CURVE FITTING WITH BOOTSTRAP ERROR ANALYSIS AND ITS APPLICATION ON TWO-DIMENSIONAL DATA

by

NUR SOFFIAH BINTI SAHUBAR ALI

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

August 2019

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and Most Merciful. Alhamdulillah and all praise to Allah and peace be upon His beloved prophet MUHAMMAD s.a.w. I am grateful for all the strength that He gave me through His blessings to complete my research. The kind support, help and motivation from many individuals enhanced my thesis progress. The struggles and obstacles I faced while completing this thesis makes me stronger and determined. Taking this opportunity, I would like to show my sincere appreciation of gratitude to all of them.

There are no words that could express my gratitude enough to my supervisors. Without their supports, I would not be able to make progress throughout my research. Firstly, my deepest gratitude to my supervisor, Dr. Ahmad Lutfi Amri bin Ramli,who spent countless hours reading the thesis, making many corrections and valuable suggestions. Constant monitoring and guidance with support from him gave me determination and encourage me to perform my best throughout the thesis since my master studies. I will be forever indebted to him for his help, enthusiasm and encouragement. Also in this PhD journey, my lovely co-supervisor, Dr Nuzlinda Abdul Rahman who gave many suggestions and corrected my mistake in thesis and guidance as well in improving the thesis.

Then, special gratitude to my lovely friends, Adila Aida Azahar, Insathe Mohd Ali, Zuliana Ridzwan, Nurul Ainina Redwan and Irfan Shah whom encourage and helped me to go through the PhD journey and supported me to achieve the fulfillment of the thesis. The advices and motivation given by them make me more confident of myself to complete this thesis. My thanks and sincere gratitude also goes to my colleagues

Nur Fatihah Fauzi, Asyraf Alvez, Yushalify Misro, Shareduwan Kasihmuddin, Nur Ardiana Amirsom, Nur Ezzah Fauzi and those who have involved directly and indirectly towards the completion of this thesis.

I would also like to thank my parents, brother and sister for giving me moral support and encouragement fully to finish up this thesis. They gave me strength by just being there always.

Not forgetting, Miss Hartini whom always give helping hand and much love to all our lab mates when encountering technical problem and also all the people who has contributed to the accomplishments of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	Two-Dimensional
3D	Three-Dimensional
CAGD	Computer Aided Geometric Design
LSF	Least Square Fitting
SSE	Sum Square Error
RMSE	Root Mean Square Error
ANFIS	Artificial Neuro Fuzzy Inference System
RBF	Radial Basis Function
TPS	Thin Plate Spline

PENYUAIAN LENGKUNG DENGAN ANALISIS RALAT CANGKUK BUT DAN APLIKASINYA PADA DATA DUA-DIMENSI

ABSTRAK

Dalam penyuaian data, para penyelidik menggunakan pelbagai cara untuk menentukan kualiti penyuaian. Visualisasi sesuatu imej adalah penting bagi memerhatikan kelakuan sesuatu data yang diperoleh. Masalah dalam menilai ketepatan hasil yang diperoleh melalui pemerhatian visual kerap dialami oleh penyelidik dalam mengendalikan data yang tercemar seperti data hingar, data hilang dan pencilan. Dalam kajian ini, penyelidikan telah dijalankan untuk menangani data hingar dan data hilang menggunakan penyuaian kuasa dua terkecil (LSF). Dalam LSF, peningkatan darjah penyuaian dapat mengurangkan ralat latihan tetapi ia boleh menyebabkan penyuaian yang visualnya tidak menyenangkan dan tidak diingini. Oleh yang demikian, kebergantungan kepada ralat latihan sebagai penilaian visual mungkin agak sukar dalam sesetengah kes. Kaedah pemurataan model, iaitu kaedah cangkuk but yang digunakan untuk mengira ralat dan menganggar ralat dengan lebih baik dalam LSF dibincangkan pada data 2-dimensi. Satu set data simulasi yang dijana secara rawak dengan menggunakan data yang ditambah sedikit hingar diaplikasikan untuk menganalisis penyuaian. Dalam kajian ini, perbandingan antara ralat latihan dan ralat cangkuk but pada darjah LSF yang berbeza menggunakan data terpilih dilakukan. Kebolehpercayaan kaedah cangkuk but dengan membandingkan ralat data sahih sepadan dengan penilaian subjektif iaitu melalui pemerhatian ditunjukkan. Dalam kajian ini, titik n digunakan secara rawak daripada titik data hingar yang disimulasi bagi melaksanakan fungsi pengang-

garan dengan fungsi asas jejari (RBF). Penilaian ralat menggunakan RBF dengan splin plat nipis daripada nilai parameter yang berbeza, c diperolehi. Pengesahan parameter pilihan optimum, c_{opt} dilakukan dengan perwakilan visual yang dihasilkan. Kemudiannya pendekatan menggunakan kaedah cangkuk but dalam LSF dan RBF diaplikasikan pada data sebenar pesakit yang mengalami kecederaan pada bahagian tengkorak dan data penggerudian terowong. Namun, keputusan tidak dapat disimpulkan dengan baik untuk data penggerudian terowong kerana ia menghasilkan dapatan seragam pada ralat cangkuk but yang besar dan tidak menunjukkan hubung kait dengan parameter lain yang terlibat. Sebaliknya, kajian ini telah memberi keputusan yang baik dalam pembinaan semula kraniofasial untuk mendapatkan lengkung terbaik bagi kawasan retakan tengkorak yang hilang. Ini telah memberikan sumbangan penting yang boleh dijadikan sebagai kaedah alternatif dalam bidang perubatan terutamanya pada pembinaan semula kraniofasial.

