METHODS OF HANDLING MISSING DATA WITH REFERENCE TO RAINFALL IN PENINSULAR MALAYSIA

HO MING KANG

UNIVERSITI TEKNOLOGI MALAYSIA

METHODS OF HANDLING MISSING DATA WITH REFERENCE TO RAINFALL IN PENINSULAR MALAYSIA

HO MING KANG

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

SEPTEMBER 2014

To My Beloved Family

ACKNOWLEDGEMENT

I would like to express my deepest appreciation and thankful to my main supervisor, PM Dr. Fadhilah Yusof and co-supervisor, PM Dr. Ismail Mohamad for their supports. Throughout the process of completing this thesis, they have given me a lot of supervision, criticism, advice and guidance. Without their encouragements and enthusiasm, this thesis would not been the same as presented here.

Apart of my both supervisors, I also would like to thank Department of Irrigation and Drainage Malaysia in providing the rainfall data. Sincere grateful and appreciation are extended for the fund and sponsorship from Bajet Mini 2009 and Zamalah Scholarship, Universiti Teknologi Malaysia. Their financial supports are gratefully acknowledged.

Last but not least, I need send my gratitude to my lovely parents and family members who continuingly given me a lot of understanding, patience, tolerating and motivating. To them, I dedicate this thesis.

ABSTRACT

Missing data is one of the issues often discussed amongst hydrologists in Malaysia. Various imputation methods were introduced to help minimize the bias and improve the accuracy of the statistical analysis. However, the performances of the imputation methods will be affected if the reason for data being missing is unidentified. Therefore, this study objectively investigates the reasons why some data is missing, known as missingness mechanism, and selects the best model to impute the missing rainfall data. A model using a combination of expectation maximization and logit (EM-Logit) is proposed and applied to a simulated data with missing values that are characterised as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Besides, homogeneous rainfall data that are coupled with temperature and humidity in Damansara and Kelantan are also used before validating the proposed model. The results indicate that the model is able to identify types of missingness mechanism which leads to a data being missing. The results of the model has also identified that the MNAR is best missingness mechanism to describe missing rainfall data in both study areas. Therefore, for the imputation purposes, a two-step approach is proposed. The first step is to analyze the rainfall events, either wet or dry day, by using weighted-average algorithm and the subsequent step is the wet-classified day with missing data is estimated by self-organizing map (SOM). The two-step approach, also known as Probability Density Function Preserving Approach with SOM (PDSOM), is then compared with SOM model alone and Multilayer Perceptron (MLP). By using the mean absolute error (MAE) and root mean square error (RMSE) criteria and comparing the statistical properties of the imputed data with the rainfall data, PDSOM is found to be performing better than SOM and MLP. The missing rainfall data from 1996 to 2004 from the two stations (Damansara and Kelantan) are also selected to validate the performance of PDSOM by comparing the estimated mean and variance of the rainfall data with missing values that are imputed by PDSOM. The imputations are found within the confidence interval that are constructed under observed rainfall data. PDSOM has shown its capability to well preserve the mean and variance of the missing rainfall data, as well as the number of rainfall events in Damansara and Kelantan. Thus, PDSOM can be an alternative imputation model in dealing with rainfall data with missing values.

ABSTRAK

Kehadiran data ketakdapatan adalah salah satu isu yang sering dibincang di kalangan para hidrologi di Malaysia. Pelbagai kaedah telah diperkenalkan untuk mengurangkan ralat dalam penganggaran dan meningkatkan kejituan data anggaran. Namun, prestasi kaedah akan terganggu jika penyebab data ketakdapatan tidak diketahui. Oleh itu, tujuan utama kajian ini adalah untuk mencari faktor-faktor kehilangan data dan mengenal pasti kaedah yang paling efektif untuk mengimput data hujan ketakdapatan. Model hibrid Pemaksimuman Jangkaan dan Logit (EM-Logit) dicadang dan diguna dalam simulasi data yang bermekanisme Ketakdapatan Secara Rawak Sepenuhnya (MCAR), Ketakdapatan Secara Rawak (MAR) dan Ketakdapatan Secara Tak Rawak (MNAR). Selain itu, data hujan yang homogen bersama dengan data suhu dan kelembapan di Damansara dan Kelantan juga digunakan dalam kajian ini. Berdasarkan keputusan yang didapati, model ini dapat membezakan jenis-jenis mekanisme yang diimplikasi dalam data dengan tepat dan juga mengesahkan penyebab data hilang di Damansara dan Kelantan adalah berkait rapat dengan mekanisme MNAR. Justeru, pendekatan dua langkah diperkenalkan untuk tujuan menganggar data ketakdapatan. Langkah pertama meramal keadaan hujan, jika basah atau kering dengan menggunakan algoritma purata berwajaran sebelum langkah menganggar data ketakdapatan berasaskan Peta Swaorganisasi (SOM). Pendekatan dua langkah yang juga dinamai sebagai Pendekatan Pemeliharaan Fungsi Ketumpatan Kebarangkalian dengan gabungan SOM (PDSOM) telah dibandingkan dengan kaedah-kaedah yang sedia ada, iaitu perceptron berlapis (MLP) dan SOM. Perbandingan model telah dijalankan dengan menggunakan min ralat mutlak (MAE), ralat punca min kuasa dua (RMSE) dan juga membandingkan ciri statistik dalam data input dengan data hujan yang lengkap. Kemampuan dan kemantapan model terbaik ditentukan apabila data hujan yang hilang dari tahun 1996 ke tahun 2004 dari kedua-dua stesen (Damansara dan Kelantan) digunakan dengan membandingkan di antara min dan varians anggaran data hujan yang telah diimput oleh PDSOM. Data imput diperolehi dalam selang keyakinan yang dibentuk di dalam data hujan yang dicerap. Keputusan telah membuktikan bahawa PDSOM adalah lebih cekap daripada MLP dan SOM. Model ini juga mampu memelihara ciri statistik dalam data hujan, terutamanya bilangan hari hujan dan kering serta taburan hujan di Damansara dan Kelantan. Oleh itu, PDSOM boleh bertindak sebagai model alternatif dalam menangani masalah data ketakdapatan.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	**ii**
	DEDICATION	**iii**
	ACKNOWLEDGEMENTS	**iv**
	ABSTRACT	**v**
	ABSTRAK	**vi**
	TABLE OF CONTENTS	**vii**
	LIST OF TABLES	**xi**
	LIST OF FIGURES	**xv**
	LIST OF ABBREVIATIONS	**xviii**
	LIST OF SYMBOLS	**xix**
	LIST OF APPENDICES	**xxi**
1	**INTRODUCTION**	**1**
	1.1 Background of Study	1
	1.2 Statement of Problem	3
	1.3 Objectives	5
	1.4 Scope of Study	5
	1.5 Significance of Study	6
	1.6 Organization of Thesis	7
2	**LITERATURE REVIEW**	**8**
	2.1 Introduction	8
	2.2 Homogeneity Tests	9

