

CERTIFICATE

Pareto Exponent Estimation using Rolling Regression and Nonparametric Methods

By

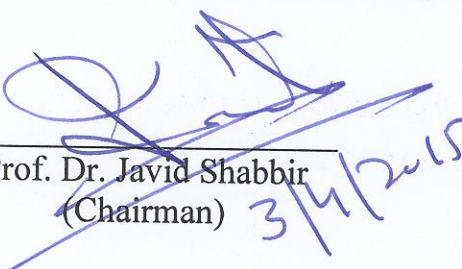
Muhammad Asif

(Reg. No. 02221311009)


A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY IN
STATISTICS

We accept this thesis as conforming to the required standards


1.


Prof. Dr. Javid Shabbir
(Chairman) 3/4/2015

2.


Dr. Zahid Asghar
(Supervisor)

3.


Dr. Masood Anwar
(External Examiner)

DEPARTMENT OF STATISTICS
QUAID-I-AZAM UNIVERSITY
ISLAMABAD, PAKISTAN
2015

**PARETO EXPONENT ESTIMATION USING
ROLLING REGRESSION AND NONPARAMETRIC
METHODS**



BY

MUHAMMAD ASIF

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2015

**PARETO EXPONENT ESTIMATION USING
ROLLING REGRESSION AND NONPARAMETRIC
METHODS**



By

MUHAMMAD ASIF

Supervised By

Dr. ZAHID ASGHAR

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2015

Declaration

I Mr. **Muhammad Asif** hereby solemnly declare that this thesis entitled "**Pareto Exponent Estimation Using Rolling Regression and Nonparametric Methods**", submitted by me for the partial fulfillment of Master of Philosophy in Statistics, is the original work and has not been submitted concomitantly or latterly to this or any other university for any other Degree.

Dated: _____

Signature: _____

Dedication

*This work is dedicated to
the fond and loving memory of my*

Late Father

"Surely to allah do we belong and into Him shall we return"

Acknowledgment

All praise and gratitude is to Almighty **Allah**, who owed upon me the potential and patience to complete this work. All respect goes to the Holy Prophet **Hazrat Muhammad (P.B.U.H)**, who enlightens our conscience with the essence of faith in Allah.

I am highly privileged to thank, **Dr. Zahid Asghar**, for taking the responsibility as my research supervisor and for his useful comments, remarks and engagement through the learning process of this research work. I am grateful to all my teachers specially **Dr. Muhammad Aslam Dr, Javid Shabbir, Dr. Fareed Khan, Dr. Zawar Hussain, Dr. M. Riaz, Dr. Ejaz Hussian, Dr. Abdul Haq** and **Mr. Manzoor Khan** for enlightening my statistical know how during my course work and research work.

Special thanks extended to all of my friends and class fellows, **Mariya Raftab, Syed Farooq Shah, M. Imran Shahid , Hina Abid, Rehma Asad, M. Shahid, Sehrab Ali, Maqbool Jan, Waheed Ullah Wazir, Wajeih Ul Jamal, Niama Mubarak, Erum Zahid, Tahir Iqbal, Shakeel Ahmad, Shoaib Iqbal, Talha Umer** for their cooperation in all respects. I like to thanks my friend **Mr. Akbar Ali Wazir** who have willingly shared their precious time during the process of write up.

My sincere thanks goes to my **Parents** for their love and support throughout my life. I owe my loving thanks to my brothers **Muhammad Arif** and **Muhammad Abid**, to my sisters **Umm E Kulsoom** and **Umm E Azeema** for their great love and support. Without their support and encouragement it would have been impossible for me to complete this work.

Muhammad Asif

Contents

Abstract	iii
List of Abbreviation	vii
List of Tables	vii
List of Figures	viii
1 Introduction	2
1.0.1 Variation Pattern in Power Law Exponent	9
1.0.2 Objectives of the study	9
1.0.3 Organization of Study	10
2 Review of Literature	11
3 Methodology	18
3.1 The Model	19
3.2 Parameters Estimation of Power Law distribution	19
3.2.1 Estimation of Scale Parameter (x_{min})	20
3.2.2 Estimating Shape Parameter ($\hat{\alpha}$)	20
3.2.3 Ordinary Least Square (OLS) Estimator	21

3.2.4	Modified Ordinary Least Square (MOLS) estimator	21
3.2.5	Hill's Estimator	22
3.2.6	Maximum Likelihood Estimator (MLE)	22
3.2.7	Minimum Variance Unbiased Estimator	23
3.3	Goodness of Fit	23
3.3.1	Kolmogorov-Smirnov test	23
3.4	Bootstrap Investigation of Pareto Index Using Different Estimators	24
3.4.1	Simulation Design	24
3.5	Elasticity of Power Law Exponent with Respect to Truncation Point and Sample Size using Recursive Sampling	25
3.5.1	Rolling Sampling Using With Replacement Sampling	26
3.6	A Nonparametric Analysis of Power Law Exponent	27
4	Empirical Evidence of Power Law distribution	28
4.0.1	Data	29
4.0.2	Descriptive Statistics	30
4.1	Minimum Threshold	31
4.2	Graphical Representations	33
4.2.1	Graphical Representation for US and China	33
4.2.2	Graphical Representation for the Pakistan and India	35
4.3	Goodness of fit	37
4.3.1	Goodness of fit test for US and India	38
4.3.2	Goodness of fit test for Pakistan and India	38
4.4	Estimate's of Power law Exponent	39
4.4.1	Estimates of Power Law Exponent for US and China	39

4.4.2	Estimates of Power Law Exponent for Pakistan and India	40
4.5	Simulation study	40
4.5.1	Simulation Results for the US and China	41
4.5.2	Simulation Results for the Pakistan and India	44
5	Rolling Sampling and Non-Parametric Analysis of Pareto Exponent	47
5.1	Rolling Sampling	47
5.1.1	Rolling sample results for US and China	48
5.1.2	Rolling Sampling Results for Pakistan and India	51
5.2	Rolling Sampling by Using With Replacement Sampling	53
5.2.1	Results of Rolling Random Sampling for US and China	53
5.2.2	Results of Rolling Random Sampling for Pakistan and India	56
5.3	Nonparametric Analysis of Power Law Exponent	58
5.3.1	Kernel Density Plots for US and China	59
5.3.2	Kernel density plots for Pakistan and India	61
6	Conclusions and Recommendations	63
6.0.3	Recommendations	64
	References	66

APPENDICES APPENDICES APPENDICES

List of Tables

4.1	Descriptive Statistics for Developed Countries (U.S, China)	30
4.2	Descriptive Statistics for Developing Countries (Pakistan, India)	31
4.3	Number of Cities After and Before Truncation	32
4.4	Number of Cities After and Before Truncation	32
4.5	Goodness of fit for U.S and China	38
4.6	Goodness of fit for Pakistan and India	38
4.7	Estimates of the Power Law Exponent for US and China	39
4.8	Estimates of the Power Law Exponent for Pakistan and India	40
4.9	Simulation Results for U.S	42
4.10	Simulation Results for China	43
4.11	Simulation Results for Pakistan	45
4.12	Simulation Results for India	46

List of Figures

4.1	plots of US Data (1990 and 2000)	33
4.2	log log plots of US Data (1990 and 2000)	34
4.3	plots of China Data (2005 and 2010)	34
4.4	log log plots of China Data (2005 and 2010)	35
4.5	plots of Pakistan Data (1981 and 1998)	36
4.6	log log plots of Pakistan Data (1981 and 1998)	36
4.7	simple Plots for India (2001 and 2011)	37
4.8	log log plots of India (2001 and 2011)	37
5.1	Rolling Sampling Plots for U.S 1990	49
5.2	Rolling Sampling Plots for U.S 2000	50
5.3	Rolling Sampling Plots for Pakistan 1981	51
5.4	Rolling Plots for Pakistan 1998	52
5.5	Rolling Random Sampling Plots for US (1990)	54
5.6	Rolling Random Sampling Plots for US (2000)	55
5.7	Rolling Random Sampling Plots for Pakistan (1981)	56
5.8	Rolling Random Sampling Plots for Pakistan (1998)	57
5.9	Kernel desity plots for US (1990)	59
5.10	Kernel desity plots for US (2000)	60

5.11	Kernel desity plots for Pakistan (1981)	61
5.12	Kernel density plots for Pakistan (1998)	62
1	Rolling Sampling Plots for China (2005)	70
2	Rolling Sampling plots for China (2010)	71
3	Rolling Sampling Plots for India (2001)	72
4	Rolling Sampling Plots for India (2011)	73
5	Rolling Random Sampling Plots for China (2005)	74
6	Rolling Random Sampling Plots for China (2010)	75
7	Rolling random sampling plots for India (2001)	76
8	Rolling Random Sampling Plots for India (2011)	77
9	Kernel Density Plots for China (2005)	78
10	Kernel Density Plots for China (2010)	79
11	Kernel Density Plots for India (2001)	80
12	Kernel Density Plots for India (2011)	81

Abstract

In this study we have checked the validity of Zipf's Law for city size data of U.S, China, Pakistan and India. Zipf's Law says that, the distribution of city sizes follows a Power Law distribution with shape parameter equal to 1. We have used two-step approach to check the validity of Zipf's Law, where in first step we test (using goodness of fit test) if the distribution of city sizes follow a Power Law distribution and in second step, we estimate the Power Law exponent whether its value is equal to unity or not. The HILL, OLS, MOLS, ML and MVU estimation techniques are considered for the estimation purposes. Graphical display is presented to overview the nature of the city size data sets. The Kolomogrove-Simirnov (KS) goodness of fit test is applied to check the distributions of the all the data sets, assuming the Power Law distribution under null hypothesis. The KS statistics is also used to estimate the minimum threshold values.

Simulation study is carried out to point out an efficient estimator for the estimation of the Power Law exponent. Base on the bootstrap simulation we conclude that minimum variance unbiased estimator (MVUE) is more efficient and unbiased. The range, in which the exponent value is one, is to be found through rolling sampling technique under the considered estimation methods. A nonparametric analysis is carried out to give a more detailed description of the Power Law exponent. It will be shown, through kernel density plots (a nonparametric technique), that the Power Law exponent distribution is uni-modal.

CHAPTER 1

Introduction

City size distribution is a broad term. A lot of work has been done in analyzing the city size distribution for both developing and developed countries. Various dimensions for modeling the size of the cities have been proposed. It has been investigated through these studies whether there exists evidence that some cities grow faster than other cities? What are the basic reasons for these particular growth patterns? What are the problems associated with these growth rates?

Cities are the center of many economic activities, commonly cities are of different sizes and the main focus of the researchers is to study the size distribution of cities. Therefore, it is necessary to put attention on the urban expansion of different countries through city size distribution. Cities are important centers for poverty reduction and development in both urban and rural areas. They play a vital role for the national economic activity and provide a link between rural areas. Urban areas enjoy properties like high literacy, better health, and greater access to social services. When the necessary infrastructure is not developed or when policies are not well planned then rapid and unplanned growth will lead unsustainable development, pollution and environment deprivations. The criteria for classifying an area are as urban

on the basis of following characteristics like population density; proportion employed in non-agricultural sectors; the presence of infrastructure such as paved roads, electricity, piped water or sewers; and the presence of education or health services. Thus urbanization is the extremely concerned phenomena that cannot be ignored.

The proper definition of the city is a burning topic in literature, while analyzing city size distribution, it is very important to consider a proper urban unit. As we know that in different time period different proportion of the country's population are living in the cities. Any study of Zipf's Law for city size distributions usually faces the problem that what is meant by city. The definition of cities varies with respect to the country as well as within the country after some period of time. Any city can be defined as:

1. Administratively defined city or politically defined city e.g. area defined by a committee of a town.
2. Economically defined cities (Metropolitan cities) where most of the people are not engaged in the agriculture and there exists high employment and there exist characteristics of the urban.

To highlight the effect of definition of cities on the Pareto exponent [Rosen and Resnick \(1980\)](#) checked the validity of Zipf's law for six countries, considering the proper city and metropolitan areas. He found that the value of the index is more close to unity for the metropolitan areas as compared to the proper city.

When compiling information on city population size, it is better to use data or estimates based on the concept of urban agglomeration. When those data are not available, population data that refer to the city were used. However, when the administrative boundaries of cities remain fixed for long periods of time, they are likely to misrepresent the actual growth of a city with respect to both its territory and its population.

Many natural phenomena's like distribution of wealth and income in a society, distribution of face book likes, distribution of football goals follows power law distribution (Zipf's Law). Like above phenomena's, distribution of city sizes also follow Power Law distribution. [Auerbach \(1913\)](#) first time gave the idea that the distribution of city size can be well approximated with the help of Pareto distribution (Power Law distribution). This idea was well refined by many researchers but [Zipf \(1949\)](#) worked significantly in this field. The distribution of city sizes is investigated by many scholars of the urban economics, like [Rosen and Resnick \(1980\)](#) , [Black and Henderson \(2003\)](#), [Ioannides and Overman \(2003\)](#), [Soo \(2005\)](#), [Anderson and Ge \(2005\)](#) and [Bosker et al. \(2008\)](#).

Zipf's law states that:

"The rank of cities with a certain number of inhabitants varies proportional to the city sizes with some negative exponent, say that is close to unit".

In other words, Zipf's Law states that the product of city sizes and their ranks appear roughly constant. This indicates that the population of the second largest city is one half of the population of the largest city and the third largest city equal to the one third of the population of the largest city and the population of n^{th} city is $\frac{1}{n}$ of the largest city population. This rule is called rank, size rule and also named as Zipf's Law. Hence Zip's Law not only shows that the city size distribution follows the Pareto distribution, but also show that the estimated value of the shape parameter is equal to unity.

According to [Auerbach \(1913\)](#), the relation between city sizes and their ranks is $R_i P_i = C$, where P_i is the population of the i^{th} city, R_i denotes the rank of the i^{th} city and A is a positive constant. The relation $R_i P_i = C$, states that when we sort cities in decreasing order with respect to their population, rank them and then the product of rank and there population is roughly a constant. We elaborate here the rank size rule in detail. Let us consider cities of a

country, count the number of inhabitants in that city (city population), then one can arrange the cities in the descending order, so that the most populated cities get rank equal to 1, the second most populated city gets rank equal to 2 and so on. After ranking the cities and then multiply rank with the population, we get a constant relation. Let us consider the city size data of ten most populated cities of Pakistan. The results of the above relation are shown in the following table.

City Name	Rank	Size	Product
Karachi	1	5208132	5208132
Lahore	2	2952689	5905378
Faisalabad	3	1104209	3312627
Rawalpindi	4	794843	3179372
Hyderabad	5	751529	3757645
Multan	6	732070	4392420
Gujranwala	7	600993	4206951
Peshawar	8	566248	4529984
Sialkot	9	301609	2714481
Sargodha	10	291362	2913620

From the above table, it is apparent that there exists approximately a constant relationship between the rank and size of the Pakistani city for the year 1981 and the range of the constant relation is between [2714481, 5905378]. We apply logarithm on both sides of the relation $R_i P_i = A$.

$$\log R_i + \log P_i = \log C \quad (1.1)$$

$$\log R_i = \log C - \log P_i \quad (1.2)$$

From equation 1.2, we can see that the Pareto curve is in linear form and we can easily plot this linear equation in double log scale. The interpretation of the slope term is given in detail

in the next section. Generalized form of Equation 1.2 can be written as:

$$R_i = CP_i^{-(\alpha+1)} \quad (1.3)$$

If $R_i = p(x_i)$ and $P_i = x_i$ then above equation can be written as:

$$p(x_i) = Cx_i^{-(\alpha+1)} \quad (1.4)$$

From equation 1.4 it is apparent that $p(x_i)$ diverges as $x \rightarrow 0$. It means that the distribution must deviate from the Power Law distribution below some minimum value called X_{min} . Hence equation 1.4 can be normalized as $\int_{x_{min}}^{\infty} p(x)d(x) = 1$ to obtain the value of constant C only if x and α obey the following condition $\alpha > 0$ and $x \geq X_{min}$. Hence equation 1.4 becomes:

$$p(x_i) = Cx_i^{-(\alpha+1)}, \alpha > 0, x \geq x_{min} > 0 \quad (1.5)$$

In this study, we have used the two step approach to check the validity of Zip'f Law following Terra (2009). In the first step, we check, by applying goodness of fit test, that whether the city size data follow Power Law distribution within a country. In the second step, we estimate the Power Law exponent and check that whether its value equals to unity (Zip'f Law) or not. It is to be noted that the usual approach to test the validity of Zip'f Law based upon the estimated value of the Power Law exponent is misleading. Hence we need to check at first step that whether the underlying distribution is Power Law or not.

Testing only for the equality of Power Law exponent, ignoring the value of the constant, does not constitute a proper test of the rank size distribution and the conclusion drawn may be misleading (Rosen and Resnick (1980)). Alperovich (1984) concluded from his study that the

confirmation of rank size distribution is just to obtain the value of the Power Law exponent equal to unity and the value of the constant is equal to the population of the largest city which is not supported by the data. Instead if the confirmation of the rank size distribution is considered in such a way that the value of the Power Law exponent is equal to unity and the value of the constant is taken equal to the average of the product of the ranks and population of cities, then this confirmation is supported by the data. This means that the condition of taking the intercept in simple linear regression line equal to the size of the largest city is misleading. Alternatively, we should take the intercept equal to the average of the product of the ranks and population of cities.

Related to the city size data, [Eeckhout \(2004\)](#) have discussed two important ideas.

1. When the city size data is considered as a whole with no restriction on their size, the Lognormal distribution best fits the data.
2. When the true distribution is Lognormal then the plot of rank size is concave and the value of the Power Law exponent decreases. This means that a sample size can be found for which the Zipf's Law holds exactly.

[Soo \(2007\)](#) pointed out that city size is too much studied due to following two main reasons

1. Zipf's Law and Power Law exponent provide useful information related to the distribution of urban system.
2. The Power Law index is closely related to the Gini coefficient. As Gini coefficient is used to measure the inequality of income in a society, similarly Power Law index is also used to study the distribution of population in the cities.

We know that the number of cities with certain population (sample size) affects the estimated coefficient.

If we use the first criterion, then we will get a small number of cities for some countries, while for the other countries we will get a large number of cities. Using this criterion, we might include different fractions of urbanized population. On the other hand, if we use the second criterion, we will get the same number of cities for each country but the limitation of this criterion is that, for the small countries, the n^{th} ranked city might be a village and for the large country the n^{th} ranked city might be a large city. In this study, we consider the first criterion but the threshold x_{min} is fixed by using Kolomogrov-Simernov (K.S) statistics following

As the logarithmic form of the Power Law distribution is $\log(R_i) = \log(C) - (\alpha + 1) \log(P_i)$, the Power Law exponent is linear in this case. So we can easily plot Log-Log plot. Here α is the elasticity function $\alpha = -\frac{d\log(y)}{d\log(x)}$ i.e. change in the number of cities having a particular size due to the change in city size. The negative sign shows that the number of cities decreases as the city size increases.

Also the Zipf's Law explains the fact of the equality of two forces, named as force of diversification and the force of unification in the economy. Force of diversification is defined as the tendency of the population to be split into many communities. This force is related to the economic location and raw material. There exists reverse tendency of the force of unification. This force is defined as the tendency of gathering of population in one community. If all persons in a society located at the same point, then there is maximum force of unification. Hence it can be concluded here that as a result of high force of diversification, the large number of small communities will be formed and as a result of the force of unification, small number of large communities will be formed.

1.0.1 Variation Pattern in Power Law Exponent

In most of the previous studies, the estimated values of the Power Law exponent are based on the fixed sample size. All these studies rule out the possibility of variation in the value of the Power Law exponent with respect to size of the sample. It is to be noted that some scholars ([Rosen and Resnick \(1980\)](#), [Black and Henderson \(2003\)](#) and [Eeckhout \(2004\)](#)) concluded that the estimated value of the Power Law exponent is very sensitive to the choice of the sample size. In this study, we have considered a rolling sampling rather than considering the fixed sample size to check the effect of the sample size on the value of the exponent.

Nonparametric analysis is also used to check the variation in the estimated value of Power Law exponent likewise recursive sampling. Kernel density plots can be used to analyze the estimated values of the Power Law exponent using nonparametric approach. As we know that kernel density method of estimation is widely used nonparametric method for estimating the probability density function of a random variable. The main advantage of constructing the Kernel density plot is that it gives us a more clear description of how the values of the Power Law exponents are distributed. Kernel density plot also describe whether the distribution of Power Law exponent is unimodal or bimodal. In this study, we have used a nonparametric analysis to get a more clear interpretation of how the values of the Power Law exponents obtained from the recursive sampling are distributed.

1.0.2 Objectives of the study

The main objectives of this study are as follows.

1. To study the validity of Zipf's Law for the Developed (U.S, China) and developing countries (India, Pakistan) using different estimation techniques.

2. To compare different estimators based on their Bias and MSE using real life data and bootstrap simulation.
3. To check the sensitivity of the Power Law exponent by using Rolling sampling.
4. To observe the distribution of Power Law exponent using nonparametric technique.

1.0.3 Organization of Study

Chapter 2 explains literature review on Zipf's Law. The methodology about the existence of the Zipf's Law for the city distribution is discussed in Chapter 3. In chapter 4, we empirically check the validity of Zipf's Law. In chapter 5, we adopt the Rolling sample technique and nonparametric analysis for checking the variation in estimated value of Power Law exponent. Chapter 6 comprises the concluding remarks and recommendations for further study.

CHAPTER 2

Review of Literature

As we know that cities are very complex systems that differ in many ways like in their size, scale and shape. Related to city size distribution, two important questions have been raised by urban economists. The first question concerns with how cities sizes grow relative to each other, the second is about which theoretical distribution better fits the city size data. Most empirical studies related to the former question suggest that the relative size and rank of cities remain stable over time. With respect to the second question, literature argue strongly that the Power Law distribution fits best.

The idea that city size distribution is well approximated by Pareto distribution (Power Law distribution) is first argued by [Auerbach \(1913\)](#). While studying the distribution of cities in Germany, he concluded that there exists an empirical relationship between Rank and Size of the cities. The linear trend exists between these two phenomena on a double logarithmic scale. Ever since this theory proposed by [Auerbach \(1913\)](#) , become widely held opinion among researchers in a variety of disciplines.

[Zipf \(1949\)](#) made a major contribution in this field. He pointed out that city size distribution can not only be described by Power Law distribution, but the size distribution of the city

can take a specific form of Power Law distribution when the value of the shape parameter is equal to 1. He also found that the constant value is equal to the population of the largest city. According to Zipf (1949) there exists an empirical relationship between size and rank of cities. When cities are ordered in decreasing pattern, the second largest city is half of the population of the largest city and third largest city is one third of the size of the largest city and so on.

Rosen and Resnick (1980) in their paper they examined the city size distribution. They used the city size data in 44 countries and found the average value of the Power Law exponent equal to 1.136 which is not much close to 1, indicating that population in most of the countries are more evenly distributed as compared to that predicted by the rank size rule. The range of the Power Law exponent lies in the interval [0.81, 1.96]. Out of 44 Countries, the value of the Power Law exponent is less than 1 for only 12 countries. They explained the variation pattern in the Power Law exponent and concluded that the value of the Power Law exponent is sensitive to the definition of city and the number of cities included in the sample. In order to check whether the larger cities grow faster than the smaller cities, the authors included a non-linear term in the Pareto equation. This inclusion yielded positive coefficient of the non-linear term which indicates that large cities grow faster than the smaller cities.