CURVE FITTING WITH BOOTSTRAP ERROR ANALYSIS AND ITS APPLICATION ON TWO-DIMENSIONAL DATA

ABSTRACT

In data fitting, researchers use various methods to determine the quality of a fitting. Visualization of images is crucial in observing the behavior of data obtained. The problem in judging the accuracy of a result obtained through visual observation are commonly faced by researchers when handling contaminated data such as noisy data, missing data and outliers. In this research, study has been conducted to deal with those noisy data and missing data using least square fitting (LSF). In LSF, increasing the degree of fitting reduces the training error but may lead to visually unpleasant and undesirable fitting. Therefore, reliability on training errors as a visual evaluation may be quite difficult in some cases. A model averaging method, namely bootstrap method that is used to compute error and better estimate the error in LSF is discussed on 2-dimensional data. A set of randomly generated simulated data with some added noise is applied to analyze the fitting. In this study, comparisons between the training error and bootstrap error for different degree of LSF on selected data are performed. Demonstration on the reliability of the bootstrap method by comparing to a ground truth error corresponding to a subjective assessment which is via observation is shown. In this research, n points are randomly used from the simulated noisy data points to implement an approximation function of Radial Basis Function (RBF). Error evaluation using RBF with thin plate spline (TPS) of different parameter value, c is obtained. Verification of optimum chosen parameter, c_{opt} is performed with the resulting visual

representation. Then, the approach is applied to real life data of a head injured patient on the skull and tunnel drilling data using bootstrap method in LSF and RBF. However, the results seem to be inconclusive for drilling data because it produces a uniform findings of large bootstrap error and did not show a correlation with other parameters that are involved. On the contrary, this study has performed well on craniofacial reconstruction to estimate the best fit curve for the missing part of fractured region of a skull. This provides a significant contribution that could be an alternative approach in medical application especially on craniofacial reconstruction.

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Applied mathematics field which deploy an algorithm to enhance the problems involving computer graphics prospect that specifically deals with curve and surfaces is known as computer-aided geometric design (CAGD). Images are handled in the form of small dots and pixels in computer graphics. The generation of curves and surfaces are denoted as curve and surface modeling in CAGD. Designers show their contributions with the aid of CAGD, corresponding to the user interface and CAGD tools in displaying a graphical finding. In this context, user interface refers to the interactions between a computer system and a user. Nevertheless, from the previous publication and review on CAGD, several applications have been expanded rapidly into various aspects of our real life including industrial applications, pharmaceutical design, and animation. Computer technology can enhance data analysis and increase our understanding of various scientific, economic, engineering and social phenomena. For detailed study on curve and surface fitting in CAGD, one may refer to Lancaster and Salkauskas (1986), Yamaguchi (2012) and Farin (2014).

In terms of visualization, Rockwood and Chambers (1996) mentioned that the design cycle is minimized via computer where the process of design alteration and tool production are simplified. In CAGD, visualization is an essential aspect to be considered. Ideally, the term visualization in computer graphics is prioritized to attain and interpret the problem of study faced by researchers for scientific data in 2-dimension

(2D) and 3-dimension (3D). In the late 1970's and beginning of 1980's, computer visualization in computer graphics and applied sciences enhanced the importance of computer graphics technology. The visualization output produced through computer graphics will provide a better understanding for the problem study. Issues related to scientific visualization have been addressed and discussed by visualization researchers, Defanti and Brown (1991) and Johnson (2004). Moreover, researchers present their ideas through programming languages in computer. C or C++ is usually used to generate algorithms to evaluate the performance of the methods. Besides that, the graphical results are also presented using programming softwares such as Mathematica and Matlab by recent researchers. Both programming languages benefits the researchers as they are more user-friendly and effective in displaying the graphical form. The results of this research are performed and generated using the Mathematica and Matlab software to compute the calculations involved as well as to display graphical results.

In our real life situations, naturally visualized images through observation using human naked eyes seem to look pleasing as well as perfectly smooth in appearance. However, most images or data that are obtained from sources such as scanners, telescope and other devices might be contaminated and disturbed. This is because the data might undergo various disturbances throughout the process of occupying. For instance, inaccurate data collections commonly occur due to the level of precision in instrumental set up or a human error such as errors that occur during the recording process of the data. These sorts of data might produce subjectively and quantitatively undesirable results which can affect the computation process. Nevertheless, from our point of view, these images or data are considered to have zero error. Shape and feature approximation via mathematical representative are also studied in CAGD. Numerical analysis and geometry and computer science are blended together in the portrayal of data visualization. Its procedures change gigantic measures of information into diagrams and figures. According to Ahmad et al. (2014), from the aspect of shape designing, the main purpose of the visualized curve illustration in research is to get a nice and "visually pleasing" curve. Similarly, this applies to surfaces. However, the visualization itself is not enough to value the presentation of study without an error analysis. As such, a review of the error is conducted to evaluate the quality of the result obtained.