2.3 Missingness Mechanism 11

2.4 Distribution of Rainfall 14

2.5 Development of the Imputation Methods 17

2.5.1 Classical Imputation Models 17

2.5.2 Modern Imputation Models 18

2.6 Summary 25

3 MISSINGNESS MECHANISM 28

3.1 Introduction 28

3.2 Study Area 29

3.3 Introduction to Missingness Mechansim 32

3.4 Expectation Maximization (EM) 33

3.4.1 General Description of EM 34

3.5 Data Simulation with Different Missingness Mechansim 36

3.6 Algorithm of EM for Imputation of Missing Data 38

3.7 EM-Logit Model 42

3.7.1 Bivariate EM-Logit Model 43

3.7.2 Multivariate EM-Logit Model 52

3.8 Standard Error 62

3.9 Homogeneity Tests 63

3.10 Application of EM-Logit Model 66

3.10.1 Box-Cox Transformation 67

3.10.2 Anderson-Darling Normality Test 68

3.7 Summary 68

4 IMPUTATION OF MISSING DATA 71

4.1 Introduction 71

4.2 Introduction to Neural Network 72

4.3 Multilayer Perceptron 75

4.4 Self-Organizing Map 78

4.4.1 Example of SOM 85

4.5 Proposed Probability Density Function Preserving Approach with Self-Organizing Map Imputation Model (PDSOM) 89

4.5.1 Example of Determination Threshold 93

4.6 L-Moments 96

4.7 Best-fit Distribution for Rainfall Patterns in Malaysia 98

4.7.1 Wakeby Distribution 98

4.7.2 Generalized Extreme Value (GEV) Distribution 99

4.7.3 Two-parameter Weibull Distribution 101

4.7.4 Kolmogorov-Smirnov Goodness-of-Fit Test 102

4.8 Evaluation of Models 103

4.8.1 Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) 103

4.8.2 Confidence Interval of Mean 104

4.8.3 Confidence Interval of Variance 105

4.9 Summary 106

5 MISSINGNESS MECHANISM RESULTS AND ANALYSIS 109

5.1 Introduction 109

5.2 Results of Homogeneity Tests 110

5.3 Simulation Results for Missingness Mechanism 115

5.4 Box-Cox Transformation and Identification of Missingness Mechanism 135

5.5 Summary 141

6 IMPUTATION OF MISSING DATA RESULTS AND ANALYSIS 145

6.1 Introduction 145

6.2 Best-fit Distribution 146

6.3 Evaluation of MLP Performance 156

6.3.1 Selection of Best Architecture of MLP 157

6.4 Evaluation of SOM Performance 159

6.4.1 Clustering Analysis 161

6.5 Imputation of Missing Data 163

6.5 Imputation by PDSOM 178

6.6 Summary 192

7 CONCLUSION AND RECOMMENDATIONS 194

7.1 Introduction 194

7.2 Conclusion for Missingness Mechanism 195

7.2 Conclusion for Imputation of Missing Data 197

7.3 Recommendations for Future Research 199

REFERENCES 200

Appendices A – E 208 – 255

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Summary for the existing models of missingness mechanism and Artificial Neural Network (ANN)	27
3.1	Location, descriptive statistics and percentage of missing data from 1996 to 2004 for stations in Damansara	30
3.2	Location, descriptive statistics and percentage of missing data from 1996 to 2004 for stations in Kelantan	30
3.3	Values for abscissa and weights in 6-points Gauss Hermite	39
4.1	Terminology of biological and artificial neural network	72
4.2	Terminology of statistics equivalent to ANN (adapted from Karlatis and Vlagianni, 2011)	74
4.3	Comparison of ANN and statistics (adapted from Karlatis and Vlagianni, 2011)	74
5.1	Results of several homogeneity tests based on annual mean rainfall for the stations in Damansara and Kelantan basins	111
5.2	Results of several homogeneity tests based on annual maximum rainfall for the stations in Damansara and Kelantan basins	112
5.3	Results of several homogeneity tests based on annual median rainfall for the stations in Damansara and Kelantan basins	113
5.4	Guideline to classify missingness mechanism	118
5.5	Estimated parameters (standard error) of EM-Logit model and types of missingness mechanism when correlation between *X* and *Y*, *X* and *Z* are 0.25, 0.5 and 0.75 with 10% missing data	120

5.6 Estimated parameters (standard error) of EM-Logit model and types of missingness mechanism when correlation between *X* and *Y*, *X* and *Z* are 0.25, 0.5 and 0.75 with 20% missing data 122

5.7 Estimated parameters (standard error) of EM-Logit model and types of missingness mechanism when correlation between *X* and *Y*, *X* and *Z* are 0.25, 0.5 and 0.75 with 50% missing data 125

5.8 Results of AD test for rainfall data in Damansara (smallest value of AD is given in bold) 137

5.9 Results of AD test for rainfall data in Kelantan (smallest value of AD is given in bold) 137

5.10 Results of AD test for temperature and humidity data in Damansara (smallest value of AD is given in bold) 138

5.11 Results of AD test for temperature and humidity data in Kelantan (smallest value of AD is given in bold) 138

5.12 Estimated parameters (standard error) of EM-Logit model with types of missingness mechanism for stations in Damansara 139

