Volume-5, Issue-5, October-2015
International Journal of Engineering and Management Research
Page Number: 304-311

Empirical Analysis of Binary Search Worst Case on Two Personal Computers Using Curve Estimation Technique

Dipankar Das[1], Arnab Kole[2], Shameek Mukhopadhyay[3], Parichay Chakrabarti[4]
[1,2,3,4]Assistant Professor, The Heritage Academy, Kolkata, INDIA

ABSTRACT

Searching is one of the most basic and fundamental algorithms of the computer science. There are many types of searching technique e.g. Binary Search, Linear Search etc. Binary Search follows divide and conquer technique. The present study is an empirical analysis of Binary Search in the worst case on two personal computer having same hardware and software configurations. In this paper the objectives of the researchers are to find out that whether the two personal computers show identical behavior when performing Binary Search in the Worst Case scenario and to identify the best curve(s) along with its mathematical model(s) that can be fitted to the data points (Average Searching Time in the Worst case versus Number of Data Elements) for both the personal computers. The researchers have used simple Graphical representation, Mann-Whitney U test, Curve estimation technique, F-test and Residual analysis for this paper and came to the conclusion that both the personal computers exhibits different behavior in terms of execution time while performing Binary Search in the Worst Case scenario but both the datasets can be best fitted to Compound, Growth and Exponential curves.

***Keywords*----** Binary Search, Mann-Whitney U test, Curve Estimation technique, F-test

I. INTRODUCTION

Experimental algorithmics is the area within computer science that uses empirical methods to study the behavior of algorithms [1]. In scientific method the word "empirical" refers to the use of working hypothesis that can be tested using observation and experiment [2]. Searching is one of the most fundamental or basic algorithm of computer science. There are different types of searching algorithms e.g. Linear Search, Binary Search, Interpolation Search etc. We know that Binary search relied on divide and conquer strategy to find a value within an already sorted collection [3]. The worst case for binary search is when the searched value is not in the set [4][5][6]. The worst case running time is given by O(log N) [6]. The present study is aimed at an empirical analysis of Binary Search in the worst case scenario on two personal computers.

II. RELATED WORK

Kumari, Tripathi, Pal & Chakraborty (2012) had done a statistical comparison between linear search and binary search for binomial inputs [7].

Sapinder, Ritu, Singh & Singh (2012) in their work had shown that though binary search has more line of code, program volume, program vocabulary etc. but it gave more optimized result as compared to linear search [8].

Das & Khilar (2013) proposed a randomized searching algorithm and did a performance analysis between the proposed algorithm and binary search and linear search algorithms [9].

Roy & Kundu (2014) in their work had given a detailed study on the working of linear search, binary search and interpolation search and gave their performance analysis on the basis of time complexity [10].

Chadha, Misal & Mokashi (2014) had proposed a modified binary search which improved the execution time over traditional binary search [11].

Parmar & Kumbharana (2015) had done a comparative study to search an element from a linear list (static array, dynamic array and linked list) using linear search and binary search [12].

Pathak (2015) had conducted an analysis and comparative study on linear and binary search and compared them on the basis of their time complexity [13].

III. OBJECTIVES OF THE STUDY

(i) To find out whether two personal computers having same hardware and software configurations show identical behavior when performing Binary Search in the Worst Case scenario.

(ii) To identify the best curve or curves that can be fitted to the data points (Average Searching Time in the Worst case versus Number of Data Elements) for both the personal computers.

(iii) To identify the mathematical model or models of the best fitted curve or curves for both the personal computers which may help us to explain the behavior of the Binary Search in the Worst Case on two personal computers.

IV. RESEARCH METHODOLOGY

Sample Dataset:

We have used Windows Operating System (Windows XP Professional, Version 2002, Service Pack 3) and Java (NetBeans IDE 7.0; Java: 1.6.0_17) for generating the experimental dataset which is tabulated below (TABLE 1). We have considered fifty (50) numbers of data series (Number of data elements 1000 to 50000 with an interval of 1000) on two (2) personal computers (named as PC1 and PC2) having the same hardware configurations (Intel(R) Core(TM)2 Duo CPU, 2.93 GHz; 2 GB of RAM), collected the 'Searching Time in the Worst Case' ten thousand (10000) times for each of these fifty (50) number series (i.e. from data size 1000 to 50000 with an interval of 1000) on both the computers (PC1 and PC2) and calculated the 'Average Searching Time' (AST) in the worst case for each of these fifty (50) number series (i.e. from data size 1000 to 50000 with an interval of 1000) on both the computers (PC1 and PC2) to avoid any inconsistencies/ variations.

