Range Estimation of Stock Market Index Using Extreme Value Approach

by

Naveen Pentakota.

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 5

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interest of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED:___________________

APPROVAL BY THESIS DIRECTOR

This Thesis has been approved on the date below

____ ___________________

Satheesh V. Aradhyula Date

Associate Professor of Agricultural and Resource Economics

ACKNOWLEDGEMENTS

I would like to thank Dr. Satheesh Aradhyula, my thesis advisor, for his timely help whenever it was needed. I thank Dr. Gary Thompson and Dr. Alan Ker for their valuable input as my other committee members.

I would like to thank the other faculty and departmental staff, especially Nancy Smith, for their co-operation all throughout my graduate program.

TABLE OF CONTENTS

List of Figures

List of Tables

Abstract

The heart of the financial/stock markets is the ability to predict future prices. The range i.e. the difference between the high and the low of the stock price or index in a day, could be a key to the amount of exposure and the possible gains from investment. So an extreme value approach for the range prediction is adopted to investigate the problem. The daily log ratio, i.e. log of the ratio of the high/low on one day to that of next day, is chosen as the variable of interest.

Traditional approaches would consider the midrange, defined as the average of the daily high and low, as the variable of interest. The predictions from this traditional model do not portray possible extreme gains or losses available from range estimation. The results thus provide some unique insight into risk assessment and mitigation, as the approach is a breakaway from the tradition.

CHAPTER 1

Introduction

The analysis of tail behavior of asset returns with a thorough understanding of large movements in asset prices is important for financial risk management. Explicit forms of the tails of the distribution provide important information to risk managers and investors.

Empirical studies have established that the distribution of speculative asset returns tend to have heavier tails than the Gaussian distribution tails (Mandelbrot, 1963; Pagan, 1996). Furthermore, very often these thick tailed distributions are found to have asymmetric tails. Such stylized features of financial returns provide useful insight into the economics of financial markets and calls for appropriate methodologies of modeling such behavior.

Conditional heteroskedasticity models of Engle (1982) and Bollerslev (1986) and their various modifications do incorporate some of these stylized features which occur due to phenomenon such as volatility clustering in financial data. Although the conditional heteroskedasticity models can explain part of the non-Gaussian features of the unconditional distribution, it is often found that features like heavy tails may persist even after accounting for conditional heteroskedasticity.

The present research is focused on the estimation of the range of prices in financial markets. The daily prices of the stock indices like the S&P 500 or the DJIA are often expressed in the range format. Display of information in this format offers information about the price changes in a day, a reflection of the price movements for that day.

Range is defined as:

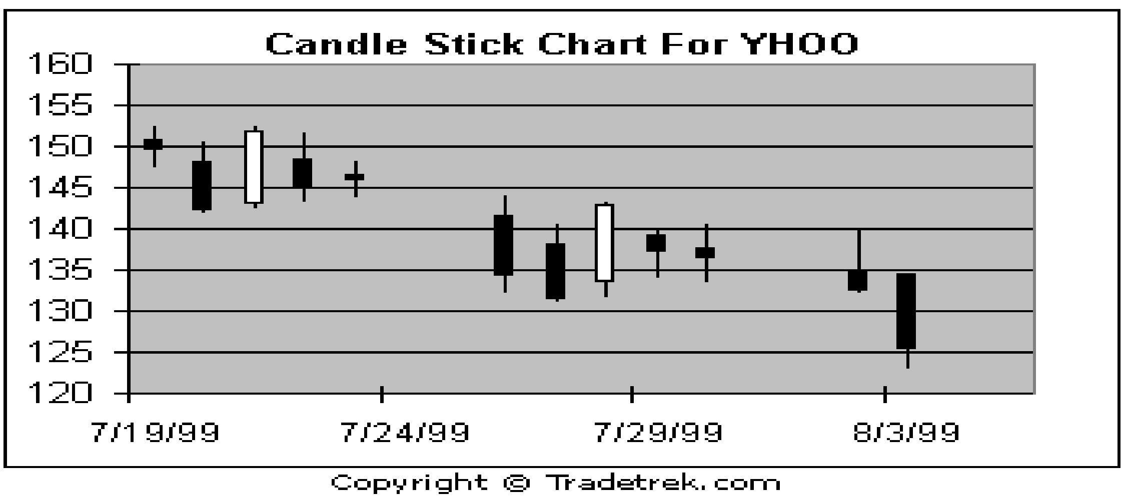
Daily range = (daily-high – daily-low).

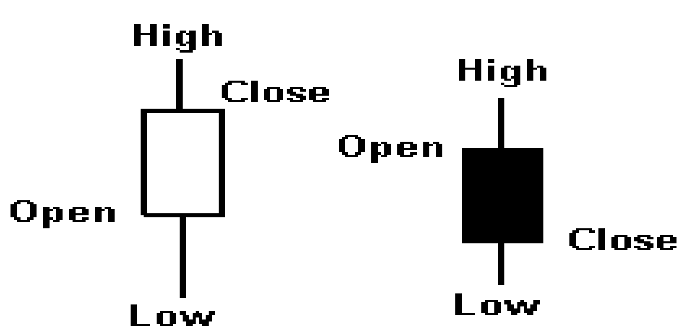
Typically, the industry estimates the midrange defined as:

Mid-Range = (daily-high + daily-low)/2

This midrange estimate with its confidence interval constitutes a major part of the risk management strategy of investors. Figure 1.1 below displays typical price range information.

Figure 1.1. A Candle Stick representation of stock prices





Another potential application of range prediction is the futures commodity markets. The daily highest and lowest are averaged out to arrive at the mean daily price. This mean daily price is used in future analysis of the markets or for further predictions for the future. There seems to be loss of information if the midrange alone is used for predicting future prices. Alternative perspectives on utilizing the available information could help investors make better and more informed decisions. As an alternative, we suggest predicting the range of prices. The essence of the present exercise has been to mould the available data into more informative terms. The range prediction of prices would provide us with information, which would help one in estimating risk better than the mean price prediction method.

The work done thus far in the field of mean prediction has employed the assumption of a normal distribution for the underlying distribution. This we perceive as a shortcoming for the following reasons. Firstly, if the underlying distribution is normal, the midrange has a very high variance for predictions purposes, making it an inefficient estimator. Secondly, the normal assumption is often violated empirically. So the midrange would typically not be a normal as it is usually assumed. To relax the assumptions of normality we propose to estimate the high and low separately, using the extreme value theory (EVT) and a generalized extreme value model is fitted to the data. This allows the data to define the parameters and nature of the tails of the original distribution would decide on the tail index. The EVT imposes fewer distributional assumptions than previously developed models.

CHAPTER 2

Literature Review

Extreme value theory has been around for some time, from the pioneering work on block maxima[1] of Fischer and Tippett (1928) and Gnedenko (1943) to the exposé by Gumbel (1958). More recently, Balkema and de Haan (1974) and Pickands (1975) have presented results for threshold-based extreme value methods. Applications of the theory have since appeared in hydrology and wind engineering. More recently extreme value theory has been applied to finance and insurance problems. Analysis of these applications can be found in Embrechts, Klueppelberg, Mikosch (1997), and in Reiss and Thomas (2001).

The demand for practical statistical tools to analyze the extreme events within time series has led to many recent advances in the theory and methods for extreme values. Historically, the demand came from environmental topics where extreme events are the only aspects of the series of practical concern (Natural Environment Research Council, 1975; Simiu and Scanlan, 1986; Dixon and Tawn, 1997). Other applications are emerging, with financial time series increasingly being analyzed to assess the risk from extreme events and to determine the capital that is required to control this risk (Longin, 1996; Embrechts et al., 1997). The standard approach for describing the extreme events

[1]Two approaches to extreme value theory are block maxima and threshold methods. The theory behind the block maxima method can be found in work by various authors (for instance in Embrecths et al., 1997; McNeil, 1998; and Kellezi and Gilli, 2000 and is discussed in detail in the next chapter. As an alternative to looking at blocks and block maxima one can collect the returns in a series that exceed a certain high threshold, say u, and model these returns separately from the rest of the distribution. This is the peaks over threshold (POT) method.

of a stationary time series is to focus on its exceedances of a fixed high threshold level, which leads to a description of extreme events that contains four components:
the probability of exceeding the threshold, the distribution of excesses of the threshold, the long-range dependence between extreme values and the local dependence within extreme events. The first two of these components are determined by the marginal distribution of the time series and the last two by its dependence structure. The marginal features of time series extremes are well understood from the study of independent and identically distributed (IID) random variables, and flexible statistical methods are available (Pickands, 1971, 1975; Davison and Smith, 1990; Smith, 1989) for their analysis. Though the theory pertains to exceedances over a threshold, the insight can be appropriate in the block maxima (i.e. maxima of a sample) methods too. The dependence structure of the maxima had been addressed in the Leadbetter and Leadbetter et.al (1983) in the theorem on the Asymptotic Independence of Maxima.

In studies of today's volatile financial markets, it is common to make a distinction between conditional and unconditional asset return distributions. While the unconditional distribution is interesting when long-term investment decisions and the occurrence of very rare (stress) events are of interest, the conditional distribution is more appropriate when day-to-day risks and short-term risk management are considered. McNeil (1998) calculates conditional VaR (Value at Risk)-measures by filtering return series with a GARCH model and then apply threshold based EVT tools to model the IID residuals. We follow a similar approach for the daily maximum and minimum of tick data of DJIA. The variations in our methodology are in adopting an AR model to filter the data, instead of

GARCH, and application of this method to the block maxima as compared to the threshold approach.