5.13 Estimated parameters (standard error) of EM-Logit model with types of missingness mechanism for stations in Kelantan 140

5.14 Summary of the model performance when identifying the missingness mechanism 143

6.1 Results of KS test for stations in Damansara from 1996 to 2004 147

6.2 Results of KS test for stations in Kelantan from 1996 to 2004 149

6.3 Learning rate selection with corresponding MAE and RMSE 157

6.4 Momentum constant selection with corresponding MAE and RMSE 157

6.5 Number of hidden nodes selection with corresponding MAE and RMSE 158

6.6 Quantization and topographic error in Damansara and Kelantan 160

6.7 Mean, variance and percentage of missing data (%) for daily complete rainfall data in Damansara and Kelantan from 1996 to 2004 163

6.8 Results of MAE and RMSE of MLP, SOM and PDSOM in Damansara 166

6.9 Results of MAE and RMSE of MLP, SOM and PDSOM in Kelantan 166

6.10 Number of wet and dry events between observed and reconstructed rainfall data by PDSOM in Damansara 167

6.11 Number of wet and dry events between observed and reconstructed rainfall data by PDSOM in Kelantan 167

6.12 Comparison of mean and variance between MLP, SOM and PDSOM for stations (denoted as Sta. No.) in Damansara 170

6.13 Comparison of mean and variance between MLP, SOM and PDSOM for stations (denoted as Sta. No.) in Kelantan 170

6.14 Results of KS test for daily rainfall amount on wet days in Damansara and Kelantan 174

6.15 The comparison between SOM and PDSOM when the actual rainfall amount in Damansara is Weibull distributed using KS test 175

6.16 The comparison between SOM and PDSOM when the actual rainfall amount in Damansara is Wakeby distributed using KS test 175

6.17 The comparison between SOM and PDSOM when the actual rainfall amount in Kelantan is Weibull distributed using KS test 176

6.18 The comparison between SOM and PDSOM when the actual rainfall amount in Kelantan is Wakeby distributed using KS test 177

6.19 Comparison of number of dry events (hit rate) between reconstructed rainfall data by PDSOM and observed rainfall data for stations in Damansara 182

6.20 Comparison of number of dry events (hit rate) between reconstructed rainfall data by PDSOM and observed rainfall data for stations in Kelantan 183

6.21 Confidence interval of mean and variance for observed rainfall data in Damansara from 1996 to 2004 185

6.22 Estimated mean, variance for stations in Damansara after the missing rainfall data was imputed by PDSOM for the period 1996 to 2004 (* represents the estimation is out of the C.I.; - represents no missing data) 187

6.23 Confidence interval of mean and variance for observed rainfall data in Kelantan from 1996 to 2004 189

6.24 Estimated mean, variance for stations in Kelantan after the missing rainfall data was imputed by PDSOM for the period 1996 to 2004 (* represents the estimation is out of the C.I.; - represents no missing data) 191

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	Types of missing data pattern ($\times$ is missing data)	11
3.1	Location of stations (station code) in Damansara and Kelantan	31
3.2	General flow chart of the EM algorithm	34
3.3	Flow chart of the theoretical framework	69
3.4	Framework of EM-Logit model	70
4.1	Biological neural network	72
4.2	Simple architecture of MLP	76
4.3	Simple architecture of SOM	79
4.4	Types of Kohonen map	80
4.5	Learning rates as functions of time	82
4.6	Neighbourhood function	83
4.7	The behaviour of the neighbourhood of BMU	84
4.8	Schematic representation of the proposed model	90
4.9	Flow chart of the theoretical framework	107
4.10	Framework of probability density function preserving approach with SOM imputation (PDSOM)	108
5.1	Results of homogeneity tests when annual median rainfall as testing variable	114
5.2	Pattern of MCAR and MNAR when true parameter of *X* is 0, 0.25, 0.5 and 1.0	117

5.3 The behavior of MNAR with degree 0.25, 0.50 and 1.00 with 10% missing values in the simulated data 128

5.4 The behavior of MCAR and MAR of *Y* with degree 0.50 and 1.00 with 10% missing values in the simulated data 129

5.5 The behavior of MNAR with degree 0.15, 0.50 and 1.00 with 20% missing values in the simulated data 132

5.6 The behavior of MAR with degree 0.50 and 1.00 and MCAR with 20% missing values in the simulated data 133

5.7 The behavior of MCAR, MNAR with degree 0.50 and 1.00, MAR of Y with degree 0.50 and 1.00 when 50% missing values in the simulated data 134

5.8 Types of missingness mechanism in Damansara and Kelantan 141

6.1 The best-fit distribution of stations in Damansara from 1996 to 2004 152

6.2 The best-fit distribution of stations in Kelantan from 1996 to 2004 154

6.3 Comparison between Generalized Extreme Value (GEV) and Wakeby (WKY) distributions for station Brook, K1 in Kelantan 156

6.4 Colour maps of the component planes from SOM for the rainfall data in Damansara and Kelantan 162

6.5 Comparison number of wet and dry events between data constructed by PDSOM and actual rainfall data for stations (denoted as Sta. No.) in Damansara 168

6.6 Comparison number of wet and dry events between data constructed by PDSOM and actual rainfall data for stations (denoted as Sta. No.) in Kelantan 168

6.7 Comparisons of mean between the ANN models and observed mean for stations (denoted as Sta. No.) in Damansara 171

6.8 Comparisons of variance between the ANN models and observed variance for stations (denoted as Sta. No.) in Damansara 171

6.9 Comparisons of mean between the ANN models and observed mean for stations (denoted as Sta. No.) in Kelantan 172

6.10 Comparisons of variance between the ANN models and observed variance for stations (denoted as Sta. No.) in Kelantan 172

6.11 Comparison of best-fit distribution to actual distribution for daily rainfall amount on wet days in Damansara when missing rainfall data was imputed by SOM and PDSOM (WBL: Weibull distribution; WKY: Wakeby distribution) 176

6.12 Comparison of best-fit distribution to actual distribution for daily rainfall amount on wet days in Kelantan when missing rainfall data was imputed by SOM and PDSOM (WBL: Weibull distribution; WKY: Wakeby distribution) 178

6.13 Colour maps of the component planes from SOM for the rainfall data from homogeneous station in Damansara and Kelantan 181

LIST OF ABBREVIATIONS

MCAR	-	Missing completely at random
MAR	-	Missing at random
MNAR	-	Missing not at random
SNHT	-	Standard Normal Homogeneity Test
BR	-	Buishand Range test
VNR	-	Von Neumann Ratio test
EM	-	Expectation Maximization
MLP	-	Multilayer Perceptron
SOM	-	Self-organizing map
PDSOM	-	Probability density function preserving approach with self-organizing map
ANN	-	Artificial Neural Network
MAE	-	Mean absolute error
RMSE	-	Root mean square error
MLE	-	Maximum likelihood estimation
SD	-	Standard deviation
BMU	-	Best-matching unit
WKY	-	Wakeby distribution
WBL	-	Weibull distribution
GEV	-	Generalized Extreme Event distribution
KS	-	Kolmogorov-Smirnaz goodness-of-fit test
AD	-	Anderson-Darling normality test

LIST OF SYMBOLS

θ	-	Vector of the parameter
N	-	Number of sample size
$\log L(\theta \mid X_{obs}, X_{mis})$	-	Log-likelihood function given data with complete and missing
r	-	Binary indicator for observed and missing data
V	-	Observed subsamples
$\overline{V}$	-	Missing subsample
$f(x \mid y, z)$	-	Conditional distribution of *X* given *Y* and *Z*
w_k, w_j	-	Weights used in Gauss Hermite quadrant
x_k, x_j	-	Abscissa used in Gauss Hermite quadrant
q	-	Number of Gauss-Hermite quadrature points
$J(\theta)$	-	Expected information used in Fishing Scoring
$S(\theta \mid x, y, z, r)$	-	Score function used in Fisher Scoring
$Cov^*(\hat{\theta}^*)$	-	Covariance matrix for the estimations of θ
ρ_{xy}, ρ_{xz}	-	Correlation between *X* and *Y*, and correlation between *X* and *Z*
$\sigma_x, \sigma_y, \sigma_z$	-	Standard deviation of *X*, *Y* and *Z*
a_1, a_2, a_3, a_4	-	Variables used to control desired degree of missingness mechanism
$T(x)$	-	Test statistic for year *x* used in standard normal homogeneity test
R_{bui}	-	Rescaled adjusted range used in Buishand Range test
R_i	-	Rank of data used in Pettitt test
V	-	Test statistic used in Von Neumann Ratio test

λ	-	Parameter used in Box-Cox transformation
A^2	-	Test statistic of Anderson-Darling test
μ_m	-	Momentum constant with the interval between 0 and 1
w	-	Weight vector
N_{map}	-	Number of map unit in SOM
α	-	Learning rate
σ	-	Radius of neighbourhood function
$h(t)$	-	Gaussian neighbourhood function
r_c, r_i	-	Location vectors of node *c* and node *i*
$V(t)$	-	Input vector
$d_i(x, y)$	-	Distance difference between target station (*x*, *y*) to the *i*-th surrounding station
n_{wet}	-	Number of wet day in PDSOM
$\xi, \alpha, \beta, \gamma, \delta$	-	Five parameters in Wakeby distribution
$k, u, \alpha > 0$	-	Shape, location and scale parameters in Generalized Extreme Values distribution
a, b	-	Scale and shape parameters in Weibull distribution
λ_r	-	Linear function of the expected order statistics for *r*-th moment
$P_r^*(F)$	-	Shifted Legendre polynomial used in L-moments
D	-	Test statistic of Kolmogorov-Smirnov test
μ	-	Sample mean
s	-	Standard deviation of the sample

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Confidence interval for the estimated parameters in EM-Logit Model	208
B	Box-plot for mean and variance of the stations in Damansara from 1996-2004	213
C	Box-plot for mean and variance of the stations in Kelantan from 1996-2004	223
D	Sample of Matlab program	233
E	Publications / proceedings	254

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Climate change is the most pressing term that has been defined as a threat to nature and human activities in the last few decades. Observational evidences such as high surface temperature, high greenhouse concentrations, high sea level and widespread melting of ice and snow has given warning signs to the world. Indeed, such multi-fold effects that have affected the statistical distribution of weather over a long period is undeniably a long-lasting crisis that has caused disastrous impacts to every aspect of human life, the ecosystem and even a country's development. Malaysia is also a victim of climate change and efforts are taken to prevent the situation from escalating.

Situated in the Southeast Asia and in between of the Pacific Ocean and Indian Ocean, climate over areas or states in Malaysia is affected by the monsoon season. The three major monsoons namely southwest monsoon, northeast monsoon and inter-monsoon have contributed the most in determining the rainfall patterns in Malaysia and subsequently has induced high variability in climatic data. Among the areas in Malaysia, Damansara and Kelantan are two distinct areas that experience unique rainfall regime. For example, rainfall in Kelantan is dominated by the northeast monsoon that occurs during November to February while rainfall in Damansara is dominated by the inter-monsoon that occurs during March to April and from

September to October. During the northeast monsoon, Kelantan receives heavy rainfall associated with thunderstorms and strong wind while Damansara receives heavy rainfall in the forms of consecutive rains during intermonsoon period. Indeed, both phenomena are sufficient for monitoring the rainfall patterns and climate over Malaysia.

Understanding the synoptic circulations over both regions is very valuable because extreme rainfall events with high rainfall amount and strong wind have boosted the frequency of flood in the last few years. This devastating flood may report a scenario where there is no complete rainfall data to forecast the rainfall events before a natural disaster happens. Apart from this, a high amount of missing rainfall data that tie up with extreme events has also reduced the reliability of the data. Therefore, a well-documented rainfall data is essential to produce intensive information regarding on the changes of rainfall behaviour and patterns in Malaysia.

