

# An Application of Extreme Value Analysis to U.S. Movie Box Office Returns

Bi, G.<sup>1</sup> and D.E. Giles<sup>1</sup>

<sup>1</sup>Department of Economics, University of Victoria, Victoria BC, Canada  
Email: dgiles@uvic.ca

*Keywords: Movie revenue, extreme values, generalized Pareto distribution, value at risk*

## EXTENDED ABSTRACT

This study uses extreme value theory (EVT) to model U.S. weekend movie box office returns. Most Hollywood movies open in theaters on a weekend, as the majority of audiences watch movies during the weekend. The weekend box office revenue therefore accounts for a major part of the total box office revenue. Weekend box office returns – *i.e.*, the percentage change in revenue from one weekend to the next - have empirical fluctuations that lend themselves naturally to being modeled by EVT. In this paper we use the Peaks over Threshold method and maximum likelihood estimation to fit the Generalized Pareto Distribution (GPD) to the tails of the distributions for both extreme positive and negative returns in box office returns. We use these results to calculate value at risk (VaR) and expected shortfall (ES) measures of return risk.

We find that we are able to model the tails of the distributions for both positive and negative returns satisfactorily with the GPD. Our estimates of VaR and ES for positive return indicate that, with probability 1%, the revenue increase from one weekend to the next could exceed 81.04%, and that when it does, the average increase is 96.32%. Also, with probability 1%, box office revenue could drop 68.72% from one weekend to the next, and that

when it does the average fall 73.21%. That is, if the first weekend's box office revenue is \$100 million, there is a one percent probability that the revenue will decrease to \$31.28 million next week and the corresponding expected revenue for all possible revenues less than \$31.28 million is \$26.79 million. These estimates are useful for film distributors in determining the number of film prints, and as a reference for potential investors in the movie industry.

## 1. INTRODUCTION

After being adjusted for the effects of seasonality, U.S. weekend box office revenue is dominated by high budget movies. According to the Internet Movie Database (IMDB), among 360 blockbusters with gross box office income of over \$100 million during their theatrical runs, 290 movies, or about 80%, had budgets above \$60 million. In most cases, the distribution of box office revenue is dominated by these high budget movies. However, this is not always the case. Some high budget movies sustain losses at the box office. Based on absolute loss on worldwide gross, for example, each of the top five money losers had budgets of over \$100 million but lost over \$90 million. When these movies were released, they dragged the weekend box office returns down. In contrast, some low budget movies are box office winners. For example, the most profitable movie based on return on investment, *The Blair Witch Project*, had a budget of only \$35,000 but worldwide gross earnings of \$248 million. Blockbusters and losers that appear in the distribution of the weekend box office can be taken as extreme events which can be analyzed *via* Extreme Value theory (EVT).

EVT has been applied in many areas where “disasters” occur, such as earthquakes, floods, and even terrorism attacks (*e.g.*, Jenkinson, 1955, Embrechts *et al.*, 1999, and Reiss and Thomas, 1997). Many studies have analyzed the variations in financial markets with EVT. The tail behavior of financial returns series has been discussed by Koedijk *et al.* (1990), Longin (1996), Danniellsson and de Vries (2000), Neiftci (2000), McNeil and Frey (2000), Gençay *et al.* (2003) and Gençay and Selçuk (2006), for example.

We use the Peaks over Threshold method and maximum likelihood estimation to fit the Generalized Pareto Distribution to the tails of the distributions for both extreme positive and negative box office returns. We also calculate value at risk and expected shortfall measures of return risk.

## 2. EXTREME VALUE THEORY

The principal result of extreme value theory relates to the asymptotic distribution of “block maxima” – *i.e.*, the maximum values of blocks, or snapshots, of data from an unknown underlying distribution. The Fisher and Tippet (1928) Theorem tells us that if these maxima are suitably normalized, they converge in distribution to one of only three forms – Gumbel, Fréchet, or Weibull. This is an extreme value analogue to conventional central limit theory. These three distributions can be encompassed by a single one – the Generalized Extreme Value (GEV) family of distributions. (Coles, 2001, pp. 45-52). If individual data values  $\{X_1, X_2, \dots\}$ , rather than blocks, are available then it is inefficient to artificially “block” them and estimate a GEV distribution. The detailed data information can be used more efficiently by modeling the distribution,  $F_u$ , of values that are “extreme” (*i.e.*, exceed some high threshold value,  $u$ ). The Conditional Excess distribution function is defined as:

$$F_u(y) = P(X - u \leq y | X > u) \\ 0 \leq y \leq x_F - u ,$$

where  $X$  is a random variable,  $u$  is a particular threshold value,  $y = x - u$  are the excesses (or “exceedances”), and  $x_F < \infty$  is the right endpoint of the unknown population distribution,  $F$ . So,