CHAPTER3

Extreme Value Theory

3.1. Extreme value distributions: Block Maxima Approach

Let $(X_{m,n}, m, n \in \mathbb{Z})$ be random variables representing tick data. We begin with the assumption that $X_{1,n},\ldots,X_{m,n}$ are iid. It is our effort to find a limit distribution for maxima $M_n = \max(X_{1,n},\ldots,X_{m,n})$. It is transformed such that the limit distribution of the new variable is a non-degenerate one. Following Fisher and Tippett's (1928) theorem, the variate, $\mathrm{M_n}$, is reduced with a *location* parameter, μ_n, and a *scale* parameter, σ_n, in such a way that $\lim_{n\to\infty} \Pr\{M_\mathrm{n}^* \le z\} = \lim_{n\to\infty} \mathrm{F^n}(\sigma_n \mathrm{z} + \mu_n) = G(z)$,

where $\mathrm{M_n}^* = (\mathrm{M_n} - \mu_n)/\sigma_n$. Assuming the existence of a sequence of such coefficients, we have three types of non-degenerated distributions for the standardized maximum, $\mathrm{M_n}^*$ (represented by the random variable z), as Gumbel (I), Frechet (II) and Weibull (III).

$$\text{I.}\quad G(z) = \exp\{-\exp[-(\frac{z-\mu}{\sigma})]\},\quad -\infty < z < \infty; \tag{3.1}$$

$$\text{II.}\quad G(z) = \begin{cases} 0, & z \le \mu, \\ \exp\{-(\frac{z-\mu}{\sigma})^{-\alpha}\}, & z > \mu; \end{cases} \tag{3.2}$$

III. $G(z) = \begin{cases} \exp\{-[-(\frac{z-\mu}{\sigma})^{\alpha}]\}, & z < \mu, \\ 1, & z \geq \mu; \end{cases}$ (3.3)

for parameters $\sigma > 0$, μ and, in the case of families II and III, $\alpha > 0$.

von Mises (1954) and Jenkinson (1955) independently proposes a generalized extreme value (GEV) distribution, which includes the three limit distributions distinguished above:

$$G(z) = \exp\{-[1+\xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\}, \{z : 1+\xi(\frac{z-\mu}{\sigma}) > 0\}, \quad (3.4)$$

where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

where ξ is a shape parameter. For $\xi > 0$, $\xi < 0$ and $\xi = 0$ we obtain the Frechet, Weibull and Gumbel families, respectively. The Frechet distribution is fat tailed as its tail is slowly decreasing; the Weibull distribution has no tail—after a certain point there are no extremes; the Gumbel distribution is thin-tailed as its tail is rapidly decreasing. The following figures show the different curves for varying values of the shape parameter. In the graphs below, the horizontal axis represents the random variable while the vertical axis represents the pdf value.

Figure 3.1-A Gumbel Distribution Curve with Mu=0, Sigma=1, Zeta=0

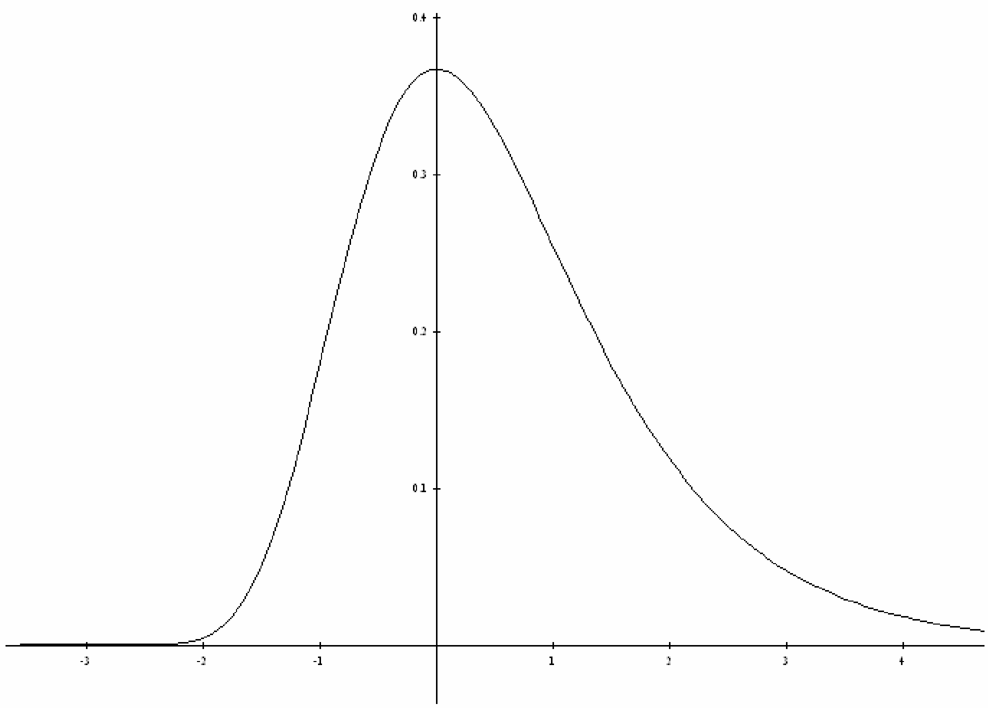


Figure 3.2-A Fretchet Distribution Curve with Mu=0, Sigma=1, Zeta=1

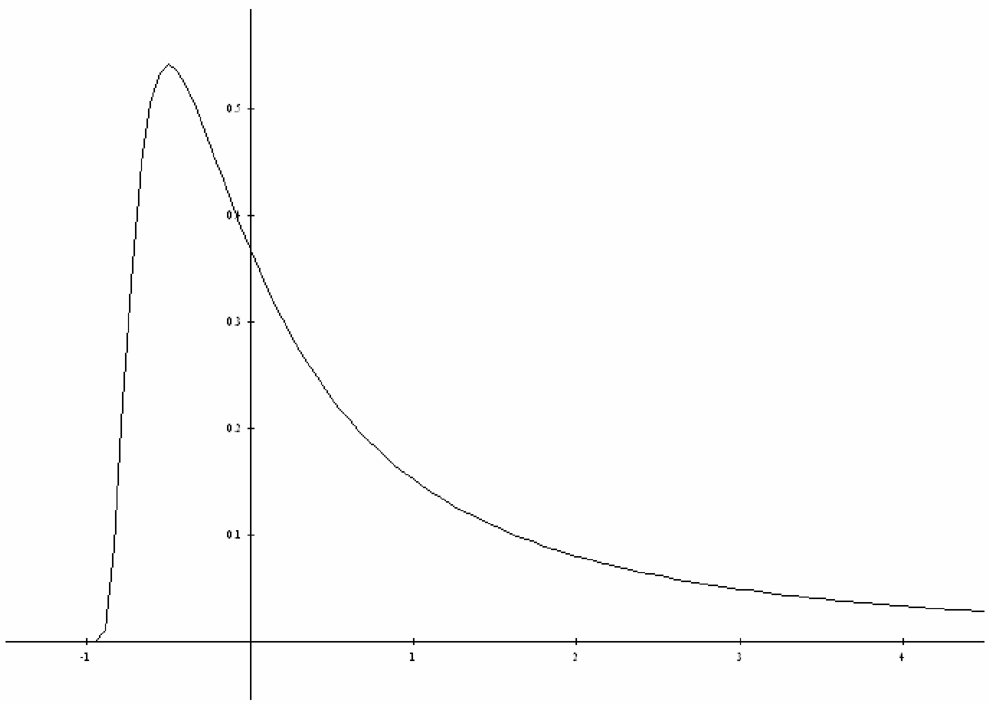
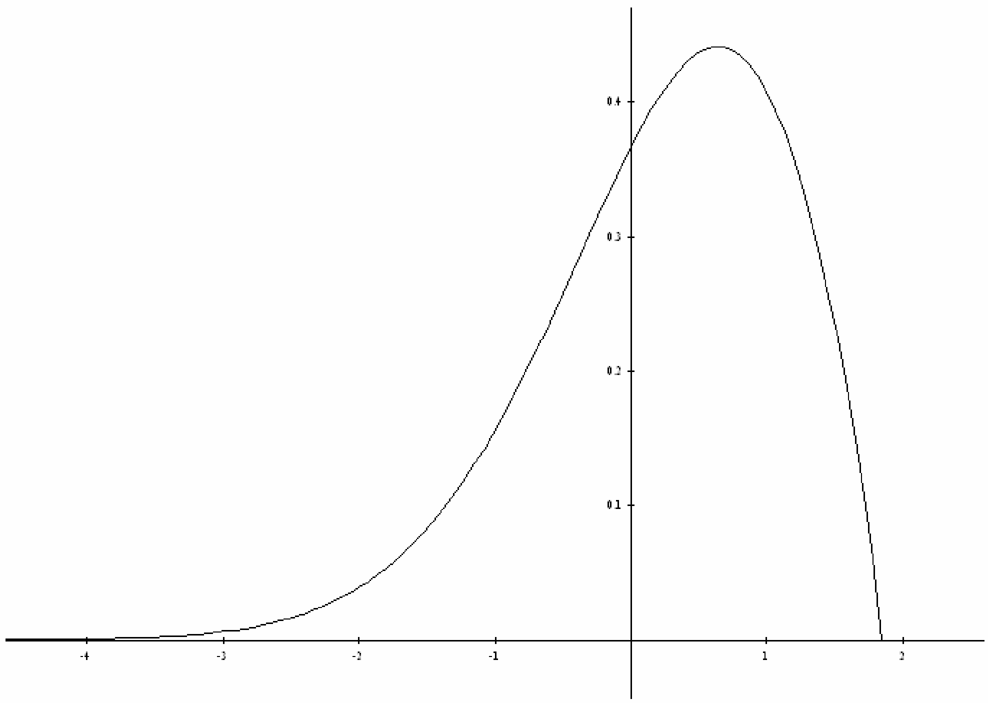


Figure 3.3-A Weibull Distribution Curve with Mu=0, Sigma=1, Zeta=-0.5



The discussion thus far in this chapter has highlighted the extreme value theory and the types of extreme distributions. This needs to be applied to the data. In the next chapter the data description and the application of the knowledge from this chapter to the data is included.

