PEAKS OVER THRESHOLD IN MODELLING OF THE CZECH HOUSEHOLD INCOME DISTRIBUTION

Adam Čabla

Abstract

The article deals with the usage of "Peaks over Threshold" (POT) method in modeling tails of distribution of incomes of the Czech households and estimating high quantiles of these incomes.

Income distributions are usually considered long-tailed distributions and the right tail is often important part of income inequality metrics, especially in ratio of percentiles measures. It is also very problematic part of the income distribution to be modeled.

The POT method as a part of extreme value theory is a theoretically well supported method for modeling tails of an unknown underlying distribution and thus estimating high quantiles. Main problem of this method is the choice of a suitable threshold, therefore the article will discuss several possibilities for choosing threshold and then resulting tail models and quantile estimates.

These estimates are done for the Czech households as whole.

Data in this work were collected in Czech Statistical Offices (CZSO) surveys in the years 1992, 1996, 2002 and 2005 through 2009.

Key words: Peaks over threshold, generalized Pareto distribution, quantile estimation, income distribution, Czech households

JEL Code: JEL Code, JEL Code, JEL Code (2 – 3)

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 generally the third approach is used as the object of interest is the distribution of the highest incomes and estimation of the very high quantiles. In the first chapter there is a brief summary of the peaks over threshold method.

1 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.

1.1 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 (X_1, X_2,…) be a sequence independent and identically distributed random variables with distribution function F. Random variables for which $X > u$ have excess distributional function

$$F_u(y) = P(X - u \le y | X > u)$$

$$for \quad 0 \le \omega_F - u, \tag{1}$$

where X is random variable, u is given threshold, $y = x - u$ are excesses and $\omega_F \le \infty$ is right point of the underlying distribution. Then:

$$F_u \to H_{\xi,0,\beta} = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$$

$$as\ u \to \infty. \tag{2}$$

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 $\xi > 0$, exponential distribution if $\xi = 0$ or beta distribution if $\xi < 0$.

1.2 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 is 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.

1.3 Parameter Estimation

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:*

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

$$for\ r = 1,2. \tag{3}$$

The estimators of ξ and β are then:

$$\hat{\xi} = M_n^{(1)} + 1 - \frac{1}{2\{1-(M_n^{(1)})^2/(M_n^{(2)})\}} \tag{4}$$

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

$\rho_1 = 1$ *for* $\xi \geq 0$ *otherwise* $\rho 1 = 1/(1-\xi)$. (5)

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 $\xi < 1$ *and* $\beta+u\xi > 0$*, then the CME vs. u plot should follow a line with intercept* $\beta/(1-\xi)$ *and slope* $\xi/(1-\xi)$*. 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.

1.4 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.

2 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.

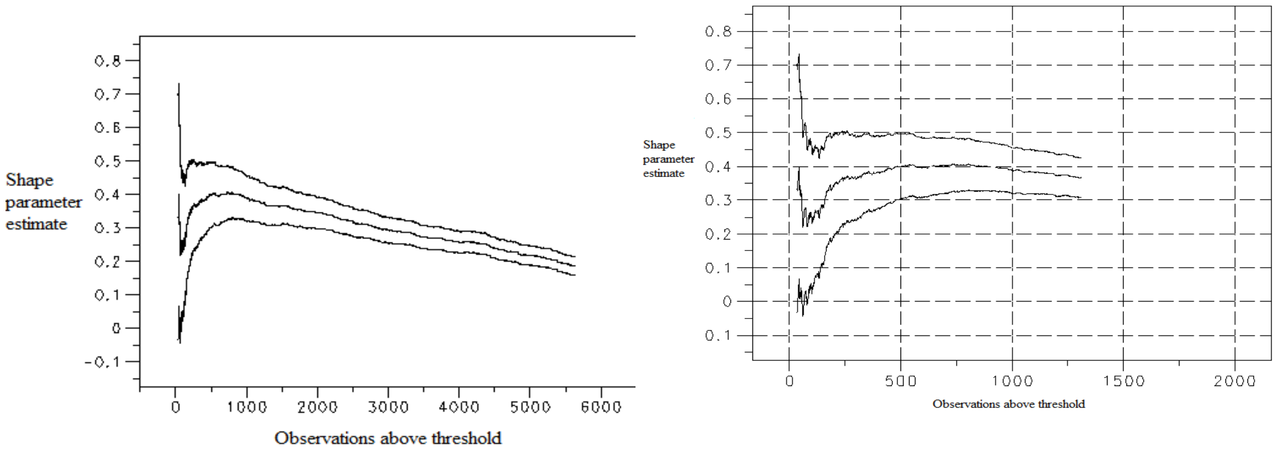
3 Example: the Year 1992

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 in chapter 1.4 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.