A curve can be represented by a series of polynomials. Parametric curves are curves having a parametric form. Particularly in this research, attention will be given in considering 2D curves. This research will include the assessment of obtaining the best fit curve. A best fit curve refers to a curve that best fits a series of data points that correspond to the constraints. As far as concerned, interpolation and approximation methods are well known and broadly used in the research field. In practice, measured data tend to face the presence of noise as well as missing data issues. Estimation of a value within two known values in a sequence is known as interpolation. For instance, interpolation used in the prediction of unknown values for rainfall, noise levels and many others. The term approximation is referred to as an estimation to predict the behavior and pattern of the data. Normally, the approximation method works well in the performance of industrial production system. The question arise to determine whether interpolation or approximation should be carried out depends on the problems. Thus, the researchers would need to make the right decision in applying interpolation or approximation scheme to study their problems. In some circumstances, researchers will tend to face overfitting problems in handling the noise during the modeling process. Overfitting problems normally happens when the model fits the data too well in the presence of contaminated data. Inversely, when the model does not fit the data well and unable to predict the behaviour and trend of data, the model is said to be under-fitting. The two patterns of fitting will result in poor predictions of new data sets. Hence, when dealing with noisy data, an approximation is probably more appropriate. According to Feng and Zhang (2013), interpolation by Radial Basis Function (RBF), interpolation by splines and the least square approximation are the most commonly used fitting methods.

Noise is known as the disturbances or more precisely the contamination present when dealing with images, specifically when collecting real-life data. Teck et al. (2008) discussed a study revealing the noise, particularly concerned with Gaussian noise in mechanical engineering. Gaussian noise is referred to as statistical noise where Gaussian noise has a probability distribution function that is equivalent to that of the normal distribution. Teck et al. (2008) applied algorithm namely median filters and its variants in cleaning the noise that appears in drawing images. However in this thesis, attention is given on evaluating the best fit curve in the presence of noise by analyzing the error which affects the computation process. Observing the idea of handling problems in the presence of noise during fitting is one of the elements discussed in this thesis. Mooney and Swift (1999) stated that, in the case of data points having a huge number of random fluctuation or noise, it is not suitable to apply an interpolation method. This is because the data points interpolate yielding an unpleasant surface with remarkable errors. Ramli and Ivrissimtzis (2009) mentioned that in computer-aided geometric modeling application, noise from the raw data obtained via physical measurements from optical devices such as laser and 3D scanners should be elucidated. This is because these devices will produce point cloud which is referred to as a set of data points in space and needed to be processed further.

Missing data issue during data collection is also one of the matters that need to be considered. A number of researchers discussing the issues of missing data and related methods in handling this issue are presented in literatures, includes Little (1992), Allison (2001), Van Buuren (2012) and Schlueter and Harris (2006). It is important to distinguish the types of missing data in one encounter. For instance, in the medical field involving patient dataset such as for the skull data. In this case, missing data can be grouped into three categories, such as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The focus in dealing with the missing data in real life data using Least Square Fitting (LSF) and RBF with a statistical approach is concerned in this research. This real life data specifically considers a problem of the fractured skull of a patient with a head injury.

In real life problems, most simulations involving complex phenomena are prone to have the error. To select the best model representation, one normally decides based on the smallest error obtained. Various methods are used by researchers to evaluate the error in predicting the accuracy and quality of a model. It can be seen that the field of statistics is extremely vital in acquiring information. A vast number of application used statistics field in interpreting and making evaluations from the gained data. There are numerous methods for handling problem related to error and fitting in statistics such as goodness-of-fit test, Kolmogorov-Smirnov test, and others. In this research, a statistical method, namely, bootstrap method has been deployed to evaluate the accuracy of fitting. Approximation methods are performed to analyze the fitting using this statistical method. Training error, sum square error (SSE), root mean square error (RMSE) and bootstrap error are computed and the results are compared. From several reviewed studies, insights in employing the attributes of the statistical method accomplished needs to be much more examined in order to demonstrate the knowledge of the applied method. This would be valuable in understanding their multiple roles of the used method in the mechanism of fitting and error evaluation. It has also been suggested through empirical evidence that the bootstrap method normally yields well in performance (Bickel and Freedman, 1981; Daggett and Freedman, 1984). In addition, theoretical justification of the valuability of bootstrap performance has also been attempted in several studies to evaluate the accuracy of the method in diverse prospects (Singh, 1981; Beran, 1982; Babu and Singh, 1984).

Additionally, the aforementioned statistical method is also applied in evaluating the selection of the optimum parameter to produce a best fit curve using thin plate spline (TPS). TPS is being considered in this research with the motivation obtained from Liew (2017) to evaluate an optimum parameter. Specifically, the visualization of fitting using different parameter values, c are demonstrated. Here, the errors are computed using the statistical method to select the best fitting with respect to the selected parameter value, c to emphasize the importance of visualization and error evaluation in order to determine the quality of fitting.