TABLE 1
SAMPLE DATASET

Number of Data Elements (N)	**Average Searching Time of PC1 (AST1)**	**Average Searching Time of PC2 (AST2)**
1000	258.16	267.6485
2000	268.5651	269.4489
3000	276.0775	272.5172
4000	275.051	273.6451
5000	282.8466	281.7805
6000	283.969	282.3507
7000	284.5535	282.9524
8000	281.4707	298.3697
9000	296.2802	295.3182
10000	290.7681	294.4005
11000	290.5184	301.8302
12000	291.8582	299.9102
13000	296.4739	303.861
14000	295.953	305.7184
15000	293.0293	323.4054
16000	294.237	302.7224
17000	302.5414	317.9148
18000	303.4936	310.648
19000	306.7111	323.6499
20000	305.2104	331.2794
21000	306.6043	332.3994
22000	314.5313	338.6976
23000	319.06	329.0409
24000	317.4743	343.6762
25000	329.2612	342.4055
26000	328.8123	355.2623
27000	322.0267	346.113
28000	326.0779	346.9218
29000	324.1115	342.0446
30000	328.5583	349.6822
31000	329.1821	350.3376
32000	328.7759	339.251
33000	343.2853	361.4481
34000	342.2839	354.6099
35000	347.9416	390.843
36000	352.283	358.8265
37000	352.226	355.1403
38000	352.226	364.5929
39000	348.5712	373.879
40000	348.3077	374.6645
41000	350.5729	382.3849
42000	349.7342	364.7573
43000	349.7342	359.247
44000	342.4499	355.8453
45000	345.4422	370.2254
46000	347.3734	356.4804
47000	344.8421	352.0978
48000	378.6777	350.6627
49000	365.9558	365.4195
50000	356.3326	374.0888

Unit of 'Average Searching Time' (AST) is in Nano-Seconds.

Data Analysis Steps:

Step1: Graphical representation of Average Searching Time of PC1 and PC2 in the worst case.

Step2: Testing the Distribution of Average Searching Time of PC1 and PC2 in the Worst Case Using Mann-Whitney U Test.

Decision rule: If the Asymptotic significance is less than .05 then the two groups are significantly different [14].

Step3: Using Curve Estimation Technique for Best Model Selection Based on Goodness of Fit Statistics. In this case we have used the following goodness of fit statistics:

(a) R Square
(b) Adjusted R Square
(c) Root Mean Square Error (RMSE)

Decision rule: The model which has the highest R Square value (close to 1), highest Adjusted R Square value (close to 1) and lowest RMSE value (close to 0) will be selected as the best model [15][16].

Step4: F-Test of the Best Models for PC1 and PC2 (Model Diagnostics 1).

Decision rule: if the significance of F-test is less than our alpha level (.05) then we can conclude that the independent variable reliably predicts the dependent variable [17].

Step5: Testing of Normal Distribution of the Residuals of the Best Models for PC1 and PC2 (Model Diagnostics 2).

Decision rule: if we observe a symmetric bell shaped curve which is evenly distributed around zero we may conclude that the residuals are normally distributed [18][16].

Step6: Mathematical Equation(s) and Graphical Representations of the Best Model(s) Selected for PC1 and PC2

Model Used:

In this study the researchers have used nine (9) models for evaluating the dataset of both the computers which are given below:

(i) Linear, (ii) Quadratic, (iii) Cubic, (iv) Logarithmic, (v) Inverse, (vi) S, (vii) Compound, (viii) Growth and (ix) Exponential.

Software Used for Data Analysis:

We have used SPSS 20 and MS Excel for doing the data analysis.