Fig. 1: Threshold determination for the year 1992 – de Haan

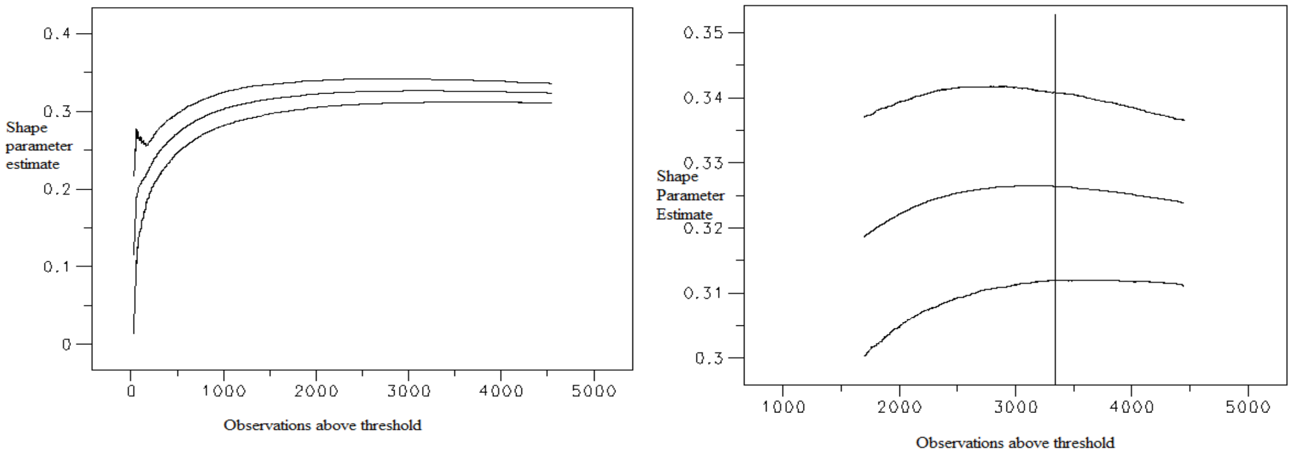


Source: CZSO, own calculations

Parameter estimates are thus $\xi = 0.3982$ and $\beta = 47\ 820$.

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 $\xi = 0.3263$ and $\beta = 32\ 839$.

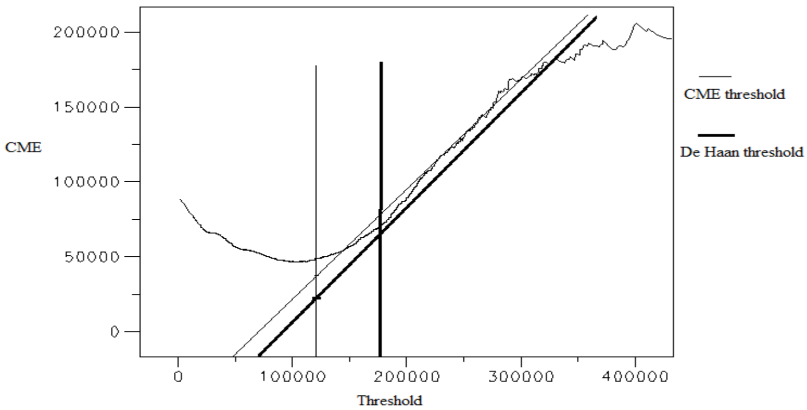
Fig. 2: Threshold determination for the year 1992 - CME



Source: CZSO, own calculations

Figure 3 shows CME vs. u plot, the second method of threshold determination described in chapter 1.5. 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.

Fig. 3: CME vs. u plot with highlighted thresholds



Source: CZSO, own calculations

Having parameters estimated obtaining high quantiles estimates is quite simple. The three quantiles to be estimated are $x_{0.95}$, $x_{0.99}$ and $x_{0.999}$ setting which means the estimated income of the 950th, 990th a 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.

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 were 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. 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.

Tab. 1: Estimated number of households in the Czech Republic

Year	1992	1996	2002	2005	2006	2007	2008	2009
No.	3 594	3 725	3 953	4 100	4 162	4 224	4 287	4 349

Source: CZSO, own calculation

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.

Tab. 2: Estimated parameters for the year 1992

		de Haan				CME			
Year	Observations in sample	Threshold	Obs. above threshold (%)	ξ	β	Threshold	Obs. above threshold (%)	ξ	β
1992	16 234	176 847	5.54	0.3982	47 820	123 504	20.33	0.3263	32 839

Source: CZSO, own calculation

Tab. 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
non-parametric	181 422	276 155	594 036	1 784 554

Source: CZSO, own calculation

4 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 quantiles with the non-parametric estimates (np). The values closest to the non-parametric estimates are highlighted. In Table 6 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.