A more detailed study on the city size distribution is done by Guérin-Pace (1995). He has considered the city size data of French for almost two centuries (1831 and 1990). He included the city which has a population more than 2000 inhabitants and checked the sensitivity of the Pareto exponent with respect to the sample selection criteria. There are many papers in which the model is developed to test directly the validity of Zipf's law. Gabaix (1999) gave a statistical explanation of Zipf's Law. In his article, he showed that if different cities grow with the same expected rate and with same variance (Gibrats Law) than, in the long run

(Steady state), the distribution of cities follows Zipf's Law.

[Ioannides and Overman \(2003\)](#) used the nonparametric method to evaluate Zipf's exponent instead of using regression of log of rank on log of the size of cities to estimate the Zipf's exponent. By using metro area of U.S for the time period 1900-1990, the Zipf's exponent is calculated from the mean and variance of growth of cities rate. The results of this paper suggest that value of the Zipf's exponent vary across the cities.

[Soo \(2005\)](#), empirically tested the validity of Zipf's Law using the city size data of 73 countries. He had used two estimation methods i.e. Hill and OLS. By using OLS method of estimation, the Zipf's Law was invalid for more than half of the sample (53 out of 73 countries (73% of the sample)) and this result is consistent with [Rosen and Resnick \(1980\)](#) who rejected Zipf's Law in 36 out of 44 countries. [Soo \(2005\)](#) also rejected Zipf's Law for 30 out of 73 countries (41% of the sample) using HILL estimation method and this results are not the same as those of the [Rosen and Resnick \(1980\)](#). The author concluded that the value of the estimate of the Power Law exponent depends upon the estimation technique we adopt. [Soo \(2005\)](#) also tried to explain variation in the Power Law exponent and argued that political economy is the main factor which affects the value of the Power Law exponent. In this paper, it is also concluded that the variation in the value of the Power Law exponent is better explained by political economic variable. One more results the author got is that the average value of the Pareto exponent for the urban agglomeration is less than 1 and thus Zipf's Law does not hold.

[Gan et al. \(2006\)](#) claimed that Zipf's Law is spurious (fake) in explaining the city size distribution. To prove their statement, the authors used the Monte Carlo simulation technique to examine the rank size relation between the dependent and independent variables. To study Zipf's Law, they considered two real data sets of US urbanized area for the years 1990 and

2000, china data set for the years 1985 and 1999. Two main findings have been obtained. First, the values of R^2 for both countries were very high indicating that Zipf's Law fits well to all these cases. Second, the estimated value of β was close to 1 but their corresponding standard deviation was showing that β is not equal to 1 statistically for all four cases. On the basis of standard deviation of β , it is concluded that rank size rule does not hold for the four cases (data sets). Does high value of R^2 imply that the city's size follows the power law? This statement was checked with the help of non-parametric (KS) test. In case of US data for the year 1990 and 2000, the Pareto distribution was not rejected while all other distributions were rejected. For the case of china, the KS test rejected all the distributions with $p= 0.000$ which implies that Chinese data does not follow a Pareto distribution. Therefore, it is concluded that the Zipf's Law with a high degree of explanatory power does not necessarily imply that city size follow a Power Law distribution.

[Moura Jr and Ribeiro \(2006\)](#) used the Zipf's law for the cities in Brazil. They collected data only from those cities whose inhabitants are more than 30,000. They considered city size data for the time period 1970, 1980, 1991 and 2001. The results reveal that the Brazilian population distribution does not follow a Power law seem like as other countries. Estimates of Power Law exponent for the census 1970 and 1980 are 2.22 ± 0.34 and for 1991 and 2000 was 2.26 ± 0.11 . The results obtained from MLE's for 1970 are 2.41 and for the other three years was 2.36.

[Nota and Song \(2007\)](#), analyzed Zipf's exponent by changing the sample size and truncation point by using rolling sampling. They estimated the values of the exponent and also checked the elasticity of the exponent with respect to the sample size. In this paper the authors considered U.S. (1990 and 2000) and China (1985 and 1999) data. Two estimation techniques were used in this paper OLS and modified OLS. By using Rolling sampling,

they concluded that rank size rule holds for the selected sub-samples. In order to check the elasticity of the Power Law exponent with respect to the sample size, the authors regressed the estimated coefficients on the sample sizes. In case of U.S data, one percent increase in sample size gave 0.15 percent decrease in the value of the estimated coefficient. Therefore, the authors concluded that Zipf's exponent depends upon the sample size and rank size rule does not holds always.

[Sarabia and Prieto \(2009\)](#), have introduced Positive Pareto Stable (PPS) distribution to model the city size data. They demonstrated the problem of selecting a correct truncation point for estimating the Power Law. For the existence of optimal points, Akaike Information Criterion (AIC) and simulation study are used under different assumptions. Different kinds of heavy tailed distributions are considered in this paper. They also described the methodology for the city size data of Spain for optimum estimation of Power Law distribution and came across the conclusion that this new distribution (PPS) out performs other (Tsallis, Pareto and Lognormal) distributions.

[Gangopadhyay and Basu \(2009\)](#), gave a new structure to analyze the city size distribution. He has used two-step approach to check the validity of Zipf's Law. In the first step, the author formally tested that the distribution of cities follows the Power Law distribution and in the second step, the exponent of the Power Law distribution is estimated. By using the Monte Carlo simulation, the author compared the performance of MVU estimator with OLS, MOLS and HILL estimators. The MVU estimator performed better and turned out to be unbiased and more efficient as compared to the other estimators. To check the validity of Zipf's Law empirically, based on the two step approach, the author has used city size data of 155 countries. It is concluded that Zipf's Law holds for 62 countries.

[Akhtar and Dhanani \(2012\)](#) first time investigated the city size data of Pakistan. This

study has been made in order to check that whether the city size data in Pakistan follow any of the three rules i.e. Law of Primate city, City size pyramid, rank size rule or any other? In order to check the above mentioned laws, the authors have considered the city size data of the years 1951, 1961, 1972, 1981 and 1998. The authors concluded that the city size data for these time periods is well approximated by City size pyramid.

[Fazio and Modica \(2012\)](#), in their article, studied the relation between city size distribution and the truncation point. The author considered the U.S census data for 2000 and 2010. They applied the recursive truncation approach to estimate the Zipf's Law. By calculating the recursive estimates, the authors showed the presence of Zipf's Law for each truncation sample of the distribution of US cities. Also the authors applied the above methods to simulated data set. The results confirmed the sensitivity of truncation point. Based upon the test results, using real and simulated data sets, they highlighted the difficulty to distinguish between the Power Law upper tail and the tail of Log-normal. By using the census data set, the authors obtained recursive estimate of the Power Law exponent which showed that as the observations in the right tail increases, the estimated value of α decreases.

[González-Val et al. \(2013\)](#) used four densities (Lognormal, q-exponential, Log-Logistic and Double Pareto Lognormal) for analyzing the city size distribution of urban economic. The data sets have been taken from the US, Italy and Spain from 1900 to 2010 with no restriction on the city size. MLE method of estimation was used for estimating the parameters of above densities. In order to check the goodness of fit for the data, Kolmogorov-Smirnov and Cramer's-von Mises test was performed. To check which distribution better fits the data, they computed AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). Double Pareto Lognormal distribution best fitted all the data sets in most of the cases i.e. (86.76%).

Luckstead and Devadoss (2014), have analyzed the city size data of China and India for 1950-2010 for three probability distributions (Pareto, Lognormal and General Pareto). The authors studied how the distribution of cities gets change during these time periods. It is apparent from this study that the world two famous countries have similar trend. The chines data set for the time period 1950-1990 is well modeled by Log normal distribution, where as it is well modeled by Pareto distribution for the time period 2010 but it does not follow Zipf's Law. Indian city size data for earlier time period is also well modeled by lognormal distribution and Zipf's Law for the time period 2000 and 2010 holds.

Amalraj et al. (2014) defined Pareto positive distribution as a new model to describe the city size distribution of a country. PPS distribution is defined as a flexible model for fitting the entire range of a country. Pareto distribution is treated as special cases of PPS distribution. The PPS distribution is compared with Pareto distribution and Log-normal distribution. They predicted number of cities for future time period by using Lagrange method of interpolation.

CHAPTER 3

Methodology

As the main objective of our study is to examine the city size distribution for the developed and developing countries i.e. we want to examine whether city size distribution for U.S (1990, 2000), China (2005, 2010), Pakistan (1981, 1998) and India (2001, 2011) follow Zipf's Law. We have used parametric methods to estimate Power Law exponent applied the goodness of fit test of the above data sets. The graphical representation e.g. histogram and log-log plots are also used in this study to examine the rank size rule. We know that when we plot the log of rank versus log of city size yields a straight line having slope equal to -1 which confirms the validity of Zipf's Law ([Zipf \(1949\)](#)).

The value of the Power Law exponent is too much sensitive to the sample size (number of cities to be included in the analysis) and the truncation point x_{min} (below which we truncate the data). In order to check this argument we have used recursive sampling and non-parametric analysis (Kernel density plots are made to check the distribution of the Power Law exponent). Detail of these techniques is given below..

3.1 The Model

Let x denotes the size of cities then the probability density function (pdf) of Power Law distribution is given by

$$f(x) = \frac{\alpha}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-(\alpha+1)}, x \geq x_{min} > 0, \alpha > 0 \quad (3.1)$$

where α is shape and x_{min} is scale parameter.

The cumulative density function of Power Law distribution is given by

$$F(x) = 1 - \left(\frac{x}{x_{min}} \right)^{-\alpha}, x \geq x_{min} > 0, \alpha > 0 \quad (3.2)$$

One main goal of this study is to investigate the distribution of city size data and to use an accurate and robust estimation method to estimate the empirical distribution of city size data. For that purpose, correct fitting of Power Law distribution to empirical data is very important. Most of the previous studies used different criteria to fix the threshold value x_{min} as well as considered different estimation methods to estimate the scaling parameter α of Power Law distribution.

3.2 Parameters Estimation of Power Law distribution

We know that most of the previous studies considered different estimation techniques to estimate the Power Law exponent. In this study we have considered five estimation techniques and compared them on the basis of their estimated values.

3.2.1 Estimation of Scale Parameter (x_{min})

The first step in the fitting Power Law distribution is to correctly estimate the value of its scale parameter x_{min} which is also known as the threshold from above which we consider the data set. It is commonly discussed that the estimated value of the shape parameter $\hat{\alpha}$ of the Power Law exponent is very much sensitive to the choice the value of x_{min} . After estimating this threshold we will have the data for which the Power law distribution is valid. We adopt here the procedure of [Clauset et al. \(2009\)](#) to estimate the threshold value. We choose value of x as a (x_{min}) which minimizes the KS statistics. The detail of this procedure is given below.

To find the best value of the x_{min} , we go through our data set and choose each value of the x as x_{min} , truncate the data below this and compute the empirical cdf for each truncation value as:

$$F(x) = 1 - \left(\frac{x}{x_{min}} \right)^{-\alpha}, x \geq x_{min} > 0, \alpha > 0 \quad (3.3)$$

On the basis of that truncated data, we calculate KS statistics which is the difference between the empirical cdf and the corresponding theoretical cdf $(1, \frac{n-1}{n}, \frac{n-2}{n}, \dots, \frac{2}{n}, \frac{1}{n})$. We choose that value of X as estimate of the threshold (x_{min}) for which the value of KS statistics is minimum.

It is to be noted that the lower population threshold for a city to be included in the sample varies from one country to another. On average, larger countries have higher thresholds, but also a larger number of cities in the sample and smaller countries have smaller threshold.

3.2.2 Estimating Shape Parameter ($\hat{\alpha}$)

After estimating the Scale Parameter (x_{min}) correctly, we are left with the data set for which the Power Law distribution gives best fit. The next step is to estimate the shape

parameter α of Power Law distribution by using a precise and accurate method. We know that the validity of Zipf's Law extensively depends upon the estimated value of its exponent. In this study we have used five different estimation techniques (OLS, MLE, MOLS, HILL and MVU) following Terra (2009) to estimate the value of the Power Law exponent.

3.2.3 Ordinary Least Square (OLS) Estimator

The Ordinary Least Square method has been widely used for estimating the Pareto index (Gan et al. (2006) and many other). The OLS estimate of the Pareto index is the slope of the following equation.

$$\ln(R_i) = \alpha - \beta \ln(P_i) + e_i, \quad i = 1, \dots, n \quad (3.4)$$

where R_i is rank of i^{th} city in decreasing order, P_i is the population of i^{th} city and e_i is error term.

3.2.4 Modified Ordinary Least Square (MOLS) estimator

The Ordinary Least square estimation method gives biased estimate for small sample size. Gabaix and Ibragimov (2011) gave a solution to this problem and suggested to use $R_i - \frac{1}{2}$ instead of $R - i$ to reduce the Bias. Thus equation 3.4 becomes

$$\ln\left(R_i - \frac{1}{2}\right) = \alpha - \beta \ln(P_i) + e_i, \quad i = 1, 2, 3, \dots, n \quad (3.5)$$

The standard errors of estimated Power Law exponent is given by $\hat{\beta} \sqrt{\frac{2}{n}}$

3.2.5 Hill's Estimator

An alternative approach for reducing the bias in Ordinary least square method in small sample size is to use [Hill et al. \(1975\)](#) estimator. This estimator is defined as

$$\beta^{\hat{HILL}} = \frac{n-1}{\sum_{i=1}^{n-1} (P_{(i)} - P_{(n)})} \quad (3.6)$$

Where, P_i is the population of i^{th} ranked city such as $P_1 \geq \dots \geq P_n$ and P_n is the population of city with rank n .

The standard error for the above estimator is given by

$$\sigma(\beta^{\hat{HILL}}) = \beta^{\hat{HILL}} \left(\sum_{i=1}^{n-1} \frac{(\xi - \frac{1}{\beta^{\hat{HILL}}})^2}{n-2} \right) (n-1)^{-\frac{1}{2}} \quad (3.7)$$

where, $\xi = i(\ln P_{(i)} - \ln P_{(i+1)})$. According to [Ioannides and Overman \(2003\)](#), under the null hypothesis that the underlying data follow a Pareto distribution, the hill estimator is the maximum likelihood estimator. In fact, the HILL estimator is almost same as ML estimator. By comparing both estimator it can be seen that, the numerator of HILL estimator is $n-1$ instead of n , which is the numerator of ML estimator.

3.2.6 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator can be determined by taking the first differential of Log of Likelihood of Power Law distribution and equating it to zero. MLE of the scale parameter is:

$$\beta^{\hat{MLE}} = n \left[\sum_{i=1}^n \ln \left(\frac{x_i}{x_{min}} \right) \right]^{-1} \quad (3.8)$$

3.2.7 Minimum Variance Unbiased Estimator

Jani and Dave (1990) have suggested to use Minimum variance unbiased estimation technique for exponential family of distribution. A minimum variance unbiased estimator is an unbiased estimator which possess a minimum variance among all unbiased estimators. Likeš (1969) initially proposed the minimum variance unbiased estimators for estimating the Power Law exponent. The MVU estimator can be defined as

$$\hat{\beta}^{MVE} = \left(1 - \frac{2}{n}\right) \hat{\beta}^{MLE} \quad (3.9)$$

where $\hat{\beta}^{MLE}$ has been define above.

3.3 Goodness of Fit

The Goodness of fit tests are used to check whether the empirical distribution of variable follow the theoretical. In literature there are many tests that are used to check the goodness of fit of the data. Kolmogorov-Smirnov (KS) test, Anderson-Darling (AD) test and Chi square are the mostly used tests. To test how well Power Law distribution fits our observed data, we have performed KS goodness of fit test.

3.3.1 Kolmogorov-Smirnov test

Kolmogorov-Smirnov (KS) test is a nonparametric and is one of the most commonly used goodness of fit test. Let X_1, X_2, \dots, X_n be i.i.d observations from some unknown distribution with cdf $S(x)$. We want to test the hypothesis that $S(x)$ comes from some specified distribution with distribution function $S^*(x)$. We have the following hypothesis

$$H_0 : S(x) = S^*(x), \text{ vs } H_1 : S(x) \neq S^*(x)$$

where, S_x is the theoretical and $S^*(x)$ is the empirical distribution functions. As we know that Kolmogorov-Smirnov test measures the distance between the theoretical distribution function $S(x)$ and the empirical distribution function $S^*(x)$. The KS test statistic is defined as.

$$K = \sup_x |S^*(x) - S(x)| \quad (3.10)$$

We reject the null hypothesis when the test statistics K is greater than the critical value.

3.4 Bootstrap Investigation of Pareto Index Using Different Estimators

We have discussed in the previous sections that OLS, MOLS and Hill estimators are biased in the small sample. In order to check the small properties of these estimators, we adopt bootstrap simulation. By using bootstrap simulation, we compare the above estimators on the basis of their Biases and Precisions.

3.4.1 Simulation Design

We assume that the considered data sets follow Power Law distribution. Samples of different sizes (20, 50, 100 and 150) from the city size data for different countries have been selected and estimated the Power Law exponent by using different estimation techniques for each sample of given size. This process is repeated 1000 times and the results have then been

averaged.

Here we consider two criteria to evaluate the properties of different estimators. The first criteria we have considered is the Percentage relative bias defined given as

$$PRB = \left(\frac{\frac{1}{R} \sum_{r=1}^R \hat{\theta}_r - \theta}{\theta} \right) \times 100 \quad (3.11)$$

$$PRB = \left(\frac{\hat{\theta}_r - \theta}{\theta} \right) \times 100 \quad (3.12)$$

where, θ denotes the true value of parameter and $\hat{\theta}_r$ is its estimated value on the basis of "r" replications ($r = 1, 2, 3, \dots, R$).

The second criteria to evaluate the performance of those proposed estimators are to find MSE,

$$MSE = \frac{1}{R} \sum_{i=1}^R (\hat{\theta}_r - \theta)^2 \quad (3.13)$$

3.5 Elasticity of Power Law Exponent with Respect to Truncation Point and Sample Size using Recursive Sampling

There are various criteria in literature to check the validity of Zipf's Law. To analyze Zipf's exponent by changing the sample size and changing truncation point, [Gabaix \(1999\)](#), [Gabaix \(1999\)](#) have discussed theoretically explanation of variation pattern in the Power Law exponent.

It is known that small sample containing big cities results in higher value of the Power Law exponent $\alpha > 1$, while large sample containing smaller cities yield lower value of Power Law exponent $\alpha < 1$. In other words Pareto exponent greater than unity indicates that the second largest city is more than half of largest city and third largest city is more than one

third of largest city and so on. Similarly if the Power Law exponent is less than unity then the second largest city is smaller than half of largest city and third largest city is less than the third largest city and so on.

We have used rolling sample regression and have repeated the estimation process with moving down truncation point i.e. for the first sample we have a specific truncation point (number of cities) and for the second sample we have a specific number of cities and so on. We have used five estimation methods (OLS estimator, MOLS estimator, HILL estimator, ML estimator, MVU estimator) to estimate the Power Law exponent for each sub sample. The starting point is the largest city of the ordered data and the first sample contains first ten largest cities which is fixed arbitrarily i.e. $n_1 = 10$ and the second sub sample is $n_2 = n_1 + 1$ and so on, we continue this process until we consider the full sample size. Hence we get estimates of the Pareto exponent. We estimate the Power Law exponent for each sub sample and plot estimate versus sub sample size. From the plots, we can point out the sample for which the value of the Power Law exponent is exactly equal to or greater than or less than unity. The main advantage of this method is that it captures the variation in the Power Law exponent which may vary either due to the sample size or due to the truncation point or both.

3.5.1 Rolling Sampling Using With Replacement Sampling

In rolling sampling, because of the ordered pattern of the sizes of the cities, the value of the Pareto exponent is over estimated for the most populated cities and under estimate for less populated cities. Rolling sampling gives the overall variation in the Power Law exponent which may be due to the sample size and/or truncation point. To separate these simultaneous effects, we will use rolling random sampling with replacement. For rolling random sampling, the data should not be in ordered form. We initially fix the size of sub-

sample arbitrarily and randomly select first sub-sample of size n , rank the observations and apply the estimation method to estimate the Pareto index. Again we draw a second sub-sample which is independent of the first sub-sample, rank the observations and estimate the Power Law exponent and so on. This process is repeated until the last sub-sample of the full sample size is selected. Since we adopt a random process therefore, we replicate each sample 100 times and obtain the Power Law for each sub-sample. Finally, we plot these estimates versus sub-sample sizes and observe the variation in the Power Law exponent.

3.6 A Nonparametric Analysis of Power Law Exponent

Several studies have been done in literature to observe the behavior of the Power Law exponent using the nonparametric methodologies. [Giesen et al. \(2010\)](#), by using non parametric analysis, concluded that city size data, at national and regional level, follows Zipf's Law. [Ioannides and Overman \(2003\)](#) considered data for metropolitan area of U.S Census for the time period 1900 to 1990 to test the validity of Zipf's Law for cities size. They have used the nonparametric method to obtain the estimate of the Power Law exponent from the mean and variance of the growth rate.

In this study, we check the distribution of the Power Law exponent by performing the nonparametric method by kernel density plots using the five considered estimators. The kernel density method of estimation is widely used nonparametric method for estimating the probability density function of a random variable. Through kernel density plots, we observe the mode of the distribution of Power Law exponent which will reflect the variation in the Power Law exponent.

CHAPTER 4

Empirical Evidence of Power Law distribution

Zipf's Law is the great interest in urban research. The term Rank size rule have been noted by many number of researchers [[Rosen and Resnick \(1980\)](#), [Alperovich \(1984\)](#), [Gabaix \(1999\)](#), [Ioannides and Overman \(2003\)](#), [Soo \(2005\)](#), [Gan et al. \(2006\)](#) etc] but a valuable empirical studies are done by Zipf's. This study is also about validity of Zipf's law and our aim is to provide the empirical evidence of Power law distribution for the developing and the developed countries and check the exponent of the Power Law distribution to its universal value 1. We have taken the city size data from the U.S, China, Pakistan and India for the Census 1990 and 2000, 2005 and 2010, 1981 and 1998 and 2001 and 2011 respectively. As we know that Power Law distribution states that there exist inverse relationship between the city sizes and there rank in a region or in a country. [Auerbach \(1913\)](#) first found this rank size relationship among the city size and their ranks. Since it can be seen from the graph of the city size data that the sizes of the city decline more rapidly and yield a graph of L-Shape (Highly Skewed nature). This highly skewed pattern can well be approximated with the help

of the Power law distribution. To see the decline slope more clearly the city size is plot at the double logarithmic scale. From the log- log plots it can be seen that the line is bent upward and yield a straight line in the bottom right and the bottom of the log-log graph is a place where more city sizes lied for the same rank. More recently it is found that many natural phenomenon's like the intensity of the earth quakes, number of family sir names etc. follow Power Law distribution.

This chapter includes seven sections. The description about real life data set is given in the first section. Section 2 describes the descriptive statistics obtain from real life data sets of various countries. In section 3 we present graphical representation including simple plots and log-log plots for the data sets. Section 4 deals with goodness of fit test results using real life data. In fifth section we provide different threshold we obtained for the various countries based on the K.S statistics. The results obtained from real life data sets are describe in sixth section. Simulation study for the estimation of Power Law exponent is conducted in seventh section and last section include some concluding remarks.

4.0.1 Data

To demonstrate the existence of the Zipf's law, we used real life data sets for the developing and developed countries. Firstly we have chosen the US data of 1990 and 2000 and China data set of 2005 and 2010. The following data sets are for the developed countries. The United States and China has many cities so these data sets are ideal for city-size distribution analysis. Second data sets are data on the city size also taken from the developing country Pakistan 1981 and 1988 and India 2001 and 2010.

