Handling of Missing Values in Static and Dynamic Data Sets

A thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

Faraj Bashir

The University of Sheffield
Department of Automatic Control and Systems Engineering

March 2019

Dedicated to my Mother and family,

thank you for your support.

Acknowledgements

First, I am thankful to the Almighty God who has given me the grace, strength and wisdom to sail through the three years of my PhD. I would like to gratefully and sincerely thank Dr. Hua-Liang Wei, Dr Viktor Fedun and Dr M. O. Tokhi for their guidance, understanding, patience, and most importantly, their friendship. I would like to express my sincere gratitude to Dr. Hua-Liang Wei for the guidance and thesis related ideas, and particularly for guiding me in this journey to the right path. My profound gratitude goes to my family for their prayers and regular encouragements. Thank you very much Renata Ashton, Matthew Ham and Darren Fox for all the help. Last but not least I would like to express my deepest appreciation to all those who provided me the possibility to complete this thesis.

Faraj ALi A Bashir

faabashir1@sheffield.ac.uk

STATEMENT OF ORIGINALITY

Unless otherwise stated in the text, the work described in this thesis was carried out solely by the candidate. None of this work has already been accepted for any degree, nor is it concurrently submitted in candidature for any degree.

Candidate (Faraj Bashir):.......................

Supervisor (Dr.Hua-Liang Wei):..................

Abstract

Despite considerable advances in missing data imputation techniques over the last three decades, research and data analysis across many fields are still affected by inferior techniques of imputation. Incorrect imputation can lead to bias, over confident intervals, and inaccurate conclusions. Many techniques have emerged in the literature as candidate solutions, including the traditional and modern methods such as listwise, regression, stochastic, maximum likelihood and multiple imputation and others. While these methods may have value in improving the data set, most of the traditional methods do introduce some level of bias but, more importantly, none of the traditional methods have been proved to be useful for handling missing data in nonlinear systems, dynamic systems and multivariate time series data sets. This thesis contributes by first, conducting a comparative study of traditional and modern classifications by highlighting the differences in their performance. Second, an algorithm to enhance the prediction of values to be used for data imputation with nonlinear models is presented. Third, a novel algorithm model selection to enhance prediction performance in the presence of missing data is presented. It includes an overview of nonlinear model selection with complete data, and provides summary descriptions of Box-Tidwell and fractional polynomial methods for model selection. In particular, it focuses on the fractional polynomial method for nonlinear modelling in cases of missing data. An analysis example is presented to illustrate the performance of this method.

Another novel technique for dealing with missing data in multivariate time series is also presented and studied. The new algorithm utilises a vector autoregressive model (VAR) to handle missing data by combining a prediction error minimization (PEM) routine with an expectation maximization (EM) algorithm. As shown in a simulation study, the proposed algorithm produces better estimates than traditional and modern methods such as listwise deletion, imputation by using sample means and variances. It also outperforms the naive approach of conducting linear regression on time series while ignoring the time dependency (i.e., treating observations at different time points as independent), K-nearest neighbour (KNN), Multivariate Autoregressive State-Space Modelling package (MARSS) and EM algorithms. An empirical example demonstrates the use of the new method showing the advantages and limitations of the proposed method. Lastly, empirical results obtained using real data provide a valuable and promising insight to the problem of missing data. Thus, this thesis has uniquely opened the doors of research to this area.

Contents

List of Figures

List of Tables

Nomenclature

A list of the variables and notation used in this thesis is defined below. The definitions and conventions set here will be observed throughout unless otherwise stated. For a list of acronyms, please consult page xvii.

e_l	likelihood estimation error
$\boldsymbol{\phi}$	Regression matrix
$\mathit{\Gamma}$	Multivariate covariance matrix
$\sum$	Univariate Covariance matrix
μ	Mean
$\overline{Q}$	Average of squared error
σ	Standard deviation
ε	White noise
φ	Indexed value
J	Jacobian matrix
L	Back shift operator

$logL$	Log likelihood
R^2	Squares sum of residual
S_T	Penalty function
SE	Total standard error
α_k	Parameters bias
δ	Threshold value
$\hat{\beta}_k$	Estimated parameters
$\hat{Q}_k$	Squared standard error
$\phi_m(x,p)$	Fractional polynomial model

Acronyms

ACE	Advance composition explorer
AIC	Akaik's information criteria
ARMA	Autoregressive moving average
BT	Box-Tidwell
CCA	Complete case analysis
corr	Correlation
DF	Degree of freedom
ECG	Electrocardiograph
ECG	Electrocardiograph
EEG	Electroencephalograph
EM	Expectation maximization
E-step	Expectation-step
FMRI	Functional magnetic resonance image
FP	Fractional polynomial
FPE	Final prediction error
HQ	Hannan-Quinn
I-step	Imputation-step
KNN	K nearest neighbour

LR	Likelihood ratio
MA	Moving average
MAR	Missing at random
MARSS	Multivariate autoregressive state space
MCAR	Missing completely at random
MI	Multiple imputation
ML	Maximum likelihood
MMSE	Minimum mean square error
MNAR	Missing not at random
M-step	Maximization-step
PEM	Prediction error minimization
P-step	Posterior-step
SC	Schwarz criteria
SPSS	Statistical Package for the Social Sciences
var	Variance
VAR	Vector autoregressive
VAR-IM	Vector autoregressive - Imputation
VARMA	Vector autoregressive moving average

Chapter 1

Introduction

1.1 Background

Virtually all scientific and research fields have suffered from data sets that are incomplete. These missing values can have tremendous impacts on the conclusions and recommendations that are made from the study. Nowhere is this more apparent than in the medical field where it is not possible to make a decision without full information about the case. Examples include certain regions of a gene microarray that may fail to yield measurements of the underlying gene expressions due to scratches, fingerprints, dust, or manufacturing defects. Also, participants in a clinical study may simply drop out during the course of the study leading to missing observations at critical time points. Similarly, a doctor may not order all applicable tests while diagnosing a patient resulting in the absence of potentially useful data. These varied reasons for missing data are sometimes referred to as the missing data mechanism.

The analysis of missing data processes leads to a theory of missing data in terms of its impact on learning, inference, and prediction [36]. This theory draws a distinction between three fundamental categories of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). An easy way to understand these categories is to look at a study of diabetic patients in which N participants are recorded in months X and Y. In the first month (X), all of the participants accomplished the test but only some of the participants continued the testing into the next month (Y).

If the missing data does not result from the measurements themselves such as a patient moving away from the study location, then this is the first category MCAR. However, if the missing data depends on observed measurements such as if, a patient drops out of a study due to poor discipline then this is MAR. And if the missing data can't be categorized into either of these categories then MNAR is used.

For example, suppose diabetes measurements for N participants are recorded in two months X and Y. In the first month X, all of the participants did the test but some of them (n) have a test in the second month and others do not. In the first category, the n participants in month Y were randomly selected from those participants in month X; this mechanism is missing completely at random (MCAR). In the second scenario, those who returned in Y measurements exceeded a normal level in month X; this is missing at random (MAR) but not missing completely at random. In the third category, those recorded in month Y were those whose month Y measurements exceeded normal level this mechanism is not missing at random (MNAR) [48], more details can be found in chapter 2.

The main question is: can the behaviour of the system be predicted when data sets have some missing values. Missing data problems are deeply related to statistical issues because most analytical methods depend on statistical theory. That means all imputed values for the missing data are depend on types of estimated models. This has made some researchers to consider missing data analysis problems to be the most significant issue within many real data analyses problems and applications [9]. In simple missing data situations, more often than not, the missing values are arbitrarily removed or the missing data value itself is simply replaced by its mean value. However, for cases where there are a significant number of missing data values, these strategies do not work well [12]. Recent research regarding modern methods of data imputation has concentrated on areas such as maximum likelihood estimators and multiple imputation techniques. These methods can produce good results for most applications [12, 35, 48, 105, 107]. Although the uses of these modern approaches still has greater interest in the literature, especially in case of static data set, there is insufficient knowledge to know if these methods can produce good results when applied to dynamic missing data set [11, 120].

1.2 Overview of the Thesis

- **Chapter 2** explores the various methods for analysing data with missing observations. Each method is explained as to how it works mathematically and a discussion of its limitations and advantages. Also represents an overview of multivariate time series. Some notions on multivariate time series analysis in time domains are succinctly introduced. Tools and conventions used herein are presented. They are essential to appreciate the contributions later in the thesis. Although they are widely available in textbooks, they have been adapted appropriately to suit this thesis.

- **Chapter 3** applies a Gauss-Newton method for nonlinear parametric estimation for the case of missing data. The primary aim is to introduce a nonlinear modelling technique for missing data analysis. Also, solving the model selection problem with missing data and providing accessible descriptions of nonlinear parametric with missing data is addressed.

- **Chapter 4** introduces improved method and a novel algorithm for handling missing values in multiple time series. An algorithm is introduced for handling missing data in multivariate time series based on a vector autoregressive (VAR) model. This is accomplished by combining an expectation and minimization (EM) algorithm with the prediction error minimization (PEM) method. A case study was conducted to compare the proposed algorithm with traditional and modern methods for imputing missing data.

- **Chapter 5** conducts two cases studies: one for space weather data and another for electrocardiogram (ECG) data. These case studies compare the $VAR - IM$ algorithm with different methods for imputing missing data. Missing data analysis, multivariate time series, and vector autoregressive models have been introduced for forecasting the electric flux from solar wind real data at geosynchronous orbit. Numerical results show that the proposed vector autoregressive models estimated by using the imputed data can produce promising prediction results for the relativistic electron flux. The ECG data set was used as a benchmark to test the performance and limitations of different missing data analysis methods.

- **Chapter 6** summarises the thesis conclusions and presents areas recommended for further research.

1.3 Publications

The author′s publications with relevance to this thesis are outlined below:

- **A. A. Bashir F.** and Wei H. (2015). Using Nonlinear Models to Enhance Prediction Performance with Incomplete Data. In Proceedings of the International Conference on Pattern Recognition Applications and Methods ISBN 978-989-758-076-5, pages 141-148. DOI: 10.5220/00051 57201410148.

- **Bashir, F.** Wei, H.L. and Benomair, A., 2015, August. Model selection to enhance prediction performance in the presence of missing data. In 2015, 20th International Conference on Methods and Models in Automation and Robotics (MMAR) (pp. 846-850). **IEEE**.

- **Bashir, F.** and Wei, H.L., 2015, October. Parametric and non-parametric methods to enhance prediction performance in the presence of missing data. In 2015, 19th International Conference on System Theory, Control and Computing (ICSTCC) (pp. 337-342). **IEEE**.

- Benomair, A.M., **Bashir, F.** and Tokhi, M.O., 2015, August. Optimal control based LQR-feedback linearization for magnetic levitation using improved spiral dynamic algorithm. In 2015, 20th International Conference on (pp.558-562). **IEEE**.

- **Bashir, F.** and Wei, H.L. 2016. Handling missing data in multivariate time series using a vector autoregressive model based imputation (VAR-IM) algorithm Part I: VAR-IM algorithm versus traditional methods. In 2016, 24th Mediterranean Conference on Control and Automation (MED). **IEEE**.

- **Bashir, F.** and Wei, H.L. (2016, August). Handling missing data in multivariate time series using a vector autoregressive model based imputation (VAR-IM) algorithm Part II: VAR-IM algorithm versus modern methods. In 2016, 19th International Conference on Computational Science and Engineering (CSE). **IEEE**.

- **Bashir, F.** and Wei, H.L., 2017. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. Neurocomputing. (in press)

- **Bashir, F.** and Wei, H.L., 2018. Missing Data Imputation on Independent Variables within Classification Models. : (To be submitted.)

1.4 List of Contributions

The contributions coming from the thesis are:

- Introducing a Gauss-Newton method for nonlinear parametric estimation to handling missing data. The primary aim is to introduce a nonlinear modelling technique for missing data imputation. Solving a model selection problem with missing data and providing new algorithm for missing data imputation (**Chapter 3**).

- A new method (MLD) for handling missing values in multiple time series is presented in (**Chapter 4**).
- A novel algorithm (VAR-MI) based on vector autoregressive (VAR) model to handling the missing values in multivariate time series introduced in (**Chapter 4**)
- A novel method was used to addressing and solving the incomplete data problems in space weather data and in ECG data. Comparing a novel method with different traditional and modern methods for imputing missing values in these data sets (**Chapter 5**).

Chapter 2

Literature Review of Missing Data Analysis in Static and Dynamic Data Sets

2.1 Introduction

This chapter discusses the concept of what missing data values are, missing data mechanisms, reviews important missing data patterns and mechanisms. Finally, a simple example to introduce and discuss the various methods that have been proposed to handle missing data in the literature is presented. Note that the first part of this chapter is limited to static data sets, and the review of dynamic data will be given later in this chapter.

Missing values, (incomplete data) simply means that observed data is not available for the output for the current response. Generally, missing data can be divided into three types: special numeric, numeric and character.

In practice, several reasons may lead to an unobserved response. For example, individuals responding to a survey sometimes fail to answer specific questions. In a measurement test, a sensor may fail to record data in some automatic process. Alternatively, the problem may be related to the output itself and some information may be purposely omitted or ignored during the work or in recording of the results [119].

Frequently researchers may be able to determine a systemic cause as to why data are missing. Typically, however the information is insufficient to give the main reasons for missing data. The ideal approach to abstain from missing information problems is to have a decent system (model) which minimizes the missing data [113].

2.2 Missing Data Mechanisms

It is important to classify the mechanisms of ″missing data″ because this would determine which missing data handling strategies would be used for specific problems. There are three important patterns of missing data which are MAR (missing at random), MCAR (missing completely at random) and MNAR (missing not at random) [72].

These patterns explain the relationships between the inputs and outputs of the system and the probability density function of missing values. In more detail, these mechanisms of missing values give the reasons why these values are missing or unobserved. For each pattern, a conceptual explanation will be given in the next paragraph, and for more details on missing data mechanisms, see [48, 105].

2.2.1 Missing Data at Random

Values are missing at random (MAR) when the probability of a missing value on an output Y (dependent variable) is related to the input (or inputs) X in the system but not to the response of the output Y itself. In other words, the probability of the missing values depends on the relation between the output Y and input (or inputs) X, that means there is no direct relationship between the probability of the missing values on Y and the values of Y variable itself [36]. MAR does not mean the value is missed in a random way. In fact, missing at random means that the probability of a missing value depends on a relationship between the output value and the input value for that variable. To give more detail, consider the data in Table 2.1 [32]. In this example, the dependent variable (Y) is the proportion of available chlorine in a certain quantity of chlorine solution and the independent variable (X) is the length of time in weeks since the product was produced. When the product is produced, the proportion of chlorine is 0.50. During the 8 weeks it takes to reach the consumer, the proportion declines to 0.49.

The first two columns in Table 2.1 show the complete values for the two variables (input X and output Y). The remaining columns represent the amount of Y, which appear in hypothetical missing data caused by three mechanisms. In the third column (”MAR”), the probability of missing values has a direct relationship with the variable X, where the values started missing after 30 weeks ($X > 28$). This mechanism is missing at random (MAR). In fact, there are many hypothetical MAR cases that can be generated from this example, depending on the probability function of the missing data. For example, if the proportion of available chlorine is unknown during the periods from 12 to 18 weeks or after 18 weeks, then the mechanism is MAR. As noted previously all missing cases occurred continuously and happened randomly. In fact there are no specific methods can prove that the probability function of values which is missed on the output Y is only a function of input X [36]. This is considered a significant problem in practice for analysing missing data because most of the modern techniques, such as multiple imputation and maximum likelihood approaches, assume the that the data is missing at random when it may be missed due to another mechanism [36].

2.2.2 Missing Data Completely at Random

The data missing completely at random (MCAR) that is what most fields consider as ″purely randomly″ missing. The basic property of MCAR is the probability density function of missing values for an output Y does not have a direct relationship with other outputs of the system or the values of the output itself. To some extent, it is similar to the MAR mechanism. On the other hand, with comparing MAR and MCAR, the latter has more restrictive random values because missing cases occur in a discrete form without considering the missing rate.

With regard to the data set of the proportion of available chlorine in Table 2.1 to mimic the MCAR case, the data was deleted or missed hypothetically in a random way. This random missing is not correlated with the output Y itself but it does have indirect relationship with the input X and It can be noted that the missing data were not isolated to a specific position in the response of the system.

Table 2.1: The proportion of chlorine and length of time in weeks with different missing data mechanism [32].

	Y			
X	Complete	MAR	MCAR	MNAR
8	0.49	0.49	0.49	0.49
8	0.49	0.49	0.49	0.49
10	0.48	0.48	0.48	0.48
10	0.47	0.47	-*	0.47
10	0.48	0.48	0.48	0.48
10	0.47	0.47	0.47	0.47
12	0.46	0.46	-	0.46
12	0.46	0.46	0.46	0.46
12	0.45	0.45	-	0.45
12	0.43	0.43	-	0.43
14	0.45	0.45	0.45	0.45
14	0.43	0.43	-	0.43
14	0.43	0.43	0.43	0.43
16	0.44	0.44	0.44	0.44
16	0.43	0.43	0.43	0.43
16	0.43	0.43	0.43	0.43
18	0.46	0.46	0.46	0.46
18	0.45	0.45	0.45	0.45
20	0.42	0.42	0.42	0.42
20	0.43	0.43	-	0.43
20	0.41	0.41	0.41	0.41
22	0.41	0.41	0.41	0.41
22	0.4	0.4	0.4	-
22	0.42	0.42	0.42	0.42
24	0.4	0.4	-	-
24	0.4	0.4	0.4	-
24	0.41	0.41	-	0.41
26	0.4	0.4	0.4	-
26	0.41	0.41	0.41	0.41
26	0.41	0.41	0.41	0.41
28	0.4	0.4	0.4	-
28	0.4	0.4	0.4	-
30	0.4	-	-	-
30	0.38	-	0.38	0.38
30	0.41	-	0.41	0.41
32	0.4	-	-	-
32	0.4	-	0.4	-
34	0.4	-	0.4	-
36	0.41	-	0.41	0.41
36	0.38	-	-	0.38
38	0.4	-	0.4	-
38	0.4	-	0.4	-
40	0.39	-	0.39	0.39
42	0.39	-	0.39	0.39

*Dashes indicate missing values.

For example, there are 11 measured values randomly selected from those were measured in 42 weeks; which means each missing data value is affected by the value of X, this method is MCAR but not MAR. By reconsidering the same data in Table 2.1, the basic meaning of MCAR is that, the missing values which are missed randomly from the measured data with a probability function is correlated with the input. This means that the cases with observed output Y has an input with average similar to the average of input that correlated to this missed output values. By testing the missing mechanism, it is possible to identify whether the values are missing completely at random or just at random [36] , more detailed information for the basic logic for such a test can be found in [101]. To apply this test, first, the missing and complete data should be separated and the mean of the data is determined for each case. If the mean for both cases has a small difference, then the data are missing completely at random. Also, the input variable should have the same mean value. To explain this, the input may be classified into two groups: observed and missing by dependence on the missing mechanism (MCAR or MAR) and comparing the mean of the groups. For example, consider a case where the mean of the observed data has an input of 22.85, and the missing data sample has a mean of 20.55. There is similarity between the group means, suggesting that the missing mechanism for the two groups is equivalent, giving evidence that the output Y is MCAR. As a contrast, the same procedure for the input in the MAR case could be done to check the contrast. The full-observed data input mean is 17.56, and the mean of incomplete cases is 34.83. This big difference shows that the missing values occur continually within a specific period. This is evidence for the MAR.

2.2.3 Missing Data Not at Random

The third missing mechanism is missing not at random (MNAR), that is, the values are missing not at random when the probability of a missing value on an output Y depends on Y itself but not on the input (or inputs) X. To illustrate, consider the previous data in Table 2.1. Values which equal to 0.40 (Y = 0.40) were unobserved, and there is not a clear direct relation between the input variable X and the missing values in the output Y. In other words, the probability of missing values depends on the variable Y only. This represents the category of MNAR. The same data set may have many different cases of this mechanism, which is determined by the probability function of missing values.

For example, if the system has missing values when the output Y< 0.40, then the missing value depends on the output Y itself, as in the case where Y > 0.40. Unlike the previous mechanism, no specific test available to check if data are MNAR without predicting the relation between the missing data and its variable [36].

2.3 Approaches to Deal with Missing Data

There are many missing data analysis methods. In general, these methods are divided into two groups: traditional and modern techniques [48]. Basically the traditional techniques can be relatively easily implemented without difficulty. On the other hand, modern methods require a high performance computer and powerful software. Both traditional and modern methods have advantages and disadvantages [36, 101, 103].

2.3.1 Traditional Missing-Data Techniques

Many missing data analysis methods are abundant throughout the literature. In this chapter, a limited selection of the widely used approaches is presented. Readers are referred to [36, 93, 104] for additional detailed information concerning missing data techniques.

2.3.1.1 Listwise Deletion

Listwise deletion simply discards data whose information is insufficient. This means that if any variable of the data is missing, then the entire record is thrown out. Listwise deletion is also known as filtering approaches and complete case analysis (CCA). This method is used in many missing data problems, but its implementation depends on the type of data mechanism [12]. If the data missing mechanism is MCAR, this technique would generate an unbiased estimation if the number of removed data records is small, but with a large number of removed data this is not true [37]. That means after applying CCA the data analysis process can deal with cases that have full observed values only. For example, in any estimation process when calculating a mean and variance for a variable Y, CCA discards any records, which have missing values on the variable Y and that may lead to a biased parametric estimation [6]. Furthermore, by omitting the missing values, a direct dramatic reduction in data size may result in data sets with large sample size.

2.3.1.2 Pairwise Deletion

Pairwise deletion is one of the commonly used missing data analysis methods (available case analysis) [6]. With this approach, the missing data are removed with an analysis by analysis principle that means any observed case may be used for some analyses but not all analyses. For example, every value in a parameters vector and matrix depends on the observed cases in each variable. Predominantly this method gives better results as compared with filtering approaches because it reduces the number of omitted cases in the observed data. In contrast, this method still works under the same central restriction as complete case analysis. Thus the data mechanism should be MCAR. Similar to filtering approaches, this technique leads to biased estimates when the data have different mechanism from MCAR [93].

To explain the principles of these deletion approaches, consider the data set in Table 2.1 for the proportion of available chlorine and length of time in weeks. A scatter plot of the complete data is shown in Figure 2.1. The negative correlation between the input X and the output Y (-0.86) means that the low proportion of available chlorine would have acquired high length of time in weeks. Figure 2.2 shows a scatter plot of the deletion approaches for the case of MAR, because there are only two variables; the scatter plot of available case analysis method is same as to that of complete case analysis [12].

This section will focus on the MAR mechanism to show how these approaches effect on the bias of parameters estimation. Because deletion approaches keep the case with full observed values of the variable Y, it systematically ignores the values from 28 weeks and on the plot also shows that there is weak nonlinear variation association between Y and X (linear relationship between X and Y). In the complete data set, the estimated value of the variable Y (mean value) is 0.425, whereas for the deletion approaches, analysis give an estimated value of 0.435. Similarly, the estimated value of the variable X is 22.27 for complete data and 17.56 for deletion methods. Even with taking the standard deviation into consideration, the proportion of available chlorine has a standard deviation 0.03053 for the complete case, as contrasted to the deletion methods yield a standard deviation of 0.02907.

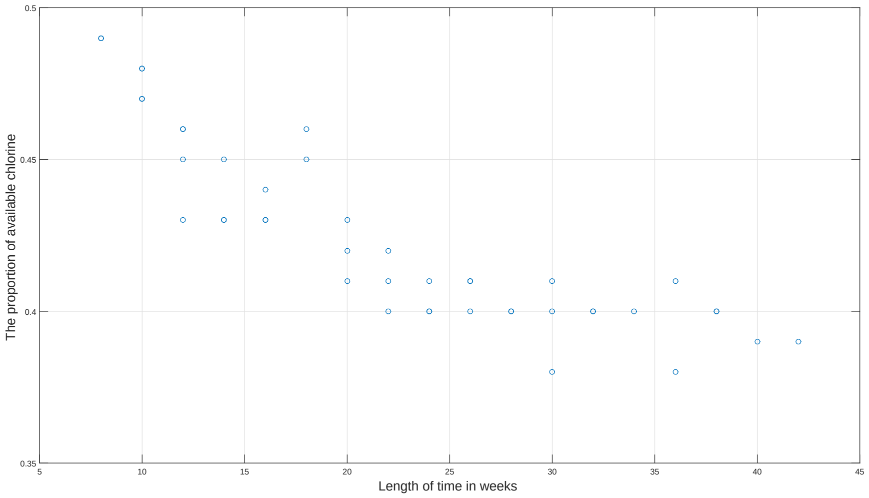

Figure 2.1: Complete-data scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution.

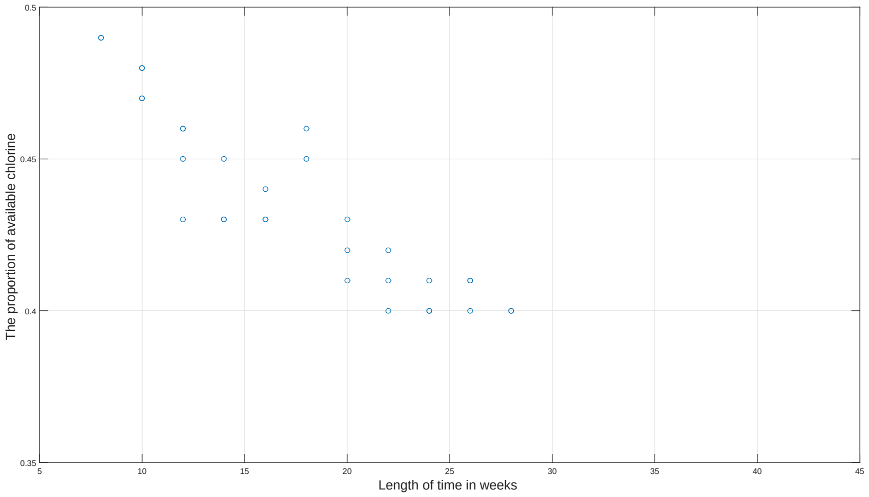

Figure 2.2: Deletion approaches scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution (MAR)

2.3.1.3 Imputation Methods

Imputation represents a group of common traditional methods where the estimator imputes (changes) the missing values with appropriate values [102]. In fact, there are many imputation approaches [12], but this study will concentrate on three of the most common methods: mean substitution imputation, linear regression imputation, and stochastic (random) regression imputation. The simplest method is mean imputation. This method imputes the missing values with the mean of the observed data [5, 35]. For example, for the data in Table 2.1 for the MAR mechanism case, the expected value of the observed output is 0.435, this value is substituted for the missing values in all records. Figure 2.3 shows that the imputed data from using mean substitution imputation are horizontally linear across the Y-axis at 0.435 with a zero slope.

In this case, the correlation between the input X and the output Y is equal to zero because the imputation of the missing data depends only on the output Y. Focusing on more features of mean imputation method, the cross correlation between the imputed output $\hat{Y}$ and input X is -0.497, in contrast to the complete data correlation is -0.86, the negative sign represents the opposite relation between the input and the output (as the input increase the output decrease). The data variability may not appear when the missing values are replaced by the average of observed data (a constant value). Considering the mean and standard deviation the mean imputation method produces a mean and standard deviation to be 0.435 and 0.025, respectively.

Regression imputation is a technique that fills missing values with expected value by using a regression model [1]. In this method, observed data of the output Y are used to estimate a regression model, which is used to impute the values of missing data.

Take the data in Table 2.1 as an example. In MAR mechanism, there are 12 unobserved values and 32 observed cases. The observed data of output Y (variable with missing data) are used with observed data on input X (variable with complete data) to impute the missing cases on output Y. In this case, linear regression model: $\hat{Y} = 0.509 - 0.0042X$ has been used. Applying the input X (complete data) on the regression model yields estimated output ($\hat{Y}$), and these estimated values impute the missing data of the output Y.

The basic idea of the regression imputation depends on a technique of borrowing information from the observed data of the output variable. This method also leads to a biased estimation, as shown in Figure 2.4. Notice that the linear regression imputation yields a correlation equal to -0.97 between the output Y and input X, in contrast with the correlation of -0.86 for the complete data case. Because the imputed data values are generated by a linear function, there are no fluctuations for the imputed values. Consequently, the imputation process will attenuate the variability of the imputed values. For example, the standard deviation estimation of the output Y from linear regression is 0.042, whereas it is equal to 0.025 in the case of complete data. Although linear regression yields a biased estimation of standard deviation and correlation for the MCAR or MAR data mechanism, it does yield unbiased estimates for the average value.

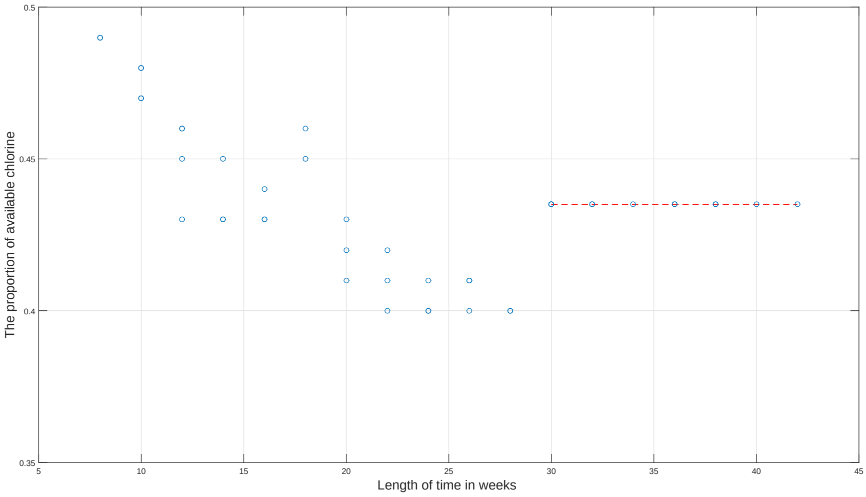

Figure 2.3: Mean imputation scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution (MAR).

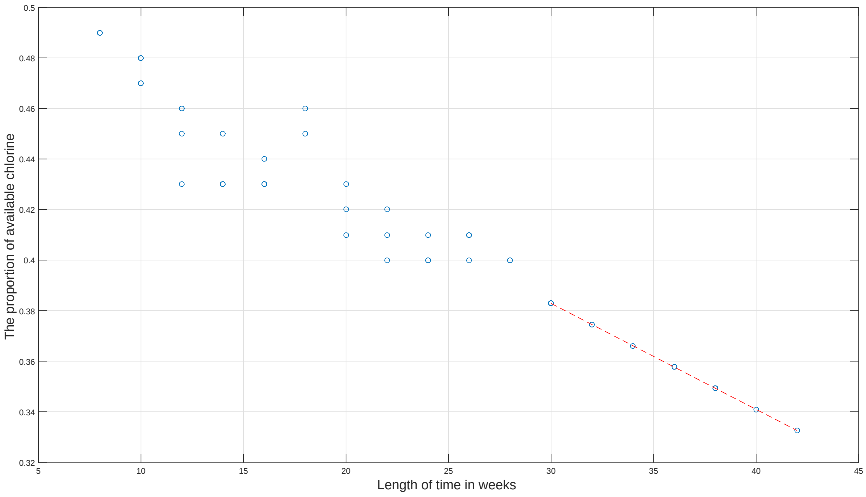

Figure 2.4: Linear regression imputation scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution (MAR).

Mean substitution imputation and linear imputation lead to a bias estimation, especially of correlating the standard deviation of both MAR and MCAR [48, 103]. Stochastic linear regression imputation can eliminate these biased estimates, it is similar to a standard regression imputation technique and it uses same regression model for imputing the missing data [12]. Stochastic linear regression is a linear regression method such that to each imputed value a random error is added. This random value is generated from a normal distribution with a variance equal to the residual variance and a mean of zero, estimated from the linear regression imputation model [5, 48, 102, 103].

Recall the data in Table 2.1, where the regression of the output Y on input X yields a residual variance of 0.000162. Then, the new random error is produced randomly from a normal distribution with a variance of 0.000162 and a mean of zero. These new error terms can then be added to the estimated output $\hat{Y}$, which is predicted from the linear regression model. Figure 2.5 shows the scatterplot of the imputed values of available chlorine data obtained from a stochastic linear imputation model.

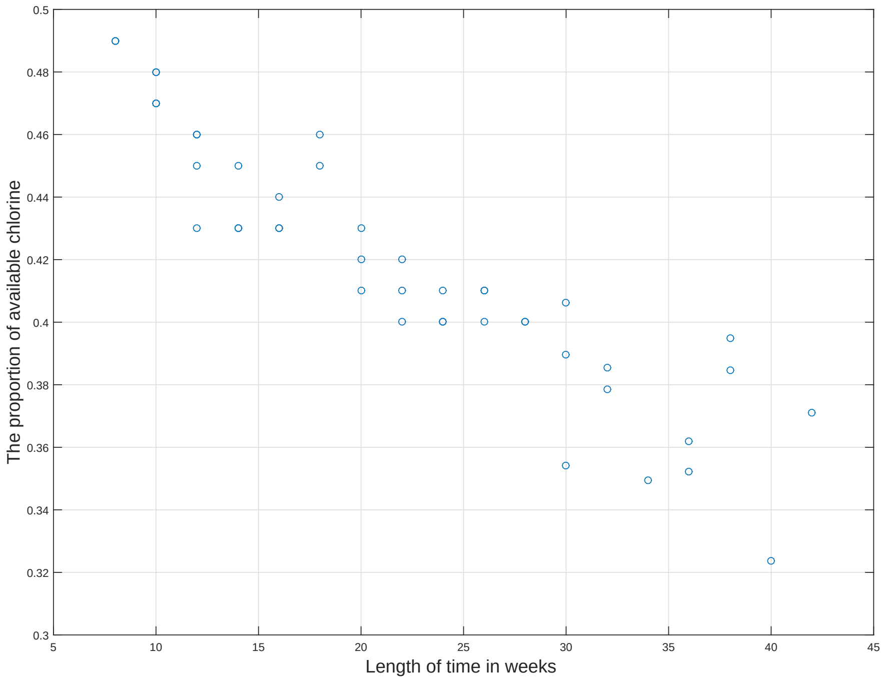

Figure 2.5: Stochastic imputation scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution (MAR).

Because there is a random error added to each imputed value, the imputed data do not represent a straight line, as that generated from a standard linear regression imputation model. Comparing Figure 2.2 with Figure 2.5 it is clear that the stochastic regression model produces a much better result. This slight adjustment to the regression model yields an unbiased parameter estimation in the case of MAR mechanisms. However, stochastic regression imputation may not be able to determine the actual error between the real and imputed values because it depends on random error values.

2.3.2 Modern Missing Data Techniques

The revolution of modern missing data techniques began in 1987 when two statisticians, Little and Rubin, published two books, Statistical Analysis with Missing Data [72], and Multiple Imputation for Nonresponse in Surveys [102]. Although some important articles were previously published e.g. [29, 59, 101], these two books for the first time represented a full background for missing data. There is powerful software coupled with these books, but new, more robust software is still needed today. Also a good book coupled with powerful software implemented with different programming languages was published by [103]: Analysis of Incomplete Multivariate Data. In addition, there are many useful articles were published recently which gave good background on the modern methods and software for missing data imputation [48, 64, 70]. The two modern missing data analysis approaches that have been suggested as the best techniques are: multiple imputation and maximum likelihood. These methods are considered better than traditional approaches because they need less assumptions and can handle most data types [5, 12, 35, 48, 103].

2.3.2.1 Maximum Likelihood

Missing data analysis with a maximum likelihood technique (sometimes referred to as "direct maximum likelihood" and "full information maximum likelihood") is an old procedure. Fifty-five years ago [33, 76], this method was applied to specific applications (e.g., bivariate time series with incomplete data) until the 1970s when statisticians developed cooperative techniques, which opened new windows for many applications of this method [29, 42, 101]. As mentioned previously, this modern routine has only been available in robust software packages from the end of the 1980s. Rather than dealing only with full observed cases, maximum likelihood uses both observed and incomplete cases to calculate the values of parameters that meet the peak of the probability density function for these parameters. Maximum likelihood estimation technique is implemented by software packages that are widely available on the internet and they are user-friendly and self-explanatory, therefore the mathematical procedures behind the parameter estimation process will not be addressed in more detail in this chapter. Unbiased estimation represents the main goal of any estimation process, and this can be achieved if Maximum likelihood is used for MAR mechanism cases [5, 103]. The following description is focused on the case that gives the most accurate estimation, the MCAR case, which needs additional assumptions is discussed later in this chapter. The estimation process starts by using a log likelihood mathematical function to identify the highest probability density function of the parameters population that are used to impute the missing values. The main goal of this method is to find the parameters that minimize the distance between the imputed and real values. In fact, this technique is similar to linear regression estimation, by using an ordinary least squares method, where the goal is to identify the parameters of a linear model that minimizes the distances between the real data (mean) and the estimated values. Applying the maximum likelihood estimation to a single variable case is simpler than applying it to a multivariable case, but it is not possible to apply it directly to a univariate or multivariate time series. To begin the process, the likelihood function (l) for a specific number of n data points, used to characterise the distribution of the data around the mean (μ) and the standard deviation (σ) for specific case k is defined as:

$$l_k = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{y_k-\mu}{\sigma}\right)^2/2} \tag{2.1}$$

The logarithm of the likelihood function for a specific number of data points (say n points) is:

$$logl = \sum_{k=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{y_k-\mu}{\sigma}\right)^2/2} \right] \tag{2.2}$$

So, the log likelihood is actually a summation of all the n individual probabilities; each single probability just simply represents a specific case of the normal distribution for the data. On the other hand, the log likelihood for a single case in a complete data set with normal distribution of a multivariate time series can be described as:

$$logl_k = -0.5[mlog(2\pi) + log|\Sigma| + (Y_k - \mu)^T \Sigma^{-1} (Y_k - \mu)] \tag{2.3}$$

where m is the order of system, y_k is the output vector at case k, μ represents the mean vector and Σ the covariance matrix of the observed values. The part $(y_k - \mu)^T \Sigma^{-1} (y_k - \mu)$ describes the distance value and is called Mahalanobis formula [36]. This formula is the squared distance that identifies the standardized space between each output measurement and the centre of normal distribution for the data. In fact, this standard distance represents the logarithm for the likelihood, which leads to produce a small deviation between the output vector and the mean vector [36].

The estimation process starts by determining Mahalanobis formula in equation (2.2) that gives the squared standardized distance for each measured value. The parameter values to determine this formula are the mean (μ) and the covariance matrix (Σ). Substituting these parameter values into equation (2.2) yields a squared distance that is in inverse proportion to the log likelihood function (i.e.,larger log likelihood value and small squared distance), this explains the theory of the maximization of likelihood function.

The main objective of this method is to calculate the exact values of the parameters of interest to yield the maximum likelihood value for each parameter, and this can be achieved by using an iterative algorithm, using the principle of substituting different mean and covariance values into the log likelihood formula until it produces the maximum value of log likelihood function. In other words, it estimates the parameters that minimize the value of Mahalanobis formula to achieve the highest log likelihood value.

Returning to the previous example taken from [32], more details about the maximum likelihood approach are given as follows, where the case of the MAR mechanism is considered, by assuming that the mean and variance values are $\mu = 0.42$ and $\sigma^2 = .0008$, respectively.

By substituting these parameter values in equation (2.3) for each observed output value, it yields different values of log likelihood function. Substituting by two different values, 0.47 and 0.43 in log likelihood function equation (2.3) yields two different values of 1.084 and 2.584, respectively.

It is clear that substituting a measured value of 0.47 gives a small log likelihood function compared with a value of 0.43. This is because the latter value is closer to the mean value. In other words, the best result should be the value that has a higher probability when the data represented by normal distribution with variance of $\sigma^2 = 0.0008$ and a mean of $\mu = 0.42$. Sometimes the value of the log likelihood function can give a negative result. In this case, the sign should be considered (the closed value to zero, becomes closest to the mean value and therefore associated with the best fit for parameter estimation).