V. DATA ANALYSIS & FINDINGS

Step1: Graphical Representation of Average Searching Time of PC1 and PC2 in the Worst Case:

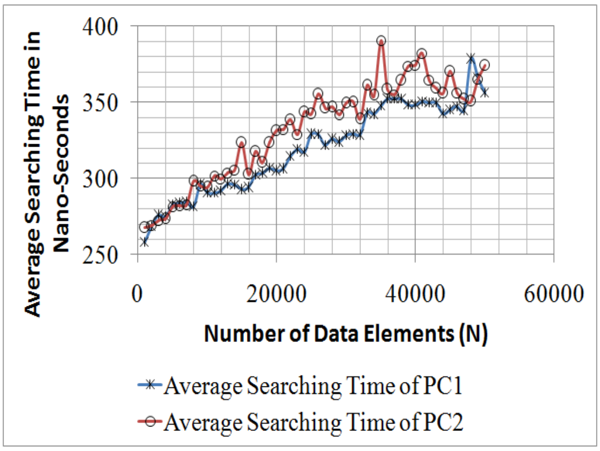


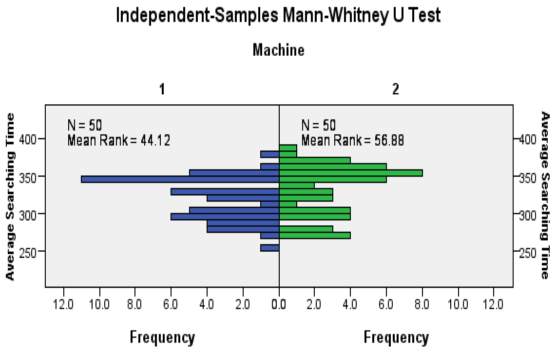
Figure 1: Average Searching Time versus Number of Data Elements Plot of PC1 and PC2

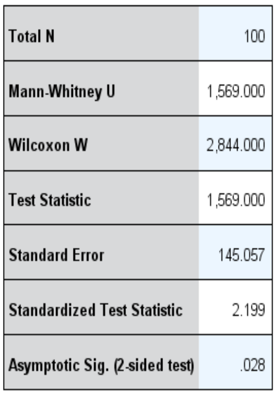
Findings: It has been observed from the above graph that the average searching time in the worst case for PC1 and PC2 are behaving differently.

Step2: Testing the Distribution of Average Searching Time of PC1 and PC2 in the Worst Case Using Mann-Whitney U Test:

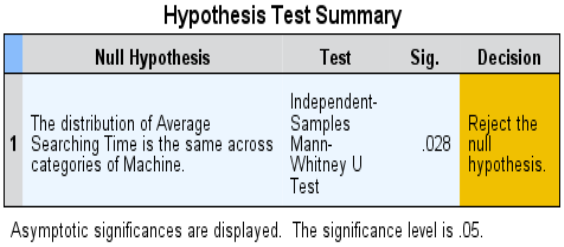
Null Hypothesis: The distribution of Average Searching Time is same across categories of machine.

The output of the Mann-Whitney U test is given below.





Total N	100
Mann-Whitney U	1,569.000
Wilcoxon W	2,844.000
Test Statistic	1,569.000
Standard Error	145.057
Standardized Test Statistic	2.199
Asymptotic Sig. (2-sided test)	.028



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Average Searching Time is the same across categories of Machine.	Independent-Samples Mann-Whitney U Test	.028	Reject the null hypothesis.

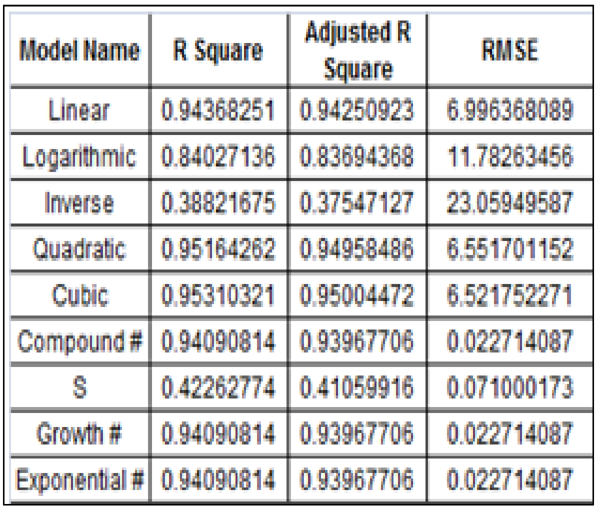
Asymptotic significances are displayed. The significance level is .05.

Findings: It has been observed from the above test that the distribution of average searching time in the worst case for PC1 and PC2 is not same.