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}$$

As the realizations of the random variable  $X$  lie mainly between 0 and  $u$ , the estimation of  $F$  in this interval is usually quite straightforward. However, the estimation of the portion,  $F_u$ , which is of interest here, can be difficult due to the fact that the number of observations above the large enough threshold might be quite limited. The following asymptotic result is a natural generalization of the GEV result for block maxima:

**Theorem** (Pickands, 1975; Balkema and de Haan, 1974): For a large class of underlying distribution functions  $F$  the Conditional Excess Distribution function  $F_u(y)$ , for  $u$  large, is well approximated by

$$F_u(y) \approx G_{\xi, \sigma}(y), \quad u \rightarrow \infty,$$

where

$$G_{\xi, \sigma}(y) = \begin{cases} 1 - (1 + \frac{\xi}{\sigma} y)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-y/\sigma} & \text{if } \xi = 0 \end{cases}$$

for  $y \in [0, (x_F - u)]$  if  $\xi \geq 0$ , and  $y \in [0, -\frac{\xi}{\sigma}]$  if

$\xi < 0$ .  $G_{\xi, \sigma}$  is the so-called Generalized Pareto Distribution (GPD).

Defining  $x = u + y$ , the GPD can be written as a function of  $x$ :

$$G_{\xi, \sigma}(x) = \begin{cases} 1 - (1 + \frac{\xi}{\sigma}(x-u))^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-(x-u)/\sigma} & \text{if } \xi = 0 \end{cases}$$

where  $u$  is the threshold,  $\xi$  is the shape parameter and  $\sigma$  is the scale parameter. Maximum Likelihood Estimation (MLE) can be used to estimate the parameters of the GPD after selecting an appropriate threshold  $u$ . Then we can fit the GPD to the exceedances. The details follow.

### 3. THE PEAK OVER THRESHOLD METHOD

The Peak over Threshold (POT) method is used to obtain the distribution of exceedances above a certain threshold. The POT method involves the following steps: select the threshold  $u$ ; fit the GPD function to the exceedances over  $u$ ; compute estimates for various risk measures. The selection of the threshold  $u$  is the key factor that decides the fraction of data belonging to the tail, and therefore affects the results of the MLE of the parameters of the GPD function. The value of  $u$  should be high enough to satisfy the Theorem in section 2, but the higher the threshold the fewer observations are left for the estimation of the parameters. There is a trade-off, and the determination of the threshold is complicated. Previous research (e.g., Danielsson *et al.*, 2001; Dupuis, 1998) has attempted to deal with this issue, but there is no unambiguous method for selecting the threshold. Graphical tools are usually adopted (e.g., Gilli and K ellezi, 2006). We use two tools - the Sample Mean Excess (SME) plot and the Shape Parameter (SP) plot - to determine the threshold,  $u$ .

The (population) mean excess function of the GPD with parameter  $\xi < 1$  is

$$e(z) = E(X - z | X > z) = (\sigma + \xi z) / (1 - \xi), \quad \sigma + \xi z > 0$$

This gives the average value of the excesses of  $X$  conditional on a value for the threshold,  $z$ . The SME-plot is defined by the points:

$$(u, e_n(u)) \quad ; \quad x_1^n < n < x_n^n$$

where  $e_n(u)$  is the *sample* mean excess function defined as:

$$e_n(u) = (n - k + 1)^{-1} \sum_{i=k}^n (x_i^n - u) \quad ;$$

$$k = \min\{i \mid x_i^n > u\},$$

and  $(n-k+1)$  is the number of observations exceeding the threshold. As an estimate of the mean excess function, the sample mean excess function should be linear. This property can be used as a criterion for the selection of the threshold,  $u$ . The selected  $u$  should be that located at the beginning of a portion of the sample mean excess plot that is roughly linear and sloping up (Angelini, 2002). This involves a subjective choice in practice.

