www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2. Issue 10 October 2013 Page No. 2985-2996

Modeling and Data testing for Indian Universities Clusters

Srinatha Karur[1], Prof. M.V. Raman Murthy[2],

[1]C.M.J University, School of Computer Science & Engineering,
Shillong, Meghalaya, India
karur_sri@yahoo.co.in

[2]School of Mathematics & Computer Science, Osmania University
Hyderabad, Andhra Pradesh, India
mv_rm@rediffmail.com

Abstract: ***This paper gives Modeling and Data testing on Indian Universities Clustering with respect to Statistical and Mathematical methods. In next phase of action we can implement with different Data mining tools for observe the difference between them. Using different modeling techniques we can easily find out the outliers in the data. In this paper we discussed only on numeric data very purely and not allowed any other type of data. Authors in this paper consider the two types of data set one is given set and another is 50% random set from given data. Authors already published about the data preparation in their published paper [51], pp 31-32.***

Keywords: ***Sampling techniques, Errors, Outliers, Data mining tools, Curves***

.

1. Introduction

This document describes about different types of methods and techniques which are available for Modeling of given or obtained set of data. Authors in this paper also described about nature and behavior of data with respect to its scope or need of application. In real world applications any is always has some error before it enters into process. Each and every domain has its own terminology on error concept and has different mechanisms for handle the outliers. Already lot of history is available on outlier's nature and detection with respect to real application [2]. Outliers may influence the analysis of a set of data in various different ways. Some practical examples are used to motivate a categorization of the different aims in handling outliers and of the different models which might be employed to reflect the presence of outliers [3]. In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers are to be expected (and not due to any anomalous condition) [4].Very good rain example is available in [1]. Quartiles are very useful to estimate the outliers or extreme values [44].All popular data mining tools are using Quartiles for outlier's estimation. The authors are already bringing notice about nature of outliers [49], pp. 132.

2. Objective

Our research objective is to model and integrate data with Statistical and Mathematical methods. After modeling the data it is once again test with data mining tools for outliers. After testing with data mining tools the results are generate and record in the form of tables or figures. We want to estimate what type of clusters is useful for Indian Universities using Modeling and Data mining tools.

3. Related Work

3.1 Nature and Scope of outliers

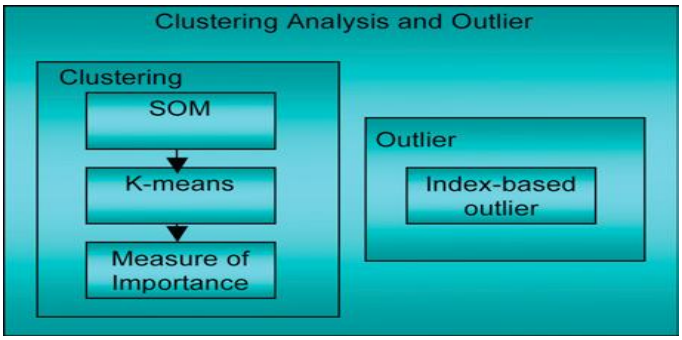
The In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model Misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis (Williams et al., 2002; Liu et al., 2004).

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. The taxonomy of outliers are available in so many publications. The

nature of outliers in Data mining is available in [5]. The original author described about Distance data, spatial data, and Cluster data in terms of outliers [5], pp. 11-12. The implementation details in terms of data mining are not covered by original author. The author discussed about neural networks outliers also. At present in these paper neural networks has no role.

The outliers are implemented in terms of Statistics, Linear regression; control charts are available in [6]. The original author only pointed out the 95% confidence interval and implemented only linear curve with single degree. The author evaluated the performances of above three methods [6], pp. 5.

The Clustering and Outlier Analysis for Data Mining (COADM) tool is one of the three key components delivered under the Systematic Data Farming (SDF) project. SDF was sponsored by the Singapore Armed Forces (SAF) Centre for Military Experimentation (SCME) and was completed in 2005[7]. The architecture is as shown in the figure-1.



. **Figure 1:** Hybrid Model of outliers estimation

The outlier's analysis is available in terms of Statistics methods, Distance methods, and deviation methods. Sequential exception and Index based algorithms are available [8].

Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection etc are the some of the applications are discussed and implemented in [9]. The authors further discussed the outliers are on the basis of graphics basis and modeling basis. The authors described convex hull method and Grubbs test as methods for outlier's estimation. Grubbs test is implemented for Univariate data. Distance based methods are also available in [9].

The Statistics methods have some limitations especially they are very nice for Univariate nature variables only. But in real time the data has multiple dimensions. The original author discussed role of 95% confidence interval in outlier detection [10].The Outliers exist in almost every real data set. Some of the prominent causes for outliers are Malicious activity, Instrumentation error(such as defects in components of machines or wear and tear),Change in the environment (such as a climate change, a new buying pattern among consumers, mutation in genes) and Human error(such as an automobile accident or a data reporting error)[10],pp 2. The author in his abstract pointed out the differences between Outliers and Anamolies. The author said in this paper [10] they focused on Survey on "outlier detection methods".

There are multiple ways of how descriptive learning handles outliers. If a summarization or data preprocessing phase is present, it usually takes care of outliers. For example, this is the case with grid-based methods. They simply rely on input thresholds to eliminate low-populated cells. Algorithms in the section Scalability and VLDB Extensions provide further examples. The algorithm BIRCH [Zhang et al. 1996; Zhang et al. 1997] revisits outliers during the major CF tree rebuild, but in general handles them separately. This approach is shared by other similar systems [Chiu et al. 2001]. The framework of [Bradley et al. 1998] utilizes a multiphase approach to outliers. Certain algorithms have specific features for outliers handling. The algorithm CURE [Guha et al. 1998] uses shrinkage of cluster's representives to suppress the effects of outliers. K-medoids methods are generally more robust than k-means methods with respect to outliers: medoids do not "feel" outliers. The algorithm DBCSAN [Ester et al.1996] uses concepts of internal (core), boundary (reachable), and outliers (non-reachable) points. The algorithm CLIQUE [Agrawal et al. 1998] goes a step further: it eliminates subspaces with low coverage. The algorithm Wave Cluster [Sheikholeslami et al. 1998] is known to handle outliers very well through its filtering DSP foundation. The algorithm ORCLUS [Agarwal & Yu 2000] produces a partition plus a set of outliers [11].

Basic approaches currently used for solving this problem are considered, and their advantages and disadvantages are discussed. A new outlier detection algorithm is suggested. It is based on methods of fuzzy set theory and the use of kernel functions and possesses a number of advantages compared to the existing methods. The performance of the algorithm suggested is studied by the example of the applied problem of anomaly detection arising in computer protection systems, the so-called intrusion detection systems. In this paper [12] author used kernel functions, Fuzzy functions and Neural networks are outliers detection. In Statistics method he used Regression methods and Smart sifter algorithm used for outliers detection.

In this paper they examined a subset of these techniques, those that have been designed to discover outliers quickly. The algorithms in question are ORCA, LOADED, and RELOADED. We have performed an empirical evaluation of these algorithms, and here present our results as guide to their strengths and weaknesses. ORCA can handled mixed data type and dynamic data also [13]. But in real time for technical data we can use dynamic data like in "C" language.

In real world applications due to nature of Market Uncertainty all methods should have their own identity. This context the original authors implemented all traditional and advanced methods. They discussed about Wavelet form also. As per context they used Fuzzy logic, Support Vector Machine (SVM), Neural networks etc. They proposed block diagram for outlier detection for outlier detection as follows [14].

