlighted thresholds




Source: CZSO, own calculations

The last estimate is done for "the highest earning household in the Czech Republic". The estimate of the number of households for the years for which the GPD estimates are done is made in a simple linear manner from number of households according to the CZSO´s LFS surveys. The result is showed in Table 1.

Table 1: Estimated number of households in the Czech Republic

| Year | 1992 | 1996 | 2002 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| House-holds | 3594 | 3725 | 3953 | 4100 | 4162 | 4224 | 4287 | 4349 |

Source: CZSO, own calculations

The income of the highest earning household in the Czech Republic in the year 1992 was then estimated as the income of the 3 594 000th highest earning household out of 3 594 001, which is around quantile $x_{0.999999722}$. It is 15 530 846 according to de Haan method or 8 266 204 according to CME method.

Table 2 gives the estimated parameters for the year 1992 by both methods and Table 3 gives the estimated quantiles by both methods and nonparametric estimates from the sample. In both tables the estimates of and by the three-parametric lognormal distribution are provided to show difference between whole-distribution estimates and specific right tails POT estimates.

Table 2: Estimated parameters for the year 1992

| | | | CME | | | |
|---|---|---|---|---|---|---|
| Year | Observations in sample | Thresh-old | Thresh-old | Obs. above threshold (%) | ξ | β |
| 1992 | 16 234 | 176 847 | 123 504 | 20,33 | 0,3263 | 32 839 |
| | De Haan | | | Lognormal (3P) | | |
| | Obs. above threshold (%) | ξ | β | σ | μ | γ |
| 1992 | 5,54 | 0,3982 | 47 820 | 0,56258 | 11,286 | -3597 |

Source: CZSO, own calculations

Table 3: Estimated quantiles for the year 1992

| Method | $x_{0.95}$ | $x_{0.99}$ | $x_{0.999}$ | Highest Earning |
|---|---|---|---|---|
| de Haan | 181 853 | 294 208 | 650 739 | 15 530 846 |
| CME | 181 917 | 291 780 | 592 920 | 8 266 204 |
| LN (3P) | 197 464 | 291 411 | 449 791 | 1 328 318 |
| non-parametric | 181 422 | 276 155 | 594 036 | 1 784 554 |

Source: CZSO, own calculations

**Results and discussion**

In the following tables there are summarized resulting estimates obtained for all years available. In Table 4 there are the number of observations in the sample and the estimated parameters. In Table 5 there are the estimated parameters of the lognormal distributions (three-parametric or where the goodness of fit tests favored it, two-parametric), Tables 6 and 7 show estimated quantiles alongside with the non-parametric estimates (np). The values closest to the non-parametric estimates are boldly highlighted. In Table 8 there are the estimations of the highest earning household´s incomes - the column np covers the highest observations in sample, the last four columns contains the estimates with the threshold set at $x_{0,9}$ and $x_{0,95}$, respectively. Highlighted are always the largest results in the given year.

Table 4: Estimated parameters of GPD

| | | | de Haan | | |
|---|---|---|---|---|---|
| Year | n | Threshold | Obs. above threshold (%) | ξ | β |
| 1992 | 16 234 | 176 847 | 5,54 | 0,3982 | 47 820 |
| 1996 | 28 148 | 349 500 | 4,44 | 0,3734 | 98 586 |
| 2002 | 7 973 | 454 165 | 6,27 | 0,3406 | 130 450 |
| 2005 | 4 351 | 477 542 | 7,47 | 0,3578 | 127 932 |
| 2006 | 7 483 | 502 291 | 6,88 | 0,3185 | 136 035 |
| 2007 | 9 675 | 384 199 | 19,64 | 0,2249 | 117 567 |
| 2008 | 11 294 | 416 187 | 19,92 | 0,2476 | 124 220 |
| 2009 | 9 911 | 627 606 | 6,56 | 0,3762 | 178 326 |
| | | | **CME** | | |
| Year | n | Threshold | Obs. above threshold (%) | ξ | B |
| 1992 | 16 234 | 123 504 | 20,33 | 0,3263 | 32 839 |
| 1996 | 28 148 | 217 700 | 21,33 | 0,2952 | 67 301 |
| 2002 | 7 973 | 429 751 | 8,15 | 0,3812 | 113 442 |
| 2005 | 4 351 | 290 731 | 28,73 | 0,281 | 107 572 |
| 2006 | 7 483 | 556 273 | 4,68 | 0,3802 | 134 708 |
| 2007 | 9 675 | unable to obtain | 19,64 | 0,3359 | 109 009 |
| 2008 | 11 294 | 416 187 | 15,94 | 0,2662 | 127 547 |
| 2009 | 9 911 | 397 007 | 27,24 | 0,294 | 131 037 |

Source: CZSO, own calculations

Table 5: Estimated parameters LN (3P)

| Year | σ | μ | γ |
|---|---|---|---|
| 1992 | 0,56258 | 11,286 | -3597 |
| 1996 | 0,62565 | 11,772 | 2666,2 |
| 2002 | 0,62327 | 12,113 | 7921,3 |
| 2005 | 0,60908 | 12,213 | 2775,2 |
| 2006 | 0,59434 | 12,259 | 2100,1 |
| 2007 | 0,59439 | 12,34 | 2493,7 |
| 2008 | 0,59206 | 12,431 | 0 |
| 2009 | 0,59705 | 12,501 | 0 |