The shape parameter (SP) plot graphs the estimates of the shape parameter  $\xi$  on the vertical axis as a function of increasing thresholds  $u$  on the horizontal axis. For a sample  $y = \{y_1, \dots, y_n\}$  of exceedances, the log-likelihood function for the GPD is:

$$L = \begin{cases} -n \log \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^n \log\left(1 + \frac{\xi y_i}{\sigma}\right); & \xi \neq 0 \\ -n \log \sigma - (1/\sigma) \sum_{i=1}^n y_i & ; \xi = 0 \end{cases}$$

We can compute the MLE's of the parameters for one sample of exceedances,  $y$ , defined by the observations exceeding a single threshold  $u$ . After generating a series of thresholds and repeating the process of computing estimates on the basis of equation (16), a series of estimates for  $\xi$  and  $\sigma$  can be computed. Estimates of  $\xi$  are plotted against the associated thresholds to find a range over which the estimates are relatively stable.

After selecting a threshold  $u$ , using the above tools the corresponding MLE's of the parameters are used with the sample of exceedances,  $y$ , to construct a series of values for  $G_{\hat{\xi}, \hat{\sigma}}(y)$ . Plotting both the theoretical and the empirical distribution function, we can observe if the GPD provides a reasonable fit to

the exceedances above the threshold.

#### 4. RISK MEASURES

Two typical risk measures are the Value at Risk (VaR) and the Expected Shortfall (ES). Value at Risk is the return sufficient to cover, in most instances, gains or losses over a fixed number of weekends. Suppose a random variable  $X$ , with continuous distribution function  $F$ , models positive or negative returns over a certain time horizon. The  $VaR_p$  is the  $p$ -th quantile of the distribution  $F$  such that  $VaR_p = F^{-1}(1-p)$ , where  $F^{-1}$  is the  $p$ <sup>th</sup> quantile function. The expected shortfall is defined as the expected size of a return that exceeds  $VaR_p$ :

$$ES_p = E(X \mid X > VaR_p)$$

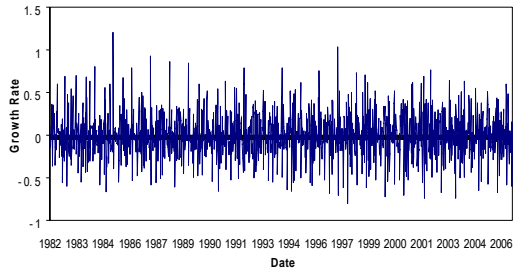
Assuming a GPD function for the tail distribution, the  $VaR_p$  and  $ES_p$  can be expressed in terms of the GPD parameters:

$$\hat{VaR}_p = u + \frac{\hat{\sigma}}{\hat{\xi}} \left( \left( \binom{n}{N_u}^p - 1 \right)^{-\hat{\xi}} - 1 \right)$$

$$\hat{ES}_p = \hat{VaR}_p + e(z) = \frac{\hat{VaR}_p}{1 - \hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi}u}{1 - \hat{\xi}}$$

#### 5. DATA

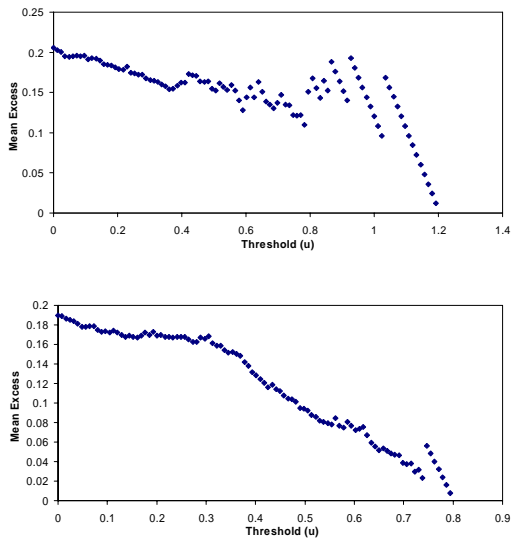
Weekend box office total revenues of the top 12 movies per week, from 8 January 1982 to 15 September 2006, have been obtained from the online movie database *Box Office Mojo* ([www.boxofficemojo.com](http://www.boxofficemojo.com)). Weekly returns in revenue are calculated as logarithmic differences. As the distributions of the positive and negative are asymmetric, we model them separately (as is usually the case with financial returns). Our sample comprises 620 positive returns and 668 negative returns.



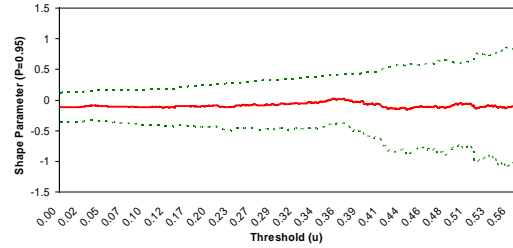
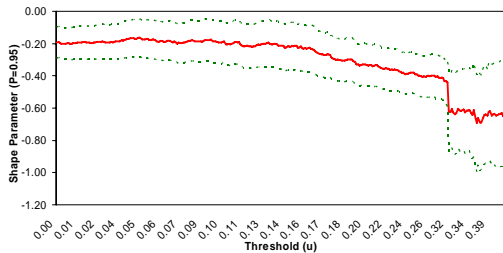
**Figure 1.** U.S. weekend box office returns.

## 6. RESULTS

Our results were obtained by writing program code for EViews 5.1. The R package, PoT, was used to verify the MLE results. Figure 2 shows the SME plots for the positive and negative returns. Values of  $u = 0.35$  ( $0.28$ ) for the positive (negative) returns locate the beginning of a portion of the SME plot that is approximately linear and sloping up.



**Figure 2.** Top (Bottom): ME plot for positive (negative) returns.

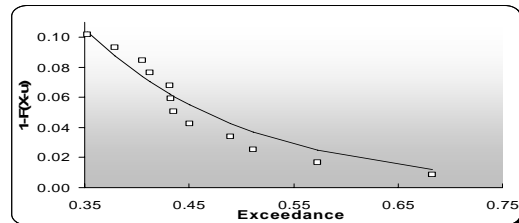
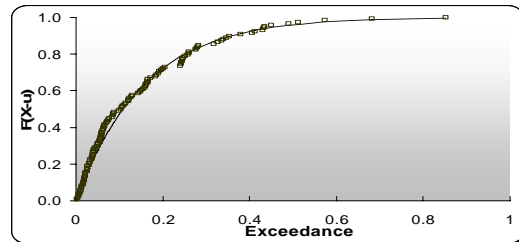


**Figure 3.** Top (Bottom): Shape parameter estimates for positive (negative) returns as a function of the threshold.

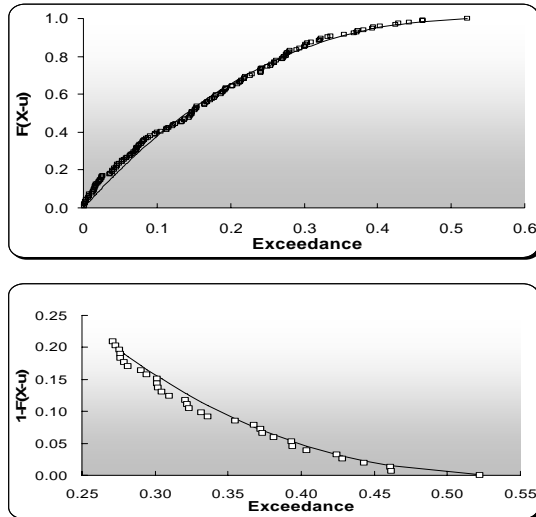
Figure 3 provides the SP plots with 95% confidence intervals. For the positive returns we choose  $u = 0.3529$  (118 exceedances). For the negative returns we choose  $u = 0.2801$  (153 exceedances). These thresholds closely match those suggested by the SME plots. Table 1 gives the MLE results.

**Table 1.** Parameter estimation of GPD.

	Positive returns	Negative returns
$u$	$=0.3529$	$=0.2801$
$\hat{\xi}$ (Std. Error)	-0.0066 (0.1136)	-0.4138 (0.0690)
$\hat{\sigma}$ (Std. Error)	1.5679 (0.2075)	2.3193 (0.2285)



**Figure 4.** Top: Positive returns. GPD fitted to 118 exceedances above  $u = 0.3529$ . Bottom: GPD fitted to the tail exceedances.



**Figure 5.** Top: Negative returns. GPD fitted to 153 exceedances above  $u = 0.2801$ . Bottom: GPD fitted to the tail exceedances.

A series of values for  $G_{\xi, \sigma}(y)$  can then be computed by substituting these estimates into the GPD function. Figures 4 and 5 show the GPD fits, which are relatively satisfactory. Table 2 gives the estimates of VaR and ES for  $p = 0.01, 0.10$ , with 95% asymptotic confidence intervals (from the Delta method).