CHAPTER4
Data Source and Methodology

4.1. Data Description

Dow Jones Industrial Average (DJIA) is the source of information for the present study. The sample consists of the daily index maximum and minimum for a period of about 7years, from April 1st 1993 to October 31st 2000. The sample size is 1908 observations of which 1884 were used in the model development, and 22 observations from the month of October 2000 were set aside for out sample validation. The source of this data is www.tickdata.com. According to the EVT, these data would be distributed as Gumbel/Frechet, depending on the underlying distributional assumption. These maxima and minima being over daily volumes of about 1000 or more ensures that they are large samples. As the underlying samples need to be large for EVT to be applicable. A normal distribution assumption would lead to a Gumbel limiting distribution for the maximum and minimum. The literature, however, indicate returns are usually fat tailed, suggesting the Frechet distribution would be appropriate.

The graphs 4.1.1 and 4.1.2. represent the daily high and low respectively of the DJIA. The increasing trend can be seen in the graphs.

Figure 4.1.1: Distribution of DJIA high from April 1st 1993 to September 29th 2000

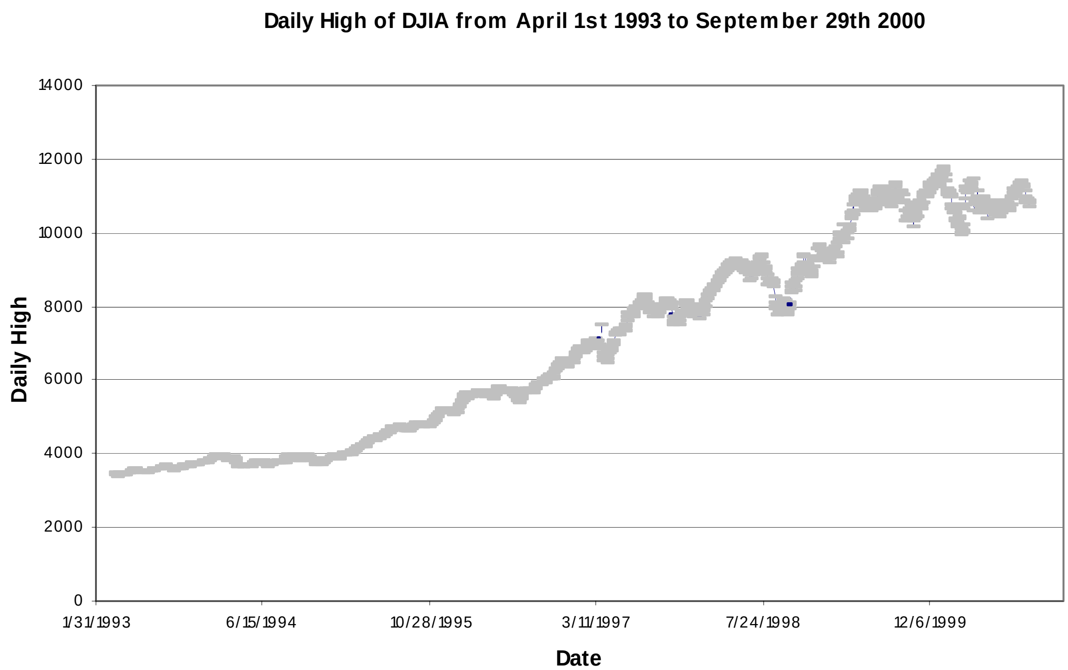
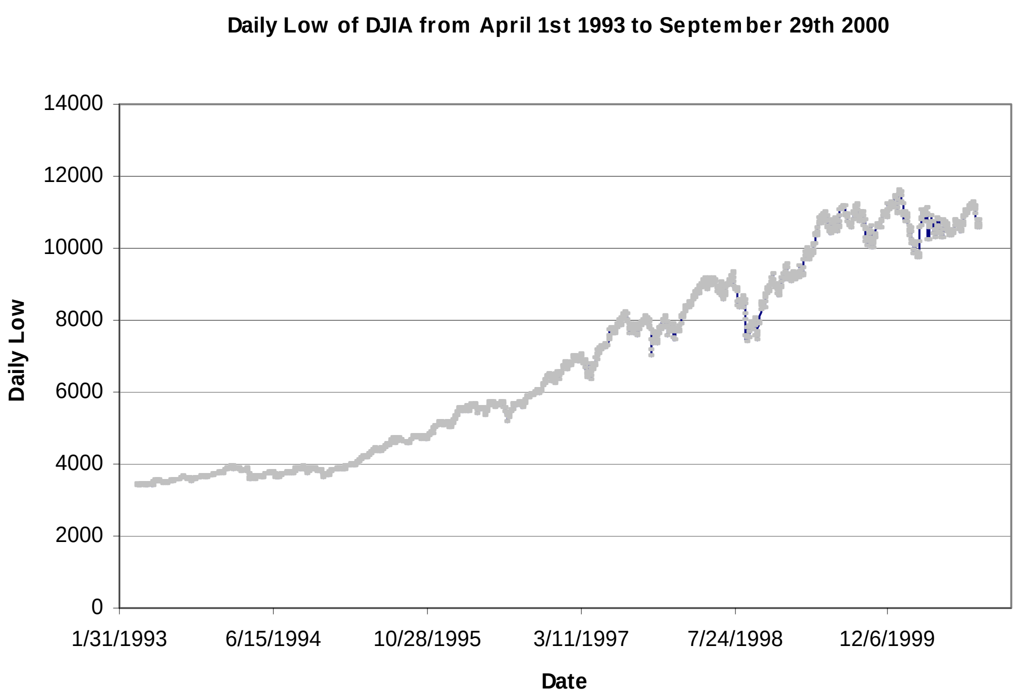


Figure 4.1.2: Distribution of DJIA low from April 1st 1993 to September 29th 2000



4.2. Model Specification:

The dynamics of model specification, very often, involves analysis of the data and certain assumptions. There are numerous avenues for exploration. The various combinations of the extreme value model parameters as functions of time or other covariates, is one of them. The underlying principle is parsimony and simplicity. For example, when modeling we would have to ascertain that the data is devoid of trend-linear or quadratic, or an exponential relation. After analyzing the DJIA sample for the present exercise, we concluded that an AR (1) process would the simplest alternative.

The model for the prediction of the daily log ratio y_t is specified as described below:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t - 0.577\sigma \tag{4.2.1a}$$

for the maxima, and

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t + 0.577\sigma \tag{4.2.1b}$$

for the minima;

where, y_t are the daily log returns of the DJIA, ε is the error term and α (intercept) and β (slope) are the two out of the three parameters to be estimated. The third parameter is the scale parameter σ (refer equations 4.2.3 and 4.2.4). The factor of 0.577 is used as the mean of a Gumbel distribution is $\mu + 0.577\sigma$, where μ is the location parameter of the Gumbel. This adjustment eliminates the bias in the estimation for the Gumbel estimation. The corresponding correction for the GEV is $\mu - \frac{\sigma}{\xi}(1 - \Gamma(1-\xi))$.

Two possible scenarios arise depending on the assumption that can be made about the distribution of ε.

Case 1: $\varepsilon \sim$ Gumbel(μ, σ)

Case 2: $\varepsilon \sim$ GEV(μ, σ, ξ).

So, the assumption for the Gumbel estimation is:

$E(y_t - (\alpha + \beta y_{t-1}) + 0.577\sigma) = 0 \cdot$

The density functions for the Gumbel and GEV are:

$$f(\varepsilon) = \frac{1}{\sigma}.\exp(-\frac{\varepsilon}{\sigma}).\exp(-\exp[-\frac{\varepsilon}{\sigma}]),\ \varepsilon \in (-\infty, \infty). \tag{4.2.2}$$

$$f(\varepsilon) = \frac{1}{\sigma}.[1+\xi(\frac{\varepsilon}{\sigma})]^{-(1+1/\xi)}.\exp(-[1+\xi(\frac{\varepsilon}{\sigma})]^{-1/\xi}),\ \varepsilon \in (-\infty, \infty). \tag{4.2.3}$$

These density equations are applicable for the maxima. The corresponding equations for the minima are obtained by replacing ε with $-\varepsilon$ in the above equations. Thus, the likelihood equations for the minima are:

$$f(\varepsilon) = \frac{1}{\sigma}.\exp(\frac{\varepsilon}{\sigma}).\exp(-\exp[\frac{\varepsilon}{\sigma}]),\ \varepsilon \in (-\infty, \infty). \tag{4.2.4}$$

$$f(\varepsilon) = \frac{1}{\sigma}.[1-\xi(\frac{\varepsilon}{\sigma})]^{-(1+1/\xi)}.\exp([1-\xi(\frac{\varepsilon}{\sigma})]^{-1/\xi}),\ \varepsilon \in (-\infty, \infty). \tag{4.2.5}$$

As would be discussed in the next chapter, Gumbel is a special case of GEV, where the shape index, ξ, is equal to zero. When the underlying distribution is normal, the maxima and minima converge to a Gumbel. Case 1, with the Gumbel error distribution, is a simpler approach to the problem in hand. The issues with the estimation of case 2, error with GEV distribution, is more tedious taking into account the estimation trouble with the shape parameter of the GEV. There is extensive literature highlighting the above observation. (Morrison and Smith, 2002). This difficulty in estimation of the shape parameter has caused us to abandon our efforts to estimate case 2.

