A COMPARATIVE STUDY ON MACROECONOMIC ANALYSIS OF MONETARY TRANSACTIONS AMONG ECONOMIC SECTORS

[1]R. RAMYA M.Sc., M. Phil., [2]C.RUKMANI M.Sc., M. Phil.,
[1]Assistant professor, [2]Assistant professor
[1]Department of Computer Science
[1]Adhiyaman Arts and Science College for Women, Uthangarai, Krishnagiri, India

ABSTRACT

National Economic Input-Output (EIO) data describes the monetary transactions among economic sectors. The analysis of monetary transactions among these sectors form a weighted bi- directional network from a supply sector to a demand sector and the weight is equivalent to the transaction value between them. In this research, the properties of this network and identify patterns of inter-sector dependence evolution by investigating the historical EIO data over the years 1947-2018. Here we make the following contributions. The first is the discovery that economic transactions (the distribution of the weight) are highly skewed, but follow the double Pareto-lognormal distribution (dPlN). The second contribution is the design of a new method, "Multiple Steps of Pattern Recognition in skewed Data" (M-SPREAD) which identifies patterns and clusters despite the skewness of the data set. Applied these methods on the EIO data and we found interesting and explainable patterns, such as correlations among sectors, various evolution patterns within different transaction scales, outlier sectors and outlier time-stamps.

Keywords: Double Pareto lognormal Distribution-spread, Economic Input-Output (EIO), GDP

I .INTRODUCTION:

The research is presented in the following chapters describes the data set and the data integration processes. In the preliminary analysis part, discover the features of the EIO data set: skewed distribution and asymmetric, hyperbolic-like log- log density curve. The pattern recognition and trend analysis process, where an effective pattern identification procedure, called Multiple Steps of Pattern

Recognition in skewed Data (M-SPREAD) is introduced that utilizes both the skewed distribution property of the data set and the effectiveness of classical clustering method to identify refined clusters and patterns in a skewed data set. The clustering results from the new method are then presented in following sections.

The results illustrate that growth patterns a data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event.

II.DATA DESCRIPTION

The US Economic Input-Output (EIO) accounts show how industries provide input to, and use output from, other industries to produce Gross Domestic Product (GDP). These accounts provide detailed information on the flows of the goods and services of industries in US dollars, such as the purchase of coal from the coal mining sector by the power generation sector. Graphically, these sectors form a weighted bi-directional network through the economic transactions between them. Individual sectors become the vertices of the network; the edges are generated by the economic transaction relationships from the supply sector to the demand sector. The weight of the edges is measured by the dollar amount of monetary transactions between them. Figure 1 shows an example of part of the economy network composed by three economic sectors and the amount of transactions between each pair of them. Learning the web properties of the economy network, including the web structure, the distribution of the size of transactions as well as the evolution of the network can benefit the understanding of the formation and movement of the interconnections among these sectors and is therefore helpful for the prediction of the change of the economic system in the future.

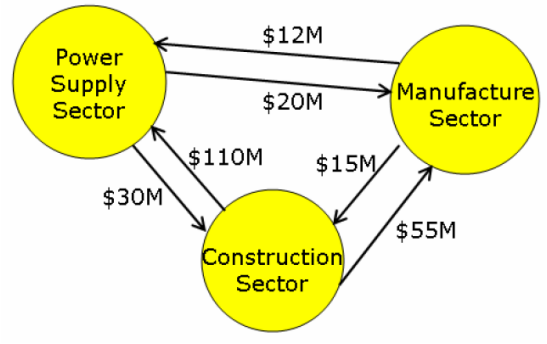


Figure 1 Economy Network Illustration

Monetary connections and commodity supply demand transactions determine the interdependence among economic sectors. The existence of supply-demand connections makes the dysfunction of one economic sector jeopardize for the normal operation of the other sector. The disruption of any sector can potentially endanger the operability of the entire economic system. Economic input output data records the amount of economic transactions among these sectors and reflects the strength of dependence among them. The dependence among these sectors is often reviewed by their direct monetary transactions. For example, large supply and demand requirements normally imply tight dependence of one sector on the other. However, indirect dependences or hidden correlations based on a third party factor, such as competition for the same resources, etc. are underestimated using this evaluation process. Understanding the indirect connections and hidden connections can help comprehend the interdependence better. One way to detect these hidden correlations is to identify sectors that have correlated performance over time. Sectors which have a similar or opposite dependence evolution patterns are thought to be correlated.