In fact, when the parameters population of the system are unknown and are required to be predicted from the measured system input and output values, as mentioned before the estimation process depends on the iteration process. The maximum likelihood method is a technique that uses different values of parameters (mean and variance) to be substituted into equation (2.2). The results for all measured output values are summed to give the total log likelihood; this process is repeated until it finds the optimum parameters, which gives the best estimation.

In summary, the maximum likelihood approach attempts trial solutions using different parameter values to find which one gives the highest log likelihood value or that meets the highest probability, and the above explanation assumes the case of complete data analysis by maximum likelihood. But this technique can be adapted to handle the missing data problem for other cases as well. Fortunately, the maximum likelihood function can work for incomplete data and it does not need full and complete observed data.

2.3.2.2 A general Case for Multivariable Estimation

With the MAR mechanism, the Mahalanobis formula can be determined by using the available parameters and observed data. The best advantage of this method is that it estimates the parameters that give a better fit without discarding any part of the data. To explain how the multivariable estimation can be implemented by using log likelihood function in case of complete data, the same data taken from Table 2.1 [32] was considered. The Mahalanobis formula for the incomplete case is determined as follows:

$$(X_k - \mu)^T \Sigma^{-1} (X_k - \mu) = (42 - \mu_X)^T \rho_X^{-1} (42 - \mu_X),$$

where $\rho_X^{-1}$and μ_X (mean and variance, respectively) represent the unknown parameters, of the input X (this input case is avoided in calculation if the measured value is missed). As a multivariable estimation case, consider another case that include full observed data, for example the first case that has input value of 8 and output measured value of 0.49, so the resultant Mahalanobis formula is determined as:

$$(Y_k - \mu)^T \Sigma^{-1} (Y_k - \mu) = \left(\begin{bmatrix} 8 \\ 0.49 \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}^T \right) \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{y,x} & \sigma_y^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 8 \\ 0.49 \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \right)$$

2.3.2.3 Multiple Imputation (MI)

A multiple imputation technique was proposed by Robin [101], which is one of the most complicated methods among existing imputation approaches [5, 65, 102]. It depends on the iteration algorithms like the EM algorithm (will be explained later), because it needs to improve the estimation process in each iteration cycle to get the best parameters into several data sets [2].

Different Copies of data require different regression models. The output of these regression models are combined into one regression model to get to the final step of the multiple imputation approach. These procedures are divided into three main stages: imputation stage, analysis stage and pooling stage. In the next section, a brief illustration of the three stages will be discussed. More information may be found in [5, 36, 48, 50, 102–104].

- **Imputation phases** Various iterative algorithms can be used in the first phase but the data augmentation approach is still the best if the data is distributed normally [12]. The imputation process in this algorithm is divided into two procedures: the imputation procedure and the posterior procedure (I and P).

  The **imputation procedure** (I-procedure) produces a number of data sets; each one contains different prediction of missing data. The number of data sets varies between 15 and 20 [104]. It resembles a data augmentation algorithm in the same way as in the stochastic imputation technique in that it uses a covariance matrix and mean vector to construct regression models. Missing data are imputed by the estimated values from these models. Normally, values of the residual matrix, with zero mean and a constant variance, are added to the new imputed values (the variability process).

  The imputation step leads to the next imputation phase, that is, the **posterior procedure** (P-step), this step depends on the Bayesian estimation method to estimate the parameters of a regression model (the unknown mean vector and the covariance matrix for these estimated values). Conceptually, the posterior step determines the parameter estimates from the data that imputed from the previous step, and then adds a residual variation to each of the estimated values. This step generates a new set of parameters values that differs from the parameters that were used to impute the missing data in the previous imputation step. Using a new mean vector and covariance matrix values that resulted from the last posterior step to produce a new block of regression models in the next imputation procedure produces a new set data. This new data set has values differ from those at the previous imputation step. By iterating these two procedures up to a hundred of times, creates a specific number of copies of the data set. Sometimes 10 data sets are quite enough [72].

The ultimate purpose of the first phase is to generate a specific number of data sets. Each data set consists of values that differ from the other data sets values. The variation between these data sets is caused from the addition of a random error value to each imputed case. However, there is an autocorrelation between these I steps, so the first phase becomes more difficult to implement especially with a large number of missing data. For example, in an imputation phase that needs to generate 15 data sets, if the I step and P step iterated 200 times, then the hall process needs to iterate 3000 times taking extra time to process. This problem makes the multiple imputation less desirable to be used in commercial environments [93].

- **Analysis phase** After the generating the required number of data sets, statistical methods are used to analyse each data set. This process is called the analysis phase. It is considered the simplest among the multiple imputation phases. The main goal of this phase is to analyse the data sets that are generated from the imputation phase to be ready for to the next phase [36].

- **Pooling phase** This phase is sometimes known as the averaging step [93]. The pooling phase combines the average of the parameter estimates and their standard errors into a single data set. Formulas exist to determine the average and standard errors of the estimated parameters [102], and the pooling phase consists of three basic steps:

  1. Averaging the squared standard errors for all of copies of the data sets.

  $$\overline{Q} = \frac{1}{m} \sum_{k=1}^{m} \widehat{Q_k} \tag{2.4}$$

  where $\widehat{Q_k\,is}$ is the squared standard errors from the k^{th} data set and m is the total number of data sets.

  1. Calculate the parameter variance of the data sets.

  $$\sigma_\beta = \frac{1}{m-1} \sum_{k=1}^{m} (\widehat{\beta}_k - \overline{\beta})^2 \tag{2.5}$$

  where σ_β is the parameters variance, $\widehat{\beta}_k$ is the parameter estimation for k^{th} data set and $\overline{\beta}$ represents mean of the parameters of the system.

  1. Calculate the total standard error of the system

  $$SE = \sqrt{\overline{Q} + \sigma_\beta + \sigma_\beta / m} \tag{2.6}$$

  Although the multiple imputation phases can be tiresome, powerful software packages are available to facilitate, this step. They exist in different programming languages and can be used to perform these calculations quickly and accurately.

2.3.2.4 Expectation and Maximization Algorithm

Anderson proposed the basic idea of the maximum likelihood function and outlined it in simple steps [7]. If there are two different systems A and B, both having the same observed input data X. System A has complete output data Y and system B misses all of the output data. To estimate the missing data on the system B first, determine the average and variance value of the input X for system A and B. Use the observed data of the output Y of system A to estimate the parameters of the system B, considered as linear system. By using these estimated parameters, the missing data in system B can be predicted. The work accomplished by Anderson [7], assumed that the data has a single variable normal monotone pattern. However, in the general case these steps do require an iteration algorithm [104]. Dempster gave a good solution for the general case of missing data problem [29], he proposed an iterative algorithm called "Expectation and maximization" (EM) algorithm, the main idea of this algorithm is to estimate the system parameters needed to predict the missing data, this approach performs iteratively to obtain a solution by determining the best mean and variance among the parameters population. In fact, this method has been updated in recent years, and detailed discussions may be found in the open literature for example [72, 82]. Most of applications of the EM algorithm have concentrated on the missing data problem, by estimating the system parameters (mean and covariance) to predict the missing value. However some researchers have used this algorithm to solve difficult problems for complete data set cases. For example, structural equation model, multilevel linear models and finite mixture [8, 71, 87, 89, 95]. The following section describes the linear regression model estimation based on mean vector and variance estimation by using EM.

The expectation maximization algorithm is an iterative method consisting of two steps: an expectation step (”E-step”) and a maximization step (”M-step”), the iterative procedures require initial values to initiate the process of estimation, the parameters vector and matrix of the measured data are used for these initial values, and they can be determined by traditional missing data techniques including those that were previously discussed. The expectation step starts by using the initial mean vector and covariance matrix to construct the linear regression model that estimates the missing values from observed data. The maximization step is a procedure that comes after the expectation step to produce new parameter values for the estimated data. The EM algorithm stores the last mean vector and covariance matrix to determine the next expectation step, where it uses the result to build a new regression model that estimates new missing values. The maximization step subsequently runs again by using the updated estimates to determine the new parameters. The algorithm iterates these steps until the mean vector and the covariance matrix converge to some constant values or no longer change, where the converged value of the EM algorithm is the same as that of the maximum likelihood estimates [19, 82, 88].

In the optimization technique, the aim of estimation is to arrive at the maximum value of the log-likelihood (i.e., locate the maximum of the curve of log-likelihood function) where the required parameter estimates are settled. In the analysis of the optimization algorithm (EM), the starting point of the log likelihood curve represents the initial values of guessed parameters (e.g. mean vector and covariance matrix), and every iteration step (expectation step and maximization step) moves the parameter values closer towards the top of the curve. In other words, the aim of each single step is to set the mean and variance values in the right path which maximizes the value of log-likelihood function to make the estimated parameters move vertically. The expectation step is just a calculation process of the points that lie on the curve of the log likelihood function. Each maximization step maximizes the distance between the old and new parameters as it generates a next log-likelihood point which is large than previous value. The closer the parameters value approach the top of the curve, the distance between the coordinates, mean and variance value, becomes very small and the change of the log-likelihood values are very small. The iteration continues until the difference between parameters value is less than some small-specified number called the convergence number. In the literature, EM algorithm is known as maximum likelihood method because it searches for parameters that maximize the log-likelihood function.

The above illustration of the EM algorithm focused on the physical meaning of the process and ignored the conceptual meaning of the mathematical process. The description below provides details for mathematical conception, especially for the two main steps, the expectation and maximization step.

To explain the EM algorithm mechanism, a single variable analysis data case is considered in this illustration. Let U represent the input of the system with complete data, and Y is the output with incomplete data. To simplify the description, this system is considered with small number of data points with single input/output variables (single variable case). In case of missing data, the following formulas are used to determine the parameters, with the maximum likelihood approach [36].

$$\mu_U = \frac{1}{N}\sum_{i=1}^{N} U_i \tag{2.7}$$

$$\sigma_U^2 = \frac{1}{N}\left(\sum_{i=1}^{N} U_i^2 - \frac{\left(\sum_{i=1}^{N} U_i\right)^2}{N}\right) \tag{2.8}$$

$$\mu_Y = \frac{1}{N}\sum_{i=1}^{n} Y_i \tag{2.9}$$

$$\sigma_Y^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}\right) \tag{2.10}$$

$$\sigma_{U,Y} = \frac{1}{N}\left(\sum_{i=1}^{N} U_i Y_i - \frac{\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_i}{N}\right) \tag{2.11}$$

These equations consist of five sufficient statistics: the input and output data average (i.e., $\sum_{i=1}^{N} U_i$ and $\sum_{i=1}^{n} Y_i$), the squared sum of the input and output data (i.e., $\sum_{i=1}^{N} U_i^2$ and $\sum_{i=1}^{N} Y_i^2$), and the cross product of the input and output data (i.e., $\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_i$) [36]. These sufficient statistics are the basic data points to determine the model parameters, and are considered as a significant part in the expectation step.

The process of estimation starts with the expectation step, which imputes the missing data by using the initial conditions. After that the maximization step, these imputed values are substituted in (2.7) to (2.11) to estimate the new values for the parameters. The expectation step uses the new parameters values, to build the linear model equations which impute the missing values by using the observed input data. In the case of single variable data that has missing data on output Y, the formulas that used to build the linear model are:

$$\hat{Y} = \beta_0 + \beta_1 U \tag{2.12}$$

$$\beta_1 = \frac{\sigma_{U,Y}}{\sigma_U^2} \tag{2.13}$$

$$\beta_0 = \mu_Y - \beta_1 \mu_U \tag{2.14}$$

$$\sigma_{U,Y}^2 = \sigma_Y^2 - \beta_1^2 \sigma_U^2 \tag{2.15}$$

Equation (2.12) is a simple linear model, where $\hat{Y}$ is the predicted output value, β_0 and β_1 represent the linear coefficients of the model, and the parameter $\sigma_{U,Y}^2$ is the variance of the residual between the input U and output Y.

For missing data, this imputation procedure is not straightforward, because of difficulty in computing the sufficient statistics [29]. The Expectation step overcomes this difficulty by using the available observed data to determine the initial conditions that can be used initially to calculate the sufficient statistics. In fact, the EM algorithm depends on the borrowing of information from the observed data to predict the missing data. This is called conditional expectation. Further, depending on the mean of output data and the cross product of the input and output data terms $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_i$, respectively, the predicted output values are determined from equation (2.12). Then the expectation step uses these predicted values to determine the sufficient statistics. A small adjustment is then added to the squared sum of the output data by way of

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} \left(\hat{Y}_i^2 + \sigma_{U,Y}^2 \right) \tag{2.16}$$

where $\hat{Y}_i^2$ represents the predicted squared output data. The expectation step replaces the squared sum of the output data in equation (2.10) with the result of (2.16).

To further clarify the EM algorithm mechanism, assume a single variable case through the data taken from [32], where U represents the length of time in weeks and Y represents the proportion of available chlorine. The first step of EM is to estimate the initial values of the model parameters, mean vector and covariance matrix, and these initial values can be determined by other simple approaches such as regression imputation and complete data analysis [36, 72]. In this example the initial parameters values are estimated by using a listwise deletion technique as follows:

$$\mu_0 = \begin{bmatrix} \mu_U \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 22.27 \\ 0.435 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} \sigma_U^2 & \sigma_{U,Y} \\ \sigma_{Y,U} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 93.13 & 0 \\ 0 & 0.00093 \end{bmatrix}$$

In the first iteration, the algorithm borrows the initial values from the parameters vector and matrix to construct a linear model. This model then imputes the missing output data (the proportion of available chlorine) by using the complete input data (the length of time in weeks). Substituting the initial parameters values from mean vector (μ_0) and covariance matrix ($\sum_0$) into parameters equations result the following parameter values:

$$\beta_1 = \frac{0}{93.13} \Longrightarrow \beta_1 = 0$$

$$\beta_0 = 0.435 - (0)\mu_U \Longrightarrow \beta_0 = 0.435$$

$$\sigma_{U,Y}^2 = 0.00093 - (0)\sigma_U^2 \Longrightarrow \sigma_{U,Y}^2 = 0.00093$$

In this case all of the imputed values ($\hat{Y}$) are equal to the mean value $\hat{Y} =$ 0.435 The main aim of the expectation step is to impute the missing data of the output Y to determine the sufficient statistic terms $\sum_{i=1}^{N} Y_i$, $\sum_{i=1}^{N} Y_i^2$,

$\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_i$ and the squared output data Y_i^2:

$$Y_i^2 = \hat{Y}_i^2 + \sigma_{U,Y}^2 = 0.435^2 + 0.00093 = 0.19007$$

The first iteration of the expectation step calculations are shown in Table 2.2. Each expectation step is followed by a maximization step; using the results from the expectation step (Sufficient Statistics) in Table 2.2 to generate the new parameters of linear model. It substitutes the results of Table 2.3 through equations (2.12) and (2.16).

$$\mu_1 = \begin{bmatrix} \mu_U \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 22.27 \\ 0.435 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} \sigma_U^2 & \sigma_{U,Y} \\ \sigma_{Y,U} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 91.02 & -0.1157 \\ -0.1157 & 0.00083 \end{bmatrix}$$

Nevertheless, the imputed values of the output Y remain equal to the mean value, this is because the intersection parameter equals to the mean of the incomplete data. On the other hand, the variance of the output Y did changed a bit, even though, the missing values were imputed, and this was caused from the sufficient statistics equations itself, because in the generation of the variance most of statistical laws use $(N-1)$, but in this case the sufficient statistics equations uses only (N). After the first iteration, the next expectation step starts again by using the new mean vector and covariance matrix elements, and a new linear regression model is produced in next maximization step.

The same procedures that were done in the previous expectation step are repeated. By substituting the new parameters values in the sufficient statistics equations, the following results are obtained:

$$\beta_1 = \frac{-0.1157}{91.02} \implies \beta_1 = -0.0013$$

$$\beta_0 = 0.435 - (-0.1157)\mu_U \implies \beta_0 = 0.463$$

$$\sigma^2_{U,Y} = 0.00083 - (-0.0013)^2\sigma^2_U \implies \sigma^2_{U,Y} = 0.00068$$

In this case, all of the predicted values ($\hat{Y}$) do not equal the mean value, because the parameter β_1 has a non-zero value. Results of second expectation step are shown in Table 2.4. As before, the expectation step is followed by the maximization step. The Sufficient Statistics that yielded from the expectation step is shown in Table 2.5. The maximization step uses this result to predict new values of the mean vector and covariance value as follows.

$$\mu_2 = \begin{bmatrix} \mu_U \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 22.27 \\ 0.4306 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} \sigma^2_U & \sigma_{U,Y} \\ \sigma_{Y,U} & \sigma^2_Y \end{bmatrix} = \begin{bmatrix} 91.02 & -0.1758 \\ -0.1758 & 0.00084 \end{bmatrix}$$

In case of full observed data, the parameters values settled after the first iteration step because the parameters of the model enabled the log-likelihood function to reach the top of the curve. In contrast, in case of incomplete data the mean vector and covariance matrix for the output did not settle even in the second iteration. The reason for this is that the existence of missing data on the output Y, taking several iterations for the parameters values to reach the settling value. However, the number of iterations depends on the size of data set and number of missing values. In this example, the EM algorithm iterated 27 times to settle at the following mean vector and covariance matrix values:

Table 2.2: First Expectation step calculations

U_i	U_i^2	Y_i	U_i^2	U_iY_i
8	64	0.49	0.2401	3.92
8	64	0.49	0.2401	3.92
10	100	0.48	0.2304	4.8
10	100	0.47	0.2209	4.7
10	100	0.48	0.2304	4.8
10	100	0.47	0.2209	4.7
12	144	0.46	0.2116	5.52
12	144	0.46	0.2116	5.52
12	144	0.45	0.2025	5.4
12	144	0.43	0.1849	5.16
14	196	0.45	0.2025	6.3
14	196	0.43	0.1849	6.02
14	196	0.43	0.1849	6.02
16	256	0.44	0.1936	7.04
16	256	0.43	0.1849	6.88
16	256	0.43	0.1849	6.88
18	324	0.46	0.2116	8.28
18	324	0.45	0.2025	8.1
20	400	0.42	0.1764	8.4
20	400	0.43	0.1849	8.6
20	400	0.41	0.1681	8.2
22	484	0.41	0.1681	9.02
22	484	0.4	0.16	8.8
22	484	0.42	0.1764	9.24
24	576	0.4	0.16	9.6
24	576	0.4	0.16	9.6
24	576	0.41	0.1681	9.84
26	676	0.4	0.16	10.4
26	676	0.41	0.1681	10.66
26	676	0.41	0.1681	10.66
28	784	0.4	0.16	11.2
28	784	0.4	0.16	11.2
30	900	0.435	0.19007	13.05
30	900	0.435	0.19007	13.05
30	900	0.435	0.19007	13.05
32	1024	0.435	0.19007	13.92
32	1024	0.435	0.19007	13.92
34	1156	0.435	0.19007	14.79
36	1296	0.435	0.19007	15.66
36	1296	0.435	0.19007	15.66
38	1444	0.435	0.19007	16.53
38	1444	0.435	0.19007	16.53
40	1600	0.435	0.19007	17.4
42	1764	0.435	0.19007	18.27

Table 2.3: The Sufficient Statistics for first Expectation step iteration.

$\sum_{i=1}^{N} U_i$	$\sum_{i=1}^{N} U_i^2$	$\sum_{i=1}^{n} Y_i$	$\sum_{i=1}^{N} Y_i^2$	$\sum_{i=1}^{N} U_i Y_i$
980	25832	19.14	8.3622	421.21

$$\mu_3 = \begin{bmatrix} \mu_U \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 22.27 \\ 0.415 \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} \sigma_U^2 & \sigma_{U,Y} \\ \sigma_{Y,U} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 91.02 & -0.3815 \\ -0.3815 & 0.0018 \end{bmatrix}$$

The previous description is an illustration about the basics of the expectation step and maximization step and is not intended to be a detailed description of the EM algorithm method. As mentioned so far, the EM algorithm is similar to the maximum likelihood method.

To explain how an EM algorithm tracks the curve of log-likelihood function, another example is considered. The EM algorithm does not compute the log-likelihood function that is why the log likelihood does not appear in the calculation [36]. Further, substituting the initial values of mean vector, covariance matrix and the predicted output data in the log likelihood function is defined by equation (2.3). It gives the log-likelihood value of 99.73833.

Identically, in the second iteration step, substituting the new parameter values in log likelihood function and continue until the result settles or has only a small change. The log likelihood function for the output data is shown in Table 2.6. Notice that, as likelihood function nears to a final log-likelihood value, the difference between the current value and the previous one becomes smaller. The same thing happens in EM algorithm estimation.

Table 2.4: Second Expectation step calculations

U_i	U_i^2	Y_i	U_i^2	U_iY_i
8	64	0.49	0.2401	3.92
8	64	0.49	0.2401	3.92
10	100	0.48	0.2304	4.8
10	100	0.47	0.2209	4.7
10	100	0.48	0.2304	4.8
10	100	0.47	0.2209	4.7
12	144	0.46	0.2116	5.52
12	144	0.46	0.2116	5.52
12	144	0.45	0.2025	5.4
12	144	0.43	0.1849	5.16
14	196	0.45	0.2025	6.3
14	196	0.43	0.1849	6.02
14	196	0.43	0.1849	6.02
16	256	0.44	0.1936	7.04
16	256	0.43	0.1849	6.88
16	256	0.43	0.1849	6.88
18	324	0.46	0.2116	8.28
18	324	0.45	0.2025	8.1
20	400	0.42	0.1764	8.4
20	400	0.43	0.1849	8.6
20	400	0.41	0.1681	8.2
22	484	0.41	0.1681	9.02
22	484	0.4	0.16	8.8
22	484	0.42	0.1764	9.24
24	576	0.4	0.16	9.6
24	576	0.4	0.16	9.6
24	576	0.41	0.1681	9.84
26	676	0.4	0.16	10.4
26	676	0.41	0.1681	10.66
26	676	0.41	0.1681	10.66
28	784	0.4	0.16	11.2
28	784	0.4	0.16	11.2
30	900	0.435	0.19007	13.05
30	900	0.435	0.19007	13.05
30	900	0.435	0.19007	13.05
32	1024	0.435	0.19007	13.92
32	1024	0.435	0.19007	13.92
34	1156	0.435	0.19007	14.79
36	1296	0.435	0.19007	15.66
36	1296	0.435	0.19007	15.66
38	1444	0.435	0.19007	16.53
38	1444	0.435	0.19007	16.53
40	1600	0.435	0.19007	17.4
42	1764	0.435	0.19007	18.27

It is clear from the log-likelihood function that, as the value approaches the top of the curve, the curve becomes smoother, making the change in the results smaller.

Table 2.5: The Sufficient Statistics for second Expectation step iteration.

$\sum_{i=1}^{N} U_i$	$\sum_{i=1}^{N} U_i^2$	$\sum_{i=1}^{n} Y_i$	$\sum_{i=1}^{N} Y_i^2$	$\sum_{i=1}^{N} U_i Y_i$
980	25832	18.95	8.1969	414.298

Table 2.6: Output log-likelihood function.

Iteration	$logL_i$	μ_Y	σ_Y^2	$\sigma_{U,Y}$
1	99.73833	0.435	0.000845	0
2	97.84869	0.435	0.000826	-0.11568
3	93.51322	0.430646	0.000838	-0.17579
4	88.86392	0.427196	0.00091	-0.22194
5	85.02352	0.424518	0.001014	-0.25774
6	82.09372	0.422441	0.001128	-0.28551
7	79.89264	0.420829	0.001236	-0.30706
8	78.23475	0.419578	0.001333	-0.32378
9	76.97863	0.418608	0.001415	-0.33675
10	76.02173	0.417855	0.001484	-0.34681
11	75.28959	0.417271	0.00154	-0.35462
12	74.72758	0.416818	0.001585	-0.36068
13	74.29508	0.416466	0.001621	-0.36538
...	...	...	...	...
38	72.82354	0.415251	0.001752	-0.38162
39	72.82297	0.415251	0.001752	-0.38163
40	72.82252	0.41525	0.001752	-0.38163
41	72.82218	0.41525	0.001752	-0.38164
42	72.82191	0.41525	0.001752	-0.38164
43	72.8217	0.41525	0.001752	-0.38164
44	72.82154	0.41525	0.001752	-0.38164
45	72.82142	0.41525	0.001752	-0.38164
46	72.82132	0.415249	0.001752	-0.38165
47	72.82124	0.415249	0.001752	-0.38165
48	72.82119	0.415249	0.001752	-0.38165
49	72.82114	0.415249	0.001752	-0.38165
50	72.82111	0.415249	0.001752	-0.38165

2.4 Multivariable Missing Data Analysis

To some extent, the previous analysis is simple because the missing values occurred on just one variable (single variable). Using the EM algorithm to analyze missing data of multivariable data set is more complex because in each expectation step a different regression equation is needed for each variable that has missing values. Nevertheless, the basic idea of the algorithm remains the same and needs just a few modifications. To explain this addition, let us assume there are two dependent variables Y_1 and Y_2, which are related to a single input U. In this case, the algorithm will deal with three variables (U, Y_1 and Y_2) rather than two. The following additions for the sufficient statistics are required for output Y_1 and Y_2:

$$\mu_{Y1} = \frac{1}{N}\sum_{i=1}^{n} Y_{1i} \tag{2.17}$$

$$\sigma^2_{Y1} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_{2i}^2 - \frac{\left(\sum_{i=1}^{N} Y_{2i}\right)^2}{N}\right) \tag{2.18}$$

$$\sigma_{U,Y_1} = \frac{1}{N}\left(\sum_{i=1}^{N} U_i Y_{2i} - \frac{\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_{2i}}{N}\right) \tag{2.19}$$

$$\mu_{Y2} = \frac{1}{N}\sum_{i=1}^{n} Y_{2i} \tag{2.20}$$

$$\sigma^2_{Y2} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_{2i}^2 - \frac{\left(\sum_{i=1}^{N} Y_{2i}\right)^2}{N}\right) \tag{2.21}$$

$$\sigma_{U,Y_2} = \frac{1}{N}\left(\sum_{i=1}^{N} U_i Y_{2i} - \frac{\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_{2i}}{N}\right) \tag{2.22}$$

$$\sigma_{Y_1,Y_2} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_{1i}Y_{2i} - \frac{\sum_{i=1}^{N} Y_{1i}\sum_{i=1}^{N} Y_{2i}}{N}\right) \tag{2.23}$$

These equations consist of several sufficient statistics, namely the input and output data average ($\sum_{i=1}^{N} U_i$, $\sum_{i=1}^{n} Y_{1i}$ and $\sum_{i=1}^{n} Y_{2i}$), the squared sum of the input and output data ($\sum_{i=1}^{N} U_i^2$, $\sum_{i=1}^{N} Y_{1i}^2$ and $\sum_{i=1}^{N} Y_{2i}^2$), and the cross product of the input and output data ($\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_{1i}$, $\sum_{i=1}^{N} U_i \sum_{i=1}^{N} Y_{2i}$, and $\sum_{i=1}^{N} Y_{1i} \sum_{i=1}^{N} Y_{2i}$).

These sufficient statistics represent basic information to determine the model parameters.

In this case, there is one linear model with two regression equations for both predicted outputs Y_1 and Y_2.

$$\hat{Y}_1 = \beta_0 + \beta_1 U \tag{2.24}$$

$$\hat{Y}_2 = \beta_2 + \beta_3 U \tag{2.25}$$

In the case of single input/multi output data set, the maximization step does not require a modification because it depends on the data that is estimated from the expectation step. Therefore the modification occurs in the sufficient statistics formula. On the other hand, the maximization step requires additional modification when the data set has multi input/single output. For example, a system with two inputs U_1 and U_2 and single output Y needs residual covariance and regression equation as follows:

$$\hat{Y}_1 = \beta_0 + \beta_1 U_1 + \beta_2 U_2 \tag{2.26}$$

$$\sigma^2_{Y_1|U_1,U_2} = \sigma^2_Y - \beta_1\beta_2\sigma^2_{U_1,U_2} \tag{2.27}$$

$$\sum_{i=1}^{N} Y_{1i}^2 = \sum_{i=1}^{N}\left(\hat{Y}_{1i}^2 + \sigma^2_{Y_1|U_1,U_2}\right) \tag{2.28}$$

Compared with other algorithms, the EM algorithm is simple, useful, and does not require derivatives. Even with large data, it takes less time to implement with a software package [57, 77, 83]. On the other hand, the basic idea of the maximum likelihood method depends on a differential process however; the EM algorithm is able to skip this step [23, 117]. When extending the EM algorithm applications to multivariable missing data analyses, it becomes more complex as the number of independent and response variables increase. The increase in the number of inputs and outputs means an increase in difficulty of determination of expectation and maximization step. However, this difficulty can be overcome with modern powerful software packages [114].

2.5 Overview of Stationary Multivariate Time Series

A time series dataset is a set of measured values arranged by their sequential time order. A time series may be a collection of observations produced from a discrete time process, or a collection of discretized values gathered from a continuous time system, or any other time ordered sequence of data measurements.

Multivariate time series processes are of considerable interest in a variety of fields of engineering, sciences, and medicine. By studying many related variables together, rather than a single variable, a better understanding of the observed process may be obtained. Nowadays, improved data collection methods permit large amounts of time series multivariate data to be collected from various application domains.

For n time series $x_{1t}, x_{2t}, \ldots., x_{nt}$, let X_t denote a multivariate time series for an n-dimensional time series vector, where each x_{it} time series represents i^{th} raw of X_t vector, that is, for any time t, $X_t = \{X_{t1}, X_{t2}, \ldots, X_{tk} \}$. One of the fundamental objectives of multivariate time series analysis of X_t is to fit the data to a mathematical model to demonstrate the dynamic relationships among the univariate time series elements. The selection of a time series model encompassing X_t, depends on the dynamic interrelationships between these time series variables, and this relationship is further described by time lags between the data points for each time series.

The multivariate time series data set X_t , is a stationary time series, if at arbitrary time intervals $t_1,\ t_2, \ldots\ , t_k$, the probability distributions of the component time series variables $X_{t1}, X_{t2}, \ldots, X_{tk}$ and $X_{t1-p}, \ \ldots, X_{tk-p}$ are the same. Here k is the number of measured values, while p represents the lag. That means cross time intervals $t_1,\ t_2, \ldots\ , t_k$, throughout the stationary multivariate time series has a random probability distribution of the observed data points with respect to the time lags. Consequently, any stationary multivariate time series should have the same mean value (M) at any time interval :

$$M = \ E\left(X_t\right) = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_3 \end{bmatrix} \tag{2.29}$$

In addition, the covariance matrix, $\sum_X$ of a stationary time series X_t, is a constant matrix [108]:

$$\sum\nolimits_X = E[(X_t - M)\,(X_t - M)^T].$$

2.5.1 Covariance and Correlation for Multivariate Time Series

For the stationary multivariate time series X_t, the covariance matrix between X_{it} and $X_{j,-p}$ does not rely on the time t, for $(i,j) = 1, \ldots.., $ n. Rather it is a function of lag p.

where

$$Cov\left(X_{it},\ X_{j,t-p}\right) = E\left[\left(X_{it} - M_i\right)\left(X_{j,t-p} - M_j\right)^T\right] = \gamma_{ij}(p)$$

with the $n \times n$ cross-covariance matrix expressed as:

$$\Gamma\left(p\right) = E\left[\left(X_t - M\right)\left(X_{t-p} - M\right)^T\right] = \begin{bmatrix} \gamma_{11}(p) & \gamma_{12}(p) & \cdots & \gamma_{1n}(p) \\ \gamma_{21}(p) & \gamma_{22}(p) & \cdots & \gamma_{2n}(p) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{n1}(p) & \gamma_{n2}(p) & \cdots & \gamma_{nn}(p) \end{bmatrix} \tag{2.30}$$

and the $n \times n$ cross-correlation matrix at lag p becomes:

$$\rho\left(p\right) = U^{-1/2}\Gamma\left(p\right)U^{-1/2} = \begin{bmatrix} \rho_{11}(p) & \rho_{12}(p) & \cdots & \rho_{1n}(p) \\ \rho_{21}(p) & \rho_{22}(p) & \cdots & \rho_{2n}(p) \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1}(p) & \rho_{n2}(p) & \cdots & \rho_{nn}(p) \end{bmatrix} \tag{2.31}$$

For $p = 0, 1,\ 2, \ldots\ldots.$, the square root of the diagonal of the cross-covariance's matrix represents U vector:

$$U^{-1/2} = Diag\{\frac{1}{\sqrt{\gamma_{11}\left(0\right)}}, \ldots.., \frac{1}{\sqrt{\gamma_{nn}\left(0\right)}}\}$$

$$\rho_{ij}\left(p\right) = Corr\left(X_{it},\ X_{j,t-p}\right) = \gamma_{ij}(p)/\sqrt{\left[\gamma_{ii}(0)\gamma_{jj}(0)\right]} \tag{2.32}$$

with $\gamma_{ii}\left(0\right) = Var(X_{it})$. Thus, for $i = j$, $\rho_{ii}\left(p\right) = \rho_{ii}\left(-p\right)$ denotes the autocorrelation function of the i^{th} time series X_{it}, and for $i \neq j$, $\rho_{ij}(p) = \rho_{ji}(-p)$ denotes the cross-correlation function between the series X_{it} and X_{jt}. Note that $\gamma_{ij}(p) = \gamma_{ji}(-p)$, so

$$\begin{cases} \Gamma(p)^T = \Gamma(-p) \\ \rho(p)^T = \rho(-p), \end{cases} \tag{2.33}$$

In addition, the cross-covariance matrices $\Gamma(p)$ and cross-correlation matrices $\rho(p)$, have the property of non-negative definiteness, in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{b}_i^T \Gamma(i-j)\, \boldsymbol{b}_j \geq 0 \tag{2.34}$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{b}_i^T \rho(i-j)\, \boldsymbol{b}_j \geq 0 \tag{2.35}$$

For all positive integers k and all n-dimensional vectors $\boldsymbol{b_1}, \ldots, \boldsymbol{b}_k$, which follows since $Var(\sum_{j=1}^{n} \boldsymbol{b}_i^T X_{t-i}) \geq 0$.

2.5.2 Filtering of Multivariate Time Series

A multivariate linear filter relating an l dimensional input series U_t to n-dimensional output series Y_t is often formulated as:

$$Y_t = \sum_{N=-\infty}^{\infty} B_N U_{t-N} \tag{2.36}$$

where B_N are $n \times l$ matrices. The filter is physically realizable or causal if $\mathrm{B_N}= 0$ for $\mathrm{N} < 0$, leading to $Y_t = \sum_{N=-\infty}^{\infty} B_N U_{t-N}$ which means that $\mathrm{Y_t}$ can be characterized by past values of the input $\mathrm{U_t}$. The filter is said to be stable if $\sum_{\mathrm{N}=-\infty}^{\infty} ||\mathrm{B_N}||<\infty$. Under the stability condition, together with an assumption that the input random vectors $\mathrm{U_t}$ have uniformly bounded second moments, the output random vector $\mathrm{Y_t}$ defined by (2.36), exists uniquely and represents the limit:

$$\lim_{r\to\infty} \sum_{N=-r}^{r} \mathrm{B_N U_{t-N}}$$

Such that as $r \to \infty$

$$\mathrm{Y_t}= \mathrm{E}\left[\left(\mathrm{Y_t}- \sum_{N=-r}^{r} \mathrm{B_N U_{t-N}}\right)\left(\mathrm{Y_t}- \sum_{N=-r}^{r} \mathrm{B_N U_{t-N}}\right)^{\mathrm{T}}\right]$$

When the filter is stable and the input series U_t is stationary with cross-covariance matrices $\Gamma_u\,(p)$, the equation (2.36) is a stationary process [96].

The cross-covariance matrices of the stationary process Y_t are then given by:

$$\Gamma_u\,(p) = Cov\left(Y_t,\ Y_{t-p}\right) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} B_i \Gamma_u\,(p+i-j)\, B_j^T$$

2.6 Multivariate Time Series Linear Models

2.6.1 Wold Representation

Modelling of multivariate time series are useful processes for many types of data analysis, applications and forecasting. These processes require knowledge of the dynamic interrelationships between different kinds of variables and will provide useful information about their behaviour.

White noise ε_t, is an dimensional vector $\varepsilon_t = [\varepsilon_{1t}, \ldots\ldots, \ \varepsilon_{nt}]$ having a normal distribution with zero mean and constant variance σ, that satisfy the relationship $E(\varepsilon_t) = 0$ and $E(\varepsilon_t {\varepsilon_t}^T) = 0$.

$$E\left[\varepsilon_t \varepsilon_{t-p}^T\right] = \begin{cases} \Sigma_X & p = 0, \\ 0 & p \neq 0, \end{cases} \tag{2.37}$$

ε_t plays a special role in the analysis of linear models of stationary multivariate time series [47].

For the stationary multivariate time series $X = \{X_t\}$ with invariant mean M, current values can be estimated by using the previous values. Under particular and specific conditions it can be proven that, the multivariate time series $\{X_t\}$ can expressed as function of $\{\varepsilon_t\}$.

$$X_t = \sum_{j=0}^{\infty} G_j \varepsilon_{t-j} + M \tag{2.38}$$

where G_j are $n \times n$ dimensional matrices (coefficient matrices) and G_0 is $n \times n$ identity matrices . Equation (2.38) is known as "Wold Representation" [96].

As mentioned above, white noise plays an important role in modelling of multivariate time series. This is because the size of the ε_t vector affects the property states within the function.

The multivariate time series $\{X_t\}$ can be represented by a moving average model (MA) expressed as:

$$X_t = \varepsilon_t + \ G_1\varepsilon_{t-1} + \ G_2\varepsilon_{t-2} + \cdots + M$$

$$\dot{X}_t = X_t - M = \varepsilon_t + \ G_1\varepsilon_{t-1} + \ G_2\varepsilon_{t-2} + \ldots$$

$$\dot{X}_t = \sum_{j=0}^{\infty} G_j L^j \varepsilon_t$$

$$\dot{X}_t = G(L)\varepsilon_t, \tag{2.39}$$

where L is the backshift operator $\varepsilon_{t-j} = L^j\varepsilon_t$ and $G\left(L\right) = \sum_{j=0}^{\infty} G_j L^j$ [47]. Consider $G_j = \left[g_{rq,j}\right], r = 1,2, \ \ldots, \ n$ and $q = 1,2, \ \ldots, \ n,$ where $g_{rq}\left(L\right) = \sum_{j=0}^{\infty} g_{rq,j} L^j$. This can be rewritten as: $G\left(L\right) = \left[g_{rq}(L)\right]$ then

$$g_{rq,j} = \begin{cases} 1 & r = q, \\ 0 & r \neq q, \end{cases} \tag{2.40}$$

For a stationary multivariate time series, the coefficient matrices G_j need to satisfy the relationship $\sum_{j=0}^{\infty}\left(g_{rq,j}\right)^2 < \infty$, for $r = 1, 2, \ldots, n$ and $q = 1, 2, \ldots, n$. This results in the expectation of a mean of zero.

$$E\left[\left(\dot{X}_t - \sum_{q=0}^{s} G_q\varepsilon_{t-q}\right)\left(\dot{X}_t - \sum_{q=0}^{s} G_q\varepsilon_{t-q}\right)^T\right]_{as\ s\ \rightarrow\infty} \xrightarrow{} \quad 0 \tag{2.41}$$

2.6.2 The Vector Autoregressive Moving Average Model

The vector autoregressive moving average model for multivariate time series $VARMA(p,q)$ has the formula:

$$A_p\left(L\right)\dot{X}_t = B_q\left(L\right)\varepsilon_t, \tag{2.42}$$

$$A_p\left(L\right) = A_0 - \ A_1L - \ A_2L^2 - \cdots - \ A_pL^p$$

$$B_q\left(L\right) = B_0 - \ B_1L - \ B_2L^2 - \cdots - \ B_qL^q$$

where $A_p\left(L\right)$ and $B_q\left(L\right)$ represent the polynomials of order p and q for the two parts, autoregressive and moving average, respectively. A_0 and B_0 are $n \times n$ invertible matrices.