In statistics, the research on error has been discussed in previous literature. However, due to a different types of data such as binary data and high dimensional data (4 or more dimension), visualizations may not always be possible. In geometric modeling, computer vision, CAGD, researchers have a tendency to rely on training error. Therefore, one of the primary aim is to depict graphical visualization using the method to observe the quality of a fitting.

The exploratory outcomes on visualization and computation of error will be mainly assessed in this thesis. The integration in the field of study between CAGD and statistics enable the researcher to examine the accuracy and quality of curve reconstruction or fitting much simpler. The demonstration of visualization of fitting as well as error evaluation is concerned throughout the study producing a smooth curve and designing a surface.

1.2 Motivation of Study

In this research, two research papers have inspired our problem of study. Firstly, the doctoral work by Ramli (2012) that investigated research on surface reconstruction using the bootstrap method. This study evaluates upon obtaining a visually high-quality model and retaining the exactness towards the original model. A statistical method is deployed in achieving the aim of the study in this research. This statement was further evaluated by the second work conducted in another doctoral study by Liew (2017) recently. Liew (2017) discussed two types of approximation method, namely, B-spline and TPS methods. Similarly, the study applied the bootstrap method on surface reconstruction approach to better estimate error. As a whole, both works concluded that the statistical method works well as a better estimate in surface reconstruction.

As far as concerned, RBFs are tremendously utilized in various fields of study, including economics, computer graphics, and data processing (Chen et al., 2014). Both interpolation and approximation methods do exist in RBFs. Particularly, in this thesis, the intention will be on an approximation method using TPS specifically. This is due to the reason that the aim is to study contaminated data in the presence of noise and missing data. Besides, the method is demonstrated functionally using simulated data to analyze the effectiveness of the method in searching an optimum parameter value for fitting.

The motivations gained from previous work inspired us to analyze 2D studies in order to discover the key attributes when applying the statistical method in the presence of noise and missing data. Therefore, by lowering the dimension setting into 2D prospects, it is possible to observe the graphical form better while minimizing the cost of computation and speeding the efficiency. Apart from that, visual observation through simulations is evaluated in this research to validate the demonstration in the presence of actual data. Thus, motivated by these work, the research is further explored in 2D using real life data, namely craniofacial data and drilling data by focussing on visualization and error analysis. This is because visualization plays an important role in CAGD.

1.3 Problem Statement

Over the past few years, a number of studies have been conducted to handle noisy data as well as missing data. Besides that, curve fitting is obtained in the presence of contaminated data using various methods to analyze a study. In order to evaluate the accuracy and quality of a fitting using contaminated data, there should be a method applied to evaluate the behavior of data to select the best fit curve. Training error seems to be used by several researchers to study the accuracy of fitting of data. However, it is not recommended to rely only on training error as it could not produce the best fit curve. Thus, the bootstrap error is evaluated to compare the fitting in this research using LSF. RBF is one of the methods used to obtain an optimum parameter value, c in this research for fitting. Only few studies apply RBF or specifically TPS to demonstrate the behavior of a data in a fitting in the presence of error. Hence, the issues in justifying the preciseness dealing with noisy data and missing data are still lacking and this gap should be dealt with. Therefore, this research will contribute to a study in demonstrating and understanding the key attributes of the bootstrap method using two different functions namely, LSF and RBF in order to choose the best fit curve.

1.4 Objective of Study

The main research objectives are:

1. To demonstrate the differences and accuracy of using bootstrap error and training error in the context of 2D data on noisy data for LSF with different degree of fitting (Chapter 4).
2. To demonstrate the difference and accuracy of using bootstrap error and training error in the context of 2D data on noisy data for RBF or specifically TPS with a free parameter that can adjust the shape of the fitting (Chapter 5).
3. To evaluate an approach in choosing the best fitting for the respective functions on a real life data for LSF, by applying the bootstrap method, demonstrating the strength and limitation. This is done on a craniofacial image to recover missing data and on a drilling data to study the relationship between drilling speed and other parameters such as drilling depth of tunnel, presence of sand and presence of gravel using LSF by finding the optimal degree (Chapter 6).

4. To evaluate an approach in choosing the best fitting for the respective functions on a real life data for RBF or specifically TPS, by applying bootstrap method, demonstrating the strength and limitation. This is done on a craniofacial image to recover missing data using RBF by finding the optimum free parameter value (Chapter 6).

By fulfilling the targeted objective of the study, better knowledge on understanding the presence of noise and missing data by applying the mechanism of the statistical method will be gained.

1.5 Research Methodology

The methodology approaches established in this research are as follows:

1. Problem Simulation.

   The analysis has been carried out to generate a simulation using 2D data for error evaluations. Two types of functions, specifically the LSF and RBF are used to illustrate the problem. The reliability of training error, RMSE, SSE are discussed and compared with boostrap method by finding the optimal degree.

2. Searching Optimum Parameter Value, c.

   An optimum parameter value, c is evaluated and the best fit curve is observed and depicted using the RBF. Limitation on applying the bootstrap method is highlighted. A separation method is proposed to deal with specific data.