In this research, interested in answering the following questions: (a) Describe the graph properties of the economy network? For example, are there any distribution and growth patterns among the transactions of these sectors? (b) Characterize the changes in the transactions over time and explain why? (c) Detect outlier sectors effectively? (d) Spot correlated sectors effectively?

The research is presented in the following chapters describes the data set and the data integration processes. In the preliminary analysis part, discover the features of the EIO data set: skewed distribution and

asymmetric, hyperbolic-like log- log density curve. The pattern recognition and trend analysis process, where an effective pattern identification procedure, called Multiple Steps of Pattern Recognition in skewed Data (M-SPREAD), is introduced that utilizes both the skewed distribution property of the data set and the effectiveness of classical clustering method to identify refined clusters and patterns in a skewed data set. The clustering results from the new method are then presented in following sections.

The results illustrate that growth patterns a data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event.

2.1 Properties

Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis.

The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values are normally all of the same kind. However, there may also be missing values, which must be indicated in some way.

In statistics, datasets usually come from actual observations obtained by sampling a statistical_population, and each row corresponds to the observations on one element of that population. Datasets may further be

generated by algorithms for the purpose of testing certain kinds of software. Some modern statistical analysis software such as SPSS still presents their data in the classical dataset fashion.

2.3 Data integration

It involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations, which include both commercial (when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains. Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved. In management circles, people frequently refer to data integration as "Enterprise Information Integration" (EII).

The United States Economic Input-Output data are kept in a square table with economic

Sectors listed in the row and column of the table. Each data cell entry shows the transaction in US dollars processed from the row sector to the column sector, aggregated during the year when the data was collected. Economic sectors are defined according to a standard classification system developed by the US Department of Commerce to categorize business activities. The Standard Industrial Classification (SIC) was originally developed in the 1930s to classify and compare the establishments by the type of activity in which they were primarily engaged. The SIC was replaced by the North American Industry Classification Standard (NAICS) in 1997. The EIO tables collected from different years have varied levels of aggregation ranging from 65 sectors to more than 500 sectors. To make the data tables comparable from year to year, select the data levels that appear the most frequently over the given set of table series and choose the sectors that are defined using the same classification scheme.

The final decision is to choose the transaction data at the industry level with around 100 sectors, among which 73 individual sectors are selected for the interdependence analysis because they have the same sector categorization definitions. The inter-transaction data are selected from the EIO table of year 1947, 1958, 1963, 1967, 1972, 1977 and 1982. Meanwhile, we also collected the total industry output data from these

years. The total industry output from any sector is the sum of its direct transactions over all the sectors plus the final consumptions on that sector.

Seven input output tables in equal size square tables reorganized and integrated the data so that they are arranged into one table and the transaction values from the same year are represented as one dimension in the new table. The EIO data and the formation of the new data tables specifically, each transaction between any two sectors is presented as one record in the new table. The transaction values over different years are listed as different attributes in the new table.

Since reorganized the transaction data from seven different years, there are seven attributes in the newly formed table. Considered as a missing data, transaction records that have zero values in one or more years are removed from the table. There are around 2950 records in the resulting inter-industry transaction data table.

III.ECONOMY NETWORK PROPERTY

The National Economy Input Output System is an example of complex networks describing the economic transactions among economic sectors at a given time. An economy network can be defined as a directed weighted graph, which has a set of S vertices representing S economic sectors and L directed links pointing from supply sector to the demand sector. The weight attached to each link is equivalent to the dollar value of transactions from the supply sector to the demand sector. This is a bi-directional graph since the sector.

Which produces supply for the other sectors might need also the services or products from those sectors as a necessary part for its operation process. In this chapter, evaluate the property of this network with a major interest on the distribution of the dollar transaction, which is also the weight distribution of the economy network.

3.1 Distribution of Transactions

First step, the histograms of the data set, including the total industry output and Inter-industry transaction from each collected year, are plotted and it is obvious to see that the data is highly skewed with only a few

huge transactions or sector outputs. There is a high density of small and middle-size transactions. The total industry output from year 1947 to year 2018 ranges from $4.2 million to over $1,100 billion with a mean of $43.5 billion and a median of $12.8 billion; the Industry-by-Industry transaction ranges from $100 to $130 billion with a mean of $168.48 million and a median of $8.1 million. Normally, highly skewed data needs further transformation, such as a logarithmic Transformation. The combination of different distribution curves (Probability Distribution Function (PDF), Cumulative Distribution Function (CDF), Negative Cumulative Distribution Function (NCDF)), together with different transformation schemes (linear-linear, log-linear, log-log), for the distribution of the total industry output, inter-industry transaction from each year. Here, the log-linear transformation takes the logarithm of the data value; log-log transformation takes the logarithm of both the data Values and the probabilities.