Step3: Using Curve Estimation Technique for Best Model Selection Based on Goodness of Fit Statistics:

We have used curve estimation technique on the dataset collected from both the computers (PC1 and PC2) for selecting the best model(s) based on the goodness of fit statistics which are tabulated below (TABLE 2 & TABLE 3).

TABLE 2
GOODNESS OF FIT STATISTICS FOR PC1

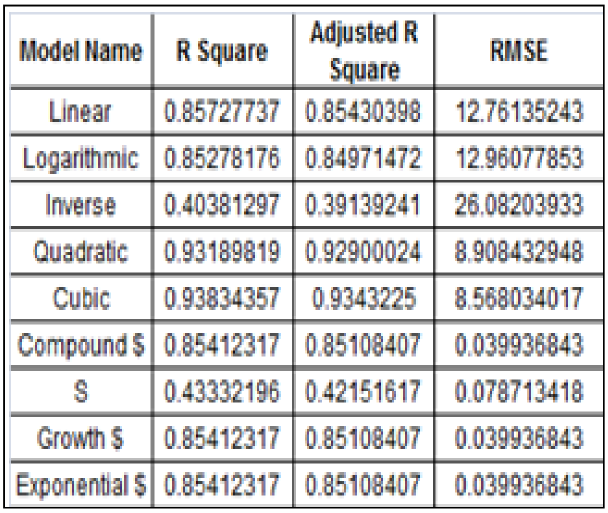


Model Name	R Square	Adjusted R Square	RMSE
Linear	0.94368251	0.94250923	6.996368089
Logarithmic	0.84027136	0.83694368	11.78263456
Inverse	0.38821675	0.37547127	23.05949587
Quadratic	0.95164262	0.94958486	6.551701152
Cubic	0.95310321	0.95004472	6.521752271
Compound #	0.94090814	0.93967706	0.022714087
S	0.42262774	0.41059916	0.071000173
Growth #	0.94090814	0.93967706	0.022714087
Exponential #	0.94090814	0.93967706	0.022714087

# Best model for PC1

Findings: It has been observed from the above table (TABLE 2) that three (3) models namely ‘Compound’, ‘Growth’ and ‘Exponential’ are having highest R Square and Adjusted R Square values and lowest RMSE values. Hence, we have identified these three (3) models as best models for PC1.

TABLE 3
GOODNESS OF FIT STATISTICS FOR PC2



Model Name	R Square	Adjusted R Square	RMSE
Linear	0.85727737	0.85430398	12.76135243
Logarithmic	0.85278176	0.84971472	12.96077853
Inverse	0.40381297	0.39139241	26.08203933
Quadratic	0.93189819	0.92900024	8.908432948
Cubic	0.93834357	0.9343225	8.568034017
Compound $	0.85412317	0.85108407	0.039936843
S	0.43332196	0.42151617	0.078713418
Growth $	0.85412317	0.85108407	0.039936843
Exponential $	0.85412317	0.85108407	0.039936843

$ Best model for PC2

Findings: It has been observed from the above table (TABLE 3) that three (3) models namely ‘Compound’, ‘Growth’ and ‘Exponential’ are having highest R Square and Adjusted R Square values and lowest RMSE values. Hence, we have identified these three (3) models as best models for PC2.

Step4: F-Test of the Best Models for PC1 and PC2 (Model Diagnostics 1):

The F – test and the significance of the F – test of the best models identified in the above two (2) tables (TABLE 3 and TABLE 4) is tabulated below (TABLE 4 and TABLE 5).

TABLE 4
F – TEST AND SIGNIFICANCE OF F – TEST OF THE BEST MODELS FOR PC1

Model Name	F Test Value	Significance
Compound	764.294607	.000
Growth	764.294607	.000
Exponential	764.294607	.000

Findings: From the above table (TABLE 4) it is evident that the p-value (significance column) for all the chosen models are less than .05, so all the models are good fit for the data.

TABLE 5
F – TEST AND SIGNIFICANCE OF F – TEST OF THE BEST MODELS FOR PC2

Model Name	F Test Value	Significance
Compound	281.044721	.000
Growth	281.044721	.000
Exponential	281.044721	.000

Findings: From the above table (TABLE 5) it is evident that the p-value (significance column) for all the chosen models are less than .05, so all the models are good fit for the data.