In most of the hydrological studies, incomplete data is an issue that is relevant to the topic of data processing and analysis. Incomplete data, also known as missing data, can affect the quality of the data. Reference from the data will also decrease due to the lost of informative data. Often, missing data is not only caused by technical difficulties such as errors in obtaining the rainfall data, insufficient samples for analysis, failure of instruments such as in rain gauges and carelessness in data entries, but also caused by environmental variables such as temperature and humidity. To reduce the loss of important information, explanation of data being missing should be explored.

As mentioned earlier, there are several reasons to explain why hydrological data go missing. With the knowledge on why the problems occur and what variables affect the missing data, the performance of an imputation model will be increased. Therefore, exploring the missing data pattern known as missingness mechanism is an important procedure before conducting an imputation method. If the missingness mechanism is missing completely at random (MCAR), traditional methods such as mean imputation and hot deck imputation are sufficient to solve the missing data problem. If the missingness mechanism is missing at random (MAR) or missing not at random (MNAR), expectation maximization and artificial neural network are chosen. However, most of the researchers may do the imputation process without knowing the types of missingness mechanism in a dataset (Junninen *et al*., 2004). The performance of an imputation model may decrease if the missingness mechanism is not expressed correctly. To best handle the missing data problem in hydrology, types of missingness mechanism is determined before an appropriate model is selected.

Ways of imputation can be divided into three, which are i) ignoring the missing values, ii) using estimation methods, and iii) imputing the values by using mean imputation or simulated randomness. The increase of missing rainfall data has led researchers to delve into the topic of imputation models as the missing rainfall data will become an obstacle in their studies. As a result, estimation models have undergone an extensive development. The methods used include normal ratio method, multiple imputation, nearest neighbour weighting method, inverse distance weighting method, expectation maximization (EM), and artificial neural network (ANN).

1.2 Statement of Problem

A complete set of rainfall data is essential in the hydrological studies to help in country development such as designing bridges, or forecasting and predicting the occurrence of floods. However, the completeness of data is not easily achieved as external factors such as climate change may contribute to the occurrence of missing data. In Malaysia, for example, the factor that contributes to missing data problem may be due to monsoon season or climate change. Besides, the cause of missing data may also be related to technical errors such as error in data entry or rain gauge malfunction. To better understand the factors that affect the missing data, missingness mechanism is introduced. The knowledge of missingness mechanism that leads to missing data should be pre-determined before the imputation process is carried out.

The task of estimating missing rainfall records ranges from traditional methods, model based imputation, spatial interpolation methods and data driven approach. Traditional methods such as listwise and pairwise deletion are the pioneers in the imputation methods. However, both methods have shown marked errors because deletion of missing data will reduce the sample size and then increase the variance of the data. Spatial interpolation methods also worked well under the assumption that the relation between target station and neighbouring stations is significant, and their performances will be disrupted if there is variability in time and space. More recently, data-driven approach that uses the evolutionary principles and biological network, namely artificial neural network (ANN), is suggested because it is superior and powerful in predicting missing values with minimum errors (Juininen *et al*., 2004; Srikalra and Tanprasert, 2006; Bustami *et al*., 2007; Kalteh and Hjorth, 2009; Piazza *et al*., 2011).

Self-organizing map (SOM) is one of the branches from ANN that do not require the desired output for the input vectors. Optimum results will be archieved if the architecture of SOM, such as map size and learning rate are well defined. More often, SOM is acknowledged as a high-performing computational model that is able to model a highly complicated system. However, SOM is not a statistical-sound model as it does not preserve the statistical properties of a data. In a hydrological study, statistical properties of rainfall data such as the distribution of rainfall process, number of wet and dry days, mean and variance of the imputed data are important and should be preserved. When SOM is applied, missing data will be highly estimated but sometimes, the statistical properties of the data may alter. For that reason, a model that can retain the statistical properties of the data is underlined in this study.

The problem of missing data is not a new topic for researchers in Malaysia, but studies focusing on the missing data analysis in tropical regions are few and far between. As mentioned earlier, river basins from Damansara and Kelantan were chosen because the rainfall patterns in both areas are different from each other. The number of missing rainfall data is also high especially during the monsoon seasons. It is also recorded that the missing observations sometimes happen for a few days consecutively. In order to recover the completeness and quality of rainfall data, the mechanism of the missing rainfall data is first identified before the process of imputation.

1.3 Objectives

The objectives of the study are:

i) To propose a generalized joint model that can identify the types of missingness mechanism in a data.
ii) To determine the missingness mechanism of the rainfall data in Malaysia.
iii) To determine the best-fit distribution to represent the rainfall patterns in Malaysia.
iv) To propose a new imputation method by hybridizing probability distributed model and self-organizing map (SOM).
v) To compare the performance of the proposed artificial neural network (ANN) model with the existing ANN methods.

1.4 Scope of the Study

This study will divide the problems of missing data into two parts, which are determination of missingness mechanism and prediction of missing data in the rainfall data of Malaysia. Daily rainfall data over a 9-year period (1996 – 2004) in Damansara and Kelantan river basins will be studied.

To accelerate the accuracy of imputation process, the mechanism of missing data needs to be recognized by introducing a model, namely as Expectation-Maximization (EM) with logit. It is a joint model where the parameter estimates obtained from the EM will determine the types of missingness mechanism existing in the data, either by missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). The performance of the model is then tested on a simulated data that reflect the real rainfall condition in Malaysia, and imposed with different missingness mechanism, different correlation among the variables and the percentage of missing data.

Daily rainfall data coupled with humidity and temperature from the stations in Damansara and Kelantan for the period of 1998 to 2004 are also included in the model assessment. Before the application of the joint EM-Logit model, homogeneity tests will be carried out to ensure the measurements of the data are taken at a time using same instruments and environments. Then, the type of missingness mechanism that lead to rainfall data being missing is determined.

The second part of the study is the comparison between imputation models. A new two-step ANN model that first analyses the occurrence of rainfall events before the imputation by self-organizing map (SOM) for the wet-classified day is constructed. The new model will be compared with two classical ANN models, namely multilayer perceptron (MLP) and self-organizing map (SOM). In order to ensure the proposed model is able to cope with the problem addressed, daily rainfall data is divided into training data, calibration data and validation data. A 4-year daily rainfall data that was extracted during the monsoon season for each river basin is used to train the SOM while another 5-year complete daily rainfall data is used to evaluate all the ANN models. The best model will then be selected and applied to the rainfall data in Damansara and Kelantan from 1996 to 2004. All the algorithms are written in MATLAB 7.

1.5 Significance of the Study

This study investigates the mechanism of missing rainfall data and looks for possible ways to improve the estimation of missing rainfall data in Malaysia. Because of the high variability of climate in the last few decades, estimation by traditional methods such as inverse distance weighted method (IDWM) becomes less convincing. Since the rainfall data obtained from Jabatan Pengaliran dan Saliran (JPS) Malaysia is used for future hydrological prediction and for better understanding of the changes of hydrological processes, the missing data problems are highlighted. By having a high quality and with correct statistical properties of hydrological data, engineers and economists can make investment decisions wisely in future infrastructure and water management systems in Malaysia.

1.6 Organization of Thesis

This thesis comprises of seven chapters that can be divided into two parts, which are missingness mechanism and imputation of missing data. For Chapter 2, a literature review of missingness mechanism and imputation models is listed. Chapter 3 presents the description for the proposed missingness mechanism model, including mathematical formulation, simulation and implementation procedures. Meanwhile, Chapter 4 outlines the details for proposed imputation model, along with existing artificial neural networks (ANN) models. To assess the performance of the proposed models, Chapter 5 and Chapter 6 provide a complete evaluation and discussion. Last but not least, all the summary, conclusions and recommendations for future research are presented in Chapter 7.

REFERENCES

Abdella, M. and Marwala, T. (2005). The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database. *IEEE Xplore*. 207-212.

Afolabi, M. O. and Olude, O. (2007). Predicting Stock Prices Using a Hybrid Kohonen Self-Organizing Maps (SOM). *Proceedings of the 40th Hawaii International Conference on System Science*.

Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton Algorithm for Maximum Likelihood in Mixture Distributions. *Statistics and Computing*. 6(2), 127-130.

Alexandersson, H. A. (1986). Homogeneity Test Applied to Precipitation Test. *J.Climatol.* 6**,** 661-675.

Aksoy, H. (2000). Use of Gamma Distribution in Hydrological Analysis. *Turk. J. Engin. Environ. Sci.* 24(6), 419-428.

Blankers, M., Koeter, M. W. J., Schippers, G. M. (2010). Missing data approaches in eHealth research: simulation study and a tutorial for nonmathematically inclined researchers. *J. Med. Internet Res*. 12(5): e54.

Bridges, T. C. and Haan, C. T. (1972). Reliability of Precipitation Probabilities Estimated from the Gamma Distribution. *Monthly Weather Review*. 100(8), 607-611.

Boudour, M. and Hellal, A. (2005). Combined Use of Supervised and Unsupervised Learning for Power System Dynamic Security Mapping. *Engineering Application of Artificial Intelligence.* 18, 673-683.

Buishand, T.A. (1982). Some Methods for Testing the Homogeneity of Rainfall Records. *J. Hydrol.* 58, 11-27.

Chen, H., Grant-Muller, S., Mussone, L., Montgomery, F. (2001). A Study of Hybrid Neural Network Approaches and the Effects of Missing Data on Traffic Forecasting. *Neural Comput. Applic.* 10, 277-286.

Costa, A. and Soares, A. (2006). Identification of Inhomogeneities in Precipitation Time Series Using SUR Models and the Ellipse Test. *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 5–7 July. Lisboa, Instituto Geográfico Português, 419-428.

Cottrell, M. and Letremy, P. (2007). Missing values: processing with the Kohonen algorithm. *CoRR*. arXiv preprint math/0701152.

Coulibaly, P. and Evora, N. D. (2007). Comparison of Neural Network Methods for Infilling Missing Daily Weather Records. *Journal of Hydrology*. 341, 27-41.

Curran, D. Molenberghs, G., Fayers, P. M., Machin, D. (1998). Incomplete quality of life data in randomized trials: missing forms. *Statistics in Medicine*. 17(5-7), 697-709.

Dao, V. N. P. and Vemuri, V. R. (2002). A Performance Comparison of Different Back Propagation Neural Networks Methods in Computer Network Intrusion Detection. *Differential Equations and Dynamical Systems*. 10(1&2), 201-214.

Dastorani, M. T., Moghadamnia, A., Piri, J., Rico-Ramirez, M. (2009). Application of ANN and ANFIS Models for Reconstructing Missing Flow Data. *Environ Monit Assess*. Doi: 10.1007/s10661-009-1012-8.

Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. C., Lott, N. J. (2000). Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *Journal of Applied Meteorology*. 39, 1580-1591.

Eltahir, E. A. and Pal, J. S. (1996). Relationship between surface conditions and subsequent rainfall in convective storms. *Journal of Geophysical Research: Atmospheres (1984-2012)*. 101,(D2), 26237-26245.

Fadhilah Yusof, Zalina Mohd Daud, Nguyen, V-T-V, Suhaila S., Zulkifli Yusof (2007). Fitting the Best-Fit Distribution for the Hourly Rainfall Amount in the Wilayah Persekutuan. *Jurnal Teknologi*. 46(C), 49-58.

Fairclough, D.L. (2002). *Design and Analysis of Quality of Life Studies in Clinical Trials*. New York: Chapman and Hall.

Forti, A. and Foresti, G. L. (2006). Growing Hierarchical Tree SOM: An Unsupervised Neural Network with Dynamic Topology. *Neural Networks*.19, 1568-1580.