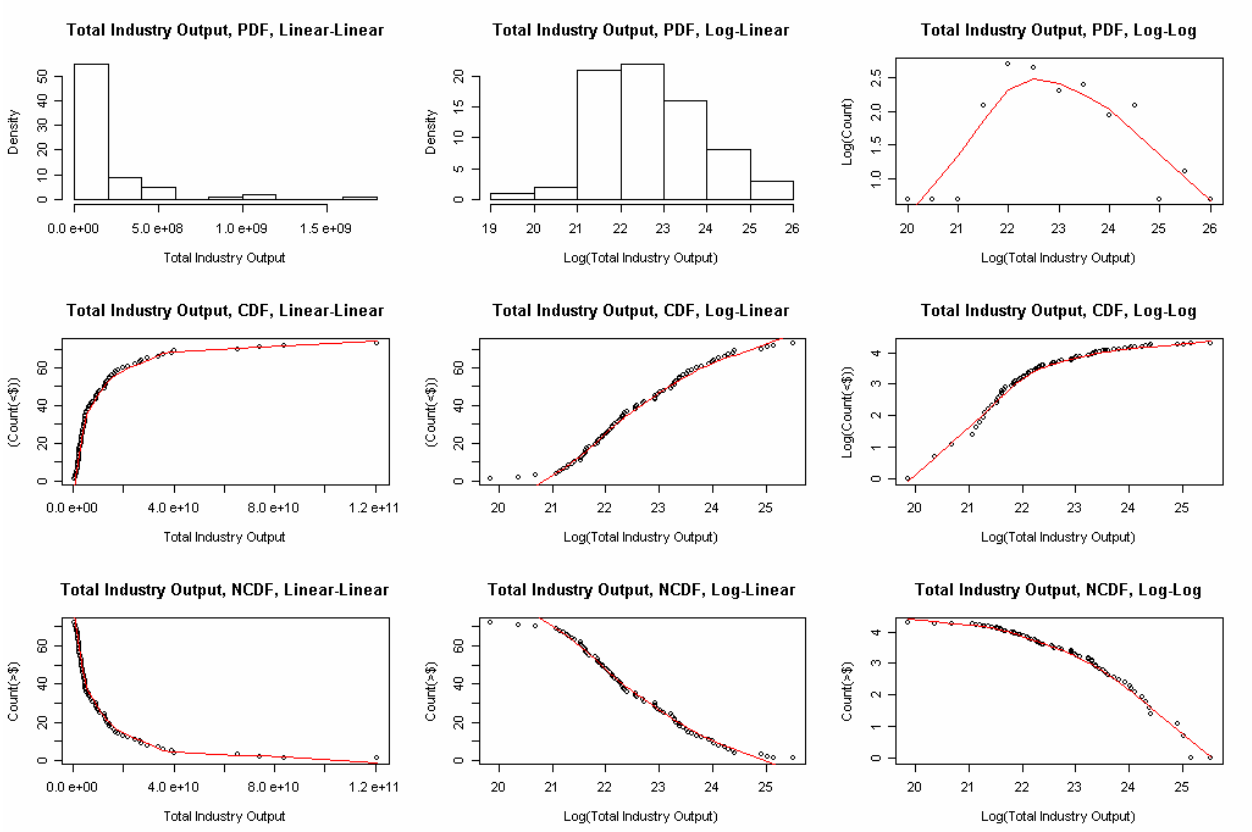
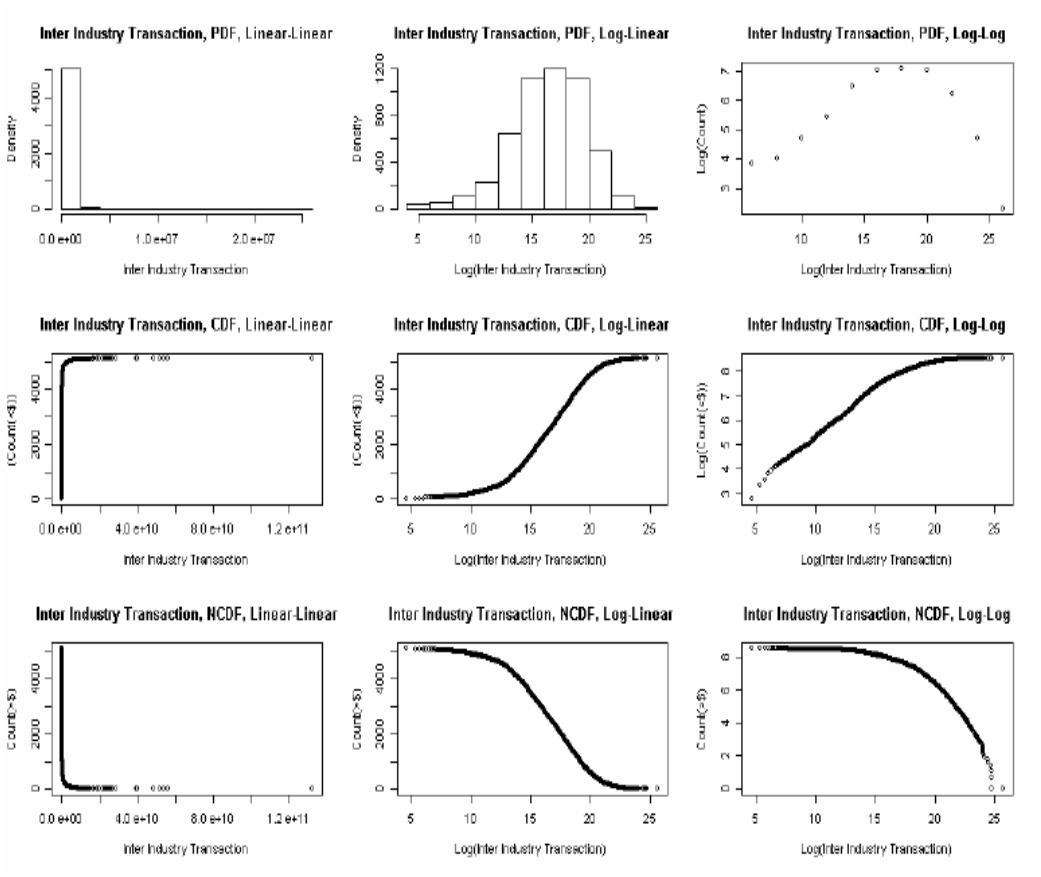