Tab. 4: Estimated parameters

		de Haan				CME			
Year	Observations	Threshold	Obs. above threshold (%)	ξ	β	Threshold	Obs. above threshold (%)	ξ	β
1992	16 234	176 847	5.54	0.3982	47 820	123 504	20.33	0.3263	32 839
1996	28 148	349 500	4.44	0.3734	98 586	217 700	21.33	0.2952	67 301
2002	7 973	454 165	6.27	0.3406	130 450	429 751	8.15	0.3812	113 442
2005	4 351	477 542	7.47	0.3578	127 932	290 731	28.73	0.2810	107 572
2006	7 483	502 291	6.88	0.3185	136 035	556 273	4.68	0.3802	134 708
2007	9 675	384 199	19.64	0.2249	117 567	unable to obtain	19.64	0.3359	109 009
2008	11 294	416 187	19.92	0.2476	124 220	416 187	15.94	0.2662	127 547
2009	9 911	627 606	6.56	0.3762	178 326	397 007	27.24	0.2940	131 037

Source: CZSO, own calculation

Tab. 5: Estimated quantiles

	$x_{0.95}$			$x_{0.99}$			$x_{0.999}$		
	deHaan	CME	Np	de Haan	CME	np	de Haan	CME	np
1992	**181 853**	181 917	181 431	294 208	**291 780**	276 518	650 739	**592 920**	607 090
1996	xxx	**339 570**	338 100	**545 881**	552 346	525 500	**1 173 420**	1 099 973	1 180 400
2002	484 859	**490 674**	495 949	**786 894**	794 306	775 428	**1 639 174**	1 724 934	1 580 772
2005	**532 773**	533 625	531 600	**854 188**	891 437	764 665	**1 793 441**	1 786 300	1 941 640
2006	**547 994**	Xxx	547 336	864 609	**839 066**	831 641	**1 718 844**	1 730 969	1 596 005
2007	572 538	**573 519**	588 701	**882 677**	941 976	898 972	**1 575 497**	1 971 806	1 600 577
2008	**620 938**	589 424	633 321	**966 804**	938 348	965 421	1 775 485	**1 785 314**	2 164 046
2009	**678 591**	684 972	676 290	**1 115 454**	1 128 902	1 043 634	**2 440 845**	2 268 684	2 886 000

Source: CZSO, own calculation

Tab. 6: Estimated highest earning household´s income in the Czech Republic

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

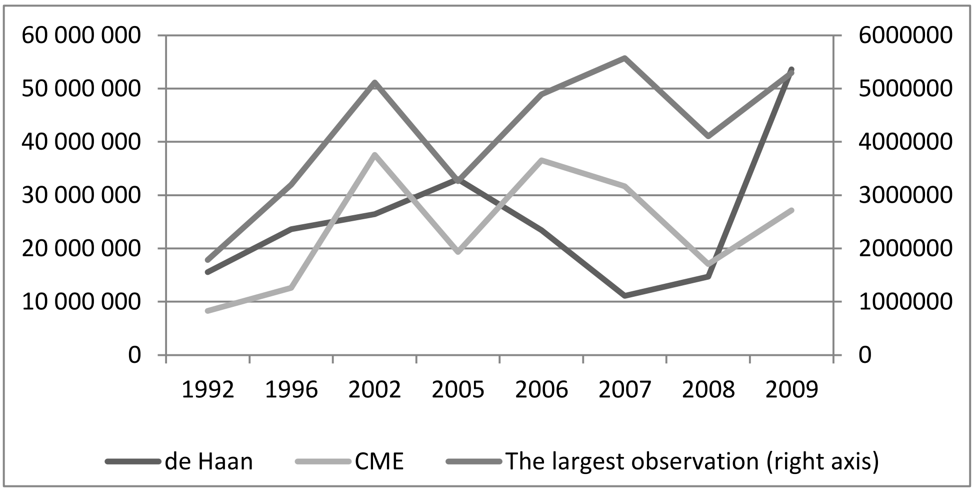
Source: CZSO, own calculation

The Figure 4 plots the estimated highest earnings obtained by the two methods and the largest value at the sample. 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.

The ratio of the rise of income between the years 1992 and 2009 is 3.73 for $x_{0.95}$, 3.79 for $x_{0.99}$ and 3.75 for $x_{0,999}$. These values come from de Haan method. The same ratio for lower quartile in the samples is 3.37, for median 3.36 and for upper quartile 3.59.

The main problem stems from the available data. 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. De Haan method seems to better fit the data especially at the highest quantiles, but if the data doesn´t cover high incomes, the whole tail is underestimated and so the CME method can produce better results for the especially highly improbable events if there is at least one large value. It is all a part of larger discussion about extreme values estimates obtained from the samples, the topic skeptically covered i.e. in Taleb (2010).

Fig. 4: Estimated highest household´s incomes in the Czech Republic



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. 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 (Times New Roman, 14 pt., bold)

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

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.

Contact

Adam Čabla

University of Economics in Prague

nám. W. Churchilla 4, Praha, Czech Republic
adam.cabla@vse.cz