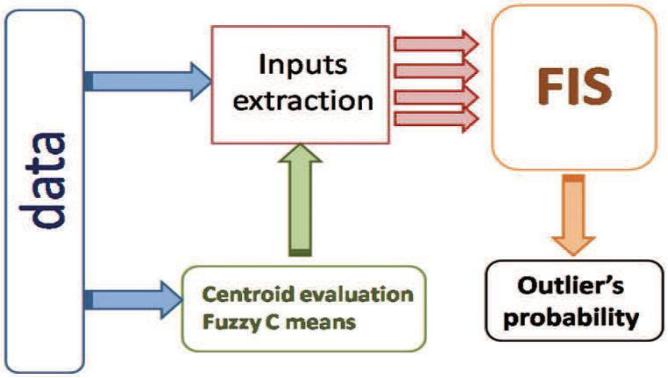


Figure 2: Block diagram of outliers with Fuzzy C means

Oracle Data mining is available as embedded along with 10g. In oracle this outliers are termed as "Anomalies" and we have separate mechanisms for these error finding [15], pp 71-75.

Hybrid model is available for outlier's detection as shown in the Figure. It uses both cluster approach and Distance approach

for outlier detection [16].he block diagram is as shown in Figure.

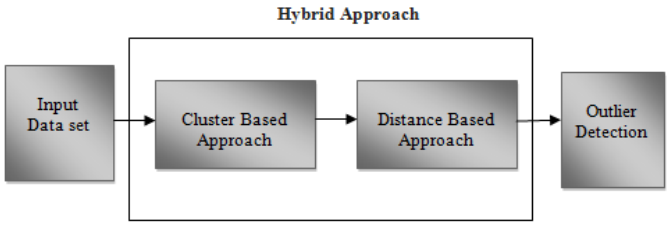


Figure 3: Hybrid Approach for Outliers detection

Analysis of scaling time and application to census using proposed algorithm. In this paper, we presented an algorithm based on randomization and pruning which finds outliers on many real data sets in near linear time. This efficient scaling allowed us to mine data sets with millions of examples. An algorithm is available in [17], pp 07.

The Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. In this paper the author had an attempt to bring together various outlier detection techniques, in a structured and generic description [18].

Quartiles are used in construction of numerical outliers. More information and different types of examples are available in[30].The step by step procedure of estimation of outliers for popular example " Monthly temperature[1] " is available in [31].

Data Mining is used to extract useful information from a collection of databases or data warehouses. In recent years, Data Mining has become an important field. This paper has surveyed upon data mining and its various techniques that are used to extract useful information such as clustering, and has also surveyed the techniques that are used to detect the outliers. This paper also presents various techniques used by different researchers to detect outliers and present the efficient result to the user .All methods are summarized in [19],pp 31.

An Empirical Bayes Approach to Detect Anomalies in Dynamic Multidimensional Arrays in the form of KDD. This paper presents mainly on spatial data but not numeric data. They have theme on multi dim data. This is once again using Statistic methods for find out the outliers [20].

3.2 Body paragraphs

Modeling is one of the technical method which is useful to describe the systems in terms of abstract or prototype or concrete methods. Unified modeling language is one the best method for all type of applications. Beside this all Research community can express their modeling in terms of either Statistics or Mathematics or both. As a core subjects both Statistics and Mathematics have very depth methods for standard procedures. Rational Rose is IBM product which is highly useful for modeling and testing also. For more details of Rational Rose visit IBM official web site.

3.2.1 Modeling with Statistical methods

Systematic sampling is a name of a wide class of sampling strategies in which selection of individual elements is following a systematic pattern. Examples are square grids of sample points laid out over an area of interest; or the selection of every 10th tree in an alley; or parallel transects.

Systematic sampling and its applications to forest inventory are best illustrated with square grids of sample points. We may imagine a transparency sheet on which this grid is printed; and this transparency is placed randomly over the map, where randomly means: randomly selected starting point and random orientation. From a sample selection point of view, it is important to state that we have only one independent selection of a sample point; after having selected the first point, all others are fixed. We defined earlier that sample size is the number of independently selected elements; an immediate conclusion is that systematic sampling is obviously a sample of size n=1. The "plot" that is being laid out then is a large cluster plot consisting of numerous sub-plots – that is, all the sample points on the systematic sample are strictly spoken sub-plots of one single cluster that is spread out over the entire class [37]. This is special case of stratified Sampling [38].

The main difference between cluster sampling and stratified sampling is that in cluster sampling the cluster is treated as the sampling unit so analysis is done on a population of clusters (at least in the first stage). In stratified sampling, the analysis is done on elements within strata. In stratified sampling, a random sample is drawn from each of the strata, whereas in cluster sampling only the selected clusters are studied. The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. This contrasts with stratified sampling where the main objective is to increase precision [39].

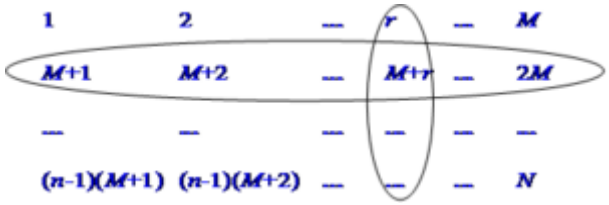


Figure 4: Cluster Sampling

The standard error is the standard deviation of the sampling distribution of a statistic. The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate. For a value that is sampled with an unbiased normally distributed error, the above depicts the proportion of samples that would fall between 0, 1, 2, and 3 standard deviations above and below the actual value. The following diagram shows about the nature of standard error [40].

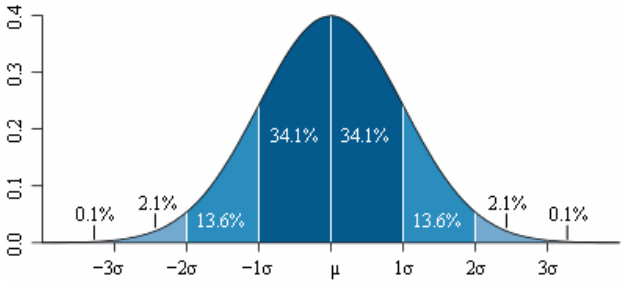


Figure 5: Shows the formation of standard error

All fundamental terms of statistics are explained in [41],[42] and [43]. In this[42] the author discussed more advanced features such as Scaling and Shifting. The author proposed theorem also is as follows.

Theorem Suppose $\{ x_1 , x_2 , \cdots , x_n \}$ is a data set with s standard deviation s_x . Then if $w_i = \frac{x_i - \bar{x}}{s_x}$, $\bar{w} =$

In descriptive statistics, the quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. A quartile is a type of quintiles. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. The second quartile (Q2) is the median of the data. The third quartile (Q3) is the middle value between the median and the highest value of the data set. In applications of statistics such as epidemiology, sociology and finance, the quartiles of a ranked set of data values are the four subsets whose boundaries are the three quartile points. Thus an individual item might be described as being "in the upper quartile"[48]. All most all data mining tools are using this method for outlier's detection especially for numeric data.

3.2.1.1 Modeling Data with Easy fit 5.5 and Stat fit

The first step is loaded into tool and observes the Statistic values for Mean, Mode, Average, Standard deviation or any other values which are necessary for our experiment.