Step5: Testing of Normal Distribution of the Residuals of the Best Models for PC1 and PC2 (Model Diagnostics 2):

The histograms of the residuals of the best models i.e. “Compound”, “Growth” and “Exponential” models for PC1 are given below (Figure 2, Figure 3 and Figure 4).

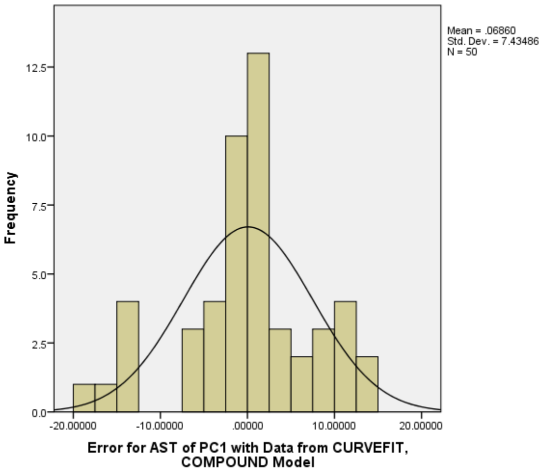


Figure 2: Histogram of the Residuals of PC1 for Compound model

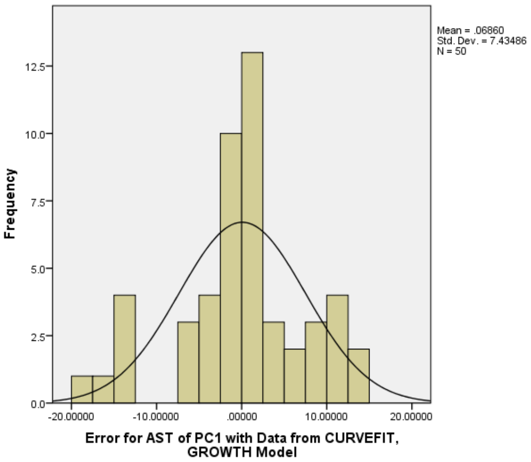


Figure 3: Histogram of the Residuals of PC1 for Growth model

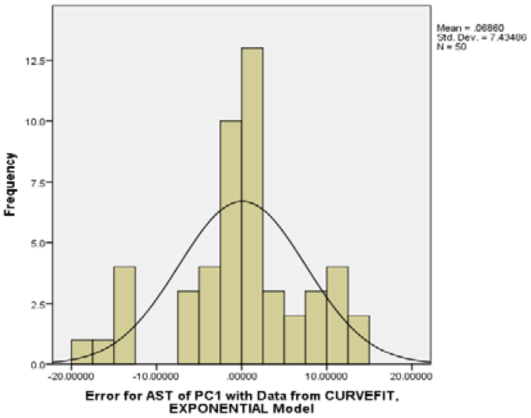


Figure 4: Histogram of the Residuals of PC1 for Exponential model

Findings: From the above histograms (Figure 2, 3 and 4) we observe, in all the cases, a symmetric bell shaped curve which is evenly distributed around zero. Therefore, we conclude that in all the cases the residuals are normally distributed.

The histograms of the residuals of the best models i.e. “Compound”, “Growth” and “Exponential” models for PC2 are given below (Figure 5, Figure 6 and Figure 7).