Source: CZSO, own calculations

Table 6: Estimated quantiles x0.95 and x0.99

| | | $x_{0.95}$ | | |
|---|---|---|---|---|
| Year | deHaan | CME | LN (3P) | np |
| 1992 | **181 853** | 181 917 | 197 464 | 181 431 |
| 1996 | xxx | **339 570** | 365 283 | 338 100 |
| 2002 | 484 859 | **490 674** | 515 897 | 495 949 |
| 2005 | **532 773** | 533 625 | 551 223 | 531 600 |
| 2006 | **547 994** | Xxx | 562 610 | 547 336 |
| 2007 | 572 538 | **573 519** | 610 344 | 588 701 |
| 2008 | **620 938** | 589 424 | 663 213 | 633 321 |
| 2009 | **678 591** | 684 972 | 717 164 | 676 290 |
| | | $x_{0.99}$ | | |
| Year | de Haan | CME | LN (3P) | Np |
| 1992 | 294 208 | 291 780 | **291 411** | 276 518 |
| 1996 | **545 881** | 552 346 | 558 086 | 525 500 |
| 2002 | 786 894 | 794 306 | **784 725** | 775 428 |
| 2005 | 854 188 | 891 437 | **833 399** | 764 665 |
| 2006 | 864 609 | **839 066** | 842 507 | 831 641 |
| 2007 | **882 677** | 941 976 | 913 913 | 898 972 |
| 2008 | **966 804** | 938 348 | 992 853 | 965 421 |
| 2009 | 1 115 454 | 1 128 902 | **1 077 276** | 1 043 634 |

Source: CZSO, own calculations

Table 7: Estimated quantile $x_{0.999}$

| | | $x_{0.999}$ | | |
|---|---|---|---|---|
| Year | deHaan | CME | LN (3P) | np |
| 1992 | 650 739 | **592 920** | 449 791 | 607 090 |
| 1996 | **1 173 420** | 1 099 973 | 898 405 | 1 180 400 |
| 2002 | **1 639 174** | 1 724 934 | 1 258 415 | 1 580 772 |
| 2005 | **1 793 441** | 1 786 300 | 1 325 492 | 1 941 640 |
| 2006 | **1 718 844** | 1 730 969 | 1 325 413 | 1 596 005 |
| 2007 | **1 575 497** | 1 971 806 | 1 437 677 | 1 600 577 |
| 2008 | 1 775 485 | **1 785 314** | 1 560 634 | 2 164 046 |
| 2009 | **2 440 845** | 2 268 684 | 1 699 803 | 2 886 000 |

Source: CZSO, own calculations

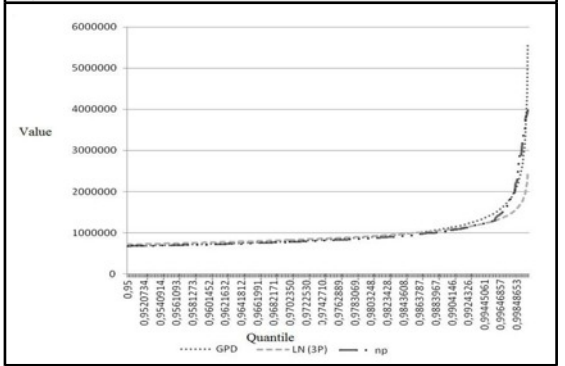
Table 8: Estimated highest earning household´s income in the Czech Republic

| | | **As above** | | |
|---|---|---|---|---|
| Year | de Haan | CME | LN (3P) | np |
| 1992 | **15 530 847** | 8 266 204 | 1 328 318 | 1 784 554 |
| 1996 | 23 611 452 | **12 567 953** | 2 984 806 | 3 192 600 |
| 2002 | 26 401 844 | **37 569 312** | 4 181 716 | 5 110 628 |
| 2005 | **32 949 702** | 19 360 507 | 4 316 494 | 3 262 118 |
| 2006 | **23 439 916** | 36 548 870 | 4 203 225 | 4 891 034 |
| 2007 | 11 067 401 | **31 638 736** | 4 566 896 | 5 569 100 |
| 2008 | **14 673 804** | 17 062 571 | 4 949 220 | 4 103 711 |
| 2009 | **53 619 530** | 27 158 513 | 5 452 193 | 5 294 482 |
| | **From last 10%** | | **From last 5 %** | |
| Year | de Haan | CME | de Haan | CME |
| 1992 | 11 993 147 | 7 225 018 | **16 311 799** | 6 173 719 |
| 1996 | **11 896 373** | 18 223 851 | 22 014 455 | 9 485 327 |
| 2002 | 19 901 819 | 37 820 742 | **28 183 472** | 36 627 761 |
| 2005 | 21 396 114 | 16 621 226 | **42 081 118** | 12 622 750 |
| 2006 | 15 511 989 | **36 974 021** | 26 349 499 | 36 356 764 |
| 2007 | **11 818 955** | 38 579 010 | 16 317 940 | 43 258 527 |
| 2008 | **18 651 283** | 16 989 021 | 28 809 551 | 14 333 214 |
| 2009 | 41 202 088 | 21 357 171 | **65 515 874** | 16 958 829 |