Descriptive Statistics

Statistic	Value
Sample Size	404
Range	0.99469
Mean	0.48215
Variance	0.07635
Std. Deviation	0.27632
Coef. of Variation	0.57311
Std. Error	0.01375
Skewness	0.14057
Excess Kurtosis	-1.0470

Percentile	Value
Min	0.00531
5%	0.03802
10%	0.10828
25% (Q1)	0.26818
50% (Median)	0.46046
75% (Q3)	0.71728
90%	0.89107
95%	0.93004
Max	1

Figure 6: Shows the total number of records

In probability theory, the normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution—a function that tells the probability of a number in some context falling between any two real numbers. For example, the distribution of income measured on a log scale is normally distributed in some contexts, as is often the distribution of grades on a test administered to many people. Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known. The authors sampled and tested their data with Stat Assistant 5.5 and Easy fit 5.5 professional. This data is already published in their publication [62],pp 26. The various functions of Normal distribution are Density function, Cumulative function, Hazard function, Cumulative Hazard function, Survival function; Probability difference, Q-Q plot, and P-Empirical plot are available. Q-Q Plot, P-P Plot and Probability difference curves are modeled with East fit 5.6 and remaining are modeled with Stat Assistant 5.5 are as shown below in the figures. The below first figure shows Histogram curve Normal Distribution with its density function.

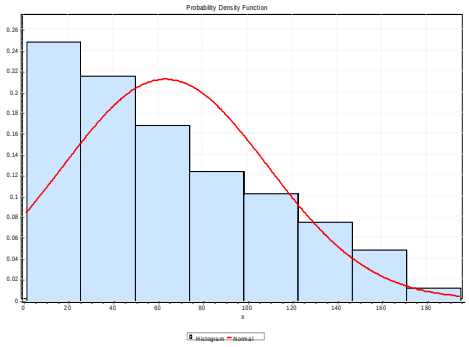


Figure 7: Shows Histogram Density Function

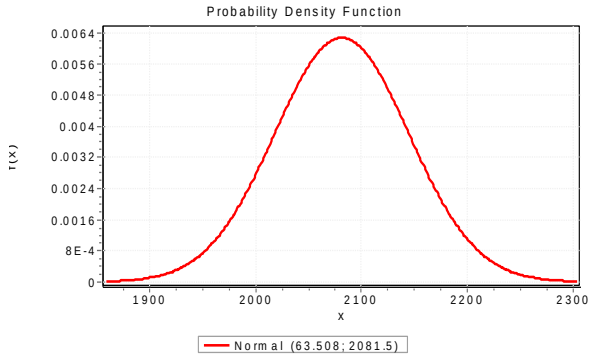


Figure 8: Shows Density function without Histogram

The authors observed all types of functions for Normal Distribution and all the pictures are not shown in the figure due to Space complexity. The different types of available curves are mentioned just above the Figure 7. The authors examined two types of data set one is full data set and another is 50% Sample data(Cluster Sample) and result is very surprise. The performance of two sets is almost same and 50% Sample data is more stable than full data set. The following figure shows the performance of two data sets.

53	Reciprocal	0.37834	54	130.23	51
54	Rice	0.16196	42	60.629	41
55	Student's t	0.93831	58	1418.3	58
56	Triangular	0.03866	7	45.617	8
57	Uniform	0.10449	33	231.99	54
58	Weibull	0.04072	11	49.451	16
59	Weibull (3P)	0.04117	12	49.396	15
60	Gen. Extreme Value	No fit			
61	Johnson SU	No fit			

Figure 9: Shows fitness test for two data sets

From the above figure it is observed that the full data set and 50 % data set sample has almost same performance. Either full data set or 50% data sample set the tests Extreme Value and Johnson are not valid or unfit.

3.2.1.2 Modeling with MS-Excel

In probability theory 95% confidence interval has very vital role, In Normal distribution values under bell curve are generally 95% confidence interval values. But in real time applications the interval may be 25%,50%,75% also available. The availability of equal size intervals are smoothes the process. The authors are applied their data for 95% confidence intervals on the basis of residuals in linear regression. In this mainly three values are available. First Upper Confidence level(UCL),second lower confidence level and third value is expected value. The authors are applied Linear, Quadra, Poly

with 3 degree and Poly with 4 degree. The following figure shows Open interval, Middle interval and Closed interval for different types of Regression curves.

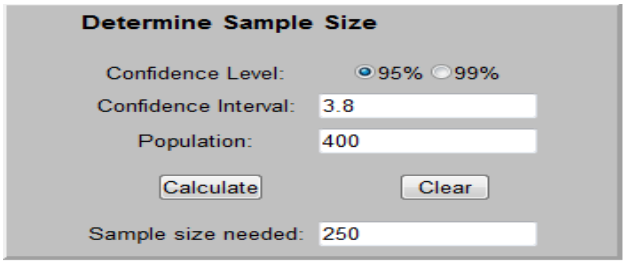


Figure 10: Shows Sample size

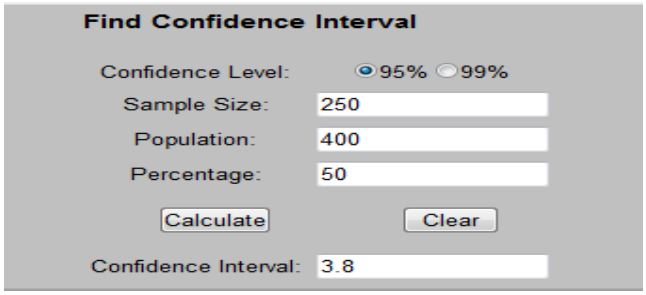


Figure 11: Shows Confidence interval

The above two figures shows estimation of Sample size and Confidence Interval for population of 400.The authors did not test for 99% confidence Interval due very rare use in real time application. All business events are generally dynamic nature. Different online helps are available for CI estimation. The authors used online help for CI estimation for data set with full population as follows. The authors took online help from http://www.surveysystem.com/sscalc.htm, http://www.danielsoper.com/statcalc3/calc.aspx?id=96. The result is as follows as shown in the figure.

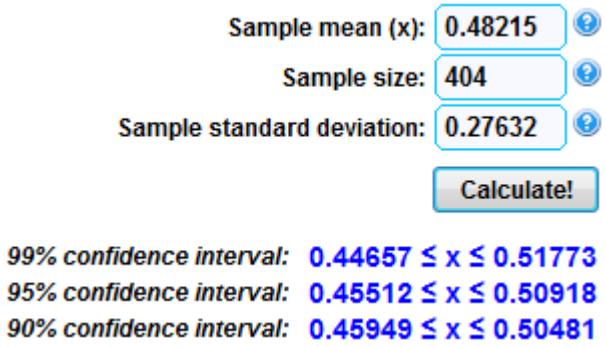


Figure 12: Show different confidence intervals

The below figure shows the 95% confidence interval for open interval of all types of curve. The maximum value available is 0.335 approximately. All values in the Figure are >0 which indicates that all values are available in first quadrant only. In Excel sheet as per our requirement we can model the Figure. Here we have three values only. So we chose this model. If we have Open stock then it is not applicable. The authors applied same procedure for Middle and Closed intervals. The Figure are as follows. The authors implemented Excel sheets for confidence interval preparation for High value, Low value and expected value. Other options are also available as follows.