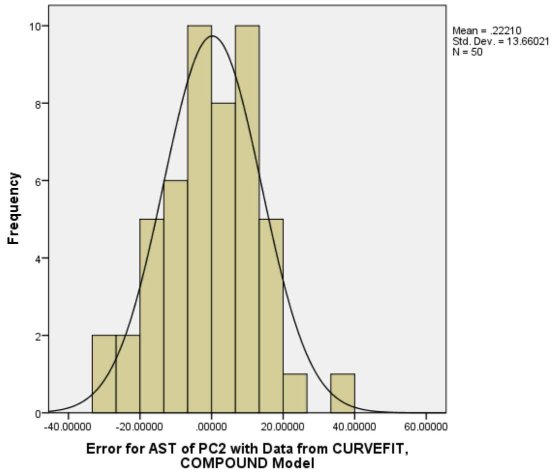


Figure 5: Histogram of the Residuals of PC2 for Compound model

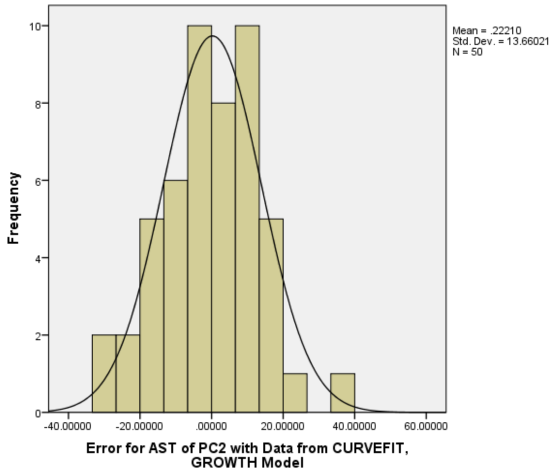


Figure 6: Histogram of the Residuals of PC2 for Growth model

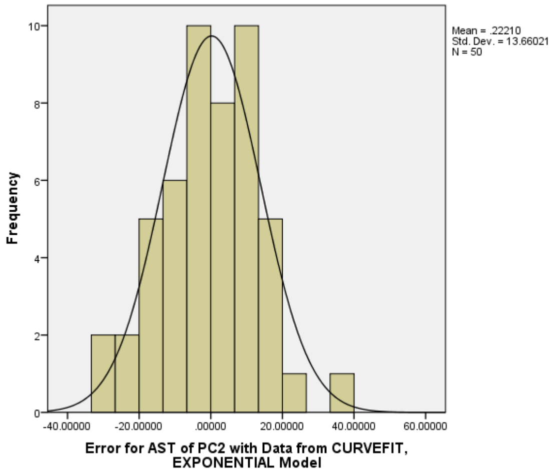


Figure 7: Histogram of the Residuals of PC2 for Exponential model

Findings: From the above histograms (Figure 5, 6 and 7) we observe, in all the cases, a symmetric bell shaped curve which is evenly distributed around zero. Therefore, we conclude that in all the cases the residuals are normally distributed.

Step6: Mathematical Equations and Graphical Representations of the Best Models Selected for PC1 and PC2:

(a) Mathematical equation of Compound curve for PC1:

AST1 = 271.889563 + (1.000006**N)

The graphical representation of the Compound curve for PC1 is shown below (Figure 8).

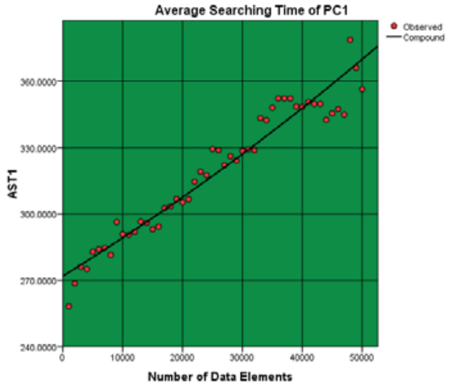


Figure 8: Compound Model for PC1

(b) Mathematical equation of Growth curve for PC1:

AST1 = e**(5.605396 + (0.000006 * N))

The graphical representation of the Growth curve for PC1 is shown below (Figure 9).

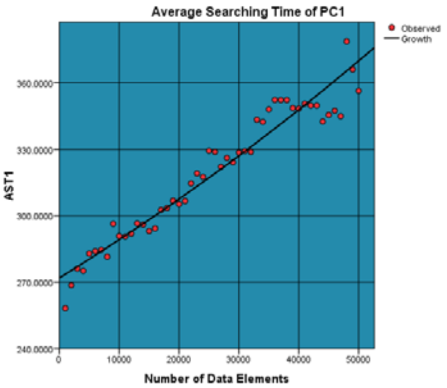


Figure 9: Growth Model for PC1

(c) Mathematical equation of Exponential curve for PC1:

AST1 = 271.889563*(e**(0.000006 * N))

The graphical representation of the Exponential curve for PC1 is shown below (Figure 10).

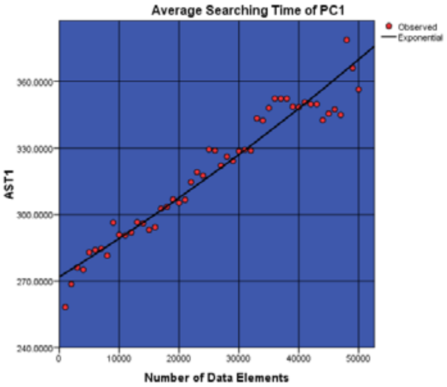


Figure 10: Exponential Model for PC1

(d) Mathematical equation of Compound curve for PC2:

AST2 = 279.752809 + (1.000007**N)

The graphical representation of the Compound curve for PC2 is shown below (Figure 11).