Figure 2 Total Industry Output Distribution Summary



(b) Inter-Industry Transaction Distribution Summary (Year 1982)

Figure 3 Example Distribution Plots (more plots)

Three major observations have been obtained from these distribution plots: (1) the log- linear histograms approximate a normal distribution in most cases. However, one side of the tail often exhibits as slightly skewed. (2) The log-log transformed density distribution function (log-log, PDF) plots approximate a hyperbola shape but, are slightly asymmetric. There is one transition point that separates the curve into two parts.

The log-log transformation of the negative cumulative distribution function (log-log, NCDF) and log- log CDF plots exhibit approximately two straight lines with one transition point in the middle.

IV.ECONOMIC DEPENDENCY EVOLUTION PATTERN

The distribution test shows that the transaction data set is highly skewed. In this chapter present the methodology for, and results of, finding patterns of change in the inter sector connections over several years. Pattern and trend analysis can be conducted using clustering methods. Since each year's transaction is treated as one dimension and there are seven years of transaction data, the number of the dimensions for this problem is seven. The way to discover the pattern of the development of inter-sector interdependency is to conduct clustering analysis using these seven years transaction attributes.

4.1 Quality of Patterns and Hypothesis Evaluation

An important issue in data mining in general and finance in particular is the evaluation of quality of discovered pattern P measured by its statistical significance. A typical approach assumes the testing of the null hypothesis H that pattern P is not statistically significant. A meaningful statistical test requires that pattern parameters such as the month(s) of the year and the relevant sectorial index in a trading rule pattern P have been chosen randomly In many tasks this is not the case. Greenstone and Oyer argue that in the summer "summer swoon" trading rule mentioned above, the parameters are not selected randomly, but are produced by data snooping – checking combination of industry sectors and months of return and then reporting only a few significant combinations.