Gonzalez-Rouco, J.F., Jimenez, J. L., Quesada, V., Valero, F. (2001). Quality Control and Homogeneity of Precipitation Data in the Southwest of Europe. *Int. J. Climate*. 14, 964-978.

Gupta, J. N. D. and Sexton, R. S. (1999). Comparing Backpropagation with a Genetic Algorithm for Neural Network Training. *Omega*. 27, 679-684.

Hagenbuchner, M. and Tsoi, A. C. (2005). A Supervised Training Algorithm for Self-Organizing Maps for Structures. *Pattern Recognition Letters*. 26, 1874-1884.

Hanson, L. S. and Vogel, R. (2008). The Probability Distribution of Daily Rainfall in the United States. *World Water & Environmental Resources Congress*. 12-16 May. Honolulu, Hawaii, United States, 1-12.

Hosking, J. R. M. (1989). Research Report: Some Theoretical Results Concerning L-Moments. IBM Research, RC. 14492(64907).

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distribution using Linear Combinations of Order Statistics. *J. R. Statist. Soc. B.* 52(1), 105-124.

Ismail Mohamad (2003). *Data Analysis in the Presence of Missing Data*. PhD Thesis, Lancaster University.

Jaeger, M. (2006). On testing the missing at random assumption. *Machine Learning: ECML 2006*. Springer Berlin Heidelberg 2006. 671-678.

Jamaludin Suhaila, Sayang Mohd Deni, Abdul Aziz Jemain (2008). Detecting Inhomogeneity of Rainfall Series in Peninsular Malaysia. *Asia-Pacific Journals of Atmospheric Sciences*. 44(4), 369-380.

Janssen, M., Donders, A. R. T., Harrell, F. E., Vergouwe, Y. (2009). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*. 63, 721-727.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M. (2004). Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmospheric Environment*. 38, 2895-2907.

Khaliq M.N. and Ouarda, T. B. M. J. (2007). On the Critical Values of the Standard Normal Homogeneity Test (SNHT). *Int. J. Climatol*. 27, 681-687.

Karlaftis, M. G. and Vlahogianni, E. I. (2010). Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. *Transportation Research Part C*. 19, 387-399.

Katteh, A. M. and Hjorth, P. (2009). Imputation of Missing Values in a Precipitation-Runoff Process Database. *Hydrology Research*. 40(4), 420 – 432.

Kim, J. W. and Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression tree and artificial neural network for SWAT streamflow simulation. *Journal of Hydrology*. 294, 305-314.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* 43, 59-69.

Kohonen, T., Kaski, S., Lappalainen, H. (1997). Self-Organized Formation of Various Invariant-Feature Filters in the Adaptive-Subspace SOM. *Neural Computation*. 9(6), 1321-1344.

Koikkalainen, P., Horppu, I. (2007). Handling Missing Data with the Tree-Structured Self-Organizing Map. *Proceedings of International Joint Conference on Neural Network*. 12-17 August. Orlando, Florida USA.

Lamrini, B., Lakhal, El-K., Le Lann, M-V., Wehenkel, L. (2011). Data validation and missing data reconstruction using self-organizing map for water treatment. *Neural Comput & Applic*. 20(4), 575-588.

Landwehr, J. M., Matalas, N. C., Wallis, J. R. (1980). Quantile estimation with more or less flood-like distributions. *Water Resources Research*. 16(3), 547-555.

Lei, Y. (2008). Evaluation of three methods for estimating the Weibull distribution parameters of Chinese pine (Pinus tabulaeformis). *Journal of Forest Science*. 54(12), 566-571.

Listing, J. and Schlittgen, R. (1998). Test if dropouts are missed at random. *Biometric*. 40, 929-35.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Unites States in America: John Wiley & Sons Inc.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. 83(404), 1198-1202.

Mahmut Firat, Fatih Dikbas, Koc, A. C., Mahmud Gungor (2010). Missing data analysis and homogeneity test for Turkish precipitation series. *Sadhana*. 35 part 6, 707-720.

Marlinda Abdul Malek, Sobri Harun, Siti Mariyam Shamsuddin, Ismail Mohamad (2008). Imputation of Time Series Data via Kohonen Self Organizing Maps in the Presence of Missing Data. *World Academy of Science, Engineering and Technology*. 41, 501-506.

Marlinda Abdul Malek (2008). *Development of Rainfall Data Infilling Model with Expectation Maximization and Artificial Neural Network*. PhD Thesis. Universiti Teknologi Malaysia, Skudai.

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques*. New York: Information Science Reference.

Merlin, P., Sorjamaa, A., Maillet, B., Lendasse, A. X-SOM and L-SOM: A double classification approach for missing value imputation. *Neurocomputing*. 73(7-9), 1103-1108.

Minai, A. A. and Williams, R. D. (1994). Perturbation Response in Feedforward Networks. *Neural Network*. 7(5), 783-796.

Modarres, R. (2006). Regional Precipitation Climates of Iran. *Journal of Hydrology (NZ)*. 45(1), 13-27.

Mooley, D. A. (1973). Gamma Distribution Probability Model for Asian Summer Monsoon Monthly Rainfall. *Monthly Weather Review*. 101(2), 160-176.

Nakai, M. (2011). Analysis of Imputation Methods for Missing Data in AR(1) Longitudinal Dataset. *Int. Journal of Math. Analysis*. 5(45), 2217-2227.

Nelwamondo. F. V., Mohamed, S., Marwala, T. (2007). Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*. 93(11), 1514-1520.

Park, J. S., Qian, G. Q., Jun, Y. (2008). Monte Carlo EM algorithm in logistic linear models involving non-ignorable missing data. *Applied Mathematics and Computation 197*, 440-450.

Park, Taesong and Davis, C.S. (1993). A Test of the Missing Data Mechanism for Repeated Categorical Data. *Biometrics*. 49, 631- 638.

Pettitt, A. N. (1979). A Non-Parametric Approach to the Change-Point Detection. *Appl. Stat*. 28, 126-135.