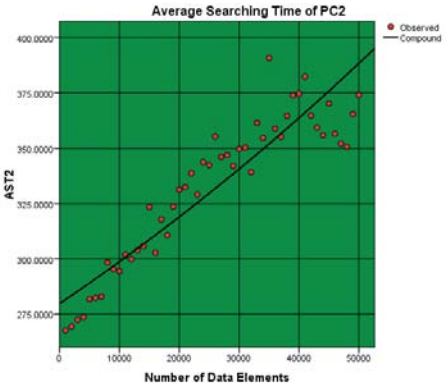


Figure 12: Compound Model for PC2

(e) Mathematical equation of Growth curve for PC2:

AST2 = e**(5.633906 + (0.000007* N))

The graphical representation of the Growth curve for PC2 is shown below (Figure 13).

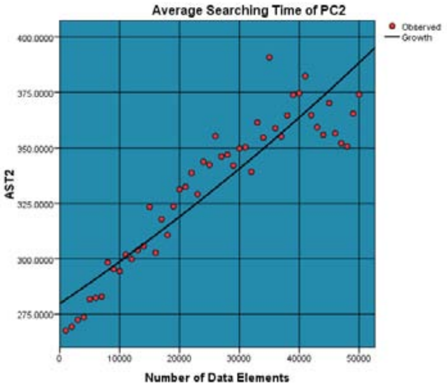


Figure 13: Growth Model for PC2

(f) Mathematical equation of Exponential curve for PC2:

AST2 = 279.752809* (e**(0.000007* N))

The graphical representation of the Exponential curve for PC2 is shown below (Figure 14).

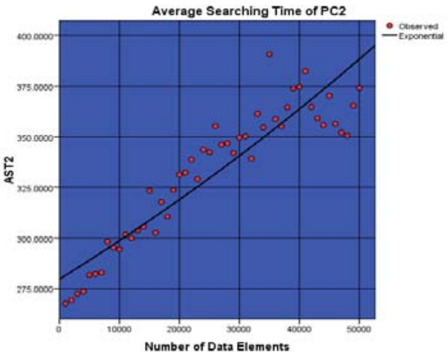


Figure 14: Exponential Model for PC2

VI. LIMITATIONS & FUTURE SCOPE

In this study the researchers have used two particular personal computers for carrying out the research work. We have used Java language for finding out the execution time. We have taken ten thousand (10000) observations for each number of data elements for both the computers (PC1 & PC2) and to avoid any inconsistencies/ variations we have calculated the 'Average Searching Time'. At the same, we have used only nine (9) families of curves to fit the data points.

Therefore, carrying out this experiment on various types of personal computers either having same hardware and/or software configurations or having different hardware and/or software configurations will definitely be our future scope. Running this study on different operating systems and programming languages are another challenge lies in front of us. In the present study we have not used any outlier identification technique (data mining technique). Hence, what would happen if the outliers are identified before starting the analysis is also an unanswered question before us. Using other types of curves other than those used in this study is also definitely our future endeavor.

VI. CONCLUSION

From the graphical representation (Figure1) we have observed that the average searching time in the worst case for PC1 and PC2 are behaving differently. It has also been observed from the Mann-Whitney U test that the distribution of average searching time in the worst case for PC1 and PC2 is not same. Therefore, these two aforesaid observations tempted us to rapidly jump to a conclusion that in this case the two personal computers under study did not show identical behavior when performing Binary Search in the Worst Case scenario and thus fulfilling the objective number 1 of this study.

In the course of identifying the best curve or curves that can be fitted to the data points (objective number 2) and proposing mathematical model or models of the best fitted curve or curves (objective number 3) strange phenomena were observed by the researchers. We found that in case of both the personal computers (PC1 & PC2) under study the data points could be best fitted to Compound, Growth and Exponential curves.

From these later findings we may conclude that though both the personal computers exhibits different behavior in terms of execution time while performing Binary Search in the Worst Case scenario but at the same time both the datasets can be best fitted to Compound, Growth and Exponential curves which may help us to explain the behavior of the Binary Search in the Worst Case.