We have obtained the above data sets for the above mentioned countries form [Brinkhoff \(2008\)](#). This site contains the record of the city population for more than 100 countries. The

data available in the web site is based upon the administratively defined cities. [Soo \(2005\)](#) also used the city size data from the above web site for the 29 countries and to check the reliability of the data set he cross checked with the official statistics of each country, from UN Demographic Year book, Statistical agencies and conclude that the data are very similar.

4.0.2 Descriptive Statistics

In this section we provide the descriptive statistics for developed countries U.S Census data for 1990 and 2000, China Census data set for 2005 and 2010 and developing countries Pakistan census data set for 1981 and 1998 and India census data set for 2001 and 2011.

Table 4.1 Descriptive Statistics for Developed Countries (U.S, China)

Country	Year	Sample size	Mean	SD	Median	Minimum	Maximum
U.S	1990	396	399643.6	1163804	114939	50066	1 6044012
U.S	2000	452	425649.8	1246796	117465	50058	17799861
China	2005	284	737370.2	1412520	350000	158800	17784200
China	2010	139	1666414	2615784	905200	550200	23019100

Table 4.1 shows the descriptive statistic for the U.S data 1990 and 2000. The numbers of cities in U.S are 396 and 452 for the census period 1990 and 2000 indicating here that there is 14.14% increase in number of cities. The average city size for 1990 is 399643 persons and for 2000 the mean value is 425649 persons. Hence there is 6.50% increase in the mean city size which is moderate increase in the city size. The minimum city size for the 1990 and 2000 are 50066 and 50058 inhabitants respectively for the US and the maximum city size are 16044012 and 17799861 inhabitants respectively. It is apparent from the 4.1 table that the mean is greater than median which indicate the positive skewness of the US data. Similarly for China data sets for census 2005 and 2010 the numbers of cities are decrease by 51% indicating that there are more migration in these time period and the number of small

cities are more. The average city size is increased by 125% indicate that there is high density of population in the urban areas.

Table 4.2 Descriptive Statistics for Developing Countries (Pakistan, India)

Country	Year	Sample size	Mean	SD	Median	Minimum	Maximum
Pakistan	1981	154	137284.8	496441.2	37844	5208132	20386
Pakistan	1998	244	161005.2	699342	45908.5	9269265	23653
India	2001	284	647300.4	1608807	293474.5	16434386	124245
India	2011	314	783987	1847788	243279	18394912	150019

Table 4.2 shows the descriptive statistic for the Pakistan data 1981 and 1998. The numbers of cities in Pakistan are 154 and 244 for the census period 1981 and 1998 indicating here that there is 58.44% increase in number of cities. The average city size for 1981 and 1998 are 137284 and 161005 persons respectively. Hence there is 17.27% increase in the mean city size which is moderate increase in the city size. The maximum city size for the 1981 and 1998 are 5208132 and 9269265 persons respectively for Pakistan and the minimum city size are 20386 and 23653 persons respectively. It is apparent from the above table that the mean is greater than median which indicate the positive skewness of the US data. Similarly for India, there is an increase in number of cities and city sizes are 10.56% and 21.11%. For both countries mean is greater than median which shows the positive skewness of the city size data.

4.1 Minimum Threshold

Many researchers have found Lognormal distribution fit well for city size data when city size data is considered as a whole e.g see [Anderson and Ge (2005), Gibrat (1931)]. However Power Law distribution well approximate the city size data above some predefined threshold value called x_{min} . To fix this minimum threshold value, different criteria's are discussed in

literature. In this study, we find the minimum threshold value following [Clauset et al. \(2009\)](#). Table 4.3 and 4.4 show the total number of cities and the number of cities after truncation for different census years. For the U.S city size data, the minimum threshold values are 5006 and 50058 inhabitants for the year 1990 and 2000, respectively. The total number of cities in complete data sets and number of cities in the truncated data sets are equal. For china, the minimum threshold values are quite high i.e. 158800 and 550000 inhabitants for the year 2005 and 2010, respectively. After the truncation, we are left with 457 and 139 cities, respectively. In 2010, irrespective of the fact that number of cities is increased, the high value of threshold results in small number of cities i.e. 139.

For Pakistan, the minimum threshold values are 20386, 23653 inhabitants for the years 1988 and 1996, respectively. The numbers of cities in the truncated data sets are 153 and 244, respectively. In Indian data sets, the minimum threshold values are 97011 and 150019 inhabitants for the year 2001 and 2011, respectively. The numbers of cities in the truncated data sets are 284 and 314, respectively.

Table 4.3 Number of Cities After and Before Truncation

Country	Year	Number of cities	X_{min}	Number of cities $\geq X_{min}$
U.S	1990	396	5006	396
U.S	2000	452	50058	452
China	2005	659	158800	457
China	2010	655	550000	139

Table 4.4 Number of Cities After and Before Truncation

Country	Year	Number of cities	X_{min}	Number of cities $\geq X_{min}$
Pakistan	1981	164	20386	153
Pakistan	1998	280	23653	244
India	2001	321	124245	284
India	2011	322	150019	314

4.2 Graphical Representations

In order to check the deviation of data from the Power Law distribution, we present simple plots and the log-log plots in this section. We are to examine the validity of the Power Law distribution as well as the validity of the Zipf's Law graphically.

As we know that the plot of city sizes (sorted) yield us L shaped curve indicating that a bulk of cities population lie in the tail of the graph. This heavy tailed graph is similar to the graph of the Power Law distribution. In order to check the slope of the graph more clearly, we will plot the rank and city sizes on the double log-log graph. The value of the slope equal to -1 confirms the validity of the Zipf's Law.

4.2.1 Graphical Representation for US and China

The following graphs are plotted using the data sets of the developed countries. The logarithmic of city sizes are scaled on the horizontal axis while the logarithmic of ranks are taken on the vertical axis.

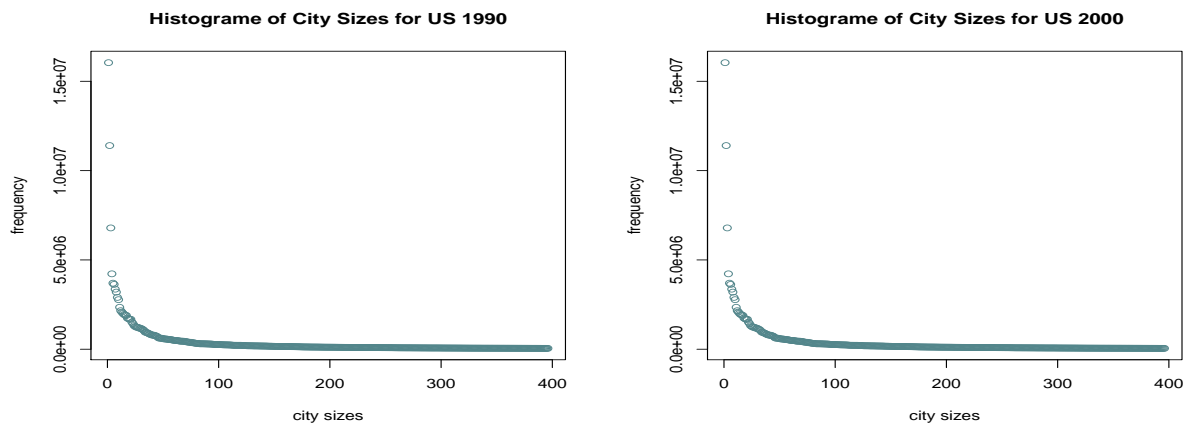


Figure 4.1 plots of US Data (1990 and 2000)

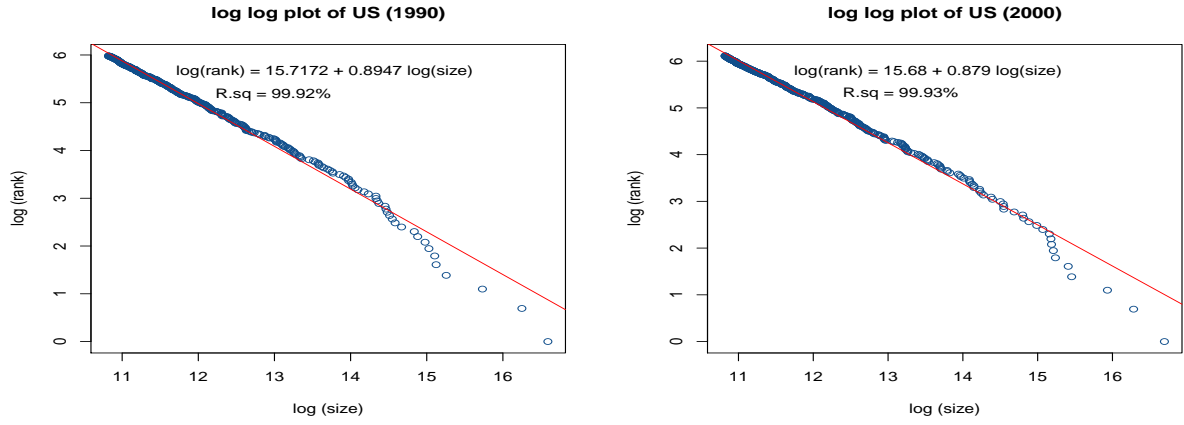


Figure 4.2 log log plots of US Data (1990 and 2000)

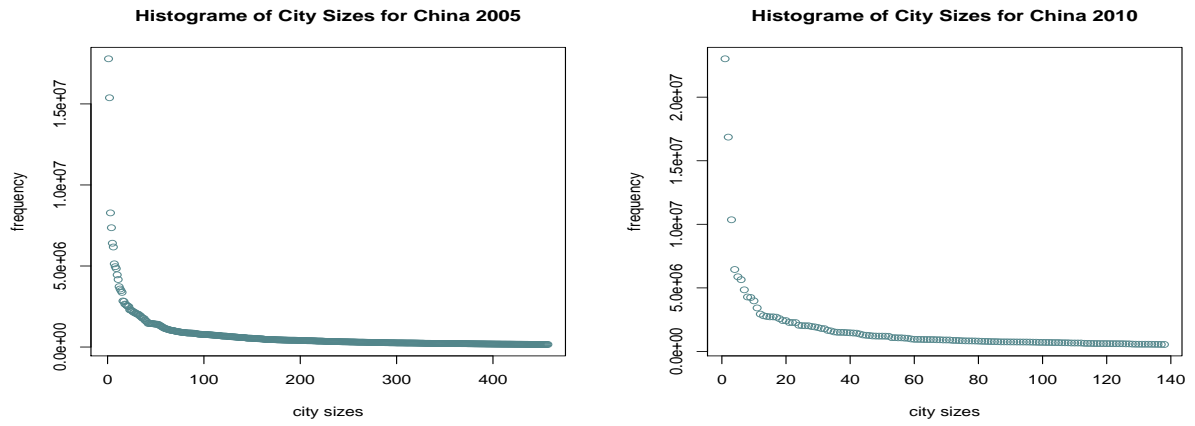


Figure 4.3 plots of China Data (2005 and 2010)

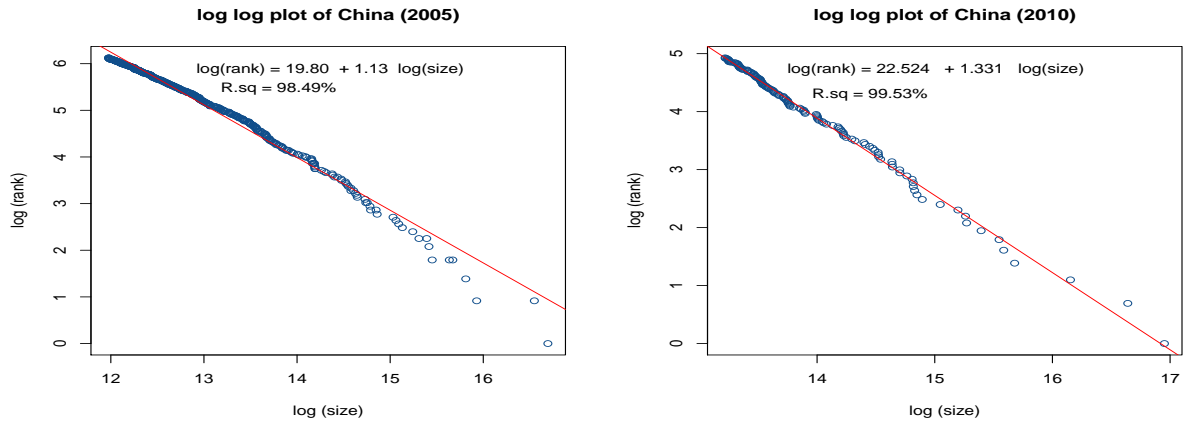


Figure 4.4 log log plots of China Data (2005 and 2010)

Plots shown in figure 4.1, 4.2, 4.3 and 4.4 are for both U.S data and China data sets are L shaped and resembles the Power Law distribution curve. The Log-Log plots for both countries data sets are straight lined graphs which indicate presence of Zipf's Law. For these two country data sets the deviation from Power Law can be observe for the top city size since the corresponding points are far from the straight line on the log-log plot.

4.2.2 Graphical Representation for the Pakistan and India

In order to check the validity of Power Law distribution graphically, we present simple plots of city size and log-log plots for the city size data of developing countries Pakistan and India. The simple plots seem near to the Power Law behavior of the both country data sets. While the log-log plots for the Pakistan and India conforms the validity of the Zipf's Law because the slope of the log-log plots are approximately equal to -1.

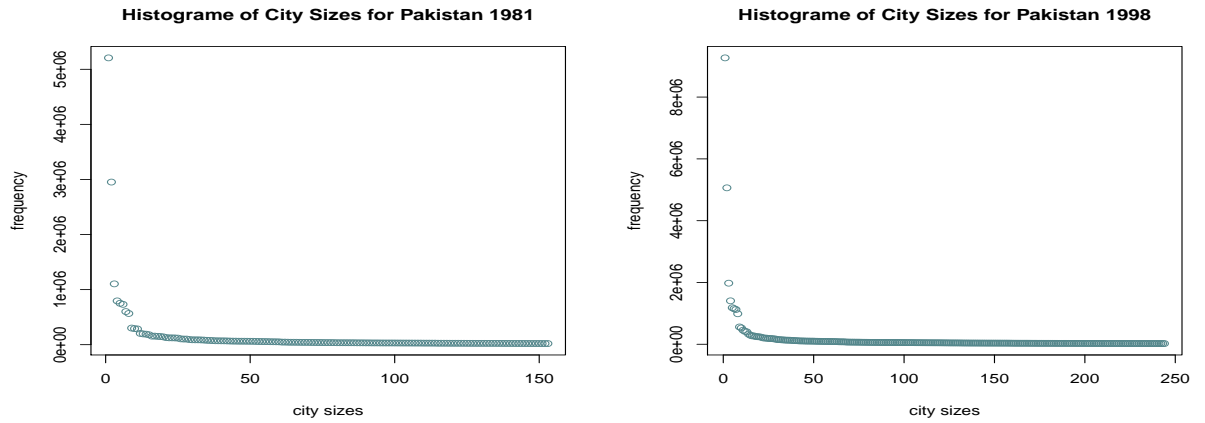


Figure 4.5 plots of Pakistan Data (1981 and 1998)

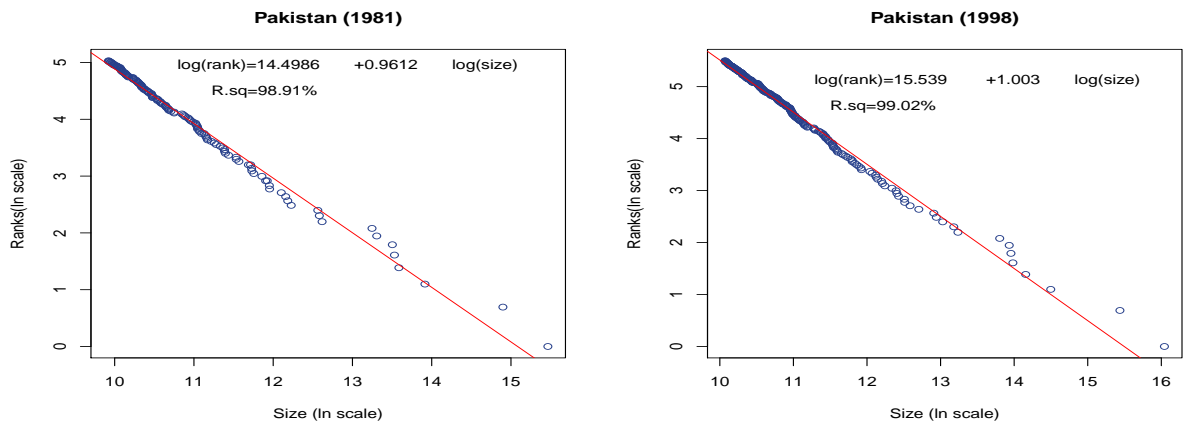


Figure 4.6 log log plots of Pakistan Data (1981 and 1998)

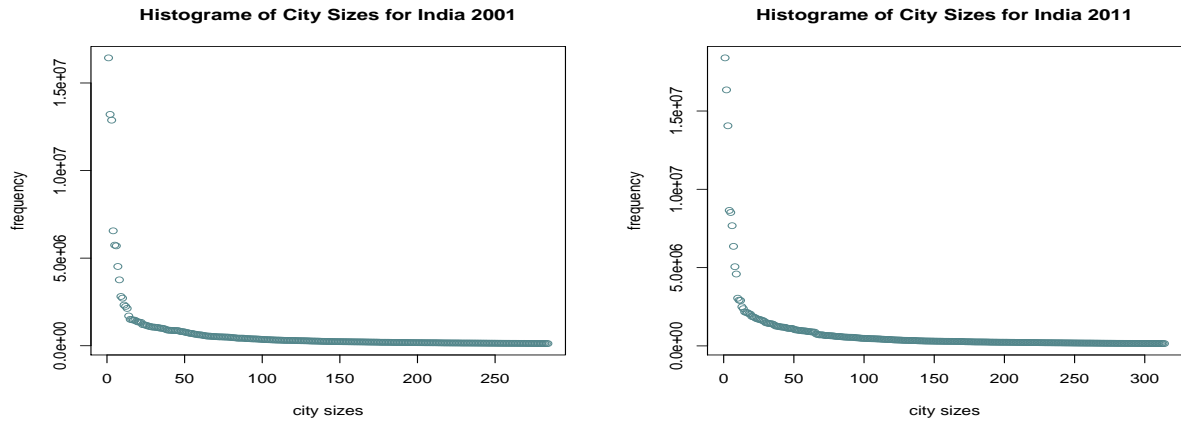


Figure 4.7 simple Plots for India (2001 and 2011)

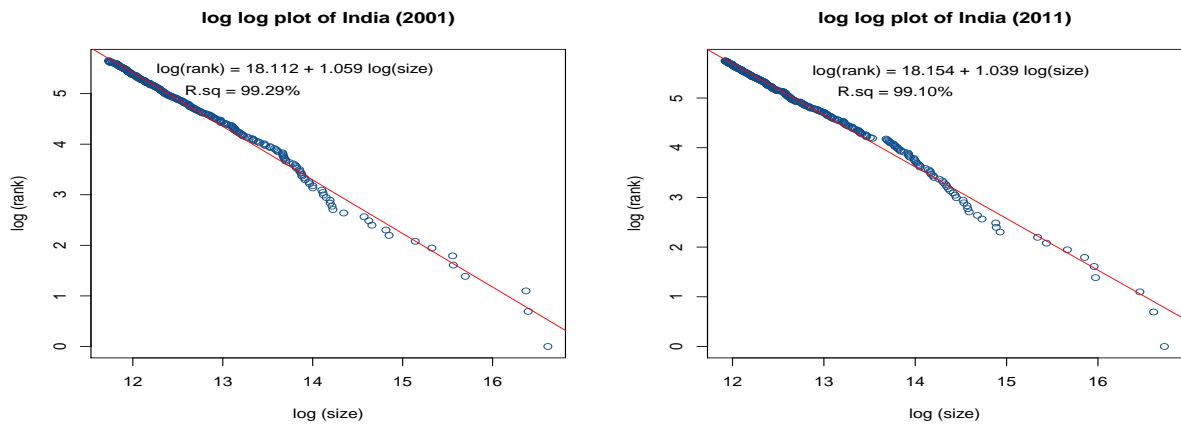


Figure 4.8 log log plots of India (2001 and 2011)

4.3 Goodness of fit

Goodness of fit tests are commonly used to investigate whether the empirical distribution of the data coincides with the theoretical distribution. Formally, conducting the goodness of fit test is very important to check the validity of the Power law distribution. The value of the Power Law exponent is used to conclude that whether the data follows Zipf's Law. But before estimating the value of the Power Law exponent it is important to check the goodness of fit of underline data in our case to check whether the data follows Power Law distribution. We use Kolmogorov-Smirnov (KS) test to check the goodness of fit of the data sets for the

developing and developed countries.

4.3.1 Goodness of fit test for US and India

The values of K-S test for the developed countries i.e. U.S and China are shown in the following table.

Table 4.5 Goodness of fit for U.S and China

Country	Year	Test	Sample size	Statistics	Critical value	Alpha	Conclusion
U.S	1990	K.S	396	0.02	0.068	0.05	Accept Ho
U.S	2000	K.S	452	0.03	0.063	0.05	Accept Ho
China	2005	K.S	457	0.05	0.063	0.05	Accept Ho
China	2010	K.S	139	0.04	0.115	0.05	Accept Ho

From these results we observe that, our null hypotheses is not rejected at 5% level of significance. It is therefore, concluded that both the country data sets follow Power Law distribution.

4.3.2 Goodness of fit test for Pakistan and India

The following table consists of the KS test results for the data sets of the developing countries, i.e. Pakistan and India.

Table 4.6 Goodness of fit for Pakistan and India

Country	Year	Test	Sample size	Statistics	Critical value	Alpha	Conclusion
Pakistan	1981	K.S	153	0.037	0.109	0.05	Accept Ho
Pakistan	1998	K.S	244	0.025	0.086	0.05	Accept Ho
India	2001	K.S	284	0.030	0.080	0.05	Accept Ho
India	2011	K.S	314	0.042	0.076	0.05	Accept Ho

It is apparent from table 4.6 that we cannot reject any of our null hypotheses at 5% significance level. This means that both the countries data sets have Pareto nature.

4.4 Estimate's of Power law Exponent

After confirming that the considered data sets follow Pareto distribution (Power Law), we move towards the estimation of the shape parameter of the Power Law distribution. The exponent of Power Law distribution measures the degree of inequality of the population among the cities. In literature, different estimation methods are used. Most of the common methods is OLS, used by many authors [[Alperovich \(1984\)](#), [Gan et al. \(2006\)](#) and others]. [Moura Jr and Ribeiro \(2006\)](#) used MLE method as the estimation method. [Soo \(2005\)](#) have used OLS and HILL estimator for the estimation purpose. [Moura Jr and Ribeiro \(2006\)](#) used three estimation methods, namely OLS, MLE and Parameter averaging method to estimate the Power Law exponent. Following [Terra \(2009\)](#), we use five estimation methods i.e. OLS, MOLS, ML, HILL and MVU.

4.4.1 Estimates of Power Law Exponent for US and China

We report the estimates of Power Law exponent for US data set 1990, 2000 and China data set 2005 and 2010 in table [4.7](#).