As a particular case, it can be assumed that $A_0 = B_0 = I$, where I is $n \times n$ identity matrix. For $p = 0$, the vector autoregressive moving average model $VARMA(0,q)$ represents a moving average model $MA(q)$,

$$\dot{X}_t = \varepsilon_t - \ B_1\varepsilon_{t-1} - \ B_2\varepsilon_{t-2} - \cdots - \ B_q\varepsilon_{t-q},$$

For $q = 0$ the vector autoregressive moving average model $VARMA(p,0)$ represents a vector autoregressive model $VAR(p)$,

$$\dot{X}_t{=}A_1\dot{X}_{t-1}{+}\ A_2\dot{X}_{t-2}{+}\cdots{+}\ A_p\dot{X}_{t-p}{+}\ \varepsilon_t$$

If the roots of $A_p\left(L\right)$ of the vector autoregressive moving average model is outside the unit circle, then the process is stationary. If the roots of the $B_q\left(L\right)$, are outside the unit circle, then the model is invertible [115].

Similarly, in the multivariate time series modelling $VARMA(p,q)$, to guarantee the unique function representation, the inevitability terms must be fulfilled. In other words, the selection of the $VARMA(p,q)$ model is specified by the values of p and q and the coefficient matrices $A_p\,(L)$ and $B_p\,(L)$ which are function in the covariance matrices of X_t.

The problem of the model selection for multivariate time models was first introduced by Wouter J. Den Haan in 1979. He stated that, to apply the model selection procedures for the stationary multivariate time series, it must fulfil the following conditions:

- For the coefficient matrices, $A_p\,(L)$ and $B_q\,(L)$, if $A_p\,(L) = \alpha(L)\beta(L)$ and $B_q\,(L) = \alpha(L)\gamma(L)$, then the determinant of $|\alpha(L)|$ should not equal zero , where $\alpha\,(L)$, $\beta(L)$ and $\gamma(L)$ are non-singular arbitrary matrices.

- The roots of the polynomials $A_p\,(L)$ and $B_q\,(L)$ must lie outside the unit circle.

Additional details for the multivariate time series model selection will be presented later in chapter 4.

$VARMA(1,1)$ Model

From equation (2.42), the first order $VARMA(1,1)$ model for the univariate time series system $\dot{X}_t$ $(n = 1)$, can be written as:

$$[I - A_1\,(L)]\,\dot{X}_t = [I - B_1\,(L)]\,\varepsilon_t \tag{2.43}$$

$$\dot{X}_t = A_1\dot{X}_{t-1} - B_1\varepsilon_{t-1} + \varepsilon_t \tag{2.44}$$

Similarly the $VARMA(1,1)$ model for the multivariate time series system $\dot{X}_t$ $(n = 2)$, can be written as:

$$[I - A_1\,(L)] \begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} = [I - B_1\,(L)] \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

$$\begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \dot{X}_{1,t-1} \\ \dot{X}_{2,t-1} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix}$$

$$\dot{X}_{1,t} = A_{11}\dot{X}_{1,t-1} + A_{12}\dot{X}_{2,t-1} - B_{11}\varepsilon_{1,t-1} - B_{12}\varepsilon_{2,t-1} + \varepsilon_{1,t} \tag{2.45a}$$

$$\dot{X}_{2,t} = A_{21}\dot{X}_{1,t-1} + A_{22}\dot{X}_{2,t-1} - B_{21}\varepsilon_{1,t-1} - B_{22}\varepsilon_{2,t-1} + \varepsilon_{2,t} \tag{2.45b}$$

In case of the univariate time series modelling, each time series $\dot{X}_{n,t}$ depends only on lagged values of the time series itself and the current and past values of the white noise. However, in the multivariate time series modelling, each time series is a function of the other lagged time series values and the current and previous values of ε_t. This dependability of each time series on the lagged values of other variables gives more advantages for multivariate than univariate modelling. For example, if $\dot{X}_{1,t}$ and $\dot{X}_{2,t}$ are the blood pressure and heart rate for a patient at time t, then from equation (2.45), the current blood pressure value depends not only on the previous blood pressure values, but also on the heart rate at that previous time period. Moreover, the heart rate will also be affected by the blood pressure measurements at the last period.

Model Average First Order Model $MA(1)$

From the equation (2.42) the moving average model $MA(1)$ for multivariate time series (two time series $\dot{X}_{1,t}$ and $\dot{X}_{2,t}$) can be represented by:

$$\dot{X}_t = (I - B_1(L))\varepsilon_t$$

$$\begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix},$$

The covariance matrix for $\dot{X}_t$ is

$$\begin{aligned}\Gamma(p) = E\left(\dot{X}_t\dot{X}_{t-p}^T\right) &= E([(I - B_1(L))\,\varepsilon_t]\left[(I - B_1(L))\,\varepsilon_{t-p}\right]^T) \\ &= E([(\varepsilon_t - B_1\varepsilon_{t-1})]\left[(\varepsilon_{t-p} - B_1\varepsilon_{t-p-1})\right]^T)\end{aligned}$$

$$\Gamma(p) = \sum B_1 \sum B_1{}^T \tag{2.46}$$

where ε_t is 2×1 vector with normal distribution of zero mean and covariance matrix $\sum$.

Transfer Function Model for $VARMA(p,q)$

By assuming $A_{12}(L) = 0$ in equation (2.45), then

$$\begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} - \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \dot{X}_{1,t-1} \\ \dot{X}_{2,t-1} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix}$$

$$\begin{bmatrix} 1 - A_{11}(L) & 0 \\ -A_{21}(L) & 1{-}A_{22}(L) \end{bmatrix} \begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} - \begin{bmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

$$\begin{bmatrix} 1 - A_{11}(L) & 0 \\ -A_{21}(L) & 1{-}A_{22}(L) \end{bmatrix} \begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} = \begin{bmatrix} 1 - B_{11}(L) & 0 \\ -B_{21}(L) & 1{-}B_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

To avoid the correlated noise, assume $B_1(L) = 0$, then

$$\begin{bmatrix} \dot{X}_{1,t} \\ \dot{X}_{2,t} \end{bmatrix} = \begin{bmatrix} 1 - A_{11}(L) & 0 \\ -A_{21}(L) & 1{-}A_{22}(L) \end{bmatrix}^{-1} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

$$\dot{X}_{1,t} = \frac{1}{1 - A_{11}(L)}\varepsilon_{1,t} \tag{2.47a}$$

$$\dot{X}_{2,t} = \frac{-A_{21}(L)}{(1 - A_{11}\,(L))(1{-}A_{22}(L))}\varepsilon_{1,t} + \frac{1}{1{-}A_{22}(L)}\varepsilon_{2,t} \tag{2.47b}$$

For a causal transfer function model, the noise series $\varepsilon_{1,t}$ must not be correlated as input to the output time series $\dot{X}_{2,t}$ (with lagged A_{21} coefficient matrix). Equation (2.47) to be a causal model, the covariance matrix between $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ must be zero's on the diagonal. Thus,

$$\sum = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$$

would have $s_{12} = s_{21} = 0$

***VARMA*(1, 1) Model Fitting**

As discussed in the first part of this chapter, the maximum likelihood method, for the invariant case, can be generalized to estimate the associated parameter matrices for a $VARMA(p,q)$ model, $A = \left(A_1, \ldots, A_q\right)$, $B = \left(B_1, \ldots, B_q\right)$ and the covariance matrices $\sum$. For the multivariate time series set $X = (X_1,\ X_2,\ \ldots,\ X_n)$, the maximum likelihood log function is represented by:

$$logL\left(A, B, \sum |X\right) = -\frac{n}{2}\left(mlog2\pi\ + log\left|\sum\right|\ +\ \sum_{i=1}^{n} \varepsilon_t{}^T \sum{}^{-1} \varepsilon_t\right),$$

where $\varepsilon_t = X_t -\ A_1 X_{t-1} -\ .\ .\ . - A_p X_{t-p} +\ B_1\varepsilon_{t-1} + \ \ldots\ +\ B_q \varepsilon_{t-q}$

and the residual sum of squares errors is $R = (A, B) = \sum \varepsilon_t \varepsilon_t{}^T$, then

$$logL\left(A, B, \sum |X\right) = -\frac{n}{2}\left(mlog2\pi\ + log\left|\sum\right|\ +\ \mathrm{tr}\sum{}^{-1}R\right)$$

[55, 90] introduced the maximum likelihood estimation for multivariate time series with different scenarios. Most researchers agree that the estimation method is quite difficult to implement and very slow to reach the

convergence state [55]. [21] recommended using the least squares estimation in the case of complete data rather than the maximum likelihood.

2.7 Multivariate Time Series and Forecasting

Most of recent multivariate time series literature uses the term “forecasting” rather than the term “prediction”. The basic tenets of the linear forecasting theories for multivariate time series were first introduced by [69, 116, 118]. They stated that the forecasting process represents one of the most important objectives in the analysis of multivariate time series. The multivariate modelling usually depends on forecasting, even if the main objective was for the control of the system. Forecasting of multivariate autoregressive models is a general case of univariate autoregressive models. To simplify the approach, the forecasting will be illustrated in univariate autoregressive process as a first step.

2.7.1 Minimum Mean Square Error Forecasting

The main aim of the forecasting process is to reach the optimum forecasting case and this occurs when the mean value of the error is at minimum. That fulfils the theorem of the least mean squares error forecasting. By satisfying this theory, the forecasting process can achieve the optimum future prediction.

For the stationary univariate time series X_t, $ARMA(p,q)$ model is:

$$A_p(L)X_t = B_p(L)\varepsilon_t$$

and the univariate $MA(q)$ model is:

$$X_t = B_q(L)\varepsilon_t$$

$$X_t = \varepsilon_t - B_1\varepsilon_{t-1} - B_2\varepsilon_{t-2} - \cdots - B_q\varepsilon_{t-q} \tag{2.48}$$

where $B_0\left(L\right) = 1$, and for $t = T + n$. Then

$$X_{T+n} = \sum_{j=0}^{\infty} B_j\varepsilon_{T+n-j}$$

where n is called the origin point of the forecasting. If $\{X_n,\ X_{n-1},\ X_{n-2},\ \ldots.\}$ are the observations at time $t = n$, then to predict T_{th} step in the future X_{T+n} as function of the observations $X_n,\ X_{n-1},\ X_{n-2},\ \ldots.$

The least mean square error forecasting $\hat{X}_n(T)$ of X_{T+n} can be expressed from equation (2.48) as:

$$\hat{X}_n\left(T\right) = B_T^*\varepsilon_n - B_{T+1}^*\varepsilon_{n-1} - B_{T+2}^*\varepsilon_{n-2} - \ldots$$

where the coefficient matrix of B_q^* is to be calculated. The average of squared errors of the prediction is:

$$E(X_{T+n} - \hat{X}_n\left(T\right))^2 = MSE$$

The main aim of the forecasting model is to drive the process leads close to each other, $X_{T+n} \approx \hat{X}_n\left(T\right)$, resulting in $MSE \approx 0$.

During the forecasting process, the predicted noise should satisfy:

$$E\left(\varepsilon_{n+j} \mid X_n,\ X_{n-1}, \ldots\right) = \begin{cases} 0 & j > 0 \\ \varepsilon_{n+j} & j \leq 0 \end{cases}$$

where

$$E\left(X_{n+T} \mid X_n,\ X_{n-1}, \ldots\right) = B_1\varepsilon_n - B_{T+1}\varepsilon_{n-1} - B_{T+2}\varepsilon_{n-2} - \ldots$$

and the forecasting values of X_{n+T} when $MSE \approx 0$ is:

$$\hat{X}_{n+T} = E\left(X_{n+T} \mid X_n,\ X_{n-1}, \ldots\right) \tag{2.49}$$

Where $\hat{X}_{n+T}$ is the T-step ahead of the forecast of X_{n+T} at the origin point and the forecasting error is expressed as:

$$z_n\left(T\right) = X_{n+T} - \hat{X}_{n+T} = \begin{cases} 0 & t \leq n \\ MSE & t > n \end{cases}$$

For the stationary univariate time series 95% forecast limits are:

$$\hat{X}_{n+T} \pm 1.96\left(s\sqrt{1+\sum_{j=1}^{T-1} B_j^2}\right)$$

where s is the standard deviation.

2.7.2 Forecasts Computation for $ARMA(p,q)$ Model

The first step of the predictions can be initiated by using the condition expectation formula shown in equation (2.49).

For $ARMA(p,q)$model:

$$A_p(L)X_t = B_p(L)\varepsilon_t$$

$$(1-A_1\left(L\right)-A_2\left(L\right)-\cdots-A_p\left(L\right))X_t = (1-B_1\left(L\right)-B_2\left(L\right)-\cdots-B_p\left(L\right))\varepsilon_t$$

For $t = n+T$, X_t can be written as:

$X_{n+T} \ = \ A_1X_{n+T-1} + A_2X_{n+T-2} + \cdots + A_pX_{n+T-p} + \varepsilon_{n+T} - B_1\varepsilon_{n+T-1} - B_2\varepsilon_{n+T-2} - \cdots - B_p\varepsilon_{n+T-q}$

By applying the conditional expectation at the origin point n, $\hat{X}_n\left(T\right)$ becomes:

$\hat{X}_n\left(T\right) = A_1\hat{X}_n(T-1) + A_2\hat{X}_n(T-2) + \cdots + A_p\hat{X}_n(T-p) + \widehat{\varepsilon}_n(T) - B_1\widehat{\varepsilon}_n(T-1) - B_2\widehat{\varepsilon}_n(T-1) - \cdots - B_p\widehat{\varepsilon}_n(T-q)$

Where

$$\hat{X}_n\left(T\right) = \begin{cases} X_{n+j} & j \leq 0 \\ E\left(X_{n+j} \mid X_n,\ X_{n-1},\ldots\right) & j > 0 \end{cases}$$

and

$$\widehat{\varepsilon}_n\,(T) = \begin{cases} 0 & j > 0 \\ \varepsilon_n(j) & j \leq 0 \end{cases}$$

For $ARMA(1,1)$model:

$$(1 - A_1\,(L))\,X_t = (1 - B_1\,(L))\varepsilon_t$$

$$X_{n+T} = A_1 X_{n+T-1} - B_1 \varepsilon_{n+T-1} + \varepsilon_{n+T}$$

and

$$\hat{X}_n\,(T) = A_1 \hat{X}_n(T-1) - B_1 \widehat{\varepsilon}_n(T-1) + \widehat{\varepsilon}_n(T)$$

$$\hat{X}_n\,(T) = {A_1}^T \hat{X}_n - {B_1}^T \widehat{\varepsilon}_n$$

$$\hat{X}_n\,(1) = A_1 \hat{X}_n - B_1 \widehat{\varepsilon}_n$$

$$\hat{X}_n\,(2) = {A_1}^2 \hat{X}_n - {B_1}^2 \widehat{\varepsilon}_n$$

Numerical example:

To clarify the concepts of the forecasting process, consider the first order autoregressive model $AR(1)$:

$$(1 - A_1\,(L))\,(X_t - 5) = \ \varepsilon_t$$

The standard deviation of the distributed data is 0.2, the coefficient $A_1 = 0.4$ and the observations of X_t are $X_{30} = 5,\ \ X_{31} = 4.5,\ \ X_{32} = 4,\ \ X_{30} = 5.4$. Using a forecast confidence limit of 95%, the forecast for 3-steps ahead of X_t is:

$$X_t = A_1\,(X_t - 5) + 5 + \varepsilon_t$$

For $j > 0$, $\varepsilon_j = 0$, and from equation (2.49) :

$$\hat{X}_n\,(T) = A_1(\hat{X}_n\,(T-1) - 5) + 5$$

$$\hat{X}_n\,(T) = {A_1}^T(X_n - 5) + 5$$

$$\hat{X}_{34}\,(1) = 0.4^1(X_n - 5) + 5 = 5.16$$

$$\hat{X}_{35}\,(2) = 0.4^2(X_n - 5) + 5 = 5.06$$

$$\hat{X}_{36}\,(3) = 0.4^3(X_n - 5) + 5 = 5.03$$

and 95% confidence intervals for the forecasts values are:

$$\hat{X}_{34}\,(1) \pm 1.96 * 0.2\sqrt{1 + (0)} \rightarrow 4.6 < X_{34} > 5.71$$
$$\hat{X}_{35}\,(2) \pm 1.96 * 0.2\sqrt{1 + (0.4)^2} \rightarrow 4.64 < X_{35} > 4.48$$

$$\hat{X}_{36}\,(3) \pm 1.96 * 0.2\sqrt{1 + (0.4)^3} \rightarrow 4.6 < X_{36} > 5.45$$

The software programs for modelling and forecasting VARMA models are not widely available. One identified by Scientific Computing Associates (SCA) is their multivariate time series package (MTC) [66], SAS program. Unfortunately, these software programs work under restricted conditions and are not easy to be implemented for the VARMA model tasking. For these reasons, these models will not be further addressed.

2.8 Causality

In considering the design of multivariate time series models, a structure is required for representing both the behaviour of each time series separately, and to address cross connections among the multivariate time series. The objective for displaying and analysing the time series together is to comprehend the dynamic connections among the time series over time. In addition, another benefit is the ability to enhance the predictions for each time series by using extra information available from the dynamic relationships among these time series. In view of these targets, the class of multivariate time series modelling (e.g. (p, q)) is designed and its properties are analysed [69, 116, 118].

The concept of the causality is not specified in general system identification procedure. It is particularly relying on the cause and effect relationship between the time series elements themselves.

Consider that $X = \{X_{1t},\ X_{2t}\}$ are stationary time series, being restively linked with each other. Under specific circumstances, it can be assumed that series X_{1t} causes X_{2t} this type of assumption is significant, especially when planning system behaviour, analysis or modelling [97].

In multivariate time series modelling, most methods dealing with causal inferences are based on the concepts of forecasting. Among these methods, the approaches developed by [51–53] are considered as the most useful and generally accepted technique in practice.

2.8.1 Granger Causality

The main notion of the Granger causality test relies on the statistical principle of specifying if one or more time series from multivariate set X_t can have an effect on the forecasting values of a specified time series X_{nt}. Generally, the correlation in regression process indicates the relationships between variables in the periodicity of measuring time (observed data). In contrast, Granger discussed that the cause and effect in multivariate time series can be determined by testing the ability to forecast the next values of the time series at time $(t + T)$ by using the observed values of other time series.

Deductions of cause and effect relationships in multivariate time series analysis rely mostly on the notion of Granger causality [51–53]. Unlike the other causality tests, this technique does not depend on the correlation between the observed values of the time series, but it relies on investigation of cause and effect relationships in the prediction period. The general definition of Granger causality test is based on two main basic assumptions:

- The causes in time-varying functions always proceeds the effect.
- The causal time series includes specific information about the affected time series.

When the first occurrence of a cause within a process is fulfilled, and it is usually considered as a primary principle driver for the other causality test techniques. On the other hand, the second basic assumption is quite difficult to be specified, as it needs provisionally unique information about the affected time series. That requires knowledge of all prior information about each time series to specifying the unique information for the affected time series. To that situation, Granger divided the specification of the available information into two sets:

- $f^*(t)$ is the available information set for the multivariate time series set X_{t^*} at specific time t^*.
- $f_n^*(t)$ is the available information for all time series, except the affected time series X_{nt^*} at the same specific time t^*.

Given that all the available information about multivariate time series at time t^* are included in the information set $f^*(t)$, if the time series X_{1t^*} cause time series X_{2t^*}, based on the above basic the conditional probability distribution of X_{2t^*+1}, then two different information set $f^*(t)$ and $f_n^*(t)$ result [97].

2.8.2 Granger Causality in the Context of $VARMA(p,q)$

In the case of $VARMA$ model, the Granger causality test starts by separating the multivariate time series X_t in two parts of time series Z_t and Y_t, then separating the $VARMA$ model into two models, $VAR(p)$ and $MA(q)$. Hence,

$$\begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} Z_t \\ Y_t \end{bmatrix} = \begin{bmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{2.50}$$

The multivariate time series X_t is assumed stationary, and then the model equation (2.50) is both stable and invertible. The MA model canonical form is

$$\begin{bmatrix} Z_t \\ Y_t \end{bmatrix} = \begin{bmatrix} G_{11}(L) & G_{12}(L) \\ G_{21}(L) & G_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

For MA model, the multivariate time series Z_t is Granger-affected by Y_t if and only if $G_{12}(L) = 0$ [79].

where

$$\begin{bmatrix} G_{11}(L) & G_{12}(L) \\ G_{21}(L) & G_{22}(L) \end{bmatrix} = \begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix}^{-1} \begin{bmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{A_{22}(L)}{A_{11}(L)A_{22}(L)-A_{21}(L)A_{12}(L)} & \frac{-A_{12}(L)}{A_{11}(L)A_{22}(L)-A_{21}(L)A_{12}(L)} \\ \frac{-A_{21}(L)}{A_{11}(L)A_{22}(L)-A_{21}(L)A_{12}(L)} & \frac{A_{11}(L)}{A_{11}(L)A_{22}(L)-A_{21}(L)A_{12}(L)} \end{bmatrix}$$

For simplicity, assume

$$K(L) = \frac{A_{22}(L)}{A_{11}(L)\,A_{22}(L) - A_{21}(L)A_{12}(L)}$$

$$\begin{bmatrix} G_{11}(L) & G_{12}(L) \\ G_{21}(L) & G_{22}(L) \end{bmatrix} = \begin{bmatrix} K(L) & \frac{-K(L)A_{12}(L)}{A_{22}(L)} \\ \frac{-K(L)A_{21}(L)}{A_{22}(L)} & \frac{1}{A_{22}(L)} + \frac{K(L)A_{12}(L)A_{21}(L)}{A_{22}(L)^2} \end{bmatrix}$$

From the Granger Causality test for MA model $G_{12}(L) = 0$, if and only if, Y_t is not causal for Z_t.

Then:

$$G_{12}(L) = K(L) B_{12}(L) - \frac{K(L) A_{12}(L)}{A_{22}(L)} = 0$$

Or

$$B_{12}(L) = \frac{A_{12}(L)}{A_{22}(L)} \tag{2.51}$$

Equation (2.51) can be generalized for $VARMA(p,q)$ model for simplicity. Assume, for example, a first order $VARMA(1,1)$ model

$$\begin{bmatrix} Z_t \\ Y_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Y_{t-1} \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

In this case equation (2.51) will be

$$B_{12} = \frac{-A_{12}(1 + B_{22})}{1 - A_{22}}$$

For Granger-causality test

$$B_{12} = -A_{12} = 0$$

Or $(1 + B_{22}) = (1 - A_{22})$

Chapter 3

Using Nonlinear Models to Enhance Prediction Performance with Incomplete Data

3.1 Introduction

Modern research on data imputation has concentrated on maximum likelihood methods such as the EM algorithm to deal with missing data problems. These methods can produce good results for most applications and generally, these approaches are much improved as compared to traditional methods. One benefit of these modern techniques is that in many particular applications, the estimate of parameters is unbiased. However, these methods do not work well for nonlinear systems, especially those exhibiting highly nonlinear behaviours. This chapter introduces nonlinear parametric imputation technique for the case of missing data. First, an overview of biased and unbiased linear parametric estimation with missing data is presented followed by descriptions of the Gauss-Newton method. In particular, This chapter explores in detail the Gauss-Newton iteration method for nonlinear parametric estimation in the case of missing data. However, the Gauss-Newton method needs initial values that are hard to obtain for missing data. To overcome this, The EM algorithm was used to estimate the initial values.

In linear model identification, the general formula of the linear model is known. It is only necessary to identify the key dependent and independent variables to be included in the model. While generally a simple and popular procedure in estimation process is to assume linear relationship between the predictor and dependent variables, the assumption may not always work well especially for severely nonlinear systems. For these systems, researchers try using nonlinear analysis techniques, but they are faced with a challenging problem, which is selecting the best model from different candidate nonlinear models or desired nonlinear representations. The form of the model needs to be specified, the parameters need to be estimated in some iterative manner, with the initial values for those parameters being provided. There are many methods for nonlinear model selection including the Box and Tidwell transformation technique [22], a modified Box and Tidwell method [86] and fractional polynomial (FP) approach [98], where these methods work well only for complete data [100]. Consequently, the overarching purpose of this chapter is to introduce some nonlinear model selection methods for incomplete data. Firstly, this chapter will present a brief overview of nonlinear model selection approaches. Then it illustrates the Box-Tidwell and fractional polynomial methods for missing data. In some detail, this chapter will focus on fractional polynomial method for missing data analysis by using a maximum likelihood and Gauss-Newton algorithm, this chapter also present analysis examples to illustrate the performance of these methods. The last part of this chapter focuses on the effect of missing data mechanism on nonlinear parametric estimation in the presence of missing data.

3.2 Gauss-Newton Algorithm

The linearization technique for nonlinear regression is an approach widely used in nonlinear regression model estimation [84]. The basic idea of nonlinear estimation by linearization method consists of two steps: the linearization of the nonlinear system and the estimation of model parameters [109]. Linearization can be implemented by a Taylor series expansion of the nonlinear model at a specific operating point. For example, for a nonlinear model $f(X, \beta)$ consisting of a number of samples i and n parameters (X is input and β is the estimated parameter vector) the linearization result with respect to the operation point β_0 is:

$$f\left(X_i, \beta\right) = f\left(X_i, \beta_0\right) + \sum_{k=1}^{n} \left[\frac{\partial f\left(X_i, \beta\right)}{\partial \beta_k}\right]_{\beta=\beta_0} \left(\beta_k - \beta_{k0}\right) \tag{3.1}$$

$$f_i^0 = f\left(X_i, \beta_0\right)$$

$$\alpha_k^0 = \left(\beta_k - \beta_{k0}\right) \tag{3.2}$$

$J_{ik}^0 = \left[\frac{\partial f(X_i,\beta)}{\partial \beta_k}\right]_{\beta=\beta_0}$ is $i \times n$ jacobian matrix.

The residual between the estimated and real values is:

$$e_i = Y_i - f_i^0 = \sum_{k=1}^{n} \alpha_k^0 J_{ik}^0 + \varepsilon_i \tag{3.3}$$

The linear model in equation (3.1) is assumed valid only around some specific operating point. While ε is the assumed white noise with zero mean and constant variance, and the initial value of parameter α_0 can be estimated by linear least squares method.

$$Y_0 = J_0\alpha_0 + \varepsilon \tag{3.4}$$

$$\widehat{\alpha}_0 = (J_0^{'}J_0)^{-1}J_0^{'}e \tag{3.5}$$

From equation (3.2), $\beta_1 = \alpha_0 + \beta_0$. The next step is replacing β_0 by β_1 in equation (3.1), where β_1 represents a new initial value for the system. Repeat the same steps for $[\beta_2 \ \ \beta_3 \ \beta_4, \ \ldots\ldots.. \ \beta_m]$, where m is the number of required iterations to get the convergence. The number of iterations m will terminate when the convergence ratio $|(\alpha_{k,m+1} - \alpha_{km})/\alpha_{km}| < \delta$ meets some pre-specified threshold (specific small value for δ) for example when the value less than 1.0×10^{-6} [84].

3.3 Gauss-Newton Algorithm for Missing Data

The above procedures are called a Gauss-Newton iteration method for nonlinear regression. Unfortunately, this technique cannot be used to estimate the parameters if the data contains missing values, because it depends on the error between the estimated and measured values. If there is a missing value on the regression variable, it is not possible to estimate the error. In this case, another optimization technique have been used to estimate the error and taking it as an initial value in the Gauss-Newton iteration technique. This approach shows that the combination of EM and Gauss-

Newton produces better results in comparison with linear analysis methods. To illustrate this, consider the same example taken from Table 2.1 [32].

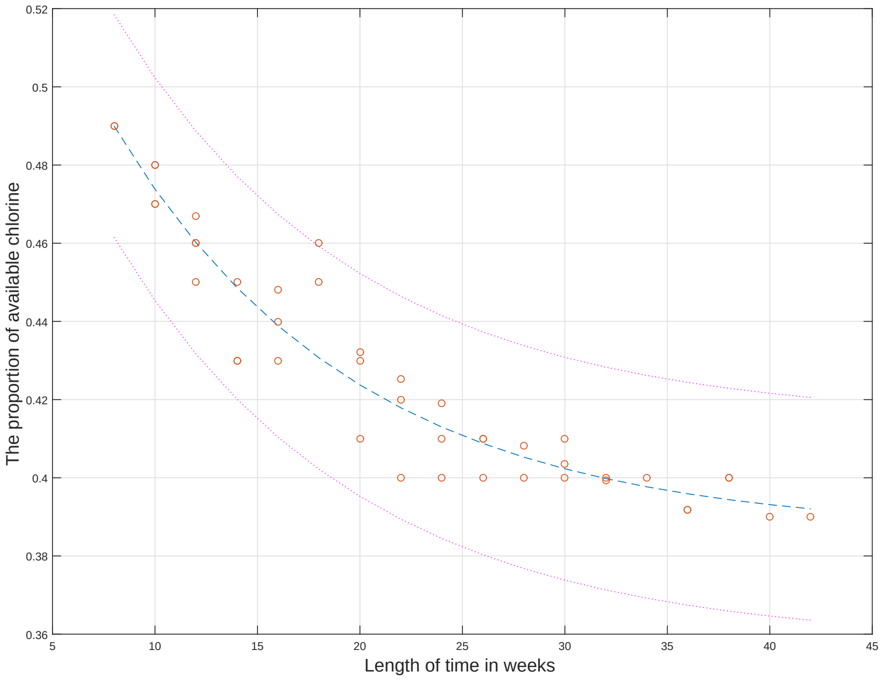

Figure 3.1: Nonlinear scatterplot of the proportion of available chlorine in a certain quantity of chlorine solution (MAR).

First, a nonlinear exponential growth model is used to fit the data $\hat{Y} = \theta_1(y_1 - \theta_1)e^{\theta_2(X - x_1)}$,

where x_1 and y_1 represent the first two initial values in the data set . The values generated by the estimated nonlinear model are shown in Figure 3.1. Comparing Figure 3.1 with Figures 2.4 and 2.5, there is similarity between the linear estimation and the nonlinear estimation. This slight modification to the nonlinear algorithm for missing data yields an unbiased parameter estimation in the MAR case. Notice that the nonlinear regression model yields a correlation -0.86 between the output $\hat{Y}$ and input X. In contrast the correlation equals to -0.94 for the complete data case. Consequently, the nonlinear model produces less variability. For example, the standard deviation of the output $\hat{Y}$ estimated from the nonlinear model is 0.029, whereas it is equal to 0.025 for the complete data case. Although the nonlinear regression model gives unbiased estimation of standard deviation and correlation, it does produce biased estimates of the mean value. In the above example, an exponential growth model was used to fit the data, showing some disadvantage in comparison with linear models.

To expanding this, consider another data set as shown in Table 3.1 (taken from [84]). In this example, the dependent variable $\hat{Y}$ is the tensile strength of Kraft paper and the independent variable X is the hardwood concentration for pulp, which produces the paper. The data set includes the following missing data mechanism MCAR with 21%, 26% and 37% missing. Note that unlike the previous examples, the ultimate purpose of this example is to compare the performance of the linear algorithm (EM algorithm) and the nonlinear algorithm (modified Gauss-Newton algorithm) in the presence of different percentages of missing data for a MCAR mechanism in term of correlations, residuals, standard deviations, and means. For illustration, the complete data is plotted in Figure 3.2. The EM algorithm is applied to estimate the parameters of the linear model: $\hat{Y} = \theta_0 + \theta_1 X$.

The modified Gauss-Newton algorithm was applied to estimate the parameters of the polynomial model: $\hat{Y} = \theta_0 + \theta_1 X + \theta_1 X^2$.

To begin the analysis, compare the imputed values generated by the linear and nonlinear models in case of MCAR 21% missing data, with the complete data, where the mean value for full-observed output Y is 34.184, and mean value for the imputed values is 34.379 and 34.178, respectively.

This result indicates that the nonlinear regression is just slightly better than the linear model for mean value estimation. By inspecting Figure 3.3 and Figure 3.4, the effect of missing values on the proposed algorithms can be seen. In Figure 3.3, the imputed values from the linear model fall directly on a straight line with a slope 1.73. The same happens with the nonlinear model shown in Figure 3.4.

Table 3.1: The input and output of the system in MCAR with missing percentage [84].

Percentage of hardwood	Strength of paper			
	Y			
X	Complete	MCAR (21%)	MCAR (26%)	MCAR (37%)
1	6.3	6.3	6.3	6.3
1.5	11.1	11.1	11.1	11.1
2	20	-	20	-
3	24	24	24	24
4	26.1	26.1	26.1	26.1
4.5	30	-	-	-
5	33.8	33.8	33.8	33.8
5.5	34	-	34	-
6	38.1	38.1	-	38.1
6.5	39.9	39.9	39.9	-
7	42	42	-	42
8	46.1	46.1	46.1	-
9	53.1	53.1	53.1	-
10	52	52	-	52
11	52.5	52.5	52.5	52.5
12	48	48	-	-
13	42.8	-	42.8	42.8
14	27.8	27.8	27.8	27.8
15	21.9	21.9	21.9	21.9

The imputed values with linear and nonlinear regression have a correlation 0.54684 and 0.53117, respectively between the imputed output $\hat{Y}$ and input X whereas the case of complete data with a correlation 0.55261. Figure 3.5 and Figure 3.6 show the effect of uncorrelated cases in terms of the residuals. For example, linear and nonlinear models give standard deviation estimates of 14.00 and 13.61, respectively. Whereas the full observed data standard deviation is 13.778. This is not surprising, because the missing values are close to the linear region area.

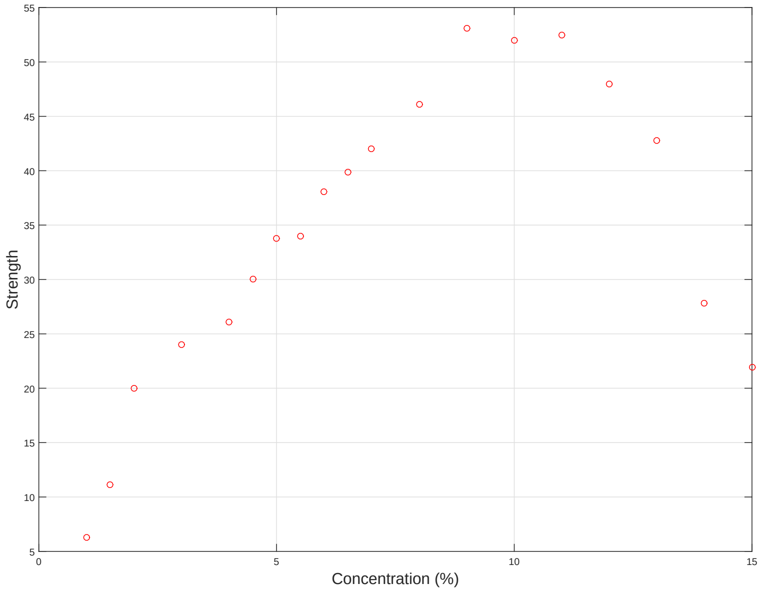

Figure 3.2: Complete Concentration/Strength data scatterplot.

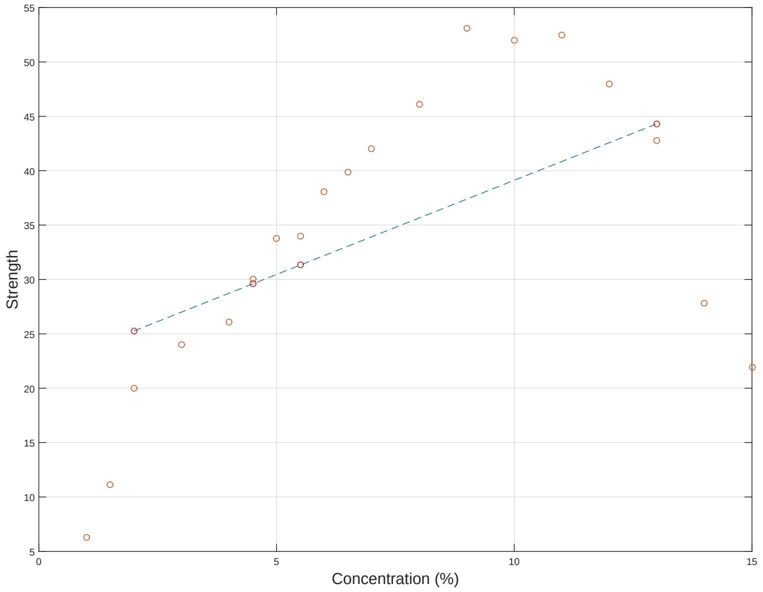

Figure 3.3: Linear regression model of Concentration/Strength data in case of 21% (MCAR) scatterplot.

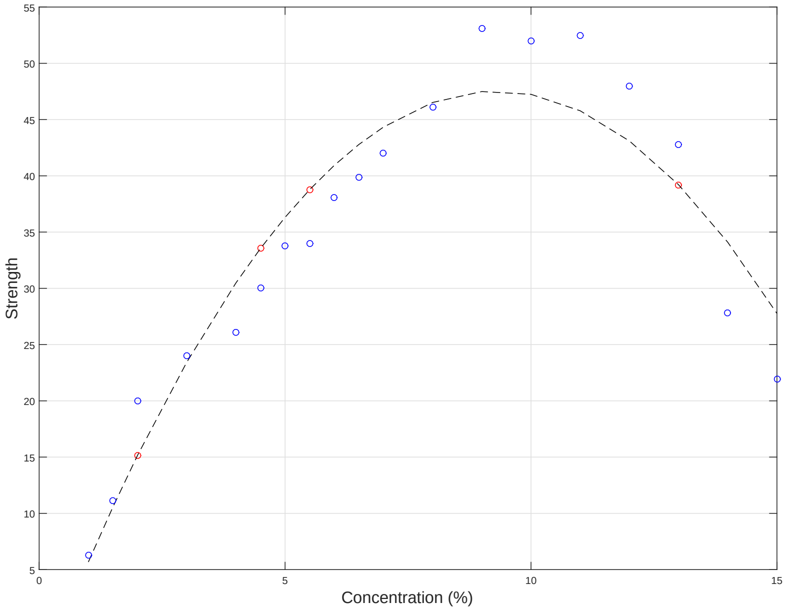

Figure 3.4: Noninear regression model of Concentration/Strength data in case of 21% (MCAR) scatterplot.

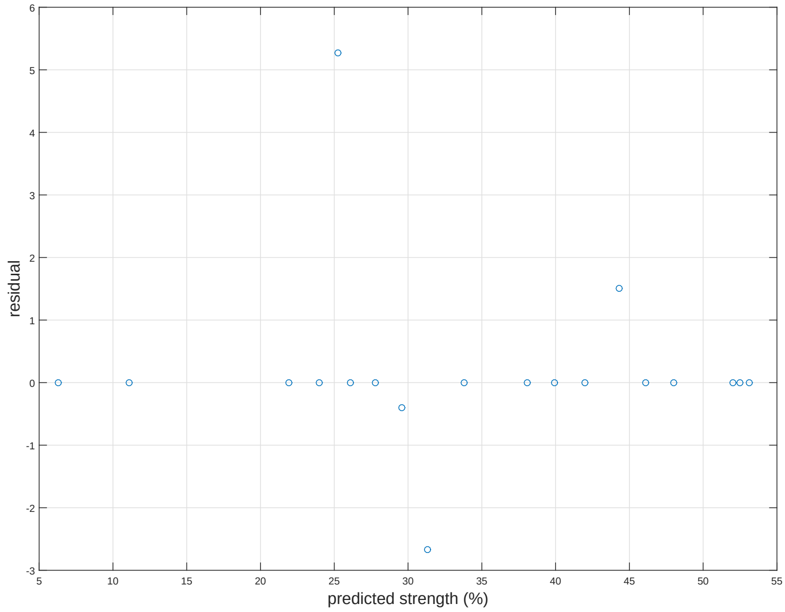

Figure 3.5: (Linear regression model) residual (e) versus predicted values scatterplot in case of 21% (MCAR).