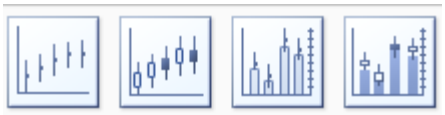


Figure 13: Shows types available in Stock chart

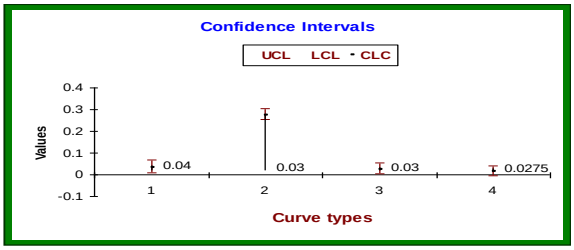


Figure14: The above figure shows Middle interval is available for Quadra and Poly-4 curve only. Highest value is 0.2 only

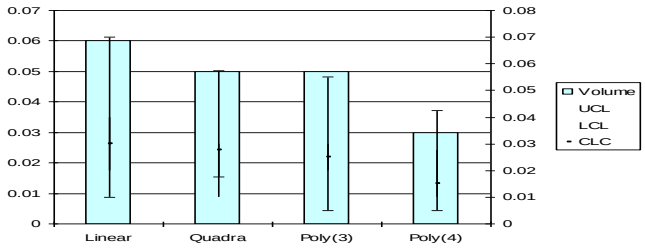


Figure 15: Shows CI for open and closed intervals

The Closed interval is available for all types of curves, ie Linear, Quadra, Poly-3 and Ploy-4 and maximum value is 0.06 for linear curve. Sampling techniques are very important in Statistical methods. The authors tested the full data and 50% sample data. The result is very surprise. The reason is 50% data sample is more fit than full data set as shown in figure. Parito, General Extreme value and Pearson test are not valid for full data set. Very surprisingly Parity test holds good for 50% sampling data. So from above testing it is verified that 50% data is highly enough for sampling testing.

Table 1: Shows the role of size of data set for fitness

Sno	Data size	Test Name	Result
1	Full	Parito Test	Fail
2	50%	Parito Test	Pass
3	Both	Extream value, Pearson	No fit
4	Both	Normal Distribution	Pass
5	Both	Johnson	No fit

There are totally 61 tests are available and only 3 tests are not suitable for given data set. It means that 96. % success and 3.3% has error data or extra data or unfit data. For sampling techniques only 50% data is act as complete representative of given data set. It is not necessary to take entire data set for testing instead we can assume or consider 50% or even 45% also enough to estimate the nature of data or fitness of data. Depend upon requirement we can choose the distribution. In complex applications or data set has very big such as "Population of Country" then we can apply more distributions for satisfy the all conditions of given problem. In this context authors used Normal distribution with all its functions as shown in the figure. Results have error with in the p-value.

Pearson's chi-squared test ($\chi 2$) is the best-known of many chi-squared tests (Yates, likelihood ratio, portmanteau test in time series, etc.) – statistical procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were

first investigated by Karl Pearson in 1900. In contexts where it is important to improve a distinction between the test statistic and its distribution, names similar to Pearson χ-squared test or statistic are used. The Density function curve is as follows

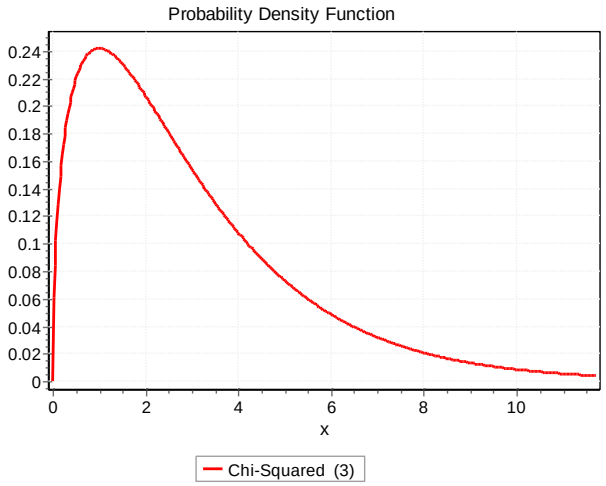


Figure 16: Red line shows failure or no fit.

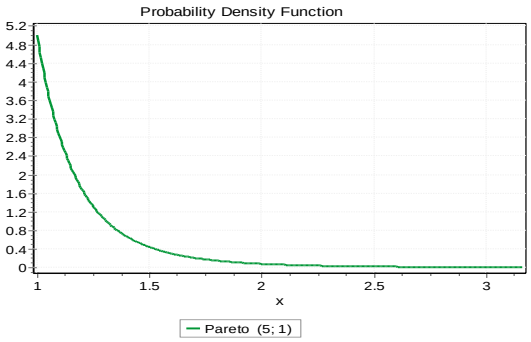


Figure 17: Green line shows success of test for 50% sample

3.2.2 Modeling with Curve fitting Methods

Mathematical modeling is one of the strong approach to describe the system and solve the system also. The core mathematics has very strong relation with all domains in the real world. Beside the Mathematics model purely we can integrate this model with Statistical model or Linear programming problem model or any other real world problem. The authors here integrate this curve fitting model with Statistical and Data mining tools. There are number of curve models are available. They are

- Linear Curve
- Quadra Curve
- Polynomial Curve with degree 3
- Polynomial Curve with degree >3
- Power Curve
- Logarithmic Curve
- Exponential Curve.
- Moving towards Curve

3.2.2.1 Modeling with Curve Expert 32 tool

As per our need of modeling we can assume or fix some intercept for maintain the uniform nature for a given problem. The authors are here fixed at 0.5 as given intercept for all curves where ever applicable. Log and Power curves do not have intercept. Moving towards curve has almost independent on curve equation, R 2 value and intercept. The authors observed the following curves as shown below. They tested for full data and 50% sample data also. The data set is loaded into Curve software with 405 data values are as shown in the figure.

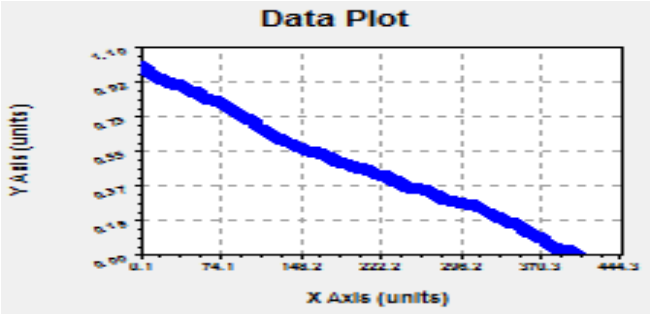


Figure 18: Shows data is loaded into tool

The authors repeated the experiment for linear curves, Quadra and Poly curves. The observed regression residuals are shown in the Figure. Poly curves with 3 rd degree and 4 Th degree are available. The standard error of various curves is ≤ 0.05 except polynomial with degree 4. Its value is 0.0135. Some of the residuals are positive and some are negative as shown in the Figure. Positive residuals are marked with green color and negative are marked with red color for ease of use and better understanding. Error is >0, so the values which are available other than first quadrant are neglected. The authors recorded coefficient values also up to 4 th degree of polynomial. The below figures gives us the residuals, Standard error and co-relation coefficient. The coefficients (a,b,c and d) also recorded for various curves up to degree 4. The authors recorded (snap-shot) all residuals of linear regression curves and their coefficents(a, b, c, d) as follows. The following snap-shot shows the residuals of linear curve with co-efficient (a , b) as follows.