3. Application on Real Life Data.

   The analysis is employed on a craniofacial reconstruction data using the functions LSF and RBF for the reconstruction of the missing fractured skull of a patient. It is validated using a non-fractured skull of a patient. Drilling data is also investigated to evaluate the correlation between drilling speed and other parameters such as drilling depth of tunnel, presence of sand and presence of gravel.

The workflow of this research is depicted in Figure 1.1.

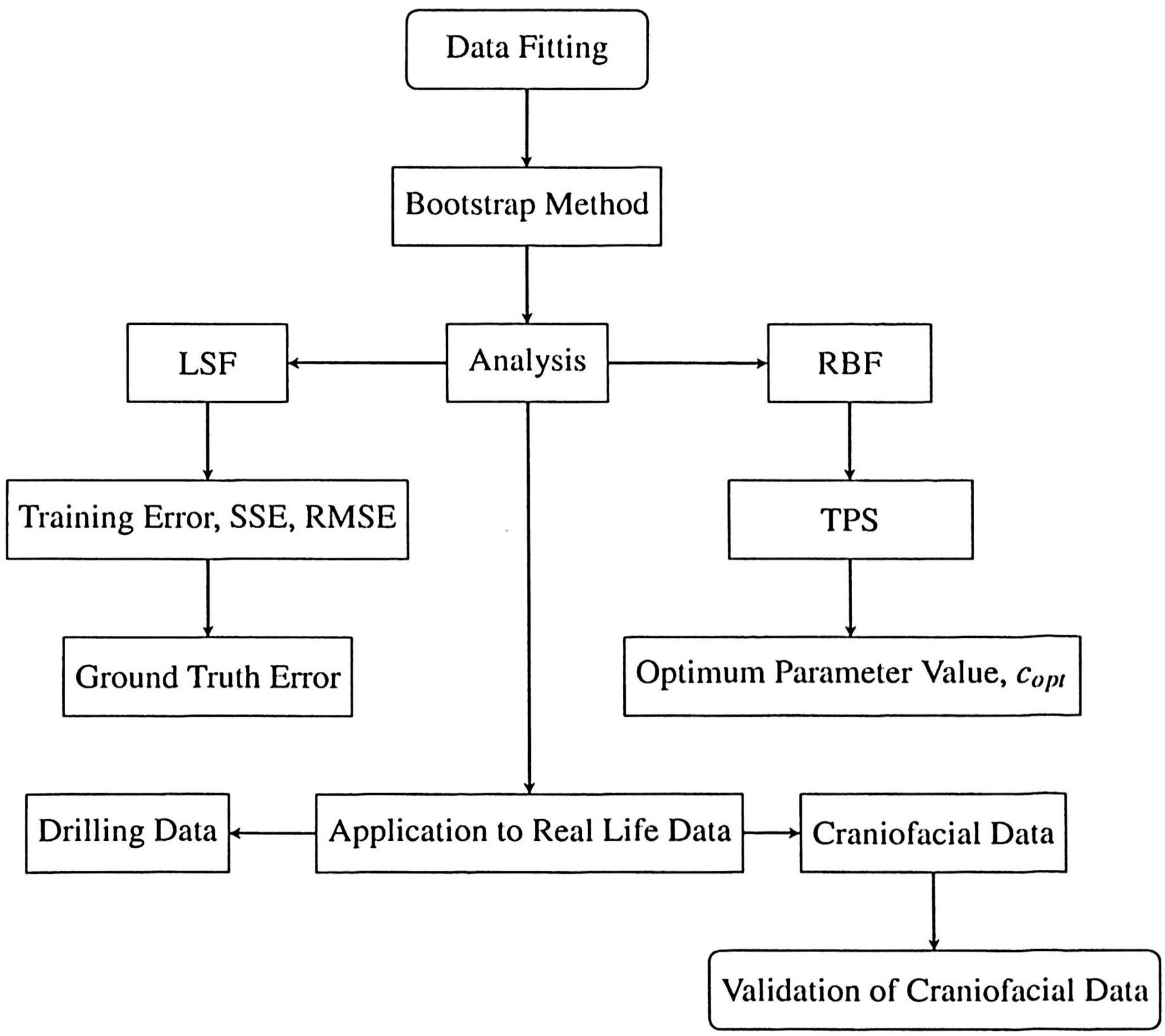

Figure 1.1: Flowchart of methodology

1.6 Summary and Organization of the Remaining Chapter

Altogether, the thesis is presented in seven chapters. As been elaborated and detailed, Chapter 1 introduces an overall view of the thesis flow and the ultimate objectives. The thesis structure of the rest of the chapters are presented. The upcoming Chapter 2 covers a survey on the literature that enhances our research scope. Most of the notable works of literature mentioned in this chapter bring its own significance in the research field.

The following Chapter 3 covers the basic concept and method used throughout the research to solve the problem of study. Here a brief discussion on the method used to evaluate the error is detailed. This will also include the different types of error such as training error, RMSE, SSE and bootstrap error which will be discussed in this study.

Chapter 4 emphasizes the analysis of LSF on different degrees via graphical interpretation as well as to select the best fitting with the bootstrap method used.