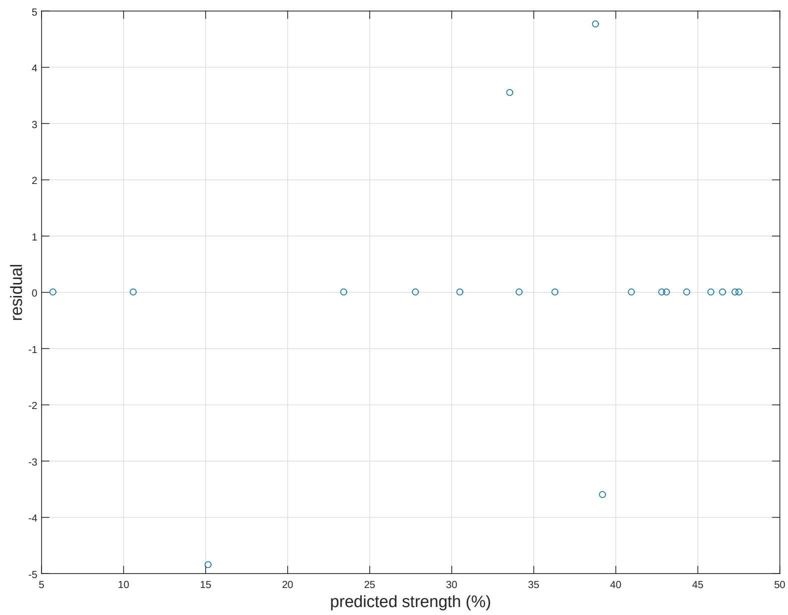

Figure 3.6: (Nonlinear regression model) residual (e) versus predicted values scatterplot in case of 21% (MCAR).

Table 3.2, summarizes the effect of two other cases of missing data percentages (MCAR 26% and MCAR 37%) for the linear and nonlinear model. Relevant results are graphically illustrated in Figures 3.7-3.14.

Table 3.2: The effect of linear and nonlinear models on the system in different MCAR missing percentage.

	Linear regression		
	MCAR 21%	MCAR 26%	MCAR 37%
Mean	34.379	31.744	31.106
Correlation	0.5468	0.543	0.5541
Standard deviation	13.61	13.778	13.778
	Nonlinear regression		
Mean	34.178	30.945	31.426
Correlation	0.5312	0.5148	0.5372
Standard deviation	14.01	12.438	12.356

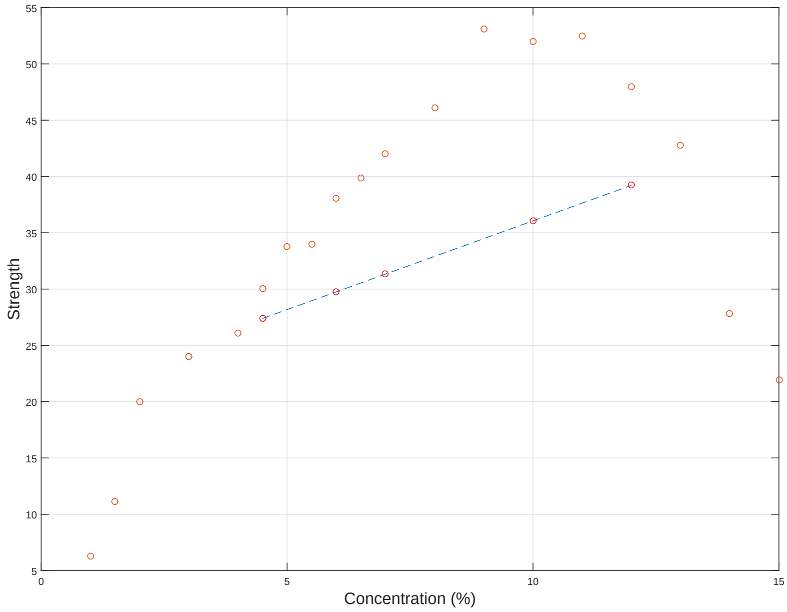

Figure 3.7: Linear regression model of Concentration/Strength data in case of 26% (MCAR) scatterplot.

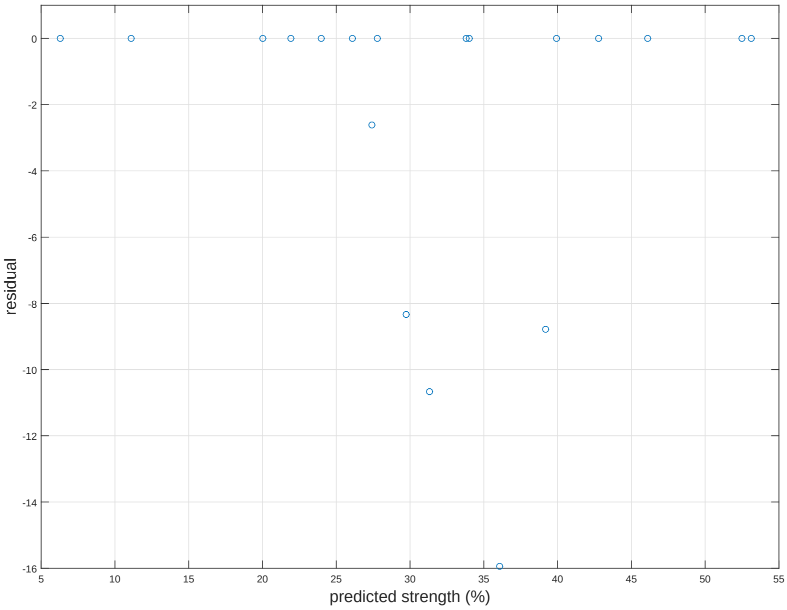

Figure 3.8: (Linear regression model) residual (e) versus predicted values scatterplot in case of 26% (MCAR).

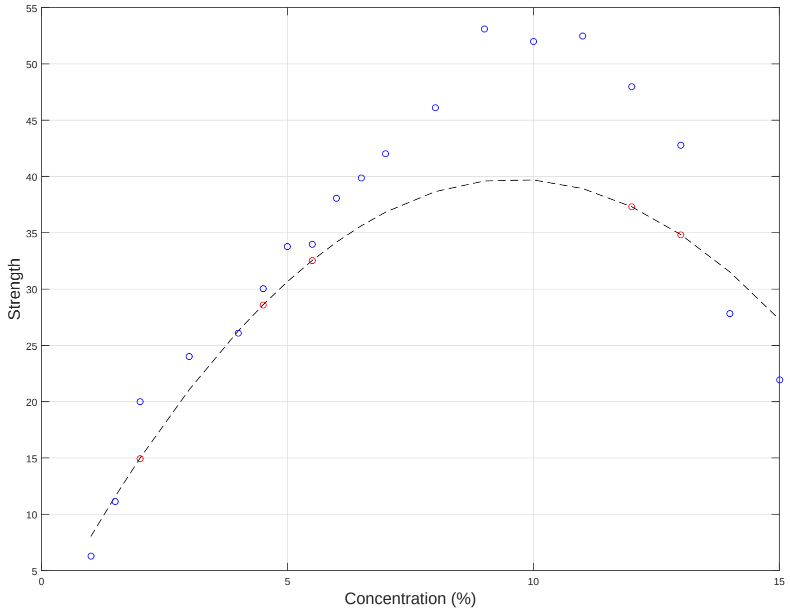

Figure 3.9: Noninear regression model of Concentration/Strength data in case of 26% (MCAR) scatterplot.

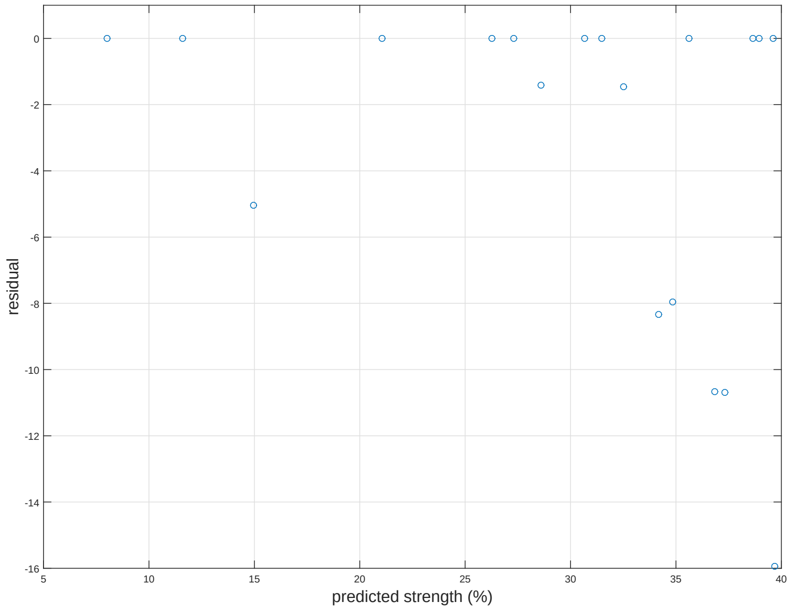

Figure 3.10: (Nonlinear regression model) residual (e) versus predicted values scatterplot in case of 26% (MCAR).

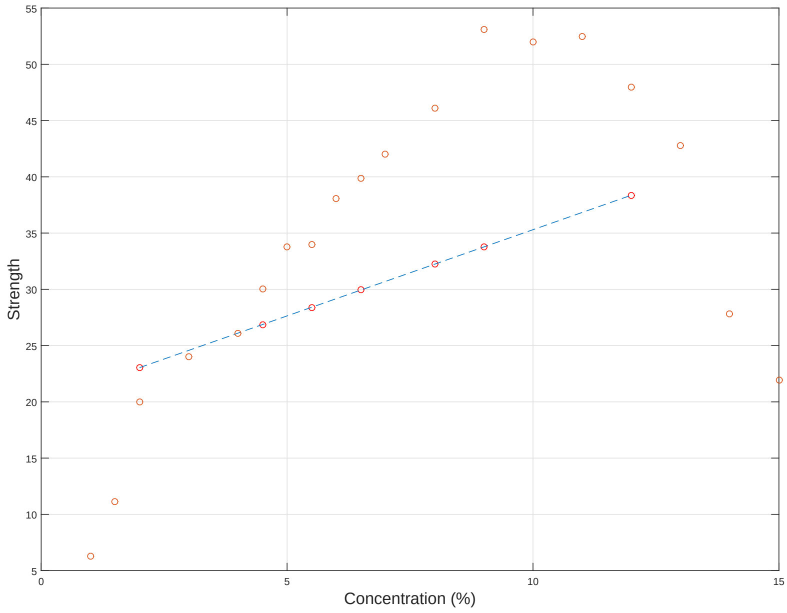

Figure 3.11: Linear regression model of Concentration/Strength data in case of 37% (MCAR) scatterplot.

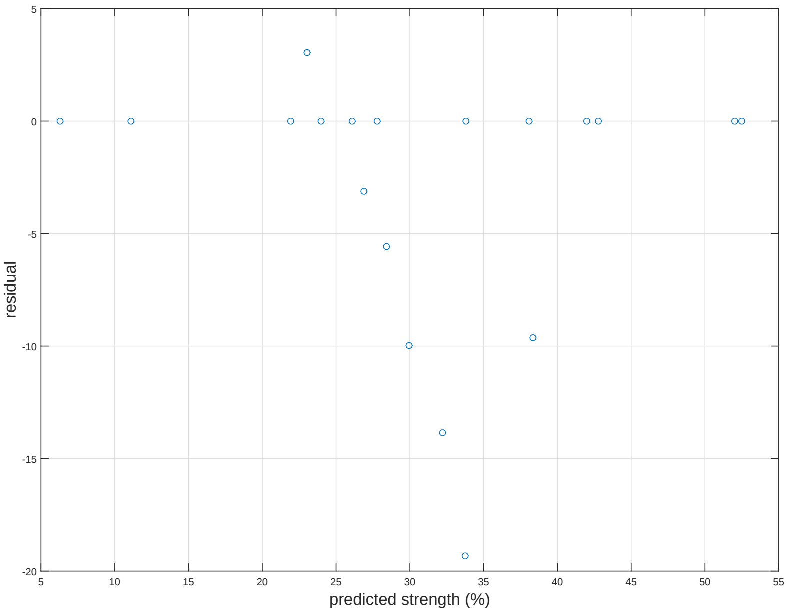

Figure 3.12: (Linear regression model) residual (e) versus predicted values scatterplot in case of 37% (MCAR).

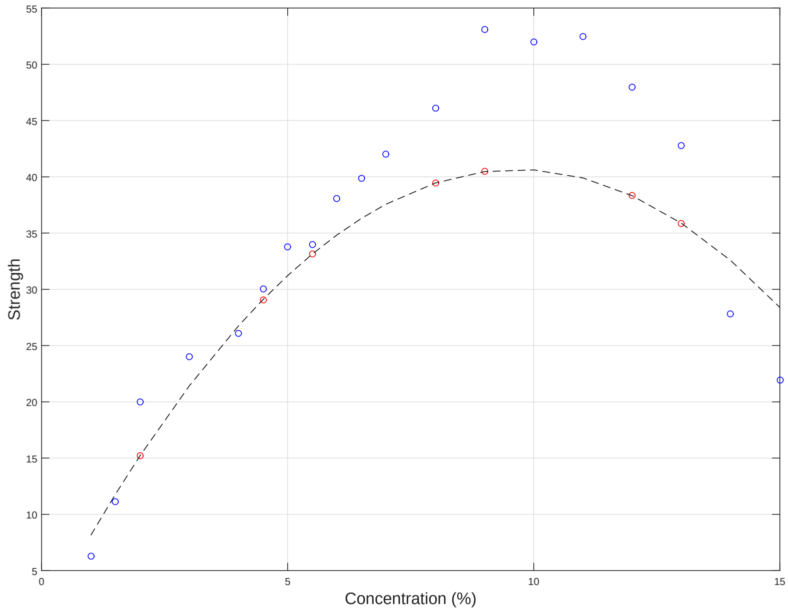

Figure 3.13: Noninear regression model of Concentration/Strength data in case of 37% (MCAR) scatterplot.

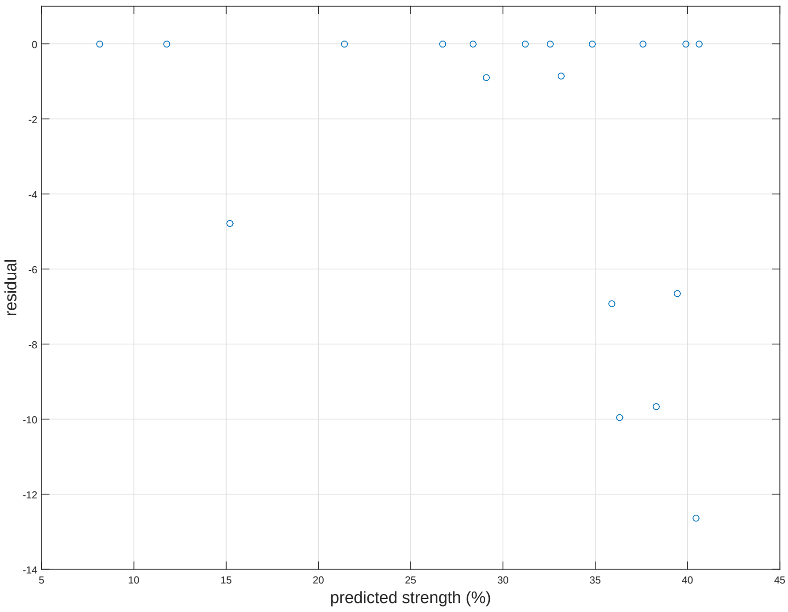

Figure 3.14: (Nonlinear regression model) residual (e) versus predicted values scatterplot in case of 37% (MCAR).

3.4 Model Selection

For linear model identification, the general formula of the linear model is known. It is only necessary to identify the key dependent and independent variables to be included in the model. While generally a simple and popular procedure for the estimation process is to assume a linear relationship between the predictor and dependent variables, this assumption may not always work well especially for severely nonlinear systems. To overcome the weakness of this assumption, researchers typically try using nonlinear analysis techniques, and this involves another challenging problem to select the best model from different candidate nonlinear models or desired nonlinear representations. Here the form of the model needs to be specified and the parameters need to be estimated in some iterative manner. Also, the initial values for those parameters must be provided [13, 16].

3.4.1 Box-Tidwell Method

Box and Tidwell introduced an iterative method for model selection [22]. It is based on calculating the best power for the polynomial model.

For the model

$$Y = \beta_0 + \beta_1 X^p + \varepsilon, \tag{3.6}$$

where β_0 and β_1 are parameters to be estimated, p is the power that needs to be determined and ε is uncorrelated white noise with zero mean and constant variance.

The Box-Tidwell transformation for a positive independent variable X in equation (3.6) is:

$$BT\,(X) = \begin{cases} \ln(X) & p = 0 \\ X^p & p \neq 0 \end{cases} . \tag{3.7}$$

The parameter p in equation (3.7) can be determined through an optimization algorithm by expanding the polynomial model in equation (3.6) using a Taylor series for p. The iteration process starts by calculating an initial value $p_{(1)}$ and iterating until $p_{(K)}$ converges [22].

3.4.2 Fractional Polynomial Model Estimation

Nonlinear regression often suffers from serious drawbacks, such as less flexibility in low order nonlinear systems (e.g. quadratic model), a lack of waviness in higher order systems, and the difficulties with model selection in specifying the relation between the input and output variables of the system [10, 81], the Fractional Polynomial (FP) method introduced by Royston and Altman gives a good solution to polynomial regression [98], and this can be achieved by finding the best model from a set of fractional polynomial models that describe the relationship between the input X and output Y. This section will illustrate the feature of this approach and how it can be used for missing data analysis.

To some extent, the fractional polynomial approach is similar to conventional polynomial based methods, where the polynomial regression has only positive integer powers of predictor variables. On the other hand, fractional polynomial methods allow non-positive integer powers, this permits the use of negative and fractional bases [98]. In many cases, fractional polynomial models give a better fit as compared with traditional polynomial models [99], and it representation is similar to traditional polynomial models.

For example, if the degree of the model $r = 1$ or 2, the model can be written, respectively as:

$$Y = \beta_0 + \beta_1 X^p, \tag{3.8}$$

$$Y = \beta_0 + \beta_1 X^{p_1} + \beta_2 X^{p_2}, \tag{3.9}$$

where p_1 *and* p_2 are powers of either integer or fractional values. The fractional polynomial model with degree r and a power vector $\mathbf{P}$ is denoted as $\phi_r\left(X, \mathbf{P}\right)$. Normally, the vector of powers, is restricted to a predefined set $\boldsymbol{s}$ as the following:

$$\boldsymbol{P} = [p_1, p_2, p_3, ...]$$

$$\boldsymbol{s} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$

This predefined set can adequately cover most practical models. This includes linear , quadratic , and cubic models, as well as other non-fractional and fractional polynomial models of degree r.

In practice, fractional polynomial models with power values up to 2 are sufficient and can give good results that are better than conventional polynomial models and the models with higher degrees are rarely used, and this is because of sensitivity to noise and small changes in data [17, 98, 100]. For this reason, it will be better to use the model family of first and second degrees $\phi_1(X, \mathbf{P})$ and $\phi_2(X, \mathbf{P})$.

An appropriate model can be selected from a predefined set of models $\{\boldsymbol{s}\}$, and all models can be estimated by using a maximum likelihood method. Each model has a power vector $\mathbf{P}$, which is associated with the model and likelihood function, and the power vector is used to calculated the deviance (D) for each model where

$$D = -2 \times \log l$$

.

The gain (G) for model $\phi_m(X, \mathbf{P})$ is defined as [98]:

$$G = G\left(\phi_m\left(X, \mathbf{P}\right)\right) = D\left(\phi_1\left(X, \mathbf{1}\right)\right) - D(\phi_m(X, \mathbf{P})), \tag{3.10}$$

where $D\left(\phi_1\left(X, \mathbf{1}\right)\right)$ and $D\left(\phi_m\left(X, \mathbf{P}\right)\right)$ are the deviance of linear fit and fractional polynomial model, respectively.

The gain (G) of each model is calculated from the deviance D of the model, defined as the difference of the deviance between the linear estimated fit $\phi_1(X, \mathbf{1})$ and fractional polynomial model $\phi_m(X, \mathbf{P})$. There is an opposite relationship between the values of D and G, the highest value of gain G and the lowest value of D results in a better fit. In general, the final procedure of model selection process depends on the appearance of the relationship between the fitted curve and data [98, 100].

3.4.3 Missing Data and Model Selection

The aforementioned model selection procedures can be used for the complete data case. Unfortunately, these techniques cannot be used directly for nonlinear model selection for the case of missing data, this is because model parameter estimation requires nonlinear least squares, and this would need the error value between the model fitted values and the real observations. If there is a missing value on the dependent variables, it is impossible to know this error directly. To overcome this problem, the combined EM-

Gauss-Newton algorithm will be used.

As an example, consider a data set, which was taken from [78]. This data represents the body mass index (BMI) which is the input X and percentage of body fat content which is the output Y, the total number of data is 327 that were taken from three different countries [78].

In this data set, two different cases of MCAR missing data mechanisms was proposed, 10% and 20% missing, respectively. The ultimate goal of this example is to examine the performance of these model selection methods in the presence of different percentages of missing data for a MCAR mechanism, and compare it with traditional models (linear, quadratic, cubic) and the Box Tidwell model selection technique.

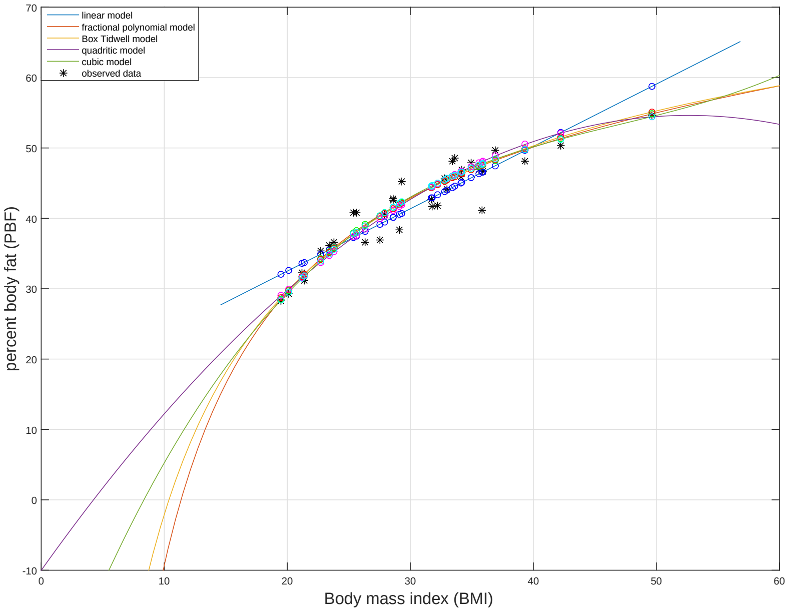

Figure 3.15: Fitted lines for models in case of 10% MCAR imputed missing data and real values.

By using Box Tidwell method for estimating the best fitting for the BMI data, among the huge number of models, the model with power p=-0.84 is the best one, having the form $Y = \beta_0 + \beta_1 X^{-0.84}$. The best fit solution from the fractional polynomial method has model order $(-2, -1)$, which has the form $Y = \beta_0 + \beta_1 X^{-2} + \beta_2 X^{-1}$, both of the two models were estimated from complete data. On the other hand, in the case of missing data, the Box Tidwell method gave two different models, with powers -0.62 and -0.5 for both cases 10% and 20% MCAR, respectively, and the fractional polynomial method generates models with power vectors $(-1, 2)$, and $(2, -1)$ for both cases 10% and 20% MCAR, respectively.

For a clearer visualization, the imputed data in case of 10% and 20% MCAR and real values are shown in Figures 3.15 and 3.16, respectively. The change of the amount of missing data affected directly on the model order in both model selection techniques, the imputed data from the proposed models are similar but the models that are proposed by the model selection techniques still have the best fit in both cases of missing data. The next section will provide more evidence concerning this.

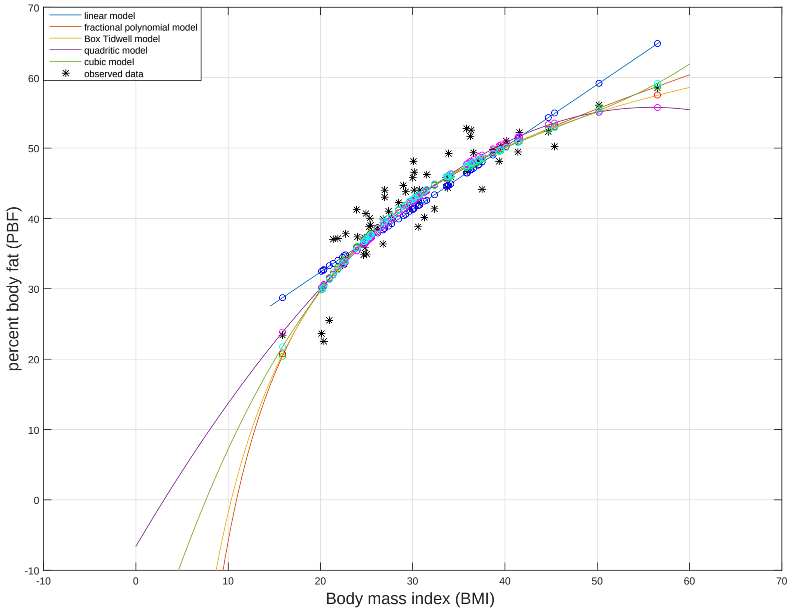

Figure 3.16: Fitted lines for models in case of 20% MCAR imputed missing data and real values.

3.4.4 Goodness of Model Fit

To ensure the optimum model is selected, a goodness fit of the models should be checked. There are many types of goodness fit tests that can be used to check the model performance [40, 56]. In this section, two simple and robust measures was used R^2 and an F test. Table 3.3 summarizes the comparison of the proposed models with three traditional models: cubic, linear, and quadratic. In terms of fit goodness, the R^2 values show little difference between the traditional models (quadratic and cubic) and the selected models of fractional polynomial and Box Tidwell.

Table 3.3: Proposed models and goodness fit statistics.

Model	R^2 (MCAR 10%)	R^2 (MCAR 20%)
Linear	0.8073	0.819
Fractional polynomial	0.8689	0.8781
Box Tidwell	0.8686	0.8779
Quadratic	0.8607	0.871
Cubic	0.867	0.8764

To implement the F test, first calculate the F-statistic, and this depends on the degree of freedom for each model. The F statistic must be determined by one of two equations:

$$F - statistic_1 = \frac{RS_1}{RS_2} \tag{3.11}$$

$$F - statistic_2 = \frac{(RS_1 - RS_2)\, DF2}{(DF1 - DF2)RS_2} \tag{3.12}$$

If both models have the same degree of freedom, then equation (3.11) is used. Otherwise equation (3.12) is used, where RS_1 is the squares sum of residual for the first model and RS_2 is the squares sum of residual for second model. $DF1$ and $DF2$ are the degree of freedom of the first and second models, respectively. After determining the F-statistic, the results can be compared with F-distribution value to extract the probability value (γ). If the $(1-\gamma)$ value is less than 0.05 (*Rejection-probability-value*) then the first model has a better fit of data, otherwise the second model is better.

Table 3.4 summarizes the results of the F-test for nonlinear models (Box Tidwell, quadratic, fractional polynomial, cubic) for the two cases of missing percentages. The F-test results show that traditional models have low F-distribution values in comparison with Box Tidwell and fractional polynomial models. The F-test results indicate that Box Tidwell still gives a better fit.

Table 3.4: F-test.

	10% MCAR	20% MCAR
Model	F-statistic	F-statistic
Box Tidwell-FP	0.99984	0.99999
Box Tidwell-quadratic	1.00000	1.00000
Box Tidwell-cubic	1.00000	1.00000
Quadratic-FP	0.75243	0.85552
Quadratic-cubic	0.77245	0.81225
FP-cubic	1.00000	1.00000

3.5 Summary

Missing data analysis plays a key role in real life data based decision making and related fields of research. The primary aim of this chapter was to introduce a nonlinear modelling technique for missing data analysis (static data). Comparative study on EM-Gauss-Newton approach has been demonstrated, EM and Gauss-Newton algorithm are advantageous over traditional approaches. In addition, in this chapter, the critical issues in choosing the best models in case of missing data was introduced, where two most popular model selection methods for incomplete data are illustrated, and the illustrations have been focused on single variable data modelling for missing data. The basic idea however can be extended to multivariable data analysis, but the modelling complexity is increasingly difficult. The key aspects of the Box Tidwell transformation and fractional polynomial methods was presented and applied these to model estimation for missing data. The comparison of the effect of different missing data mechanism (10% MCAR and 20% MCAR) on the fractional polynomial, Box Tidwell and traditional models gave good indications about the use of fractional polynomial and Box Tidwell methods. As evidenced by the F-test, the cubic, Box Tidwell, and fractional polynomial models are better and imputed the missing values about equally well, but the fractional polynomial model still give the highest R^2 value. However, complex models are generally less tractable and are less robust than simple ones.

Chapter 4

Handling Missing Data in Multivariate Time Series Using a Vector Autoregressive Model-Imputation

4.1 Introduction

Datasets involving multivariate time series are present in nearly every scientific field. Examples include economic, engineering, medicine, science, finance, and climatology [34, 44, 74]. Problems with missing data routinely occur while conducting research in these fields especially with large datasets. This is particularly apparent during the data acquisition phase. However, modelling and analysis of most of multivariate time series datasets often require complete data. Therefore, missing data is a very serious problem, especially those involving multivariate time series data modelling and analysis. To correct the problem, care must be taken to impute missing data with reasonable and accurate values to ensure valid models and accurate study results.

Within the research field of missing data analysis, traditional data imputation methods can appropriately handle missing values for static data. Approaches such as multiple imputation (MI) and maximum likelihood (ML) featured in standalone software (e.g., NORM; [49, 103]) or statistical packages (e.g., SPSS and MARSS) can easily impute good values for missing data. However, imputing values into multivariate time series presents special challenges, and these software packages cannot handle missing data for dynamic systems modelling especially when the data is missing at random [48], these packages have limitations or simply may not work for dynamic systems modelling [73]. For example, many dynamic models involving autoregressive variables produce outputs that normally are linear or nonlinear combinations of lagged variables, and the estimation of autoregressive models requires that the data be fully observable. When these autoregressive models have missing values present, estimation of the output is simply not possible [104]. Most statistical packages either do not allow missing data in time series analysis or only allow ad-hoc procedures limiting the options available. Examples include: the MARSS Package (Multivariate Autoregressive State Space) [61], and K-nearest neighbour method [18, 73]. Also, these methods often lead to bias in the output estimates.

Furthermore, most of these methods are used for static data sets and become increasingly difficult to implement when both the dependent and independent variables have missing information [5]. For this reason, it may not be appropriate to directly apply these methods to deal with dynamic models. Until now, only a limited number of algorithms have been adapted to be used for missing data imputation for cases of multivariate time series [28]. While these methods can handle some situations of missing data in multiple variable modelling (static data), they still lack robustness for multivariate time series modelling tasks [85]. Therefore, there is a need to address this issue, this chapter introduces a new methods to improve this situation and presents a suitable solution for missing data imputation in multivariate datasets.

4.2 Vector Autoregressive Model (VAR)

The vector autoregressive model (VAR) is a commonly used model for the analysis of multivariate time series. In many applications, where the variables of interest are linearly related to each other, the VAR model has shown to be a good choice for representing and predicting the behaviour of dynamic multivariate time series [121]. It primarily provides good forecasts as compared to models from univariate time series and others . The forecasts from the VAR are relatively easy to derive because the model can make conditions on the prediction paths of specified time series within the model itself [121]. In addition to time series analysis and prediction, the VAR model is additionally utilized for causality inference and strategy investigation of the multiple time series. In causality analysis, specific hypotheses of the causality of the time series under analysis are assumed, and the subsequent causal effects of each time series are outlined. This chapter concentrates on the use of the VAR model to analyse stationary multiple time series datasets with missing data.

4.3 The VAR Model for Stationary Time Series

Let $\boldsymbol{X_t} = [x_{1t},\ x_{2t},\ ..\ .,\ x_{mt}\ ]^T$ be an $(m \times 1)$ time series vector, a $VAR(p)$ model for the multiple time series can be represented by:

$$\boldsymbol{X_t = A_0 + A_1X_{t-1} + A_2X_{t-2} + \cdots + A_pX_{t-p} + \varepsilon_t},$$

$$\boldsymbol{X_t = A_0 + \sum_{i=1}^{p} A_iX_{t-i} + \varepsilon_t}, \tag{4.1}$$

where $t = 1,\ \ldots,\ T,\ \boldsymbol{A_i}$ are $(m \times p)$ coefficient matrices and $\boldsymbol{\varepsilon_t} \in (0, \Sigma)$ denotes an $(m \times 1)$ vector of white noise.

Equation (4.1) can be written in lagged notation:

$$\boldsymbol{A_p\,(L)\,X_t = A_0 + \varepsilon_t},$$

where $\boldsymbol{A_p\,(L) = I_m - A_1L - \cdots - A_pL^p}$,

For a stationary multivariate time series the mean (M) satisfies:

$$\boldsymbol{M = inv(I_m - A_1 - \cdots - A_p)A_0},$$

and the mean-adjusted form for $VAR(p)$ model is:

$$\boldsymbol{X_t - M = A_1\,(X_{t-1} - M) + A_2\,(X_{t-2} - M) + \cdots + A_p\,\left(X_{t-p} - M\right)},$$

The stability of the VAR model is dependent on the roots of equation (4.2), and $(z_1, z_2, z_3, ...)$ are eigenvalues of $\boldsymbol{A}$.

$$|\boldsymbol{A} - \boldsymbol{zI_m}| = 0. \tag{4.2}$$

4.3.1 VAR (p) Model Estimation

This section briefly reviews the least squares estimation technique for estimating $VAR(p)$ model coefficients in equation (4.1).

In many cases, the coefficients matrices $\boldsymbol{A_0},\ \boldsymbol{A_1},\ \boldsymbol{A_2},\ \ldots,\ \boldsymbol{A_p}$ are unknown, and need to be estimated from the available multivariate data set.

It is assumed that the entire time series $x_{1t},\ x_{2t},\ \ldots,\ x_{mt}$ data set is available (no missing data). Hence, the sample size for the all-time series are same: $t = 1,\ \ldots,\ T$. Furthermore, the specified p lagged values for each time series X_{t-p} are assumed to be exist.

For the m time series with sample length T ($t = 1,\ \ldots,\ T$), the $VAR(p)$ model is written as [79]:

$$\hat{\boldsymbol{X}}_t = \boldsymbol{\phi A} + \boldsymbol{e} \tag{4.3}$$

where $\boldsymbol{e}$ is error with covariance matrix $\sigma^2 \boldsymbol{I_m}$, $\boldsymbol{\phi}$ is the regression matrix and $\boldsymbol{A}$ is the coefficients matrix

$$\boldsymbol{A = (\phi^T \phi)^{-1} \phi^T X}$$

Then,

$$\boldsymbol{X} = \begin{pmatrix} x_1(p+1) & x_2(p+1) & \cdots & x_m(p+1) \\ x_1(p+2) & x_2(p+2) & \vdots & x_m(p+2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(T) & x_1(T) & \cdots & x_m(T) \end{pmatrix} ((T-p) \times m)$$

$$\boldsymbol{A} = \begin{pmatrix} a_{01} & a_{(11)1} & \cdots & a_{(1m)p} \\ a_{02} & a_{(21)1} & \vdots & a_{(2m)p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{0m} & a_{(m1)1} & \cdots & a_{(mm)p} \end{pmatrix} ((mp+1) \times m)$$

$$\boldsymbol{\phi} = \begin{pmatrix} 1 & x_1(p) & \cdots & x_m(p) & x_1(p-1) & \cdots & x_m(p-1) & x_1(1) & \cdots & x_m(1) \\ 1 & x_1(p+1) & \vdots & x_m(p+1) & x_1(p) & \vdots & x_m(p) & x_1(2) & \vdots & x_m(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(T-1) & \cdots & x_m(T-1) & x_1(T-2) & \cdots & x_m(T-2) & x_1(T-p) & \cdots & x_m(T-p) \end{pmatrix}$$

$$((T-p) \times (mp+1))$$

4.3.2 Model Order Selection

The $VAR(p)$ model selection is usually accomplished by specifying the model selection criteria. The basic idea is to identify models with different lag values $\boldsymbol{p} = \{0, 1, 2, \ldots.., p_{max}\}$ and select the p lag value that minimizes the model selection criteria [79]. A commonly used model order selection formula is represented by:

$$IC\,(p) = ln\left|\widehat{\sum}\,(p)\right| + S_T.\varphi(m,p)$$

where $\widehat{\sum}\,(p) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e_t e_t'}$ is the covariance matrix of the residual error $\boldsymbol{e}$. S_T is the indexed values sequence $\{1, \ldots, T\}$, and the penalty function $\varphi(m,p)$ which impedes the large model's order. The term $ln\left|\widehat{\sum}\,(p)\right|$, is a non-growing function while the $\varphi(m,p)$ function grows with the order of p, and the basic idea of the model order selection depends on balancing these two functions.

There are five techniques for model order selection in the applied $VAR(p)$ model literature generally broadly utilized:

- Akaike's information criterion (AIC) [4].

$$AIC\left(p\right) = ln\left|\widehat{\sum}\left(p\right)\right| + \frac{2}{T}pm^2,$$

where the penalizing function $\varphi\left(m,p\right) = pm^2$ and $S_T = \frac{2}{T}$.

- Schwarz criterion (SC) [106].

$$SC\left(p\right) = ln\left|\widehat{\sum}\left(p\right)\right| + \frac{lnT}{T}pm^2,$$

where the penalizing function $\varphi\left(m,p\right) = pm^2$ and $S_T = \frac{lnT}{T}$.

- Hannan-Quinn criterion (HQ) [54].

$$SC\left(p\right) = ln\left|\widehat{\sum}\left(p\right)\right| + \frac{2ln(lnT)}{T}pm^2,$$

For which the penalizing function $\varphi\left(m,p\right) = pm^2$ and $S_T = \frac{2ln(lnT)}{T}$.

For the previous three techniques, in each case the penalizing function $\varphi\left(m,p\right)$ has the same formula.

- Final Prediction Error (FPE) [3].

$$FBE\left(p\right) = \left[\frac{T+mp+1}{T-mp-1}\right]^m \left|\widehat{\sum}\left(p\right)\right|,$$

- Likelihood ratio test (LR test) [68].

$$LR\left(j\right) = m(ln\left|\widehat{\sum}\left(p-j\right)\right| - ln\left|\widehat{\sum}\left(p-j+1\right)\right|),$$

where $j = 1, 2, \ldots,\ (p-1)$

Other techniques are within the literature. However, they were not mentioned here, because they are not widely used in the application of *VAR* models, for more details, see [94].