REFERENCES

[1] Empirical algorithmics. (n.d.). Retrieved October 1, 2015, from https://en.wikipedia.org/wiki/Empirical_algorithmics

[2] Empirical Research. (n.d.). Retrieved October 1, 2015, from https://explorable.com/empirical-research

[3] Binary Search. (n.d.). Retrieved October 1, 2015, from http://algorithms.openmymind.net/search/binarysearch.html

[4] 1 Time Complexity of Binary Search in the Worst Case. (n.d.). Lecture. Retrieved October 3, 2015, from http://www.csd.uwo.ca/Courses/CS2210a/slides/binsearch.pdf

[5] Analysis of Binary Search. (n.d.). Retrieved October 1, 2015, from http://www2.hawaii.edu/~janst/demos/s97/yongsi/analysis.html

[6] Rao, R. (n.d.). CSE 373 Lecture 4: Lists. Lecture. Retrieved October 1, 2015, from https://courses.cs.washington.edu/courses/cse373/01sp/Lect4.pdf

[7] Kumari, A., Tripathi, R., Pal, M., & Chakraborty, S. (2012). Linear Search Versus Binary Search: A Statistical Comparison For Binomial Inputs. International Journal of Computer Science, Engineering and Applications (IJCSEA), 2(2). Retrieved October 10, 2015, from http://www.airccse.org/journal/ijcsea/papers/2212ijcsea03

[8] Sapinder, Ritu, Singh, A., & Singh, H.L. (2012). Analysis of Linear and Binary Search Algorithms. International Journal of Computers & Distributed Systems, 1(1).

[9] Das, P., & Khilar, P. M. (2013). A Randomized Searching Algorithm and its Performance analysis with Binary Search and Linear Search Algorithms. International Journal of Computer Science & Applications (TIJCSA), 1(11). Retrieved October 10, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.7058&rep=rep1&type=pdf

[10] Roy, D., & Kundu, A. (2014). A Comparative Analysis of Three Different Types of Searching Algorithms in Data Structure. International Journal of Advanced Research in Computer and Communication Engineering, 3(5). Retrieved October 10, 2015, from http://www.ijarcce.com/upload/2014/may/IJARCCE6C%20a%20arnab%20A%20Comparative%20Analysis%20of%20Three.pdf

[11] Chadha, A. R., Misal, R., & Mokashi, T. (2014). Modified Binary Search Algorithm. arXiv preprint arXiv:1406.1677.

[12] Parmar, V. P., & Kumbharana, C. (2015). Comparing Linear Search and Binary Search Algorithms to Search an Element from a Linear List Implemented through Static Array, Dynamic Array and Linked List. International Journal of Computer Applications, 121(3). Retrieved October 10, 2015, from http://search.proquest.com/openview/a9b016911b033e1e8dd2ecd4e7398fdd/1?pq-origsite=gscholar

[13] Pathak, A. (2015). Analysis and Comparative Study of Searching Techniques. International Journal of Engineering Sciences & Research Technology, 4(3), 235-237. Retrieved October 10, 2015, from http://www.ijesrt.com/issues pdf file/Archives-2015/March-2015/33_ANALYSIS AND COMPARATIVE STUDY OF SEARCHING TECHNIQUES.pdf

[14] Cramming Sam's Tips for Chapter 6: Non - parametric models. (n.d.). Retrieved October 10, 2015, from https://edge.sagepub.com/system/files/Chapter6.pdf

[15] Evaluating Goodness of Fit. (n.d.). Retrieved October 10, 2015, from http://in.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html

[16] Das, D., Chakraborty, A., & Mitra, A. (2014). Sample Based Curve Fitting Computation on the Performance of Quicksort in Personal Computer. International journal of scientific and engineering research, 5(2), 885-891. Retrieved October 10, 2015, from http://www.ijser.org/paper/Sample-Based-Curve-Fitting-Computation-on-the-Performance.html

[17] Annotated SPSS Output. (n.d.). Retrieved October 10, 2015, from http://www.ats.ucla.edu/stat/spss/output/reg_spss_long.htm

[18] 5.2.4. Are the model residuals well-behaved? (n.d.). Retrieved October 10, 2015, from http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm