2009-01-01

Novel Approach To Predict The Crash Testing Data Using Multiple Regression Analysis And Principal Component Analysis

Ravi Lochan Kallur
University of Texas at El Paso, rlkallur@miners.utep.edu

NOVEL APPROACH TO PREDICT THE CRASH TESTING DATA USING MULTIPLE REGRESSION ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

RAVI LOCHAN KALLUR

Department of Industrial Engineering

APPROVED :

_____
Tzu-Liang (Bill) Tseng, Ph.D., Chair

_____
Jianmei Zhang, Ph.D.

_____
Wenming Chung, Ph.D.

_____
Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

Dedicated to Lord Venkateswara

NOVEL APPROACH TO PREDICT THE CRASH TESTING DATA USING MULTIPLE REGRESSION ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

by

RAVI LOCHAN KALLUR, B.Tech.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2009

Acknowledgements

Guru Brahma Guru Vishnu

Guru Devo Maheshwara

Gurusakshat Param Brahma

Tasmai Sri Gurave Namah

This means Guru is Brahma the Creator, Guru is Vishnu the Preserver, Guru is Shiva the Transformer, Guru is verily the visible Supreme Being. I bow to that Guru.

A Guru is a person who is regarded as having great knowledge, wisdom and authority in a certain area and who uses these abilities to guide others. Dr. Bill Tseng is one such mentor to me. I sincerely thank my advisor, Dr. Bill Tseng for choosing me as his student and having belief in me to accomplish this research. His faith in my ability coupled with his positive outlook was the biggest source of strength at times of despair and supporting me throughout the course of action for completing this work.

I would like to extend my gratitude to my committee members Dr. Jianmei Zhang, Dr. Wenming Chung for spending their time in participation of thesis defense even within very short notice of time.

I would like to thank Mr. Abhilash Reddy Kothi for his valuable suggestions and guidance in my thesis preparation. I also owe sincere thanks to my roommates and my colleagues of Intelligent Systems Engineering Laboratory (ISEL) for creating such a conduce environment around me to accomplish this work. I thank my friends Mr. Yugandhar Reddy and Ms. Sirisha for extending their moral support and providing me all the necessities to achieve my goal.

I also want to thanks the Industrial Engineering faculty at The University of Texas at El Paso for guiding me throughout my master's program and giving me the right knowledge and experience to excel in my areas of interest.

I cannot end this without mentioning my family without whose support I could not have fulfilled all my goals.

Finally, I thank the *Almighty* for showering all the blessings necessitated for the successful completion of this thesis work.

Abstract

In the design process of vehicles, crash tests are very critical to determine the safety measures. Every motor manufacturing company has to maintain certain standards for their autos. This will lead to the design optimization of safety measures utilizing the data available from crash tests. The engineers perform various experiments to generate data from crash testing of vehicles in their manufacturing facility. With the help of simulators; they create a virtual environment to perform design changes. Hence the data obtained from the crash tests is vital in design optimization of safety systems. The present study deals with the technologies involved in analyzing data obtained from these experiments to ensure the prediction of data from crash tests being accurate. Current approach compares Multiple Regression Analysis and Principal Component Analysis for the prediction of data. The present work successfully derived methods for predicting data mere accurately to help the engineers reducing their efforts in conducting real time crash tests.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1.Background

Every motor company conducts hundreds of crash tests each year which helps the engineers to improve the design and to determine the vehicle restraint systems such as seat belts, air bags and perform to the safety requirement. Large volumes of data are collected by engineers in these crash tests to calculate the overall safety rating, develop a best approach to design the air bag system, seat belt system as well the car structure, understand the injury mechanisms and performance of child restraints [1]. They need to check these measurements to meet safety standards and federal requirements. The engineers measure different types of data like, chest deflection of dummy at particular time, safety acceleration on the head of the dummy, head damage according to the safety/crash requirements. These types of data are often summary measures of time functions such as maximum acceleration, deflection during a specified time period, acceleration at a specified time instance, etc. for design improvisation these summary measures are calculated from the time functions and are used as data to fit response surfaces. Hence it is required to analyze the data collected and validate accurately [2]. In data analysis, it is very difficult to analyze the whole data as the volume of the data is more. So, there should be redundancy in the data which can be eliminated or ignored while doing analysis. The validation of time functions has been difficult especially when there are more variables. Usually different techniques in Regression analysis are used to analyze the data with large number of variables to get interpretable results depending on the nature of data [1, 3]. The Regression analysis is used to fit time dependent RSMs for time functions as outputs of a non-linear distributed and dynamic system [4, 5]. To develop efficient methods with computer codes to predict data by reducing the dimensionality in applications such as crash/safety, noise/vibration yields implementation of

Principle Component Decomposition (PCD), Singular Value Decomposition (SVD) and Principle Component Analysis (PCA) in the current research [5]. Many Data mining tools are used for classification of data, PCA and SVD methods that are presently being used for dimensional reduction of data, as the advantage of PCA is that it reduces the dimensions of the data and retains the variables which account for maximum variation in the data [6].

1.2 Motivation of research

The main motivation of the current research is to develop computer codes for dimensional reduction and prediction of data consisting of time functions obtained from crash test data, validation of this data by testing the major regression techniques obtain better approach to reduce data and also to approximate it. The data is tested with Multiple regression analysis and then with Principle Component Analysis (PCA) using Singular Value Decomposition (SVD) theorem. Principle component Analysis methodology is selected over regression analysis in this research as PCA reduces the dimensions of the data which contains large variable, it also gives variables which account to maximum variability of dataset and produces output without losing the critical information contained in the whole dataset (i.e. without losing the main information in the dataset).

1.3 Research overview

The aim of this research is to bring out an improved approach so as to decrease the effort of the engineers to go for a crash test data analysis every time by replacing a common methodology with computer codes found in the research, so that most of the redundancy can be avoided .This helps in saving cost, time and risk involved in the setup. In this research the data is analyzed using Multiple regression analysis and Principal Component Analysis, suggesting that

PCA can be a better method to predict and fit time-dependent non-linear, distributed system. The present study deals with the experimental data collected and analyzed from crash testing of a vehicle at a major motor manufacturing facility. The prediction of data is considered after performing different regression analysis to prove the implementation of SVD and PCA which can be used as an efficient method, generating computer codes to fit response surface models of time functions, for applications such as crash/safety, noise/vibration.

1.4 Identification of problem

The collected crash testing data consists of many variables and time functions. The accurate prediction of this data is necessary for analyzing the design optimization of safety measures for an automotive and prediction accuracy of the resulting data fits needs improvement. The engineers can reduce their burden in running numerous real time simulations after doing required analysis over the data i.e. by reducing mean square errors of the approximated data and predicting it to the original design data. Hence the engineers can use the previous recorded data as well as the reduced and approximated data for design optimization purposes. The method shall now be programmed into Matlab compatible codes. The new method that provides significant improvement in reducing noise in prediction and fitting data compared to the original data while decreasing the computational load for the engineers in developing computer codes. The research needs to apply the improved method to a real-life crash/safety problem, and make refinement on the method as necessary. The merit on accuracy and efficiency needs to be demonstrated with the problem.

1.5 Research organization

This thesis is organized into four chapters. Chapter 1 gives the introduction problem identification and thesis overview. It discusses about an improved analysis which is required. The following chapter (Chapter 2: A Formal Literature) gives a brief description about the different techniques which have been used, i.e. Multiple Regression Analysis using Least Squares Method, Singular Value Decomposition Theorem and its extension as Principal Component Analysis. Chapter 3 explains about the methodology used explaining the experimental setup as the real time crash testing, how the data is collected and for what purpose it is used and the computer code generation according to the requirements. Results and interpretations from the analysis are presented in Chapter 4. And conclusions are explained in Chapter 5.

Chapter 2: Literature Review

This chapter presents literature review about the data analysis methods used in the present research. The review of these articles related to the application about the two methods analyzed in the present work which gives a better idea on the approaches. Usually the data extracted from crash tests of vehicles consists of large amount of time dependent variables. To analyze this kind of huge data, engineers use various data mining techniques.

Data mining has been used to refer the process of finding interesting information in large sets of data. Precisely, the term data mining refers to the application of finding patterns in a process which are interrelated numerous disciplines including statistics, artificial intelligence, machine learning, database science, and information retrieval (Han & Kamber, 2001). Data mining techniques can be performed on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data [7]. Data mining is being used for getting a clear insight of data and it is also used as data reduction technique as in case of dimensional reduction. The dimensionality reduction comes down to basically reducing the number of variables to few variables, newly definable variables or categorizations of huge dimensional spaces into smaller partitioned spaces, with appropriate discounting of unusual dimensions, variables [8].

In the present research the data obtained from crash testing of a vehicle from a leading car manufacturing company is to be analyzed and dimensional reduction is required to predict pattern from the data. Since the data is huge, we may implement techniques involved in dimensional reduction. Based on the purpose of extracting decision rules out of the data, interpretation and prediction of data involved, complexity of data type certain situations require

specific methods. For example, rough set techniques are applied in machine learning processes, medical systems, fuzzy logics are widely used in surface machining examining techniques, and neural networks are applied in complex engineering structural analysis [9].

Pedro Furtado and H. Maderia [10] noticed the requirement of incorporating appropriate data reduction techniques in the process of data approximation and derived guidelines for data reduction tools on the basis of skew and sparseness which are relevant to the approximation accuracy.

Lilien & Rangaswamy, implemented multiple regression analysis to predict data matrices of sales forecast model building, which are screened for potential multicollinearity among the predictor variables using correlation analysis [11].

George Kontaxakis [12] investigated use of Principal Component Analysis and Similarity Mapping techniques as the data reduction methods involved in image data, Antti Niemisto, Terry Speed [13] stated usage of SVD techniques in the implementation of Principal Component Analysis in discovering microarray patterns in gene expressions. These proposed methodologies yielded significant results.

From all data reduction techniques of data mining principal component analysys and multiple regression analysis are two powerful techniques being used for data interpretation and data reduction being used in recent researches.

The following table summarizes the works done by different people in the implementation of adequate techniques of data reduction according to the requirement [11].

Table 2.1: Summary of Characteristics of different Data reduction methods.

Technique	Variable/Space	Robustness	Importance of variable	Computational method available
Factor analysis	Variable Reduction	Less (Linearity in data)	Known	No
Cluster Analysis	Space Partitioning	More	No	Yes
Regression	Variable Regression	More (Non Linearity explained)	Known	Yes
Neural networks	Variable and Space reduction	More	No	Yes
Genetic algorithm	Variable and space reduction	Local optima are likely obtained	Known but of little use	Yes
Principle Components	Variable Reduction	More	Known	Yes

Based on this and the properties of available data, Regression and Principle Component methods are suited more than other techniques used for data reduction. To extract the relation between the variables and finding patterns, in the time dependent multi variable data obtained from crash tests, it is implemented to use Multiple Regression analysis and Principal Component Analysis.

2.1 Multiple regression analysis

Regression analysis is a statistical tool which is used for investigating data to determine the existence of relationship between the variables present in the data [14]. The presence of more than one predictor variable makes it Multiple Regression. Usually the data investigators seek the effect of every variable over the whole data so as to co relates the statistical significance of estimated relationships between the parameters and variables. The original relationship is close to their approximation or estimation. The effort of quantifying the relation among the factors that affect the variations in data is the basic concept of regression. Simple study of the data can be done by a scatter diagram [15]. For any data which consists of parameters and variables, there must be a relationship between them, which is explained by standard regression equation which is a function of variables X and β.

$$Y = f(X, \beta)$$

Where the unknown parameters are β, independent variables are X and dependent variables are Y which are bounded with a relation called regression function f. The relation could be linear or non-linear depends on which the equations are built. An error 'e' may be present while constructing the equations and therefore added to the equation. The significant difference between these error values 'Δe' for different variables is called as residual. The intension of the regression analysis is to minimize the error which is estimated by the sum of squared residuals. Regression analysis chooses amongst all possible functions by selecting the function which has the sum of the squares of the estimated error at a minimum [16].

Multiple Regression allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is important for estimating the impact of various simultaneous

parameters upon a single dependent variable. Multiple Regression is often essential even when the data analysis is only focused in the effects of one of the independent variables.

The regression creates a sequence of dataset which contains the same data points with the same values and properties, differing with those variables and data sets which have a different set of error values [16]. The minimum sum of squares of residuals characterizes consistency of the assumptions and their consequences. The calculation of root mean square error of the regression helps the parametric estimation of the data. The final fit which is obtained from the regression analysis is compared to the original and checked for the mean square error (MSE) to predict the consistency of the analysis [17].

Regression results indicate the direction, size, and statistical significance of the relationship between a predictor and response by knowing the sign of each coefficient. Regression generally uses the ordinary least squares method which derives the equation by minimizing the sum of the squared residuals [18].

2.1.1 Applications of multiple regression analysis:

Multiple Regression Analysis techniques are widely used in data analysis applications. The usage of regression analysis makes the investigator play with the data easily to determine the relationship between the variables present, the application of regression analysis expanded to various fields of study where data reduction is involved. The following is a brief summery table to show the fields of applications of Multiple Regression Analysis.

Table: 2.2: Applications of Multiple Regression Analysis

Field of Application	Researchers
Banking Sector	Creamer, Noe and Spindt. [19]
fMRI/MRA/MRI	Rowe and Hoffmann, R.G [20]
Signal Processing Engineering	Dolecek and South, H.M. [21]
Environmental Quality Control	Alden and Sokolowski. [22]
Transportation	Yamada et al. [23]
Human Pathology	Castellanos, et al. [24]
Wireless Solutions	Youssef Abdullah, Ashok Agrawal [25]

2.2 Principal component analysis:

In statistics, principal components analysis (PCA) is a technique that can be used to simplify a dataset. Now it is being used as tool for classification and reduction of the data in exploratory analysis when it is comprises of large data sets. PCA can be calculated using a correlation matrix, covariance matrix, Singular value decomposition and Eigen value decomposition for the analysis of a data [26]. The goal of PCA is to reduce the variables, if 'n' is the number of data variables then by using the principle component analysis we get 'x' variables which is less than the 'n' without losing the original variability of the data. High dimensional data set is reduced to two dimensional graphs where the maximum variability of the data is caused by only few variables which are represented in two dimensional principle components. First principal component gives the linear combination with maximum variance. Second

component is linear combination with maximal variance in orthogonal to first principal component analysis and so on.

Singular value decomposition, Principal component analyses are two techniques which are used commonly in the data mining methods [28]. Principal component analysis gives the smaller set of variables with less redundancy, which gives the good representation of the whole data set when considering the less number of the data variables.

The PCA concepts are used to visualize a lower dimensional space and hence recognizing its patterns which was introduced by Pearson and Hotelling and SVD by Eckart and Young [29]. Numerous problems in data approximation, computational algebra and multivariate statistical analysis have been solved with classical and modified PCA [28,29] and SVD. Hence the concepts of PCA is implemented here to analyze the crash test data since it has non-linearity and vast number of variables. So the implementation of these techniques yielded better results in approximating the response surface model rather than using the conventional regression analysis. These techniques include transformative decomposition of multidimensional matrices and clustering with in decomposition matrix factorization.

2.2.1 Applications of PCA:

Principal Component Analysis is a vast evolving technique which is used in many fields for data analysis and data reduction as well approximation. PCA is used for analysis of manufacturing data, PCA can be used for Quality Control, PCA is been used as a tool environmental impact assessment and transportation problems, in process engineering field and image recognition, image compression. Following is a summary table that shows the applications of PCA.

Table: 2.3 Application of PCA

Field of Application	**Researchers**
Manufacturing field	Wafik Hachichaa et al. [29]
Online Banking	Chien BruceHoa, Desheng DashWub. [30]
Process Engineering	Ricardo Dunia et al. [31]
Bio Informatics	Xi Chen and Lilly Wang. [32]
Environmental and Transportation	Nagendra Mukhesh Khare [33]
Image compression	Catalina Lucia Cocianu [34]
Fault detection and Isolation Techniques	Wei Rong et al. [35]

Chapter 3: Methodology

In this chapter the structure of experiment (crash testing of a vehicle) is explained and the data is obtained from the experiment for analysis. The approach of two techniques used for data analysis and prediction is explained. Both Multiple Regression Analysis and Principal component Analysis are used for finding out the interrelation between the variables present in the data. The analysis is also focused interpretation of data analysis using these techniques. The approach of both methodologies are focused on prediction of data and compared with each other to evaluate better approach towards the problem. The conceptual frame work of methodology is briefed as follows:

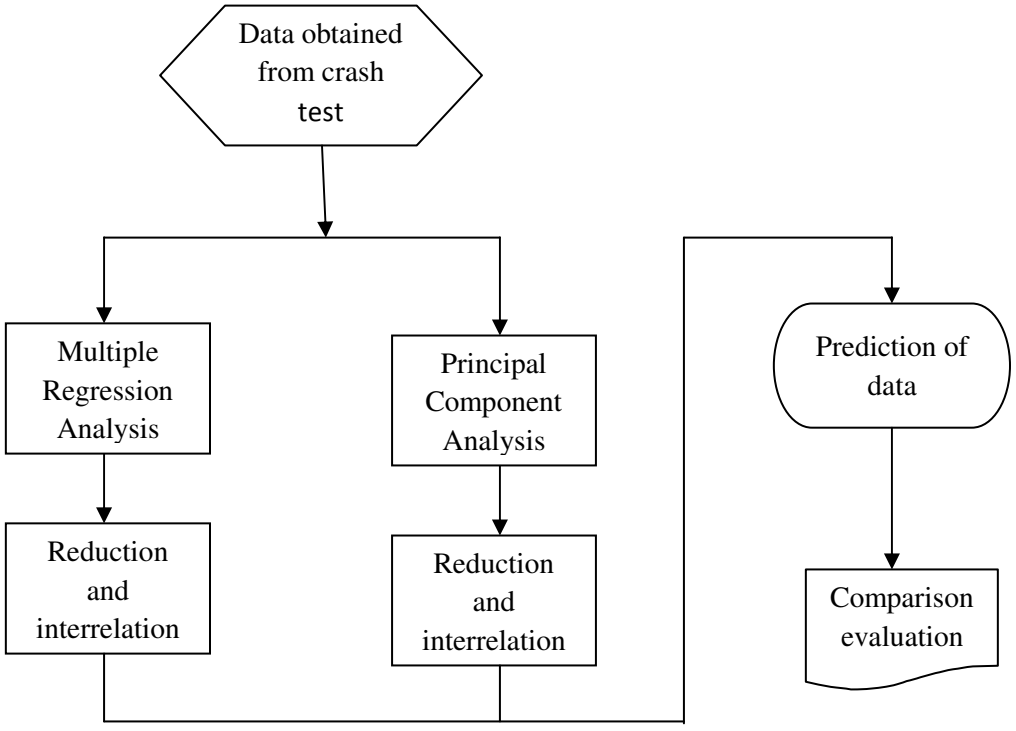


Fig 3.1: Conceptual frame work.

The experimental setup is located at one of the major car manufacturing facility. The data obtained from the crash tests done in their facility is collected and used for the present research. The crash tests are very essential in designing process of vehicles before the final design of vehicle enters the actual assembly plant.

3.1 Crash tests on vehicles:

In the designing process the engineers perform various experiments on the vehicles. The engineers can simulate collisions and study the effects caused by these collisions. These virtual simulations typically occur in early design stages. After the designing process the vehicle is out of the assembly line another set of engineers would be waiting for these toys to play. They conduct hundreds of crash tests every year which help them determine the safety system of the vehicles in their facilities using real time crash testing simulators. The simulator can recreate the deformation and tipping or pitch that a car undergoes in a real-life collision without destroying the entire car body. Engineers conduct repeated tests from the front, rear, and side of the vehicles. The engineers take immense effort to collect various types of data from these crash tests and use them in their design process. The collected data is used to develop the design to attain the safety requirements of federal laws. The engineers use various types of transducers, video, audio equipments, photographs, accelerometers, slow motion video capturing devices (digital cameras) and computer software to collect data from crash test [36].

3.1.1 Crash test preparation:

It may take great amount of time for the engineers to prepare a single dummy for a crash test by attaching hundreds of instrumentation wires and calibrating the dummy. They use transducers and chips with which they collect data as position vectors at a particular time and

speeds of the dummy at the time of impact. They paint the dummies with different colors so that after the impact they can know the actual collision position and collision dimensions in the vehicle. After numerous preparations they are all set for a real time crash test.

Fig 3.2: Preparation of dummies for crash test impact [37].

3.1.2 Crash testing setup:

The engineers build distinct facilities for crash testing with good amount of safety features. The engineer specifies the objective of the test required. The test specification covers right from speed, acceleration, point of impact, the positioning of camera and accelerometers all down to minute details so that each test is repeatable.

Fig 3.3: Set up and preparation for crash test [37].

Crash test facility can be set up to run number of tests, such as frontal and near-frontal impact, side impacts, side and rear impacts. Then, the engineers do the actual crash testing in their facility using these dummies and placing them in the vehicles.

3.1.3 Control room:

As the crash test begins, the vehicle is accelerated to impact speed on a runway track, in a full frontal crash test; the vehicle hits a fixed barrier at speeds of up to 35 mph. The transducer gives the position variables of the dummy. The digital cameras record the crash at about 1000 frames per second. During the test the vehicle acceleration and speed have been calculated accurately. The speed is controlled to ensure that the collision process should not be repeated multiple times. With the help of laser sensors cars can be collided into each other with a great accuracy. The engineers collect data such as the chest deflection at a particular time, the safety speed of the dummy, the dummy positions at various time intervals, head damage, other damages to the dummy, air bag deployment etc. the air bag deployment is a very instantaneous so that the observations made at this point of impact are critical.

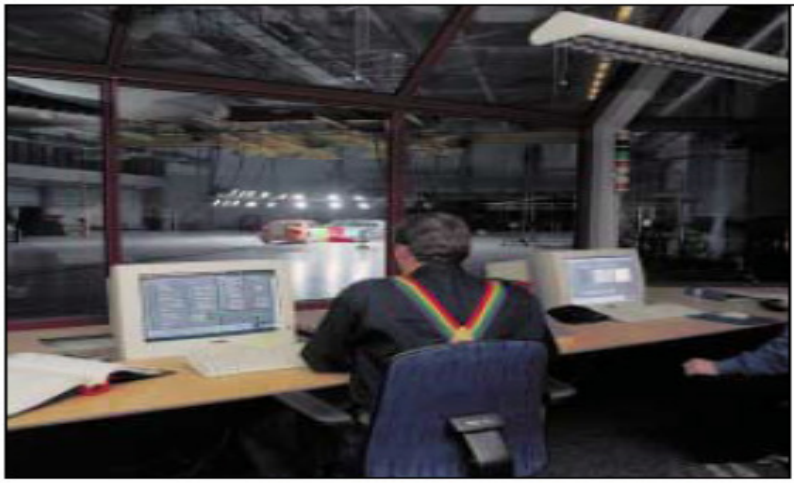


Fig 3.4: Recording of observations at control room [37].

3.1.4 At the time of crash:

At the moment of impact the digital cameras record about 1000 frames per second and various electronic instruments will calculate the time and positions of the vehicle during the collision. Accelerometers accurately calculate the accelerations and speeds of the vehicles. All these observations are made by the engineers sitting in the observation room. Observations made at each crash test give the engineers required datasets at the pre impact stage, pre deployment, at current and post impact stages. The data is collected in every millisecond of the crash test resulting huge sets of data available for the engineers to research on.

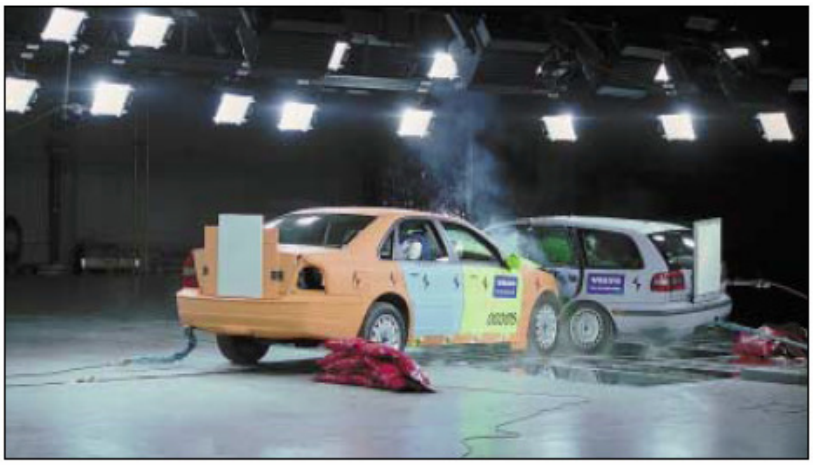


Fig 3.5: Crash test -Impact of vehicles [37].

3.1.5 After the crash:

The collected data is used to simulate the whole process by changing variables like speeds, accelerations and different dummy sizes, shapes and weights. To obtain the accuracy of these simulations the data should be approximated in order to fit the RSMs appropriately to perform design optimization. The extracted data is utilized in further studies and simulations so

as to improve the design of the vehicle and safety system. The extracted data is also helpful for crash reconstruction and developing road side safety feature design system.

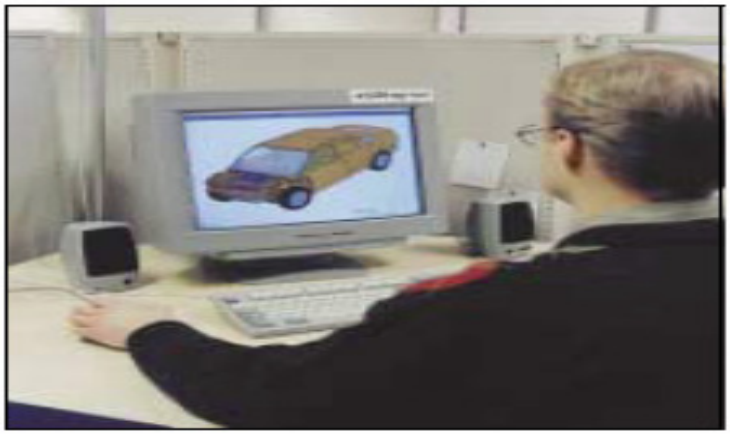


Fig 3.6: Engineers doing simulations using gathered data after the crash test [37].

3.1.6 Potential use of crash test data:

Designers can check the accuracy of computer modeling and prototype crash tests by running simulations. Mechanisms of structural collapse can be determined from post-crash inspection of the vehicle and by viewing the high-speed video of the crash test. This makes the engineers to improve the safety system of the vehicle. By studying the dummies after the crash test, the engineers understand the accident injury mechanism so that they can develop safety standards by adding extra equipment. The data further can be used for reconstruction of future crash tests and can also be used to simulate the crash tests by a computer program so that the test engineers can save time and money. Hence the data analyzed from the crash tests is important for designing a vehicle [1].Therefore the present research motivates to reduce and approximate the collected data by using the Multiple analysis and develop a computer code for the process [36].

The engineers optimize the design by performing simulation over the obtained data by constructing a response surface model and test the design data by changing variables. To fit this response surface model first the data needs to be reduced and approximated. Hence describing the objective of the research i.e. to reduce the data and predict the data approximately.

3.2 Approach towards the problem:

The data collected in crash test of vehicles is mostly time dependent. From this data, the engineers are required to fit the response model to simulate the tests. The objective of their responsiveness model is to optimize the design from the approximated data. Hence the engineers need the data applicable for simulations by giving design variables as inputs and extract the data values. If the input changes the time functions changes. Hence the motive of the research to help the design engineers provide data which has been approximated to the requirements of design variables to better fit the response model. This will save the engineer's time for the experimental set up by utilizing virtual simulations to cut the actual cost involved in the experiments. Multiple regression and Principal Component analyses are performed to the data to get the required results.

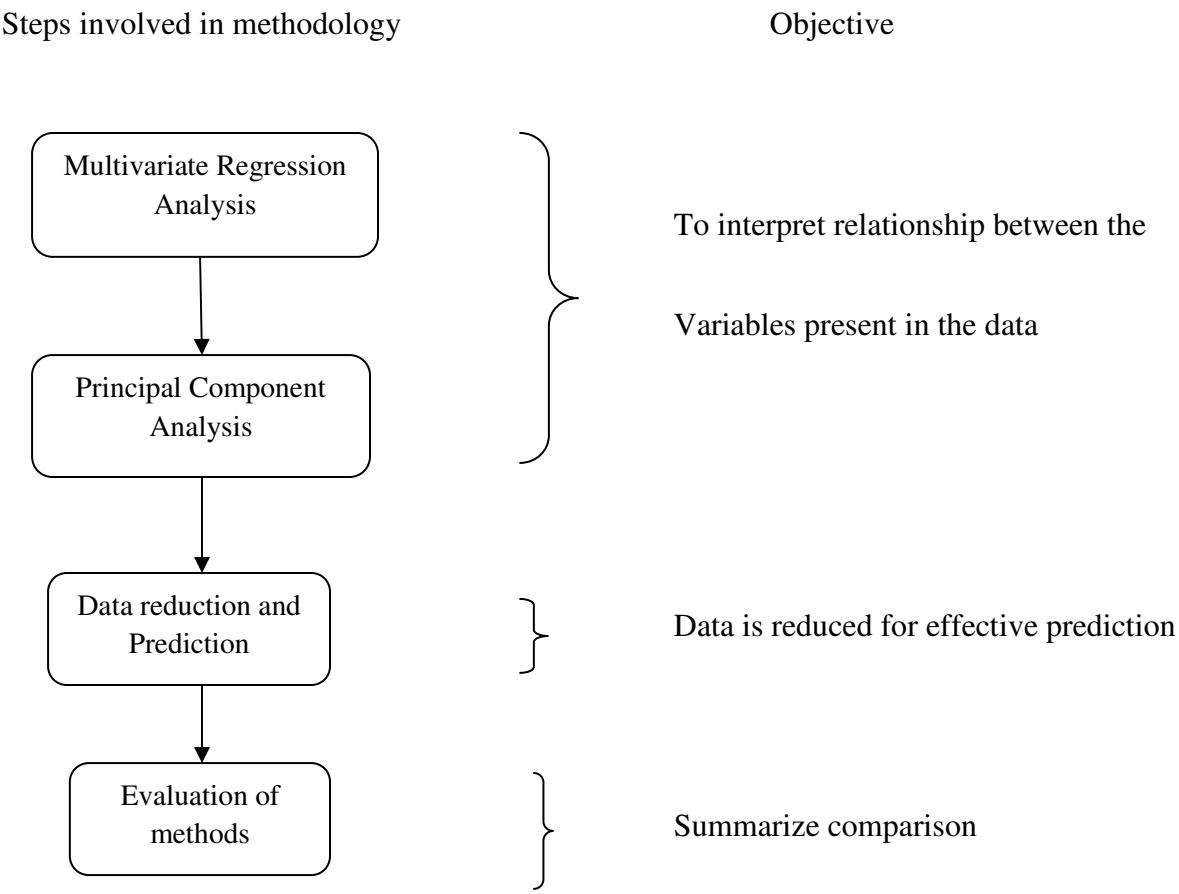


Fig 3.7: Objective of methodology sequence.

We are provided with data matrix whose dimension is 1202 * 90 in which the data corresponds to time values noted in 1202 observations and the columns corresponds to the number of simulations. The data which we have consists of values of time in micro seconds of positions of the dummy used in the crash test at certain given points of position. All the values are time functions.

3.3 Multiple regression analysis on crash test data:

The data collected from crash test of vehicle is first analyzed using Multiple regression analysis. Multiple Regression Methods are frequently used to analyze data which has large number of variables. It is one of the most popular data mining and data analysis techniques recently. In general, suppose that there is a single dependent variable 'y' or predictor variable which depends on 'n' independent variables x_1 , x_2 , x_3 , ……x_n . The relationship between these variables is characterized by a mathematical model called as regression model. The regression model is fit to a set of sample data. The relationship between independent variables and dependent variables can be sometimes exactly derivable. This is given by a functional relationship between them.

$$y = \Phi (x_1 , x_2 , x_3 , \ldots\ldots x_n)$$

However, in most cases the exact relationship is unknown. Hence an appropriate function has to be chosen to approximate Φ. These functions are generally lower order polynomial functions and linear functions. A model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots.. + \beta_n x_n + e$$

where 'y' is dependent variable and '$x_{i=1}^{n}$' are independent variables which are in relation with 'y' as functions of unknown variables $\beta_{i=1}^{n}$ and 'e' is the error or residual. The above equation is called a multiple linear regression model with 'n' regressed variables. The aim of the regression is to minimize the error values using Least Squares method to fit the data.

In our current research the data obtained from crash testing of a vehicle contains time functions and number of simulations. This data is analyzed using Multiple regression method

with the help of two different software packages Minitab and Matlab. By doing regression analysis in Minitab the interpretation of the relationship between the variables is done. After the interpretation Matlab codes are generated to check the fit of data and graphical representation to check for errors. This is explained by least squares method.

3.3.1 Least squares method

In Regression Analysis, we create a sequence for given data set. The procedure to determine best fit line or curve to the data is derived from Least Squares Method. The method of least squares assumes that the best-fit curve of given data is the curve which has a minimal sum of the deviations squared (least square error) from a given set of data [37].

Suppose that the data points are (x_1, y_1), $(x_2, y_2)\ldots (x_n, y_n)$ where 'x' is independent variable and 'y' is dependent variable. The fitting curve f(x) has the error 'e' from each data point, Then, $e_1 = [y_1 - f(x_1)]$, $e_2 = [y_2 - f(x_2)]\ldots e_n = [y_n - f (x_n)]$,

According to the method of least squares, we have the below equation.

$$\prod = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_n - f (x_n)]^2 = \text{minimum.}$$

This means the least squares method requires that a curve is to be fitted to a set of data points such that the sum of the squares of the errors or deviations from the points to the curve or a line is minimized. If the regression is on Y, or the line be fitted to a set of data points such that the sum of the squares of the horizontal deviations from the points to the line is minimized, if the regression is on X [36].

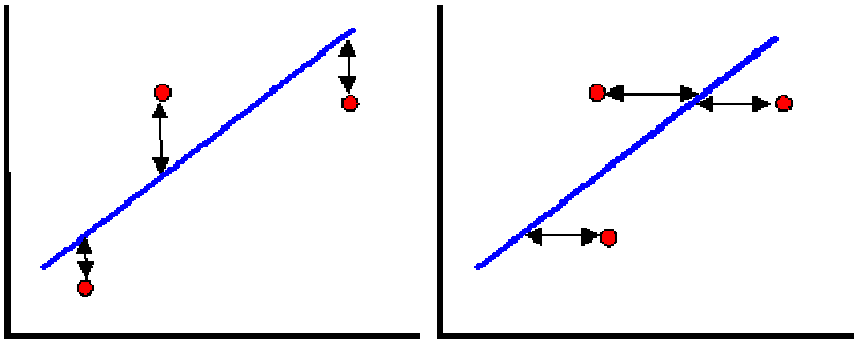


Fig 3.8: Regression using least squares on Y and on X [38].

After obtaining multiple relations between the variables, the equations are tested using Least Squares Method to check the regression of the error value. The data points which are scattered away from this line are with high error values and those which are nearer are data points with less error. This states the relation between the variables. Those variables which are having less error values have more effect on variation of the data than those of larger error values. A finalized equation is constructed to show the mathematical relation between the independent and dependent variables. This equation is considered to be the regression equation.

3.4 Principal component analysis on crash test data:

PCA is a mathematical procedure that transforms the data to new coordinate system such that the greatest variance by any projection of the data lies on the first coordinate called the first principal component, the second greatest variance on the second coordinate, and so on. The advantage of using this method is to reduce the dimensionality by retaining most of the original variability of the given data. Performing Principal Component Analysis is equivalent of performing Singular Value Decomposition on the data. First we need to decompose our data using SVD theorem.

3.4.1 Singular value decomposition theorem:

Singular value decomposition (SVD) is the technique behind the Principal Component Analysis (PCA) in Multivariate regressions. The SVD theorem explains the mathematical description of PCA. These concepts are widely used in statistics for multivariate data reduction [11]. The SVD theorem is a major tool for data analysis, data processing and compression, dimensionality reduction. The major purpose of this concept is not only data reduction but also in extracting main features in the data by finding dependencies among the variables present in the data set.

3.4.1a Singular value:

For a matrix with m rows and n columns $A_{m\times n}$ which has a rank r, the eigenvalues of A^T A are

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ \ldots\ldots \lambda_r > \lambda_{r+1} = \lambda_n = 0.$$

Then $\sigma_i = \sqrt{\lambda_i}$ is called the singular value of A, where i = 1, 2, 3 … n.

3.4.1b SVD theorem:

Let A be an m × n matrix with m ≥ n, whose rank is r, there exists two orthogonal matrices

$$U_{m\times n} = (u_1, u_2 \ldots\ u_n) \text{ and } V_{n\times n} = (v_1, v_2, \ldots.\ v_n)$$

Then one form of the singular-value decomposition of A is

$$A = U \sum V^T = \sum_{i=1}^{r} u_i\ \sigma_i\ v_i^T$$

where $\sum$ = diag ($\sigma_1, \sigma_2, \sigma_3 \ldots \sigma_n$) and σ_i is the singular value of A.

3.4.2 Principal component decomposition:

Principal component decomposition (PCD) can be computed using singular value decomposition. Singular value decomposition of a data X can be represented as,

$$X = U \times S \times V^T$$

where, U is an m x n matrix and the columns of U are called the left singular vectors.

S is an n x n diagonal matrix and the elements of S are only nonzero on the diagonal, and are called the singular values.

V^T is also an n x n matrix and the rows of V^T contain the elements of the right singular vectors.

The matrix U × S contains the principal component scores.

Finally dimension of the data can be decreased by taking large values in the S matrix and U matrix. Then U, S, V^T are partitioned depending upon the relative importance of the singular values contained in the diagonal entries of matrices S and U, V^T.

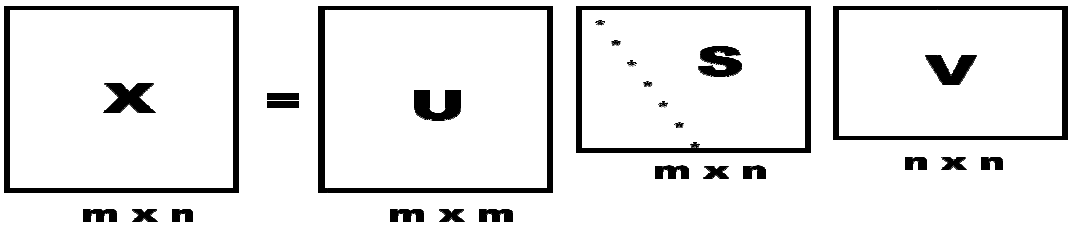


Fig: 3.9: Syntax of Singular Value Decomposition.

where, U is an m x n matrix and the columns of U are called the left singular vectors. S is an n x n diagonal matrix and the elements of S are only nonzero on the diagonal, and are called the

singular values. V is also an n x n matrix and the rows of V contain the elements of the right singular vectors. The matrix U×S contains the principal component scores. Finally, dimension of the data can be decreased by taking the large values in the S matrix and U matrix. The given response matrix is reduced using this theorem where matrix S is the decomposed matrix. Then by selecting first k singular values, the singular matrix is further decomposed.

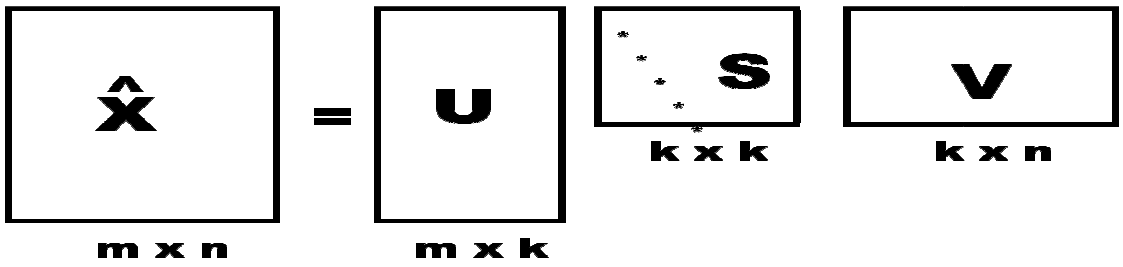


Fig: 3.10: Selecting first k singular vectors.

Then U, S, V are partitioned depending upon the relative importance of the singular values contained in the diagonal entries of matrices S and U, V.

The data obtained after SVD is

S matrix: 80*80

U matrix: 1202*80

V matrix: 1202*1202

Then Φ is calculated by using S and U matrix. $\Phi = U \times S$

After reducing the dimension of the dataset, next step is to approximate the data. Normally, it is not possible to approximate the data when both design and response matrices are likely to be high dimensional. In order to perform this task, we are using software package called

DACE (Design and Analysis of Computer Experiments), which is a Matlab toolbox. The typical use of this software package is to construct approximation model based on data from computer experiment and to use this approximation model as a substitute for the computer model.

The decision must be made to retain the principal component analysis for effectively summarize the data. To retain sufficient components to account for specified percentage of maximum variance, the components whose eigenvalues are greater than average of eigenvalues are considered in correlation matrix. The significance of large principle components are the components of corresponding larger eigenvalues.

3.5 Comparison and evaluation of multiple regression and principal component analysis

Table 3.1: Comparison of Multiple Regression and PCA

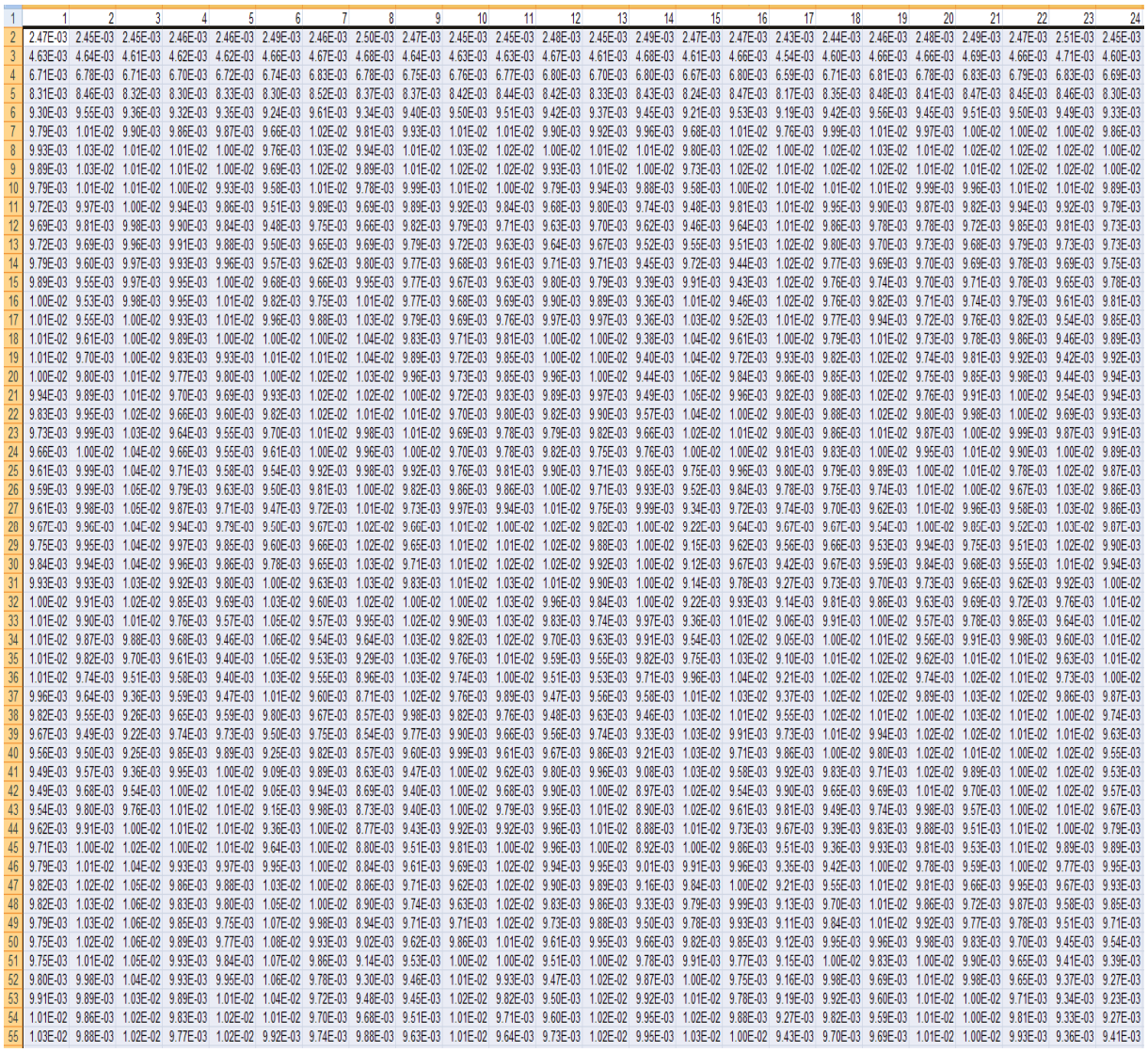
	Multiple Regression Analysis	**Principal Component Analysis**
Method	Polynomial Modeling	Bilinear Modeling
Theorem	Least Squares Method	Singular Value Decomposition Theorem
Data Reduction	Less	More
Prediction of data	Possible	Possible
Purpose	Finding interrelationship between variables	Finding interrelationship between variables

Chapter 4: Results And Comparisons

This chapter presents the analysis of obtained results from applying the methodology described in previous chapter. Multiple Regression Analysis and Principal Component Analysis are performed on the Crash testing data and compared for validation. The analysis is divided into two stages. In stage one data is analyzed by both techniques using Minitab and the interpretations are done based on the results obtained by this statistical methodology. By using these interpretations in stage two, required computer codes (programs) to predict the data are generated using Matlab. The results obtained from this methodology are the graphical interpretation of the Multiple Regression and Principal Component Analysis. Results obtained from these two stages are compared and conclusions are made at the end.

4.1 Analysis and results from multiple regression:

The data analysis using Multiple regression analysis is to be done to determine the interrelationships between the variables present in the data time dependent. The data we are provided from the crash tests is in two sets. One is response matrix which consists of 1202 observation values using 90 simulations. The whole data is entered in Minitab work sheet for statistical interpretation using Multiple Regression analysis. The following is a screen shot of data matrix in which the columns are total number of runs performed in the crash test and rows are the observations taken in each run or simulation. The observations are in micro seconds of each position of a dummy taken at 1202 different positions.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	2.47E-03	2.45E-03	2.45E-03	2.46E-03	2.46E-03	2.49E-03	2.46E-03	2.50E-03	2.47E-03	2.45E-03	2.45E-03	2.48E-03	2.45E-03	2.49E-03	2.47E-03	2.47E-03	2.43E-03	2.44E-03	2.46E-03	2.48E-03	2.49E-03	2.47E-03	2.51E-03	2.45E-03
3	4.63E-03	4.64E-03	4.61E-03	4.62E-03	4.62E-03	4.66E-03	4.67E-03	4.68E-03	4.64E-03	4.63E-03	4.63E-03	4.67E-03	4.61E-03	4.68E-03	4.61E-03	4.66E-03	4.54E-03	4.60E-03	4.66E-03	4.66E-03	4.69E-03	4.66E-03	4.71E-03	4.60E-03
4	6.71E-03	6.78E-03	6.71E-03	6.70E-03	6.72E-03	6.74E-03	6.83E-03	6.78E-03	6.75E-03	6.76E-03	6.77E-03	6.80E-03	6.70E-03	6.80E-03	6.67E-03	6.80E-03	6.59E-03	6.71E-03	6.81E-03	6.78E-03	6.83E-03	6.79E-03	6.83E-03	6.69E-03
5	8.31E-03	8.46E-03	8.32E-03	8.30E-03	8.33E-03	8.30E-03	8.52E-03	8.37E-03	8.37E-03	8.42E-03	8.44E-03	8.42E-03	8.33E-03	8.43E-03	8.24E-03	8.47E-03	8.17E-03	8.35E-03	8.48E-03	8.41E-03	8.47E-03	8.45E-03	8.46E-03	8.30E-03
6	9.30E-03	9.55E-03	9.36E-03	9.32E-03	9.35E-03	9.24E-03	9.61E-03	9.34E-03	9.40E-03	9.50E-03	9.51E-03	9.42E-03	9.37E-03	9.45E-03	9.21E-03	9.53E-03	9.19E-03	9.42E-03	9.56E-03	9.45E-03	9.51E-03	9.50E-03	9.49E-03	9.33E-03
7	9.79E-03	1.01E-02	9.90E-03	9.86E-03	9.87E-03	9.66E-03	1.02E-02	9.81E-03	9.93E-03	1.01E-02	1.01E-02	9.90E-03	9.92E-03	9.96E-03	9.68E-03	1.01E-02	9.76E-03	9.99E-03	1.01E-02	9.97E-03	1.00E-02	1.00E-02	1.00E-02	9.86E-03
8	9.93E-03	1.03E-02	1.01E-02	1.01E-02	1.00E-02	9.76E-03	1.03E-02	9.94E-03	1.01E-02	1.03E-02	1.02E-02	1.00E-02	1.01E-02	1.01E-02	9.80E-03	1.02E-02	1.00E-02	1.02E-02	1.03E-02	1.01E-02	1.02E-02	1.02E-02	1.02E-02	1.00E-02
9	9.89E-03	1.03E-02	1.01E-02	1.01E-02	1.00E-02	9.69E-03	1.02E-02	9.89E-03	1.01E-02	1.02E-02	1.02E-02	9.93E-03	1.01E-02	1.00E-02	9.73E-03	1.02E-02	1.01E-02	1.02E-02	1.02E-02	1.01E-02	1.01E-02	1.02E-02	1.02E-02	1.00E-02
10	9.79E-03	1.01E-02	1.01E-02	1.00E-02	9.93E-03	9.58E-03	1.01E-02	9.78E-03	9.99E-03	1.01E-02	1.00E-02	9.79E-03	9.94E-03	9.88E-03	9.58E-03	1.00E-02	1.01E-02	1.01E-02	1.01E-02	9.99E-03	9.96E-03	1.01E-02	1.01E-02	9.89E-03
11	9.72E-03	9.97E-03	1.00E-02	9.94E-03	9.86E-03	9.51E-03	9.89E-03	9.69E-03	9.89E-03	9.92E-03	9.84E-03	9.68E-03	9.80E-03	9.74E-03	9.48E-03	9.81E-03	1.01E-02	9.95E-03	9.90E-03	9.87E-03	9.82E-03	9.94E-03	9.92E-03	9.79E-03
12	9.69E-03	9.81E-03	9.98E-03	9.90E-03	9.84E-03	9.48E-03	9.75E-03	9.66E-03	9.82E-03	9.79E-03	9.71E-03	9.63E-03	9.70E-03	9.62E-03	9.46E-03	9.64E-03	1.01E-02	9.86E-03	9.78E-03	9.78E-03	9.72E-03	9.85E-03	9.81E-03	9.73E-03
13	9.72E-03	9.69E-03	9.96E-03	9.91E-03	9.88E-03	9.50E-03	9.65E-03	9.69E-03	9.79E-03	9.72E-03	9.63E-03	9.64E-03	9.67E-03	9.52E-03	9.55E-03	9.51E-03	1.02E-02	9.80E-03	9.70E-03	9.73E-03	9.68E-03	9.79E-03	9.73E-03	9.73E-03
14	9.79E-03	9.60E-03	9.97E-03	9.93E-03	9.96E-03	9.57E-03	9.62E-03	9.80E-03	9.77E-03	9.68E-03	9.61E-03	9.71E-03	9.71E-03	9.45E-03	9.72E-03	9.44E-03	1.02E-02	9.77E-03	9.69E-03	9.70E-03	9.69E-03	9.78E-03	9.69E-03	9.75E-03
15	9.89E-03	9.55E-03	9.97E-03	9.95E-03	1.00E-02	9.68E-03	9.66E-03	9.95E-03	9.77E-03	9.67E-03	9.63E-03	9.80E-03	9.79E-03	9.39E-03	9.91E-03	9.43E-03	1.02E-02	9.76E-03	9.74E-03	9.70E-03	9.71E-03	9.78E-03	9.65E-03	9.78E-03
16	1.00E-02	9.53E-03	9.98E-03	9.95E-03	1.01E-02	9.82E-03	9.75E-03	1.01E-02	9.77E-03	9.68E-03	9.69E-03	9.90E-03	9.89E-03	9.36E-03	1.01E-02	9.46E-03	1.02E-02	9.76E-03	9.82E-03	9.71E-03	9.74E-03	9.79E-03	9.61E-03	9.81E-03
17	1.01E-02	9.55E-03	1.00E-02	9.93E-03	1.01E-02	9.96E-03	9.88E-03	1.03E-02	9.79E-03	9.69E-03	9.76E-03	9.97E-03	9.97E-03	9.36E-03	1.03E-02	9.52E-03	1.01E-02	9.77E-03	9.94E-03	9.72E-03	9.76E-03	9.82E-03	9.54E-03	9.85E-03
18	1.01E-02	9.61E-03	1.00E-02	9.89E-03	1.00E-02	1.00E-02	1.00E-02	1.04E-02	9.83E-03	9.71E-03	9.81E-03	1.00E-02	1.00E-02	9.38E-03	1.04E-02	9.61E-03	1.00E-02	9.79E-03	1.01E-02	9.73E-03	9.78E-03	9.86E-03	9.46E-03	9.89E-03
19	1.01E-02	9.70E-03	1.00E-02	9.83E-03	9.93E-03	1.01E-02	1.01E-02	1.04E-02	9.89E-03	9.72E-03	9.85E-03	1.00E-02	1.00E-02	9.40E-03	1.04E-02	9.72E-03	9.93E-03	9.82E-03	1.02E-02	9.74E-03	9.81E-03	9.92E-03	9.42E-03	9.92E-03
20	1.00E-02	9.80E-03	1.01E-02	9.77E-03	9.80E-03	1.00E-02	1.02E-02	1.03E-02	9.96E-03	9.73E-03	9.85E-03	9.96E-03	1.00E-02	9.44E-03	1.05E-02	9.84E-03	9.86E-03	9.85E-03	1.02E-02	9.75E-03	9.85E-03	9.98E-03	9.44E-03	9.94E-03
21	9.94E-03	9.89E-03	1.01E-02	9.70E-03	9.69E-03	9.93E-03	1.02E-02	1.02E-02	1.00E-02	9.72E-03	9.83E-03	9.89E-03	9.97E-03	9.49E-03	1.05E-02	9.96E-03	9.82E-03	9.88E-03	1.02E-02	9.76E-03	9.91E-03	1.00E-02	9.54E-03	9.94E-03
22	9.83E-03	9.95E-03	1.02E-02	9.66E-03	9.60E-03	9.82E-03	1.02E-02	1.01E-02	1.01E-02	9.70E-03	9.80E-03	9.82E-03	9.90E-03	9.57E-03	1.04E-02	1.00E-02	9.80E-03	9.88E-03	1.02E-02	9.80E-03	9.98E-03	1.00E-02	9.69E-03	9.93E-03
23	9.73E-03	9.99E-03	1.03E-02	9.64E-03	9.55E-03	9.70E-03	1.01E-02	9.98E-03	1.01E-02	9.69E-03	9.78E-03	9.79E-03	9.82E-03	9.66E-03	1.02E-02	1.01E-02	9.80E-03	9.86E-03	1.01E-02	9.87E-03	1.00E-02	9.99E-03	9.87E-03	9.91E-03
24	9.66E-03	1.00E-02	1.04E-02	9.66E-03	9.55E-03	9.61E-03	1.00E-02	9.96E-03	1.00E-02	9.70E-03	9.78E-03	9.82E-03	9.75E-03	9.76E-03	1.00E-02	1.00E-02	9.81E-03	9.83E-03	1.00E-02	9.95E-03	1.01E-02	9.90E-03	1.00E-02	9.89E-03
25	9.61E-03	9.99E-03	1.04E-02	9.71E-03	9.58E-03	9.54E-03	9.92E-03	9.98E-03	9.92E-03	9.76E-03	9.81E-03	9.90E-03	9.71E-03	9.85E-03	9.75E-03	9.96E-03	9.80E-03	9.79E-03	9.89E-03	1.00E-02	1.01E-02	9.78E-03	1.02E-02	9.87E-03
26	9.59E-03	9.99E-03	1.05E-02	9.79E-03	9.63E-03	9.50E-03	9.81E-03	1.00E-02	9.82E-03	9.86E-03	9.86E-03	1.00E-02	9.71E-03	9.93E-03	9.52E-03	9.84E-03	9.78E-03	9.75E-03	9.74E-03	1.01E-02	1.00E-02	9.67E-03	1.03E-02	9.86E-03
27	9.61E-03	9.98E-03	1.05E-02	9.87E-03	9.71E-03	9.47E-03	9.72E-03	1.01E-02	9.73E-03	9.97E-03	9.94E-03	1.01E-02	9.75E-03	9.99E-03	9.34E-03	9.72E-03	9.74E-03	9.70E-03	9.62E-03	1.01E-02	9.96E-03	9.58E-03	1.03E-02	9.86E-03
28	9.67E-03	9.96E-03	1.04E-02	9.94E-03	9.79E-03	9.50E-03	9.67E-03	1.02E-02	9.66E-03	1.01E-02	1.00E-02	1.02E-02	9.82E-03	1.00E-02	9.22E-03	9.64E-03	9.67E-03	9.67E-03	9.54E-03	1.00E-02	9.85E-03	9.52E-03	1.03E-02	9.87E-03
29	9.75E-03	9.95E-03	1.04E-02	9.97E-03	9.85E-03	9.60E-03	9.66E-03	1.02E-02	9.65E-03	1.01E-02	1.01E-02	1.02E-02	9.88E-03	1.00E-02	9.15E-03	9.62E-03	9.56E-03	9.66E-03	9.53E-03	9.94E-03	9.75E-03	9.51E-03	1.02E-02	9.90E-03
30	9.84E-03	9.94E-03	1.04E-02	9.96E-03	9.86E-03	9.78E-03	9.65E-03	1.03E-02	9.71E-03	1.01E-02	1.02E-02	1.02E-02	9.92E-03	1.00E-02	9.12E-03	9.67E-03	9.42E-03	9.67E-03	9.59E-03	9.84E-03	9.68E-03	9.55E-03	1.01E-02	9.94E-03
31	9.93E-03	9.93E-03	1.03E-02	9.92E-03	9.80E-03	1.00E-02	9.63E-03	1.03E-02	9.83E-03	1.01E-02	1.03E-02	1.01E-02	9.90E-03	1.00E-02	9.14E-03	9.78E-03	9.27E-03	9.73E-03	9.70E-03	9.73E-03	9.65E-03	9.62E-03	9.92E-03	1.00E-02
32	1.00E-02	9.91E-03	1.02E-02	9.85E-03	9.69E-03	1.03E-02	9.60E-03	1.02E-02	1.00E-02	1.00E-02	1.03E-02	9.96E-03	9.84E-03	1.00E-02	9.22E-03	9.93E-03	9.14E-03	9.81E-03	9.86E-03	9.63E-03	9.69E-03	9.72E-03	9.76E-03	1.01E-02
33	1.01E-02	9.90E-03	1.01E-02	9.76E-03	9.57E-03	1.05E-02	9.57E-03	9.95E-03	1.02E-02	9.90E-03	1.03E-02	9.83E-03	9.74E-03	9.97E-03	9.36E-03	1.01E-02	9.06E-03	9.91E-03	1.00E-02	9.57E-03	9.78E-03	9.85E-03	9.64E-03	1.01E-02
34	1.01E-02	9.87E-03	9.88E-03	9.68E-03	9.46E-03	1.06E-02	9.54E-03	9.64E-03	1.03E-02	9.82E-03	1.02E-02	9.70E-03	9.63E-03	9.91E-03	9.54E-03	1.02E-02	9.05E-03	1.00E-02	1.01E-02	9.56E-03	9.91E-03	9.98E-03	9.60E-03	1.01E-02
35	1.01E-02	9.82E-03	9.70E-03	9.61E-03	9.40E-03	1.05E-02	9.53E-03	9.29E-03	1.03E-02	9.76E-03	1.01E-02	9.59E-03	9.55E-03	9.82E-03	9.75E-03	1.03E-02	9.10E-03	1.01E-02	1.02E-02	9.62E-03	1.01E-02	1.01E-02	9.63E-03	1.01E-02
36	1.01E-02	9.74E-03	9.51E-03	9.58E-03	9.40E-03	1.03E-02	9.55E-03	8.96E-03	1.03E-02	9.74E-03	1.00E-02	9.51E-03	9.53E-03	9.71E-03	9.96E-03	1.04E-02	9.21E-03	1.02E-02	1.02E-02	9.74E-03	1.02E-02	1.01E-02	9.73E-03	1.00E-02
37	9.96E-03	9.64E-03	9.36E-03	9.59E-03	9.47E-03	1.01E-02	9.60E-03	8.71E-03	1.02E-02	9.76E-03	9.89E-03	9.47E-03	9.56E-03	9.58E-03	1.01E-02	1.03E-02	9.37E-03	1.02E-02	1.02E-02	9.89E-03	1.03E-02	1.02E-02	9.86E-03	9.87E-03
38	9.82E-03	9.55E-03	9.26E-03	9.65E-03	9.59E-03	9.80E-03	9.67E-03	8.57E-03	9.98E-03	9.82E-03	9.76E-03	9.48E-03	9.63E-03	9.46E-03	1.03E-02	1.01E-02	9.55E-03	1.02E-02	1.01E-02	1.00E-02	1.03E-02	1.01E-02	1.00E-02	9.74E-03
39	9.67E-03	9.49E-03	9.22E-03	9.74E-03	9.73E-03	9.50E-03	9.75E-03	8.54E-03	9.77E-03	9.90E-03	9.66E-03	9.56E-03	9.74E-03	9.33E-03	1.03E-02	9.91E-03	9.73E-03	1.01E-02	9.94E-03	1.02E-02	1.02E-02	1.01E-02	1.01E-02	9.63E-03
40	9.56E-03	9.50E-03	9.25E-03	9.85E-03	9.89E-03	9.25E-03	9.82E-03	8.57E-03	9.60E-03	9.99E-03	9.61E-03	9.67E-03	9.86E-03	9.21E-03	1.03E-02	9.71E-03	9.86E-03	1.00E-02	9.80E-03	1.02E-02	1.01E-02	1.00E-02	1.02E-02	9.55E-03
41	9.49E-03	9.57E-03	9.36E-03	9.95E-03	1.00E-02	9.09E-03	9.89E-03	8.63E-03	9.47E-03	1.00E-02	9.62E-03	9.80E-03	9.96E-03	9.08E-03	1.03E-02	9.58E-03	9.92E-03	9.83E-03	9.71E-03	1.02E-02	9.89E-03	1.00E-02	1.02E-02	9.53E-03
42	9.49E-03	9.68E-03	9.54E-03	1.00E-02	1.01E-02	9.05E-03	9.94E-03	8.69E-03	9.40E-03	1.00E-02	9.68E-03	9.90E-03	1.00E-02	8.97E-03	1.02E-02	9.54E-03	9.90E-03	9.65E-03	9.69E-03	1.01E-02	9.70E-03	1.00E-02	1.02E-02	9.57E-03
43	9.54E-03	9.80E-03	9.76E-03	1.01E-02	1.01E-02	9.15E-03	9.98E-03	8.73E-03	9.40E-03	1.00E-02	9.79E-03	9.95E-03	1.01E-02	8.90E-03	1.02E-02	9.61E-03	9.81E-03	9.49E-03	9.74E-03	9.98E-03	9.57E-03	1.00E-02	1.01E-02	9.67E-03
44	9.62E-03	9.91E-03	1.00E-02	1.01E-02	1.01E-02	9.36E-03	1.00E-02	8.77E-03	9.43E-03	9.92E-03	9.92E-03	9.96E-03	1.01E-02	8.88E-03	1.01E-02	9.73E-03	9.67E-03	9.39E-03	9.83E-03	9.88E-03	9.51E-03	1.01E-02	1.00E-02	9.79E-03
45	9.71E-03	1.00E-02	1.02E-02	1.00E-02	1.01E-02	9.64E-03	1.00E-02	8.80E-03	9.51E-03	9.81E-03	1.00E-02	9.96E-03	1.00E-02	8.92E-03	1.00E-02	9.86E-03	9.51E-03	9.36E-03	9.93E-03	9.81E-03	9.53E-03	1.01E-02	9.89E-03	9.89E-03
46	9.79E-03	1.01E-02	1.04E-02	9.93E-03	9.97E-03	9.95E-03	1.00E-02	8.84E-03	9.61E-03	9.69E-03	1.02E-02	9.94E-03	9.95E-03	9.01E-03	9.91E-03	9.96E-03	9.35E-03	9.42E-03	1.00E-02	9.78E-03	9.59E-03	1.00E-02	9.77E-03	9.95E-03
47	9.82E-03	1.02E-02	1.05E-02	9.86E-03	9.88E-03	1.03E-02	1.00E-02	8.86E-03	9.71E-03	9.62E-03	1.02E-02	9.90E-03	9.89E-03	9.16E-03	9.84E-03	1.00E-02	9.21E-03	9.55E-03	1.01E-02	9.81E-03	9.66E-03	9.95E-03	9.67E-03	9.93E-03
48	9.82E-03	1.03E-02	1.06E-02	9.83E-03	9.80E-03	1.05E-02	1.00E-02	8.90E-03	9.74E-03	9.63E-03	1.02E-02	9.83E-03	9.86E-03	9.33E-03	9.79E-03	9.99E-03	9.13E-03	9.70E-03	1.01E-02	9.86E-03	9.72E-03	9.87E-03	9.58E-03	9.85E-03
49	9.79E-03	1.03E-02	1.06E-02	9.85E-03	9.75E-03	1.07E-02	9.98E-03	8.94E-03	9.71E-03	9.71E-03	1.02E-02	9.73E-03	9.88E-03	9.50E-03	9.78E-03	9.93E-03	9.11E-03	9.84E-03	1.01E-02	9.92E-03	9.77E-03	9.78E-03	9.51E-03	9.71E-03
50	9.75E-03	1.02E-02	1.06E-02	9.89E-03	9.77E-03	1.08E-02	9.93E-03	9.02E-03	9.62E-03	9.86E-03	1.01E-02	9.61E-03	9.95E-03	9.66E-03	9.82E-03	9.85E-03	9.12E-03	9.95E-03	9.96E-03	9.98E-03	9.83E-03	9.70E-03	9.45E-03	9.54E-03
51	9.75E-03	1.01E-02	1.05E-02	9.93E-03	9.84E-03	1.07E-02	9.86E-03	9.14E-03	9.53E-03	1.00E-02	1.00E-02	9.51E-03	1.00E-02	9.78E-03	9.91E-03	9.77E-03	9.15E-03	1.00E-02	9.83E-03	1.00E-02	9.90E-03	9.65E-03	9.41E-03	9.39E-03
52	9.80E-03	9.98E-03	1.04E-02	9.93E-03	9.95E-03	1.06E-02	9.78E-03	9.30E-03	9.46E-03	1.01E-02	9.93E-03	9.47E-03	1.02E-02	9.87E-03	1.00E-02	9.75E-03	9.16E-03	9.98E-03	9.69E-03	1.01E-02	9.98E-03	9.65E-03	9.37E-03	9.27E-03
53	9.91E-03	9.89E-03	1.03E-02	9.89E-03	1.01E-02	1.04E-02	9.72E-03	9.48E-03	9.45E-03	1.02E-02	9.82E-03	9.50E-03	1.02E-02	9.92E-03	1.01E-02	9.78E-03	9.19E-03	9.92E-03	9.60E-03	1.01E-02	1.00E-02	9.71E-03	9.34E-03	9.23E-03
54	1.01E-02	9.86E-03	1.02E-02	9.83E-03	1.02E-02	1.01E-02	9.70E-03	9.68E-03	9.51E-03	1.01E-02	9.71E-03	9.60E-03	1.02E-02	9.95E-03	1.02E-02	9.88E-03	9.27E-03	9.82E-03	9.59E-03	1.01E-02	1.00E-02	9.81E-03	9.33E-03	9.27E-03
55	1.03E-02	9.88E-03	1.02E-02	9.77E-03	1.02E-02	9.92E-03	9.74E-03	9.88E-03	9.63E-03	1.01E-02	9.64E-03	9.73E-03	1.02E-02	9.95E-03	1.03E-02	1.00E-02	9.43E-03	9.70E-03	9.69E-03	1.01E-02	1.00E-02	9.93E-03	9.36E-03	9.41E-03

Fig 4.1: Screen shot of data matrix.

First the Multiple Regression Analysis is performed using Minitab. This is stage one of our methodology, from which statistical interpretations are made so as to develop Matlab code to predict data accordingly. From the equation obtained after Regression Analysis using Minitab, it is observed that about 90 variables have more significance over the data. It means without considering these less predominant variables, the data can be reduced in its dimensionality by not loosing significant properties of original data. This helps us to write Matlab code accordingly.

The Regression Equation obtained from Matlab analysis by giving time values as 1202 independent predictors (X values) and each run (Y value) as a response value.

Y = 6.04 + 44.4 X1 - 25.2 X2 - 76.9 X3 + 75.8 X4 - 69.1 X5 + 8.0 X6 + 188 X7 + 56.0 X8 + 149 X9 - 190 X10 + 159 X11 + 67.0 X12 + 30.3 X13 + 81.0 X14 + 16.2 X15 - 72.4 X16 + 17.4 X17 - 124 X18 - 5.1 X19 - 68.5 X20 + 165 X21 - 22.1 X22 + 131 X23 + 150 X24 + 66.3 X25 - 5.9 X26 - 244 X27 + 5.4 X28 + 83.7 X29 + 214 X30 + 11.7 X31 + 69.7 X32 - 191 X33 - 140 X34 + 38.3 X35 + 55.1 X36 - 75.0 X37 + 276 X38 - 18.5 X39 - 55.9 X40 - 120 X41 - 16.5 X42 + 76.4 X43 - 0.1 X44 - 28.1 X45 - 4.0 X46 - 50.9 X47 - 57.6 X48 - 45.1 X49 - 67.3 X50 + 118 X51 - 118 X52 + 18.2 X53 + 199 X54 - 17.8 X55 + 117 X56 - 104 X57 - 98.9 X58 + 31.8 X59 - 248 X60 + 18.0 X61 - 32.3 X62 - 30.4 X63 + 156 X64 - 111 X65 + 61.9 X66 - 116 X67 + 265 X68 + 13.1 X69 - 155 X70 - 3.6 X71 + 200 X72 + 43.5 X73 + 11.7 X74 - 153 X75 + 151 X76 + 73.2 X77 - 260 X78 + 11.9 X79 - 165 X80 - 51.5 X81 + 188 X82 + 120 X83 + 150 X84 - 283 X85 + 139 X86 - 403 X87 - 19.2 X88 + 113 X89 - 13.2 X90

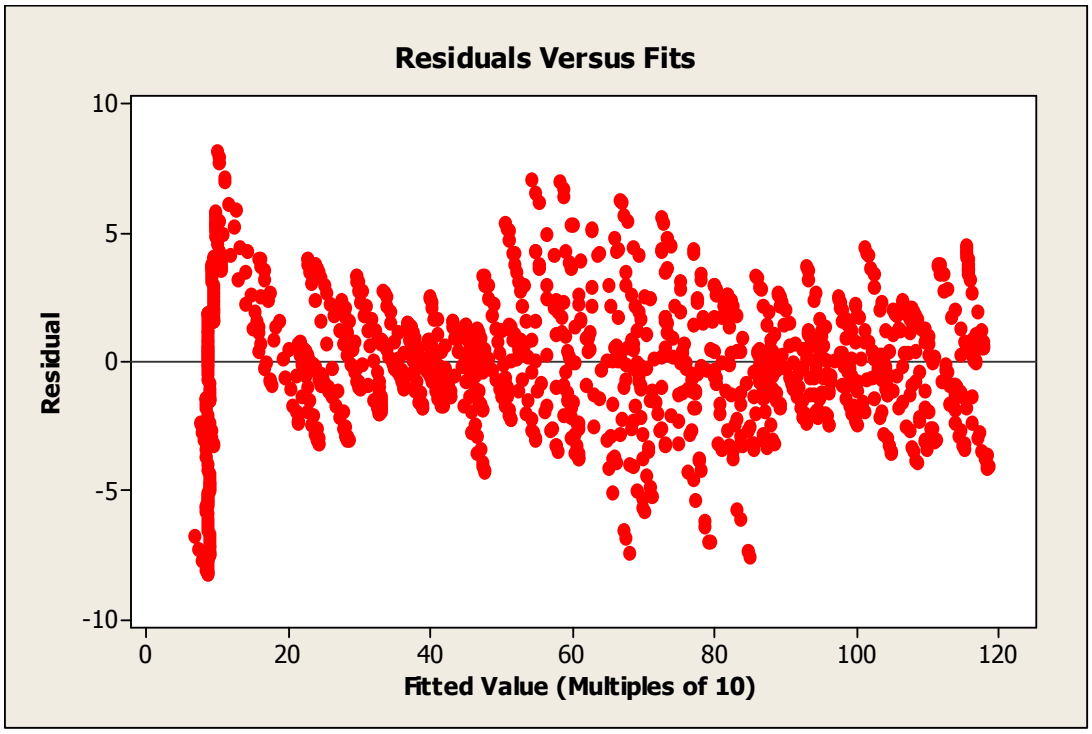


Fig 4.2: Plot of residuals versus fits.

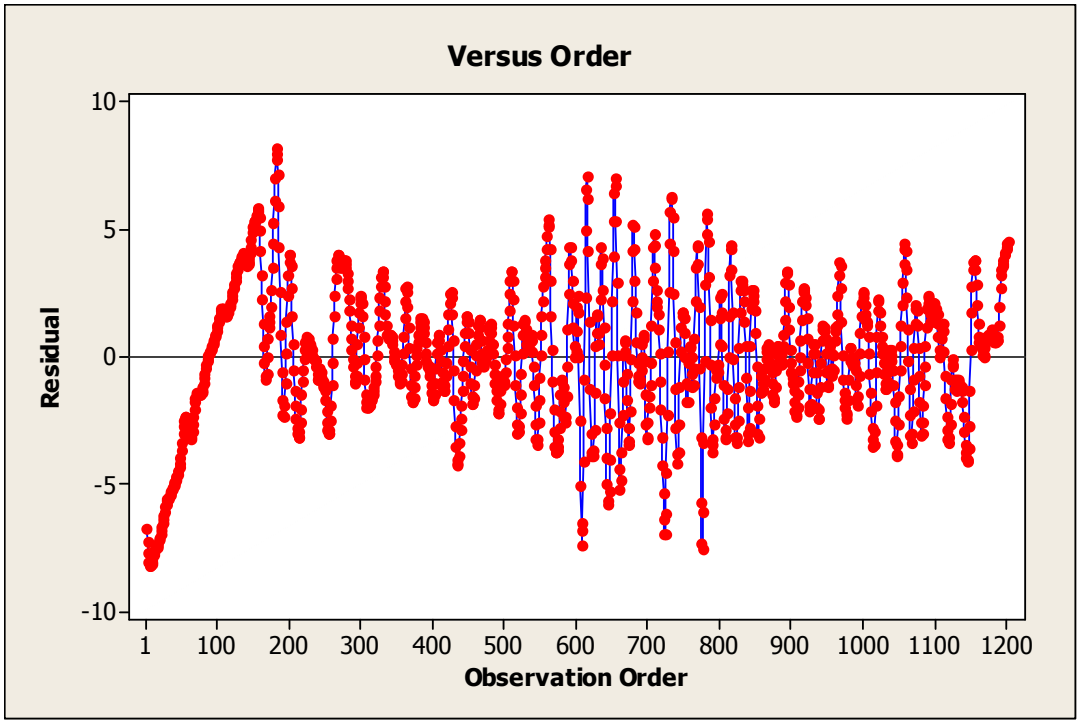


Fig 4.3: Plot of residuals versus order.

From the above graphs is it observed that a lot of variations present at each ends of the plots. Residuals versus fits plot should show a random pattern of residuals on both sides of 0. If a point lies far from the majority of points, it may be an outlier. Also, there should not be any recognizable patterns in the residual plot. Residuals versus order plot of all residuals in the order that the data was collected and can be used to find non-random error, especially of time-related effects. A positive correlation is indicated by a clustering of residuals with the same sign. A negative correlation is indicated by rapid changes in the signs of consecutive residuals. Important variables left out of the model to see if they have critical additional effects on the response. This interpretation gives an idea to build Matlab code to predict data using regression.

4.2 Analysis and results from PCA:

Principal Component Analysis with the use of Singular Value Decomposition Theorem is performed to the data to check the variability. First the PCA of the data is done using Minitab so as to interpret the statistical variation present in the data. From the results it is inferred that the first principal component has variance (eigenvalue) 80.858 and accounts for 90.9% of the total variance. The coefficients listed under Principal Component 1 (PC1) show how to calculate the principal component scores. It should be noted that the interpretation of the principal components is subjective; however, obvious patterns emerge quite often because the coefficients of these terms have the same sign and are not close to zero. The second principal component has variance 1.982 and accounts for only 2.2% of the data variability. It is calculated from the original data using the coefficients listed under PC2. Together, the first two and the first four principal components represent 95.7% of the total variability. Thus, most of the data structure can be captured in two or three underlying dimensions. The remaining principal components account for a very small proportion of the variability and are probably unimportant. The Scree plot provides this information visually. Hence using PCA the dimensionality can be reduced to the maximum without losing the properties of whole data. This interpretation gives idea to write Matlab code so as to reduce the data also for prediction of the data.

Principal Component Analysis: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Eigenanalysis of the Correlation Matrix

Eigenvalue	80.858	1.982	1.310	0.993	0.876	0.575	0.446	0.307	0.252
Proportion	0.909	0.022	0.015	0.011	0.010	0.006	0.005	0.003	0.003
Cumulative	0.909	0.931	0.946	0.957	0.967	0.973	0.978	0.981	0.984

Eigenvalue	0.167	0.153	0.139	0.132	0.114	0.101	0.087	0.069	0.054
Proportion	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001
Cumulative	0.986	0.988	0.989	0.991	0.992	0.993	0.994	0.995	0.996

Eigenvalue	0.039	0.035	0.031	0.028	0.025	0.022	0.019	0.017	0.016
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	0.996	0.997	0.997	0.997	0.997	0.998	0.998	0.998	0.998

Eigenvalue	0.013	0.012	0.011	0.010	0.009	0.008	0.007	0.006	0.006
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	0.998	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Eigenvalue	0.006	0.005	0.005	0.004	0.004	0.004	0.003	0.003	0.003
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	0.999	0.999	0.999	0.999	0.999	1.000	1.000	1.000	1.000

Eigenvalue	0.003	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Fig 4.4: PCA for Response matrix.

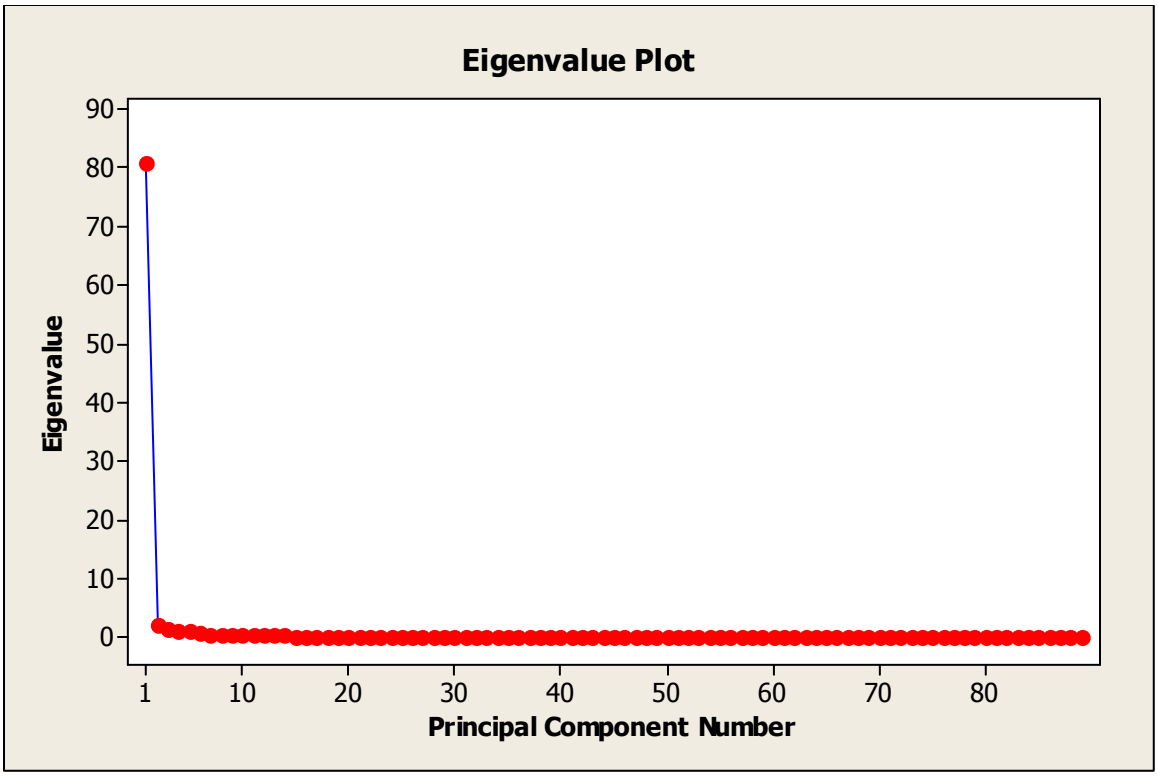


Fig 4.5: Plot of eigen values.

4.2.1 Generating Matlab codes to predict data:

Originally the dimension of the response matrix is 90×1202. But, here we are considering 80×1202 and remaining data is used for validating the results. Thus, the dimension of data1 is 80 ×1202. The design matrix is renamed as data2mod.

The data obtained after SVD is

S matrix: 80*80

U matrix: 1202*80

V matrix: 1202*1202

According to the concepts of PCA, a matrix Φ is calculated by using S and U matrix.

$\Phi = U \times S$

After reducing the dimension of the dataset, next step is to approximate the data. Normally, it is not possible to approximate the data when both design and response matrices are likely to be high dimensional. In order to perform this task, we are using software package called DACE (Design and Analysis of Computer Experiments), which is a Matlab toolbox. The typical use of this software package is to construct kringing approximation model based on data from computer experiment and to use this approximation model as a substitute for the computer model.

The matrix Φ is predicted using DACE tool box. A function ' *dmodel* ' is used for predicting this matrix, which is an inbuilt function. Matlab function *dacefit* in dace toolbox is used to model this data. Dace tool box has different regression and correlation functions. *regr* and *corr* are regression model and correlation functions are provided by user. Dacefit uses a non-linear least-squares algorithm based on correlation and regression model to model the given data set. Modeling the data of the large data set can be generated by an equation obtained using this method. This is similar to data compression. Also, data can be predicted using this equation.

```
load data1.txt;

data1=data1';

load data2mod.txt;

load data1mod.txt;
```

data1mod is original response matrix and data2mod is design matrix.

Input arguments are defined as follows:

data2, y - data sets

regr - Function handle to a regression model

corr - Function handle to a correlation function

theta0 - Initial guess on theta, the correlation function parameters

lob,upb - If present, then lower and upper bounds on theta. Otherwise, theta0 is used for theta

Output arguments are defined as follows:

dmodel - is a structure with elements. Model parameters are one of the elements in the strutcture.

perf - structure with performance information.

The obtained equation from the regression analysis using Matlab is used to develop a code for prediction of the data. Here the least squares method for Multiple Regression analysis of the given data is implemented in developing the code. The Least squares method iteratively generates the coefficients for polynomial equation with order given by us and calculates the error square between the estimated and the actual curve. This method returns the coefficients of the polynomial equation for which the error square between the actual curve and the estimated curve is minimum. The code generated using assumptions from regression analysis gives plot between predicted values and original values represented by blue and red curves successively.

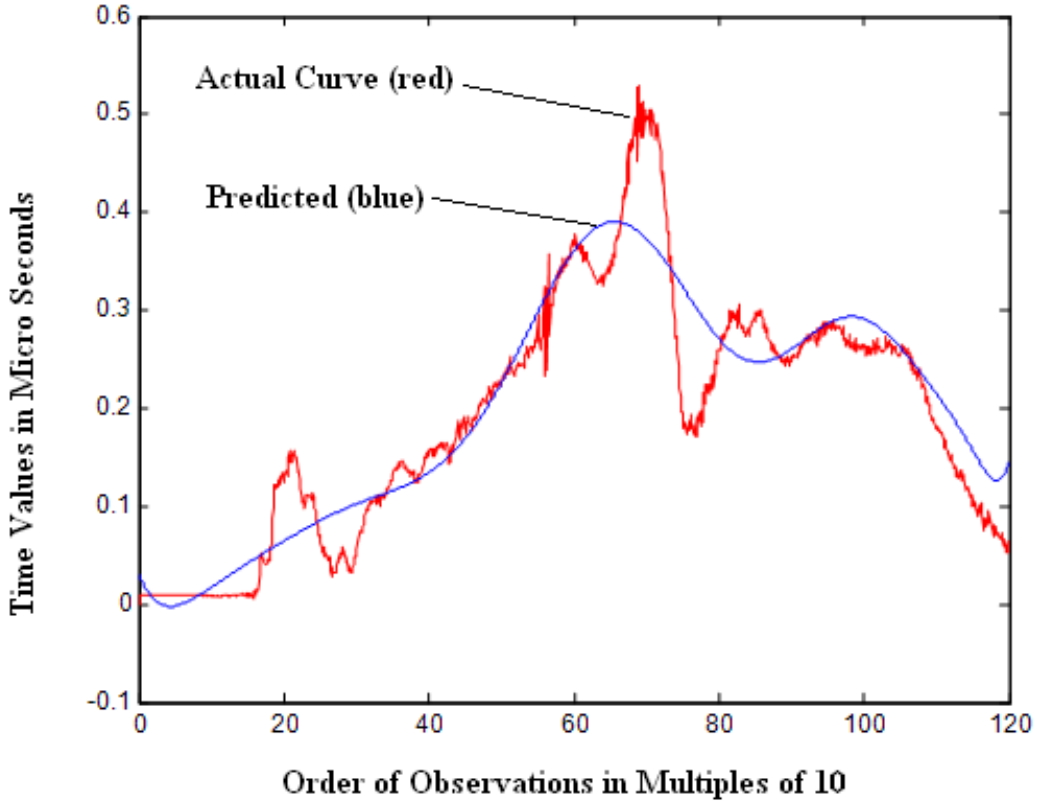


Fig 4.6: Plot between predicted data and actual data.

From the above plot, it is assumed that the curve fit for dataset1-smooth curve in blue (predicted) and dataset 2-polynomial curve in red (original) which are obtained by polynomial functions are being corrupted by some noise (error).

The computer code in Matlab which uses the least square method to do multiple regression analysis calculates the noise or error in the fit, which are measured as Mean Square Errors (MSE) of the curve fitting. The last 10 rows are left for validation of the results and hence for each row of predicted curve fitting the MSE values are obtained.

The calculated mean square error values are observed as followed.

Table 4.1: MSE values for 10 vaidating rows after Multiple regression.

Row value	Corresponding MSE
Validating Row 1	0.3212126
Validating Row 2	0.3013785
Validating Row 3	0.29040899
Validating Row 4	0.3013569
Validating Row 5	0.10184021
Validating Row 6	0.2008367
Validating Row 7	0.301311
Validating Row 8	0.204465
Validating Row 9	0.2075593
Validating Row 10	0.27070113

The obtained MSE values are relatively high. So it showed the necessity of adaptation of another method for approximation.

Since the response matrix available to us is very large it has become very difficult to approximate the data to fit responsiveness. Hence it is decided to reduce the dimension of data without losing redundancy in it keeping the properties of the data. It is noticed that the Singular Value Decomposition (SVD) could help to resolve reducing MSE values.

The PCD is applied using this equation

[U, S, V] = svd (data1);

From S matrix, the number of dominant singular values is 20. Hence we choose 20 principal components. Thus d matrix is chosen from S which is 20×20 diagonal matrix, accordingly U and V are partitioned. Before partitioning V matrix it is transposed by the definition of PCD.

```
d=S(1:20,1:20) ;

phi=U(1:80,1:20) ;

Vtrans=V' ;

Eta=Vtrans(1:80,1:1202);

y=phi*d;
```

Here y is calculated according to PCD definition used for prediction.

```
xnew=data2mod(81:90,1:13) ;
```

xnew is last 10 rows of design matrix used for validation purpose.

In the Matlab code, *theta0* are the initial guesses, this is the number parameters in the equation that models the data set. *lob* and *upb* are the lower bound and upper bound for the parameters. Parameters in the equation should not be less than the value provided by the lower bound and not greater than the upper bound value. Function *dacefit* returns a model for the data set data2 and y using the correlation and regression model functions mentioned by the user.

Output *dmodel* returned from the function *dacefit* is a structure that has model parameters as one of its elements. This output argument *dmodel* along with design matrix *xnew* is passed to a function predictor as input arguments to predict y values other than the values available.

As discussed above using DACE the data is predicted and approximated. The algorithm for predicted data is calculated according to the definition of PCD. In order to compare the predicted values and original values, we are considering last 10 rows, we are taking last 10 rows of the response matrix as Ztrue

Plots are generated for each row of predicted data versus actual data after refining the data using PCA in Matlab code from the Regression Polynomial fit. The dataset1 which is the predicted value of the data represented by blue curve and the actual data, as dataset 2 is represented as red curve.

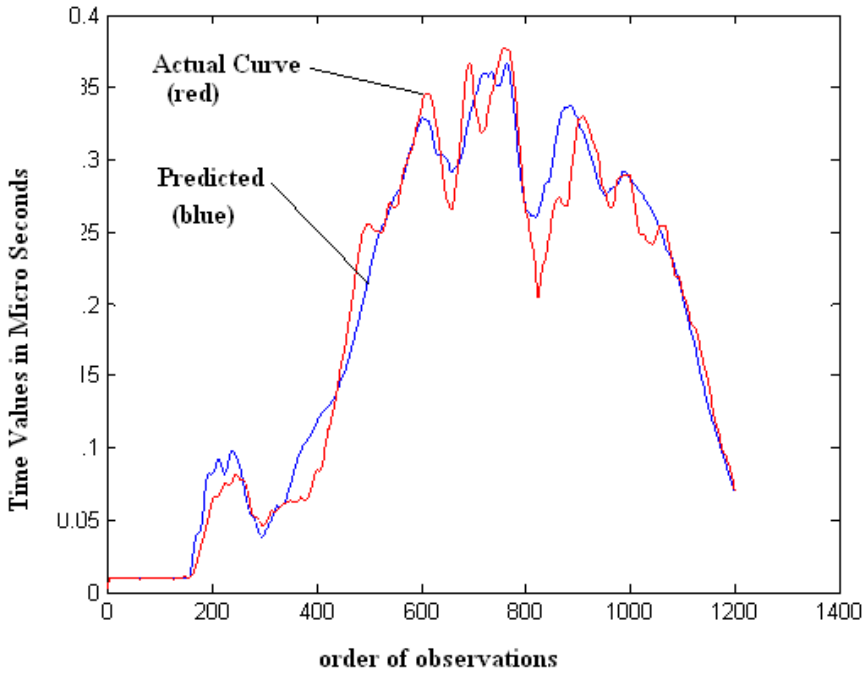


Fig 4.7: Plot between true and predicted data.

After plotting Zpredict and Ztrue matrices, their mean square error values are computed to check the fit of the responsiveness.

Mean square Error values are calculated by taking 20 Principal components as well as 80 principal components.

Table 4.2: MSE values for 10 vaidating rows after PCA.

Row value	Corresponding MSE With 20 PCs	Corresponding MSE With 80 PCs
Validating Row 1	0.0012126	0.0012444
Validating Row 2	0.0013785	0.0013757
Validating Row 3	0.0040899	0.0041155
Validating Row 4	0.0013569	0.0013763
Validating Row 5	0.0018402	0.0018583
Validating Row 6	0.0008367	0.00083896
Validating Row 7	0.001311	0. 0013079
Validating Row 8	0.0004465	0. 00044759
Validating Row 9	0.00075593	0. 00075575
Validating Row 10	0.00070113	0. 00069407

Chapter 5: Conclusions

Up to this point in the present research, the two technologies Multiple Regression Analysis and Principal Component Analysis are investigated the interrelationship between the time dependent variables present in the data obtained from crash testing of a vehicle and prediction of the experimental data. This chapter finalizes the present research deriving conclusions from the analysis and results obtained from chapter four. It has been presented in this work a practical comparison of two approaches mentioned by the literature to predict the data obtained from crash testing of vehicles so that the prediction of data is maintained to be accurate with less mean square errors.

Table 5.1: Comparison between MSE values for 10 validating rows.

Row value	Corresponding MSE After PCA	Corresponding MSE After Multiple regression
Validating Row 1	0.0012126	0.3212126
Validating Row 2	0.0013785	0.3013785
Validating Row 3	0.0040899	0.29040899
Validating Row 4	0.0013569	0.3013569
Validating Row 5	0.0018402	0.10184021
Validating Row 6	0.0008367	0.2008367
Validating Row 7	0.001311	0.301311
Validating Row 8	0.0004465	0.204465
Validating Row 9	0.00075593	0.2075593
Validating Row 10	0.00070113	0.27070113

It clearly shows there is a significant redundancy error in the predicted values using Principal Component Analysis when compared to the Multiple Regression Analysis. The data reduction carried out in Principal Component Analysis is more effective than Multiple regression Analysis.

The reduced matrix is approximated to compare with the original design matrix. The mean square error values obtained after PCA are compared to the mean square error of regression analysis and are proven that PCA is more efficiently done the approximation compared to the regression.

The main advantage is the data is reduced to a lower order matrix 80×20 with eigenvalues. The predicted matrix is compared to the design matrix stating that by selecting less number of actual experimental runs, data is adequate to run simulations. Instead of running 90 experimental runs and taking 1200 observations it is proposed by taking only 80 observations in 20 experimental runs. This helps the engineers use predicted data obtained after PCA to create simulations for design optimization. This saves money and time by decreasing number of experimental rounds in crash testing.

Data reduced more effectively using PCA with very less redundancy in its properties. Prediction of data is very significant with lesser mean square error values when applied PCA technique and a significant number of reduction in actual number of simulations carried out in crash testing of vehicles is proposed.

The presence of independent components and variability of independent variables in the data can be studied with the future research study with the help of Independent Component Analysis (ICA)

References

1. Potential Use of Crash Test Data for Crashworthiness Research by M Paine and M Griffiths

2. A Frame Work for Validation of Computer Models by M.J. Bayarri, James O. Berger, David Higdon, Marc C. Kennedy, A. Kottas, Rui Paulo, Jerome Sacks, James A. Cafeo, James C. Cavendish, C.H. Lin, and J. Tui

3. Statistical validation of system models by Patrick Barney Carlos Ferregut Luis E. Perez Norman F. Hunter Thomas L. Paez

4. Methods of Multivariate Analysis second edition by Alvin C. Rencher

5. Applied Multivariate Research by Lawrence S. Meyers, Glenn Gamst, A. J. Guarino

6. Principle Component Analysis for Nonlinear Model Correlation Updating and Uncertainty Evaluation by K Hasselman, Mark C. Anderson, Wenshui Gan.

7. Frawley, Piatetsky-Shapiro, & Matheus, 1991; Hearst, 1999; Roddick & Spiliopoulou, 1999; Zaïane, O.R., Han, J., Li, Z., & Hou, J, 1998

8. CRM Data Mining: Methods of Dimensionality Reduction and Choosing A Right Technique Dr. Nethra Sambamoorthi, Ph.D CRMportals Inc., 11 Bartram Road Englishtown, NJ 07726

9. Modeling Complex Manufacturing Process Activities Using Data Minng – A Rough Set Approach, by Alexander Joseph Nadackal.