Next, Chapter 5 presents RBF implementation with the presence of noise in curve fitting. The function generated will be shown using TPS to choose the optimum parameter which yields the best fitting. The limitation issues faced using missing data are also discussed.

In Chapter 6, application of the bootstrap method using two types of real life data will be illustrated to show the reliability of the method used. In this chapter, the data of a patient with a head injured skull will be used for the analysis. Besides that, probe drilling method data is also discussed.

The final chapter will provide a conclusion of this research. In this Chapter 7, the significance of the study, limitation faced and future studies will be mentioned briefly.

CHAPTER 2

LITERATURE REVIEW

In this chapter, recent and previous research that is in correlation with this study will be discussed. The ideas of study are presented in accordance with the research topic to understand the insights of the thesis.

2.1 Noise and Missing Data

The presence of noise and missing data are usually the problem faced by researchers when handling data tabulation or evaluation. In the existence of noisy data, the overfitting problem encountered during the modeling process due to the minimization of the distance between the original data points and the estimated data points of the model. In this case, the noisy data from the modeled data affect the computation process. As concerned, the real life data are disturbed by noise. Various techniques have been conducted in the literature in handling noise problem. The noise removal and noise reduction techniques are used to tackle those problems in response to various kinds of noise in image data. As described, the techniques are evaluated to enhance the performances of data during the tabulation of results and applications. In literature, methods such as removal of noise using median filter (Kohli and Kaur, 2015; Hambal et al., 2017) linear filter and adaptive filter are the filtering techniques used to ensure the quality of image. These aforementioned methods are able to handle noise in image data. However, the framework of this research will be towards the fitting of a curve in the presence of noise.

A number of researchers encountered missing data issues during data collection. Missing data or information is a very wide range of topic to be discussed. Missing data during data entry and experimental data analysis recording can result in invalid conclusions. Thus, this will complicate the analysis of study. Prevention and method of handling the noise in images are considered in literature. Imputation, likelihood, and weighting are some of the statistical method addressed in recent years. For further information, one can refer Little (1992), Chen et al. (2000), Allison (2001) and Van Buuren (2012). However, this thesis is mainly emphasizing on the attributes of bootstrap error in LSF and RBF in dealing with noisy and missing data.

2.2 Fitting Technique

In geometric modeling, computer graphics and images processing computer reconstruction from a 2D prospect is crucial in visualizing the curve output. Curve fitting is often applied in the reconstruction of a curve to produce a pleasant curve. There are several algorithms for curve reconstructions presented in the literature. Mukhopadhyay and Das (2007) proposed an algorithm for 2D curves with sharp features in the presence of unorganized noise and outliers. In the field of solar radiation, curve fitting methods are deployed to analyze the result in correspondence with the error evaluation. This includes Jalil et al. (2018) and Pareek and Gidwani (2015). Curve fitting is evaluated by Farayola et al. (2017) to investigate the accuracy of degrees of polynomial and compare the results with the Artificial Neuro Fuzzy Inference System (ANFIS) results. Their findings indicate that using the polynomial curve fitting of the 6th order and the ANFIS methods could detect the maximum peak power point than the curve methods of the lower order.

There are various approaches to verify the quality of fitting. For certain applications, the best fit curve can be subjectively selected through visual observation. Nevertheless, it is difficult to select the best fit curve by just visually observing a curve in graphical form. On the other hand, relying on values such as training error, or error, as introduced above and mentioned in some literature, can also lead to a wrong selection. This is because overfitting or underfitting may occur, resulting in a low training error value but not producing the desired best fit curve. Mamoon and Rahman (2017) studied on the selection of the best fit probability distribution in rainfall frequency analysis for Qatar. The most frequently-used fitting methods are an interpolation by spline, interpolation by RBF and the least square approximation (Feng and Zhang, 2013). In data fitting, LSF is commonly used to fit parameters and function that fits data. Particularly, Richter (1995) has revealed the study on estimating the error in LSF. Many research has used LSF to fit a straight line and this includes York (1968) and Draper and Smith (1998).

2.3 Error

The average distance between the original training data and the approximation of the model is defined as the training error of the model. Interpolation of the data by passing through all the training data points yields a training error that equals to zero. Thus, low-quality models may result from the minimization of training error. According to Ramli and Ivrissimtzis (2009), training error is expected to monotonically decrease when the model complexity increases. They applied the bootstrap technique to estimate the test error for polynomial fittings of locally parametrized 3D point sets. One may refer for more considerable research related to data and error analysis in

Hughes and Hase (2010), Berendsen (2011) and Good and Hardin (2012). There are also several error analysis methods in literature such as cross validation method and K-fold cross validation.

On the other hand, ground truth error is defined as the difference between the original data before the addition of noise and estimated value. In literature, ground truth is subjectively inferred under various guises. Moreover, ground truth error should give a good portrayal of a fittings quality. Noise is added to simulate real-life behavior, in which the data is assumed to be contaminated by noise. As such, a good estimating function or a good fitting will produce the best model representation that is close to the original function before noise was added. In this study, further research was done to demonstrate the reliability of a fitting using simulated data with the aid of ground truth error in Chapter 4.