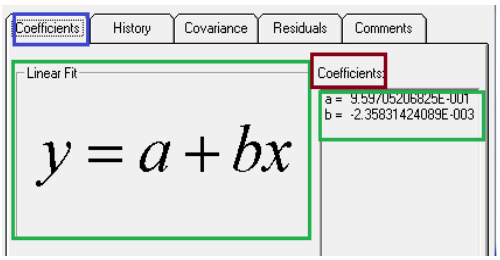


Figure 19: Shows data is loaded as linear curve

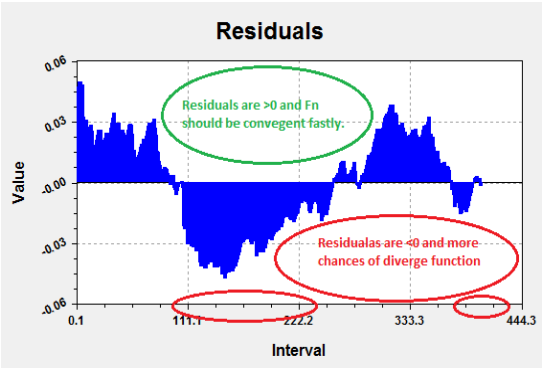


Figure 20: Shows data is loaded as linear curve

The authors repeated the same procedure up to polynomial degree-4 and they observed the residuals are available at different intervals with both positive and negative values. The authors repeated the experiment with intercept 0.05 as per uniform and standardize the procedure. At three intervals the

different residual values are available as shown in the figure 20. The authors considered the peak value is highest value and next peak as minimum value for more consistent in the numeric digits. All the results are recorded in table-4(Section 5.1). After residual estimation the authors took snap-shot of Standard Error and Correlation Coefficient as follows.

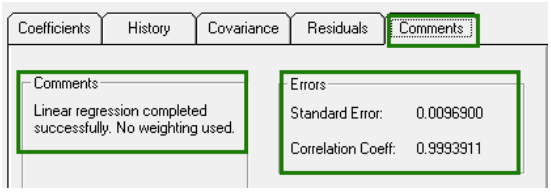


Figure 21: Shows data is loaded as linear curve

More details are available about Standard Error in figure-5 and Correlation Coefficient is very near to 1 means all variables have strong relation with each other and their values are highly stable. The authors repeated the same procedure for all types of curves for SE and CC and their values are noted as table form as shown in Table-2. Except Quadra curve
All curves have near by 0.05(p-standard value).

TABLE-2: Shows the Standard Error and Co-relation coefficient of curves.

No	Curve Name	SE	CC
1	Linear Curve	0.0229	0.9966
2	Quadra Curve	0.1822	0.9978
3	Polynomial-3rd degree	0.0097	0.9994
4	Polynomial-4th degree	0.0135	0.9988

All CC values have 0.99 as two decimals accuracy in cases and indicates that values and relation between them is consistent and standard. So we can expect consistent results.

3.2.2.2 Modeling with Graphs (MS-Excel)

The authors are obtained the following graphs and equations when implemented. They are as follows. Authors tested with full data and 50% sample data. In both of the cases they recorded the values as follows. For full data implementation the default (represented with blue color diamond) line name Series1 is very thick due to big data size (400 records). Authors used Excel for graphs generation with Stock option from Chart object. The Curve options are shown in the below figure.

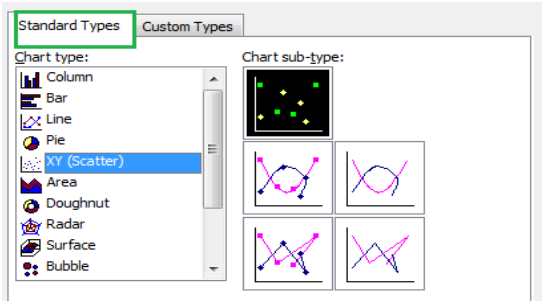


Figure22: Shows chart sub types

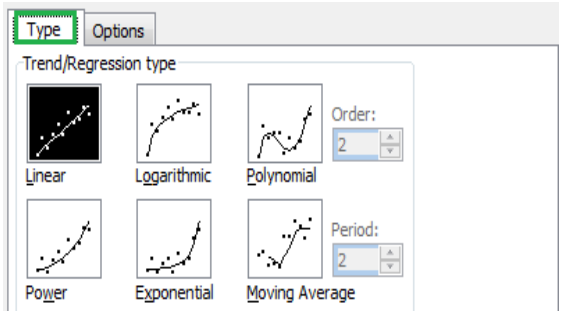


Figure 23: Shows chart sub types

The authors repeated the experiment for all available curves as shown in figure-23 with Chart sub-type as shown in the figure-24 they obtained different equations and and different curves. The authors repeated the above procedure for full data set and 50% sample data set.. All the outputs are not shown in the figure. But sample figures are only recorded.

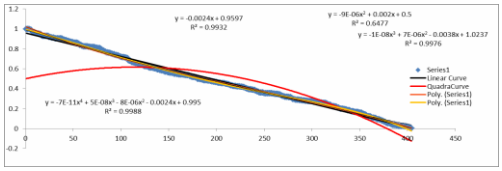


Figure 24: Shows full data testing (Linear,Qudra and Poly higher-orders)

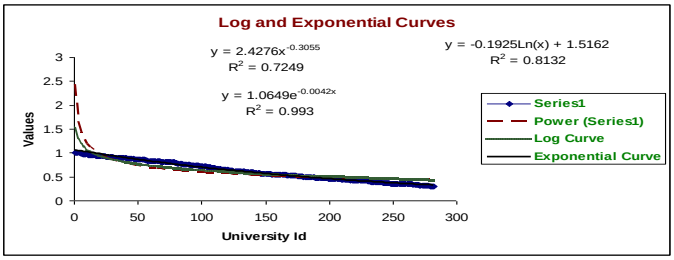


Figure 25: Shows about Log, Power and Exponential curves (Full data set)

The authors repeated the same procedures for 50% Sample data and recorded the values and graphs are not shown in the figure and only equations are tagged here. They are as follows. **y = -0.0024x + 0.9597 (1)**
$y = -9E-06x^2 + 0.002x + 0.5$ ((2)
$y = -1E-08x^3 + 7E-06x^2 - 0.0038x + 1.0237$ (3)
$y = -7E-11x^4 + 5E-08x^3 - 8E-06x^2 - 0.0024x + 0.995$ (4)
$y = 0.001x$ (5)
$y = -9E-06x^2 + 0.002x + 0.5$ (6)
$y = 6E-08x^3 - 4E-05x^2 + 0.006x + 0.5$ (7)
$y = -6E-11x^4 + 4E-08x^3 - 5E-06x^2 - 0.0027x + 1.0007$ (8)
The first four equations are obtained with full data set, for different curves(dark blue color). The next four equations from 5-8 light blue color are obtained with 50% Sample data set.

3.2.3 Modeling with Data Mining tools

There are lot of Data mining tools are available in the market. Some tools are freeware and some are shareware. In shareware some are server oriented and some are client oriented. Lot of online help also available for Data mining Modeling. Especially you tubes are very nicely working for educational purposes. Authors observed lot of video tutorials on these Data mining tools and some are available in audio and video and some are in general files such as .PDF, doc,.docx, and .ppt form. Some of the Data mining tools are available along with server as embedded form. Oracle 10g Data mining add-in, MS-

SQL Data mining add-in are the examples of embedded form. Along with this pure MS-Excel add-in also available for very reasonable prices and they are giving 30 days trail period for this type of software's. Lot of details are available in www.kdnugget.com web site. In this site all types of Data mining tools are available with details. The following Figure shows this concept very nicely. The authors referred online material and tools manuals for implementation. The authors are successfully implemented all these above said tools in their published papers [62], [63],[64]. Angoss has data mining tool called Knowledge STUDIO and they have outliers handling methods in their documentation (Refer Knowledge STUDIO Manual, pp 183).