4.3. Parameter Estimation by the Maximum Likelihood Method:

4.3.1. Inference for the GEV Distribution-General Considerations:

The GEV provides a model for the distribution of block maxima. Its application comprises partitioning the data into blocks of equal length, and fitting the GEV to the set of block maxima, e.g.: maximum rainfall from a set of yearly rainfall data points. But in implementing this model for any particular dataset, the choice of block size can be critical. The choice means a trade-off between bias and variance: blocks that are too small mean that approximation by the limit model is likely to be poor, leading to bias in estimation and extrapolation; large blocks afford too few block maxima.

We now simplify notation by denoting the block maxima $Z_1, \ldots, Z_t$. These are assumed to be independent variables from a GEV distribution whose parameters are to be estimated. If the X_i are independent, then the Z_i are also independent. However, independence of the Z_i is likely to be a reasonable approximation even if the X_i constitute a dependent series (discussed in detail in section 4.4).

Many techniques have been proposed for parameter estimation in extreme value models. These include graphical techniques based on versions of probability plots; moment-based techniques in which functions of model moments are equated with their empirical counterparts; procedures in which the parameters are estimated as specified functions of order statistics; and likelihood based methods. Each technique has its pros and cons, but the flexibility and adaptability to complex model-building of likelihood-based techniques makes maximum likelihood estimation (MLE) particularly attractive.

A potential difficulty with the use of likelihood methods for the GEV concerns the regularity conditions necessary for the MLE to be valid. Such conditions are not satisfied by the GEV models because the end points of the GEV distribution are functions of the parameter values: $\mu - \sigma / \xi$ is an upper end point of the distribution when ξ<0, and a lower end-point when ξ>0. This violation of the usual regularity conditions means that the standard asymptotic likelihood results are not automatically applicable. Smith(1985) studied this problem in detail and obtained the following results:

- when ξ>-0.5, MLE yields regular estimates, in the sense of having usual asymptotic properties;
- when –1<ξ<-0.5, MLE yields results, but they might not have the standard asymptotic properties;
- when ξ<-1, MLE results are not likely to be obtained.

The case $\xi \leq$-0.5 corresponds to distributions with a very short bounded upper tail. This situation is rarely encountered in applications of extreme value modeling, so the theoretical limitations of the MLE are usually no obstacle in practice.

4.3.2.Maximum Likelihood Estimation:

Under the assumption that $Z_1,\ldots,Z_t$ are independent variables having the GEV distribution, the log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$l(\mu,\sigma,\xi) = -t\log\sigma - (1+1/\xi).\sum_{i=1}^{t}\log[1+\xi(\frac{z_i-\mu}{\sigma})] - \sum_{i=1}^{t}[1+\xi(\frac{z_i-\mu}{\sigma})]^{-1/\xi}$$

provided that $1+\xi(\frac{z_i-\mu}{\sigma}) > 0$, for i=1,………,t. (4.3.1)

At parameter combinations for which equation (3.8) is violated, corresponding to a configuration for which at least one of the observed data falls beyond an end-point of the distribution, the likelihood is zero and the log-likelihood equals -∞.

The case ξ=0 requires separate treatment using the Gumbel limit of the GEV distribution. This leads to the log-likelihood

$$l(\mu,\sigma) = -t\log\sigma - \sum_{i=1}^{t}(\frac{z_i-\mu}{\sigma}) - \sum_{i=1}^{t}\exp\{-(\frac{z_i-\mu}{\sigma})\} \tag{4.3.2}$$

Subject to the limitations on ξ discussed in the previous section (Eq. 4.3.1), the approximate distribution of $(\hat{\mu},\hat{\sigma},\hat{\xi})$ is multivariate normal with mean (μ,σ,ξ) and variance-covariance matrix equal to the inverse of the observed information matrix evaluated at the ML estimate. Though this matrix can be evaluated analytically, it is easier to use numerical differencing techniques to evaluate the second derivatives, and standard numerical routines to carry out the inversion. Confidence intervals and other forms of inference follow immediately from the approximate normality of the estimator.

4.4. Dependence in Data:

The approach for this problem so far has been one based on the IID assumption of the data. This assumption is likely not met in the real world scenario. Hence, basing one's inference on such analysis could be erroneous. The IID assumption is relaxed and replaced it with a more acceptable assumption of stationarity. This replacement is not far-fetched and can be realized in the real world scenario. To achieve stationarity, the log ratio values were considered, instead of raw returns. The figures below present the approach. Figures 4.1.1 and 4.1.2 are those of the data. It can be seen that there has been an increase in the value of the DJIA. Return ratios assisted in transforming it into the stationary type as can be seen in figures 4.4.1 and 4.4.2.

The log ratio in the daily high case is defined as:

$y_t = \frac{\log(high_t)}{\log(high_{t-1})}$, while that in the daily low case is defined as:

$y_t = \frac{\log(low_{t-1})}{\log(low_t)}$.

Figure 4.4.1: Daily High Log Ratios of DJIA from April 1st 1993 to September 29th 2000.

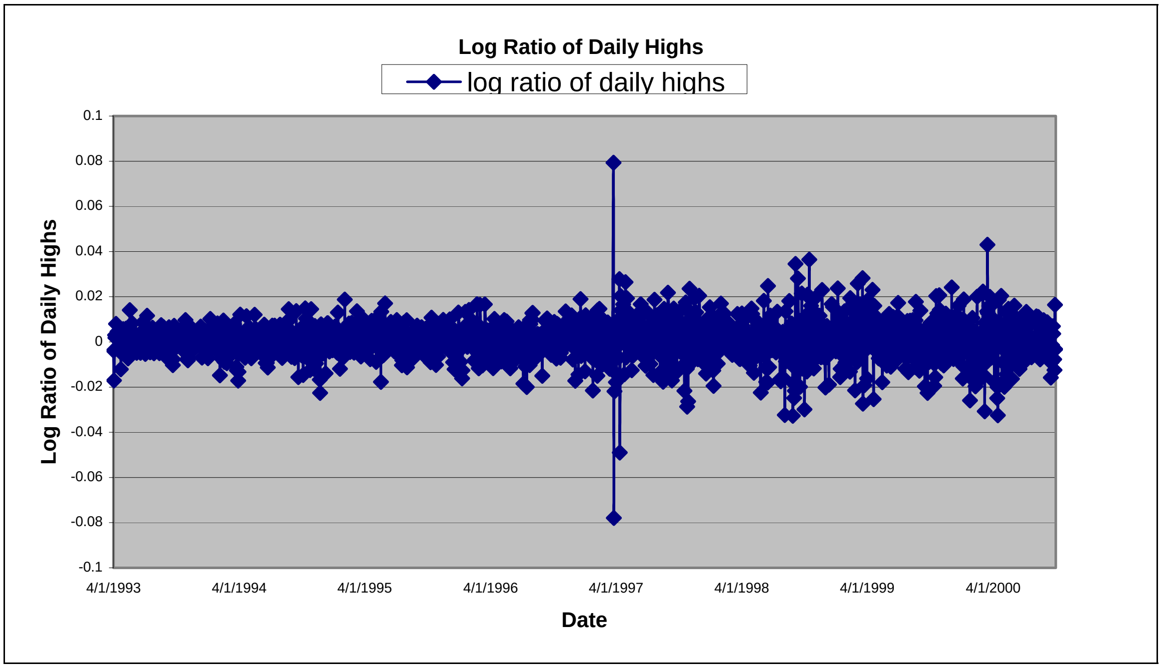
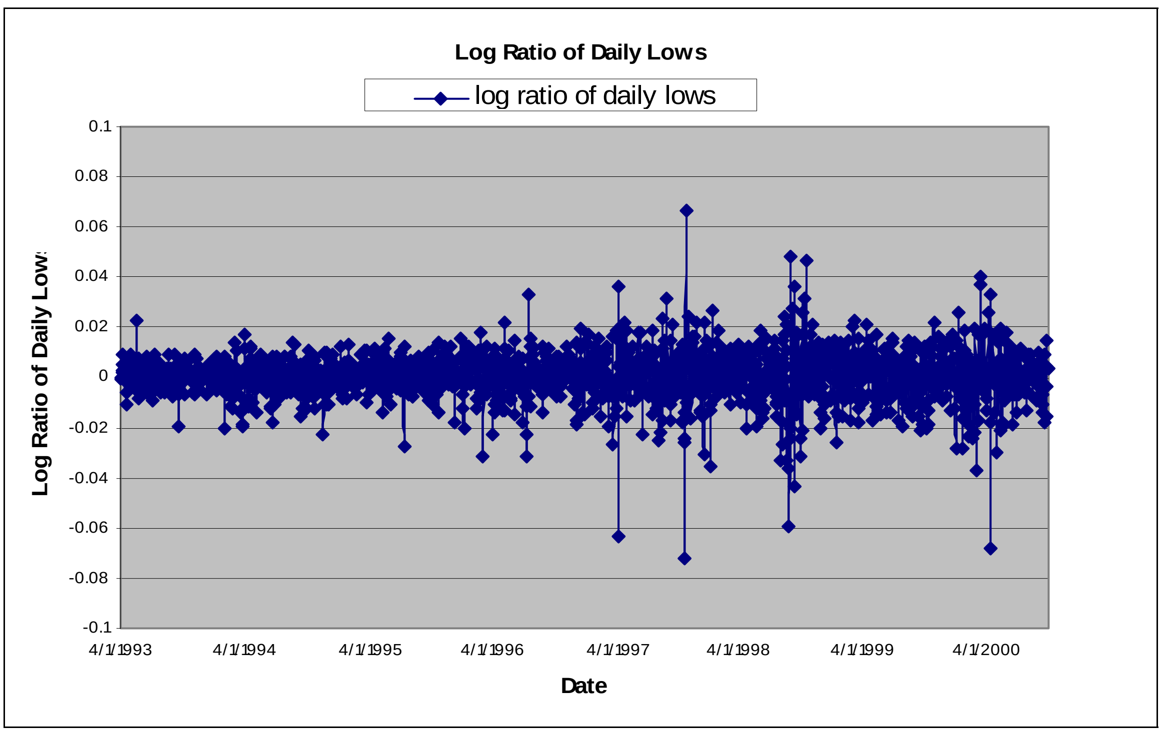


Figure 4.4.2: Daily Low Log Ratios of DJIA from April 1st 1993 to September 29th 2000.



Now the question would be how far are we justified in employing the EVT which was developed on the IID assumption. Leadbetter and Leadbetter et.al (1983) have proposed the following theorem called the "Theorem on the Asymptotic Independence of Maxima".

Theorem:

Asymptotic Independence of Maxima(AIM)

Let $M_{i,j} = \max(X_i, \ldots\ldots X_j)$ and $u_n = a_n x + b_n$ for some a_n and b_n, and x any real number.

The AIM(u_n) condition is that there exists a sequence q_n of positive integers with $q_n = O(n)$ such that for all i and j ,

$$\max |\Pr(M_{1,i} \le u_n; M_{i+q_n, i+q_n+j} \le u_n) - \Pr(M_{1,i} \le u_n) \Pr(M_{1,j} \le u_n)| \to 0$$
as $n \to \infty$.

This condition ensures that separated groups of extreme point become increasingly close to being independent as their separation and level both increase at appropriate rates.

CHAPTER 5
Estimation Results

5.1. Summary of Results:

The daily high and low from the tick data from DJIA (04/01/1993- 10/29/2000) is summarized in table 5.1.1.

Table 5.1.1: Summary statistics of DJIA Data

Variable	Sample Period	Sample Size	Sample Minima	Sample Maxima	Sample Average
DJIA Daily Low	04/01/1993- 10/29/2000	1884	3363.79	11614.40	6801.87
DJIA Daily High	04/01/1993- 10/29/2000	1884	3381.35	11749.70	6894.12

The raw data for daily highs and lows were transformed by taking the daily log returns. An AR(1) model with Gumbel error distribution was estimated, both for both the high and the low. Prior to estimation, daily log returns were divided by a factor of 1000 to magnify the parameter estimates. The following models are estimated for daily highs and lows, respectively:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t - 0.577\sigma$$

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t + 0.577\sigma$$

where y_t is daily log returns of the DJIA daily high (or low); α, β and σ are parameters to be estimated; and ε is an error term with Gumbel distribution.

Maximum likelihood estimates of parameters for daily maximum and minimum are presented in Tables 5.1.2 and 5.1.3, respectively. Results in Table 5.1.2 indicate that all three parameters for the daily high model are significant at 95% level of confidence.

For the daily minimum model in Table 5.1.3, α, β and σ are significant at 95% level of confidence.

Table 5.1.2: Maximum likelihood estimates for the DJIA daily high model

Coefficient		**Estimate**	**Standard Error**	**t-stat**
α		1.4257	0.303	4.70
β		0.3120	0.014	21.81
σ		8.2178	0.112	73.31
Summary Statistics:				
Sample Size:	1884			
Log-Likelihood:	-6857.20			
R-Square:	0.9996			
RMSE(%)	0.102			

Note: R-Square is calculated as squared correlation between actual and predicted y. RMSE is Root Mean Square Error.

Table 5.1.3: Maximum likelihood estimates for the DJIA daily low model

Coefficient		**Estimate**	**Standard Error**	**t-stat**
α		9.4912	0.254	37.36
β		0.1899	0.008	23.37
σ		9.9388	0.127	78.25
Summary Statistics:				
Sample Size:	1884			
Log-Likelihood:	-7164.15			
R-Square:	0.9994			
RMSE(%)	0.117			

Note: R-Square is calculated as squared correlation between actual and predicted y. RMSE is Root Mean Square Error.

For both the models, β is estimated to be positive indicating that current and lag returns are positively related. This positive correlation is consistent with fact that DJIA was on the increasing over the estimation period.

R-squares for both models are high indicating that the extreme value models do a good job of estimating daily highs and lows. Further validation of the model can be examined by examining how well the model tracks the actual data. One such goodness of fit measure is the root mean square error (RMSE). RMSE is defined by:

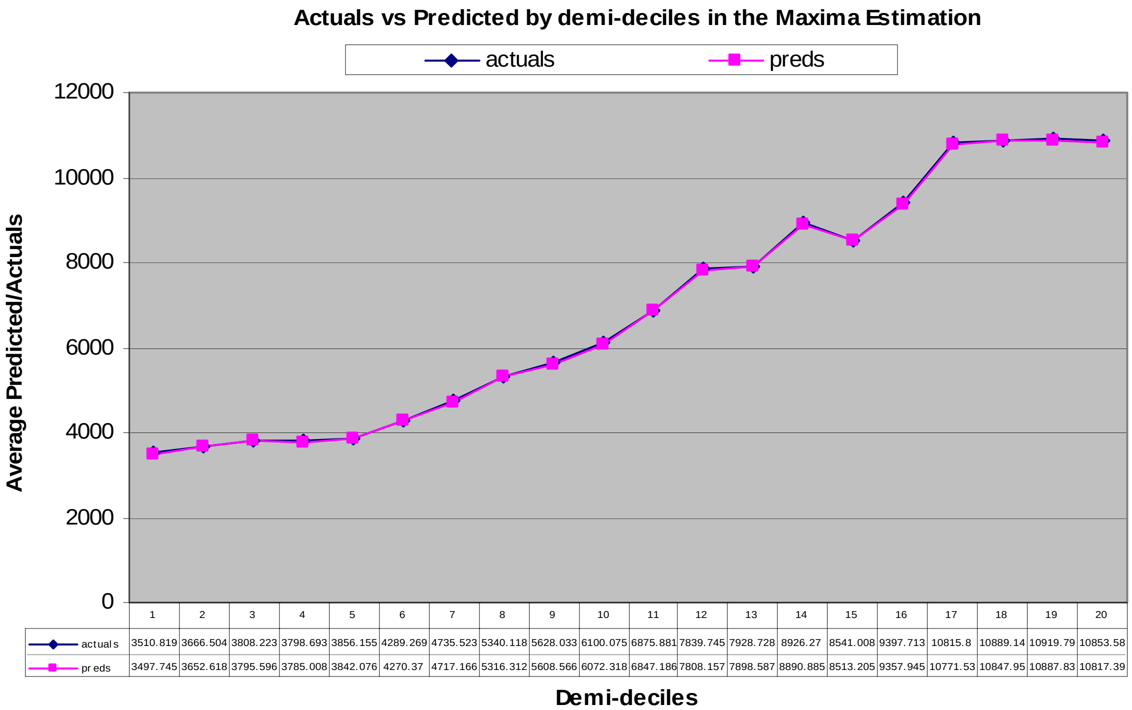
$$\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \hat{y}_t)^2}{y_t^{\,2}}}\ ,$$

where, y_t is the actual value of the log of daily maxima (or the minima) , $\hat{y}_t$ is the estimated log of daily maxima (or minima) and n is the sample size. RMSEs for models are presented in Tables 5.1.2 and 5.1.3. The RMSE for the daily maxima model is 0.102%, while that for the minima model is 0.117%. These RMSE values are very low and indicate that both models do an excellent job of tracking the actual values.

5.2. In-Sample Predictions:

The data are split into demi-deciles(20 groups) based on actual y and the means for y_t and $\hat{y}_t$ across these segments were plotted to view the model performance. Figures 5.2.1 and 5.2.2 present these in-sample predictions for daily maxima and minima models by demi-deciles. Figures 5.2.1 and 5.2.2 indicate that for both models, the predicted values follow the actual data very closely over all the do-deciles. Though there is a consistent under-prediction across all the segments, indicating a bias.

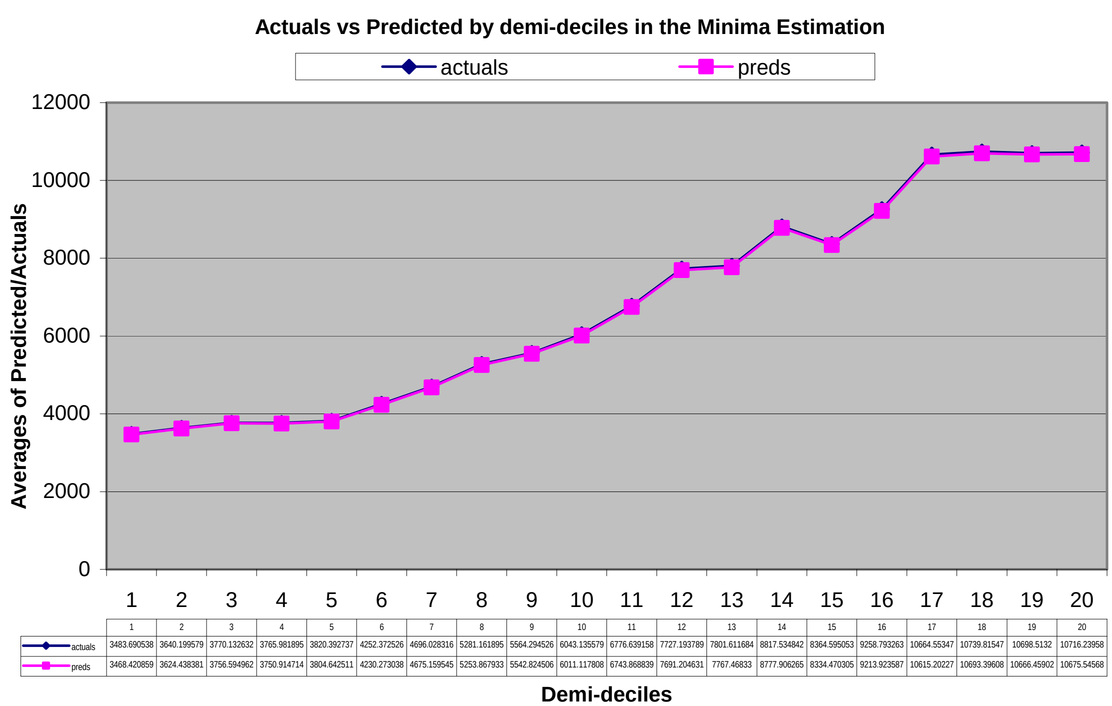
Figure 5.2.1: A plot of the actuals and predicted daily maxima



	1	2	3	4	5	6	7	8	9	10
actuals	3510.819	3666.504	3808.223	3798.693	3856.155	4289.269	4735.523	5340.118	5628.033	6100.075
preds	3497.745	3652.618	3795.596	3785.008	3842.076	4270.37	4717.166	5316.312	5608.566	6072.318

	11	12	13	14	15	16	17	18	19	20
actuals	6875.881	7839.745	7928.728	8926.27	8541.008	9397.713	10815.8	10889.14	10919.79	10853.58
preds	6847.186	7808.157	7898.587	8890.885	8513.205	9357.945	10771.53	10847.95	10887.83	10817.39

Figure 5.2.2: A plot of the actuals and predicted daily minima

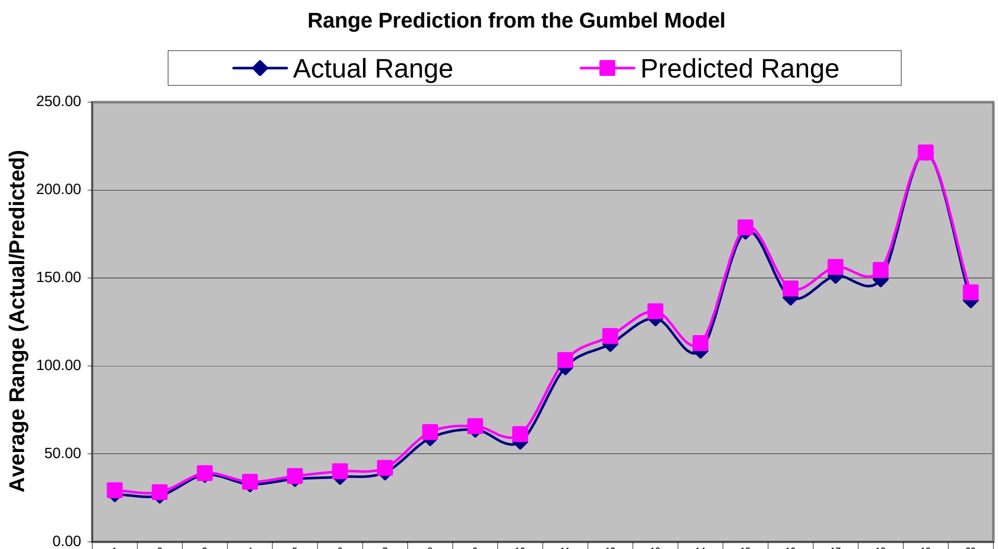


	1	2	3	4	5	6	7	8	9	10
actuals	3483.690538	3640.199579	3770.132632	3765.981895	3820.392737	4252.372526	4696.028316	5281.161895	5564.294526	6043.135579
preds	3468.420859	3624.438381	3756.594962	3750.914714	3804.642511	4230.273038	4675.159545	5253.867933	5542.824506	6011.117808

	11	12	13	14	15	16	17	18	19	20
actuals	6776.639158	7727.193789	7801.611684	8817.534842	8364.595053	9258.793263	10664.55347	10739.81547	10698.5132	10716.23958
preds	6743.868839	7691.204631	7767.46833	8777.906265	8334.470305	9213.923587	10615.20227	10693.39608	10666.45902	10675.54568

One of the objectives of the thesis is to predict daily ranges. While both models may be performing well individually, it is important to assess how the models, taken together, predict that daily range. Figure 5.2.3 combines the predictions from both models into predictions of the daily ranges, where predicted daily range is obtained as the difference between predicted daily high from the daily high model and predicted daily low from the daily low model. As with individual models, the data are sorted and presented by demi-deciles. Predictions in Figure 5.2.3 indicate that the model does an excellent job of tracking daily ranges. Of course, unlike previous studies, the present models can predict both the magnitude and the location of daily ranges. Figure 5.2.3 illustrated only the magnitudes of daily ranges.

Figure 5.2.3: A plot of the actuals and predicted of daily range.

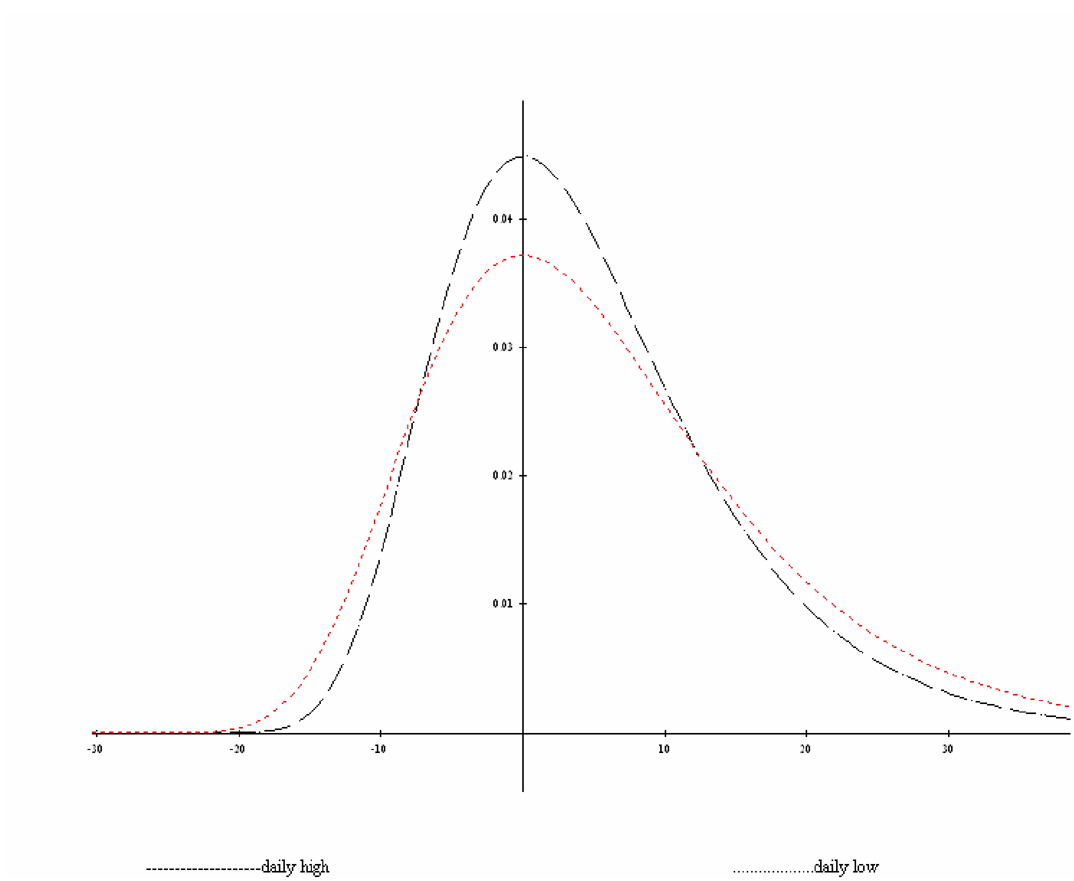


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Actual Range	27.13	26.30	38.09	32.71	35.76	36.90	39.50	58.96	63.74	56.94	99.24	112.55	127.12	108.74	176.41	138.92	151.25	149.33	221.28	137.34
Predicted Range	29.32	28.18	39.00	34.09	37.43	40.10	42.01	62.44	65.74	61.20	103.32	116.95	131.12	112.98	178.74	144.02	156.32	154.55	221.38	141.85

Demi-Deciles

The parameters from the estimation are used to plot the following distribution curves in figure 5.2.4. They are location parameter has been set to a zero in these curves.

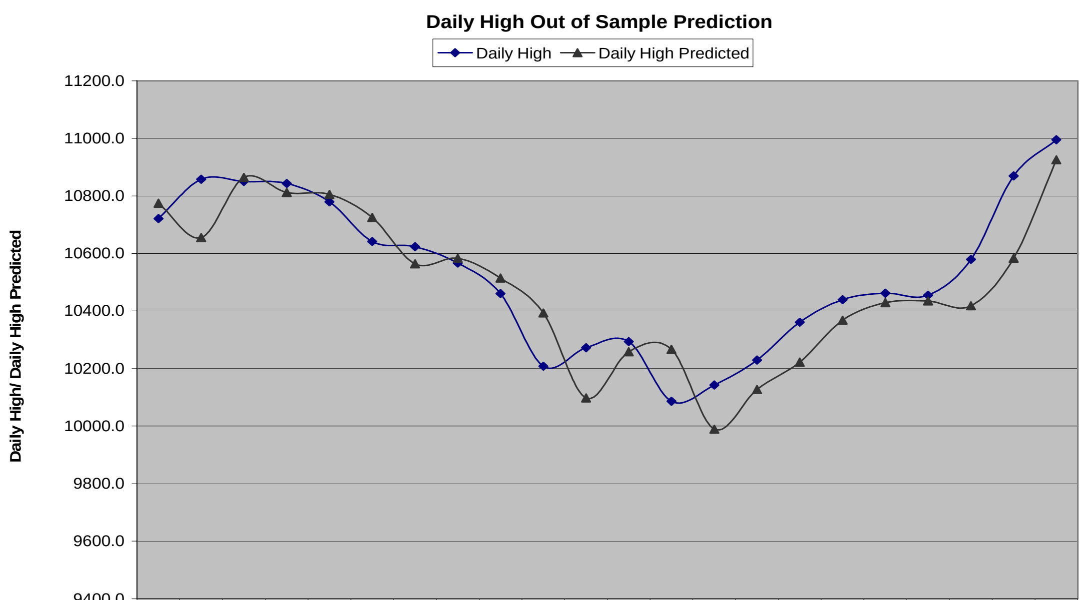
Figure 5.2.4: Gumbel distribution with the estimated parameters for the daily high ($\hat{\mu}$ =0 and $\hat{\sigma}$ =8.2) & daily hlow($\hat{\mu}$ =0 and $\hat{\sigma}$ =9.9)



5.3. Out of Sample Prediction:

Out of Sample validation was done on 22 observations from the month of October 2000. The predicted values are closer in the daily high case as compared to the daily low case. The variation in the daily low is higher than in the daily low case. This variation is not being captured by the AR model executed so far. So an alternative approach of an ARCH or a GARCH model could perform more satisfactorily. While figures 5.3.1 and 5.3.2 show the out of sample validation on the daily high and daily low respectively, figure 5.3.3 is a candle stick representation of the range for the out-sample data. In figure 5.3.3 the dotted boxes are the predicted ranges(daily high predicted -daily low predicted) and the solid lines are the actual ranges(daily high-daily low) .

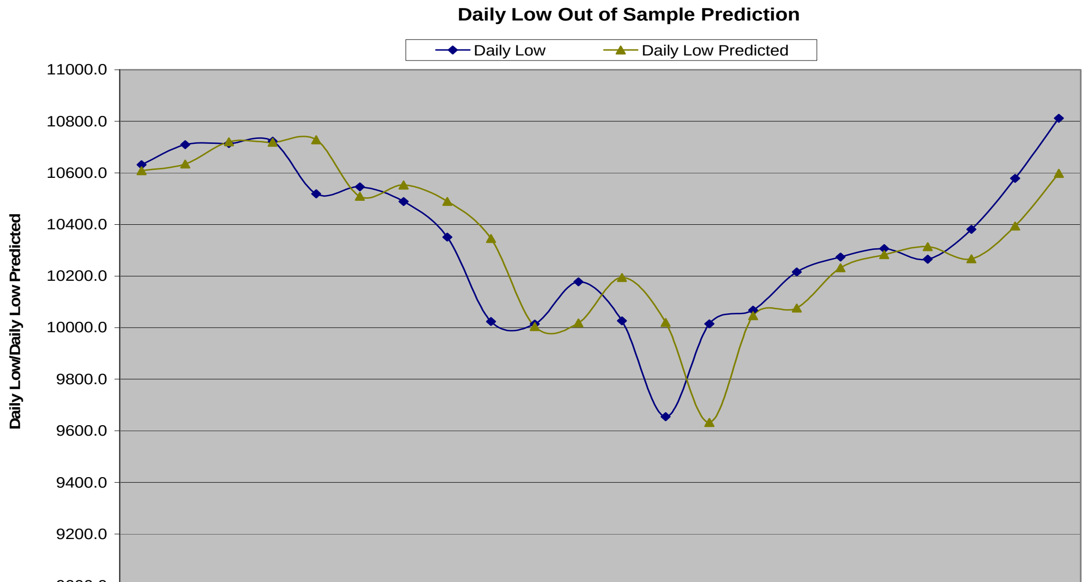
Figure 5.3.1: A plot of the Out of Sample Validation (actuals vs predicted) for the daily highs



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Daily High	10721.2	10857.7	10850.0	10842.9	10779.8	10641.3	10623.5	10566.6	10460.4	10208.1	10272.1	10293.9	10086.0	10143.0	10229.5	10361.3	10439.3	10462.0	10454.8	10579.2	10869.5	10995.4
Daily High Predicted	10774.5	10655.2	10864.4	10811.8	10804.9	10724.8	10563.8	10583.0	10514.2	10393.3	10097.7	10258.1	10266.7	9989.3	10127.2	10222.6	10368.2	10429.1	10434.5	10418.1	10583.0	10924.9

OBS(1=10/01/2000 to 22=10/31/2000)

Figure 5.3.2: A plot of the Out of Sample Validation (actuals vs predicted) for the daily lows

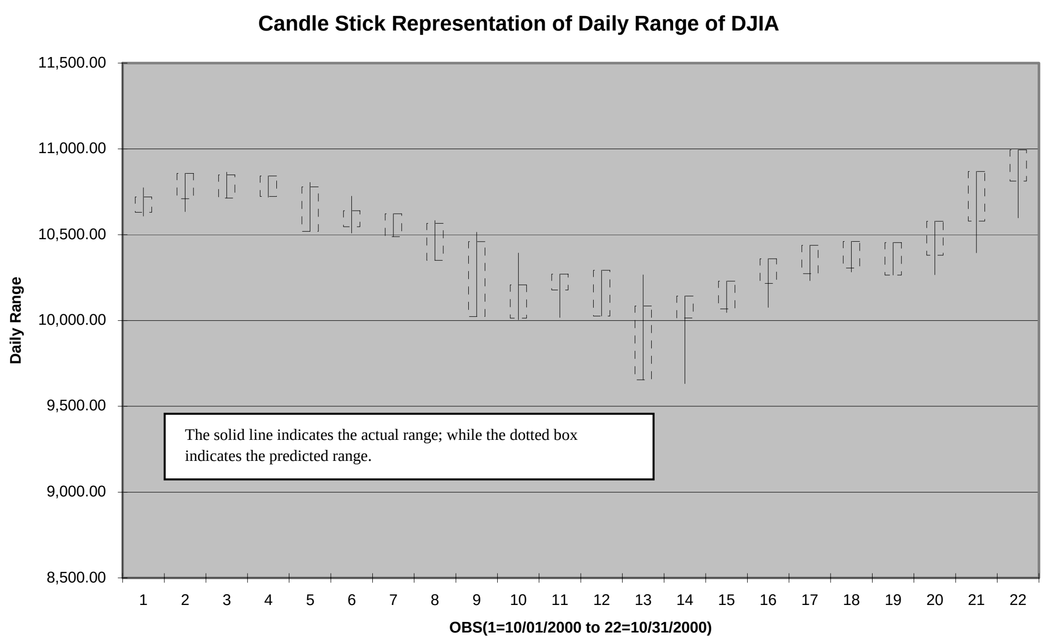


	1	2	3	4	5	6	7	8	9	10	11
Daily Low	10631.3	10709.8	10713.6	10722.7	10518.8	10546.2	10488.9	10350.9	10023.5	10014.2	10177.8
Daily Low Predicted	10608.9	10634.2	10720.5	10718.8	10728.2	10508.7	10553.0	10489.5	10345.5	10004.4	10018.1

	12	13	14	15	16	17	18	19	20	21	22
Daily Low	10026.5	9654.6	10014.6	10067.5	10216.2	10273.6	10306.6	10265.1	10380.9	10578.5	10812.3
Daily Low Predicted	10194.6	10020.0	9632.2	10046.2	10076.0	10231.9	10282.5	10313.7	10266.7	10394.2	10598.0

OBS(1=10/01/2000 to 22=10/31/2000)

Figure 5.3.3: A Candle Stick Representation of the daily range (10/01/2000-10/31/2000)



CHAPTER 6
Conclusions and Future Developments

The results, presented in the previous chapter, show that the prediction of range via Extreme Value Theory can be a practical approach. The prediction accuracy has been consistent over the entire data. Although, the results thus far have been satisfactory, there are still possible avenues for further exploration and improvement. A few of them are discussed below.

The bias in estimation needs to be accounted for which would mean an additional adjustment to be made the predicted values. The normal assumption of the log returns can be challenged in favor of other fat tailed distributions. This scenario would lead to the GEV model where the data defines the tail index, as our initial trials have been. The answer to successfully implement this approach would be in investigating estimation methodology concerned with the tail index of the GEV distribution. In this study Gumbel models for maxima and minima are estimated separately. A joint estimation of these two models can lead to more efficient parameter estimates.

Another possibility for improvement would be with the exploration of advanced model specifications like ARCH or GARCH, which would improve the applicability of the model to situations with volatile markets or varying variance scenarios. This would enhance the sensitivity of the model to small changes.

Apart from the stock market, this approach can be employed in commodity pricing and meteorological application. In commodity price modeling, midrange based estimation is commonly employed. The present approach would lead to more informative predictions, where the high and the low of the commodity can be predicted.

REFERENCES

Ancona-Navarrete, M. and Tawn, J. A. “A Comparison of Methods for Estimating the Extremal Index.” *Extremes* 3 (2000):5–38.

Balkema, A. and de Haan, L. “Residual Life Time at Great Age.” *Annals of Probability* 2 (1974):792-804.

Beirlant, J., Vynckier, P. and Teugels, J. L. “Tail Index Estimation, Pareto Quantile Plots and Regression Diagnostics.” *Journal of American Statistical Association* 91 (1996):1659–1667.

Berman, S. M. “Limit Theorems for the Maximum Term in Stationary Sequences.” *Annals of Mathematical Statistics* 35 (1964): 502–516.

Bingham, N. H., Goldie, C. M. and Teugels, J. L. *Regular Variation*. Cambridge: Cambridge University Press, 1987.

Bollerslev, T. “Generalized Autoregressive Conditional Heteroskedasticity.” *Journal of Econometrics* 31 (1986): 307-327.

Bortot, P. and Tawn, J. A. “Models for the Extremes of Markov Chains.” *Biometrika* 85 (1998): 851–867.