Table 4.7 Estimates of the Power Law Exponent for US and China

Country	Year	OLS Est	MOLS Est	ML Est	HILL Est	MVU Est
U.S	1990	0.895	0.913	0.856	0.854	0.853
U.S	2000	0.879	0.896	0.840	0.838	0.836
China	2005	1.13	1.151	0.990	0.988	0.986
China	2010	1.331	1.396	1.374	1.355	1.354

The results in [4.7](#) emphasize that the Power Law exponent for U.S data sets lie in the range (0.85, 0.91) and (0.836, 0.896) for the years 1990 and 2000, respectively. The MOLS estimate, for both the years, is closer to unity. So we conclude from the above results that Zipf's Law does not hold for both the US data sets. From the results of above table , it is also

apparent that the estimated value, using OLS and MOLS of the Power Law exponent for the year 2005 is greater than 1. On the other hand, values for the other three estimation methods are less than 1 with negligible difference from 1. For the data set of the year 2010, all the five methods give values greater than 1. So we conclude that Zipf's Law does not hold for both the data sets, except for the year 2005 by using the ML, HILL and MVU estimation methods.

4.4.2 Estimates of Power Law Exponent for Pakistan and India

We report the estimates of Power Law exponent for Pakistan and India for the Census period 1981, 1998 and 2001, 2011 respectively in the following table.

Table 4.8 Estimates of the Power Law Exponent for Pakistan and India

Country	Year	OLS Est	MOLS Est	ML Est	HILL Est	MVU Est
Pakistan	1981	0.961	1.007	1.059	1.052	1.045
Pakistan	1998	1.003	1.039	1.082	1.077	1.073
India	2001	1.059	1.089	1.025	1.022	1.018
India	2011	1.039	1.065	1.024	1.021	1.018

It is emphasized from table 4.8 that the Zipf's law holds for both the data sets of Pakistan using all of the estimation methods except under OLS for 1981. It can also be seen from the above table that the estimated value of the Power Law exponent using all the estimation methods is exactly 1. This indicates that Zipf's Law holds for both the data sets of India.

4.5 Simulation study

In bootstrap simulation we assume given data set as a population with unknown distribution. We compute only one statistic from data set, assume the statistic as true population parameter. In bootstrap simulation, we generate a large number of data sets, compute statistics for each data set. Thus we get distribution of the statistics whose sample value should be

equal to the true parametric value for large sample sizes. Bootstrap simulation allows us how a statistic varies from sample to sample and with increasing number of re-samples and how the sampling distribution evolves over time with increasing number of sample or re-sample observations. In this section, we have considered the bootstrap simulation design to study the behavior of the Power Law exponent for different sample sizes and evaluate the performance of estimation methods based upon their PRB and MSE.

4.5.1 Simulation Results for the US and China

We have simulated sample of sizes 20, 50, 100 and 150 from the data set of US for the year 1990 by assuming the true value of the Power Law exponent as 0.895, 0.913, 0.854, 0.857 and 0.853 for OLS, MOLS, HILL, ML and MVU estimation methods respectively. For US data for the year 2000, the true parametric values are 0.879, 0.896, 0.838, 0.840 and 0.836, for the estimation methods OLS, MOLS, HILL, ML and MVU respectively. The process is repeated 1000 times and then averaged the results. The results of the simulation study are reported in table 4.9.

Table 4.9 Simulation Results for U.S

Estimator	Year 2000			Year 1990		
	Estimate	PRB	MSE	Estimate	PRB	MSE
n=20						
HILL	0.915	7.131	0.049	0.863	2.996	0.038
OLS	0.797	6.708	0.042	0.778	11.507	0.050
MOLS	0.926	-8.419	0.056	0.902	-0.634	0.049
ML	0.883	3.385	0.039	0.879	4.682	0.039
MVU	0.795	-6.836	0.034	0.791	-5.310	0.032
n=50						
HILL	0.872	2.090	0.015	0.850	-0.411	0.013
OLS	0.833	2.462	0.014	0.816	4.405	0.015
MOLS	0.904	-5.860	0.020	0.888	-3.939	0.017
MLE	0.872	2.070	0.015	0.855	0.155	0.013
MVU	0.837	-1.896	0.014	0.821	-3.734	0.013
n=100						
HILL	0.865	1.315	0.006	0.849	1.334	0.012
OLS	0.862	-0.986	0.007	0.849	3.443	0.008
MOLS	0.909	-6.445	0.011	0.895	0.160	0.007
ML	0.864	1.138	0.006	0.846	0.738	0.006
MVU	0.846	-0.768	0.006	0.829	-0.800	0.006
n=150						
HILL	0.861	0.787	0.005	0.842	0.455	0.004
OLS	0.876	-2.537	0.005	0.858	2.419	0.005
MOLS	0.906	-6.138	0.008	0.893	0.361	0.005
ML	0.863	1.014	0.005	0.844	0.515	0.004
MVU	0.851	-0.216	0.004	0.833	-0.349	0.004

From the results given in table 4.9 for Years 1990 and 2000, we conclude that, for large sample size $n=150$, MVU estimator out performs as compared to all the other estimator as it possesses minimum PRB and MSE. For US 1990 data set MVU estimator contains PRB (0.092%) and MSE (0.004) and for data 2000 the PRB and MSE are 0.010%, 0.004 respectively.

We have simulated sample of sizes 20, 50, 100 and 150 from the data set of China for the year 2005 assuming the true value of the Power Law exponent as 1.13, 1.151, 0.988, 0.990

and 0.986 for OLS, MOLS, HILL, ML and MVU estimation methods respectively. For China data for the year 2010, the true parametric values are 1.331, 1.396, 1.355, 1.374 and 1.354, for the estimation methods OLS, MOLS, HILL, ML and MVU respectively.

Table 4.10 Simulation Results for China

Estimator	Year 2005			Year 2010		
	Estimate	PRB	MSE	Estimate	PRB	MSE
OLS						
20	0.977	13.523	0.071	1.223	8.098	0.126
50	1.045	7.526	0.027	1.285	3.489	0.048
100	1.082	4.209	0.013	1.319	0.927	0.023
150	1.100	2.641	0.008	1.345	-1.050	0.017
MOLS						
20	1.130	1.802	0.062	1.420	-1.721	0.150
50	1.135	1.385	0.023	1.399	-0.235	0.053
100	1.141	0.895	0.012	1.392	0.274	0.025
150	1.144	0.602	0.007	1.400	-0.305	0.018
HILL						
20	1.027	3.929	0.045	1.432	5.716	0.120
50	1.000	1.170	0.015	1.386	2.271	0.038
100	0.993	0.506	0.007	1.370	1.128	0.017
150	0.994	0.637	0.005	1.370	1.095	0.012
ML						
20	1.018	2.847	0.042	1.423	3.571	0.111
50	1.003	1.277	0.015	1.388	0.995	0.036
100	0.996	0.654	0.008	1.373	-0.089	0.017
150	0.997	0.706	0.005	1.373	-0.049	0.012
MVU						
20	0.916	-7.062	0.038	1.281	-5.409	0.093
50	0.963	-2.379	0.014	1.332	-1.613	0.033
100	0.977	-0.959	0.007	1.345	-0.641	0.016
150	0.984	-0.234	0.005	1.355	0.075	0.011

From table 4.10, we conclude that, for the year 2005, ML and MVU estimator possesses minimum MSE (0.005) as compared to the other estimators for large sample size (n=150) and the PRB of MVU estimator in minimum as compared to other estimators. Hence we conclude that MVU estimator is best for estimating the Power Law exponent. For the year 2010 MSE of MVU estimator is minimum compared to the other estimators.

4.5.2 Simulation Results for the Pakistan and India

We have simulated sample of sizes 20, 50, 100 and 150 from the data set of Pakistan for the year 1981 assuming the true value of the Power Law exponent as 0.961, 1.007, 1.052, 1.059 and 1.045 for OLS, MOLS, HILL, ML and MVU estimation methods, respectively. For the year 1998, the true parametric values are 1.003, 1.039, 1.077, 1.082 and 1.073, for the estimation methods OLS, MOLS, HILL, ML and MVU, respectively. The simulation results for the above data sets are presented in table [4.11](#).

Table 4.11 Simulation Results for Pakistan

S.size	Year 1981			Year 1998		
	Estimate	PRB	MSE	Estimate	PRB	MSE
OLS						
20	0.929	3.357	0.107	0.953	4.940	0.102
50	0.935	2.696	0.038	0.975	2.805	0.045
100	0.966	-0.527	0.019	0.992	1.136	0.022
150	0.972	-1.103	0.011	0.999	0.445	0.015
MOLS						
20	1.081	-7.375	0.144	1.111	-6.884	0.134
50	1.024	-1.651	0.044	1.068	-2.807	0.052
100	1.023	-1.589	0.021	1.052	-1.268	0.024
150	1.014	-0.698	0.012	1.044	-0.516	0.017
HILL						
20	1.127	7.099	0.089	1.129	4.793	0.077
50	1.080	2.648	0.026	1.103	2.407	0.029
100	1.068	1.541	0.012	1.091	1.257	0.014
150	1.062	0.983	0.008	1.086	0.824	0.009
ML						
20	1.124	6.120	0.084	1.140	5.330	0.079
50	1.086	2.528	0.026	1.112	2.741	0.030
100	1.073	1.366	0.012	1.096	1.339	0.014
150	1.067	0.738	0.008	1.090	0.777	0.009
MVU						
20	1.011	-3.212	0.065	1.026	-4.408	0.063
50	1.042	-0.254	0.024	1.067	-0.541	0.026
100	1.052	0.670	0.011	1.075	0.145	0.014
150	1.053	0.727	0.007	1.076	0.267	0.008

From table Table 4.11, we conclude that, for both the data sets of Pakistan, MVU estimator performs better as compared to the other estimators. The PRB and MSE of MSE is minimum for large sample size ($n = 150$) for both data sets.

Similarly we have simulated sample of sizes 20, 50, 100 and 150 from the data set of India for the year 2001 by assuming the true value of the Power Law exponent as 1.059, 1.089, 1.022, 1.025 and 1.018 for OLS, MOLS, HILL, ML and MVU estimation methods respectively. For the year 2011, the true parametric values are 1.039, 1.065, 1.021, 1.024 and 1.018, for the estimation methods OLS, MOLS, HILL, ML and MVU, respectively. The simulated results are contained in table Table 4.12.

Table 4.12 Simulation Results for India

Estimator	Year 2001			Year 2011		
	Estimate	PRB	MSE	Estimate	PRB	MSE
OLS						
20	0.958	9.503	0.075	0.937	9.784	0.070
50	0.995	6.050	0.031	0.967	6.960	0.028
100	1.024	3.320	0.014	1.010	2.768	0.011
150	1.044	1.417	0.008	1.020	1.852	0.007
MOLS						
20	1.112	-2.080	0.086	1.086	-2.004	0.078
50	1.084	0.421	0.031	1.052	1.184	0.027
100	1.082	0.638	0.014	1.067	-0.144	0.012
150	1.088	0.064	0.009	1.062	0.244	0.008
HILL						
20	1.090	6.654	0.072	1.079	5.647	0.068
50	1.046	2.312	0.021	1.036	1.442	0.020
100	1.030	0.823	0.009	1.038	1.634	0.010
150	1.027	0.504	0.006	1.028	0.676	0.006
ML						
20	1.083	5.623	0.064	1.079	5.379	0.065
50	1.046	2.028	0.020	1.037	1.280	0.019
100	1.033	0.748	0.009	1.038	1.405	0.010
150	1.029	0.359	0.006	1.028	0.434	0.006
MVU						
20	0.974	-4.286	0.051	0.971	-4.600	0.053
50	1.004	-1.380	0.018	0.996	-2.198	0.018
100	1.012	-0.588	0.008	1.018	-0.038	0.009
150	1.015	-0.298	0.006	1.015	-0.321	0.006

From table Table 4.12, we conclude that, for large sample ($n=150$), HILL, ML and MVU estimator possess minimum MSE for both the data sets (2001, 2011). But the PRB of MVU estimator is least as compared to all the other estimators. Hence, on the basis of of simulation study, we conclude here that MVU estimator is best for the estimation of the Power Law exponent.

CHAPTER 5

Rolling Sampling and Non-Parametric Analysis of Pareto Exponent

In previous chapter, we have used real data sets to examine the validity of Zipf's Law. In this chapter, using rolling sampling, we estimate the values of the exponent and check its elasticity with respect to the sample size and truncation point. We will also explore the range of the sample data for which Power Law exponent is exactly equal to unity (rank size rule holds), greater than unity and less than unity. We shall also study the behavior of the Power Law exponent non-Parametrically.

5.1 Rolling Sampling

In rolling sampling, the sample size changes with changing the truncation point. The basic logic of this sampling is that a constant coefficient, through this technique, will lead us to the conclusion that the Power Law exponent is one which guarantees the existence of Zipf's Law. The rolling sampling exhibits the variation pattern in the Power Law exponent with respect to the truncation point. Using this technique, one can extend his/her research to

the entire data distribution instead of the upper tail distribution only. All of these analysis will be made by graphical displays.

5.1.1 Rolling sample results for US and China

We have considered city size data of the U.S for the time period 1990 and 2000. For the U.S there are 396 and 452 cities for Census period 1990 and 2000, respectively. By using rolling sampling, the first sub sample contains first 10 largest cities, the second sub sample contains the first 11 largest cities and so on. The process is continued until the last sub sample contains 396 and 452 cities for the respective time periods. Hence we get 387 ($= 396 - 10 + 1$) estimated values of the Power Law exponent for the time period 1990 and 443 ($= 452 - 10 + 1$) estimated values of the Power Law exponent for 2000. The following graphs show the variation in the estimated value of the Power Law exponent for with respect to the truncation points for US data sets.

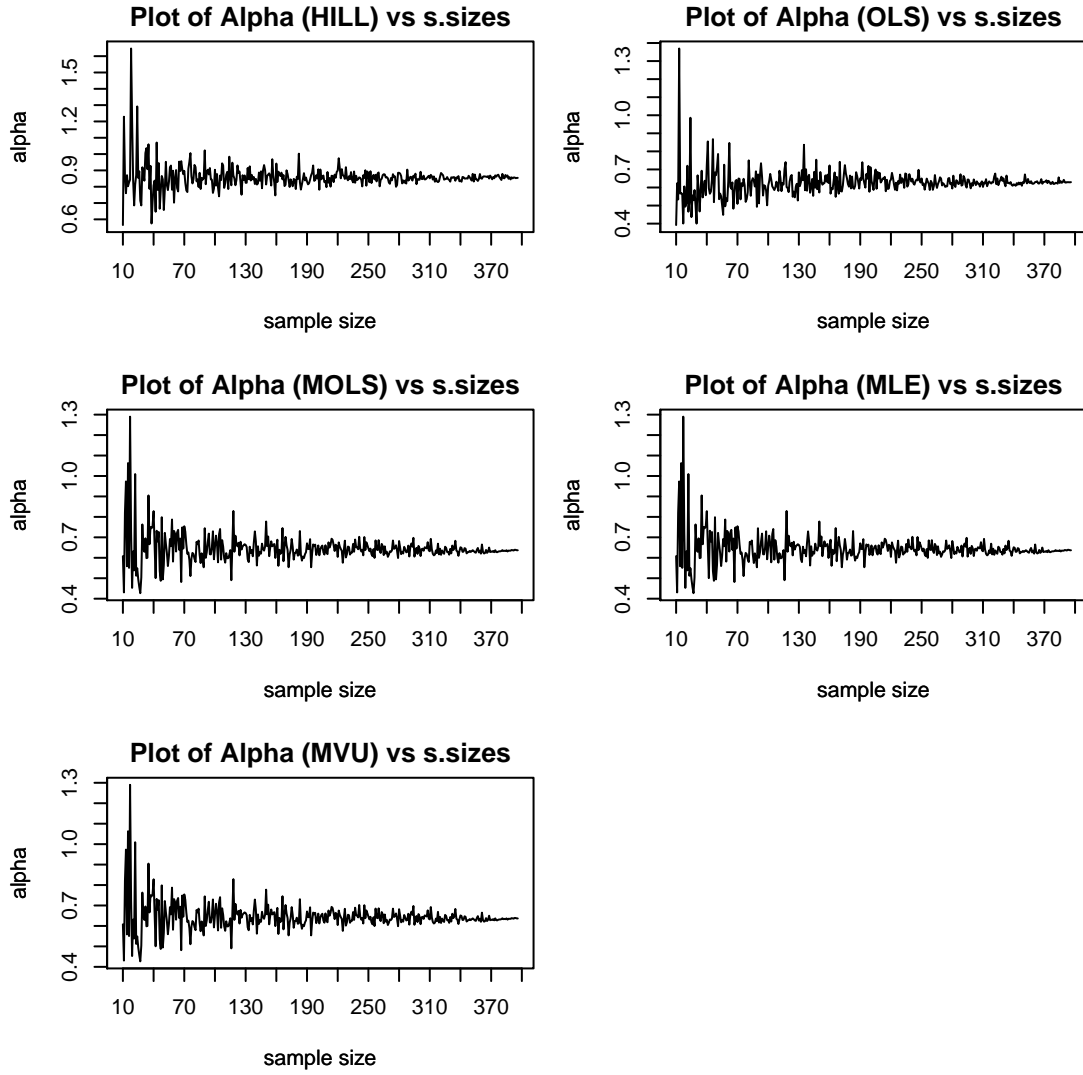


Figure 5.1 Rolling Sampling Plots for U.S 1990

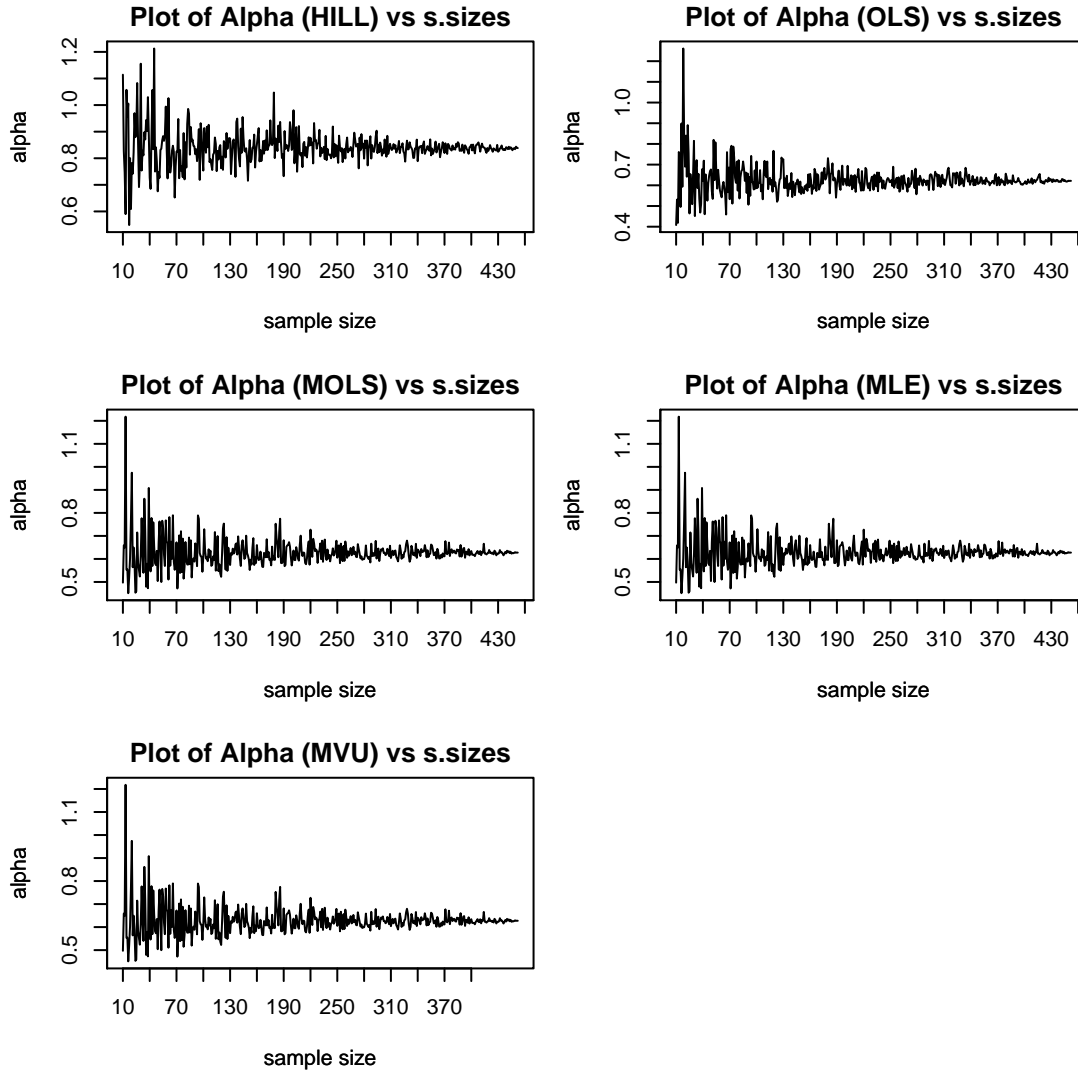


Figure 5.2 Rolling Sampling Plots for U.S 2000

Figures 5.1 and 5.2 show that the Rank size rule does not hold for both the US data sets. All the estimation methods under estimates the value of the Power Law exponent even in the large sample sizes. For both the data sets, the estimated value of the Power Law exponent moves toward the steady state as the sample contains cities greater than 100. The graphs for both the data sets of China are shown in the appendix A.

5.1.2 Rolling Sampling Results for Pakistan and India

The following graphs show the distribution of rank size rule for the city size data of Pakistan for the year 1981 and 1998.