4.4 Forecasting with VAR (p) Model

Assume that there are two time series datasets with a length of sample data T. The objective is to predict their values as $\{T+1,\ T+2,\ \ldots,\ etc.\}$. For simplicity, assume the first order $VAR(1)$ model

$$x_1(t) = b_{10} + a_{(11)1}x_1(t-1) + a_{(12)1}x_2(t-1) + \varepsilon_{1t}$$

$$x_2(t) = b_{20} + a_{(21)1}x_1(t-1) + a_{(22)1}x_2(t-1) + \varepsilon_{2t}$$

For one step prediction value $(t = T+1)$, the $VAR(1)$ model is

$$x_1(T+1) = b_{10} + a_{(11)1}x_1(T) + a_{(12)1}x_2(T) + \varepsilon_{1(T+1)}$$
$$x_2(T+1) = b_{20} + a_{(21)1}x_1(T) + a_{(22)1}x_2(T) + \varepsilon_{2(T+1)}$$

The conditional expectation value for both time series is

$$\begin{aligned} E\left(x_{1(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) &= b_{10} + a_{(11)1}x_1(T) + a_{(12)1}x_2(T) \\ &\quad + E\left(\varepsilon_{1(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) \end{aligned} \tag{4.4a}$$

$$\begin{aligned} E\left(x_{2(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) &= b_{20} + a_{(21)1}x_1(T) + a_{(22)1}x_2(T) \\ &\quad + E\left(\varepsilon_{2(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) \end{aligned} \tag{4.4b}$$

The expectation values $E\left(\varepsilon_{1(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right)$ and $E\left(\varepsilon_{2(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right)$ must be zero. In the forecasting process equation (4.4) become

$$F\left(x_{1(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) \equiv \hat{x}_{1(T+1|T)} = \hat{b}_{10} + \hat{a}_{(11)1}x_1(T) + \hat{a}_{(12)1}x_2(T) \tag{4.5a}$$

$$F\left(x_{2(T+1)} \,\middle|\, x_{1(T)}, x_{2(T)}\right) \equiv \hat{x}_{2(T+1|T)} = \hat{b}_{20} + \hat{a}_{(21)1}x_1(T) + \hat{a}_{(22)1}x_2(T) \tag{4.5b}$$

The most important term in equation (4.5) is the prediction error:

$$x_{1(T+1)} - \hat{x}_{1(T+1|T)} = \left(b_{10} - \hat{b}_{10}\right) + \left(a_{(11)1} - \hat{a}_{(11)1}\right) x_1\left(T\right) + (a_{(12)1} - \hat{a}_{(12)1}) x_2\left(T\right)$$

$$x_{1(T+1)} - \hat{x}_{2(T+1|T)} = \left(b_{10} - \hat{b}_{10}\right) + \left(a_{(21)1} - \hat{a}_{(21)1}\right) x_1\left(T\right) + (a_{(22)1} - \hat{a}_{(22)1}) x_2\left(T\right)$$

The prediction error is approximately zero if the estimated coefficients in equation (4.5) are consistent and the white noise ε_t is uncorrelated. The variance of the prediction error is [79]:

$$var\left(x_{1(T+1)} - \hat{x}_{1(T+1|T)}\right) = \sigma^2_{\varepsilon_{1(t)}}$$
$$var\left(x_{1(T+1)} - \hat{x}_{2(T+1|T)}\right) = \sigma^2_{\varepsilon_{2(t)}}$$

Similarly for two steps ahead ($t = T + 2$):

$$E\left(x_{1(T+2)} \;\middle|\; x_{1(T+1)}, x_{2(T+1)}\right) = b_{10} + a_{(11)1} E\left(x_{1(T+1)} \;\middle|\; x_{1(T)}, x_{2(T)}\right)$$
$$+ a_{(12)1} E\left(x_{2(T+1)} \;\middle|\; x_{1(T)}, x_{2(T)}\right)$$

$$E\left(x_{2(T+2)} \;\middle|\; x_{1(T+1)}, x_{2(T+1)}\right) = b_{20} + a_{(21)1} E\left(x_{1(T+1)} \;\middle|\; x_{1(T)}, x_{2(T)}\right)$$
$$+ a_{(22)1} E\left(x_{2(T+1)} \;\middle|\; x_{1(T)}, x_{2(T)}\right)$$

$$F\left(x_{1(T+2)} \;\middle|\; x_{1(T)}, x_{2(T)}\right) \equiv \hat{x}_{1(T+2|T)} = \hat{b}_{10} + \hat{a}_{(11)1} \hat{x}_{1(T+1|T)} + \hat{a}_{(12)1} \hat{x}_{2(T+1|T)}$$
$$F\left(x_{2(T+2)} \;\middle|\; x_{1(T)}, x_{2(T)}\right) \equiv \hat{x}_{2(T+2|T)} = \hat{b}_{20} + \hat{a}_{(21)1} \hat{x}_{1(T+1|T)} + \hat{a}_{(22)1} \hat{x}_{2(T+1|T)}$$

$$var\left(x_{1(T+2)} - \hat{x}_{1(T+2|T)}\right) = \left(1 + (a_{(11)1})^2\right)\sigma^2_{\varepsilon_{1t}} \\ +(a_{(12)1})^2\sigma^2_{\varepsilon_{2t}} + 2a_{(12)1}a_{(11)1}\sigma^2_{\varepsilon_{1,2t}} \tag{4.6a}$$

$$var\left(x_{2(T+1)} - \hat{x}_{2(T+2|T)}\right) = \left(1 + (a_{(22)1})^2\right)\sigma^2_{\varepsilon_{2t}} \\ +(a_{(21)1})^2\sigma^2_{\varepsilon_{1t}} + 2a_{(21)1}a_{(22)1}\sigma^2_{\varepsilon_{1,2t}} \tag{4.6b}$$

For X_t multivariate time series data set and $VAR(p)$ model, the $n - step$ predictions can be calculated utilizing the chain rule of prediction as;

$$\hat{X}_{(T+n|T)} = \hat{B}_0 + \hat{A}_1\hat{X}_{(T+n-1|T)} + \hat{A}_2\hat{X}_{(T+n-2|T)} + \cdots + \hat{A}_p\hat{X}_{(T+n-p|T)},$$

and the $n - step$ prediction errors can be written as

$$X_{T+n} - \hat{X}_{(T+n|T)} = \sum_{r=0}^{n-1} \Psi_r\varepsilon_{(T+n-r)}$$

With $\Psi_0 = I_m$ and $\hat{A}_j = 0$ for $j > p$, the Ψ_r matrices are calculating as

$$\Psi_r = \sum_{j=1}^{p-1} \Psi_{r-j}\hat{A}_j$$

where the expectation values for the prediction error is zero, the mean squares error (MSE) matrix for the $\hat{X}_{(T+n|T)}$ is

$$MSE\left(X_{T+n} - \hat{X}_{(T+n|T)}\right) = \Psi_r = \sum_{r=0}^{n-1} \Psi_r \sum \Psi'_r \tag{4.7}$$

where $\sum$ is the covariance matrix.

It also can be seen that as the steps to prediction increases, the complexity of calculating the variance also increases. Equation (4.7) becomes more complex, if the number of time series (m) increases and the order of the model (p) becomes larger. However, by using powerful modern software packages such as Eviews, Stata and MATLAB, this task becomes straightforward.

4.5 Goodness of VAR (p) Model

When a model has been developed to represent a multivariate time series data set, its structure and parameters need to be validated by testing the model behaviour, and testing the goodness of $VAR(p)$ models can be implemented through a wide range of techniques, forecasting accuracy is usually an intuitive method of validating a model. However, one-step-ahead prediction techniques do not account for the accumulation of prediction errors, therefore other prediction methods are needed to validate a model [79].

A model is said to be good enough, if it can predict, not only the observed data which is used for the estimation process, but also other unseen experimental data. Therefore, when commencing a modelling task, it is prudent to split the available experimental data into two sets: the training data that is used for the estimation process and the test data, which is used for the final assessment of the model estimation performance, this process is called cross-validation.

For a more reliable cross-validation process, a simulation prediction is used, where the mean of squared error (MSE) is computed to assess the model performance (using (4.7)).

The other simple and useful method is using the R^2 statistic, this method measures the success of the regression in predicting the values of the dependent variable within the sample. In standard settings, R^2 may be interpreted as the fraction of the variance of the dependent variable explained by the independent variables.

From equation (4.3), for multivariate time series X_t the R^2 statistic is computed as:

$$R^2 = 1 - \frac{e^{'}e}{(X-\overline{X})^{'}(X-\overline{X})},$$

where

$$\overline{X} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t,$$

One issue with using the R^2 statistic as a test of the goodness of $VAR(p)$ models is that R^2 will never reduce as more time series is added. In most cases, by including as many time series as sample observations, then the R^2 statistic is always 100% [38]. The adjusted R^2 , generally signified as $\hat{R}^2$, penalizes the R^2 for the addition of time series which do not contribute to the explanatory power of the model [92]. The adjusted R^2 is computed as:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-m}.$$

4.6 Granger Causality with VAR (p) Model

The $VAR(p)$ model is considered to be one of the most convenient formworks for testing the Granger causality. Based on the definition of Granger causality from Chapter 2 and equation (4.6), the Granger causality only implies prediction ability [121]. Now, assume two time series represented by the $VAR(p)$ model in equation (4.8), the first time series model, x_{1t}, has a linear relationship with its own previous measures and past measures of x_{2t}. x_{2t} Granger causality x_{1t} ($x_{2t}{\Rightarrow}x_{1t}$), if most of the past x_{2t} measures have non-zero impact: past x_{2t} affects x_{1t} depending on the impact of previous x_{1t}. Examining the Granger causality in equation (4.8) relies on the values of the coefficients $a_{(12)1} \ldots\ldots . a_{(12)p}$ and $a_{(21)1} \ldots\ldots\ldots a_{(21)p}$.

$$x_1(t) = b_{10} + a_{(11)1}x_1(t-1) + \cdots + a_{(11)p}x_1(t-p) + a_{(12)1}x_2(t-1) + \ldots + a_{(12)p}x_2(t-p) + \varepsilon_{1t} \tag{4.8a}$$

$$x_2(t) = b_{20} + a_{(21)1}x_1(t-1) + \cdots + a_{(21)p}x_1(t-p) + a_{(22)1}x_2(t-1) + \ldots + a_{(22)p}x_2(t-p) + \varepsilon_{2t} \tag{4.8b}$$

Therefore x_{2t} does not affect the Granger causality x_{1t} ($x_{2t} \nRightarrow x_{1t}$) if:

$$a_{(12)1} = a_{(12)2=} \ldots . = a_{(12)p} = 0$$

Similarly, x_{1t} does not affect the Granger causality x_{2t} ($x_{1t} \nRightarrow x_{2t}$) if:

$$a_{(21)1} = a_{(21)2=} \ldots . = a_{(21)p} = 0$$

Non-diagonal coefficients can result from four types of Granger Causality tests, as shown in Table 4.1.

Table 4.1: Granger Causality test.

	$a_{(21)1} = a_{(21)2} = \ldots . = a_{(21)p} = 0$ (Fail)	$a_{(21)1} = a_{(21)2} = \ldots . = a_{(21)p} = 0$ (Pass)
$a_{(12)1} = a_{(12)2} = \ldots . = a_{(12)p} = 0$ (Fail)	$x_{1t} \nRightarrow x_{2t}$ $x_{2t} \nRightarrow x_{1t}$	$x_{1t} \nRightarrow x_{2t}$ $x_{2t} \Rightarrow x_{1t}$
$a_{(21)1} = a_{(21)2} = \ldots . = a_{(21)p} = 0$ (Pass)	$x_{1t} \Rightarrow x_{2t}$ $x_{2t} \nRightarrow x_{1t}$	$x_{1t} \Rightarrow x_{2t}$ $x_{2t} \Rightarrow x_{1t}$

Note that the diagonal coefficients restrictions implied by Granger causality may be examined utilizing the Wald test [79].

There are many techniques for examining Granger causality which produce various results. For a two time series's $VAR(p)$ model, if the order of the model p changed, then the Granger Causality test yields different results and similarly, if the number of time series changes [80]. There are many software packages that can be used to implement the Granger causality tests such as S-Plus, Stata, and Eveiws.

4.7 Modified Listwise Deletion

The basic idea of the listwise deletion method is based on dropping all values, if there is just a single unknown value in at least one of the specified variables, this means that only cases with a complete data set can be used in the analysis. For a dynamic data set, the application of listwise deletion depends on ignoring time dependency, and this can lead to a significant standard error value, because in dynamic data the current value directly depends on the past value(s). The modified listwise deletion (MLD) technique is an extension of the listwise deletion technique, it aims to reduce the time dependency error in missing data imputation for multivariate time series. The application of the MLD is different from listwise deletion and pairwise deletion, it considers the first encountered complete case as the first measured case in the time series, without ignoring the cases that include missing values. That means MLD is a special case of the pairwise deletion technique. MLD utilizes a selected VAR model and a moving window approach to impute the missing values, based upon the previous observed values [45]. As a first step, the method starts by scanning the full data to specify the first case(s) with complete data at time t. Then uses the available complete cases for selecting the appropriate $VAR(P)$ model, the $VAR(P)$ model uses the observed data in the specified window $[t,\ t-n]$, where n represents number of complete cases which are used to impute one-step-ahead imputations. To examine the utility of the MLD, a simulation study was conducted to compare it with two other popular traditional methods, the mean imputation and listwise deletion.

4.7.1 Simulation Study

To compare the three methods, two time series $y_2(t)$ and $y_2(t)$ as shown in Figure 4.1 were simulated in MATLAB by using a first order VAR model with the general formula as shown in equation 5.8. To satisfy the causality conditions, the data sets were generated from a model with bidirectional effect between the two time series.

The length of time series is 200 time points. Each 10 time points represent one hour. The $MCAR$ mechanism was generated by randomly dropping values with three different proportions being 10%, 15% and 25%.

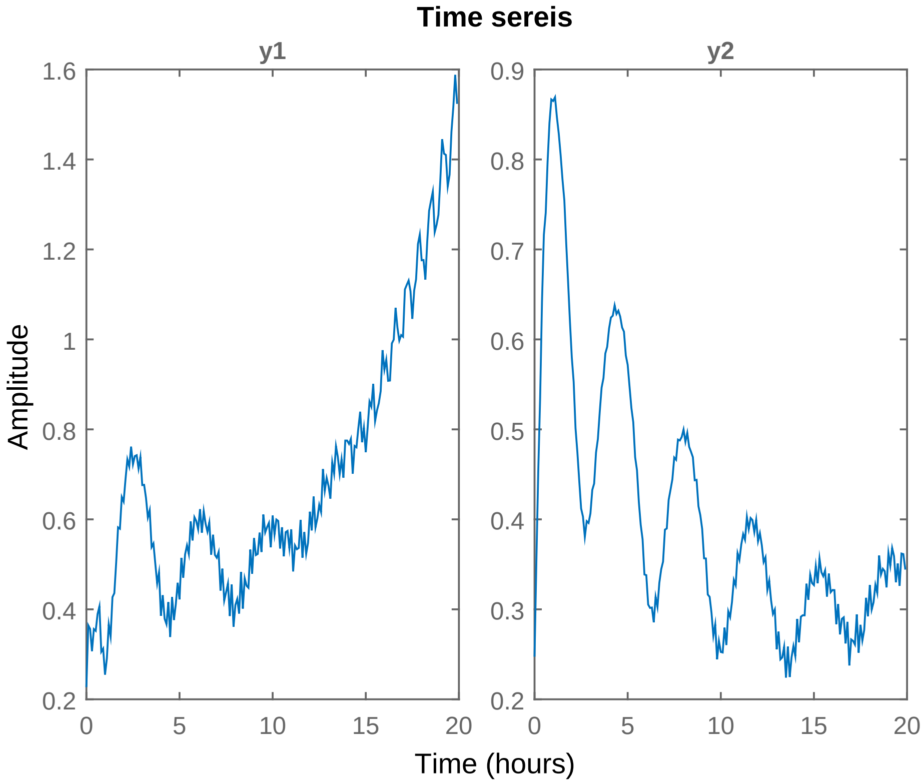

Figure 4.1: Time series with complete data.

To avoid the similarity of results, a different model other than the *VAR* model, was used for data estimation to examine the performance of the proposed technique, where this simulation study aimed to examine the ability of the proposed techniques to handle the missing values in multivariate time series, it compared the effect of missing data imputation on the behaviour of *VARMA* models. Model predictions were conducted in MATLAB utilising autoregressive process in polynomial form that introduced earlier in Chapter 2. For more details see [24, 30].

The general form of the *VARMA* model for the two time series $\boldsymbol{y_1(t)}$ and $\boldsymbol{y_2(t)}$ is [111]:

$$A_{01}(z)y_1\,(t) = A_1(z)y_2\,(zt) + C_1(z)e_1(t)$$

$$A_{02}(z)y_2\,(t) = A_2(z)y_1\,(zt) + C_2(z)e_2(t)$$

The data are split into two parts, the first half was used for identifying *VARMA* models and the second half was used to select model order and to validate the prediction results. Table 5.2 summarizes the effect of imputed data by the three techniques on the behaviour of the model, four metrics were used to measure the imputation performance: mean squared error (MSE), percentage of the fit to estimation, final prediction error and number of free parameters (model order). Among the three techniques, *MLD* produced results that are closer to the complete data as compared with the other two methods.

To validate the quality of the estimated models, the predicted responses were compared to measured data, Figure 4.2 shows the behaviour of the model for case of complete data, when 10 steps (1 hour) estimated response compared to measured data. In time series $\boldsymbol{y_1(t)}$, it is clear that there is no noticeable difference between the mean value of measured and estimated data. On the other hand, for the time series $y_2(t)$, it is evident that the difference in the means between the measured and estimated data is insignificant. Generally, the model fits the data very well compared with the measured data.

Figures 4.3 - 4.5 show the effect of missing data imputation, for the case of 10% missing data. Among the three proposed techniques (*MLD*, mean imputation and listwise deletion), *MLD* gave the best results to fit with the measured data, when The proportion of missing data was increased from 10% to 15%, as shown in Figures 4.6 - 4.8.

For listwise deletion, there was a dramatic change in the behaviour. On the other hand, the mean imputation produced different results, giving a poor estimation for the first time series and a better estimate for the second time series, and the *MLD* technique was affected slightly when the proportion of missing data changed from 10% to 15%. Overall, for the complete case, the MLD method had the best results.

Table 4.2: Statistical test result.

	MSE	Fit percentage		FPE	Number of parameters
		$y_1(t)$	$y_2(t)$		
			10% missing		
Complete	0.0001192	90.68	98.02	1.85×10^{-10}	55
MLD	0.001224	71.03	92.97	1.49×10^{-07}	25
Mean-sub	0.007662	51.18	57.39	1.86×10^{-05}	22
List-wise	0.006396	63.69	60.59	1.05×10^{-05}	31
			15% missing		
MLD	0.001331	67.59	93.26	1.61×10^{-07}	37
Mean-sub	0.0375	22.15	-11.01	1.50×10^{-05}	40
List-wise	0.003482	69.57	79.54	2.73×10^{-06}	4
			25% missing		
MLD	0.001325	70.45	91.17	2.07×10^{-07}	25
Mean-sub	0.07308	16.57	-86.69	3.26×10^{-05}	25
List-wise	-	-	-	-	-

Table 4.3: Statistical test result (R^2 and adjusted R^2)

	10% missing			
	R^2		Adjusted R^2	
	$y_1(t)$	$y_2(t)$	$y_1(t)$	$y_2(t)$
Complete	0.976445	0.970733	0.975957	0.970126
MLD	0.977074	0.971677	0.976599	0.97109
Mean-sub	0.859839	0.814202	0.856935	0.810352
List-wise	0.973700	0.965533	0.626271	0.546798
	15% missing			
MLD	0.975974	0.968912	0.975476	0.968267
Mean-sub	0.764613	0.761644	0.759734	0.756704
List-wise	0.969716	0.962359	0.968845	0.968845
	25% missing			
MLD	0.974978	0.964713	0.974459	0.963981
Mean-sub	0.63386	0.556	0.626271	0.546798
List-wise	0.957184	0.93884	0.955569	0.936533

Lastly, the case for 25% of missing measured data was considered. The Listwise deletion model reduced the model order dramatically, for that reason no estimation result is shown, the mean imputation was affected more when the missing proportion was increased to the quarter comparing with last case, and again, the *MLD* affected slightly comparing with the other two methods.

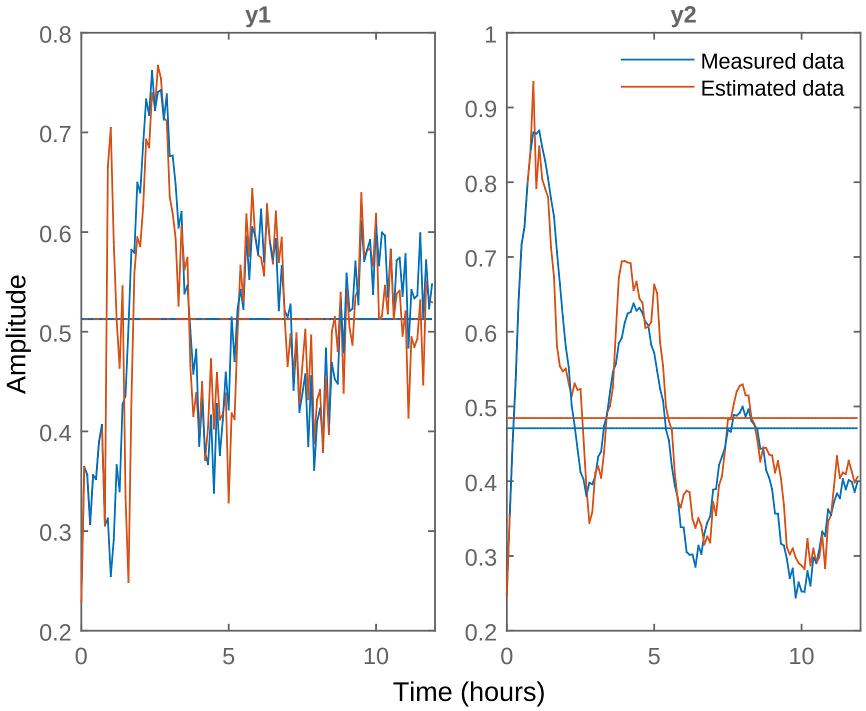

Figure 4.2: 10-step predicted response compared to measured data (complete data).

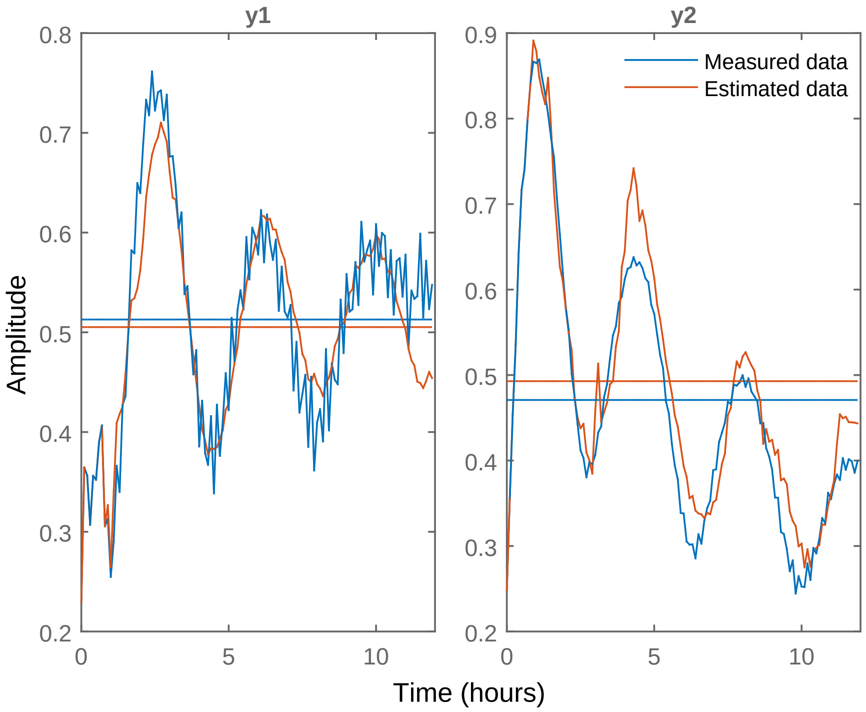

Figure 4.3: 10-step predicted response compared to measured data (MLD 10% missing).

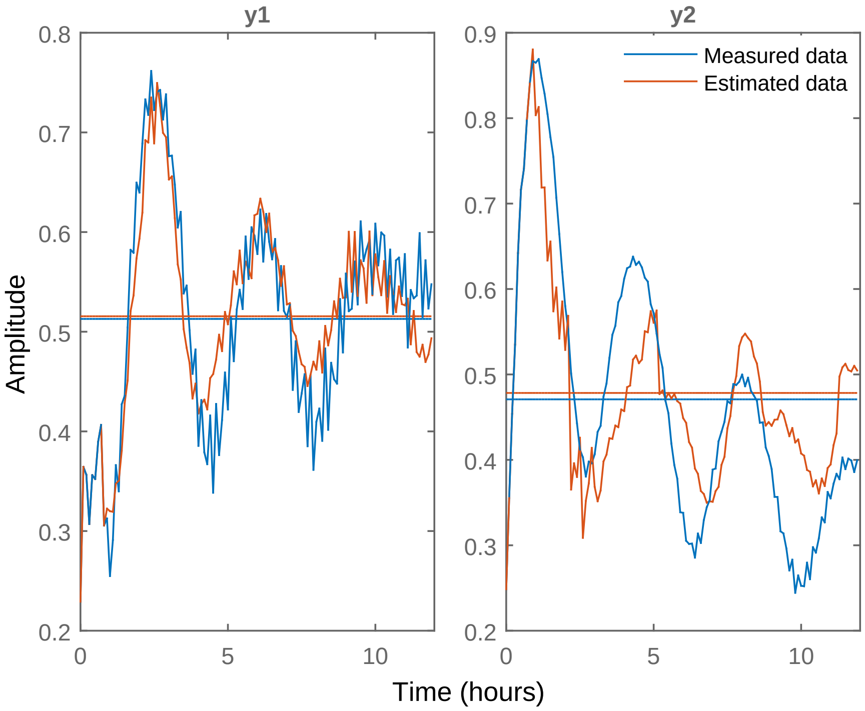

Figure 4.4: 10-step predicted response compared to measured data (Mean Imputation 10% missing).

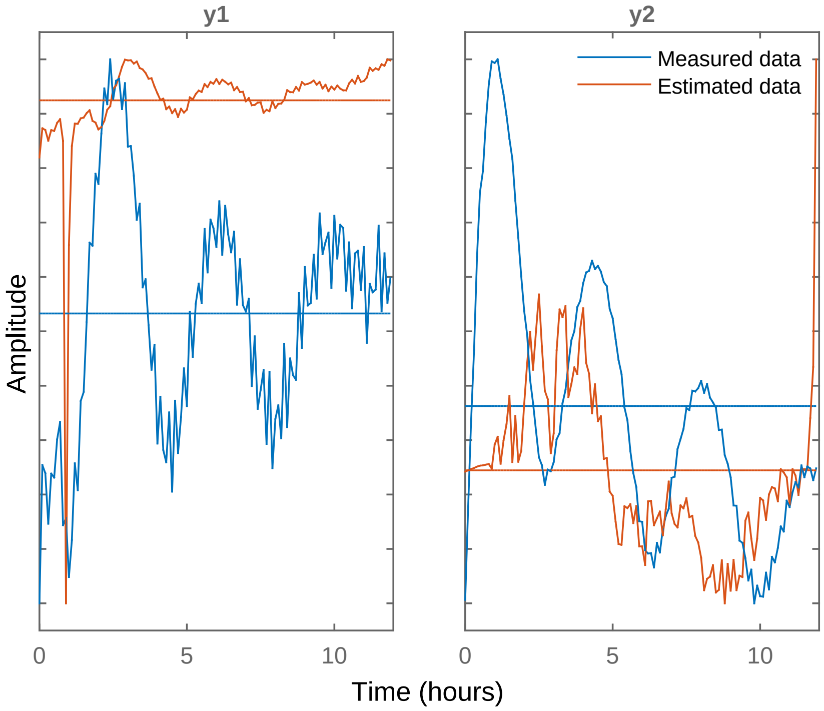

Figure 4.5: 10-step predicted response compared to measured data (Listwise deletion 10% missing).

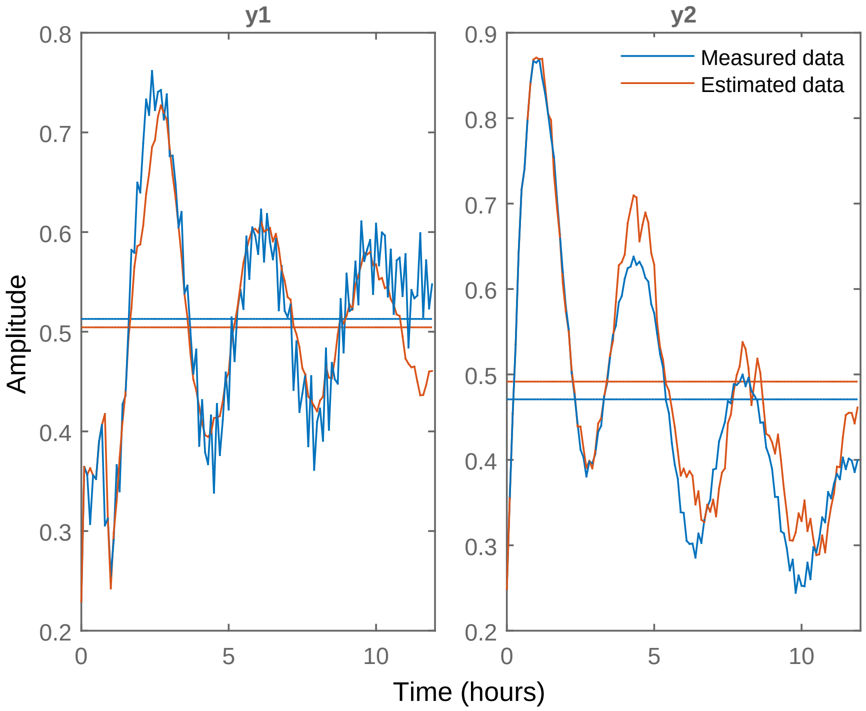

Figure 4.6: 10-step predicted response compared to measured data (MLD 15% missing).

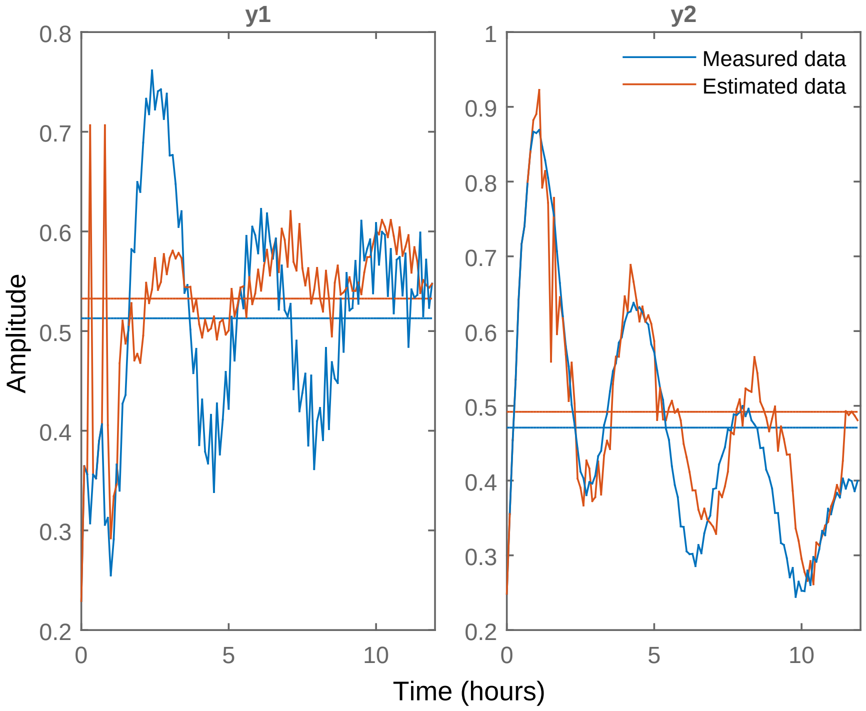

Figure 4.7: 10-step predicted response compared to measured data (Mean imputation 15% missing).

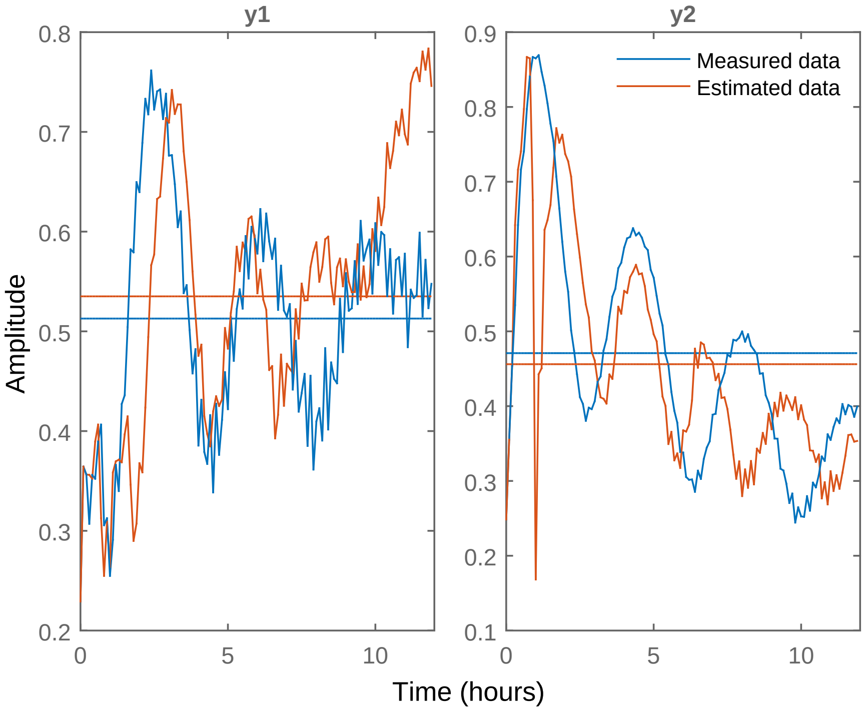

Figure 4.8: 10-step predicted response compared to measured data (List-wise deletion 15% missing).

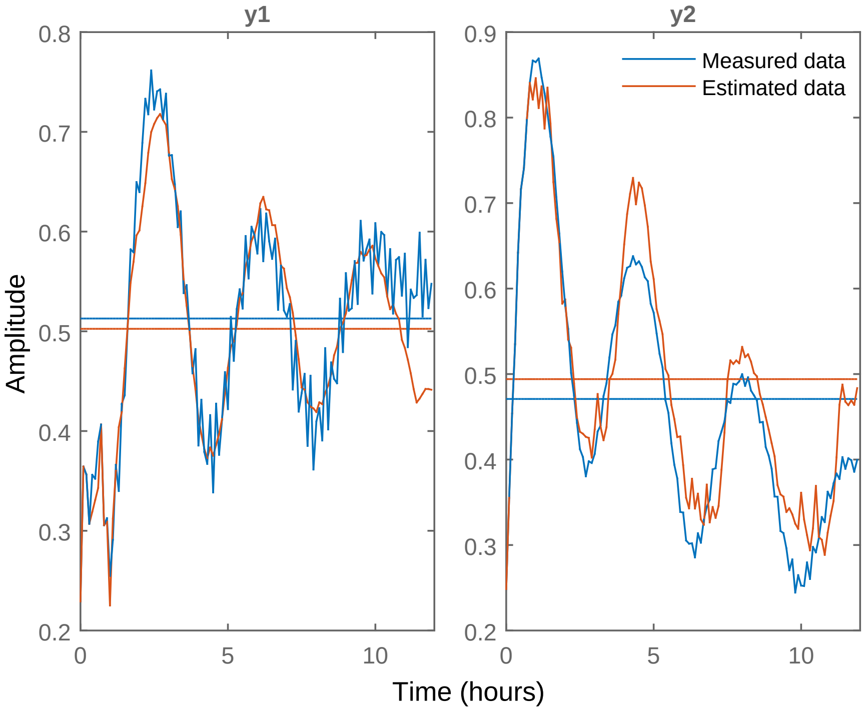

Figure 4.9: 10-step predicted response compared to measured data (MLD 25% missing).

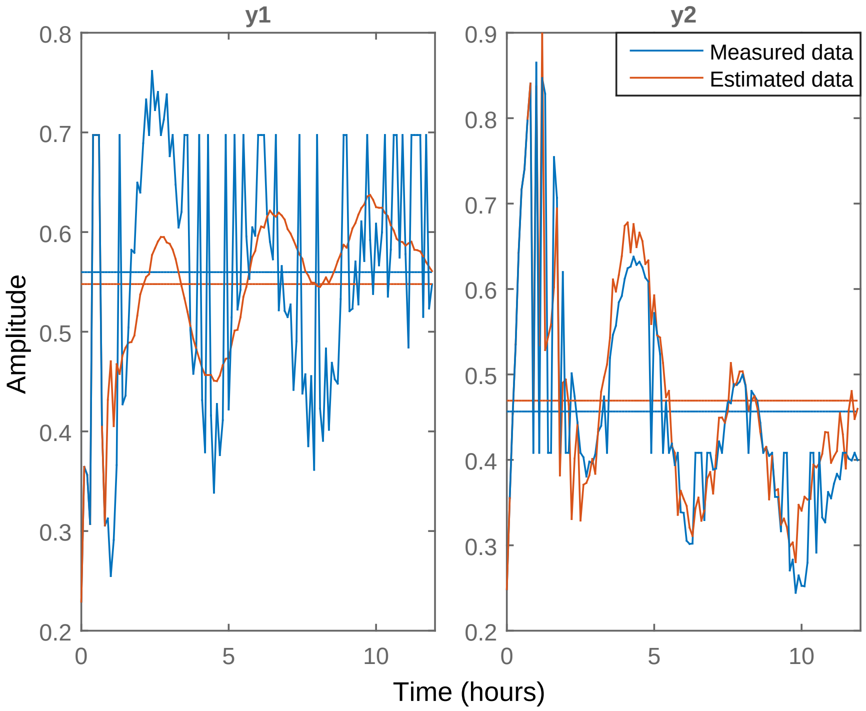

Figure 4.10: 10-step predicted response compared to measured data (Mean Imputation 25% missing).

Figure 4.11 shows the comparison between the measured and imputed values for *MLD* for three cases of missing data. Generally, the results from the proposed method for the three cases were close to the real data, this is because the *MLD* depends on an autoregression model (*VAR* model) to impute the missing values, whereas the identification of the *VAR* model depends on the availability of search windows where the specified search windows need to include complete case of observed data.

This is not possible if the length of time intervals of missing values is greater than the time interval of the search window. From Figure 4.11 it is clear, when the proportion of missing data is increased the ability of *MLD* in imputing the missing values decreased, that means the number of unimputed values will increase. To overcome these problems a new algorithm was developed and proposed in next section.