Source: CZSO, own calculations

Based on the tables 6 to 8, the fit by lognormal distribution seems to overestimate the lower quantiles of the right tail ($x_{0,90}$) and underestimate higher quantiles ($x_{0,999}$) and to be effectively of at least the same quality as the POT estimates somewhere between these two, given that in the four of eight years the estimates of $x_{0,99}$ where closest to the data available. This interesting development is plotted in Figure 4.

Figure 4: Inverse cumulative distribution function




Source: CZSO, own calculations

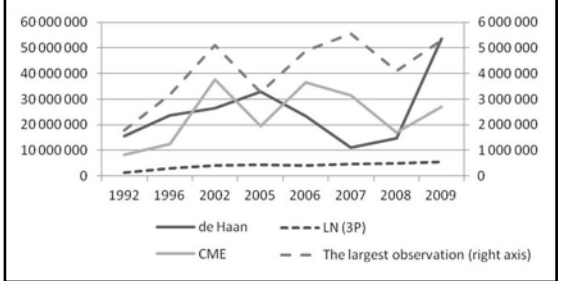
The Figure 5 plots the estimated highest earnings obtained by the three methods and the largest value at the sample on the right axis. The morale is quite obvious that there is a strong correlation between the largest value and the estimate obtained by the CME method – it is the problem of the linear regression estimate being affected by the outlier. The correlation coefficient is always between 0.8 and 0.9 as it is shown in the Table 9.

The criterion for highlighting in the table 8 is somewhat discussable as the actual highest income of the Czech households is not known, but based on the common knowledge it is still good assumption, that it is higher.

The better estimate of the rare occurrences is nevertheless done by POT method and deHaan estimates seems to be better off a bit as it is not that affected by the highest observation. But still there is no objective method of assessing threshold or estimate method.

The main problem stems from the data available. If the highest earnings are not sufficiently covered, as it seems to be the case at least for the year 2008, the estimation of the tail is underestimated even with the POT method, but still much less than by standard one-distribution approach. It is all a part of broader philosophical discussion about extreme values estimates obtained from the samples, the topic skeptically covered i.e. in Taleb (2010).

Figure 5: Estimated highest household´s incomes in the Czech Republic




Source: CZSO, own calculations

| Table 9: Correlation between estimates of the highest income in the Czech Republic and the highest observed income | | | |
|---|---|---|---|
| Method | As above | 10% | 5% |
| de Haan | 0,243762 | 0,388204 | 0,307518 |
| CME | 0,877723 | 0,843584 | 0,808949 |
| Source: CZSO, own calculations | | | |

## Conclusion

The paper covered the topic of POT method trying to obtain estimates for the right tail of the income distribution of Czech households and compared it to the classic approach of modeling incomes by three-parametric lognormal distribution. Estimates, especially those by de Haan method, seem to make a good fit to the sample data, but the problem arises with the genuine extremes. Nevertheless the fit in the right tail is still much better than the fit done by simple distributional fitting to whole data set. It is almost necessary ad-on to this approach.

## Acknowledgment

The article was supported by grant IGS 24/2010 from the University of Economics, Prague.

## References

Bílková, D. (2009). Pareto Distribution and Wage Models. Aplimat [CD-ROM], roč. II, č. III, 37–46. ISSN 1337-6365.

Čabla, A. (2011) Modelování příjmových rozdělení pomocí čtyřparametrického logaritmicko-normálního rozdělení. In: Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze [CD]. Praha: Oeconomica, 136–140. ISBN 978-80-245-1761-2.

Gross, J.L., Heckert, N.A, Lechner, J.A. & Simiu, E. (1995). Extreme Wind Estimates by the Conditional Mean Exceedance Procedure. Journal of Structural Engineering.

Malá, I. (2010). Generalized Linear Model and Finite Mixture Distributions. Demänovská Dolina 25.08.2010 – 28.08.2010. In: AMSE 2010 [CD]. Banská Bystrica : Občianske združenie Financ, 225–234. ISBN 978-80-89438-02-0.

Perline, R. (2005). Strong, weak and false inverse power laws. Statistical Science, 20(1), 68-88.

Simiu, E., & Heckert, N.A. (1996). Extreme Wind Distribution Tails: A "Peaks Over Threshold Approach". Journal of Structural Engineering.

Taleb, N.N., (2010). The Black Swan: The Impact of the Highly Improbable. Random House Trade Paperbacks. New York.

Tanaka, S., & Takara, K. (2002) A study on threshold selection in POT analysis of extreme floods. The Extremes of the Extremes: Extraordinary Floods, 271, 299 – 304.

Vojtěch, J. (2011). Využití teorie extrémních hodnot při řízení operačních rizik (Dissertation). Vysoká škola ekonomická v Praze.