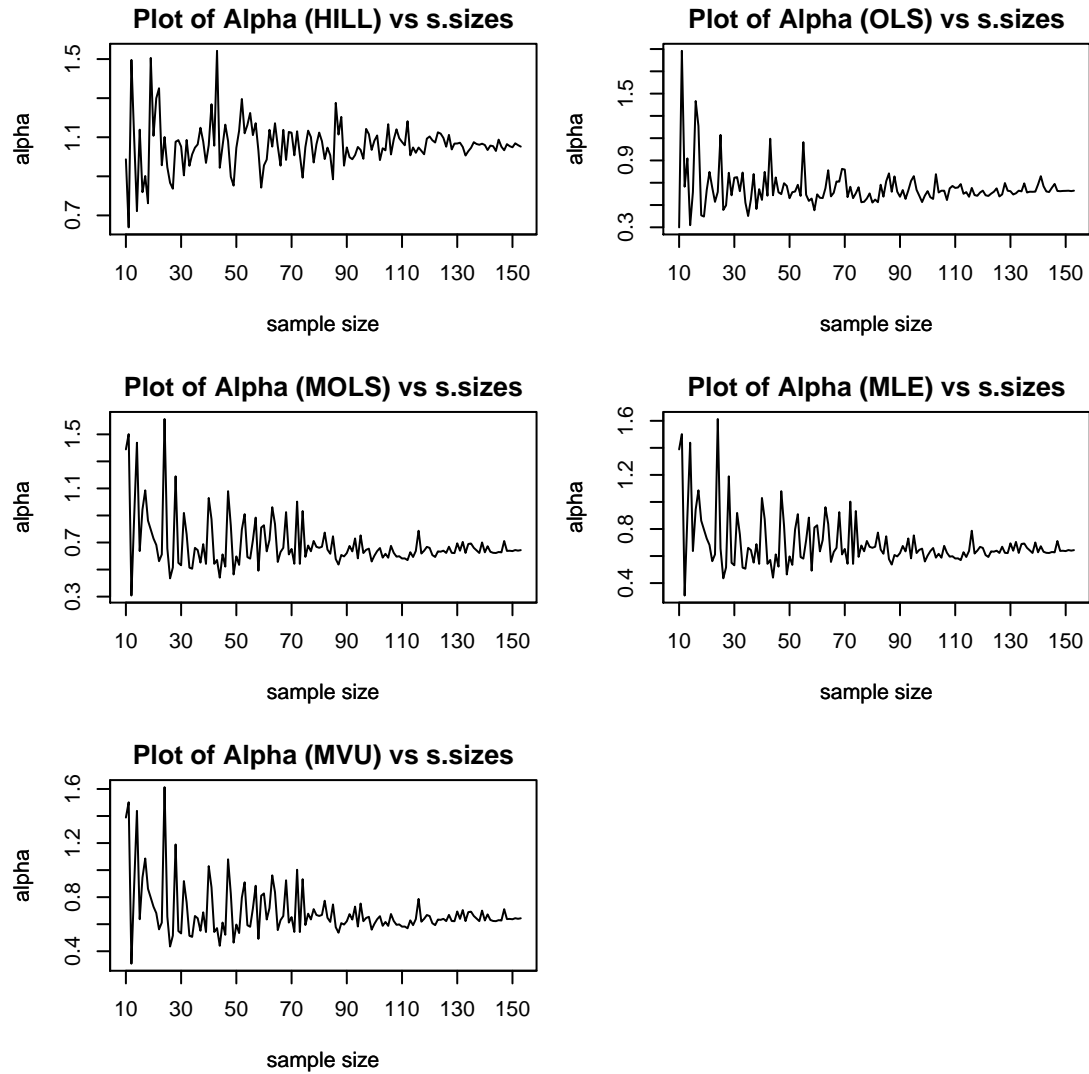


Figure 5.3 Rolling Sampling Plots for Pakistan 1981

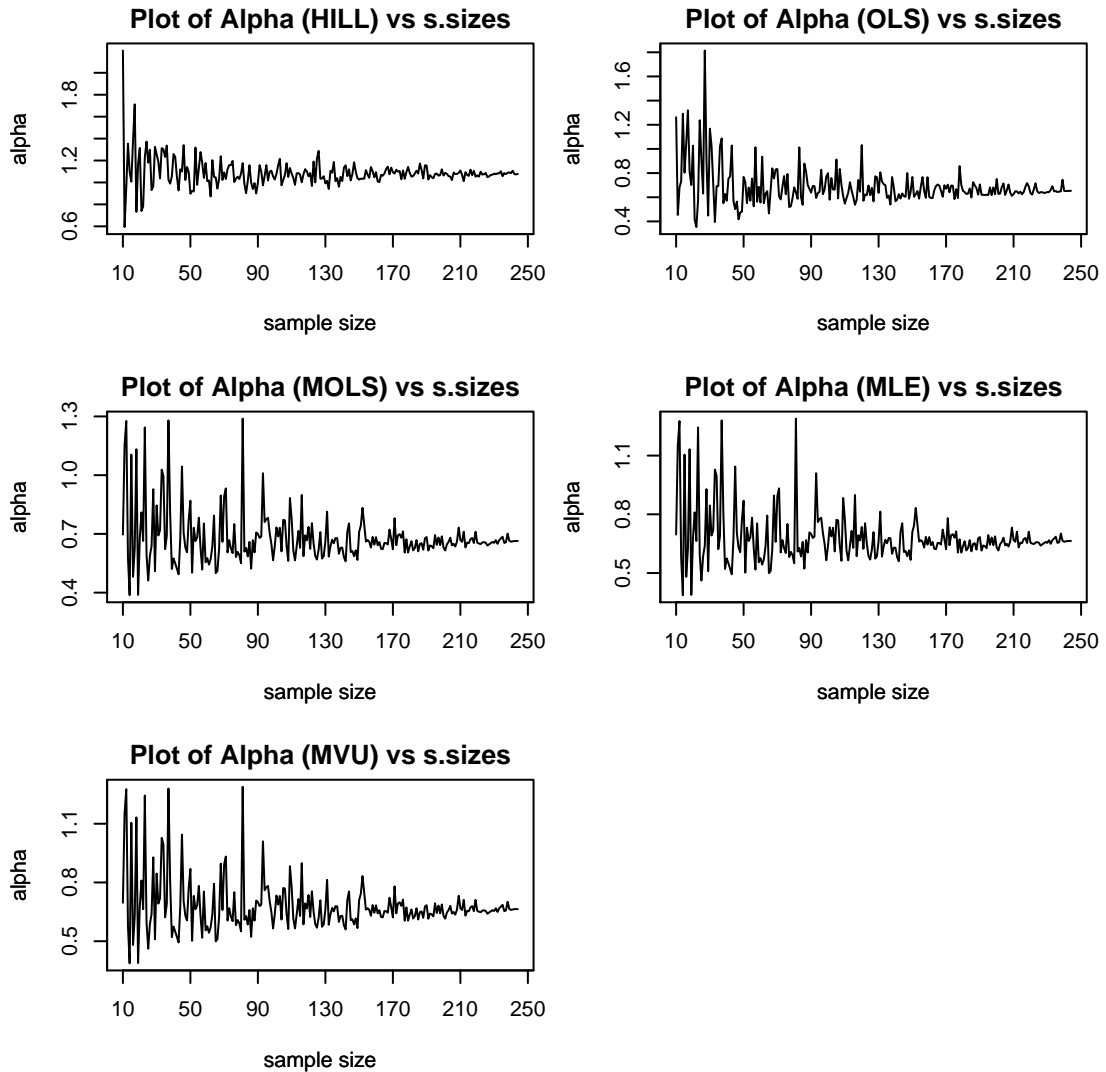


Figure 5.4 Rolling Plots for Pakistan 1998

From the graphs for the data set of 1981, we observe that the estimated value of Power Law exponent is equal to unity for the sample of sizes greater than 80 for the ML, HILL and MVU estimation methods. By using OLS and MOLS estimators, the value of the Power Law exponent is under estimated even in the large samples.

By using the Census data 1998, the value of the Power Law exponent is under estimated using OLS estimator. All the other estimation methods give the value of the exponent one for large sample size. The graphs for the data sets of India are given in Appendix B.

5.2 Rolling Sampling by Using With Replacement Sampling

The rolling sampling gives us the overall variation in the Power Law exponent that may be due to the sample size and/or due to the truncation point. To separate these simultaneous effects, we are going to use random Rolling sampling with replacement for which our data should not be in any ordered form. We initially fix the size of sub sample arbitrary and randomly select first sub sample of size $n_1 = 5$, rank the observations and apply the estimation method to estimate the Power Law exponent. Again draw a second sub sample of size $n_2 = n_1 + 1$ which is independent of the first sample. We estimate the Power Law exponent for the second sub sample. This process is continued until the last sub sample of the full sample size. Because we adopt a random process, so we replicate each sample 100 times and obtain the Power Law exponent for each replication and the result is then averaged. Finally, we plot these estimates versus sub sample sizes and observe the variation in the Power Law exponent.

5.2.1 Results of Rolling Random Sampling for US and China

In this section, we present the graphical representation of Rolling random sampling with replacement results based on the different estimation procedures for the developed countries U.S and China.

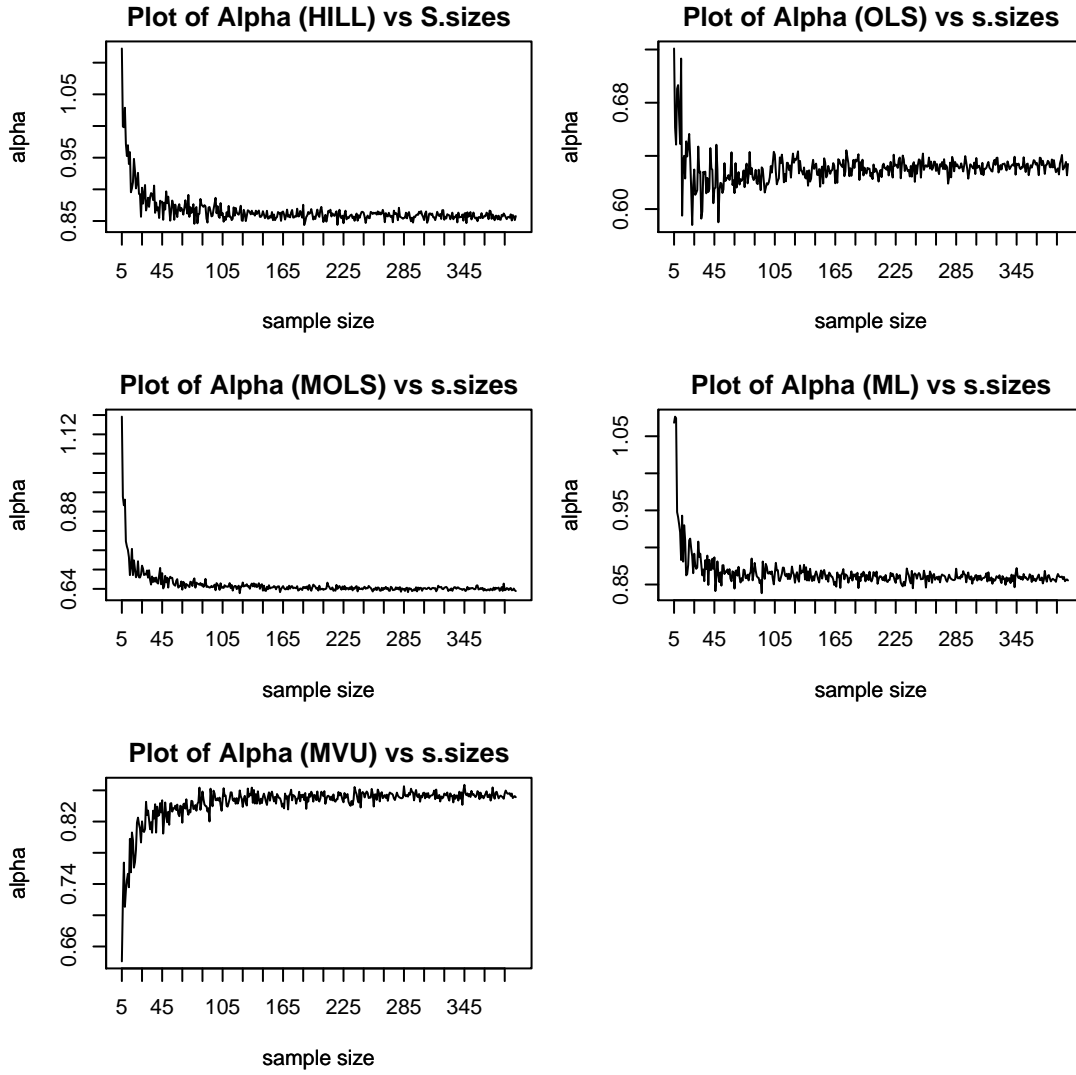


Figure 5.5 Rolling Random Sampling Plots for US (1990)

The following graphs are plotted for the US data set of the year 2000.

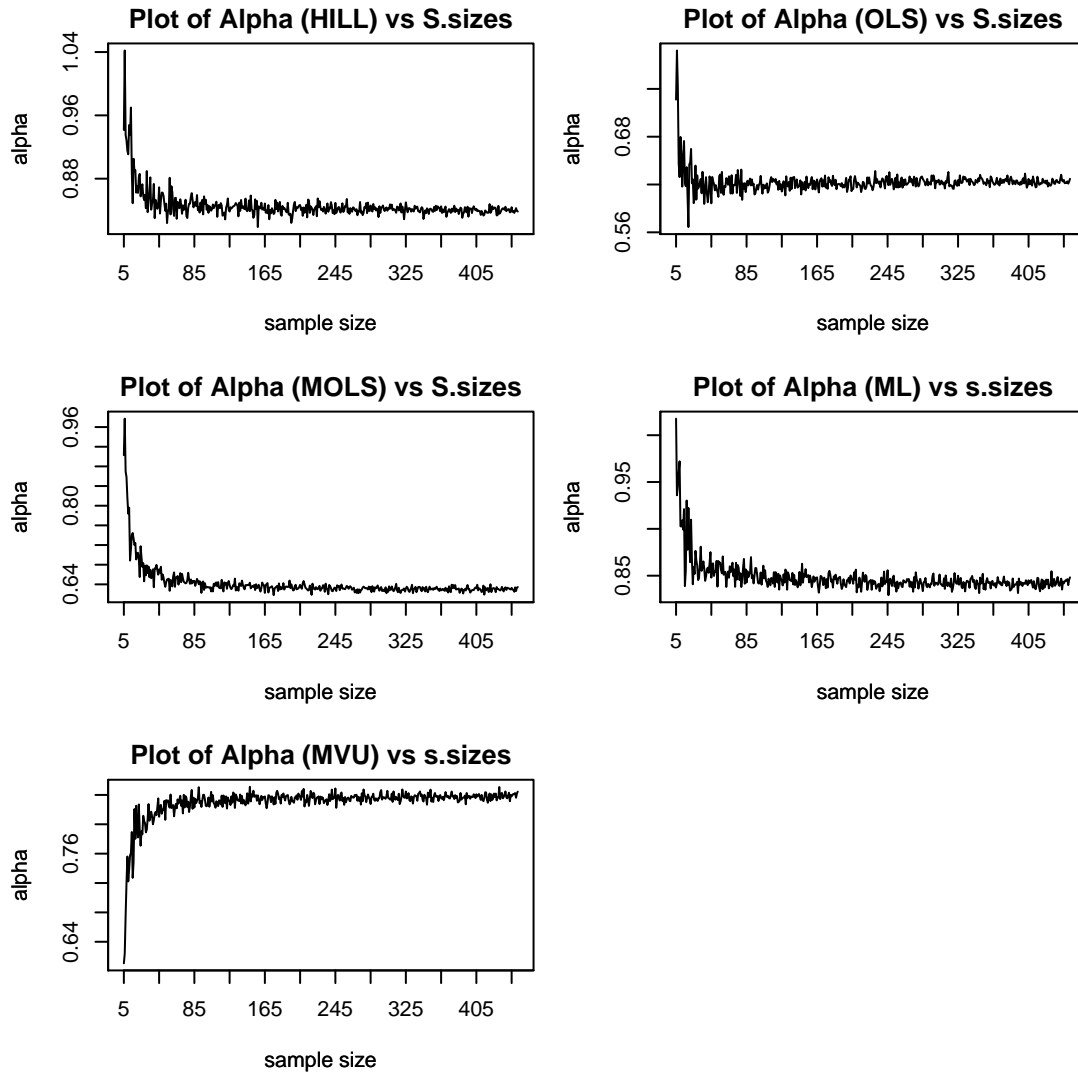


Figure 5.6 Rolling Random Sampling Plots for US (2000)

From the graphs for the data set of 1990, we observe that, except for HILL, all the estimators under estimate the Power Law exponent. The estimated value of the Power Law exponent attains the steady state for sample size greater than 80.

For the year 2000, all of the estimation techniques under estimate the Power Law exponent. There exists significant variation for small sample size for all the estimators. For sample of size greater than 80, the estimated Power Law exponent gets the constant behavior. Graphs for both the data sets for China are included in Appendix B

5.2.2 Results of Rolling Random Sampling for Pakistan and India

The graphical representation of Rolling random sampling with replacement results based on the different estimation procedures for the developing countries, Pakistan and India are discussed.

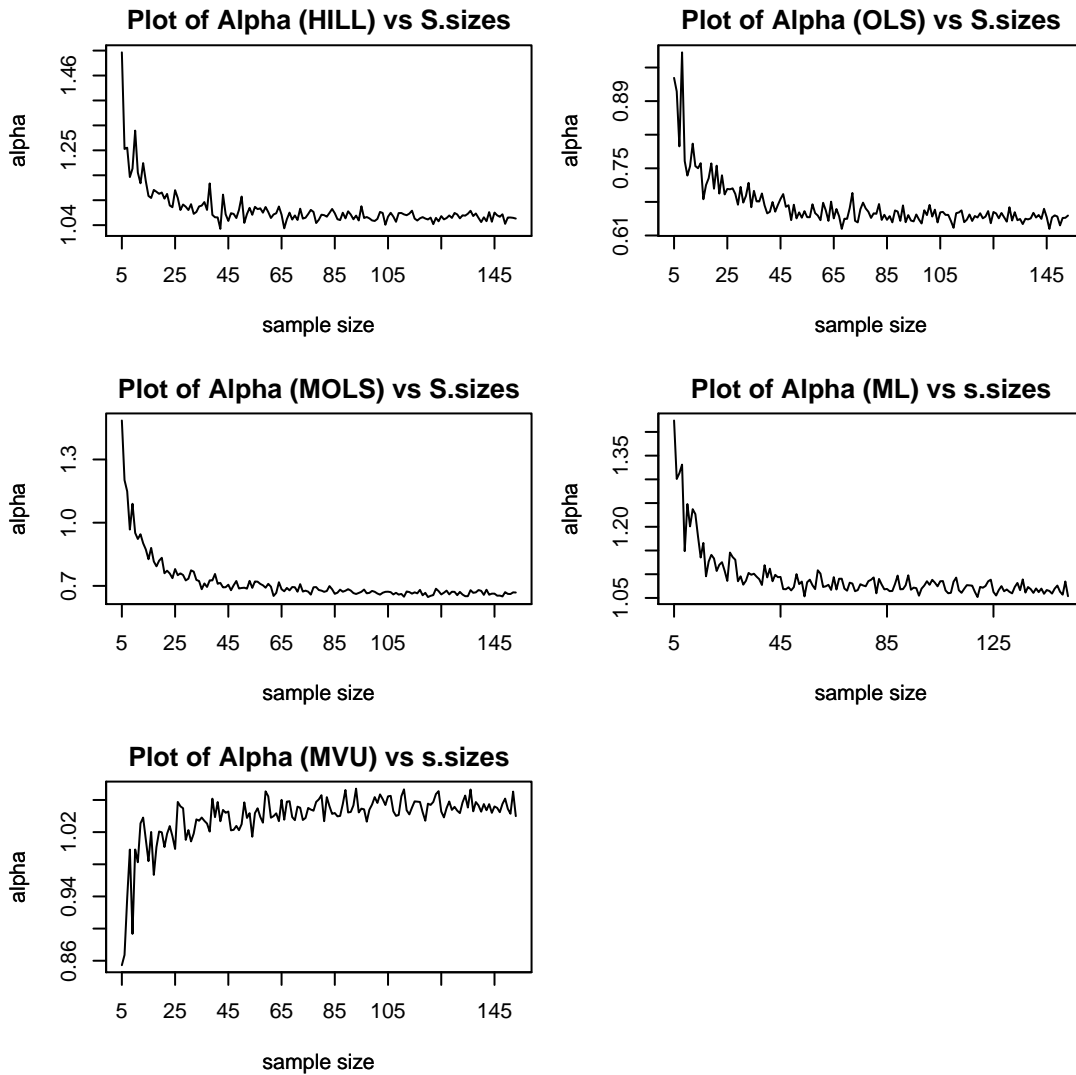


Figure 5.7 Rolling Random Sampling Plots for Pakistan (1981)

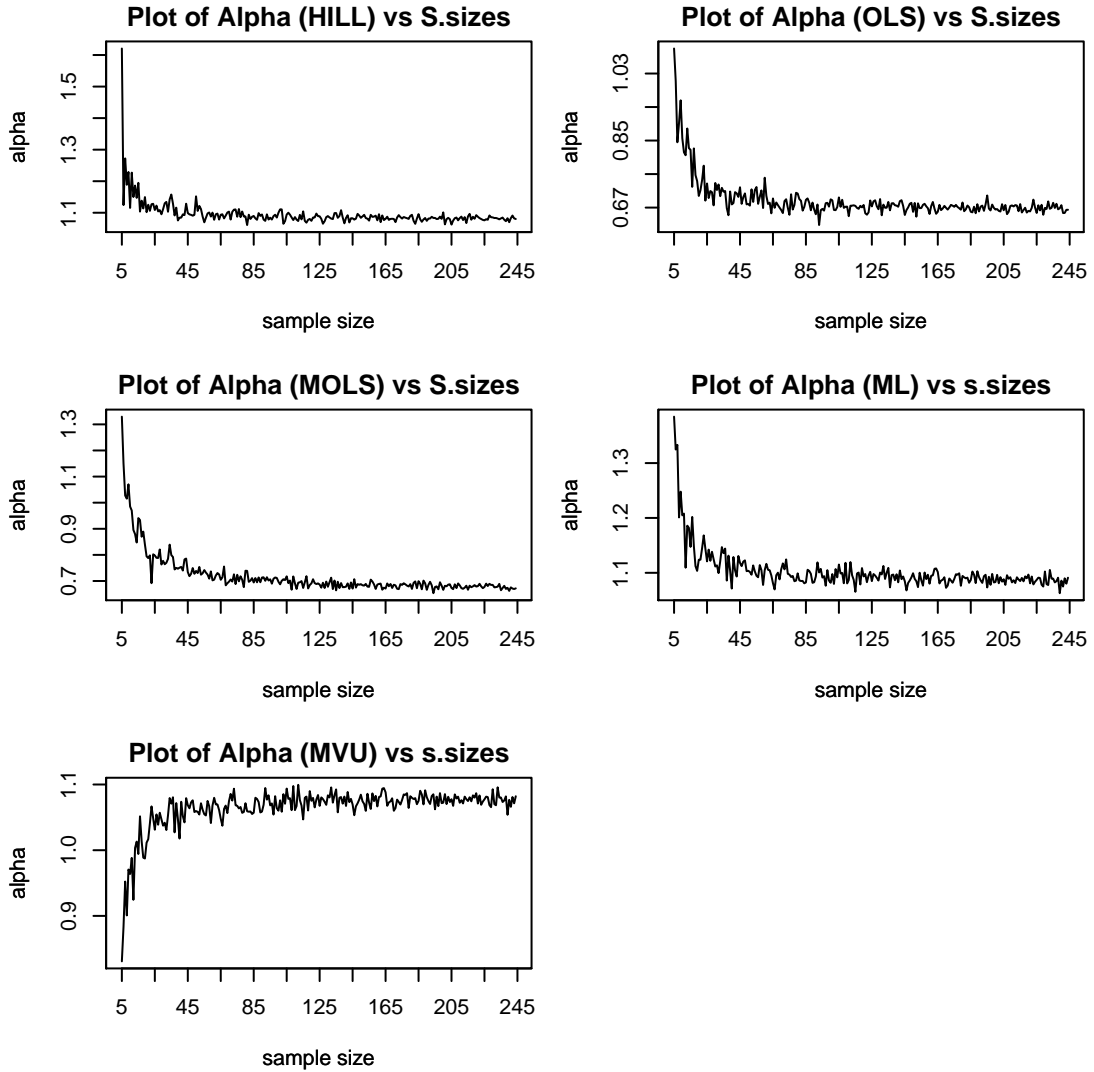


Figure 5.8 Rolling Random Sampling Plots for Pakistan (1998)

Graphs in figure and , show that the Power Law exponent is exactly equal to 1 under HILL, MLE and MVU estimators for sample size greater than 80 for both data sets of Pakistan. OLS and MOLS underestimate the Power Law exponent even in large samples. Graphs for both the data sets for India are included in Appendix B.