10. Analysis of Accuracy of Data Reduction Techniques by Pedro Futado and H. Madeira, University of Colombia, Portugal.

11. Online Sales Forecasting With Multiple Regression Analysis Data Matrices Package, by Aspy Plia, University of Hawaii at Manoa, Developments in Business Simulation and Experiential Learning, Volume 31, 2004

12. Data Reduction Techniques for the Analysis and Interpretation of Dynamic FDG-PET Oncological Studies by George Kontaxakis, Member, IEEE, Trias Thireou, Sotiris Pavlopoulos, Member, IEEE.

13. Clustering Microarry data using SVD and Gene Shaving by Terry Speed, Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC, 2003, pp. 190–200. April 13, 2005.

14. An Introduction to Regression analysis by Alan O. Sykes.

15. A multivariate regression analysis for deriving engineering Parameters of expansive soils from spectral reflectance by Fekerte Arega Yitagesu, Freek van der Meer, Harald van der Werff, Wolter Zigterman. International Institute for Geo-information Sciences and Earth Observation, ESA Department, Hengelosestraat, 7500AA Enschede, The Netherlands.

16. Regression Analysis by example by S. Chatterjee, A.S. Hadi. Wiely Publications.

17. Tutorial on Multivariate Logistic RegressionJavier R. Movellan July 23, 2006.

18. N.R. Draper and H. Smith (1981). Applied Regression Analysis, Second Edition. John Wiley & Sons, Inc.

19. Efficiency, performance and value-at-risk of Latin American banks in a process of economic integration Creamer, G.; Noe, T.; Spindt, P.

20. Multivariate statistical analysis in fMRI by Rowe, D.B.; Hoffmann, R.G.

21. Signal processing and engineering analysis on personal computers South, H.M.; Dolecek, Q.E.

22. Spatio-Temporal Patterns of Water Quality in Lower Chesapeake Bay (1984-85) Alden, R., III; Butt, A.; Sokolowski, S

23. Discrimination of the road condition toward understanding of vehicle driving environments by Yamada, M.; Ueda, K.; Horiba, I.; Sugie, N

24. Feature Selection in Pathology Detection using Hybrid Multidimensional Analysis by Castellanos, G.; Delgado, E.; Daza, G.; Sanchez, L.G.; Suarez, J.F

25. Multivariate analysis for probabilistic WLAN location determination systems by Youssef, M.; Abdallah, M.; Ashok Agrawala

26. Parallel analysis: method for determining significant principal components by Franklin,Scott B. Gibson ,David J. Robertson, Philip A., Pohlmann, jhonT.& Fralish, James S.

27. Methods of Multivariate Analysis second edition by Alvin C. Rencher.

28. PCA and SVD with nonnegative loadings by Stan Lipovetsky, GfK Custom Research North America, 8401 Golden Valley Rd., Minneapolis, MN 55427, USA

29. Principal component analysis model for machine-part cell formation problem in group technology Wafik Hachichaa , Faouzi Masmoudi a,b and Mohamed Haddar.

30. Online Banking performance evaluation using data envelopment and PCA by Chien Ta Bruce Hoa, Desheng Dash Wub, Instituteof Electronic Commerce, National Chung Hsing University, Taiwan, Risk Lab, University of Toronto, 19Borden ST, Toronto, ON, Canada.

31. Process systems engineering identification of faulty sensors using principal component analysis by Ricardo Dunia, S. Joe Qin Thomas F. Edgar Thomas J. McAvoy Fisher-Rosemount Systems, Austin, TX 78754 Dept. of Chemical Engineering, University of Texas at Austin, Austin, TX 78712 [3]Dept. of Chemical Engineering, University of Maryland, College Park, MD.

32. Principle component analysis of urban traffic characteristics and metrological data by Nagendra,Mukesh Khare.

33. Neural Network for Principal Component Analysis with Applications in Image compression Luminita STATE Dept. of Computer Science, University of Pitesti, Pitesti, Romania Catalina Lucia Cocianu Dept. of Computer Science, Academy of Economic Studies, Bucharest, ROMANIA.

34. Process fault detection and diagnosis based on principal component analysis by Tao He, Wei-Rong Xie, Qing-Hua Wu, Tie-lin Shi.

35. Supervised Principal Component Analysis for gene enrichment of microarray data with continuous survival outcomes, Chen X, Wang L, Smith JD, Zhang B. Department of

Quantitative Health Sciences, The Cleveland Clinic, 9500 Euclid Ave. Cleveland, OH 44195, USA.

36. Recording Automotive Crash Event Data by (Augustus "Chip" Chidester, National Highway Traffic Safety Administration) John Hinch, National Highway Traffic Safety Administration, Thomas C. Mercer, General Motors Corporation, Keith S. Schultz, General Motors Corporation.

37. Case Study – Volvo Cars Safety Centre, HEADQUARTERS: DK-2850 Nærum · Denmark.

38. Classical and Modern Regression with Applications by Myers, R.H. (1990)

39. Regression analysis of multiple source and multiple informant data from complex survey samples by Nicholas J. Horton and Garrett M. Fitzmaurice. Department of Mathematics; Smith College; College Lane; Northampton; MA 01063; U.S.A. Department of Biostatistics; Harvard School of Public Health; Boston; MA; U.S.A.

40. Introduction to Linear Regression Analysis. 2nd edition by D.C. Montgomery, E.A. Peck (1992)

41. A Tutorial on Principle Component Analysis by Lindsay I Smith.

42. Lophaven S.N, Nielson H.B, Sondergaard J, "Dace, a Matlab Kriging Toolbox", Version2.0, 2002.

Appendices

Appendix A: Mat lab codes

1. Matlab code for regression

```
clear all;
load data1.txt
figure; plot(data1(:,1),data1(:,2))
load data2.txt
hold on;
plot(data2(:,1),data2(:,2),'r')
xdata=data1(:,1);
ydata=data1(:,2);
% regression using least squares
order = 15;
xdata;
h=[];
for i = 0:order,
   h = [h xdata.^i];
end
thetacap = inv(h'*h)*h'*ydata;
lim(1)=min(xdata); lim(2)=max(xdata);
p = [lim(1):0.01:lim(2)]';
%h1 = [ones(length(p),1) p];
```

```
h1=[];
for i = 0:order,
   h1 = [h1 p.^i];
end
y3 = h1 * thetacap;
figure; plot(xdata,ydata,'r','linewidth',1);
hold on; plot(p,y3,'b','linewidth',0.1);
```

2. Matlab code using 20 principal components:

```
load data1.txt
data1=data1';
[U,S,V] = svd(data1);
d=S(1:20,1:20);
phi=U(1:80,1:20);
Vtrans=V';
Eta=Vtrans(1:20,1:1202);
y=phi*d;
load data2.txt
data2=data2;
load data2mod.txt
xnew=data2mod(81:90,1:13);
load data1mod.txt
data1mod=data1mod';
```

```
theta = [repmat(5,1,13)];

lob = [repmat(1e-1,1,13)];

upb = [repmat(20,1,13)];

[dmodel,perf] = dacefit(data2,y,@regpoly0 , @corrgauss, theta, lob ,upb);

[YX MSE1] = predictor(data2, dmodel);

 [Ypred MSE2] = predictor(xnew, dmodel);

Zpred=Ypred*Eta;

%Ztest=ytrue*Eta1;

Ztrue=data1mod(81:90,1:1202);

Error=(Zpred-Ztrue);

MSE20 =1/length(Zpred(1,:))*sum(Error'.^2);

 for i=1:10

    figure,

    plot(Zpred(i,:));

    hold on

    plot(Ztrue(i,:),'r');

    hold off

end
```

3. Matlab code using 80 principal components:

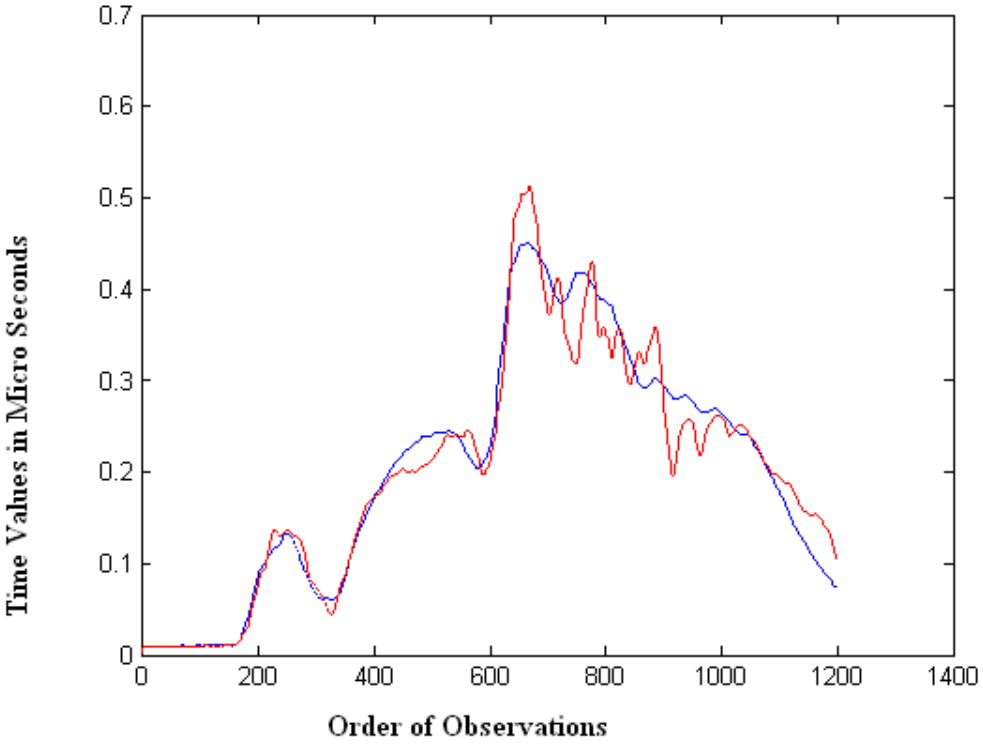
```
load data1.txt

data1=data1';

[U,S,V] = svd(data1);

d=S(1:80,1:80);

phi=U(1:80,1:80);

Vtrans=V';

Eta=Vtrans(1:80,1:1202);

y=phi*d;

load data2.txt

data2=data2;

load data2mod.txt

xnew=data2mod(81:90,1:13);

load data1mod.txt

data1mod=data1mod';

theta = [repmat(5,1,13)];

lob = [repmat(1e-1,1,13)];

upb = [repmat(20,1,13)];

[dmodel,perf] = dacefit(data2,y,@regpoly0 , @corrgauss, theta, lob ,upb);

[YX MSE1] = predictor(data2, dmodel);

 [Ypred MSE2] = predictor(xnew, dmodel);

Zpred=Ypred*Eta;

Ztrue=data1mod(81:90,1:1202);
```

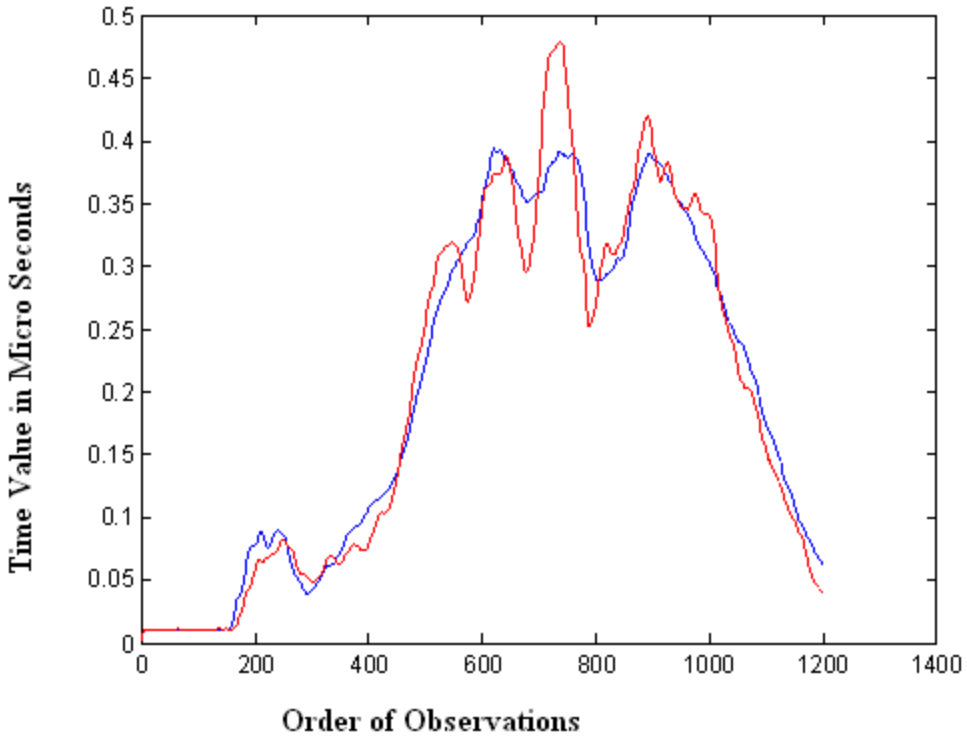
```
Error=(Zpred-Ztrue);
MSE80 = 1/length(Zpred(1,:))*sum(Error'.^2);
for i=1:10
   figure,
plot(Zpred(i,:));
hold on
plot(Ztrue(i,:),'.');
hold off
end
```

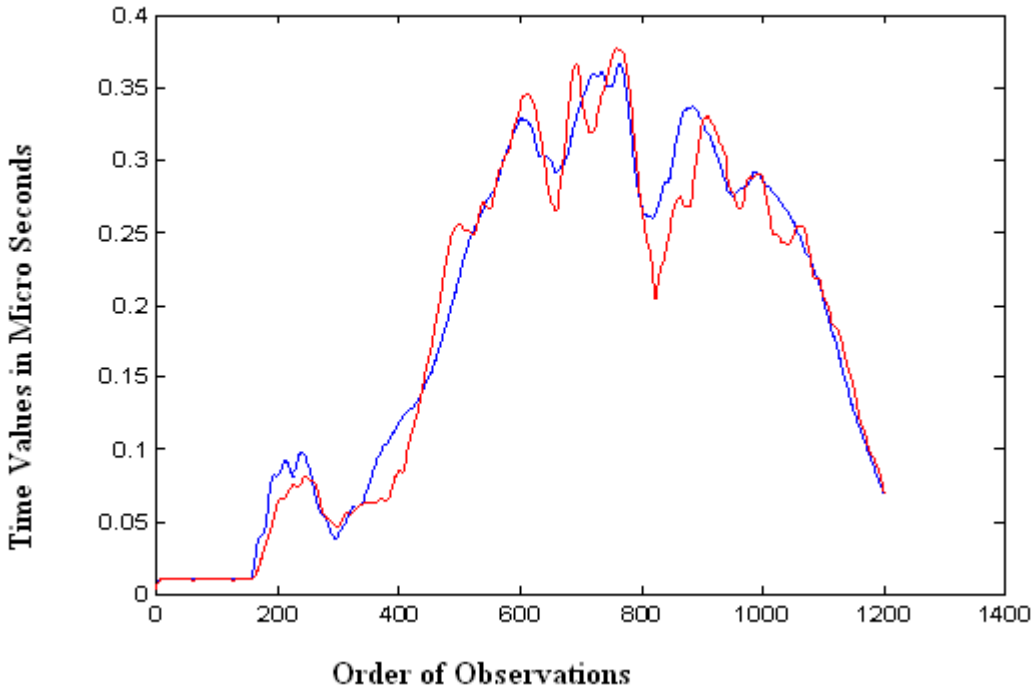
Appendix B:

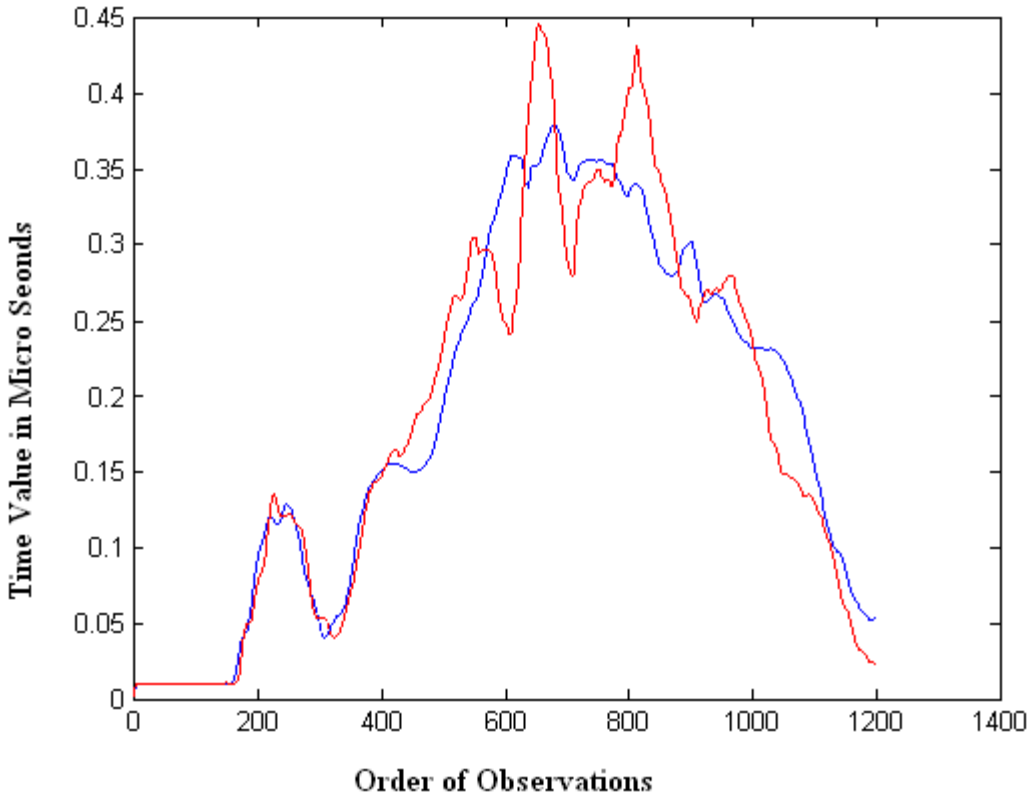
Following are the plots between predicted data and actual data for each row after PCA.