This means that rigorous test would require to test a different null hypothesis not only about one "significant" combination, but also about the "family" of combinations. Each combination is about an individual industry sector by month's return. In this setting the return for the "family" is tested versus the overall market return. Several testing options are available. Sullivan et. al. (1998, 1999) use a bootstrapping method to evaluate statistical significance of such hypotheses adjusted for the effects of data snooping in "trading rules" and calendar anomalies. Greenstone and Over (2018) suggest a simple computational method –combining individual-test results by using the Bonferroni inequality that given any set of events A1, A2, . . ., An, the probability of their union is smaller than or equal to the sum of their probabilities:

P(A1 & A2&: : : & Ak) · §i=1:kP(Ai)

Where Ai denotes the false rejection of statement I from a given family with k statements. One of the techniques to keep the family-wide error rate at reasonable levels is "Bonferroni correction" that sets a significance level of ®/k for each of the k statements. Another option would be to test whether the statements are jointly true using the traditional F-test. However if the null hypothesis about a joint statement is rejected it does not identify the profitable trading strategies (Greenstone and Oyer, 2018).The sequential semantic probabilistic reasoning that uses F-test addresses this issue able to identify profitable and statistically significant patterns for SP500 index using this method. Informally the idea of semantic probabilistic reasoning is coming from the principle of Occam's razor (a law of simplicity) in science and philosophy. Informally for trading it was written as follows: When two competing trading theories which make exactly the same predictions, the one that is simpler is the better & more profitable one. If two trading/investing theories which both explain the observed facts then you should use the simplest one until more evidence comes along. The simplest explanation for a commodity or stock price movement phenomenon is more likely to be accurate than more complicated explanations. If two equally likely solutions to a trading or day trading problem, pick the simplest. The price movement explanation requiring the fewest assumptions is most likely to be correct.

V.RESULTS AND DISCUSSION

A main interest in understanding the properties of economy network and the pattern of inter-sector transaction evolvement, this research discovers that the distribution of economic transactions is highly skewed but follows the double Pareto lognormal distribution. Moreover, it designs the procedure of Multiple Steps for Pattern Recognition in Skewed Dataset (MSPREAD) that can handle the skewness of the data set and handle the effect of various magnitude of the data set. On top of this, the research has designed effective visualization methods, such as M-Plane, M-Slice, M-Sub Setting and so on. Applying the M-SPREAD procedure and utilizing these visualization methods, the discovery patterns of transaction evolution using the Economic Input Output data set, which includes the effect of the various scale of the transaction on transaction evolution pattern; finding correlated and anti-correlated sectors and identifying outlier sectors

and outlier time stamp.

Skewed data set appears quite often, for example, any network data set that follows power law rules. In these cases, this methodology would be very helpful in terms of handling both the skewness of the data set and the discovery of patterns that are related to the skewness properties of the data set. The work in this research has broader applicability.

In addition, it can be applied to finance and business settings to help locate correlated and anti-corrected entities, such as companies, products, stocks, etc that have competing relationships or cooperation relationships.