5.3 Nonparametric Analysis of Power Law Exponent

In order to investigate Power Law exponent non-parametrically, we make kernel density plots of the estimated value of the Power Law exponent using Rolling sampling. The kernel density method of estimation is widely used nonparametric method for estimating the probability density function of a random variable. The main advantage of constructing the Kernel density plot is that it gives more clear picture that how the values of the Power Law exponent are distributed. Using the kernel density plots, we check whether the distribution of Power Law exponent is uni-modal or bimodal. In kernel density plots, we observe the mode of the distribution of Power Law exponent which will reflect the variation in the Power Law exponent. To construct the Kernel density plots, we plot the density of the Power Law exponent versus the estimated value based on the rolling sampling.

5.3.1 Kernel Density Plots for US and China

Following are the kernel density plots of estimated values of the Power Law exponent for US city size data 1900 and 2000.

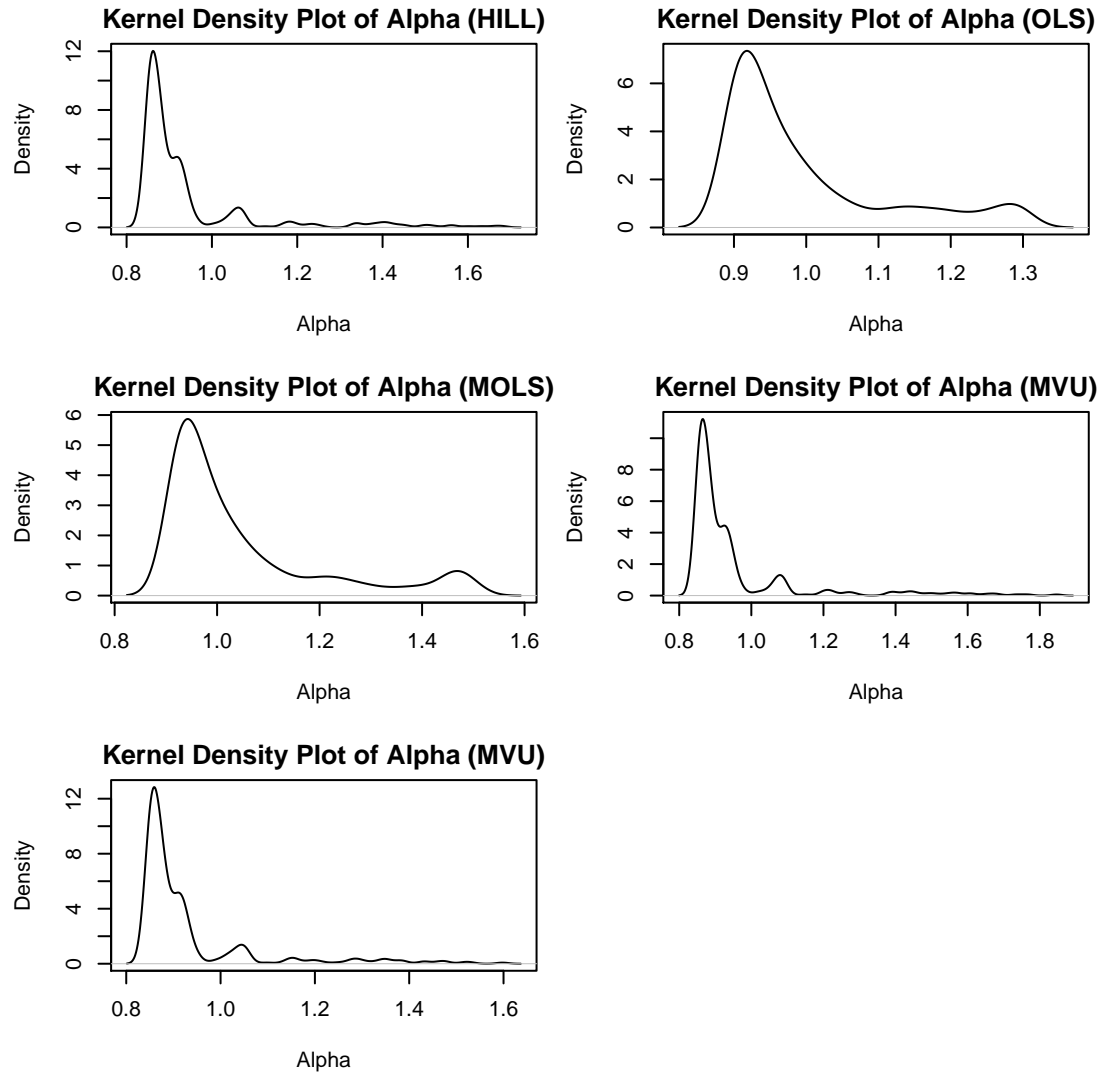


Figure 5.9 Kernel density plots for US (1990)

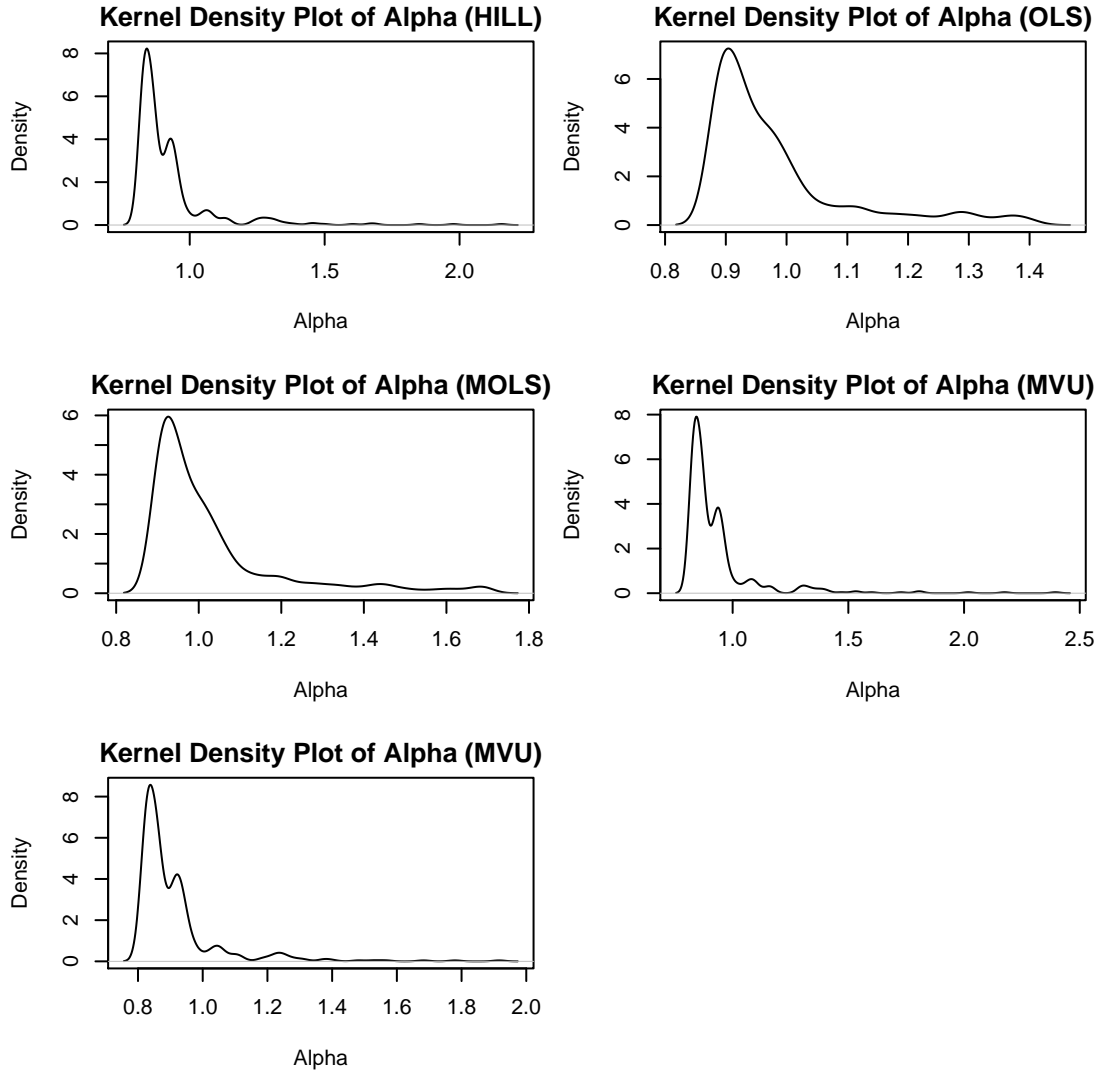


Figure 5.10 Kernel density plots for US (2000)

It can be seen from figure 5.9 and 5.10, for both the data sets of U.S, estimated value of mode is approximately 0.9 for all the assumed estimation techniques. There is considerable variation around the model value which reflects high fluctuation in the value of the Power Law exponent using rolling sampling. The distribution of the Power Law exponent is uni-modal as can be seen from figure 5.9 and 5.10. The graphs for both the data sets of China are contained in Appendix D.

5.3.2 Kernel density plots for Pakistan and India

Following are kernel density plots of estimated value of Power Law exponent for both data sets of Pakistan.

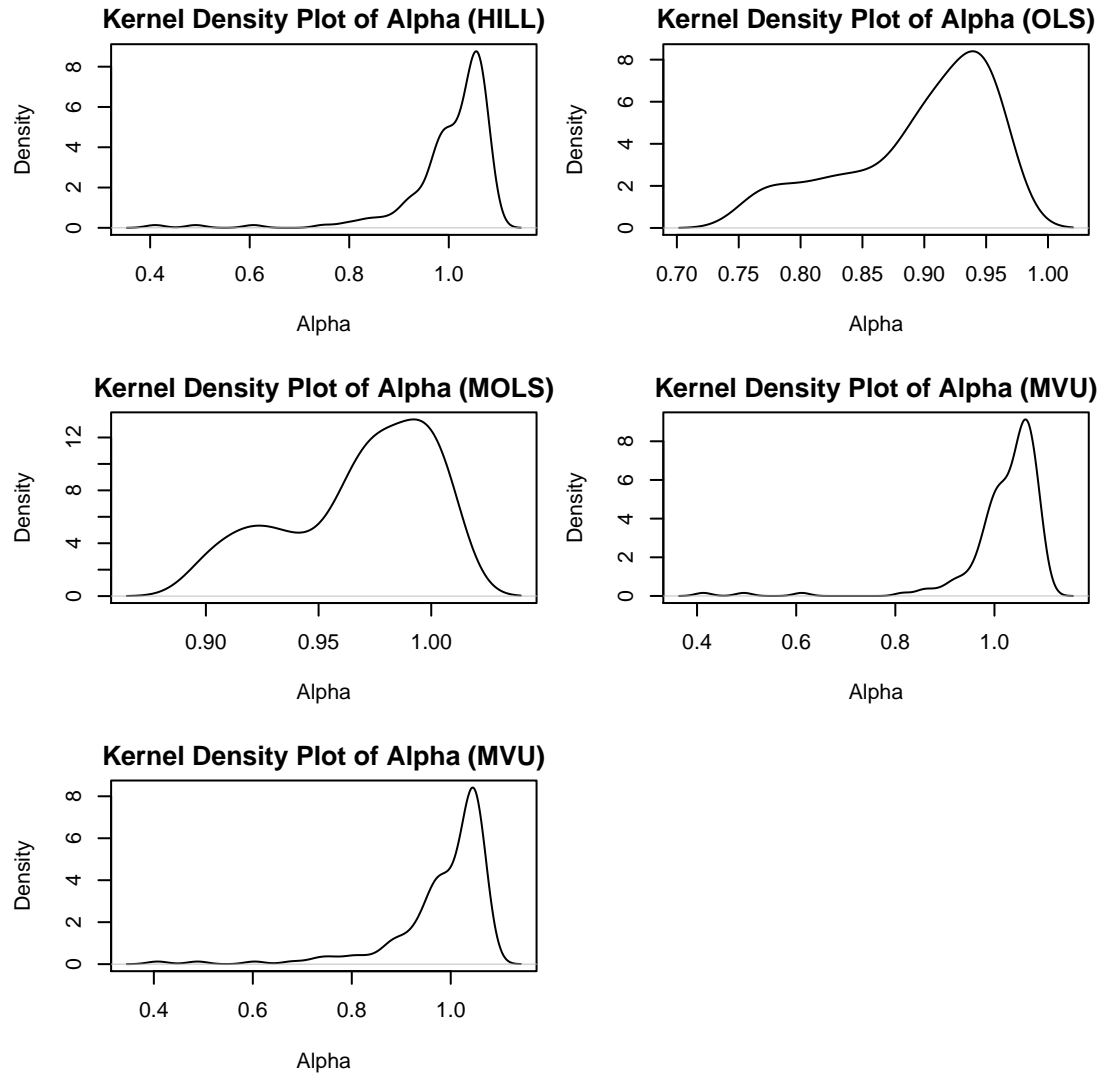


Figure 5.11 Kernel density plots for Pakistan (1981)

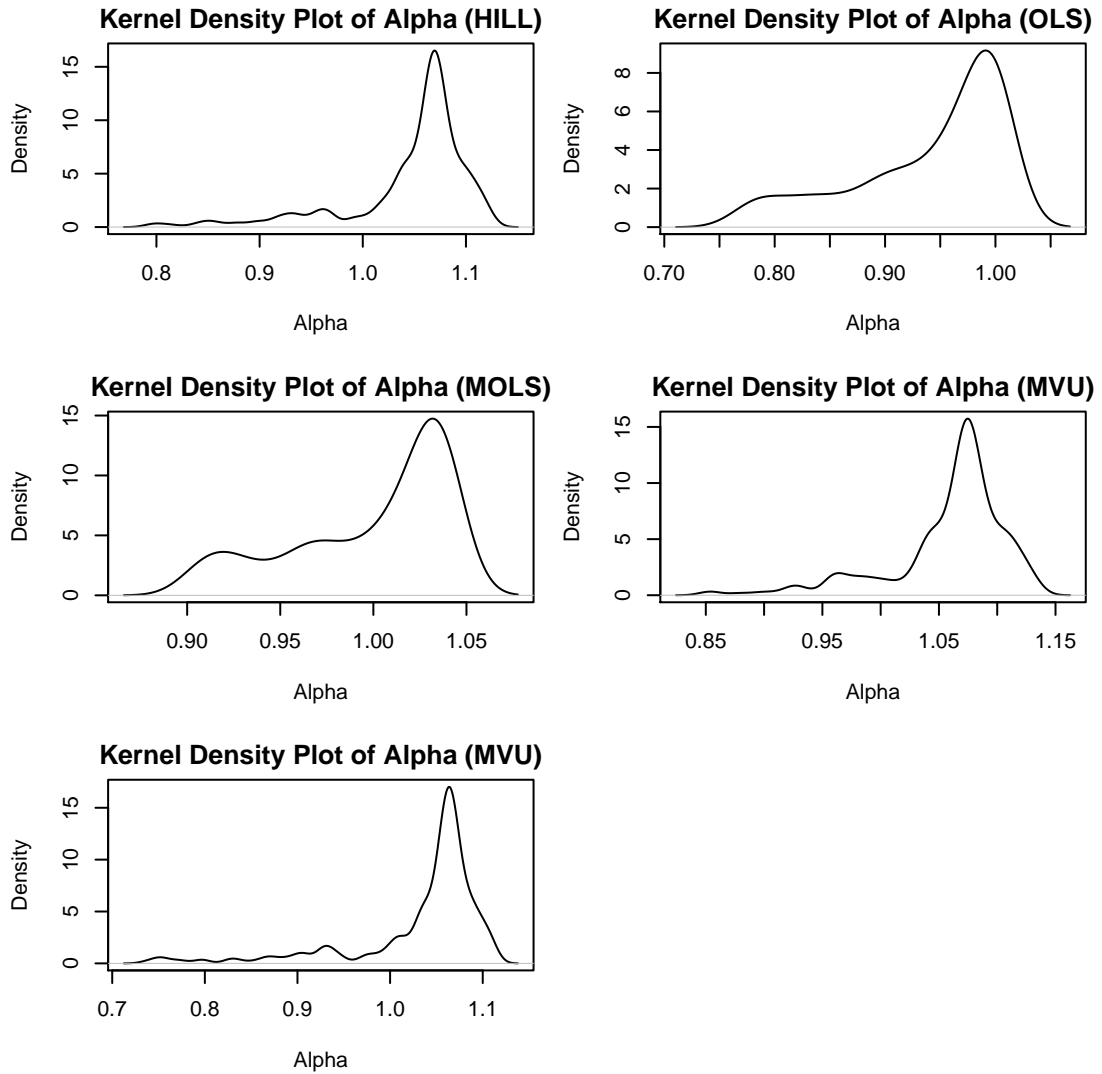


Figure 5.12 Kernel density plots for Pakistan (1998)

From the figure 5.11 and 5.12, it is learned that the distribution of Power Law exponent is uni-modal for both the data sets under all the assumed estimators. There is significant variation in the Power Law exponent around the modes of all the graphs. Graphs of both the data sets of India are shown in Appendix E.

CHAPTER 6

Conclusions and Recommendations

In this study, we have considered the city size data sets of the US, China, Pakistan and India. The main focus of this study was to check the validity of Zipf's Law for these data sets. The Power law distribution was checked for all the data sets through simple plots and KS goodness of fit test. Five different estimation methods have been used for the estimation of the Power Law Exponent through. The validity of Zipf's Law has been analyzed graphically and by estimating the Power Law exponent for all the considered data sets. A simulation study was performed to look for an efficient estimation technique. Through rolling sampling technique, range for each data set was found within which the estimated value of the Power Law Exponent was equal to one. The validity of Zipf's Law was also examined non-parametrically.

From the simple plots of the data sets, it was observed that the plots were L-shaped and seemed to follow some type of the Power law distribution. The Log-Log plots for all the city size data sets are straight line graphs with slope coefficient approximately equal to -1, which indicates the presence of Zipf's Law. It is shown that the rank size rule holds beyond a specific threshold for all the countries data sets.

The KS goodness of fit test is applied to check how best the Power law distribution fits the city size data sets. It is concluded from the KS statistics results that the Power law distribution gives best fit to all the data sets.

To examine the validity of Zipf's Law, the Power Law exponent is estimated through five different estimators, namely, HILL, OLS, MOLS, ML and MVU. For both the US data sets, all the five estimators under estimated the Power Law exponent which indicates that Zipf's Law does not hold for both the data sets. The Power Law exponent is also estimated for both the data sets of China. The results showed that all the estimators overestimated the Power Law exponent for the data set of 2010 while ML, HILL and MVU under estimated this exponent for the data set of 2005. The estimated value of the Power Law exponent was found to be approximately 1 for both the city size data sets of Pakistan which means that Zipf's law holds for both the data sets of Pakistan. Similar conclusions are made for both the city size data sets of India. Through simulation results, it is concluded that MVU possesses minimum PRB and MSE in all the cases and hence considered as the efficient estimator for the Power Law exponent.

The graphical displays, based on the rolling sampling, showed that the fluctuation in the values of the Power Law exponent for small sub sample size containing big cities was higher. When the size of the sub sample was increased, the Power Law exponent attained a constant value. The kernel density plots of all the countries city size data sets revealed that the distribution of the Power Law exponent was uni-modal for all the countries data sets. Considerable variation around the modal value is observed from those plots.

6.0.3 Recommendations

The following recommendations are suggested for future studies in this field.

1. We have considered the city size data sets of US, China, Pakistan and India for various time periods, the same study can be done for the most recent time periods.
2. In this study we have considered the city size to check the validity of Zipf's Law. Many other phenomenon's also follow power law distribution, so validity of Zipf's Law can be check for those phenomenon's.
3. We have utilized five estimation methods for the estimation of the Power Law exponent, other efficient estimators can be used in future studies.
4. We have considered KS test for model selection, other model selection criterion's such as AIC and BIC can also use for this purpose.
5. We have used the rolling sampling to analyze Power Law exponent, other sampling designs can be worked with.
6. The Bayesian estimation techniques can be applied to estimate the Power Law exponent in further studies.

References

- Akhtar, S. and Dhanani, M. R. (2012). City size distribution in Pakistan. *Sindh University Research Journal*, 44(4):699–702.
- Alperovich, G. (1984). The size distribution of cities: On the empirical validity of the rank-size rule. *Journal of Urban Economics*, 16(2):232–239.
- Amalraj, V., Subbarayan, A., and Balamuralitharan, S. (2014). The size distribution of cities in a region: An evaluation of pareto, lognormal and pps distributions. *International Journal of Pure and Applied Mathematics*, 92(2):265–278.
- Anderson, G. and Ge, Y. (2005). The size distribution of chinese cities. *Regional Science and Urban Economics*, 35(6):756–776.
- Auerbach, F. (1913). *Das gesetz der bevölkerungskonzentration*. Petermann Geogr Mitt.
- Black, D. and Henderson, V. (2003). Urban evolution in the usa. *Journal of Economic Geography*, 3(4):343–372.
- Bosker, M., Brakman, S., Garretsen, H., and Schramm, M. (2008). A century of shocks: the evolution of the german city size distribution 1925–1999. *Regional Science and Urban Economics*, 38(4):330–347.
- Brinkhoff, T. (2008). City population. <http://www.citypopulation.de>.

- Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.
- Eeckhout, J. (2004). Gibrat’s law for (all) cities. *American Economic Review*, pages 1429–1451.
- Fazio, G. and Modica, M. (2012). Pareto or log-normal? a recursive-truncation approach to the distribution of (all) cities. <http://www.gallbladder-research.org/media/media-238147-en.pdf>.
- Gabaix, X. (1999). Zipf’s law for cities: an explanation. *Quarterly journal of Economics*, pages 739–767.
- Gabaix, X. and Ibragimov, R. (2011). Rank- $1/2$: a simple way to improve the ols estimation of tail exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.
- Gan, L., Li, D., and Song, S. (2006). Is the Zipf law spurious in explaining city-size distributions? *Economics Letters*, 92(2):256–262.
- Gangopadhyay, K. and Basu, B. (2009). City size distributions for India and China. *Physica A: Statistical Mechanics and its Applications*, 388(13):2682–2688.
- Gibrat, R. (1931). *Les inégalités économiques*. Recueil Sirey.
- Giesen, K., Zimmermann, A., and Suedekum, J. (2010). The size distribution across all cities—double pareto lognormal strikes. *Journal of Urban Economics*, 68(2):129–137.
- González-Val, R., Ramos, A., Sanz-Gracia, F., and Vera-Cabello, M. (2013). Size distributions for all cities: Which one is best? <http://mpira.ub.uni-muenchen.de/45019/1/MPRA-paper-45019.pdf>.

- Guérin-Pace, F. (1995). Rank-size distribution and the process of urban growth. *Urban Studies*, 32(3):551–562.
- Hill, B. M. et al. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5):1163–1174.
- Ioannides, Y. M. and Overman, H. G. (2003). Zipf’s law for cities: an empirical examination. *Regional science and urban economics*, 33(2):127–137.
- Jani, P. and Dave, H. (1990). Minimum variance unbiased estimation in a class of exponential family of distributions and some of its applications. *Metron*, 48(1-4):493–507.
- Likeš, J. (1969). Minimum variance unbiased estimates of the parameters of power-function and pareto’s distribution. *Statistical Papers*, 10(2):104–110.
- Luckstead, J. and Devadoss, S. (2014). A nonparametric analysis of the growth process of indian cities. *Economics Letters*, 124(3):516–519.
- Moura Jr, N. J. and Ribeiro, M. B. (2006). Zipf law for brazilian cities. *Physica A: Statistical Mechanics and its Applications*, 367:441–448.
- Nota, F. and Song, S. (2007). Further analysis of the Zipf’s law: Does the rank-size rule really exist? <http://www.business.unr.edu/econ/wp/papers/UNRECONWP07006.pdf>.
- Rosen, K. T. and Resnick, M. (1980). The size distribution of cities: an examination of the pareto law and primacy. *Journal of Urban Economics*, 8(2):165–186.
- Sarabia, J. M. and Prieto, F. (2009). The pareto-positive stable distribution: A new descriptive model for city size data. *Physica A: Statistical Mechanics and its Applications*, 388(19):4179–4191.