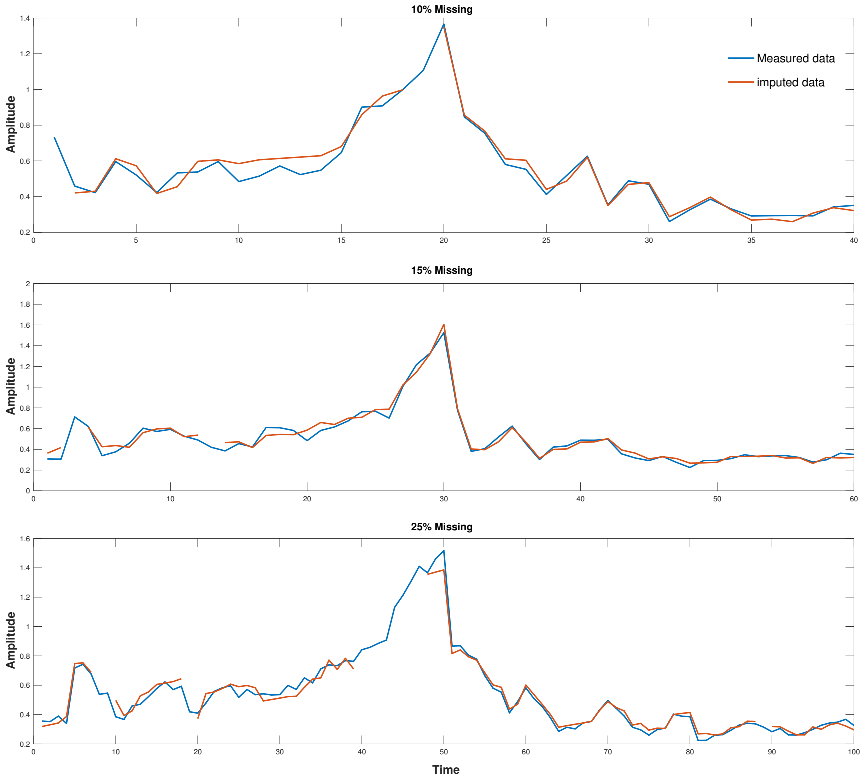

Figure 4.11: Measured and imputed data in 10%, 15% and 25% missing.

4.8 Vector Autoregressive Imputation Algorithm (VAR-IM)

The proposed algorithm for imputing missing data into a multivariate time series dataset is to use a Vector Autoregressive model combined with an EM algorithm and a prediction error minimization (PEM) algorithm [75]. This method, based on a combination of these algorithms, can significantly improve the imputation performance for dealing with missing data problems.

Specifically, in the first step, a traditional linear interpolation estimate is performed as an initial guess of the missing data. Next a $VAR(p)$ model is estimated by selecting the best lag value p. Finally, the parameters of the $VAR(p)$ model are estimated by alternatively using EM and PEM algorithms resulting in an improved value for the data imputation.

Basically, the alternation of the two algorithms between imputing missing data and estimating models, improves the model performance by applying the PEM algorithm in a way similar to the EM algorithm. The flow chart for the proposed $VAR - IM$ algorithm is shown in Figure 4.12.

The $VAR - IM$ technique formalizes an intuitive idea for identifying the best VAR model for imputing missing data:

1. Calculate the initial values to start the algorithm.

2. Select the order of the identified VAR^* Model.

3. Check the causality of time series.

4. Impute the missing values by using VAR^*.

5. Identify the new VAR model.

6. If convergence fails, return to step 4, otherwise, proceed to step 7.

7. Update the missing values with the PEM algorithm.

8. Impute the missing values.

For more details, assume that $\boldsymbol{X_t}$ represents a multivariate data set and that a set of VAR models can simulate X_t with different lags $p = 1, 2, 3, \dots$. and parameters $\boldsymbol{A}_p$. If there are no missing values, then calculate the least squares estimate of $\boldsymbol{A}_p$ based on equation (4.3).

$$X_t = \boldsymbol{\phi} \boldsymbol{A}_p + \boldsymbol{E} \tag{4.9}$$

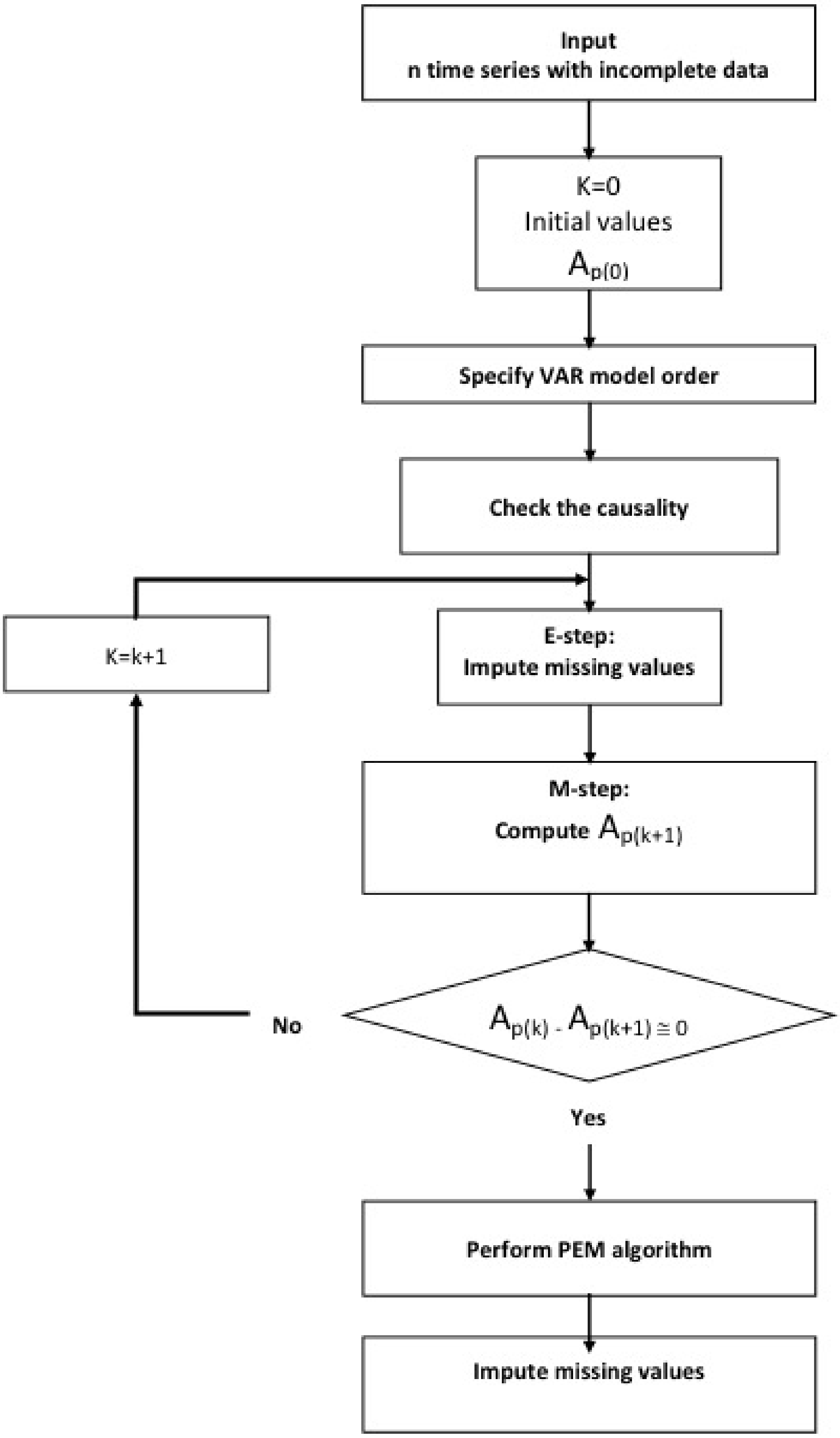

Figure 4.12: *VAR − IM* algorithm flow chart.

For dynamic systems the auto-regression process depends on the past values of the targeted data point, if the time series includes missing values, then past values will also be missed and the auto-regression cannot be applied in equation (4.9). In this case, the traditional approaches such as listwise will not work, because ignoring the missing values will effect the properties of the dynamic system. To begin the estimation process correctly, initial values are required, and the simple way to determine these initial values is to use a simple traditional method such as linear interpolation, this will be denoted by expressing X_t as $(X_{tmiss}\ ,\ \ X_{t0})$, where X_{tmiss} denotes the multivariate data set with missing values, and X_{t0} represents the multivariate data set with replaced missing values by initial values (imputed by interpolation technique [27]).

Consequently, equation (4.9) becomes:

$$\hat{X}_t = \boldsymbol{\phi}_k \boldsymbol{A}_{p_k} + \boldsymbol{E} \Rightarrow \hat{X}_{t0} = \boldsymbol{\phi}_0 \boldsymbol{A}_{p_0} + \boldsymbol{E}$$
$$\boldsymbol{A}_{p_k} = \left(\boldsymbol{\phi}_k{}^T \boldsymbol{\phi}_k\right)^{-1} \boldsymbol{\phi}_k{}^T X_k \boldsymbol{A}_{p_0} = \left(\boldsymbol{\phi}_0{}^T \boldsymbol{\phi}_0\right)^{-1} \boldsymbol{\phi}_0{}^T \boldsymbol{X}_0$$

Where $\boldsymbol{\phi}_0$ is the initial regression matrix, $k = 0, 1, 2, \ldots\ldots,$ and $\boldsymbol{A}_{p_0}$ is the initial coefficient matrix of the select $VAR(p_k)$ model.

The order of the model p_k is updated until the difference $\boldsymbol{A}_{p_k} - \boldsymbol{A}_{p_{(k+N)}}$ is less than ξ, where ξ is a prescribed small value.

4.8.1 Numerical Example

Eight time series $\boldsymbol{y_1}\,(\boldsymbol{t}), \ldots\ldots, \boldsymbol{y_8}\,(\boldsymbol{t})$ were created using the *'timeseries'* function in MATLAB. Each time series consists of 173 time points to represent a variable in a stationary multivariate time series $\boldsymbol{Y_t}$ as shown in Figure 4.13. To simulate a real scenario of missing data, the MCAR mechanism was generated by randomly dropping measured values to simulate three different proportions 15%, 20% and 30%.

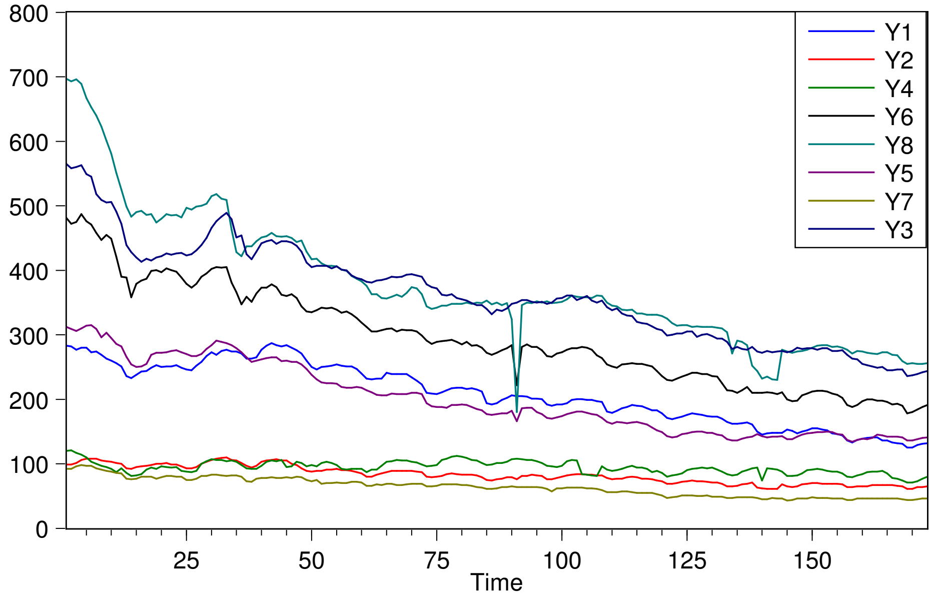

Figure 4.13: Generated time series.

This numerical example will examine the utility of the proposed algorithm by comparing it with a popular modern algorithm for handling missing data, the algorithm is a modified EM algorithm for dealing with missing values in dynamic data set [63]. The first step of $VAR - IM$ algorithm is to select the appropriate VAR model for imputing the missing values, where Table 4.4 shows the results of five tests of model selection criteria, the model selection techniques produced similar results when missing data proportions were 15% and 20%, and this suggested $VAR(1)$ model as the best model for missing data imputation.

However, the LR criteria, suggested a model with lag three to be the best model, when the missing data proportion was increased to more than a quarter of the measured data, the model selection criteria produced different results and this time the selected model was $VAR(2)$.

Statistical tests were applied to examine the performance of the proposed algorithm, Table 4.5 shows MSE and mean values (M) for complete data and three cases of missing data. The $VAR - IM$ produced closest results to complete data comparing with EM algorithm.

Table 4.4: *VAR* Model order selection.

		10% Missing			
Model order	AIC	SC	LR	HQ	FPE
1*	2.08	2.7745	9.3144	2.3607	7.57×10^{-05}
2	2.2914	3.5415	19.3114	2.7967	8.92×10^{-05}
3	2.2475	4.0532	39.6084	2.9774	8.67×10^{-05}
4	2.386	4.7473	23.2924	3.3405	9.72×10^{-05}
		15% Missing			
1*	2.24	2.9879	10.0309	2.5423	8.20×10^{-05}
2	2.4677	3.8139	20.7969	3.0118	9.60×10^{-05}
3	2.4204	4.365	42.6552	3.2064	9.30×10^{-05}
4	2.5696	5.1125	25.0841	3.5974	1.05×10^{-04}
		25% Missing			
1	3.5201	4.6954	15.7628	3.9952	1.28×10^{-05}
2*	1.9389	2.9967	15.3404	2.3664	7.54×10^{-05}
3	2.9391	5.3004	51.7956	3.8935	1.13×10^{-04}
4	2.386	4.7474	23.2924	3.3405	9.72×10^{-05}

Table 4.5: *VAR* Mean and MSE for the imputed data.

	15% missing		20% missing		30% missing	
	M	MSE	M	MSE	M	MSE
	(211.97)		(205.85)		(211.45)	
VAR-IM	206.78	0.2701	194.4	0.2169	170.61	1.0066
EM	168.2	0.4403	161.29	0.4386	136.62	1.0554

Figure 4.14 shows a comparison between the measured and imputed data for the $VAR - IM$ and the EM algorithm. In three cases of missing data, the $VAR - IM$ algorithm still remained the best choice even with the changes in proportion of missing data.

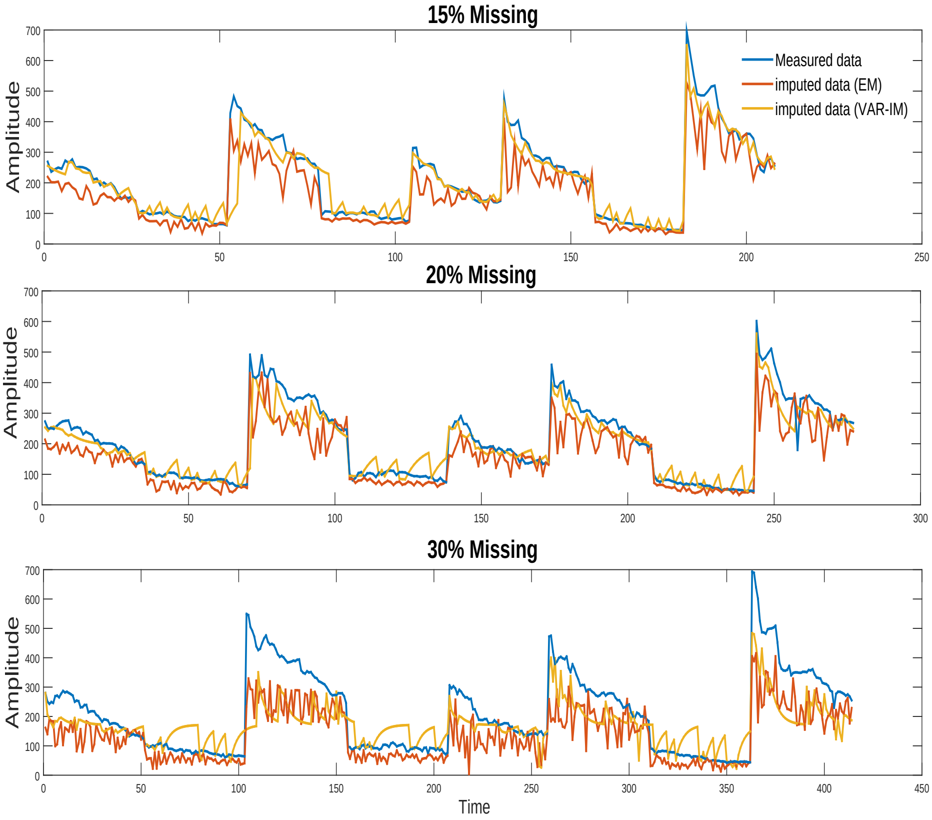

Figure 4.14: Measured and imputed data in 15%, 20% and 30% missing.

There are several advantages of the $VAR - IM$ technique. First, it is straightforward and can handle different missing data mechanisms (e.g. MAR and MCAR). Second, a steady fluctuation estimation is achieved as the missing data percentage increases. Third, it is quite robust against increasing percentages of missing data. In addition, $VAR - IM$ is straightforward to apply to the complex structure of multivariate time series, for more details the utility of this algorithm will be discussed more in Chapter 5.

4.9 Summary

Effectively handling multivariate observations containing missing data is extremely important. This is especially true in medical research, which typically includes a great number of variables, and the outcome has significant impacts to people's health. The proposed MLD and $VAR - IM$ methods provide fast and accurate approaches to impute missing values for multivariate time series datasets. It outperforms the commonly used methods such as Listwise deletion, mean substitution and EM algorithm. The positive results of the simulation study and analysis example discussed in this paper demonstrate that the MLD and $VAR - IM$ methods provide an effective alternative for the imputation of missing values in multivariate time series. When considering an increasing percentage of missing data, the other proposed methods become less effective, while the $VAR - IM$ shows a smaller deterioration in performance. In addition, the $VAR - IM$ method is more robust than the other proposed techniques and performs better on static and noisy data. However, the $VAR - IM$ method does have some limitations. Firstly, the validity of $VAR - IM$ requires that the time series must be stationary. Secondly, the $VAR - IM$ method is less effective when the percentage of missing data is quite low (less than 10%). . Finally, the contained example only considered a scenario in which data were missing completely at random (MCAR). A less stringent assumption of missing data mechanism, such as missing at random (MAR), may be more realistic in practice. Despite these limitations, $VAR - IM$ provides an important alternative to existing methods for handling missing data in multivariate time series. Furthermore, a part from this chapter was published as a journal paper in Neurocomputing journal [14]. Further extension of $VAR - IM$ to include other types of methods will be considered in the next chapter.

Chapter 5

Case Studies of the Application of $VAR - IM$ Algorithm for Dealing with Missing Values to Space Weather and ECG Data.

This chapter shows how the $VAR - IM$ algorithm deals with missing data in multivariate time series in real data sets. It presents the imputation procedure for multivariate time series data of two different real data sets, space weather and ECG data.

5.1 Space Weather Data

One of the important branches of aeronomy science is space weather conditions that focus on time-variant variables inside the Copernican system. This includes phenomena as solar wind, but typically pertains to the area outside the atmosphere but surrounding the Earth, including conditions inside the three layers (thermosphere, magnetosphere and ionosphere), where space weather conditions differ from the earthly atmosphere. “space weather conditions” is considered to be primary employed in the 1950s, however, it has become commonly used since the 1990s [26]. The solar wind is a component of plasma particles released through the atmosphere from the Sun. It includes generally energized electrons and protons varies between 1.5 and 10 keV. The stream of particles varies in time with density, heat, and velocity as well as over solar longitude. Such particles can breakout the Sun’s gravity, and it goes outward supersonically over huge distances, covering an area referred to like the heliosphere, a huge bubble shaped size flanked by the interstellar medium.

The solar wind is divided to two components, characterized the quiet solar wind and the speedy solar wind. The of quiet solar wind is around 248 $mile/s$, a heat of $1.3 - 1.5 \times 10^6\ ^0C$. Its composition closely matches the solar corona. In comparison, the fast solar wind has a typical speed of 466 $mile/s$ and a heat of $7.7 \times 10^6\ ^0C$. The speedy wind composition nearly matches that of the Sun’s photosphere [41]. The quiet solar wind is double as intensive and much more changing in strength compared to the speedy wind. The quiet wind as well own much more complex composition, with turbulent zones and great-scale structure [112].

The quit wind generally, seems to outward from the equatorial region belt. Speedy wind is assumed to result from coronal holes that are funnel-like parts of open field lines inside the magnetic field of the sun [58]. These open lines are especially diffuse over the magnetic poles of the sun. The plasma origin is short magnetic fields generated by convection cells inside the solar weather. Such fields confine the plasma and carry it to the tight channels in the coronal centre that are existing $20,000$ *kilometres* above the photosphere. The plasma is released to the centre when these magnetic field lines reconnect [62].

Nowadays, a great deal of data for space weather and the solar wind system can be acquired through satellites. Unfortunately, for various reasons, much important data are lost during transmission to earth. Because of this, much of the data become useless when performing relevant system identification and information modelling tasks. Consequently, the overarching purpose of this part is to introduce the $VAR - IM$ algorithm as a solution to the space weather missing data problem. Hopefully, researchers in the field of space weather modelling will benefit from this and be able to employ this method in their own research.

Specifically, this section will explain the problem of missing data on modelling space weather systems, with attention given to selecting and fitting models, checking stability, and comparing forecasts with forecast period data. In addition, examples are presented from the solar wind parameters rated real data measured from the NASA Advanced Composition Explorer (ACE) satellite and wind spacecraft [110].

5.1.1 $VAR - IM$ Algorithm for Solar Wind Data

With this part of the case study, this chapter will focus on the performance of the new algorithm ($VAR - IM$ algorithm), as proposed in chapter 4, to handle missing values in solar wind data. Because this algorithm is often simpler to implement than other modern methods and is suitable for multivariate time series data, this section will benefit many researchers. The $VAR - IM$ algorithm will be compared with another technique, which has been used before with similar a data set and yielded good results.

The data set is a sample of 8,664 samples extracted from the solar wind parameters rated data measured from ACE and WIND spacecraft. (OMNI Web Results FTPWeb Browser Results Listing for omni2 data set from 01/01/1995 to 01/01/1996). The data set contains information on solar wind parameters, which is divided into 12 time series. This work will focus on three of these: the solar wind magnetic field, B_z, B_x, and B_y. B_x lies along the Sun-Earth line, with B_z and B_y defining a vertical plane. Additional information about regarding this data set can be found in [67, 112].

One of the main difficulties in recovering missing solar wind parameters, is related to the numerous long data numbers. As a result, only two data intervals were used for missing data analysis. The two data sets are:

- Complete dataset 1: Consisting of 240 hours' observations, from 05 to 15 Jan 1995.
- Real incomplete dataset 2: Consisting of 240 hours' observations, from 19 to 29 Jan 1995.

The dataset 1, containing the complete information, was used to verify whether the imputed data is sensible or not. Here, some of the data were intentionally deleted to mimic the MCAR mechanism and then imputed it using available missing data methods then compare the results with the real data.

Consequently, the best method can be applied on real incomplete data and check the performance of the proposed models. The complete data set is changed to the following incomplete variables: B_x (18% missing), B_y (14% missing) and B_z (18% missing). All missing data mechanisms are assumed as MCAR. The datasets are shown in Figure 5.1. The ultimate goal of this analysis is to determine the performance of the proposed models in terms of system stability, adequacy and forecasts. These performance parameters will be a measure of the sensitivity of the different imputation methods to the performance of the proposed model. First, a comparison of the imputed values with real values is needed before applying the proposed methods directly on real incomplete data.

Consider a comparison the proposed algorithm with the mean imputation method. The comparative results for both methods include the three time series are shown in Figure 5.2. $VAR - IM$ Algorithm usually performed better than mean imputation (except for the statistic missing values which have values closed to observed data), because the $VAR - IM$ algorithm based on the information that borrowed from the observed data to impute the missing values.

On the other hand, the mean imputation method depends only on the data distribution of each time series. That means the imputed values by the mean imputation method will not automatically regressed (all imputed values will constitute a straight line).

This can be seen in Figure 5.2. Note that, the curve of imputed data by $VAR - IM$ algorithm is closer to the curve of real values. For these reasons, the use of mean imputation method will be ignored and all of remaining analysis will based on using the $VAR - IM$ algorithm.

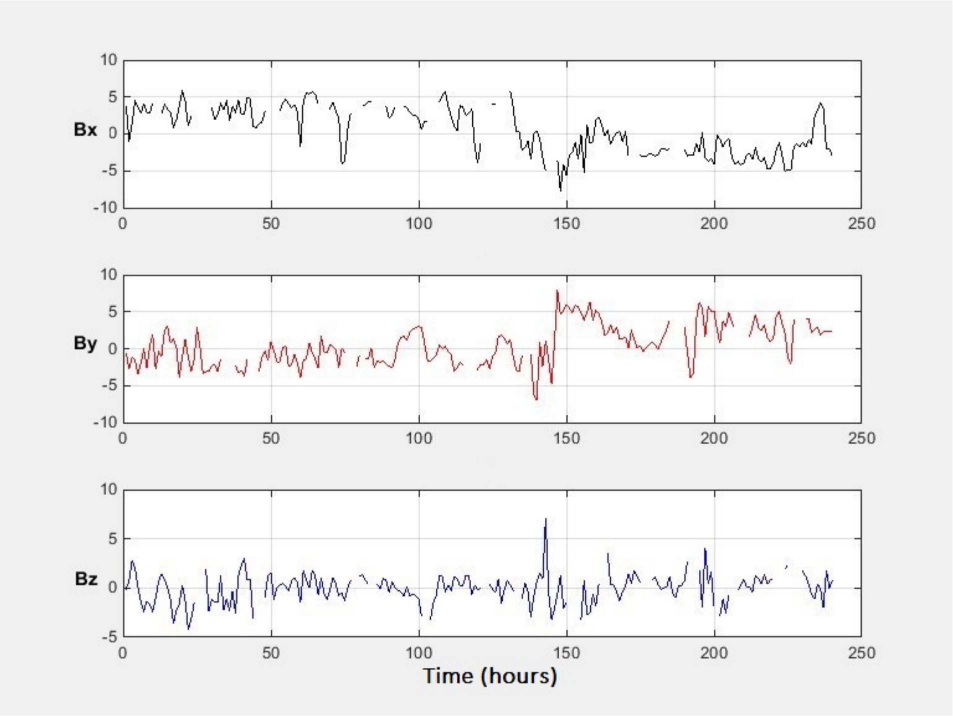

Figure 5.1: The solar wind magnetic field time series with missing values.

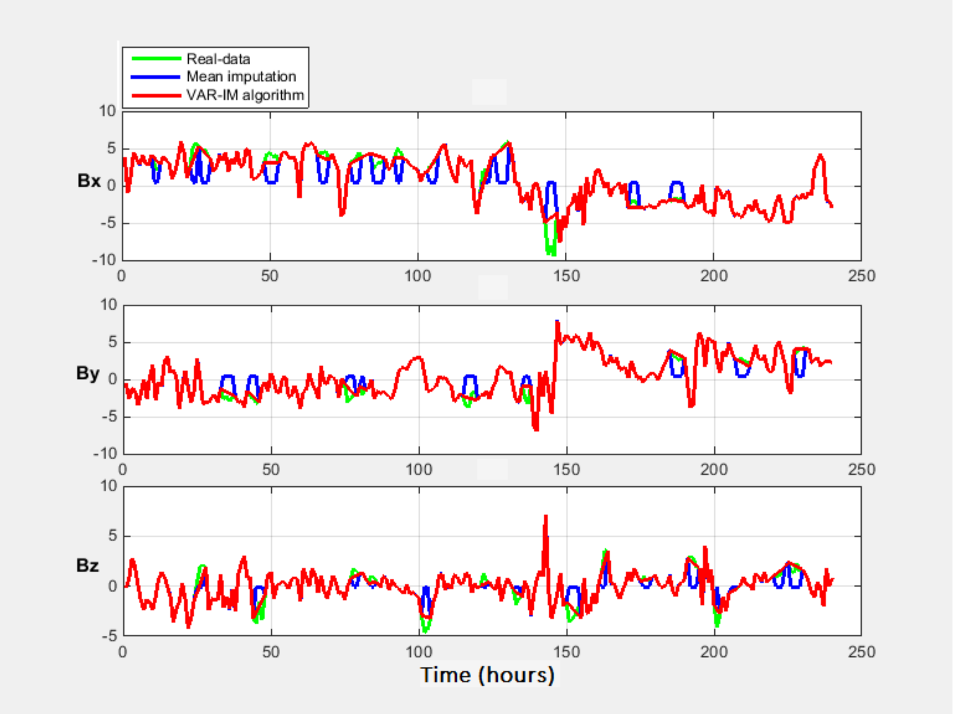

Figure 5.2: The solar wind magnetic field time series with imputed data.

5.1.2 Selecting and Fitting Models

After imputing the missing values by the $VAR - IM$ algorithm, the next step is select a model to fit the data. The choice were suggested from the model selection step in $VAR - IM$ algorithm. In this case, four models were arbitrarily chosen to be used for both the imputed data case and the real data:

- VAR second order with diagonal form
- VAR second order with full form.
- VAR forth order with diagonal form
- VAR forth order with full form

To determine a model's adequacy, a first step is to test the models for stability and inevitability. The test results indicates that, in both the complete case and the imputed case the predicted models are stable. The next step for model selection, as introduced in Chapter four, is to apply the likelihood ratio test and the Akaike information criterion.

To implement the likelihood ratio test, it is necessary to know the log of likelihood values and the number of parameters for each model to use them in comparing the AR models to their models using special MATLAB code [15], where the test refuses or be unsuccessful to refuse the hypothesis to show that the models with full form are suitable, for this test the results for both cases of the datasets were similar.

The likelihood ratio test indicated that the $VAR(4)$ models with diagonal and full form are rejected in favour of the corresponding $VAR(4)$ models with diagonal and full form. Therefore, based on this test, the $VAR(4)$

models with diagonal and full form are selected. The test did not refuse the $VAR(4)$ with diagonal form model in favour of the $VAR(4)$ with full form model.

The Akaike information criterion test requires the same inputs as the Likelihood Ratio test. It checks the Akaike information criterion values, where the models with smaller values are preferable.

The Akaike information criterion uses log likelihoods and model parameters, to determines values of Akaike information criteria. The model with the lowest value of the Akaike information criterion can be chosen as the most suitable model. To apply this test for the proposed models the MATLAB function *aicbic* was used. This gave two different results for complete and imputed data set, respectively.

- Complete data set: (2.4409 2.4309 2.4294 2.4278) × 10^3
- Imputed data: (2.3810 2.3806 2.3753 2.3867) × 10^3

According to this criteria, the best model is the $VAR(4)$ model with full form for the case of complete data set. For the case of imputed data, the $VAR(4)$ model with diagonal form has the lowest value, making it the best model. Also of note, is that the $VAR(4)$ model with diagonal form, in the case of complete data, has lower Akaike information than either of the other models. Based on this criteria, the $VAR(4)$ model with diagonal form is the best, and the $VAR(4)$ with full form model stands next in line preference. The estimated specification structures for the best models are shown in Table 5.1.

Where the number of time series is specified by n, the number of model lags is specified by nAR, nX represents the number of model lags cell array of $n \times n$ matrices of AR models and the covariance matrix $Qsolve$ is represented by $n \times n$ matrix. The parameters value for these models are shown in Table 5.2. Generally, all models give a similar data fit that can be seen in Figures 5.3 and 5.4.

Figures 5.5 and 5.6 indicate the sum of squares error between (SSE) the estimates and the real data for the four proposed models for cases of complete and imputed data set models, respectively. From the plots, the predictive performance of the four models is different in both cases, and the fourth proposed model appears to be the preferable and most powerful fit in both cases, its models parameters are shown in Table 5.3.

Table 5.1: The estimated specification structures for the best models.

Complete data set	Imputed data set
Model: 3-D $VAR(4)$ with Additive Constant Series: {’Bx’ ’By’ ’Bz’} n: 3 nAR: 4 nMA: 0 nX: 0 a: [0.0780314 0.254502 -0.0296171] additive constants asolve: [1 1 1 logical] additive constant indicators AR: {4x1 cell} stable autoregressive process ARsolve: {4x1 cell of logicals} autoregressive lag indicators Q: [3x3] covariance matrix Qsolve: [3x3 logical] covariance matrix indicators	Model: 3-D $VAR(4)$ with Additive Constant Series: {’Bx’ ’By’ ’Bz’} n: 3 nAR: 4 nMA: 0 nX: 0 a: [0.0566299 0.0654517 -0.0960825] additive constants asolve: [1 1 1 logical] additive constant indicators AR: {4x1 cell} stable autoregressive process ARsolve: {4x1 cell of logicals} autoregressive lag indicators Q: [3x3] covariance matrix Qsolve: [3x3 logical] covariance matrix indicators

Table 5.2: The parameters values for selected models.

Complete data set	Imputed data set
$Constant = \begin{bmatrix} 0.1169 & 0.2617 & 0.0445 \end{bmatrix}$	$Constant = \begin{bmatrix} 0.1438 & 0.2851 & 0.0731 \end{bmatrix}$
$VAR\,(1) = \begin{bmatrix} 0.7242 & 0.0398 & 0.0869 \\ -0.1021 & 0.6194 & -0.0920 \\ -0.0673 & 0.0620 & 0.4986 \end{bmatrix}$	$VAR\,(1) = \begin{bmatrix} 0.7347 & -0.0080 & 0.0361 \\ -0.0881 & 0.6687 & -0.1044 \\ -0.0142 & 0.0692 & 0.4432 \end{bmatrix}$
$VAR\,(2) = \begin{bmatrix} 0.1759 & 0.0440 & -0.1289 \\ -0.0296 & -0.1312 & 0.0467 \\ -0.0392 & -0.0950 & 0.0967 \end{bmatrix}$	$VAR\,(2) = \begin{bmatrix} 0.1209 & 0.0617 & -0.0676 \\ -0.0247 & -0.1500 & 0.0396 \\ -0.0576 & -0.1329 & 0.1209 \end{bmatrix}$
$VAR\,(3) = \begin{bmatrix} -0.1356 & -0.2171 & 0.0653 \\ -0.0100 & 0.2043 & -0.0529 \\ -0.2002 & -0.0464 & -0.0785 \end{bmatrix}$	$VAR\,(3) = \begin{bmatrix} -0.0895 & -0.1688 & 0.0366 \\ 0.0310 & 0.1439 & -0.0272 \\ -0.1725 & -0.0546 & -0.0617 \end{bmatrix}$
$VAR\,(4) = \begin{bmatrix} 0.0785 & -0.0303 & -0.0840 \\ 0.0388 & 0.0361 & 0.0078 \\ 0.1333 & 0.1175 & -0.0379 \end{bmatrix}$	$VAR\,(4) = \begin{bmatrix} 0.0636 & -0.0531 & -0.0831 \\ -0.0370 & 0.0517 & -0.0013 \\ 0.0821 & 0.1383 & -0.0312 \end{bmatrix}$

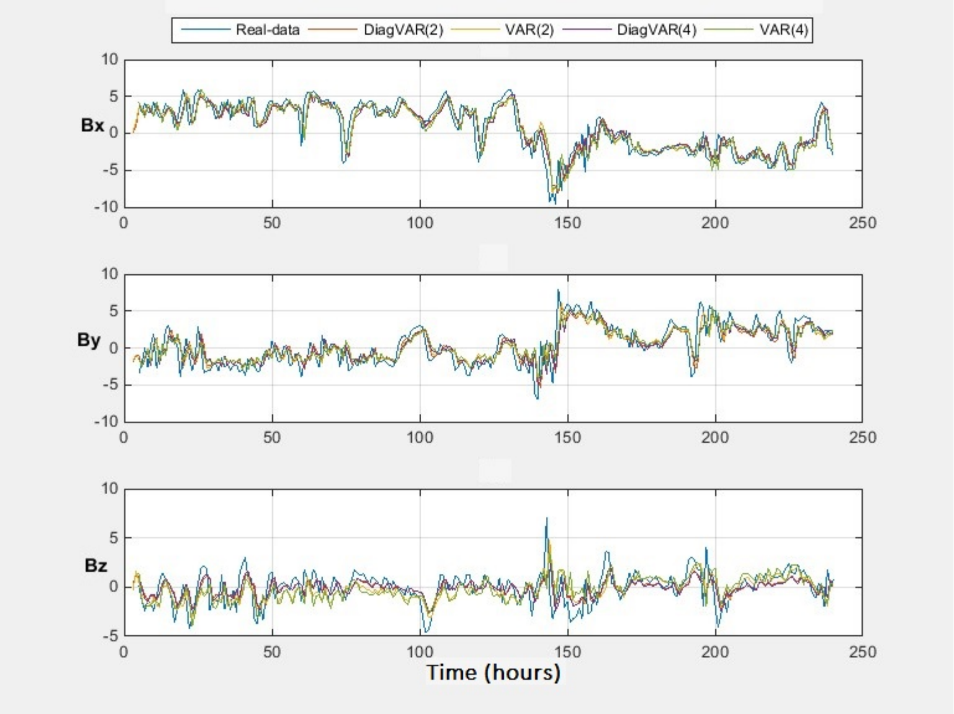

Figure 5.3: The proposed models of the solar wind system with complete data.

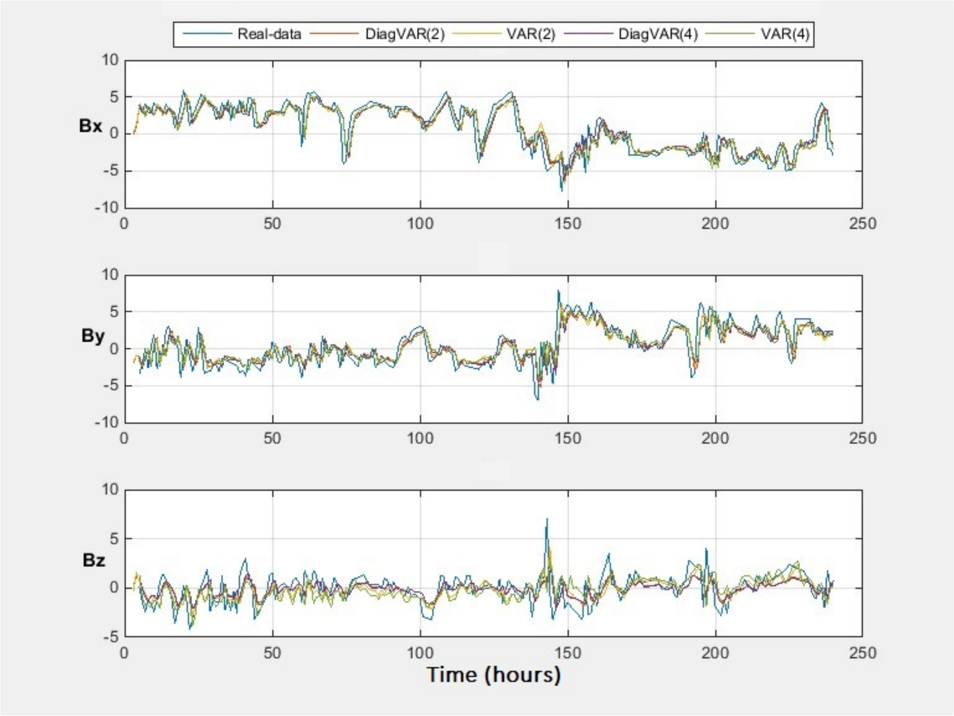

Figure 5.4: The proposed models of the solar wind system with imputed data.

Table 5.3: The parameters values for lag 4 models.

