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

\[
F_u(y) = P(X - u \leq y | X > u)
\]

for \( 0 \leq \omega_F - u \),

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

\[
F_u \rightarrow H_{\xi,0,\beta} = 1 - \left( 1 + \frac{x}{\beta} \right)^{-1/\xi}
\]
as \( u \to \infty \). \hspace{1cm} (2)

Parameter \( \xi \) 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 first 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 \( \lambda = k/n \) 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:

\[
M_n^{(r)} = \frac{1}{k} \sum_{i=0}^{k-1} \left( \log(X_{n-i,n}) - \log(X_{n-k,n}) \right)^r
\]

for \( r = 1, 2 \). \hspace{1cm} (3)

The estimators of \( \xi \) and \( \beta \) are then:

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

\hspace{1cm} (4)
\[ \hat{\beta} = \frac{uM_u^{(i)}}{\rho_1} \]
\[ \rho_1 = 1 \text{ for } \xi \geq 0 \text{ otherwise } \rho_1 = \frac{1}{1-\xi}. \]  

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 \( \xi \) and \( \beta \) 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 \( \xi \) 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**

![Graph showing threshold determination for de Haan method](image)

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 \( \beta = 32 839 \).
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.

Figure 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 95th, 99th and 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**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>3 594</td>
<td>3 725</td>
<td>3 953</td>
<td>4 100</td>
<td>4 162</td>
<td>4 224</td>
<td>4 287</td>
<td>4 349</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Year</th>
<th>Observations in sample</th>
<th>de Haan</th>
<th></th>
<th></th>
<th></th>
<th>CME</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs. above threshold (%)</td>
<td>Threshold</td>
<td>ξ</td>
<td>β</td>
<td>Obs. above threshold (%)</td>
<td>ξ</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>16 234</td>
<td>176 847</td>
<td>5.54</td>
<td>0.3982</td>
<td>47 820</td>
<td>123 504</td>
<td>20.33</td>
<td>0.3263</td>
</tr>
</tbody>
</table>

Source: CZSO, own calculation

**Tab. 3: Estimated quantiles for the year 1992**

<table>
<thead>
<tr>
<th>Method</th>
<th>$x_{0.95}$</th>
<th>$x_{0.99}$</th>
<th>$x_{0.999}$</th>
<th>Highest Earning</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Haan</td>
<td>181 853</td>
<td>294 208</td>
<td>650 739</td>
<td>15 530 846</td>
</tr>
<tr>
<td>CME</td>
<td>181 917</td>
<td>291 780</td>
<td>592 920</td>
<td>8 266 204</td>
</tr>
<tr>
<td>non-parametric</td>
<td>181 422</td>
<td>276 155</td>
<td>594 036</td>
<td>1 784 554</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Observations</th>
<th>Threshold</th>
<th>Obs. above threshold (%)</th>
<th>de Haan</th>
<th>CME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>16 234</td>
<td>176 847</td>
<td>5.54</td>
<td>0.3982</td>
<td>47 820</td>
</tr>
<tr>
<td>1996</td>
<td>28 148</td>
<td>349 500</td>
<td>4.44</td>
<td>0.3734</td>
<td>98 586</td>
</tr>
<tr>
<td>2002</td>
<td>7 973</td>
<td>454 165</td>
<td>6.27</td>
<td>0.3406</td>
<td>130 450</td>
</tr>
<tr>
<td>2005</td>
<td>4 351</td>
<td>477 542</td>
<td>7.47</td>
<td>0.3578</td>
<td>127 932</td>
</tr>
<tr>
<td>2006</td>
<td>7 483</td>
<td>502 291</td>
<td>6.88</td>
<td>0.3185</td>
<td>136 035</td>
</tr>
<tr>
<td>2007</td>
<td>9 675</td>
<td>384 199</td>
<td>19.64</td>
<td>0.2249</td>
<td>117 567</td>
</tr>
<tr>
<td>2008</td>
<td>11 294</td>
<td>416 187</td>
<td>19.92</td>
<td>0.2476</td>
<td>124 220</td>
</tr>
<tr>
<td>2009</td>
<td>9 911</td>
<td>627 606</td>
<td>6.56</td>
<td>0.3762</td>
<td>178 326</td>
</tr>
</tbody>
</table>