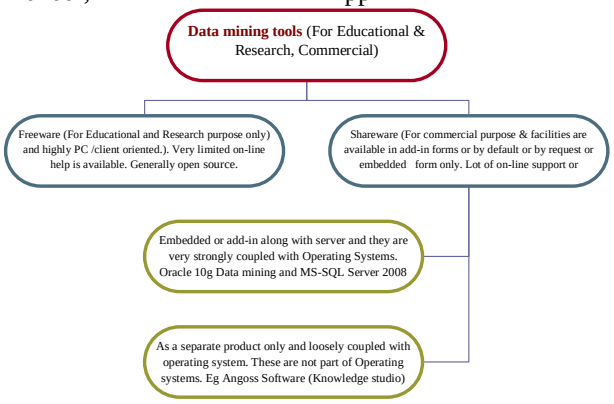


Figure 25: types of data mining tools

Authors tested the data with different data mining tools. They are as follows.

- Tanagra
- Weka
- Orange
- R-GUI(Rattle) and
- Rapid miner

The authors observed different output for outlier's estimation by using above said tools. The results are recorded in the form of snap-shots and shows different outputs. Authors implement all these tools using online help sequentially [26],[21],[27],[23],[22].

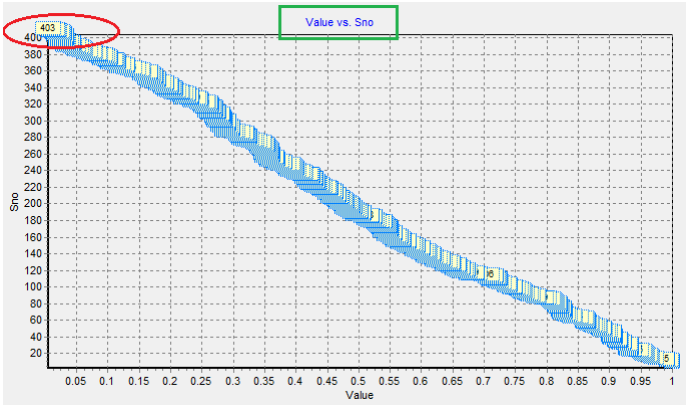


Figure 26:Tanagra implementation for outliers

The authors observed that there are only 3 outliers are available in the form of extreme boundary values. The maximum scale considered by Tanagra tool is only 400 and there are 403 elements are available in data set. The extreme boundary values are marked with red color and two columns are available. One column for University Id and another is its bi-cluster value as described in their publication [62], pp. 31-32.

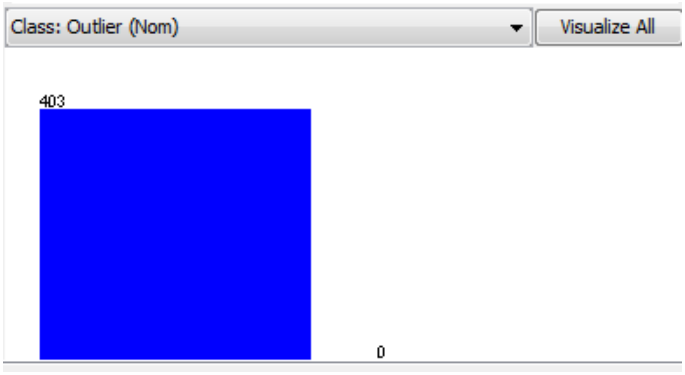


Figure 27: Shows 0 outliers with Weka

The authors observed that there are zero outliers are available with Week implementation and data size is 403 records. The authors repeat the experiment for extreme values as follows.

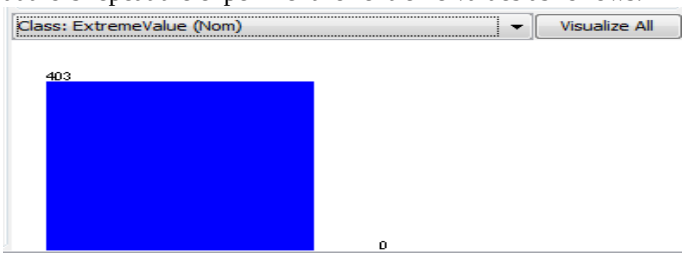


Figure 28: Shows 0 extreme values.

The authors observed that there are no outliers in a given data set. They tested for both outliers and extreme values classes and the results are shown as above figure.

table (Outliers) | table (Data with z-score) | table (Inliers)

	Sno	Value	Z score
1	2	1.000000	13.431
2	3	1.000000	4.417
3	403	0.011466	4.417
4	404	0.005310	13.431

Figure 29: Orange tool shows 4 outliers are available

Authors are observed that during the experiment there are four outliers are formed and they are at initial and boundary values and it is quite nature outliers or errors are generally available at either initial point or final point of an interval. As per software testing concepts they are very they have very low intensity error and eliminate in future plan as higher version. For more details on this testing the values visit any testing tool such as Win Runner, Load Runner, and QTP etc. Error rate by Orange tool is 0.0066 and it is very reliable error and we can assume, the data is almost free from error. Along with the values of outliers z-values are also available. More details are available in [24],[25] on Z-score concept. Authors did not use IBM SPPSS tool[25] for Z-score implementation since they want to use Orange tool.

No.	Variable	Data Type and Number Missing
1	1	Numeric [2.00 to 404.00; unique=403; mean=203.00; median=203.00].
2	0#997671	Numeric [0.01 to 1.00; unique=376; mean=0.48; median=0.46; ignored].
3	RRC_0#997671	Numeric [-1.73 to 1.88; unique=376; ; median=-0.08]. No code export.
4	RRK_0#997671	Numeric [1.00 to 403.00; unique=376; mean=202.00; median=202.00]. No code export.
5	IZR_0#997671	Numeric [0.01 to 1.00; unique=376; mean=0.48; median=0.46]. No code export.
6	IMN_0#997671	Numeric [0.01 to 1.00; unique=376; mean=0.48; median=0.46]. No code export.
7	IMD_0#997671	Numeric [0.01 to 1.00; unique=376; mean=0.48; median=0.46]. No code export.
8	IMO_0#997671	Numeric [0.01 to 1.00; unique=376; mean=0.48; median=0.46]. No code export.
9	BQ4_0#997671	Categorical [4 levels]. No code export.

Figure 30: Data set loaded and processed with R

The authors are observed all options with "Rattle" GUI of R and observed that No code export takes place for data modification. There is no change in mean or median even though the new classes are added to dataset. It indicates that modification is not necessary and we can assume data set is free from error. Authors did not use R –Command line mode. There is lot of information about numerical outliers is available on r-bloggers.com, and in that path we can find outliers implementation with Quartiles. Lot of resources is available in the form of R-journal, online help, r-bloggers.com, and R Manual Guide [45],[46],[47].