Complete data set	Imputed data set
$Constant =$ $\begin{bmatrix} 0.0780 & 0.2545 & -0.0296 \end{bmatrix}$	$Constant =$ $\begin{bmatrix} -0.2767 & 0.0998 & 0.0745 \end{bmatrix}$
$VAR\,(1) =$ $\begin{bmatrix} 0.6872 & -0.0918 & -0.0707 \\ 0.0368 & 0.6027 & 0.0853 \\ 0.0951 & -0.1126 & 0.4950 \end{bmatrix}$	$VAR\,(1) =$ $\begin{bmatrix} 0.7726 & 0.1660 & -0.0484 \\ 0.0301 & 0.7662 & 0.0351 \\ 0.0612 & -0.0100 & 0.5860 \end{bmatrix}$
$VAR\,(2) =$ $\begin{bmatrix} 0.1912 & -0.0139 & -0.0404 \\ 0.0422 & -0.1104 & -0.1176 \\ -0.1301 & 0.0682 & 0.0843 \end{bmatrix}$	$VAR\,(2) =$ $\begin{bmatrix} -0.1469 & -0.0562 & 0.0153 \\ -0.0948 & -0.0018 & 0.0883 \\ 0.0029 & -0.0011 & 0.0683763 \end{bmatrix}$
$VAR\,(3) =$ $\begin{bmatrix} -0.1070 & -0.0170 & -0.2106 \\ -0.2031 & 0.1992 & -0.0391 \\ 0.0629 & -0.0758 & -0.1049 \end{bmatrix}$	$VAR\,(3) =$ $\begin{bmatrix} 0.1300 & -0.1346 & -0.0896 \\ 0.0391 & 0.1056 & -0.0687 \\ -0.1096 & 0.0779 & -0.1749 \end{bmatrix}$
$VAR\,(4) =$ $\begin{bmatrix} 0.0910 & 0.0287 & 0.1053 \\ -0.0373 & 0.0521 & 0.1243 \\ -0.0768 & 0.0089 & -0.0380 \end{bmatrix}$	$VAR\,(4) =$ $\begin{bmatrix} 0.0353 & -0.0120 & 0.0874 \\ -0.1059 & -0.0878 & -0.0343 \\ 0.0123 & -0.0122 & 0.0820 \end{bmatrix}$

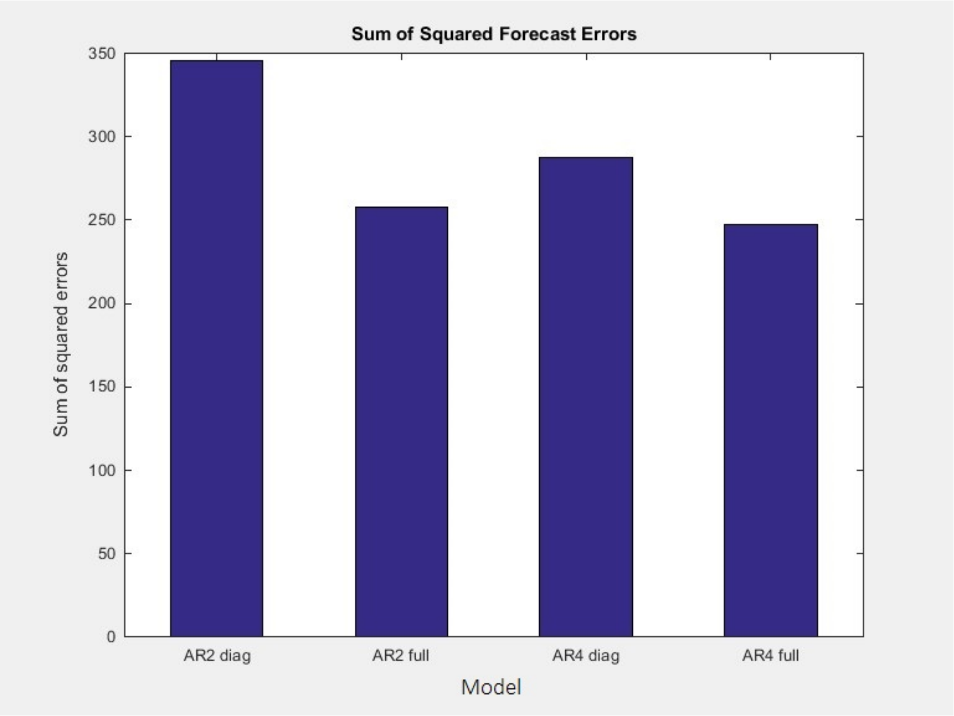

Figure 5.5: The par plot of sum of squares of four proposed models for complete data.

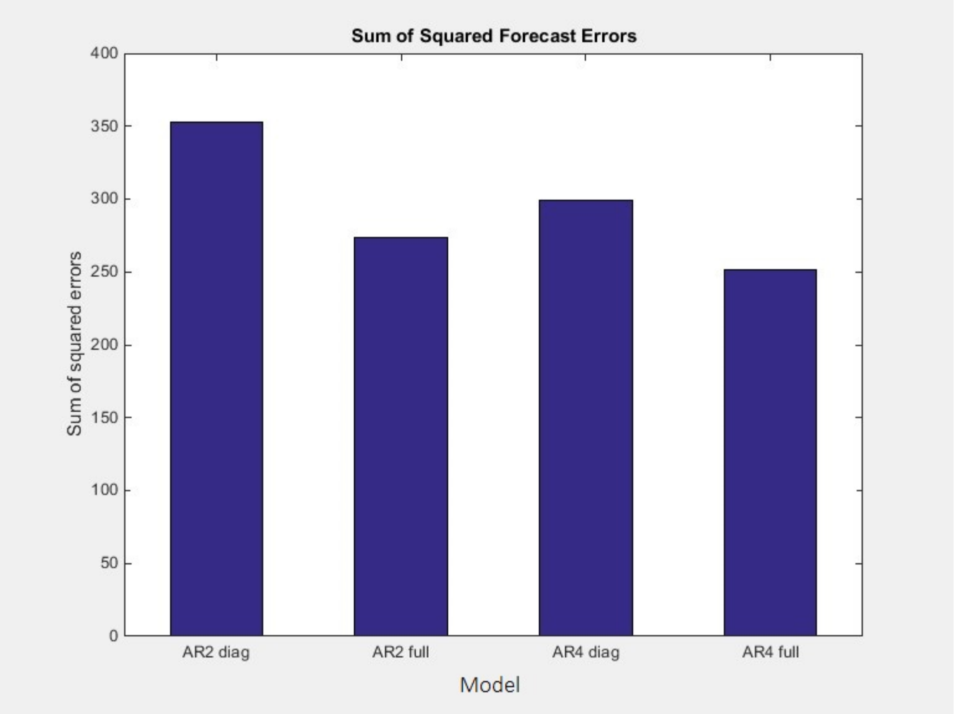

Figure 5.6: The par plot of sum of squares of four proposed models for imputed data.

5.1.3 Data Forecasts

After the model parameters have been estimated, the predictions from the models can be examined. MATLAB functions are used to match the forecasts of the selected models with the forecasted data [15], where these functions return both a forecast of the average of time series, and an error covariance matrix which shows confidence band around the average value.

Figures 5.7 and 5.8 illustrate the confidence bands overlayed on the forecasts in the shaded region to the right, for complete and imputed data set models, respectively. The model predictions are within the confidence intervals showing a good indication of the models.

It is clear, from the shaded region on the right hand side of the Figure 6.8; the fitted model for the imputed data is inside the confidence intervals giving a good indication about the quality of the $VAR - IM$ algorithm to impute the missing data in these time series.

Figures 5.9 and 5.10 show predictions of 50 hours into the future for complete and imputed data set, respectively. The dotted red line represents the extrapolations, and the solid black line indicates the real data, exploring the last few hours of these figures reveals a sense of how the forecasts relate to the latest hours.

The forecast shows little increase in B_x, a slight decline in By, and B_z remaining stable around zero. It is clear that because the models yield similar results in both cases, that the $VAR - IM$ algorithm recovers the missing values perfectly.

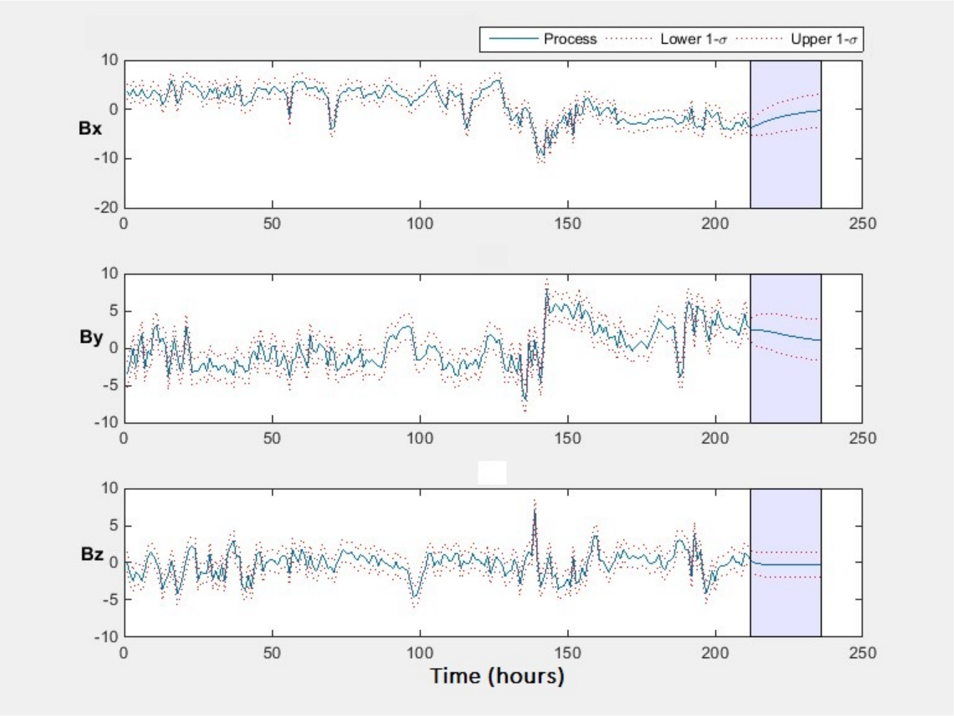

Figure 5.7: Forecasts with forecast period data of complete data.

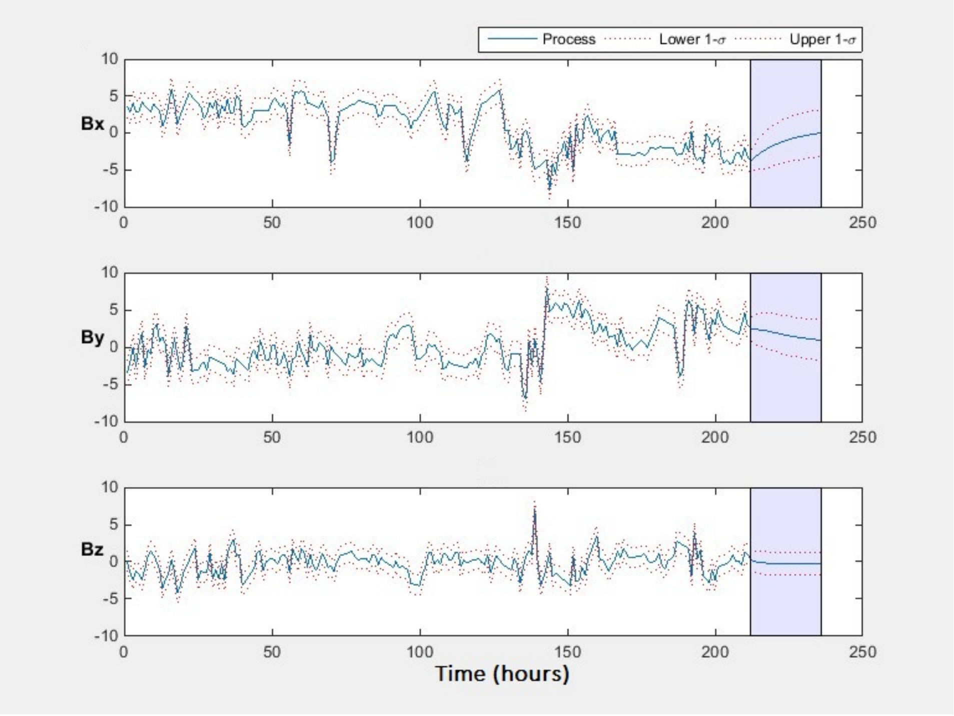

Figure 5.8: Forecasts with forecast period data of imputed data.

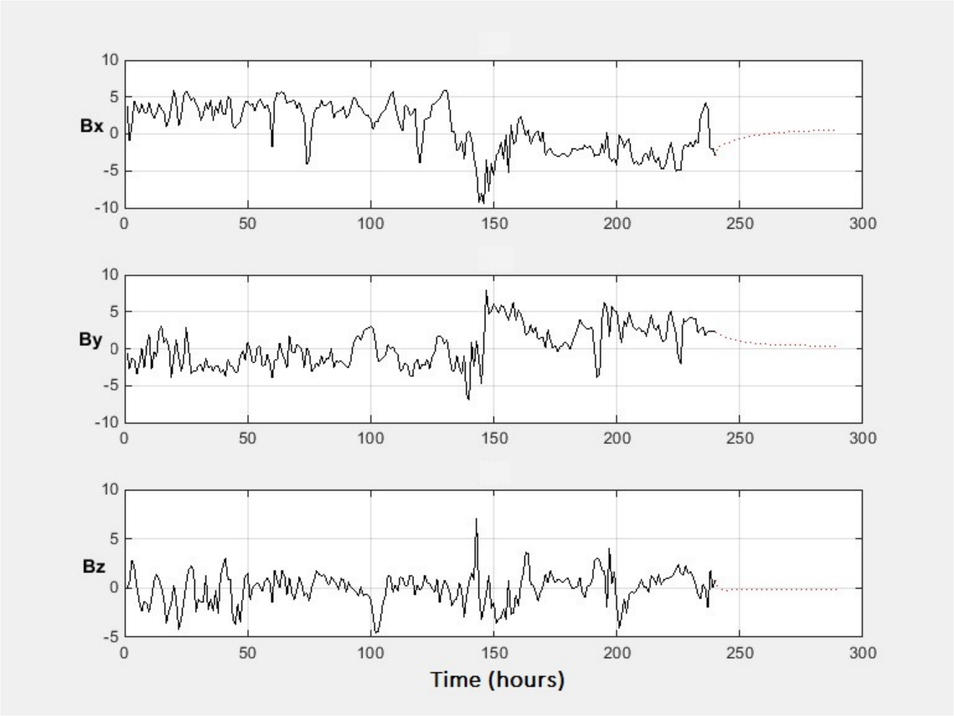

Figure 5.9: Predictions 50 hours into the future for complete data.

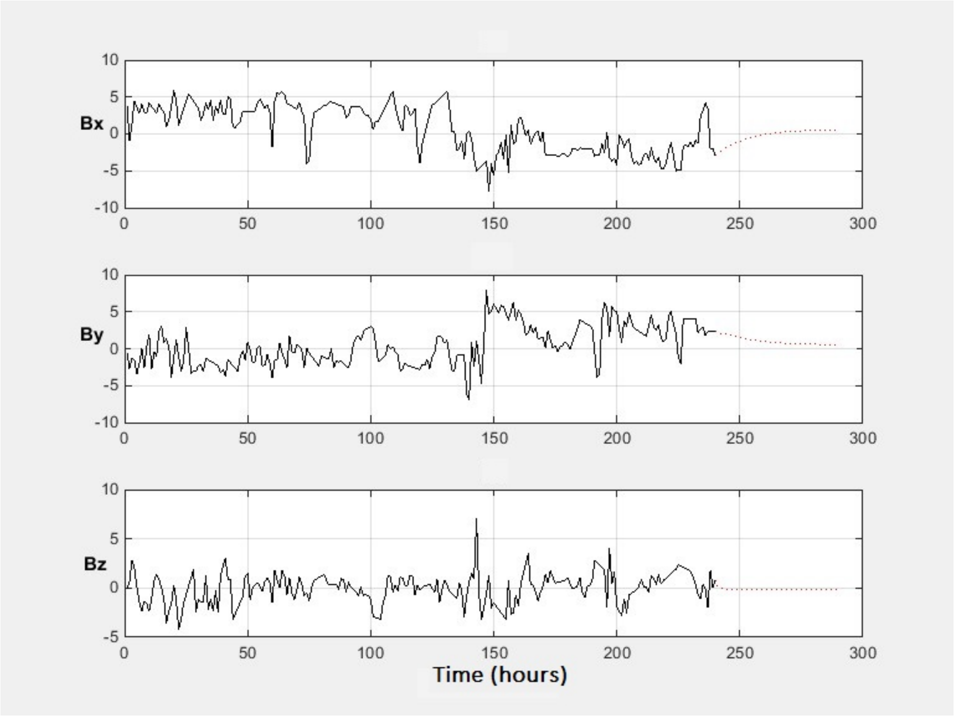

Figure 5.10: Predictions 50 hours into the future for imputed data.

5.1.4 Case Study on the Incomplete Data 2

The work above was based on a virtual missing data situation. A part of the data was deliberately removed to compare the performance of the proposed missing data analysis method for multivariate time series. The results shown by comparing the fitted models of the imputed and real values indicated that most of the models were acceptable, with only minor variations in the performance of these models, in this part the attempt was to apply the proposed method on a real data set and checked the performance of the proposed models, where the same proposed models were used for a real case of missing data (Incomplete data set 2). Stability and inevitability tests indicated that the estimated models are stable and invertible. The likelihood ratio test indicated that both models, $VAR(2)$ and $VAR(4)$ with diagonal form, were rejected in favour of the corresponding models, $VAR(2)$ and $VAR(4)$ with full form. Therefore, based on this test, the models $VAR(4)$ and $VAR(2)$ with full form, are the best. However, the test did not refuse the model $VAR(2)$ with full form in favour of the model $VAR(4)$ with full form. (The nominated model is $VAR(2)$ with full form as a model $VAR(4)$ with full form with restrictions in that the autoregression matrices for models $VAR(4)$ with diagonal and full form equals 0). Thus, it shows that the model $VAR(2)$ with full form will be the selected model. The nominated model depending on the criterion of Akaike information is the $VAR(2)$ with full form. Notice, too, the model $VAR(2)$ with full form has higher value than either of the remaining models. Based on the results of the test, the $VAR(4)$ model with diagonal form will be the selected model, and the $VAR(2)$ with full form coming next in preference. In this case, one of the nominated models can be chosen, which is $VAR(4)$ model with diagonal form and the model parameters are shown in Table 5.4.

- Imputed data set: (2.4319 2.4309 2.4307 2.4341) $\times 10^3$

Figure 5.11 shows the par plot of SSE between the predictions and the imputed data for the four proposed models. It can be noted that the of the four models nearly have the same performance. The first and third proposed models seem to be the best and most parsimonious fits. In this case, the $VAR(4)$ model with diagonal form will be chosen to be the best model to fit the data.

Table 5.4: The parameters values for lag 4 models.

Imputed data set
$Constant = \begin{bmatrix} -0.343376 & 0.295209 & 0.190591 \end{bmatrix}$
$VAR\,(1) = \begin{bmatrix} 0.791564 & 0 & 0 \\ 0 & 0.799212 & 0 \\ 0 & 0 & 0.617079 \end{bmatrix}$
$VAR\,(2) = \begin{bmatrix} -0.151809 & 0 & 0 \\ 0 & -0.00458309 & 0 \\ 0 & 0 & 0.0607771 \end{bmatrix}$
$VAR\,(3) = \begin{bmatrix} -0.0861887 & 0 & 0 \\ 0 & 0.0995244 & 0 \\ 0 & 0 & -0.207236 \end{bmatrix}$
$VAR\,(4) = \begin{bmatrix} 0.0612202 & 0 & 0 \\ 0 & -0.0794082 & 0 \\ 0 & 0 & 0.106987 \end{bmatrix}$

Figure 5.12 shows the comparison of forecasts with forecast period data, the forecasts in the shaded region to the right. The result shows that the forecasts still fall in the error bands of the forecasts period data, which give a good indication of the proposed algorithm in imputing the missing values. Predictions 50 hours into the future are shown in Figure 5.13, the extrapolations in dotted red, and the original data series in solid black. By looking at the last few hours in this plot to get a sense of how the

predictions relate to the latest data points, the forecast shows little growth in B_x, a slight decline in B_y, and uncertainty about the direction of B_z.

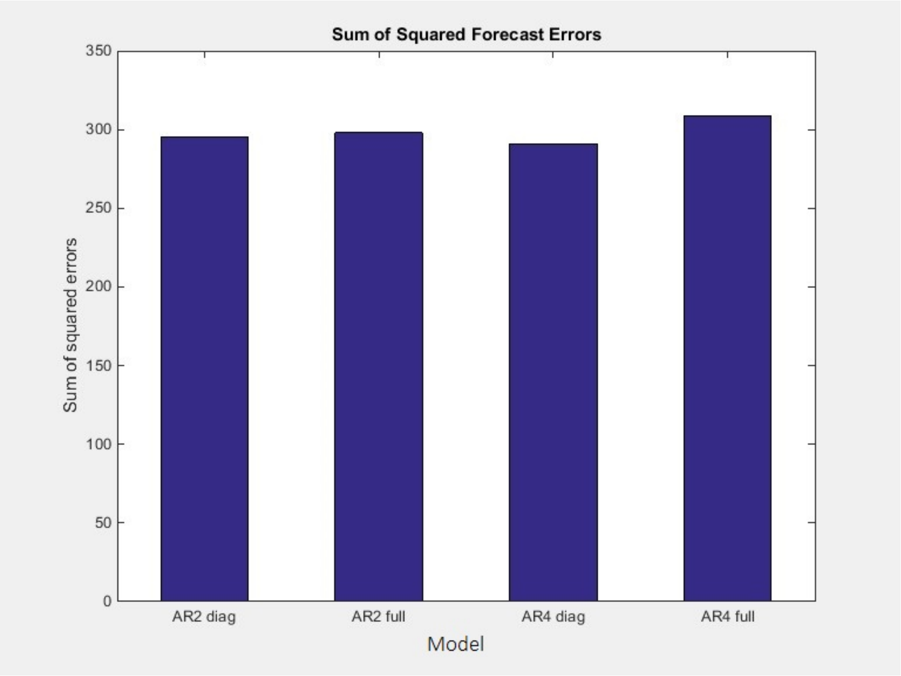

Figure 5.11: The par plot of sum of squares of four proposed models for imputed data.

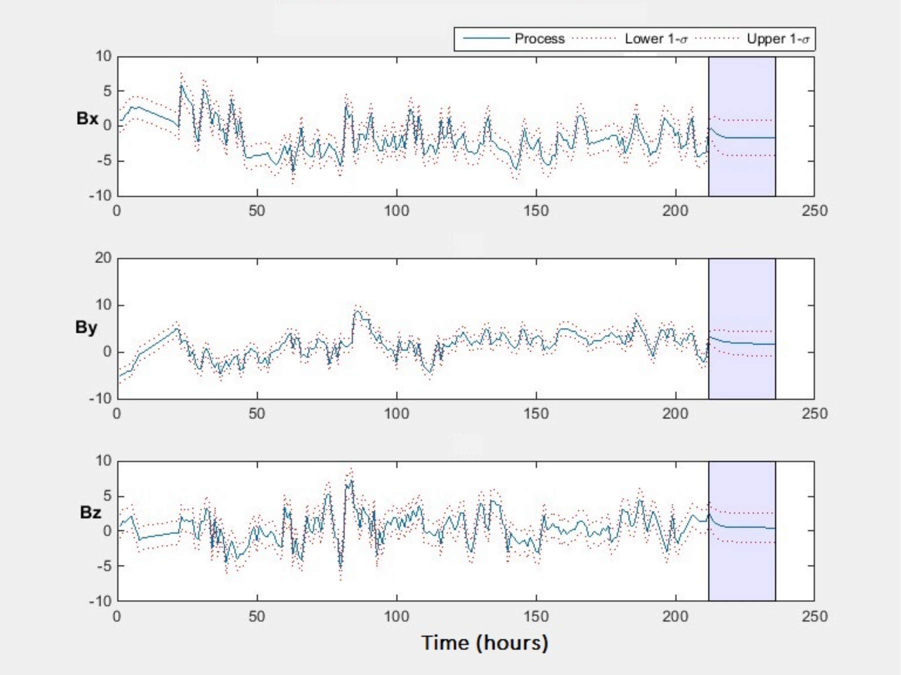

Figure 5.12: Forecasts with forecast period data of imputed data set.

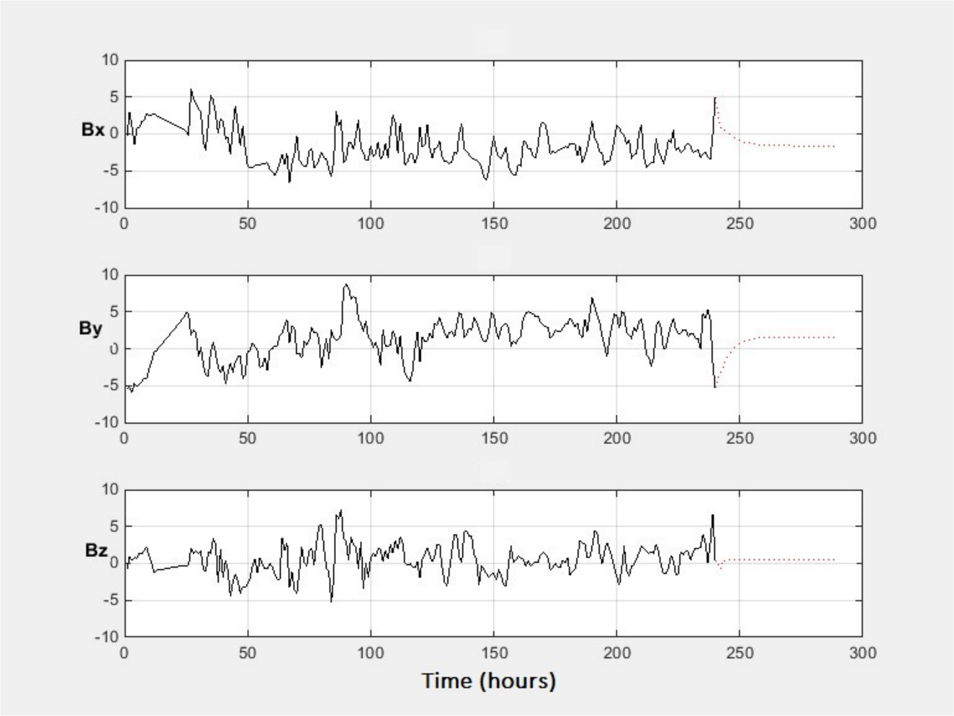

Figure 5.13: Predictions 50 hours into the future for imputed data.

5.2 ECG Data

In medical field, effective modelling using multivariate time series data is important. However, for various reasons the measured data may contain instances of absent data occurring either during or after the data collection process. Therefore, an effective method of handling missing data is important for this field. Especially in visual diagnosis, an effective process addressing missing data is of utmost importance, where disease diagnosis is typically based on measured data, which are represented by multivariate time series. Examples include functional magnetic resonance imaging (FMRI), Electroencephalography (EEG), Galvanic skin response and Electrocardiography (GSR) and electrocardiogram (ECG).

A case study involving ECG data have been selected because of the importance of handling missing data in this type of data sets, and this section is divided into two parts, in the first part; a review of the ECG data and comparing the $VAR - IM$ algorithm with three traditional methods for imputing missing data: Mean substitution, list-wise deletion and linear regression substitution. In the second part, the proposed algorithm method is compared with more powerful modern techniques: MARSS Package, nearest neighbor, and the modified EM algorithm.

To further examine the performance of the proposed algorithm with its ability to deal with real world missing data problems, a complete real dataset of ECG signals (without missing values) is considered and used as a case study. The dataset is available at the Physionet website $http://www.physionet.org/physbank/database/ptbdb$. This data set includes 290 patients with 549 measured values (total population 290 patients: aged between 17 and 87, mean 57.2; 209 men, and 81 women, mean age 61.6). Each subject is represented by one to five measured values. There are no patient numbered 124, 132, 134, or 161.

Each case contains 15 simultaneously records: the conventional 12 ECG leads (*i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6*) together with three Frank ECG leads (*vx, vy, vz*) [43]. Each signal is digitized at 1000 samples per second, with 16-bit resolution over a range of 16.384 mV. On special request to the contributors of the database, recordings may be available at sampling rates up to 10 KHz. The diagnostic classes of the patients are divided into nine types.

This case study considered signals from 12 ECG leads for two diagnostic classes: myocardial infarction and healthy control, a more detailed discussion is available at [20, 46]. Two cases of MCAR missing data mechanism with two different percentages 10% and 20% were generated.

5.2.1 Heart Rate

The diagnosis of various heart diseases has become easy, fast and efficient, thanks to the development of the ECG technique, where one of the most important features of ECGs is heart rate. From the leads, a time series graph showing significant heart rates can be measured to give good indications about the condition of the patient. One of these rates is ventricular rate, which can be measured by determining the number of QRS waves in each period. Unfortunately, the measured values can be affected and miss some important information, and that can be result from several conditions.

For instance, sometimes skin conductivity for electricity is insufficient to allow the electrodes to pass the pure signal through the electrical circuit, or the electrodes themselves lack the quality to sense the electrical signals. Any of these reasons can lead to missing data, which causes distortion of ECG signal [91].

A common problem in ECG signal processing is the removal of unwanted artefacts, noise and the appearance of missing values, and these situations can lead to problems in process the ECG signal. Such as the presence of a low frequency component, an irregular distance between QRS waves, or wave peaks appearing at irregular locations. Whereas one of the basic tasks of ECG signal processing QRS peaks, it is not possible to record pure ECG signal directly in existing of these problems.

Another concern is that the filtering process requires the removal of impacted noise from the original signal, but it is not possible to apply the filtering processing if there are missing values. As a basic step, after the imputation of the missing values, a filter can be used to remove the noise from the original signal. The $VAR - IM$ method was used to impute 10%

and 20% missing completely at random data from 38,400 samples for the conventional 12 leads of a myocardial infarction patient.

Tables 5.5 and 5.6 show the effect of missing values imputation on the heart rate in each of the 12 leads in both cases of missing data mechanism, respectively. In both cases, the proposed method $VAR - IM$, shows an improvement as compared with the other methods.

Table 5.5: Proposed methods for Heart-rate 10% MCAR.

	The conventional 12 leads											
Method	i	ii	iii	avr	avl	avf	v1	v2	v3	v4	v5	v6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing -data	73.9	63.8	66.78	47.8	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	79	67.08	70.13	54.58	55.08	65.48	37.73	43.58	50.48	70.97	73.51	72.13
Mean-sub	73.96	63.8	66.833	47.95	50	60.93	30.55	37.95	45.83	66.32	66.4	64.93
List-wise	87.79	74.05	76.07	49.82	58.84	72.17	34.71	43.92	52.03	74.4	74.96	72.85
Linear-reg	76.82	67.8	70.28	50.07	56.57	100.4	76.32	52.65	57.55	85.47	83.57	69.87

Table 5.6: Proposed methods for Heart-rate
20% MCAR.

	The conventional 12 leads											
Method	i	ii	iii	avr	avl	avf	v1	v2	v3	v4	v5	v6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing -data	73.9	63.8	66.78	47.8	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	80.87	68.2	70.55	52.05	56.78	68.15	38.72	44.43	50.53	69.82	73.3	72.22
Mean-sub	71.68	62.55	64.92	42.4	48.25	58.15	28.5	35.72	44.42	63.68	63.38	61.9
List-wise	71.63	62.55	64.85	42.27	48.17	58.17	28.43	35.63	44.35	63.57	63.35	61.85
Linear-reg	74.48	66.47	68.35	44.23	57.317	101.32	73.97	56.28	59.45	84.73	82.92	67.03

5.2.2 QRS Waves

The ventricular depolarization effect can be represented by three waves in ECG signal: Q, R and S waves, (known as QRS complexes). A QRS complex with a measured duration (time interval) of between 0.08 and 0.1 seconds is considered normal. While a QRS complex with an interval between 0.10 and 0.12 seconds is rated intermediate and abnormal if the interval is more than 0.12 seconds, and the QRS has long duration when the electrical signal needs more time to pass through ventricular myocardium, where the amplitudes of QRS represent the polarization and depolarization of the ventricular, and QRS duration is the required time for the signal to pass [25].

The important QRS properties include rise level (Lr), fall level (Lf), rise duration (Tr), and fall duration (Tf), these factors represent the quality of a QRS wave in terms of specifying the ventricular depolarization. The rise and fall levels represent length of edges of R peak on the right and left hand side, respectively, where the rise and fall durations are the required time to move from the Q peak to the R peak and from the R peak to the S peak, respectively [39].

$$Lr = Amplitude\ R\ peak - Amplitude\ Q\ peak$$

$$Lf = Amplitude\ S\ peak - Amplitude\ R\ peak$$

$$Tr = Time\ point\ R\ peak - Time\ point\ Q\ peak$$

$$Tf = Time\ point\ R\ peak - Time\ point\ Q\ peak$$

$$Mean\ Error = mean\ (noisy - ECG\ (QRS\ locations) - ((filtered\ (QRS\ locations))$$

The performance of the $VAR - IM$ method is first evaluated by comparing the effectiveness of missing data imputation on QRS wave properties using both cases of missing data (10% and 20% MCAR) and the complete dataset. Furthermore, the efficacy of missing data imputation is considered in the filtering processing. Figure 5.14 shows the QRS complex rise level, fall level, rise time and fall time in the case of complete data. In comparison, Figures 5.15-5.17 show various results with respect to the case of 10% MCAR. The three methods, namely mean substitution, linear regression imputation and $VAR - IM$ methods were applied to solve the missing data problem here.

Clearly, the three methods obviously generated different results. The mean substitution and $VAR - IM$ methods can impute the missing data with similar results, which are similar to the real data especially the QRS peaks locations. On the other hand, linear regression imputation only gives good results for Lr, Lf, Tr and Tf. List-wise deletion method was excluded from the comparison because it reduces the number of peaks which makes specifying QRS properties impossible.

Table 5.7 and 5.8 summarize the results of the effectiveness of missing data imputation of the four methods for the QRS wave properties in both cases of missing data 10% and 20%, respectively.

As the amount of the missing data is increased from 10% to 20%, the proposed method ($VAR - IM$) gave the best results among all the methods. To some extent as can be noted in both cases of missing data (MCAR 10% and 20%) the mean substitution and linear regression imputation, have similar results.

Table 5.7: Q-R-S wave properties in case of 10% MCAR.

	Data		Imputed data 20% MCAR missing			
	Complete data	Missing -data	VAR-IM	Mean-sub	List-wise	Linear-reg
MeanError_Qwave	-0.004	-	-0.0109	0.0052	-	-0.0417
MeanError_Rwave	0.021	-	0.0243	0.189	-	0.142
MeanError_Swave	-0.0155	-	-0.342	-0.0147	-	-0.0576
avg_riseTime	29	-	28	27.5	-	30
avg_fallTime	56	-	59	59	-	56
avg_riseLevel	1.4419	-	1.424	1.4048	-	1.0171
avg_fallLevel	1.9204	-	1.9165	1.8534	-	1.5867

Table 5.8: Q-R-S wave properties in case of 20% MCAR.

	Data		Imputed data 20% MCAR missing			
	Complete data	Missing -data	VAR-IM	Mean-sub	List-wise	Linear-reg
MeanError_Qwave	-0.004	-	-0.0109	0.0219	-	0.0131
MeanError_Rwave	0.021	-	0.0243	0.2745	-	0.063
MeanError_Swave	-0.0155	-	-0.342	-0.0966	-	-0.0802
avg_riseTime	29	-	28	397.5	-	27
avg_fallTime	56	-	59	55	-	429.5
avg_riseLevel	1.4419	-	1.424	1.1655	-	0.9402
avg_fallLevel	1.9204	-	1.9165	1.5306	-	1.4009

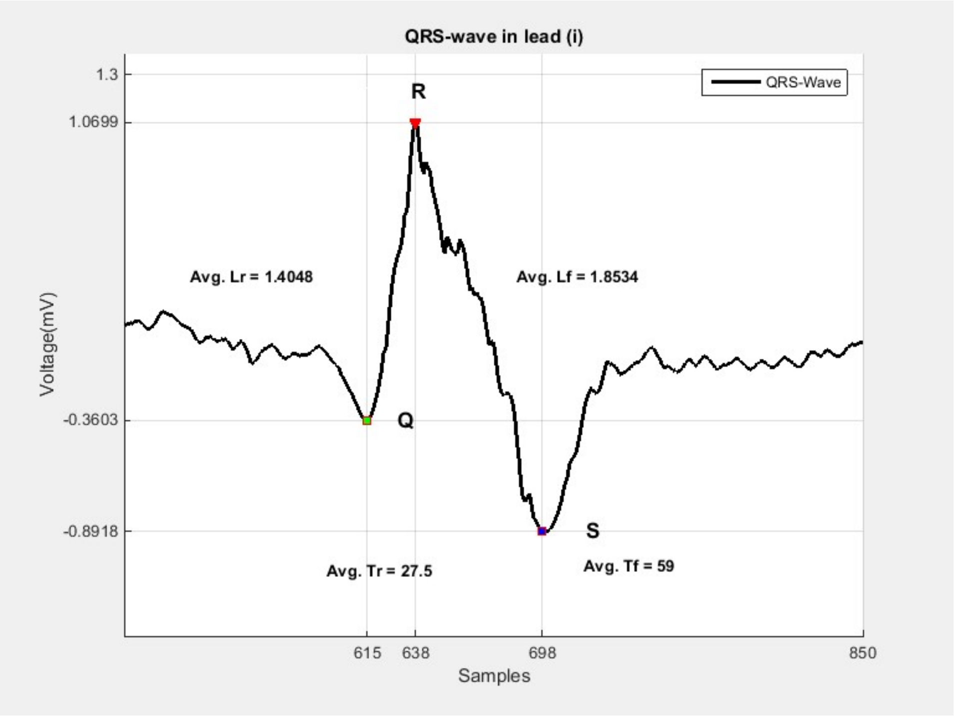

Figure 5.14: QRS wave properties in case of Mean-sub imputed data (10% MCAR).

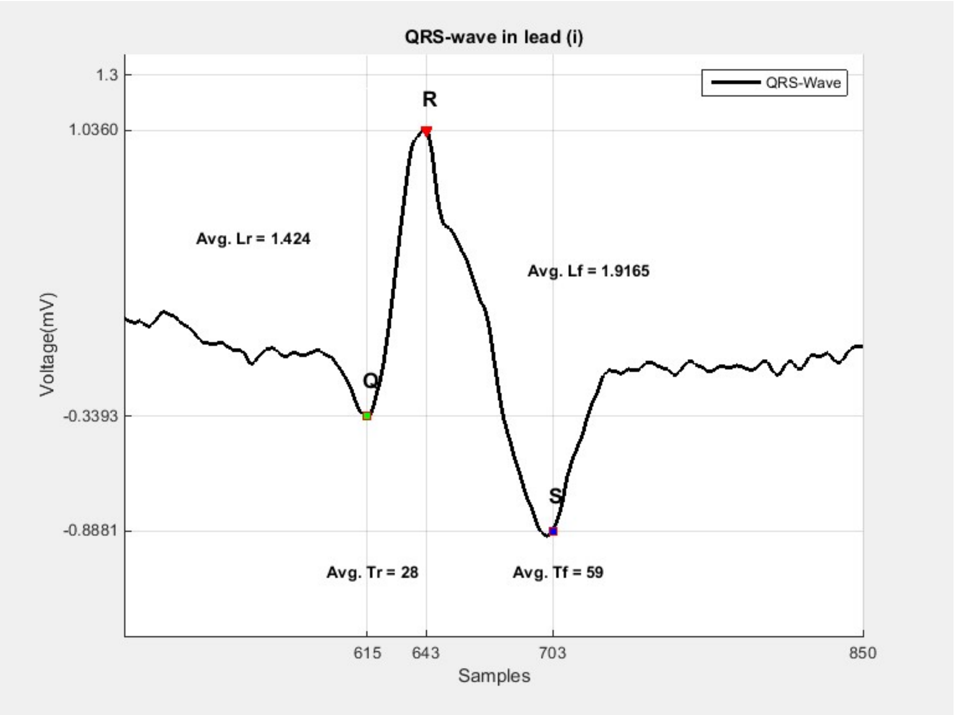

Figure 5.15: QRS wave properties in case of Linear-reg imputed data (10%) MCAR).