2.4 Bootstrap Method

Over the years, the nature of statistics is being applied in an immense number of various fields tremendously. Ideally, statistics can be declared as the methodology used when handling collection, interpretation and making inferences from data. By expressive it differently, our mathematics and scientist have utilized the statistics as the methodology to analyze and evaluate the gathered information. As a whole, all the process that handles the evaluation of the data performances has a major relation in statistics. The process of data performances is anticipated step by step from the collection, processing, interpretation, and presentation of data (Isotalo, 2001). There always exist some inaccuracies and mistakes occurs when measuring apparatus or displayed devices that yields an error (Berendsen, 2011). One may refer Rabinovich (2006) for detailed study on error analysis.

In this thesis, the reliability of the technique used to evaluate the fitting of simulated data as well application towards real-life data are discussed. The technique used in analyzing the error in this study is known as the bootstrap method. This model averaging method is a typical method used in statistics. Reviews on the related scope are presented using various studies. The bootstrap method that was introduced by Efron (1979) performs a simulation to increase the number of samples by replicating the sample in a random manner. Therefore, it is able to tackle the issue of a sample with small size effectively. Lately, diverse range of applications used bootstrap method to handle issues as it is present as a standard tools in some software packages in relevance to statistics (Zoubir and Boashash, 1998; Liu et al., 2012).

Ramli and Ivrissimtzis (2009) have discussed the efficiency of the bootstrap method in polynomial fitting for surface approximation. They computed bootstrap error estimation to choose the best fit curve in the polynomial fitting. A similar method was also used by Cabrera and Meer (1996) to estimate fitting of ellipses. Besides, Liguori et al. (2017) used the bootstrap method to estimate the uncertainty of traffic noise measurements.

Recently, Zhu and Kolassa (2018) have conducted research in assessing and comparing the accuracy of various bootstrap methods. They focussed on constructing confidence intervals and assessing the performance of bootstrap methods in two types of model, namely the Cox Proportional Hazard model and Accelerated Failure Time model to estimate the parameters. They found that for the situations in which the characteristics of the population distribution is not known, it is safe to apply the studentized bootstrap method for making confidence intervals for the mean. The studentized bootstrap is based on the bootstrap distribution of the statistic adjusted by its mean and standard error and is common in ecological data. Inversely, parametric intervals are preferred in a situation where the population distribution of data is known. This is because exact coverage of the population parameter is given.

Taking into account the attributes of the methods, it is applied in our research to demonstrate the key attributes as such using simulation and real-life data.

2.5 Radial Basis Function (Thin Plate Spline)

In the functional representation, the consideration of interpolation and extrapolation are essential. RBF method has been used widely by researchers as there exists a free shape parameter that attributes in simplicity and connects to the accuracy of the method. However, the choice of an optimal value of the shape parameter was carried out by trial and error or without solid verification by researchers in literature. Specification of the parameter value, c by a user is inherent rather than random selection. Several papers suggested the specified parameter values, c to a certain value with respect to their problem of study, includes Hardy (1971) and Franke (1982).

The implementation of RBF interpolation was originally introduced by Hardy (1971). Many researchers have conducted various approach to evaluate the optimum parameter, includes Rippa (1999), Fasshauer and Zhang (2007) and Scheuerer (2011). For functional approximations, the applications of their general purpose methods are significant. Dealing with the function to be approximated, one should determine whether the function depends on many variables or parameter. The functions ability to be defined with diverse data which are "scattered" in their domain are also discussed in detail by Buhmann (2003).

In statistics literature, cross-validation method has been used tremendously and the special case of leave-one-out cross-validation proposed by Rippa (1999) in applications with RBFs on scattered data interpolation yields the basis of the algorithm for searching an optimal value of shape parameter. Fasshauer and Zhang (2007) extended Rippa's approach to find an optimal number of iterations and the optimal shape parameter. This could be found by applying iterated approximate moving least square approximation. They discovered that this method can be implemented in setting iterated approximate moving least squares approximation of function value data and for RBF pseudo-spectral techniques for the solution of partial differential equations by modifying Rippa's algorithm to find the optimal shape parameter. Scheuerer (2011) conducted a method to obtain a good parameter value, c through an alternative technique, which can be interpreted as refinement of Rippa's algorithm for a cost function based on the Euclidean norm. They found that their algorithm compares favorably with cross validation in many cases when using the same test function as Rippa's.

Bayona et al. (2011) have applied the RBF method in computing the optimal value of shape parameter by focusing on the solution of partial differential equations. An algorithm that minimizes error is performed to compute the case on the optimal constant shape parameter value. Later on, they extended their research on optimal variable shape parameter value by allowing the changes of shape parameter for all the nodes by yielding an improvement in the accuracy of the numerical technique (Bayona et al., 2012). They found that for both structured and unstructured nodes, an accurate solution can be achieved using the optimal value of the constant shape parameter as compared to finite difference.