**Table 2.** Estimated values of VaR and ES.

Positive Returns		
Probability	VaR	ES
= 1%	0.811	0.963
Confidence Interval	0.320~1.300	0.222~1.705
= 10%	0.454	0.609
Confidence Interval	0.320~0.587	0.028~1.189
Negative Returns		
Probability	VaR	ES
= 1%	0.687	0.732
Confidence Interval	0.584~0.791	0.245~1.220
= 10%	0.443	0.559
Confidence Interval	0.373~0.513	0.1534~0.965

The estimates of VaR and ES for positive return indicate that, with 1% probability, the returns from one weekend to the next could exceed 81.04%, and that the average returns above this level will be 96.32%. So, if a weekend's box office revenue is \$100 million, there is one percent probability that revenue will increase to \$181.04 million next weekend, and the expected value for all revenues over \$181.04 million is \$196.32 million, *etc.* Such estimates can be used in different ways. For example, the VaR results imply that, given the same amount of investment the possibility of loss for an investment in the movie industry is relatively lower than the possibility of gain. In addition, the difference between the VaR and ES for the positive returns is bigger than that for the negative returns. This means that the expected gain over the VaR under the situation of gain is more than the expected loss over the VaR under the situation of loss.

The risk measures could also help movie producers forecast the required number of prints of new movies. By calculating VaR and ES with a given probability and estimating the total box office revenue at the releasing weekend, the producer could then divide the estimated revenue by the average ticket price to get the total audience, which could be used to estimate the number of prints of the new movie. This would be most useful for high budget movie producers as these movies usually dominate the box office upon release.

## 7. CONCLUSIONS

We illustrate how extreme value theory can be used to model the tails of the distribution for weekend box office returns in the U.S.. One implication of our estimates of Value at Risk and Expected Shortfall is that the possibility of loss for an investment in the movie industry is

lower than the possibility of gain. These measures can also be used to estimate the number of prints of movies that are likely to be needed, thus avoiding surpluses or shortages of film copies for high budget losers and low budget winners at the box office. Ongoing research considers returns for individual companies in the industry, and returns based on net, rather than gross, revenue.

## 8. REFERENCES

- Balkema, A.A. and L. de Haan (1974), Residual life time at great age, *Annals of Probability*, 2, 792-804.
- Coles, S. (2001), An Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London.
- Danielsson, J. and C.G. de Vries (2000), Value-at-risk and extreme returns, *Annales d'Economie et de Statistique*, 60, 239-270.
- Danielsson, J., C.G. de Vries, L. de Haan and L. Peng (2001), Using a boot-strap method to choose the sample fraction in tail index estimation, *Journal of Multivariate Analysis*, 76, 226-248.
- Dupuis, D. J. (1998), Exceedances over high thresholds: a guide to threshold selection, *Extremes*, 1, 251-261.
- Embrechts, P., C. Kläupelberg and T. Mikosch (1999), Modelling extremal events for insurance and finance, Applications of Mathematics, Springer, Berlin.
- Fisher, R.A. and L.H.C. Tippett (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, 24, 180-190.
- Gençay, R. and F. Selçuk (2006), Overnight borrowing, interest rates and extreme value theory, *European Economic Review*, 50, 547-563.
- Gençay, R., F. Selçuk, and A. Ulugülyağcı (2003), High volatility, thick tails and extreme value theory in value-at-risk estimation. *Insurance: Mathematics and Economics*, 33, 337-356.
- Gilli, M. and E. Këllezzi (2006), An application of extreme value theory for measuring financial risk, *Computational Economics*, 27, 1-23.
- Jenkinson, A.F. (1955), The frequency distribution of the annual maximum (minimum) values of meteorological events, *Quarterly Journal of the Royal Meteorological Society*, 81, 158-172.
- Jondeau, E. and M. Rockinger (1999), The tail behaviour of stock returns: emerging versus mature markets, mimeo., HEC and Banque de France.
- Koedijk, K.G., M. Schafgans and C.G. de Vries (1990), The tail index of exchange rate returns. *Journal of International Economics*, 29, 93-108.
- Longin, F.M. (1996), The asymptotic distribution of extreme stock market returns, *Journal of Business*, 69, 383-408.
- McNeil, A.J. and R. Frey (2000), Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach, *Journal of Empirical Finance*, 7, 271-300.
- Neiftci, S.N. (2000), Value at risk calculations, extreme events, and tail estimation, *Journal of Derivatives*, 1, 23-37.
- Pickands, J.I. (1975), Statistical inference using extreme value order statistics, *Annals of Statistics*, 3, 119-131.
- Reiss, R. D. and Thomas, M. (1997). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser Verlag, Basel.