Role	Name	Type	Statistics	Range	Missings
outlier	outlier	binominal	mode = false (398), least = true (5)	false (398), true (5)	0
regular	1	integer	avg = 13.500 +/- 7.071	[2.000 ; 25.000]	379
regular	1(1)	numeric	avg = 0.956 +/- 0.026	[0.919 ; 1.000]	379
regular	Name	polynominal	mode = ABABPP (6), least = IUYTWRE (	ABABOI (6), ABABPP	0
regular	Id	integer	avg = 203 +/- 116.480	[2.000 ; 404.000]	0
regular	Value	real	avg = 0.481 +/- 0.275	[0.005 ; 1.000]	0

Figure 31: Shows implementation with Rapid miner

Good data preparation is very mandatory for minimize the outliers. All data preparation methods in real time are generally numeric or Yes or No forms only. We can find good example and implementation with Rapid miner [33]. Market trends are very well examine for outliers estimation and implement as Neural networks with Rapid miner[34],[35]. The below figure shows their implementation with Rapid Miner

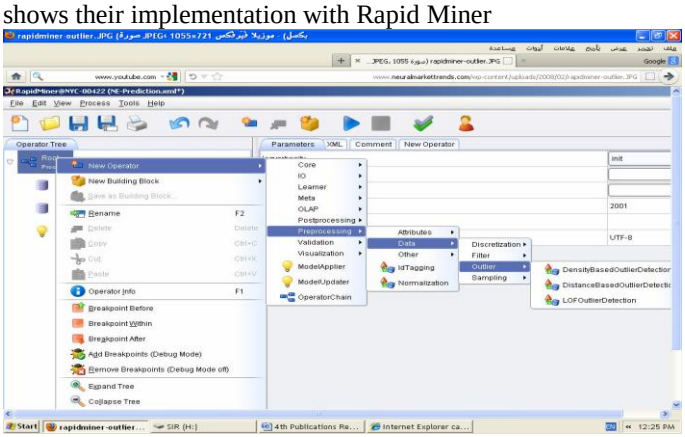


Figure 32: Shows outliers' implementation of Rapid miner

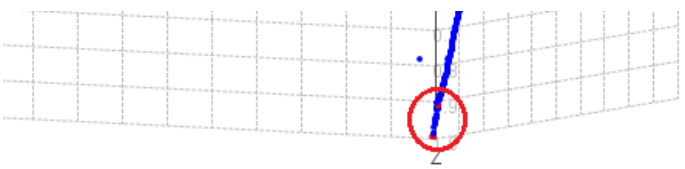


Figure 33: Shows the five outliers are available on Z-axis.(Distance Method)

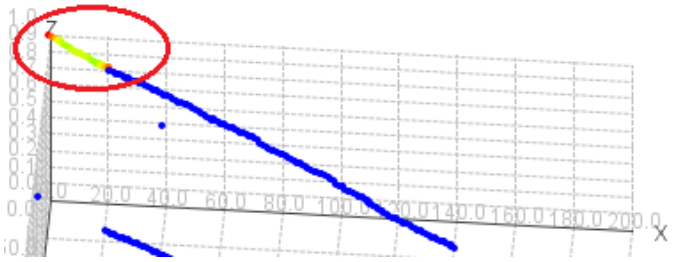


Figure 33: Shows the five outliers on Z-axis (Local Cluster Method)

The above experiments with different Data mining tools are summarized in below table as follows. The authors observed that each and every tool implementation method is different and not shown in any figure. Table-3 is performance table.

Table-3. Performance of various tools for outliers

Sno	Tool Name	Outliers
1	Tanagra	3
2	**Weka**	**0**
3	Orange	4
4	**R(Rattle)**	**0**
5	Rapid-Miner	5

4. SUMMARY

The authors are used Statistical, Mathematical and Data Mining tools. The modeling for outliers estimation can be represented in the above said methods. Each and every method has its own advantages and disadvanges. The implementation details are available in limited manner due to space limit. Overall summary of modeling is represented in tree structure as follows

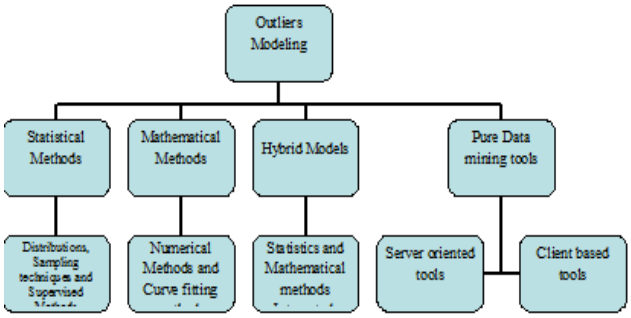


Figure 34: Shows overall modeling for outlier's estimation

5. RESULTS

The authors are tabulated all results in terms of all types of modeling techniques. The results are available in table form or figure form as per requirement and flexibility. All numeric values are rounded and consider up to 4 decimal points accuracy for maintain the uniform result.

5.1 Results for Confidence Interval construction (Using Excel sheet)

The authors are used MS-Excel for confidence interval as described in 3.2.1.2. Authors are applied only first type of sub chart as described in 3.2.1.2 and other options are neglected on the basis of model. Authors used MS-Office 2003 for

Statistical and Mathematical modeling**.**

Table-4: Confidence Interval estimation at Interval level

Confidence Interval Estimation			
Type	Open Interval	Middle Interval	Closed Interval
Linear	[0,110]	NA	[275,333]
Quadra	[0,121]	[25,111]	[222,333]
Poly(3)	[25,111]	NA	[275,333]
Poly(4)	[0,25]	[50,75]	[76,111]

Table-5: Confidence Interval Values Preparation for Open Interval

Confidence Interval Values Preparation for Open Interval				
Type	UCL	LCL	CLC	Max
Linear	0.04	0.02	0.03	0.06
Quadra	0.03	0.01	0.0275	0.05
Ploy(3)	0.03	0.02	0.025	0.05
Poly(4)	0.0275	0.01	0.015	0.03

Table -6: Confidence Interval Values Preparation for Closed Interval

Confidence Interval Values Preparation for Closed Interval				
Type	UCL	LCL	CLC	Max
Linear	0.04	0.02	0.03	0.06
Quadra	0.03	0.01	0.0275	0.05
Poly(3)	0.03	0.02	0.025	0.05
Poly(4)	0.0275	0.01	0.015	0.03

Table-7: Confidence Interval Values Preparation for Middle Interval

Confidence Interval Values Preparation for Middle Interval				
Type	UCL	LCL	CLC	Max
Linear	NA	NA	NA	NA
Quadra	0.035	0.02	0.025	0.05
Poly(3)	NA	NA	NA	NA
Poly(4)	0.025	0.175	0.180	0.03

- Linear and Quadra curves have near values for open and close intervals.
- Polynomial (3) and Polynomial (4) have near values for open and close intervals.
- Quadra and Polynomial (4) have close values for middle values**.**

Table 8: R^2 VALUES FOR 50% Sampling and Full Dataset

Dataset	Linear	Quadra	Poly-3	Poly-4
50%	-1.803	0.6414	0.764	0.989
100%	0.9933	0.6477	0.9976	0.9988

- For Quadra Curve sample size has no effect.
- For Polynomial (4) curve also sample size has no effect.
- For Log curve also the values are very near.
- For Exponential Curve the values are different.
- Polynomial (3) has very far values.

5.2 Results for Data testing (Easy fit 5.5 Professional and Stats Assistant 5.5)

Authors used Normal Distribution with all its functions are as described in 3.2.2.1 and conduct fitness test with full data and 50% Sample data. Very surprisingly 50% Sample data is highly enough for sampling test and full data set is not necessary as shown in 3.2.21 as Goodness of fit-test.

- For Sampling testing 50% data is enough and full data set is not necessary and full data has same effect of 50% sample testing.
- Extreme value, Johnson tests are failed tests by both 50% sample and full data. Parity test is fit for 50% sample data but not full data5.3 Results for Distribution functions

Authors used Noramal Distribution and its all its functions as described in 3.2.2.1 and all functions have fitness.For diagram refer 3.2.1.1.