There are several types of RBF include Gaussian, multiquadric, inverse quadric, inverse multiquadric, polyharmonic spline and TPS. It is found that TPS is commonly used in literature for problem study. In the prospect of medical imaging, cranial implant on surface interpolation using TPS has been deployed by Carr et al. (1997). The application of cranial implant suits in their problem thus inspiring us to demonstrate the visualization on 2D by applying the bootstrap method to evaluate the performance of TPS on the missing region of a skull. In data fitting, Liew et al. (2017) applied the bootstrap method mainly bootstrap leave-one-out approach in searching the optimum value of the smoothing parameter for RBF surface reconstruction specifically TPS. Bootstrap leave-one-out is used to estimate and evaluate the test error from the bootstrap sets. Based on their research, they summarized that bootstrap leave-one-out error computed from a sample of data points can be applied to search for a better parameter value compared with training error.

Here, the research is extended to run model adequacy checking to validate and analyze the reliability of fitting using simulated data points in 2D. This study analyzes the training error and bootstrap error evaluation with different parameter values for RBF, specifically TPS on noisy data and missing data. Observation will be performed on curve fitting using different parameter values, c. From bootstrap error evaluation, an optimum parameter value, c will be obtained. The cases of different parameter values,

c will be displayed and how visual observation may not be reliable for verification will also be discussed.

2.6 Real Life Applications

The art of managing questionable phenomenon and occasions are also conducted in statistics. For example, the effectiveness of therapeutics medications, evaluation of ground condition in a tunnel and considerably more are being considered with the connectivity aid of statistics, includes Jalil et al. (2018), Pareek and Gidwani (2015), Farayola et al. (2017) and Hassan et al. (2016). Lately, statistics are utilized as a part of each field of science.

Most real-life measured data are prone to problems such as noise, missing data, redundancy, and outliers. Numerous studies have been conducted to deal with such problems. For instance, a statistical approach, the Bayesian model, was proposed by Qian et al. (2006) to reconstruct a 3D mesh model from a set of unorganized and noisy data points. The exploratory outcomes demonstrated that their method can be used to remove outliers, smooth noisy data, reconstruct mesh, and enhance features. A similar approach was applied by Jenke et al. (2006) to perform reconstruction of a piecewise-smooth surface. Furthermore, they also examined the algorithm using two types of data such as real world and synthetic. To manage the noise level, users are requested to input the noise when applying these two Bayesian models. The statistical method is adopted to handle issues related to the stochastic nature of noise. Besides, to handle noise estimation of 3D point sets, a Variational Bayesian method was suggested by Yoon et al. (2009). There is various type of noise that had been surveyed in Farooque and Rohankar (2013). They also discussed on the technique for denoising the color images. Noise is often assumed to be normally distributed, whereby the amount of noise is determined by variance and by setting the mean as zero. Despite in several cases, Gaussian distribution, Poisson distribution are also been employed by researchers. Apart from that, for surface reconstruction, Ramli and Ivrissimtzis (2009) proposed bootstrap method, to approximate the test error of the model, which can be directly used to compare models and applicable in real life data.

Recently, RBF has been implied in analyzing soil characteristics by hydrologist and researchers. They have illustrated that EasyRBF is capable to obtain much accurate results of infilling the soil dataset with bigger-scale and continuous data missing Shao et al. (2017). This EasyRBF is an approach of a novel missing value infilling for soil dataset.

2.6.1 Craniofacial Reconstruction

In this thesis, the craniofacial fractured region has been chosen and pointed out to present the idea of the bootstrap method as an attribute to nature problem. Researchers have analyzed issues concerning craniofacial reconstruction via different prospects of study. To diagnose fractured regions, surgeons have used Computed Tomography scan, X-rays, and Magnetic Resonance Imaging, but it is a challenging process. Nonetheless, the availability of computer vision technology has enabled mathematicians to apply their approaches in 3D craniofacial reconstruction. For example, several works like Majeed and Piah (2016), Majeed and Piah (2014), Majeed et al. (2016) and Majeed (2016) have proposed utilizing Bezier-like functions to reconstruct the missing input data of a fractured region.

Miyasaka et al. (1995) conducted a study in reproducing the 2D manual reconstruction by proposing a computer imaging system. They build 2D parts database involving the nose, eye, hairstyle and contours. Computer based craniofacial method is also discussed in Claes et al. (2010). Besides, the mirroring method and interpolation method are also used in craniofacial reconstruction. For instance, Sauret et al. (2002) applied mirroring technique that could assist in allowing the use of digital images and manufacture the titanium implant. Deformation or surface interpolation method is used by Min and Dean (2003) to reconstruct the craniofacial reconstruction. Both mirroring and interpolation method is useful in craniofacial reconstruction.

In addition, Shui et al. (2010) employed craniofacial reconstruction using reference skull database, whereas Carr et al. (1997) discussed surface interpolation with RBF to reconstruct the unidentified fractured segment of a skull. In order to estimate the outlook of the fractured joints of a skull, the prompt in joining it using a method that requires low cost and high efficiency will be analyzed. This would be an alternative approach and enable the surgeons to reconstruct the fractured joints easily.

2.6.2 Drilling of Tunnel

Geological surveys at planned tunnel locations are needed. However, owing to technical constraints and the desire to minimize expenses, it is not unusual for such studies to be incomplete. The survey must be carried out during real tunneling activities in such instances. Exploration drilling is one technique that can be used during tunneling to monitor the condition of the ground. This also enable the researchers to