Peyre, H., Coste, J., Leplege, A. (2010). Identifying type and determinants of missing items in quality of life questionnaires: Application to the SF-36 French version of the 2003 Decennial Health Survey. *Health and Quality of Life Outcomes*. 8(16), 1-6.

Piela, P. (2003). Exploitation of Neural Methods for Imputation. *Fin-00022 Statistics Finland, Finland*. 49-52.

Pommeret, D. (2012). *Testing the mechanism of missing data*. Preprint, Institut de Mathématiques de Luminy (IML).

Qu, A. and Song, P. X. K. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*. 89(4), 841-850.

Rao, A. R., Hamed, K. H. (2000). *Flood Frequency Analysis*. United States: CRC Press.

Ridout, M.S. (1991). Testing for Random Dropouts in Repeated Measurement Data, *Biometrics*. 47, 1617-162.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*. 63(3), 581-592.

Rubin, L. H., Witkiewitz, K., St. Andre, J., Reilly, S. (2007). Methods for handling missing data in the behavioural neurosciences: Don't throw the baby rat out with the bath water. *The Journal of Undergraduate Neuroscience Education*. 5(2), A71-A77.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician*. 41, 169-178.

Sayang Mohd Deni, Abdul Aziz Jemain, Ibrahim, K. (2008). The Spatial Distribution of Wet and Dry Spells over Peninsular Malaysia. *Theor. Appl. Climatol*. 94, 163-173.

Simolo, C., Brunetti, M., Maugeri, M., Nanni, T. (2009). Improving Estimation of Missing values in Daily Precipitation Series by a Probability Density Function-Preserving Approah. *Int. J. Climatol*. 30(10), 1564-1576.

Sinharay, S., Stern, H. S., Russell, D (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods*. 6(4), 317 – 329.

Sommer, D., Grimm, T., Golz, M. (2003). Processing missing values with self-organizing map. *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (EUNITE 2003)*. July 2003, 275-279.

Sorjamaa, A., Corona, F., Miche, Y., Merlin, P., Maillet, B., Severin, E., Lendasse, A. (2009). Sparse linear combination of SOMs for data imputation: Application to financial database. *Advances in Self-Organizing Maps*. 5629, 290-297.

Sharma, M. A. and Bhagwan Singh, J. (2010). Use of Probability Distribution in Rainfall Analysis. *New York Science Journal*. 3(9), 40-49.

Stepanek P., Zahradnicek P., Skalak P. (2009). Data Quality Control and Homogenization of Air Temperature and Precipitation Series in the Area of the

Czech Republic in the period of 1961-2007. *Advances in Science and Research*. 3, 23-26.

Suhaila Jamaludin and Abdul Aziz Jemain (2007). Fitting the Statistical Distributions to the Daily Rainfall Amount in Peninsular Malaysia. *Jurnal Teknologi*. 46(C), 33-48.

Tang, W. Y., Kassim, A. H. M., Abubakar, S. H. (1996). Comparative Studies of Various Missing Data Treatment Methods - Malaysian Experience. *Atmospheric Research*. 42, 247-262.

Teegavarapu, R. S. V. and Chandramouli, V. (2005). Improved Weighting Methods, Deterministic and Stochastic Data-Driven Models for Estimation of Missing Precipitation Records. *Journal of Hydrology*. 312(1), 191-206.

Train, K. E. (2008). EM Algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*. 1(1), 40-69.

Udomboso, C. G. and Amahia, G. N. (2011). Comparative analysis of rainfall prediction using statistical neural network and classical linear regression model. *Journal of Modern Mathematics and Statistics*. 5(3), 66-70.

Valsson, S. and Bharat, A. (2011). *Impact of Air Temperature on Relative Humidity – A Study*. Architecture – Time Space & People (February 2011). 38-41.

Von Neumann, J. (1941). Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Ann. Math. Stat*. 13, 367-395.

Wang, X. L. L., Wen, Q. Z. H., Wu, Y. H. (2007). Penalized Maximal *t* Test for Detecting Undocumented Mean Change in Climate Data Series. *Journal of Applied Meteorology and Climatology*. 46, 916-931.

Weerasinghe, K. D. N. (1989). The Rainfall Probability Analysis of Mapalana and its Application to Agricultural Production of the Area. *J. Natn. Sci. Coun. Sri Lanka*. 17(2), 173-186.

Wijngaard, J. B., Kleink Tank, A. M. G, Konnen, G. P. (2003). Homogeneity of 20th Century European Daily Temperature and Precipitation Series. *Int. J. Climatol*. 23, 679-692.

Wilheit. T. T. and Chang, A. T. C. (1990). Retrieval of Monthly Rainfall Indices from Microwave Radiometric Measurements Using Probability Distribution Functions. *Journal of Atmosphericand Oceanic Technology*. 8, 118-136.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. 30, 79-82.

Wilson, D. R. and Martinez, T. R. (2003). The General Inefficiency of Batch Training for Gradient Descent Learning. *Neural Networks*. 16, 1429-1451.

Wong, T. S. T. and Li, W. K. (2006). A note on the estimation of extreme value distributions using maximum product of spacings. *IMS Lecture Notes-Monograph Series Time Series and Related Topics*. 52, 272-283.

Xia, Y. Fabian, P., Stohl, A., Winterhalter, M. (1999). Forest Climatology: Estimation of Missing Values for Bavaria, Germany. *Agricultural and Forest Meteorology*. 96, 131-144.

Young, K. C. (1992). A three-way model for interpolating for monthly precipitation values. *Monthly Weather Review*. 120, 2561-2569.

Zalina Mohd Daud, Amir Hashim Mohd Kassim, Mohd Nor Mohd Desa, Vand-Thanh-Van Nguyen (2002). Statistical Analysis of at-site Extreme Rainfall Processes in Peninsular Malaysia. *Regional Hydrology: Bridging the Gap Between Research and Practice (Proceeding of the Fourth International FRIEND Conference)*. March 2002. Cape Town, South Africa: Publication no. 274, 61-68.