Source: CZSO, own calculation

### Tab. 5: Estimated quantiles

<table>
<thead>
<tr>
<th></th>
<th>deHaan</th>
<th>CME</th>
<th>Np</th>
<th>de Haan</th>
<th>CME</th>
<th>Np</th>
<th>CME</th>
<th>Np</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{0.05}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{0.99}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{0.999}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>181 853</td>
<td>181 971</td>
<td>181 431</td>
<td>294 208</td>
<td>291 780</td>
<td>276 518</td>
<td>650 739</td>
<td>592 920</td>
</tr>
<tr>
<td>1996</td>
<td>490 674</td>
<td>495 949</td>
<td>786 894</td>
<td>794 306</td>
<td>775 428</td>
<td>1 173 420</td>
<td>1 099 973</td>
<td>1 180 400</td>
</tr>
<tr>
<td>2002</td>
<td>532 773</td>
<td>533 625</td>
<td>531 600</td>
<td>854 188</td>
<td>891 437</td>
<td>764 665</td>
<td>1 793 441</td>
<td>1 786 300</td>
</tr>
<tr>
<td>2005</td>
<td>547 994</td>
<td>547 336</td>
<td>864 609</td>
<td>839 066</td>
<td>831 641</td>
<td>1 718 844</td>
<td>1 730 969</td>
<td>1 596 005</td>
</tr>
<tr>
<td>2007</td>
<td>572 538</td>
<td>573 519</td>
<td>588 701</td>
<td>882 677</td>
<td>941 976</td>
<td>898 972</td>
<td>1 575 497</td>
<td>1 971 806</td>
</tr>
<tr>
<td>2008</td>
<td>620 938</td>
<td>589 424</td>
<td>633 321</td>
<td>966 804</td>
<td>938 348</td>
<td>965 421</td>
<td>1 775 485</td>
<td>1 785 314</td>
</tr>
<tr>
<td>2009</td>
<td>678 591</td>
<td>684 972</td>
<td>676 290</td>
<td>1 115 454</td>
<td>1 128 902</td>
<td>1 043 634</td>
<td>2 440 845</td>
<td>2 268 684</td>
</tr>
</tbody>
</table>

Source: CZSO, own calculation

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

<table>
<thead>
<tr>
<th>Year</th>
<th>As above</th>
<th>From last 10 %</th>
<th>From last 5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de Haan</td>
<td>CME</td>
<td>de Haan</td>
</tr>
<tr>
<td>1992</td>
<td>15 530 847</td>
<td>8 266 204</td>
<td>1 784 554</td>
</tr>
<tr>
<td>1996</td>
<td>23 611 452</td>
<td>12 567 953</td>
<td>3 192 600</td>
</tr>
<tr>
<td>2002</td>
<td>26 401 844</td>
<td>37 569 312</td>
<td>5 110 628</td>
</tr>
<tr>
<td>2005</td>
<td>32 949 702</td>
<td>39 360 507</td>
<td>3 262 118</td>
</tr>
<tr>
<td>2006</td>
<td>23 439 916</td>
<td>36 548 870</td>
<td>4 891 034</td>
</tr>
<tr>
<td>2007</td>
<td>11 067 401</td>
<td>31 638 738</td>
<td>5 569 100</td>
</tr>
<tr>
<td>2008</td>
<td>14 673 804</td>
<td>17 062 571</td>
<td>4 103 711</td>
</tr>
<tr>
<td>2009</td>
<td>53 619 530</td>
<td>27 158 513</td>
<td>5 294 482</td>
</tr>
</tbody>
</table>

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

![Graph showing estimated highest household’s incomes in the Czech Republic](image)

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.
References


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.


88.

Threshold Approach". *Journal of Structural Engineering*.


Vysoká škola ekonomická v Praze.

Contact

Adam Čabla

University of Economics in Prague

nám. W. Churchilla 4, Praha, Czech Republic

adam.cabla@vse.cz