5.4 Results for Curve fitting and Regression Analysis

5.4.1 Curve fit 32 tool

All residuals are recorded as shown in 3.3.2.1 the residuals are available in the form of graph values but not in numeric values. Some residuals have negative and some have positive values. It is very difficult to find out all residuals by trail and error method. It is possible to find out the maximum value point of residual by finding the peak values of graph. Some intervals have positive or negative values but some intervals have both values.

- All co-efficient values of all types of are available with scientific notation.
- For Standard Error all curves have different values where as for co-efficient all values are almost same as shown in Table-2.

5.4.2 Excel Sheet using Scatter plot graph

The equations for curve fitting methods for all types of curve are already available in 3.2.2.2.The R^2 values for different curves are as shown in the below table.

- Linear curves are highly affected by Sample size.

Table 9: Error Estimation between full data set and 50% Sample set

Sno	Curve	50%Data	Full Data set	Error
1	Linear	-.1.803	0.9933	**2.7963**

2	Quadra	0.6414	0.6477	0.0063
3	Poly-1	0.764	0.9976	0.2336
4	Ploy-2	0.989	0.9988	0.0098
5	Power	0.4301	0.7249	0.2948
6	Exponential	0.7699	0.933	0.1631
7	Logarithmic	0.7764	0.8312	0.0368

- Red color values are strongly error and out of list.
- Green color values are feasible and optimal solution.
- Yellow shaded is warning and may be lead to error.
- Blue color values are feasible but not optimum solution

6. CONCLUSION

Practically or in real application higher degree curves are more complicated and need more studies for estimate the optimum solution. From above table it is observed that Quadra curves have best fit when compare to other curves. So the clusters which are available at above said value has good stability with degree 2.Logarithmic curves are not advisable due to zero or negative values in the curves. Quadra Curves are most suitable for available data in the view of Bi-clusters of Indian Universities.

ACKNOWLEDGEMENTS

Thanks to C.M.J University, Shilling, India for giving the opportunity to do the Research in their organization. Special thanks to all teachers from school level to Research level.

REFERENCES

[1] Mercedes Andrade-Bejarano and Nicholas T. Longford," Outliers in Mixed Models for Monthly Average Temperatures", AUSTRIAN JOURNAL OF STATISTICS Volume 39 (2010), Number 3, pp. 203–221.

[2] www.jstor.org/stable/2347159

[3] http://en.wikipedia.org/wiki/Outlier

[4] http://en.wikipedia.org/wiki/Outlie

[5] http://www.researchgate.net/publication/224673269_A_Comparative_Study_for_Outlier_Detection_Techniques_in_Data_Mining/file/d912f513697367bc17.pdf

[6] http://www3.ntu.edu.sg/SCE/pakdd2006/tutorial/chawla_tutorial_pakddslides.pdf

[7] http://harvest.nps.edu/scythe/Issue3/Scythe3-COADataMining-Article.pdf

[8] http://www.rdatamining.com/examples/outlier-detection

[9] http://www.cse.yorku.ca/~jarek/courses/6412/lectures/Outliers.ppt

[10] http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap10_anomaly_detection.pdf

[11] http://www.cs.sfu.ca/CourseCentral/741/jpei/slides/OutlierDetection.pdf

[12] http://www.bradblock.com.s3-website-us-west-1.amazonaws.com/Outlier

[13] http://dms.stat.ucf.edu/STA6714/Lecture04/SUPP.pdf

[14] http://cdn.intechopen.com/pdfs/4666/InTech-Outlier_detection_methods_for_industrial_applications.pdf

[15] http://www.cse.ohio-state.edu/dmrl/papers/kddws05.pdf

[16] http://www.ijarcsse.com/docs/papers/June2012/Volume_2_issue_6/V2I6001.pdf

[17] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129.pdf

[18] http://epic.org/privacy/airtravel/nasa/study.pdf

[19] http://ijcsi.org/papers/IJCSI-9-1-3-307-323.pdf

[20] http://research.ijcaonline.org/volume67/number19/pxc3887223.pdf

[21] http://www.dmargineantu.net/AD-KDD05/DMMAD-2005.WorkshopNotes.pdf

[22] http://www.youtube.com/watch?v=WrjpO7CmUoQ

[23] http://www.youtube.com/watch?v=C1KNb1Kw-As

[24] http://www.youtube.com/watch?v=ckxEZDN1iok

[25] http://www.youtube.com/watch?v=_86q-hn_3DQ

[26] http://www.youtube.com/watch?v=DDpym2j_ILY

[27] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Outliers_Detection.pdf

[28] http://orange.biolab.si/docs/latest/reference/rst/Orange.data.outliers/#

[29] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.570&rep=rep1&type=pdf

[30] http://www.r-bloggers.com/finding-outliers-in-numerical-data/

[31]http://wiki.answers.com/Q/What_is_an_outlier_in_math

[32] http://www.wikihow.com/Calculate-Outliers

[33] http://www.dfki.de/web/forschung/publikationen/renameFileForDownload?filename=slides.pdf&file_id=uploads_1635

[34] http://www.simafore.com/blog/bid/101387/A-simple-example-to-show-value-of-good-data-preparation-for-analytics

[35] http://www.siam.org/meetings/sdm10/tutorial3.pdf

[36] http://wiki.awf.forst.uni-goettingen.de/wiki/index.php/Systematic_sampling

[37] http://www.neuralmarkettrends.com/2008/02/trimming-outliers-in-rapidminer.html

[38] http://www.neuralmarkettrends.com/wp-content/uploads/2008/02/rapidminer-outlier.JPG

[39] http://en.wikipedia.org/wiki/Stratified_sampling#Practical_example

[40] http://en.wikipedia.org/wiki/Cluster_sampling

[41] http://en.wikipedia.org/wiki/Standard_error

[42] www.stat.tugraz.at/AJS/ausg093/093Al-Saleh.pdf .

[43] classweb.gmu.edu/tkeller/HANDOUTS/Handout2.pdf

[44] http://www.pitt.edu/~super7/43011-44001/43911.ppt

[45] http://en.wikipedia.org/wiki/Quartile
[46] journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdF
[47] books.google.com.om/Outliers with R and Rattle.
[48] http://zoo.cs.yale.edu/classes/cs445/misc/mar13lae08.pdf
[49] Mr. Srinatha Karur, Prof. Ramana Murthy, "Survey and Analysis of University Clustering", IJAIA, vol 4, No 4, 2013 July.
[50] Mr. Srinatha Karur, Prof. Raman Murthy, "Local Clusters formation for Indian Universities", IJAIA, vol4, No 5, 2013 September.
[51] Mr.Srinatha Karur, "Prof. Ramana Murhty"," Data Preparation and Analysis for Andhra Pradesh Clusters", IJSBAR, ISSN 2307-4531.

Mr.Srinatha Karur received the M.C.A. and M.Tech (IT) degrees in Core Computer Applications and IT from Gulbarga University,Gulbarga and Punjabi University, Punjab, India, in 1997 and 2004, respectively. During 1997-2004 he joined as IT faculty and worked as HoD also.In 2005 onwards the author entered into technical line and completed Oracle DBA in 2007-2008 period as a both OCA and OCP. At present he is working in Government Engineering College, Ibra, Sultanate of Oman

. **Dr.M.V.Raman Murhty** is working as a Sr.Professor and Director of School of Mathematics & Computer Science, Osmania University, and Hyderabad, India. He has lot of international publications and multi intelligence in core Mathematics and Computer Science Engineering. His profile shows his all-rounder skills and knowledge in multi core subjects.

.