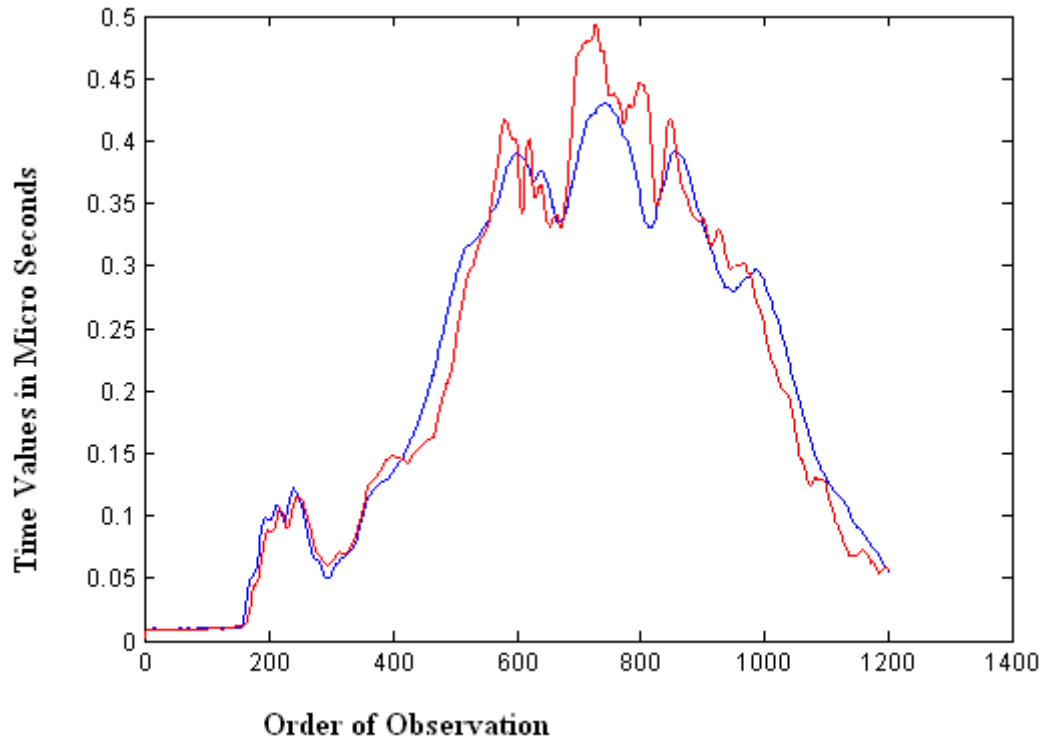
Zpred Vs Ztrue: Blue color represents Zpred and red color represents Ztrue.

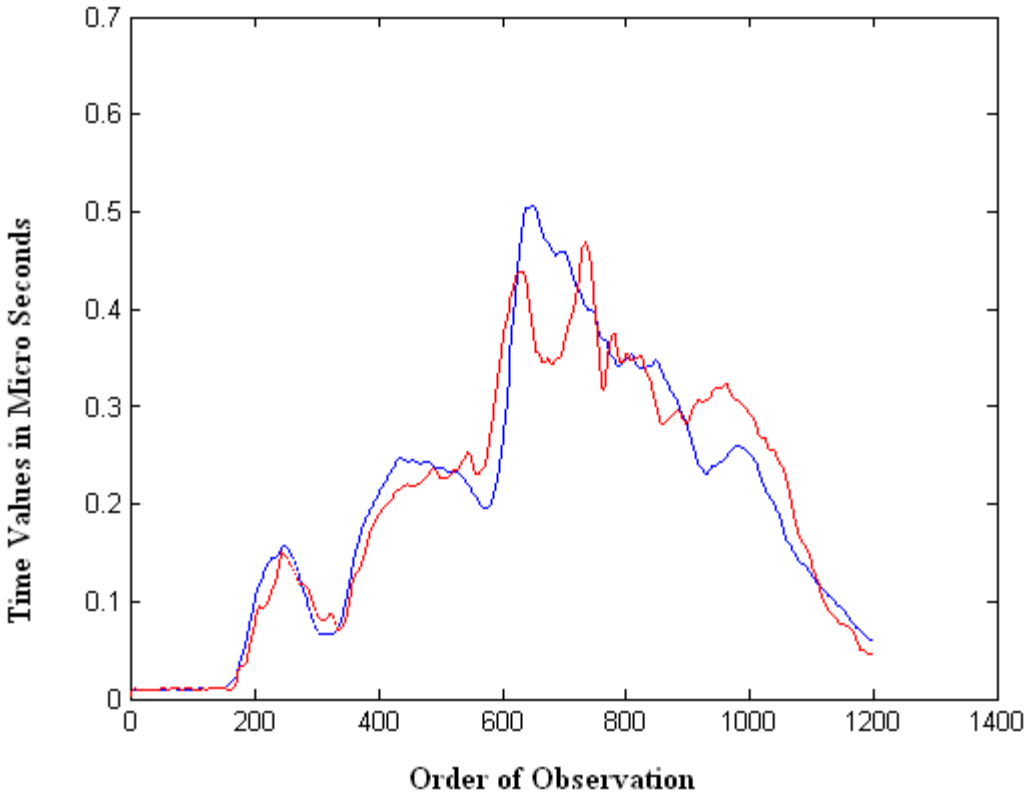


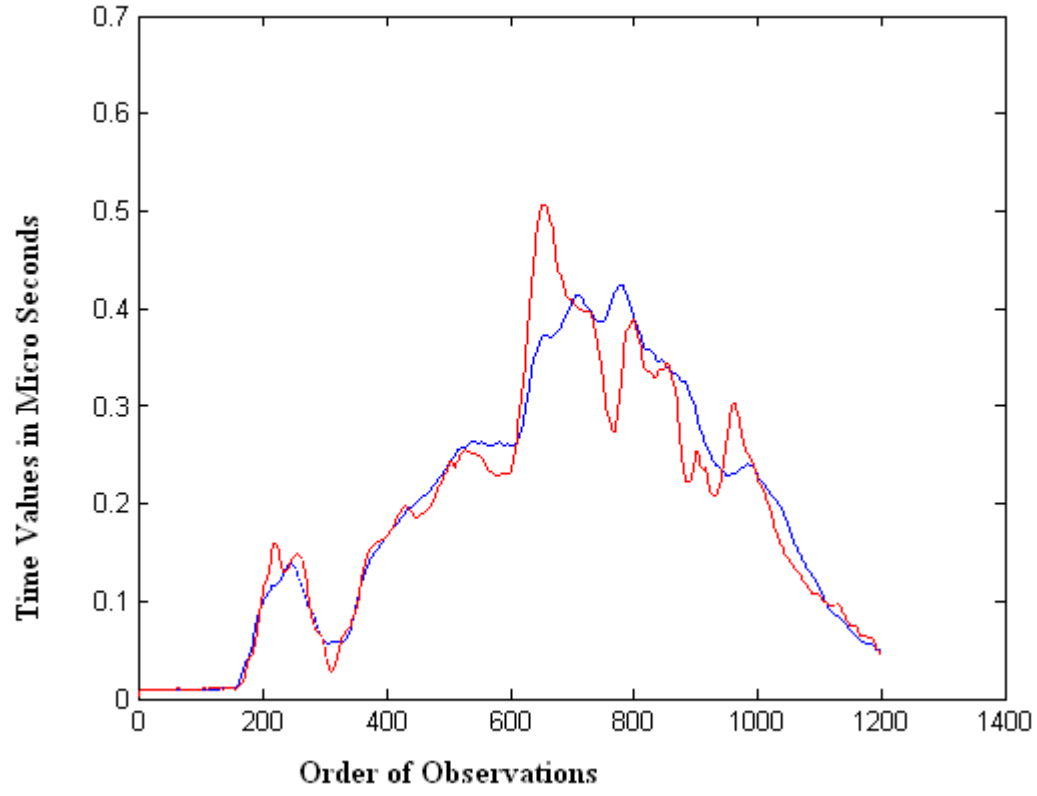

Time Value in Micro Seconds
Order of Observations
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0
200
400
600
800
1000
1200
1400

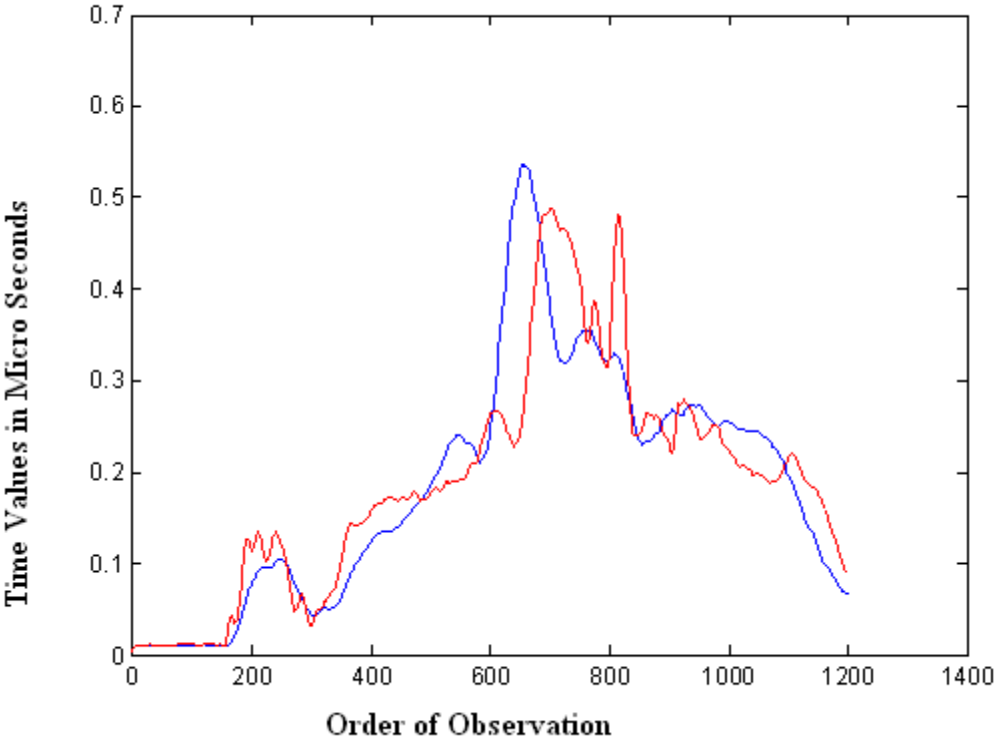

Time Values in Micro Seconds
Order of Observations
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0
200
400
600
800
1000
1200
1400

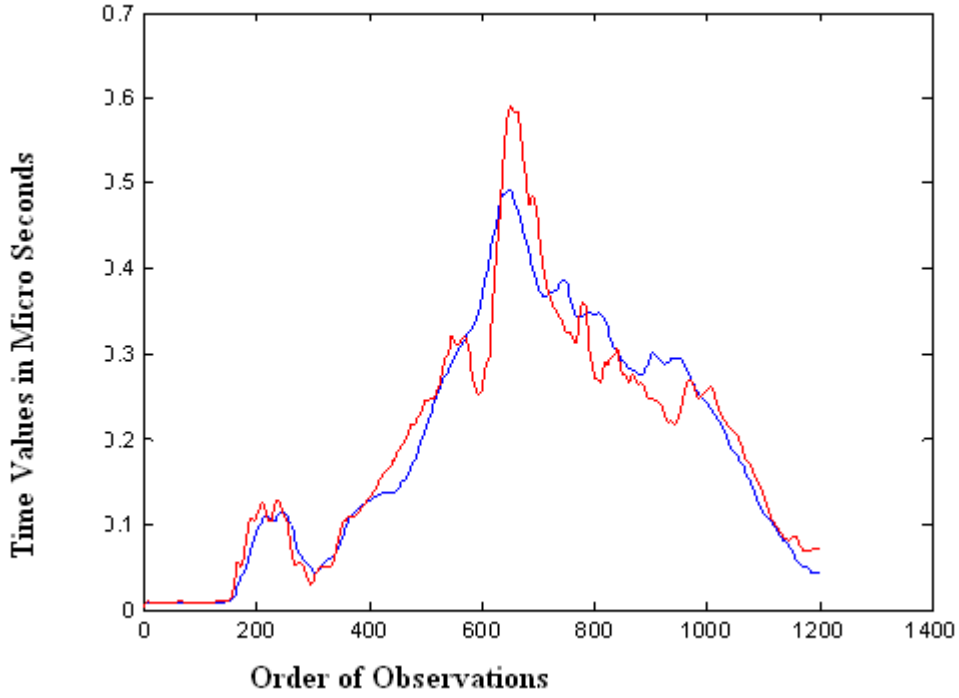

Time Value in Micro Seonds
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
200
400
600
800
1000
1200
1400
Order of Observations


Time Values in Micro Seconds
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
200
400
600
800
1000
1200
1400
Order of Observation


Time Values in Micro Seconds
Order of Observation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
200
400
600
800
1000
1200
1400


Time Values in Micro Seconds
Order of Observations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
200
400
600
800
1000
1200
1400


Time Values in Micro Seconds
Order of Observation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400


Time Values in Micro Seconds
Order of Observations
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400

Curriculum Vita

Ravi Lochan Kallur was born in Kurnool, India on April 9, 1982. The eldest son of Krishna Murthy Kallur and Sujatha Kallur, he graduated from G. Pulla Reddy Engineering College, Kurnool (Affiliated to Sri Krishna Devaraya University) in the spring of 2004. He entered the University Of Texas at El Paso in spring 2006 to pursue his Master of Science in Manufacturing Engineering. While at the university, he worked as Teaching Assistant for Industrial Engineering Department and Research Assistant at Intelligent Systems Engineering Laboratory (ISEL) in the Industrial Department during his stay at University of Texas at El Paso.