Soo, K. T. (2005). Zipf's law for cities: a cross-country investigation. *Regional science and urban Economics*, 35(3):239–263.

Soo, K. T. (2007). Zipf's law and urban growth in malaysia. *Urban Studies*, 44(1):1–14.

Terra, S. (2009). Zipf's law for cities: On a new testing procedure.
<http://www.cerdi.org/uploads/ed/2009/2009.20.pdf>.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison Wesley Press.

Appendix A

Rolling Sampling Plots for China (2005)

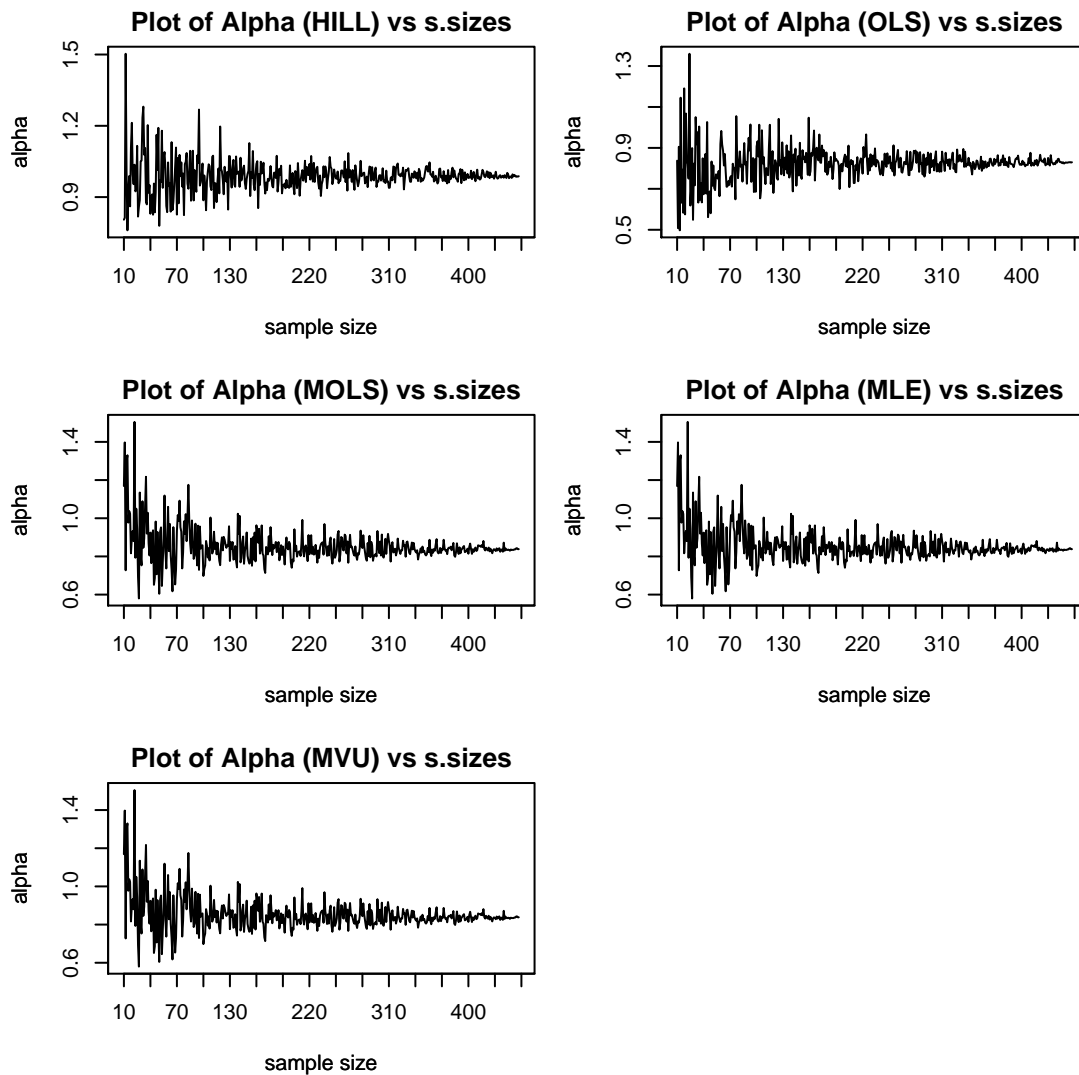


Figure 1 Rolling Sampling Plots for China (2005)

Rolling Sampling plots for China (2010)

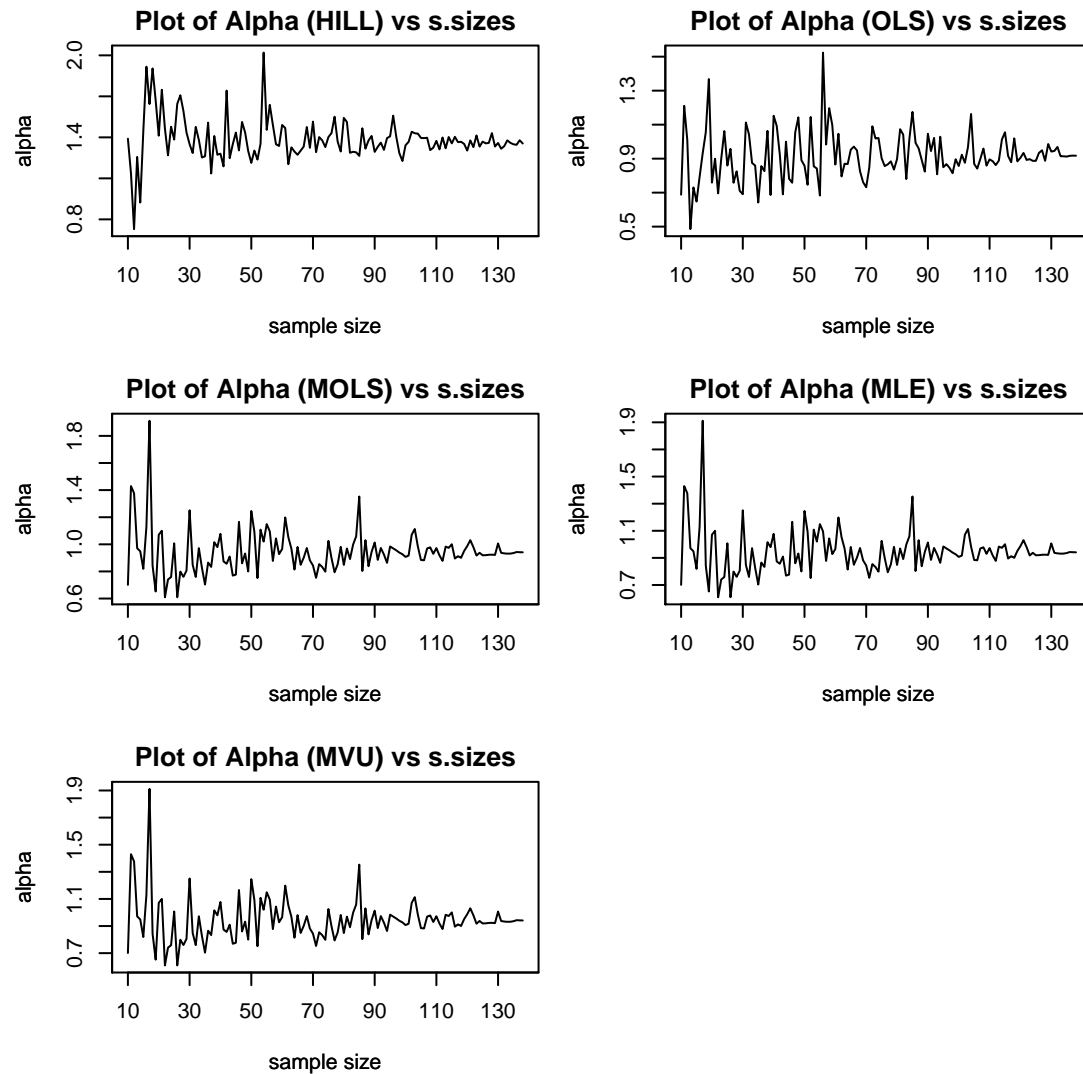


Figure 2 Rolling Sampling plots for China (2010)

Rolling Sampling Plots for India (2001)

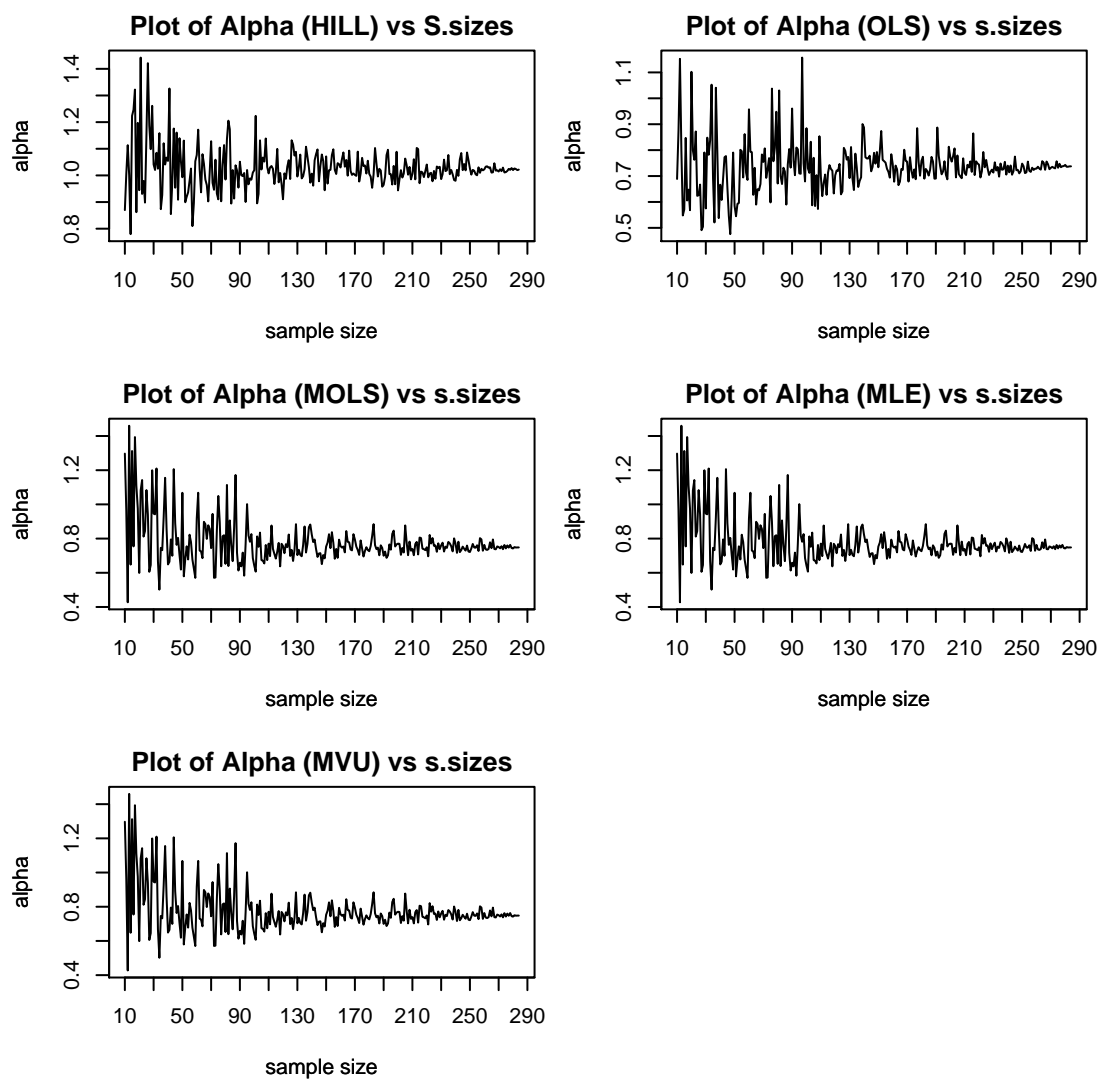


Figure 3 Rolling Sampling Plots for India (2001)

Rolling Sampling Plots for India (2011)

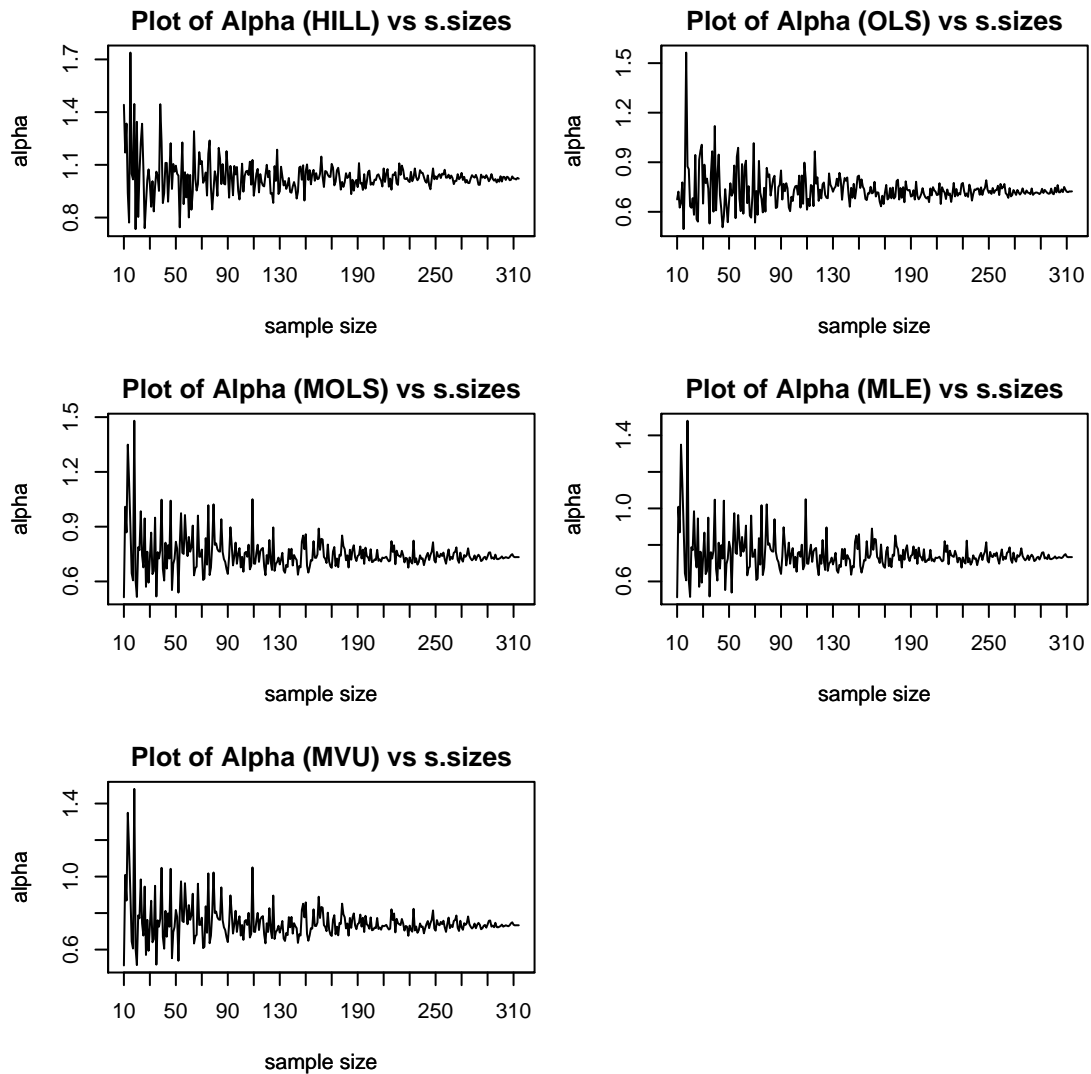


Figure 4 Rolling Sampling Plots for India (2011)

Appendix B

Rolling Random Sampling Plots for China (2005)

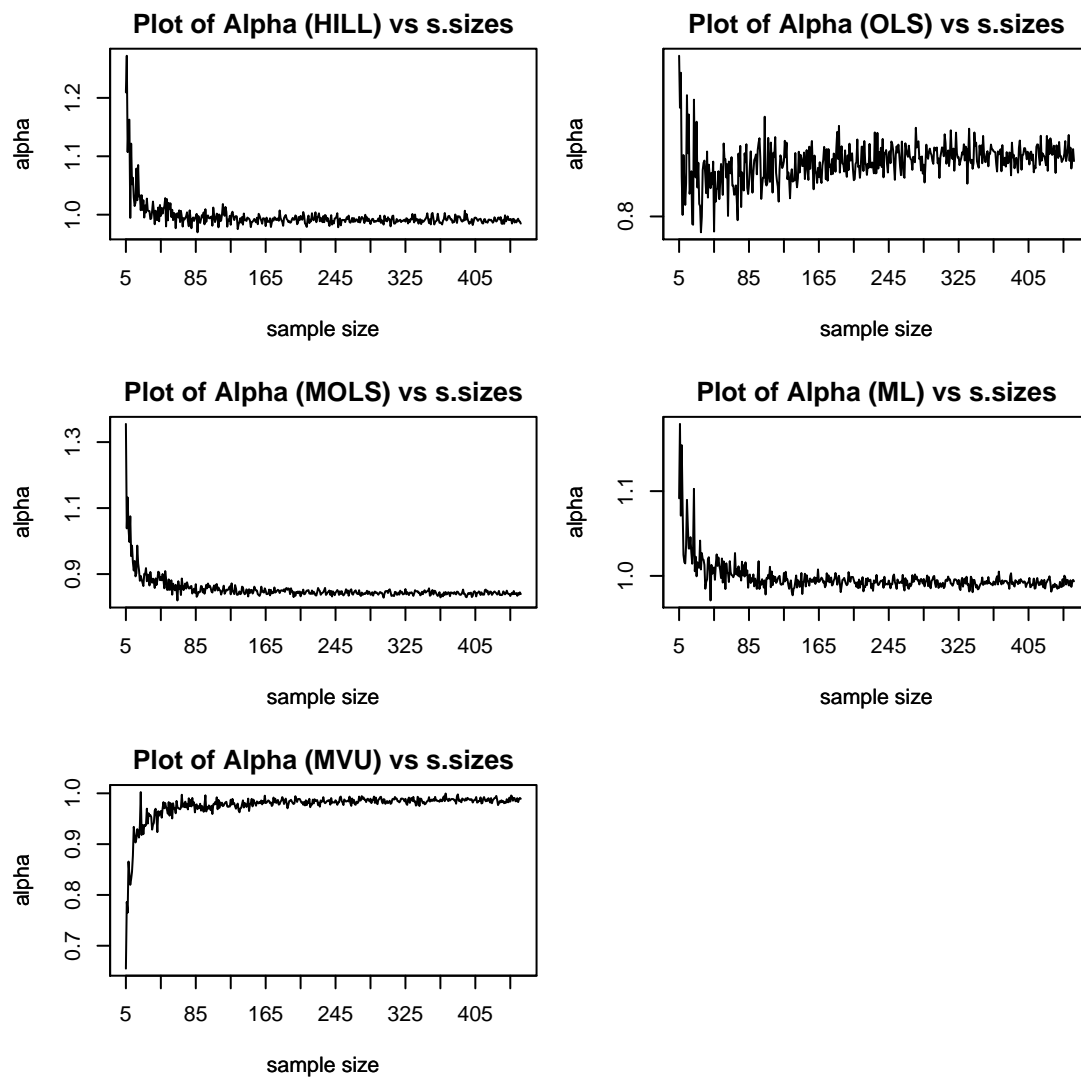


Figure 5 Rolling Random Sampling Plots for China (2005)

Rolling Random Sampling Plots for China (2010)

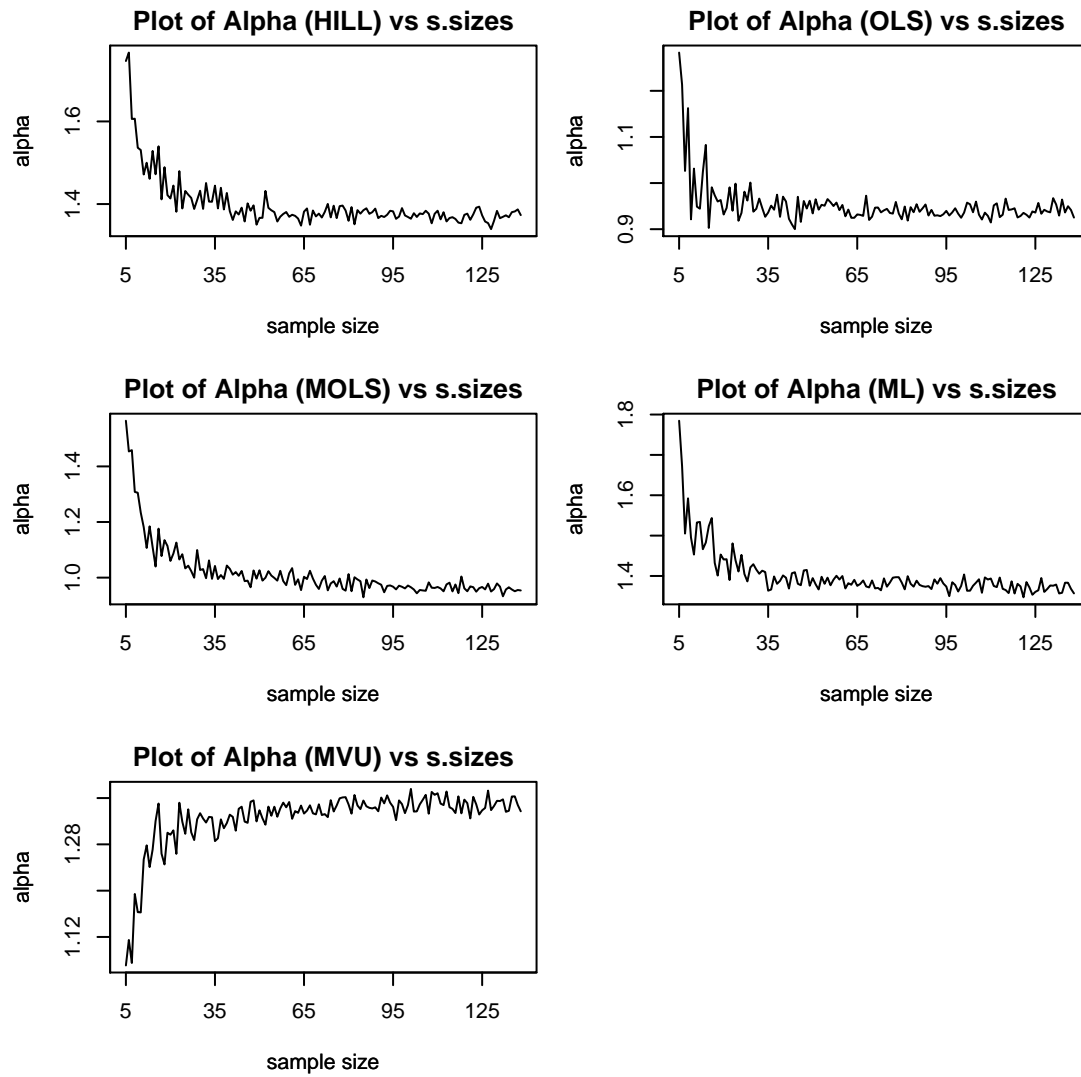


Figure 6 Rolling Random Sampling Plots for China (2010)