Coles, S.G. *An Introduction to Statistical Modeling of Extreme Values.* London:Springer Verlag. 2001.

Coles, S. G. and Tawn, J. A. “Modelling Extreme Multivariate Events.” *Journal of Royal Statistical Society B* 53 (1991): 377–392.

Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G. “Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation.” *Journal of Multivariate Analysis* 76 (2001):226–248.

Davis, R. A. and Resnick, S. I. “Basic Properties and Prediction of Max-ARMA Processes.” *Advanced Applied Probability* 21 (1989): 781–803.

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press,1997.

Davison, A. C. and Smith, R. L. “Models for Exceedances Over High Thresholds (with discussion).” *Journal of Royal Statistical Society B* 52 (1990): 393–442.

Deheuvels, P. “Point Processes and Multivariate Extreme Values.” *Journal of Multivariate Analysis* 13 (1983): 257–272.

Dekkers, A. L. M., Einmahl, J. H. J. and de Haan, L. “A Moment Estimator for the Index of an Extreme Value Distribution.” *Annals of Statistics* 17 (1989): 1833–1855.

Dixon, M. J. and Tawn, J. A. “Spatial Analyses for the UK Coast.” *Proudman Oceanographic Laboratory* Report 112, Bidston, 1997.

Drees, H. “Weighted Approximations of Tail Processes for Beta-Mixing Random Variables.” *Annals of Applied Probability* 10 (2000): 1274–1301.

Embrechts, P., Kluppelberg, C. and Mikosch, T. *Modelling Extremal Events*. New York: Springer, 1997.

Embrechts, P., McNeil, A., and Straumann, D.”Correlation and Dependency in Risk Management: Properties and Pitfalls.” *Risk Management: Value at Risk and Beyond.* (editors, Dempster, M. and Moffatt, H.).Cambridge: Cambridge University Press, 2001.

Engle, R. “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation” *Econometrica* 50:(July 1982) 987–1007.

Ferro, C.A.T, and Segers, J. “Inference for Clusters of Extreme Values.” *Journal of Royal Statistical Society B* 65 Part 2 (2003):548-556.

Fisher, R.A., Tippet, L.H.C. “Limiting Form of the Largest or Smallest Numbers of a Sample.” *Proceedings of Cambridge Philosophical Society* 24 (1928): 180-190.

Frey, R. and McNeil, A.J. “Modelling Dependent Defaults.” ETH E-Collection. (2001): http://e-collection.ethbib.ethz.ch/show?type=bericht&nr=273

Gnedenko, B. “Sur la Distribution Limite du Terme maximum d’une Série Aléatoire” *Annals of Mathematics* 44 (1943): 423-453.

Guillou, A. and Hall, P. “A Diagnostic for Selecting the Threshold in Extreme Value Analysis.” *Journal of Royal Statistical Society B* 63 (2001): 293–305.

Gumbel, E. *Statistics of Extremes.* New York: Columbia University Press, 1958.

Heffernan, J. E. “A Directory of Coefficients of Tail Dependence.” *Extremes* 3 (2000): 279–290.

Hill, B. M.”A Simple General Approach to Inference about the Tail of a Distribution.” *Annals of Statistics* 3 (1975): 1163–1174.

Hsing, T. “On the Characterisation of Certain Point Processes.” *Stochastic Processes and their Applications* 26 (1987): 297–316.

Hsing, T., Hüsler, J. and Leadbetter, M. R. “On the Exceedance Point Process for a Stationary Sequence.” *Probabability Theory Related Fields* 78 (1988): 97–112.

Hsing, T., Hüsler, J. and Reiss, R. “The Extremes of a Triangular Array of Normal Random Variables.” *Annals of Applied Probability* 6 (1996): 671–686.

Jenkinson, A.F. “The Frecuency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements.” *Quarterly Journal of Royal Meteorological Society* Vol 81 (1954): 158-171.

Kellezi, E., and Gilli, M. “Extreme Value Theory for Tail-Related Risk Measures.” Preprint, Department of Econometrics and FAME, University of Geneva, Geneva, 2000.

Kratz, M.F. and Rootz´en, H. “On the rate of Convergence for Extremes of Mean Square Differentiable Stationary Normal Processes.” *Journal of Applied Probability* 34 (1997): 908–923.

Künsch, H. R. “The Jackknife and Bootstrap for General Stationary Observations.” *Annals of Statistics* 17 (1989):1217–1241.

Leadbetter, M. R. “Extremes and Local Dependence in Stationary Sequences.” *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* 65 (1983): 291–306.

Leadbetter, M. R., Lindgren, G. and Rootz´en, H. *Extremes and Related Properties of Random Sequences and Series*. New York: Springer, 1983.

Leadbetter, M. R.,Weissman, I., de Haan, L. and Rootz´en, H. “On Clustering of High Levels in Statistically Stationary Series.” *Proceedings of 4th International Meeting in Statistics and Climate* (ed. J. Sansom). Wellington: New Zealand Meteorological Service, 1989.

Ledford, A.W. and Tawn, J. A. “Statistics for Near Independence in Multivariate Extreme Values.” *Biometrika* 83 (1996): 169–187.

Ledford, A. W. and Tawn, J. A. “Modelling Dependence within Joint Tail Regions.” *Journal of Royal Statistical Society B* 59 (1997): 475–499.

Ledford, A.W. and Tawn, J. A. "Concomitant Tail Behaviour for Extremes." *Advanced Applied Probability* 30 (1998):197–215.

Liang, K.Y. and Self, S. G. "On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic." *Journal of Royal Statistical Society B* 58 (1996):785–796.

Longin, F. "The Asymptotic Distribution of Extreme Stock Market Returns." *Journal of Business* 69 (1996):383–408.

Mandelbrot,B.B. "The Variation of Certain Speculative Prices." *Journal of Business* 36 **(**1963):392– 417.

McNeil, A.J. "Calculating Quantile Risk Measures for Financial Time Series Using Extreme Value Theory." ETH E-Collection. (1998): http://e-collection.ethbib.ethz.ch/show?type=bericht&nr=85

Morrison, J. E., and Smith,J.A."Stochastic Modeling of Flood Peaks Using the Generalized Extreme Value Distribution" *Water Resource Research* 38(12) (December 13, 2002).

O'Brien, G. L. "Extreme Values for Stationary and Markov Sequences." *Annals of Probability* 15 (1987): 281–291.

Pagan A. "The Econometrics of Financial Markets." *Journal of Empirical Finance* 3 **(**1996): 15 –102.

Parkinson, Michael. "The Extreme Value Method for Estimating the Variance of the Rate of Return." *Journal of Business* 53 (1980): 61-65.

Pickands, J. "The Two-Dimensional Poisson Process and Extremal Processes." *Journal of Applied Probability* 8 (1971): 745–756.

Pickands, J. "Statistical Inference using Extreme Order Statistics." *Annals of Statistics* 3 (1975):119–131.

Pickands, J. "Multivariate Extreme Value Distributions." *Proceedings of 43rd Session in International Statistical Instruction* (1981): 859–878.

Raab, M. "Compound Poisson Approximation of the Number of Exceedances in Gaussian Sequences." *Extremes* 1 (1998): 295–321.

Reed, D. W. and Stewart, E. J. "Inter-Site and Inter-duration Dependence on Rainfall Extremes."*Statistics for the Environment 2:Water Related Issues* (editors, Barnett, V. and Turkman, K.F.). Chichester:Wiley (1994):125–143.

Reiss, R.D. and Thomas, M. “Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields.” Basel:Birkhauser Verlag, 2001.

Resnick, S. I. and Starica, C. “Consistency of Hill’s Estimator for Dependent Data.”*Journal of Applied Probability* 32 (1995): 139–167.

Resnick, S. I. and Starica, C. “Tail Index estimation for Dependent Data.” *Annals of Applied Probability* 8 (1998):1156–1183.

Rootz´en, H. “The Rate of Convergence of Extremes of Stationary Normal Sequences.” *Advanced Applied Probability* 15 (1983): 54–80.

Shephard, N. “Statistical Aspects of ARCH and Stochastic Volatility.” In *Time Series Models in Economics, Finance and Other Fields* (editors Cox, D. R., Hinkley, D. V. and Barndorff-Nielson, O. E.), London: Chapman and Hall. (1996):1–67.

Simiu, E. and Scanlan, R. H. *Wind Effects on Structures*. New York: Wiley, 1986.

Smith, R. L. “Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground Level Ozone.” *Statistical Science* 4 (1989): 367–393.

Smith, R. L. “Regional Estimation for Spatially Dependent Data.” *Technical Report*. University of North Carolina: Chapel Hill, 1991.

Smith, R. L. “The Extremal Index for a Markov-Chain.” *Journal of Applied Probability* 29 (1992): 37–45.

Smith, R. L., Tawn, J. A. and Coles, S. G. “Markov Chain Models for Threshold Exceedances.” *Biometrika* 84 (1997):249–268.

Smith, R. L. and Weissman, I. “Estimating the Extremal Index.” *Journal of Royal Statistical Society B* 56 (1994): 515–528.

Tawn, J. A. “An Extreme Value Theory Model for Dependent Observations.” *Journal of Hydrology* 101 (1988a):227–250.

Tawn, J. A. “Bivariate Extreme Value Theory: Models and Estimation.” *Biometrika* 75 (1988b): 397–415.

Yun, S. “ The Extremal Index of a Higher-Order Stationary Markov-Chain.” *Annals of Applied Probability* 8 (1998): 408–437.

Yun, S. "The Distributions of Cluster Functionals of Extreme Events in a *d*th-Order Markov Chain." *Journal of Applied Probability* 37 (2000): 29–44.

Weblinks:
Tradetrek. http://www.tradetrek.com/education/stock_charts.asp