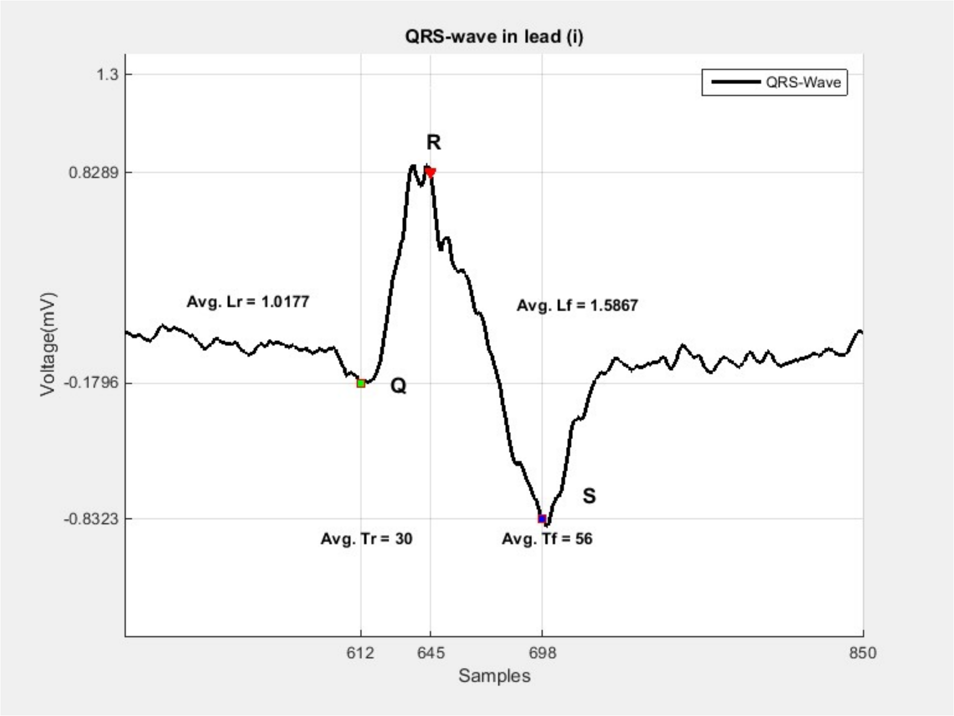

Figure 5.16: QRS wave properties in case of VAR-IM imputed data (20% MCAR).

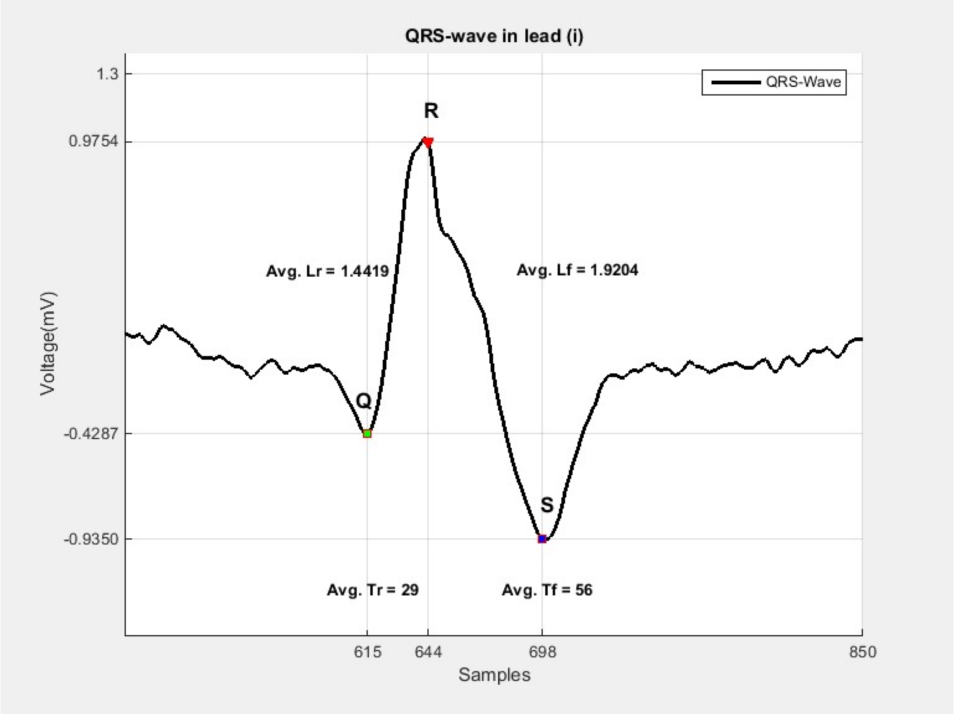

Figure 5.17: QRS wave properties in case of Mean-sub imputed data (20% MCAR).

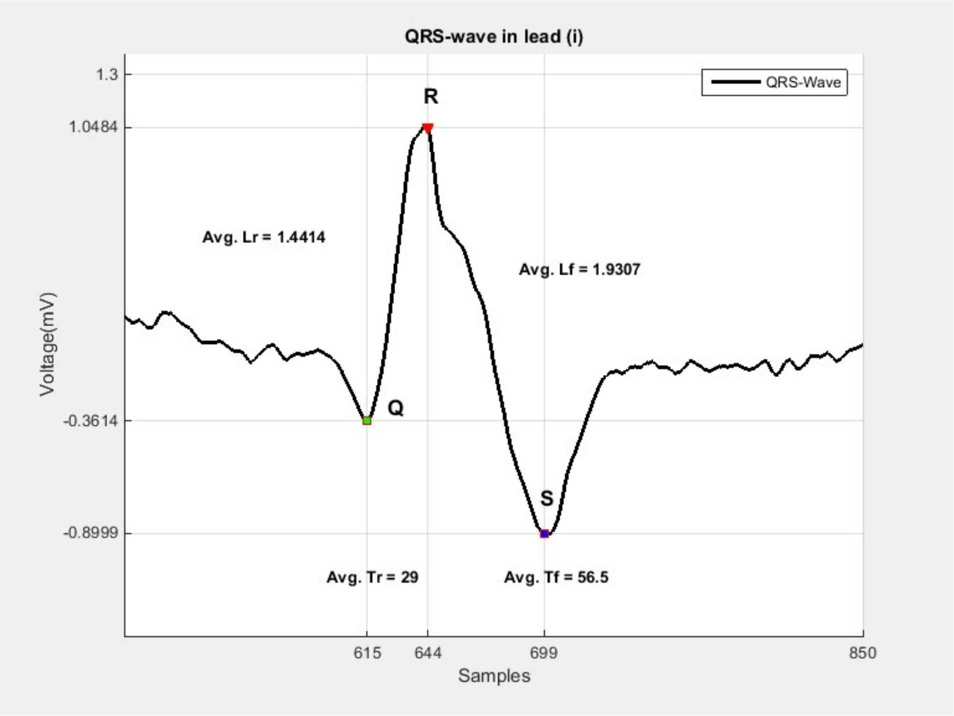

Figure 5.18: QRS wave properties in case of VAR-IM imputed data (20% MCAR).

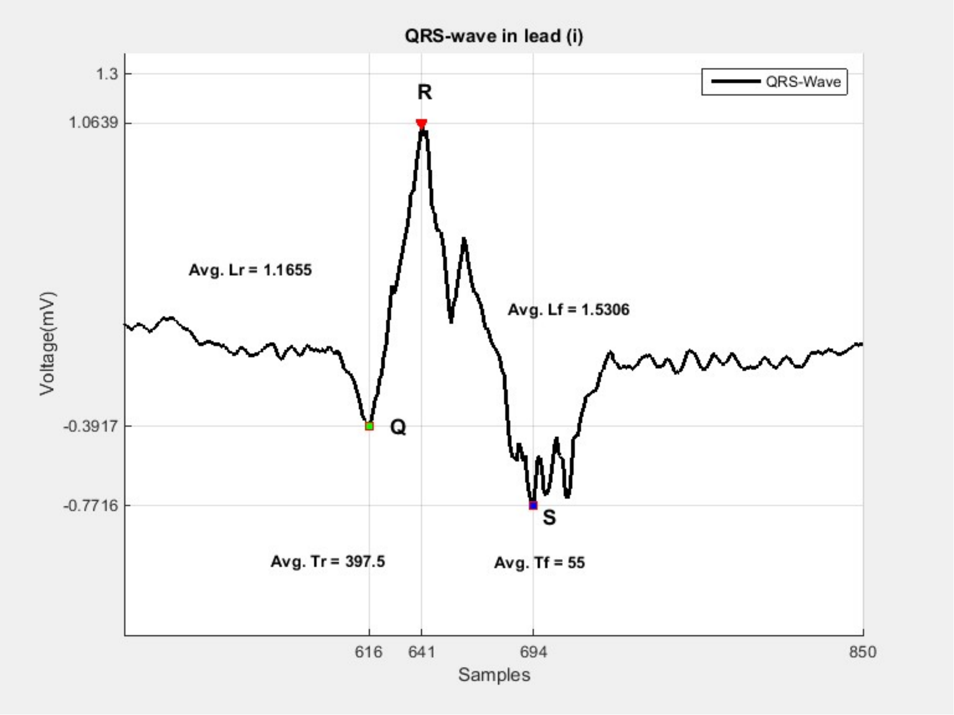

Figure 5.19: QRS wave properties in case of VAR-IM imputed data (20% MCAR).

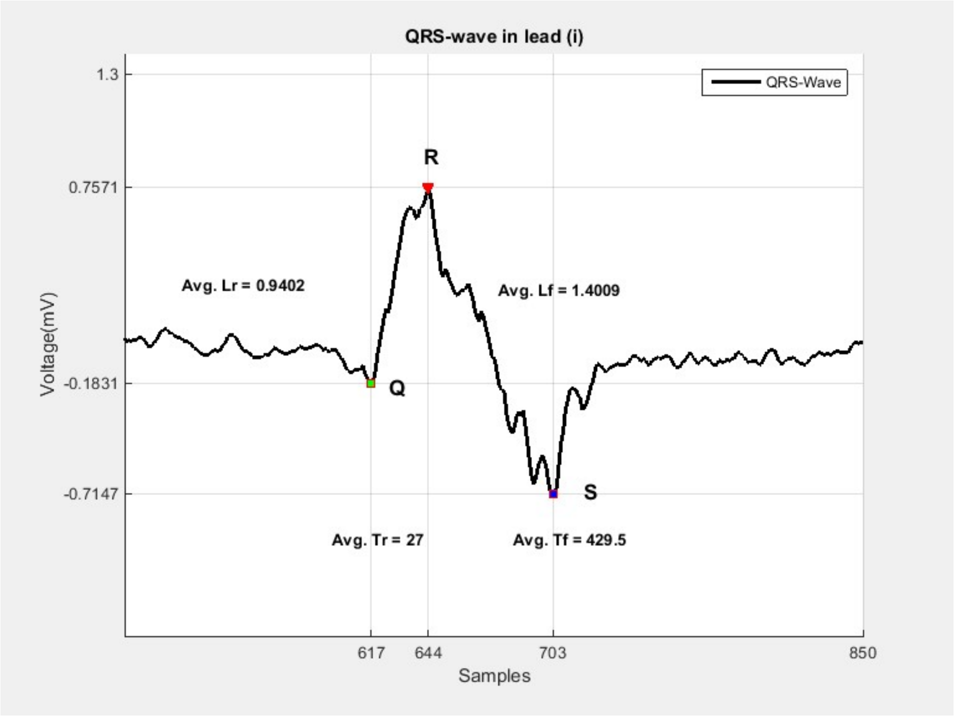

Figure 5.20: QRS wave properties in case of Mean-sub imputed data (20% MCAR).

5.2.3 $VAR - IM$ Versus Modern Method

As a second part of this case study, is to examine the performance of the proposed algorithm in terms of scalability and quality, an evaluation of its effectiveness in recovering missing values is considered. The same dataset of ECG signals as used previously is used, and the proposed algorithm is compared with three modern methods: MARSS, EM, and K-nearest neighbour (KNN).

5.2.3.1 Multivariate Auto-Regressive State-Space

The Multivariate Auto-Regressive State Space (MARSS) model was introduced in 2012 as the first complete package for handling missing data in multivariate time series data [62]. MARSS incorporates an expectation-maximization (EM) algorithm. It is an R package employing a special formula of vector autoregressive state-space models to fit multivariate time series with missing data via an EM algorithm. A MARSS model has the following matrix structure:

$$\begin{cases} x_t = A_t x_{t-1} + B_t b_t + \varepsilon_t \\ \\ y_t = C_t x_{t-1} + D_t d_t + \mu_t \end{cases} \tag{5.1}$$

where $\varepsilon_t \sim \mathcal{N}(0, Q_t)$, $\mu_t \sim \mathcal{N}(0, R_t)$
and $x_1 \sim \mathcal{N}(\pi, \Lambda)$ or $x_0 \sim \mathcal{N}(0, \Lambda)$

The state vector is represented by x_t and the measured value is designated by y_t.

Driven by data, the model evolves but it is possible that some value may be missing when measuring y. The variables b_t and d_t are inputs representing for example some indicators or exogenous variables. A_t, B_t, C_t, and D_t are system matrices, ε_t and μ_t are process and non-process error, respectively, Q_t and R_t are $m \times m$ and $n \times n$ variance-covariance matrices, respectively, where m is number of states and n the number of time series. Compared with the traditional approaches, MARSS can generate better results especially for multivariate time series modelling [60].

5.2.3.2 K-nearest neighbour

The K-nearest neighbour (KNN) imputation method for handling missing values was introduced by [31]. KNN uses the observed values of nearest neighbour time series to fill the corresponding missing values in the time series. The nearest neighbour time series is the closest time series in Euclidean distance. The next nearest time series is utilized, if the corresponding value from the nearest time series was also missing, that means this method does not reduce the length of the time series, which results in a decreased sample size, and does not need to estimate a model to impute the missing value. In contrast, in multivariate time series modelling in which the interaction and variation between data points is important, KNN cannot maintain this property.

Despite these disadvantages, many researchers still extensively use this technique, and in MATLAB, KNN is one of the best options for imputing missing values when estimating dynamic models.

Tables 5.9 and 5.10 show the accuracy for recovering missing data in the heart rate signal using different imputation methods. Table 5.9 shows the 10% MCAR and Table 5.10 shows 20% MCAR.

Tables 5.11 and 5.12 summarize the results of the performance in recovering the missing data using the four imputation methods for the QRS wave properties for both cases of missing data, 10% and 20%, respectively. In both cases, the proposed method $VAR - IM$ gives better results as compared with the other methods.

Table 5.9: Proposed methods for Heart-rate
10% MCAR.

	The conventional 12 leads											
Method	i	ii	iii	avr	avl	avf	v1	v2	v3	v4	v5	v6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing –data	73.9	63.8	66.78	47.8	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	79	67.08	70.13	54.58	55.08	65.48	37.73	43.58	50.48	70.97	73.51	72.13
MAARS	73.98	63.8	66.83	47.87	49.98	60.93	30.51	37.92	45.82	66.32	66.38	64.9
EM	75.37	64.05	67.33	49.3	51.27	61.38	31.31	38.6	48.95	70.57	66.36	69.6
K-nearest	75.75	64.08	67.22	49.95	50.78	61.13	31.06	37.93	45.8	69.93	66.35	66.37

Table 5.10: Proposed methods for Heart-rate 20% MCAR.

Method	The conventional 12 leads											
	i	ii	iii	avr	avl	avf	v1	v2	v3	v4	v5	v6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing -data	73.9	63.8	66.78	47.8	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	80.87	68.2	70.55	52.05	56.78	68.15	38.72	44.43	50.53	69.82	73.3	72.22
MAARS	71.67	62.55	64.93	42.32	48.22	58.15	28.5	35.68	44.42	63.63	63.38	61.9
EM	74.8	63.05	65.77	44.28	51.7	59.13	29.68	36.87	49.83	69.58	64.33	67.95
K-nearest	75.58	63.17	65.87	44.9	49.9	58.25	29.58	35.78	44.38	68.85	63.35	63.23

Table 5.11: Q-R-S wave properties in case of 10% MCAR.

	Data		Imputed data 20% MCAR missing			
	Complete data	Missing -data	VAR-IM	MARSS	EM	K-nearest
MeanError_Qwave	-0.004	NAN	-0.0109	-0.006	-0.0068	0.036
MeanError_Rwave	0.021	NAN	0.0243	0.0277	0.0212	0.3838
MeanError_Swave	-0.0155	NAN	-0.342	-0.0147	-0.018	-
avg_riseTime	29	NAN	28	28.5	28	-
avg_fallTime	56	NAN	59	56.5	57	-
avg_riseLevel	1.4419	NAN	1.424	1.4332	1.4312	-
avg_fallLevel	1.9204	NAN	1.9165	1.9199	1.9296	-

Table 5.12: Q-R-S wave properties in case of 20% MCAR.

	Data		Imputed data 20% MCAR missing			
	Complete data	Missing -data	VAR-IM	MARSS	EM	K-nearest
MeanError_Qwave	-0.004	NAN	-0.0067	0.0016	0.0014	0.0016
MeanError_Rwave	0.021	NAN	0.0191	0.022	0.0163	0.022
MeanError_Swave	-0.0155	NAN	-0.018	-0.0096	-0.0191	-0.0096
avg_riseTime	29	NAN	29	28.5	28.5	28.5
avg_fallTime	56	NAN	56.5	55.5	56.5	55.5
avg_riseLevel	1.4419	NAN	1.4414	1.4391	1.4445	1.4391
avg_fallLevel	1.9204	NAN	1.9307	1.9243	1.9341	1.9243

5.3 Summary

This Chapter has examined using the $VAR - IM$ algorithm with missing data from multivariate time series datasets. The $VAR - IM$ algorithm has been introduced for forecasting the electric flux from solar wind real data at geosynchronous orbit. Numerical results showed that the proposed algorithm for the imputed missing data can produce promising prediction results for the relativistic electron flux. A further extension to this study would be to introduce a relatively complicated non- stationary multivariate series models to improve forecasting performance. It is extremely important to effectively handle multivariate datasets that contain missing values. This is especially true for medical data, which could involve a great number of critical variables that could adversely affect diagnosis of critical health conditions.

The proposed $VAR - IM$ method provides improvements to speed and accuracy for imputing missing values of multivariate time series datasets. It outperforms the commonly used methods such as list wise deletion, linear regression imputation, MARSS and EM algorithms. The results of the case study show that the $VAR - IM$ method provides an effective alternative for the imputation of missing values in multivariate time series. While the other proposed traditional and modern methods become less effective with the increase of the proportion of missing data, $VAR - IM$ shows less deterioration in performance with increasing percentages of missing entries. In addition, the $VAR - IM$ method is more robust than the other proposed techniques when applied to the data types discussed in the case study, and performed better on static and noisy data. Furthermore, a part from this chapter was published as a journal paper in Neurocomputing journal [14].

Chapter 6

Conclusions and Future work

6.1 Conclusions

In conclusion, this work presented new algorithms for imputation and analysis of missing data in static and dynamic formats for univariate and multivariate time series datasets. The proposed methods are applicable to solving missing data problems in many different fields such as medical studies, financial applications, space weather forecast, and chemical process modelling.

The missing data problem occur frequently requiring researchers to handle on a regular basis. Numerous specialists sometimes neglect attention to missing values of time series datasets in their analysis. They revert to ad-hoc techniques or even not considering the effect of the missing data at all. Techniques for missing data analysis are widely available in the case of static data (non- autoregression). On the other hand, methods for handling missing data in dynamic systems are not widely available. This thesis builds a statistical methodology to handle missing data in both cases: static and dynamic data.

This study has explained, introduced and explored particular techniques for handling missing data in time series data sets. It began by reviewing the available methods for dealing with missing values in static data. Highlighting limitations of these methods over the other types of data sets. It then developed and applied new algorithms on static data.

Many methods were presented and their advantages and disadvantages were discussed. In the case of static data and from the literature review, the maximum likelihood method was the preferable method. One of the findings from the review is that there are many successful techniques for handling missing data within static datasets. The basic idea behind my research was to compare these techniques with the developed algorithms in this thesis to verify, if indeed, the proposed methods can solve the problem better. The contribution in my thesis was regarding to develop new algorithms to deal with missing data problems in terms of nonlinear modelling, model selection, parametric and non-parametric estimation. As mentioned above, the main aim was to check the performance of the proposed techniques. It was found that the proposed methods do have better ability to solve the missing data problem involving different missing data mechanisms (MAR and MCAR).

A preponderance of recent practical research on missing data analysis has focused on model parameter estimation using modern statistical methods such as maximum likelihood and multiple imputation. These approaches are superior to traditional methods, such as listwise deletion and mean imputation methods. One benefit of these modern techniques is that they can lead to unbiased parametric estimation in many particular application cases. However, when applied to nonlinear systems, especially those with highly nonlinear behaviour, these methods do not work well. The beginning of Chapter 3 explains the linear parametric estimation method applied to missing data. The chapter includes an overview of biased and unbiased linear parametric estimation with missing data. It also provides accessible descriptions of expectation maximization (EM) algorithm and the Gauss-Newton method. In particular, it was proposed to use a Gauss-Newton method for nonlinear parametric estimation in the case of missing data. Since the Gauss-Newton method needs initial values that are hard to obtain in the presence of missing data, the EM algorithm is thus coupled with Gauss-Newton method to estimate these initial values.

The primary aim of Chapter 3 was to introduce a nonlinear modelling technique for missing data analysis. Comparative studies on both the EM and Gauss-Newton approaches have been carried out. Although EM and Gauss-Newton algorithms offer advantages over traditional approaches, they produce different results specifically in systems exhibiting high nonlinearity with different missing data mechanism (i.e., MAR and different MCAR cases). Most studies in the literature have focused on the use of linear techniques because of their simple assumptions and ease of implementation especially with computers. As mentioned previously, with systems that have high nonlinearity, EM does not always give good results. On the other hand, the Gauss-Newton does need initial values to start the iteration process, and this is a disadvantage in terms of computing time.

Most nonlinear modelling approaches solve the model selection problem with complete data by incorporating nonlinear transforms such as Box-Tidwell and fractional polynomial transformation. Often these approaches can lead to models that are better than traditional models (for example, logistic model and quadratic model). However, in the case of missing data, it is not easy to predict the relationship between the independent and dependent variables. The result is that traditional nonlinear models, as applied to cases of missing data analysis, give poor results. The second part of Chapter 3 explained nonlinear model selection techniques for missing data. It presented the critical issues in choosing the best models for cases of missing data. Two of the most popular model selection methods for incomplete data were illustrated. The illustrations were focused on single variable data modelling for missing data. The basic idea, however, can be extended to multivariable data analysis but the modelling complexity is increased. The key aspects of the Box Tidwell transformation and fractional polynomial methods have been presented and applied these to model estimation for missing data. The comparison of the effect of different missing data mechanisms (10% MCAR and 20% MCAR) on the fractional polynomial, Box Tidwell and traditional models give good indications about the use of fractional polynomial and Box Tidwell methods. As evidenced by the F-test, the cubic, Box Tidwell, and fractional polynomial models are all better, and they imputed the missing values about equally well. The fractional polynomial model did fare better by giving the highest R^2 value. However, complex models are generally less tractable and less robust than simple ones [78, 98–100]

While excellent work has been done in missing data imputation, most available approaches have focused on some particular applications, such as static data and univariate time series. Another unique contribution of this thesis was to develop new algorithms for handling missing data in multivariate time series datasets. An improved technique for handling missing values in multiple time series were presented in Chapter 4, and it introduced an novel algorithms for handling missing data in multivariate time series datasets based on a vector autoregressive (VAR) model by combining an expectation and minimization (EM) algorithm with the prediction error minimization (PEM) method. Case studies were conducted to compare the proposed algorithm with traditional and modern methods for imputing missing data.

The newly proposed $VAR - IM$ method provides a fast and accurate approach of imputing missing values for multivariate time series. The $VAR - IM$ approach outperforms the commonly used methods such as mean substitution, list wise deletion and linear regression imputation. It achieves this by taking advantage of the correlation structure of the data for imputing missing values. From the results of the case study, the $VAR - IM$ method provides an effective alternative for the imputation of missing values in multivariate time series. While mean substitution, list wise deletion and linear regression imputation methods can become less effective with the increase of the proportion of missing data, $VAR - IM$ shows less deterioration in performance with increasing percentages of missing entries. In addition, the $VAR - IM$ method is more robust than the other three methods when applied to the data types discussed in the case studies, and performed better on static and noisy data. However, there are some limitations of the proposed method. First, Chapter 4 only considered the scenario in which data were missing completely at random, that is, the cause of the missing data was independent of both the observed and missing values. A less stringent assumption of the missing data mechanism, missing at random (MAR), may be more realistic in practice. Second, the validity of

$VAR - IM$ requires that the time series should be stationary. Finally, the percentage of missing data has a significant impact on most missing data analysis methods, $VAR - IM$ does not have the priority to be used if the percentage of missing data is quite low (say less 10%). Despite these limitations, $VAR - IM$ provides an important alternative to existing methods for handling missing data in multivariate time series.

Two cases studies were conducted in Chapter 5: one for space weather data and another for ECG data to compare the proposed algorithm with different methods for imputing missing data. Missing data analysis, multivariate time series, and vector autoregressive models have been introduced for forecasting the electric flux from solar wind real data at geosynchronous orbit. Numerical results show that the proposed vector autoregressive models estimated by using the imputed data can produce promising prediction results for the relativistic electric flux. The ECG data set was used as a benchmark to test the performance and limitations of the proposed methods. For these case studies, the first decision is to determine whether or not if data imputation is even necessary at all. If there is no strong evidence that data imputation can improve the data analysis result, then simply choose not to impute. Although imputed values are usually well behaved and appear to be consistent with other attribute values, an imputation procedure can be potentially harmful because even the most advanced imputation method is only able to approximate the actual missing value. Missing data imputation should be carefully applied to reduce the risk of oversimplifying the problem of missing data mechanism.

6.2 Future work

This work has produced a significant contribution towards imputation of missing data of static and dynamic systems. It also gives a starting point for further work in the field. Although the contributions in this thesis can be applied to many fields, there are still several important questions to be answered.

- The missing data mechanisms were not completely covered. This thesis assumed the data to be missing at random and missing completely at random. To what extent do these assumptions can affect the imputation processes needs to be examined. In other words, if the missing data mechanism were something other than missing at random or missing completely at random, would the proposed algorithm provide benefit?

- The other extension can be concluded from chapter 3, where missing data analysis for static data were used for model selection and parametric regression. The proposed algorithms need to be updated to be used for the case of dynamic data, particularly with multivariate time series.

- The primary aim in Chapter four was to present multivariate time series analysis for the case of incomplete data. The $VAR(p)$ model was nominated and chosen as the best model for missing data imputation for that case. The use of the other models for missing data imputation should be investigated.

- Chapter four introduces a new method for handling missing data in multivariate time series ($VAR - IM$). Although the VAR-IM approach outperforms the commonly used methods and it provides an effective alternative for the imputation of missing values in multivariate time series, there are some challenges and limitation's need to be overcome. First, this thesis only considered the scenario in which data were missing completely at random. A less stringent assumption of missing data mechanism, such as missing at random, may be more realistic in practice. Second, the validity of $VAR - IM$ requires that the time series should be stationary. Finally, while the percentage of missing data has significant impact on most missing data analysis methods, $VAR - IM$ should have a low priority to be used if the percentage of missing data is quite low (say less 10%).

Bibliography

[1] A. Abadie and G.W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[2] A.C. Acock. Working with missing values. *Journal of Marriage and Family*, 67(4):1012–1028, 2005.

[3] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

[4] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[5] P.D. Allison. Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55 (1):193–196, 2002.

[6] P.D. Allison. Handling missing data by maximum likelihood. In *SAS global forum*, volume 312, 2012.

[7] T.W. Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the american Statistical Association*, 52(278):200–203, 1957.

[8] J.L. Arbuckle and W. Wothke. *Amos 4.0 user's guide.* SmallWaters Corporation Chicago, IL, 1999.

[9] B. Azar. Finding a solution for missing data. *Monitor on Psychology*, 33(7):70–1, 2002.

[10] V. Bagnardi, A. Zambon, P. Quatto, and G. Corrao. Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. *American Journal of Epidemiology*, 159(11):1077–1086, 2004.

[11] M. Bańbura and M. Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160, 2014.

[12] A.N. Baraldi and C.K. Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

[13] F. Bashir and H.L. Wei. Parametric and non-parametric methods to enhance prediction performance in the presence of missing data. In *2015 19th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 337–342. IEEE, 2015.

[14] F. Bashir and H.L. Wei. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. *Neurocomputing*, 2017.

[15] F. Bashir and H.L. Wei. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. *Neurocomputing*, 276:23–30, 2018.

[16] F. Bashir, H.L. Wei, and A. Benomair. Model selection to enhance prediction performance in the presence of missing data. In *2015 20th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 846–850. IEEE, 2015.

[17] F.A. Bashir and H.L. Wei. Using nonlinear models to enhance prediction performance with incomplete data. In *ICPRAM (1)*, pages 141–148, 2015.

[18] G.E. Batista and M.C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

[19] J.A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):21–97, 1998.

[20] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318, 1995.

[21] G.E. Box and G.C. Tiao. A canonical analysis of multiple time series. *Biometrika*, 64(2):355–365, 1977.

[22] G.E. Box and P.W. Tidwell. Transformation of the independent variables. *Technometrics*, 4(4):531–550, 1962.

[23] R.A. Boyles. On the convergence of the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1(1):47–50, 1983.

[24] P.J. Brockwell and R.A. Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

[25] N. Burns. Cardiovascular physiology. retrieved from school of medicine, trinity college, dublin, 2013.

[26] W.B. Cade and C. Chan-Park. The origin of a space weather. *Space Weather*, 13(2):99–103, 2015.

[27] G.C. Chow and A.l. Lin. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics*, 1(1):372–375, 1971.

[28] Y.G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, A. Ait-Bachir, and V. Strijov. Time series forecasting using rnns: an extended attention mechanism to model periods and handle missing values. *arXiv preprint arXiv:1703.10089*, 2017.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1(1):1–38, 1977.

[30] K. Deng, A.W. Moore, and M.C. Nechyba. Learning to recognize time series: Combining arma models with memory-based learning. In *1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings,*, pages 246–251. IEEE, 1997.

[31] J.K. Dixon. Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(10):617–621, 1979.

[32] N.R. Draper and H. Smith. Applied regression analysis 2nd ed. 1981.

[33] G.L. Edgett. Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association*, 51(273):122–131, 1956.

[34] E. Ekheden and O. Hössjer. Multivariate time series modeling, estimation and prediction of mortalities. *Insurance: Mathematics and Economics*, 65(1):156–171, 2015.

[35] C.K. Enders. A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic medicine*, 68(3): 427–436, 2006.

[36] C.K. Enders. *Applied missing data analysis*. Guilford Press, 2010.

[37] C.K. Enders and D.L. Bandalos. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3):430–457, 2001.

[38] R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

[39] O. Escalona, G. Passariello, and F. Mora. An algorithm for microprocessor-based qrs detection. *Journal of Clinical Engineering*, 11(3):213–220, 1986.

[40] J. Fan and L.S. Huang. Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association*, 96(454):640–652, 2001.

[41] U. Feldman, E. Landi, and N. Schwadron. On the sources of fast and slow solar wind. *Journal of Geophysical Research: Space Physics*, 110 (A7):140–152, 2005.

[42] C. Finkbeiner. Estimation for the multiple factor model when data are missing. *Psychometrika*, 44(4):409–420, 1979.

[43] E. Frank. An accurate, clinically practical system for spatial vectorcardiography. *circulation*, 13(5):737–749, 1956.

[44] Y.C. Gao, Y. Zeng, and S.M. Cai. Influence network in the chinese stock market. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(3):P03017, 2015.

[45] L. Gardes and S. Girard. A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99(10):2368–2388, 2008.

[46] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000.

[47] C. Gourieroux and A. Monfort. *Time series and dynamic models*, volume 3. Cambridge University Press, 1997.

[48] J.W. Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60(1):549–576, 2009.

[49] J.W. Graham. *Missing data: Analysis and design*. Springer Science & Business Media, 2012.

[50] J.W. Graham and S.M. Hofer. *Multiple imputation in multivariate research.* Lawrence Erlbaum Associates Publishers, 2000.

[51] C.W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 1(1):424–438, 1969.

[52] C.W. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2(1):329–352, 1980.

[53] C.W. Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2):199–211, 1988.

[54] E.J. Hannan and B.G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1(1):190–195, 1979.

[55] E. Hannan, W. Dunsmuir, and M. Deistler. Estimation of vector armax models. *Journal of Multivariate Analysis*, 10(3):275–295, 1980.

[56] J.D. Hart. *Nonparametric smoothing and lack-of-fit tests*. Springer, 1997.

[57] H. Hartley and R. Hocking. The analysis of incomplete data. *Biometrics*, pages 783–823, 1971.

[58] D.M. Hassler, I.E. Dammasch, P. Lemaire, P. Brekke, W. Curdt, H.E. Mason, J.C. Vial, and K. Wilhelm. Solar wind outflow and the chromospheric magnetic network. *Science*, 283(5403):810–813, 1999.

[59] J.J. Heckman. *Statistical models for discrete panel data*. Department of Economics and Graduate School of Business, University of Chicago Chicago, IL, 1979.

[60] E. Holmes, E. Ward, and M. Scheuerell. Analysis of multivariate time-series using the marss package. *User guide: http://cran. r-project. org/web/packages/MARSS/vignettes/UserGuide. pdf*, 2014.

[61] E.E. Holmes. Derivation of an em algorithm for constrained and unconstrained multivariate autoregressive state-space (marss) models. *arXiv preprint arXiv:1302.3919*, 2013.

[62] E.E. Holmes, E.J. Ward, and K. Wills. Marss: Multivariate autoregressive state-space models for analyzing time-series data. *R journal*, 4(1), 2012.

[63] J. Honaker and G. King. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2): 561–581, 2010.

[64] N.J. Horton and K.P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

[65] N.J. Horton and S.R. Lipsitz. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254, 2001.

[66] R. Hyndman and A. Zeileis. Cran task view: Time series analysis, 2016.

[67] T. Iyemori and D. Rao. Decay of the dst field of geomagnetic disturbance after substorm onset and its implication to storm-substorm relation. 1996.

[68] S. Johansen. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2-3):231–254, 1988.

[69] R.E. Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[70] G. King, J. Honaker, and M. Blackwell. Amelia ii: A program for missing data, 2007.

[71] J. Liang and P.M. Bentler. An em algorithm for fitting two-level structural equation models. *Psychometrika*, 69(1):101–122, 2004.

[72] R.J. Little and D.B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[73] S. Liu and P.C. Molenaar. ivar: A program for imputing missing data in multivariate time series using vector autoregressive models. *Behavior research methods*, 46(4):1138–1148, 2014.

[74] Y. Liu. Scalable multivariate time-series models for climate informatics. *Computing in Science & Engineering*, 17(6):19–26, 2015.

[75] L. Ljung. Prediction error estimation methods. *Circuits, Systems and Signal Processing*, 21(1):11–21, 2002.

[76] F.M. Lord. Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, 50(271):870–876, 1955.

[77] T.A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[78] A. Luke, R. Durazo-Arvizu, C. Rotimi, T.E. Prewitt, T. Forrester, R. Wilks, O.J. Ogunbiyi, D.A. Schoeller, D. McGee, and R.S. Cooper. Relation between body mass index and body fat in black population samples from nigeria, jamaica, and the united states. *American journal of epidemiology*, 145(7):620–628, 1997.

[79] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[80] H. Lütkepohl and M. Krätzig. *Applied time series econometrics*. Cambridge university press, 2004.

[81] P. McCullagh and J.A. Nelder. Generalized linear models. 1989.

[82] G.K. McLachlan. *The EM Algorithm and Extensions*. John Wiley, 1996.

[83] X.L. Meng and D.B. Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

[84] D. Montgomery, E. Peck, and G. Vining. *Diagnostics for Leverage and Influence: Introduction to Linear Regression Analysis . Hoboken*. NJ: Wiley-Interscience, 2006.

[85] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork. Comparison of different methods for univariate time series imputation in r. *arXiv preprint arXiv*, 1(1):1–20, 2015.

[86] F. Mosteller and J.W. Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[87] B. Muthén and K. Shedden. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2):463–469, 1999.

[88] R.M. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse and other variants. (1):355–368, 1998.

[89] M.C. Neale, S.M. Boker, G. Xie, and H.M. Maes. Statistical modeling. *Richmond, VA: Department of psychiatry, virginia commonwealth university*, 1999.

[90] D.R. Osborn. Exact and approximate maximum likelihood estimators for vector moving average processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1(1):114–118, 1977.

[91] C. Oster. Improving ecg trace quality. *Biomedical instrumentation & technology*, 34(3):219–222, 1999.

[92] B. Pesaran and M.H. Pesaran. *Time series econometrics using Microfit 5.0: A user's manual*. Oxford University Press, Inc., 2010.

[93] J.L. Peugh and C.K. Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, 74(4):525–556, 2004.

[94] P.C. Phillips and W. Ploberger. Posterior odds testing for a unit root with data-based model selection. *Econometric Theory*, 10(3-4):774–808, 1994.

[95] S.W. Raudenbush and A.S. Bryk. *Hierarchical linear models: Applications and data analysis methods*. Sage, 2002.

[96] G.C. Reinsel. *Elements of multivariate time series analysis*. Springer Science & Business Media, 2003.

[97] J. Roy. Causality: Statistical perspectives and applications by carlo berzuini, philip dawid, luisa bernardinelli, 2014.

[98] P. Royston and D.G. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 1(1):429–467, 1994.

[99] P. Royston and D.G. Altman. Approximating statistical functions by using fractional polynomial regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):411–422, 1997.

[100] P. Royston and W. Sauerbrei. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons, 2008.

[101] D.B. Rubin. Inference and missing data. *Biometrika*, 1(1):581–592, 1976.

[102] D.B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[103] J.L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 2010.

[104] J.L. Schafer and J.W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[105] G.L. Schlomer, S. Bauman, and N.A. Card. Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, 57(1):1, 2010.

[106] G. Schwarz. Estimating the dimension of a model the annals of statistics. *The Annals of Statistics*, 2(6):461–464, 1978.

[107] S.R. Seaman and I.R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013.

[108] R.H. Shumway and D.S. Stoffer. Time series analysis and its applications. *Studies In Informatics And Control*, 9(4):375–376, 2000.

[109] G.K. Smyth. Nonlinear regression. *Encyclopedia of environmetrics*, 2002.

[110] E.C. Stone, A. Frandsen, R. Mewaldt, E. Christian, D. Margolies, J. Ormes, and F. Snow. The advanced composition explorer. *Space Science Reviews*, 86(4):1–22, 1998.

[111] B.G. Tabachnick, L.S. Fidell, and S.J. Osterlind. *Using multivariate statistics*. Allyn and Bacon Boston, 2001.

[112] A. Ukhorskiy, M. Sitnov, A. Sharma, B. Anderson, S. Ohtani, and A. Lui. Data-derived forecasting model for relativistic electron intensity at geosynchronous orbit. *Geophysical research letters*, 31(9), 2004.

[113] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2012.

[114] M. Watanabe and K. Yamaguchi. *The EM algorithm and related statistical models*. CRC Press, 2003.

[115] W.W. Wei et al. *Time series analysis: univariate and multivariate methods*. Pearson Addison Wesley, 2006.

[116] P. Whittle. *Prediction and regulation by linear least-square methods*. English Univ. Press, 1963.

[117] C.J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, 55(4):95–103, 1983.

[118] A.M. Yaglom. *An introduction to the theory of stationary random functions*. Courier Corporation, 2004.

[119] Y.C. Yuan. Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49(1):1–11, 2010.

[120] S.B. Zaghlool and C.L. Wyatt. Missing data estimation in fmri dynamic causal modeling. *Frontiers in neuroscience*, 8(191):1–12, 2014.

[121] E. Zivot and J. Wang. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*, 1(1):385–429, 2006.