Rolling Random Sampling Plots for India (2001)

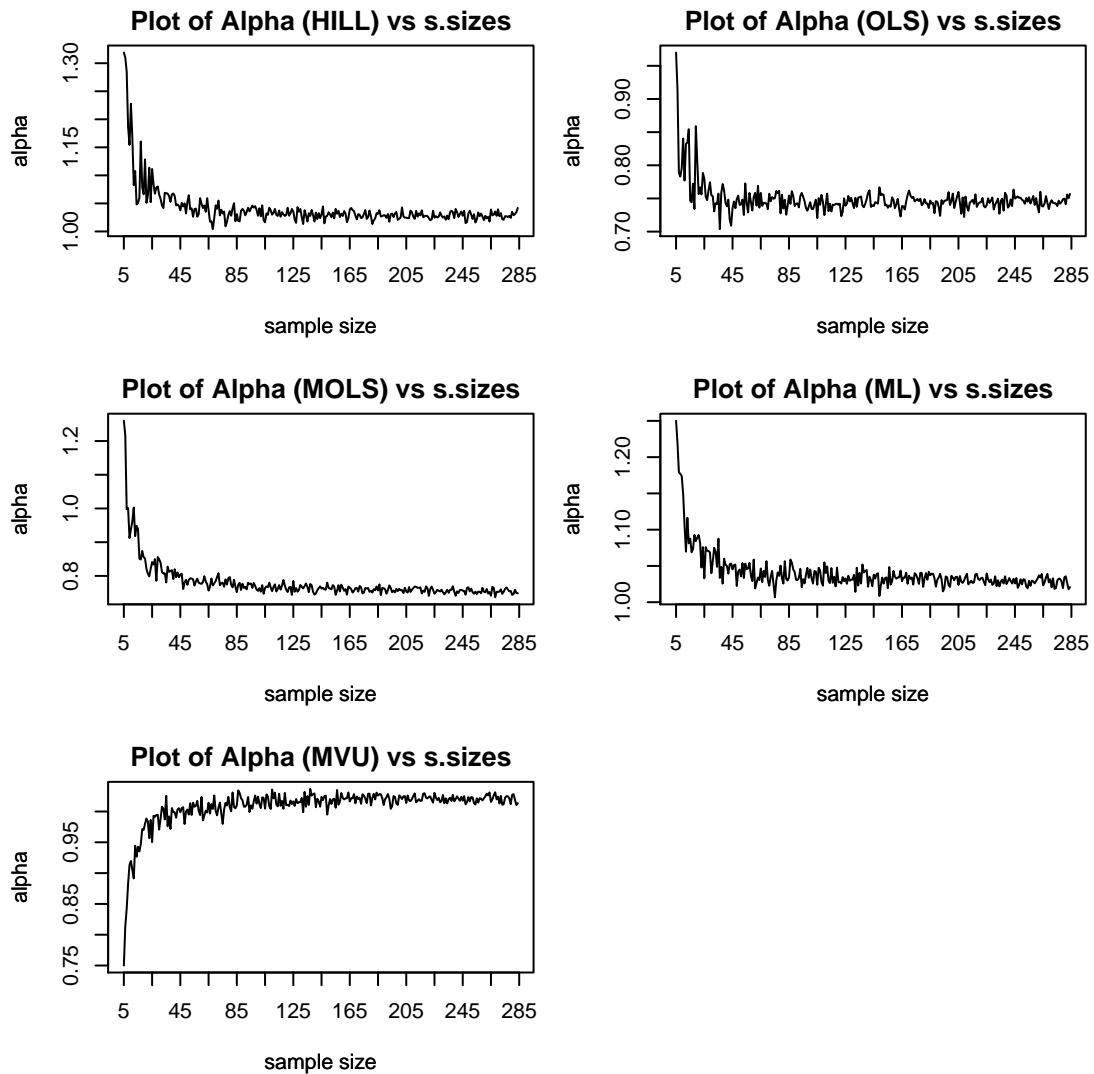


Figure 7 Rolling random sampling plots for India (2001)

Rolling Random Sampling Plots for India (2011)

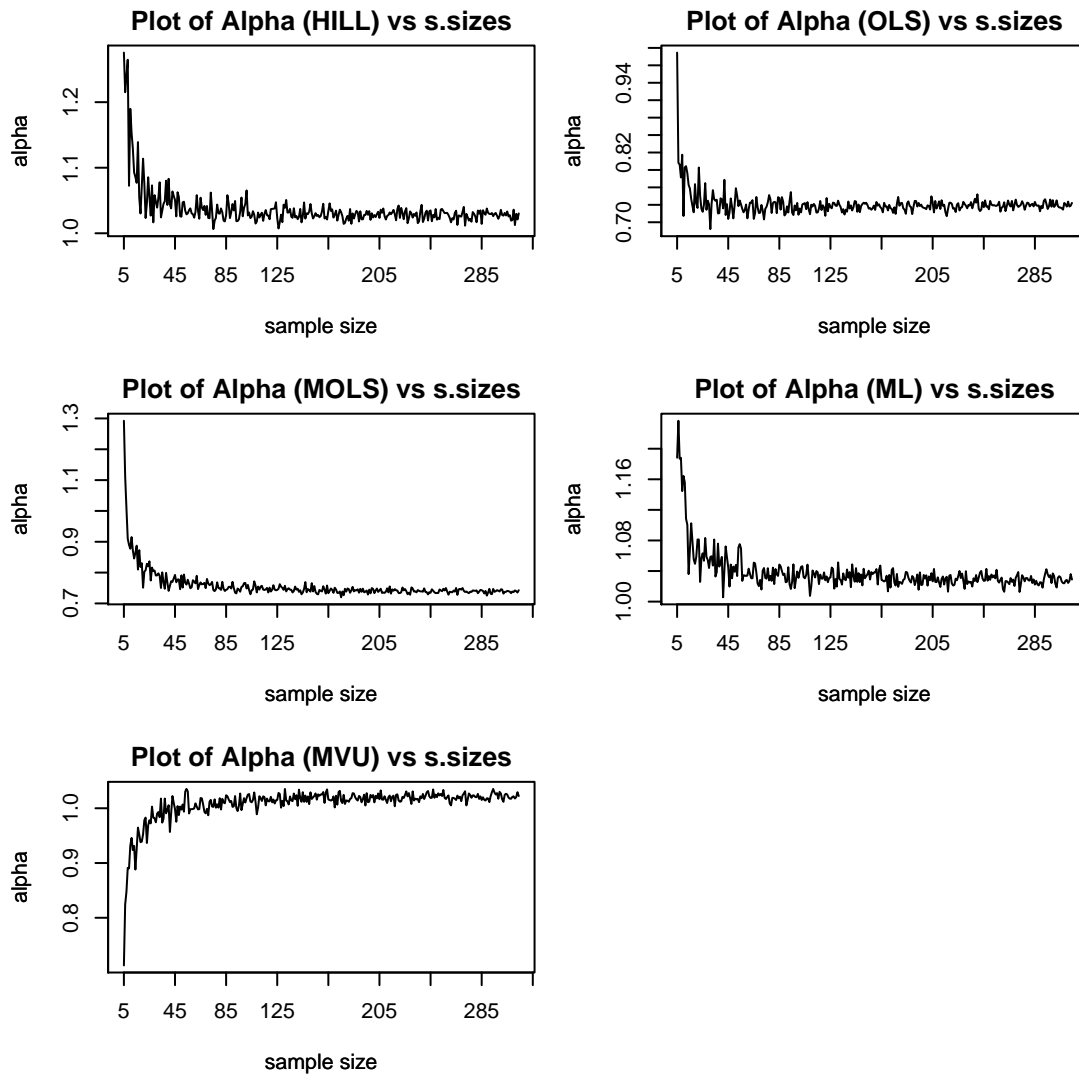


Figure 8 Rolling Random Sampling Plots for India (2011)

Appendix C

Kernel Density Plots for China (2005)

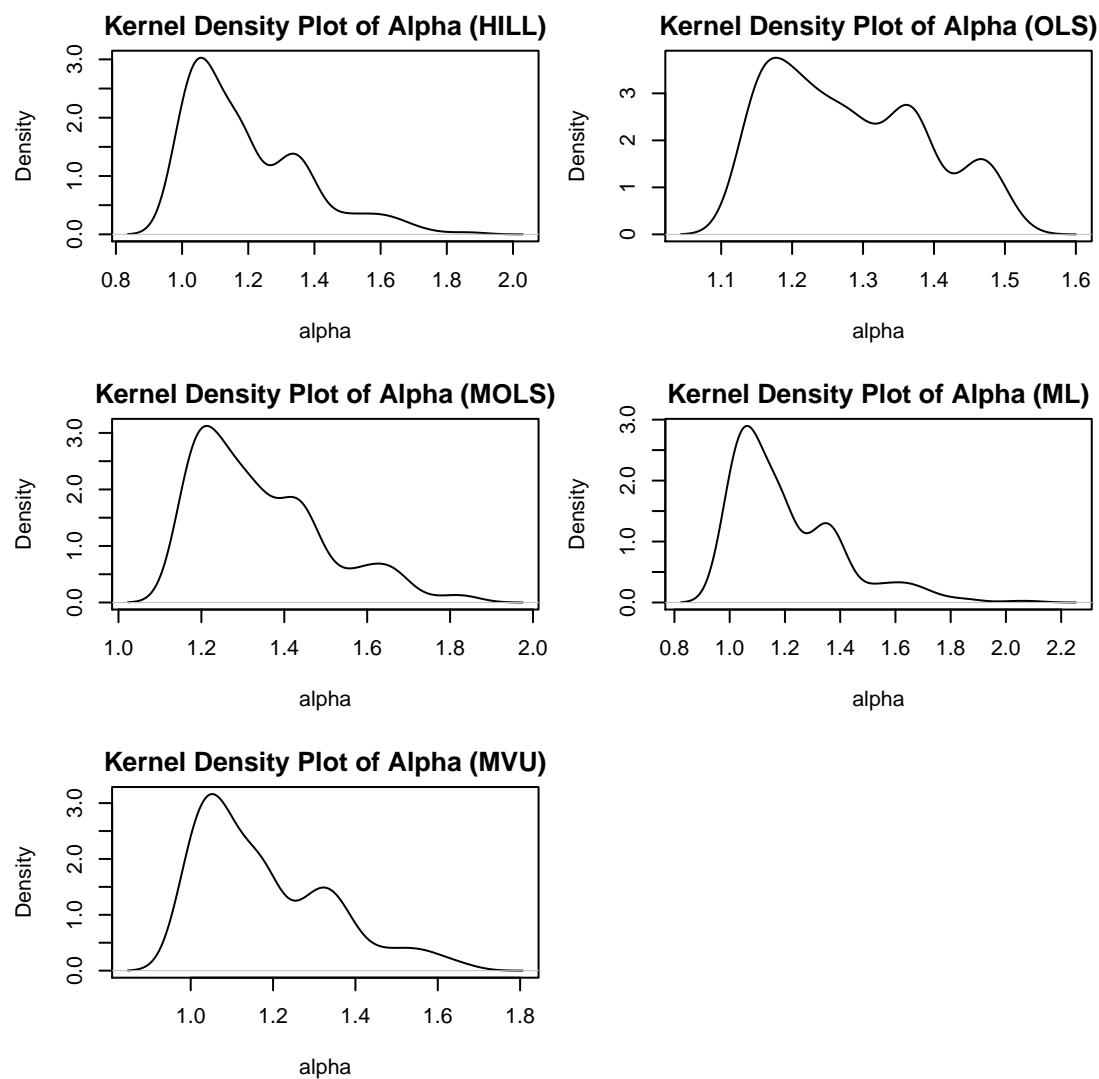


Figure 9 Kernel Density Plots for China (2005)

Kernel Density Plots for China (2010)

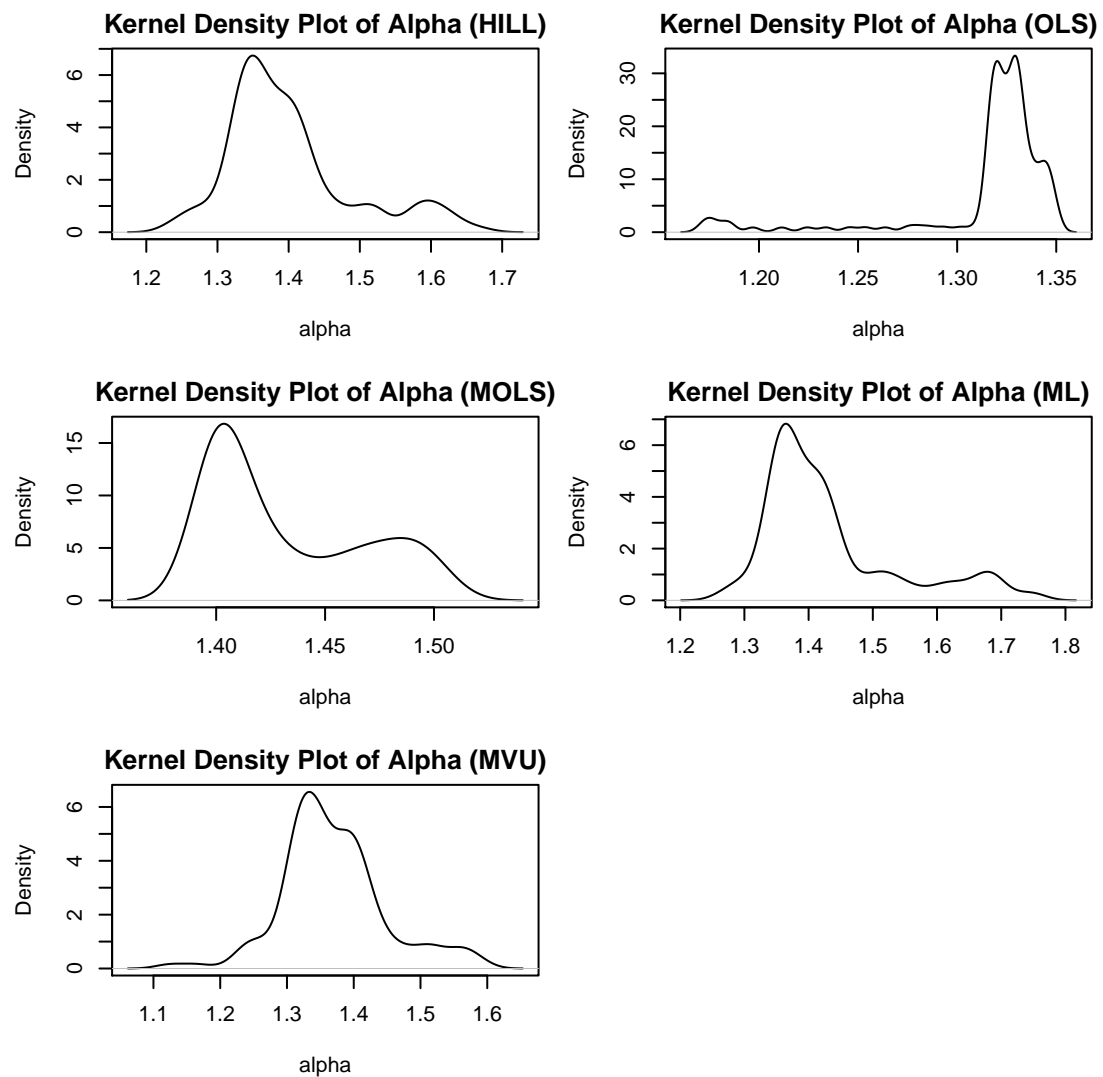


Figure 10 Kernel Density Plots for China (2010)

Kernel Density Plots for India (2001)

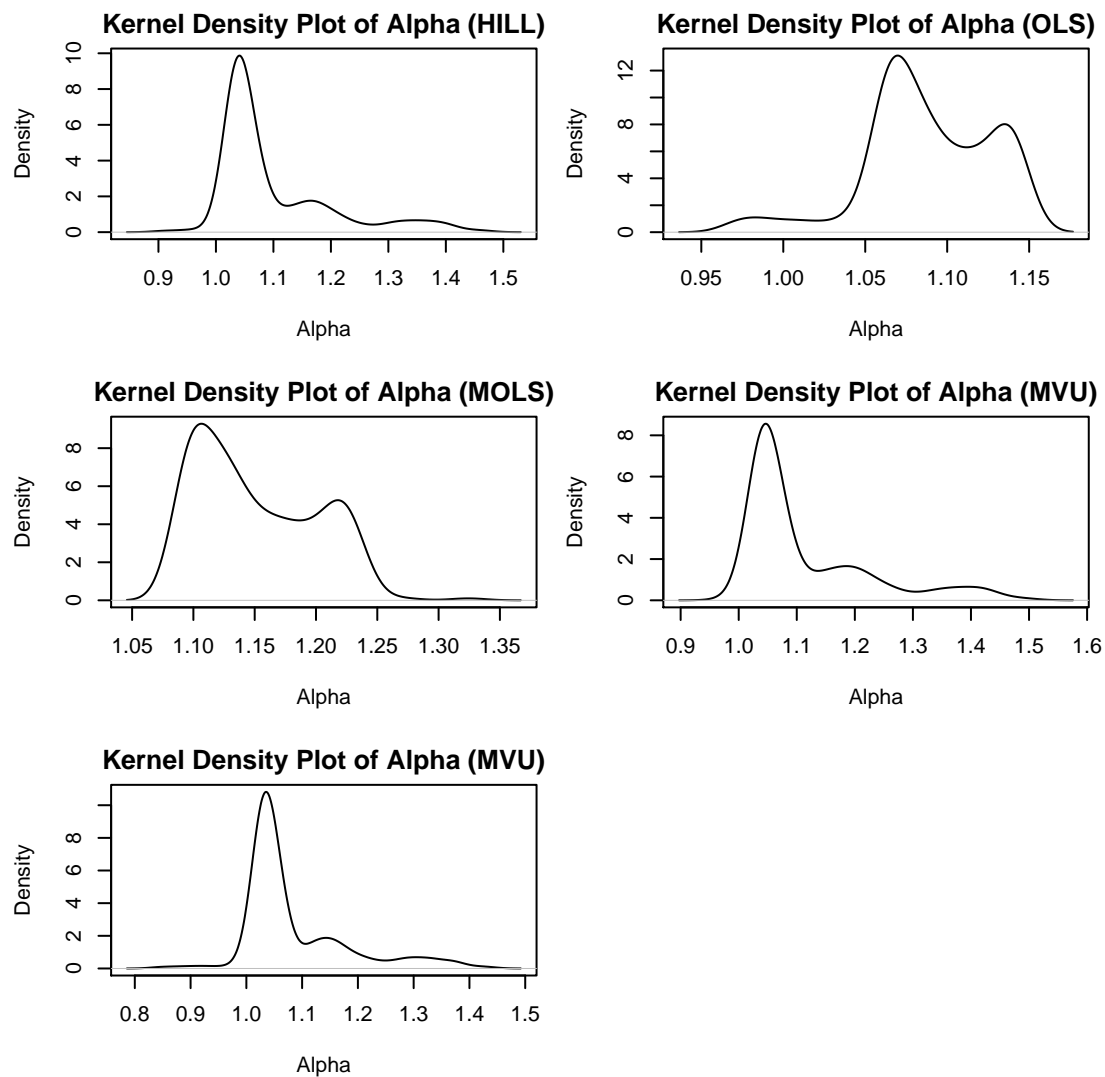


Figure 11 Kernel Density Plots for India (2001)

Kernel Density Plots for India (2011)

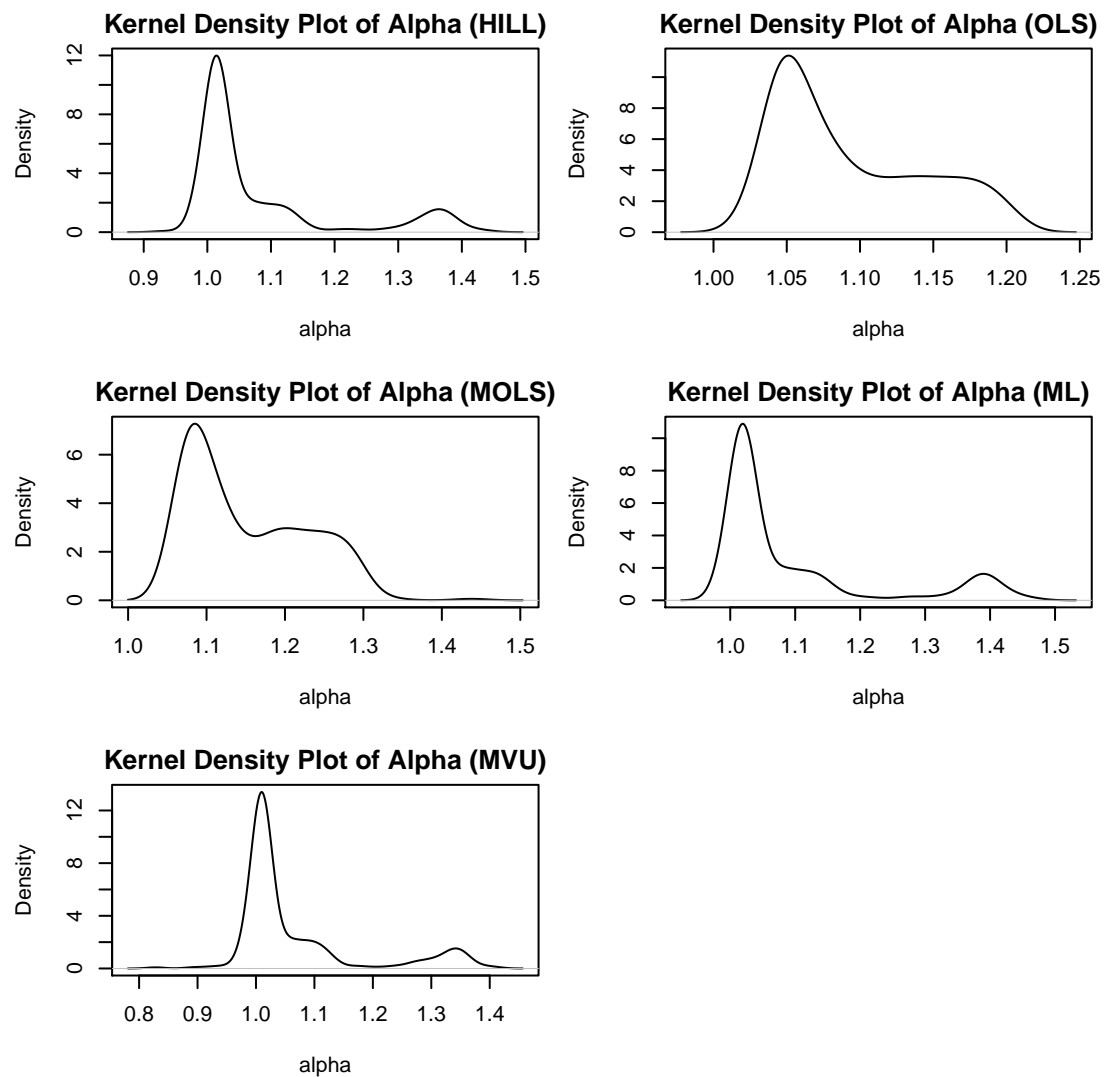


Figure 12 Kernel Density Plots for India (2011)