Tables and figures

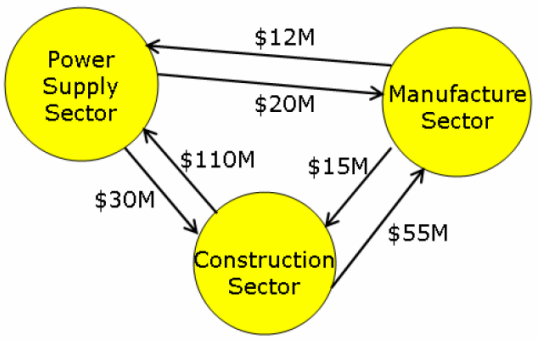


Figure 1 Economy Network Illustrations

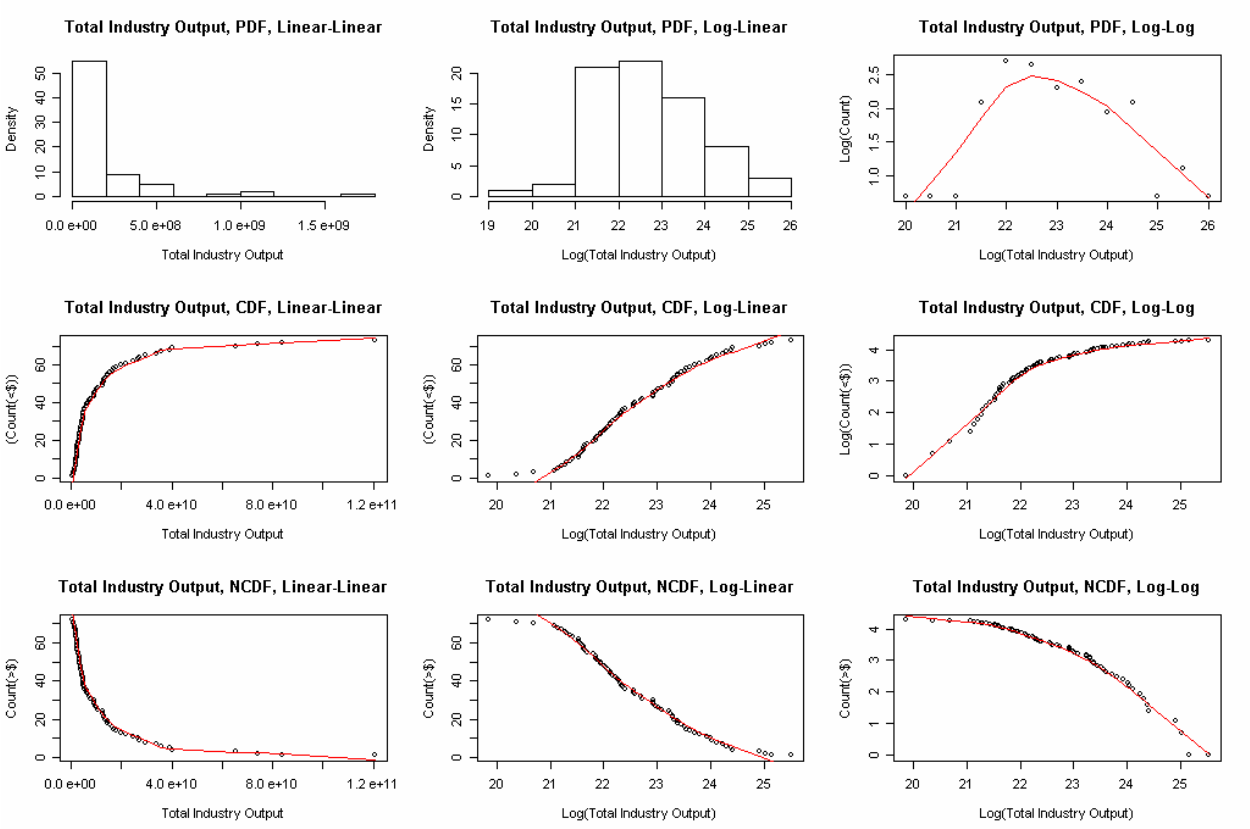


Figure2 Total Industry Output Distribution Summary

References

[1] Adamic, L. A., Zipf, Power-laws, and Pareto - a ranking tutorial, Information Dynamics Lab, HP Labs

[2] Crow, E.L., and Shimizu, K.,(editors), Lognormal Distributions: Theory and Applications, Markel Dekker, Inc.,New York, 1988

[3] Faloutsos, M., Faloutsos, P., Faloutsos, C., On Power-law Relationships of the Internet Topology, ACMSIGCOMN, Cambridge, MA, USA 1999.

[4] Guilmi, C. D., Gaffeo, E., Gallegati, M., Power Law Scaling in the World Income Distribution, EconomicsBulletin,15,6 (2003),1-7

[5] MacQueen, J. B., (1967): "Some Methods for classification and Analysis of Multivariate Observations,Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[6] Maltseva, E., Pizzuti, C., Talia, D., Mining High-Dimensional Scientific Data Sets Using Singular Value Decomposition, Data Mining for Scientific and Engineering Applications, Kluwer, 2001.

[7] Mitzenmacher, M., A Brief History of Generative Models for Power Law and Lognormal Distributions, Internet Mathematics, 1, 2 (2003), 226-251

[8] Mitzenmacher, M., Dynamic Models for File Sizes and Double Pareto Distributions. Internet Mathematics,1,3 (2003), 305-333.

[9] Reed, W. J., The Pareto law of incomes – an explanation and an extension, PHYSICA A, 319 (2003) P469-486O0] Reed, W. J., Jorgensen. M., The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions, October, 2003