



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

# SPATIOTEMPORAL RAINFALL PATTERNS RECOGNITION USING ROBUST PRINCIPAL COMPONENT ANALYSIS AND FUZZY C-MEANS



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

SITI MARIANA CHE MAT NOR

UNIVERSITI PENDIDIKAN SULTAN IDRIS

2021



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

SPATIOTEMPORAL RAINFALL PATTERNS RECOGNITION USING ROBUST  
PRINCIPAL COMPONENT ANALYSIS AND FUZZY C-MEANS

SITI MARIANA CHE MAT NOR

DISSERTATION PRESENTED TO QUALIFY FOR A MASTERS IN SCIENCE  
(RESEACRH MODE)

FACULTY OF SCIENCE AND MATHEMATICS  
SULTAN IDRIS EDUCATION UNIVERSITY

2021



Please tick (✓)

Project Paper

Masters by Research

Master by Mixed Mode

PhD

✓


## INSTITUTE OF GRADUATE STUDIES

### DECLARATION OF ORIGINAL WORK

This declaration is made on the .....6.....day of.....7.....20..21..

#### i. Student's Declaration:

I, Siti Mariana Che Mat Nor, M20181001458, Faculty of Science and Mathematics (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled Spatiotemporal Rainfall Patterns Recognition using Robust Principal Component Analysis and Fuzzy C-Means is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.


  
Signature of the student

#### ii. Supervisor's Declaration:

I Dr. Shazlyn Milleana Shaharudin (SUPERVISOR'S NAME) hereby certifies that the work entitled Spatiotemporal Rainfall Patterns Recognition using Robust Principal Component Analysis and Fuzzy C-Means (TITLE) was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a \* ~~partial~~ /full fulfillment for the conferment of Master in Science (Statistics) (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

6 Julai 2021

Date

  
Signature of the Supervisor



**INSTITUT PENGAJIAN SISWAZAH /  
INSTITUTE OF GRADUATE STUDIES**

**BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK  
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM**

Tajuk / Title: Spatiotemporal Rainfall Patterns Recognition using Robust Principal  
Component Analysis and Fuzzy C-Means

No. Matrik / *Matric's No.*: M20181001458

Saya / I: Siti Mariana Che Mat Nor

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)\* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-

*acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-*

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.  
*The thesis is the property of Universiti Pendidikan Sultan Idris*
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.  
*Tuanku Bainun Library has the right to make copies for the purpose of reference and research.*
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.  
*The Library has the right to make copies of the thesis for academic exchange.*
4. Sila tandakan ( ☒ ) bagi pilihan kategori di bawah / *Please tick ( ☒ ) for category below:-*

☐

**SULIT/CONFIDENTIAL**

Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*

☒

**TERHAD/RESTRICTED**

Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restricted information as specified by the organization where research was done.*

☐

**TIDAK TERHAD / OPEN ACCESS**

(Tandatangan Pelajar/ *Signature*)

DR SHAZLYN MILLEANA BINTI SHAHARUDDIN  
Pensyarah Kanan  
Jabatan Matematik  
Fakulti Sains dan Matematik  
Universiti Pendidikan Sultan Idris

(Tandatangan Penyerah / *Signature of Supervisor*)  
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Tarikh: 9 Julai 2021

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

*Notes: If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.*

## ACKNOWLEDGEMENT

First of all, I am beyond grateful to Allah S.W.T for the strength that I've been blessed with throughout this journey full of challenges and filled with tears.

My deepest appreciation and millions of thanks to my supervisor, Dr. Shazlyn Milleana Shaharudin for all the knowledge shared, time spent and energy contributed to help me complete this study. All her support and sacrifices are surely unrequited.

I would also like to thank my mother, Rohani Senik for always faithfully listening to the ups and downs in finishing this journey. Not forgetting also for fathers, siblings and friends who are endlessly giving me words of encouragement and praying for this success. I believe I could not reach this point without the prayer and support from all of them.

I would also like to thank the Fundamental Research Grants Scheme (FRGS/1/2019/STG06/UPSI/02/4) provided by the Ministry of Education Malaysia for supporting my study. Many thanks to the Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris for the chance and facilities provided and also a special thanks to the Department of Irrigation and Drainage Malaysia (DID) for the data and informations provided for this study.



## ABSTRAK

Objektif kajian ini ialah untuk mengenalpasti corak ruang-masa hujan lebat menggunakan Analisis Komponen Utama Teguh dan C-purata Kabur (RPCA-FCM) bagi hujan lebat di Pantai Timur Semenanjung Malaysia. Sebagai satu metodologi, model RPCA-FCM telah dicadangkan bagi menyelesaikan beberapa isu dalam mengenalpasti hujan lebat. Umumnya, kebanyakan data hujan mengalami kehilangan disebabkan oleh pelbagai punca. Mekanisme data yang hilang dikenalpasti bagi memilih kaedah imputasi yang sesuai. RF-MLR telah dipilih sebagai kaedah imputasi yang terbaik bagi mengendalikan data hujan yang hilang. Kaedah pengurangan dimensi serta pendekatan pengelompokan digunakan bagi mengurangkan dimensi data dan menjalankan pembahagian kluster. RPCA berpusatkan fungsi *Tukey's biweight* dan titik pemecahan optimum untuk mengekstrak nombor komponen dalam RPCA juga diperkenalkan. Data kajian ini dihasilkan menggunakan simulasi *Monte Carlo* untuk menilai prestasi model statistik yang dicadangkan. Hasil kajian mendapati titik pemecahan 0.4 pada 85% peratus kumulatif varians mengekstrak nombor komponen dengan baik bagi mengelakkan variasi frekuensi rendah atau skala ruangan bagi kelompok tidak signifikan. Kajian ini juga menunjukkan bahawa terdapat peningkatan di mana RPCA berjaya mendekatkan data terpencil dari pusat dan memperbaiki pembahagian kluster. Namun, K-purata membenarkan setiap elemen untuk dimiliki oleh satu kelompok secara eksklusif. Satu penyelesaian dapat dicapai di mana pengelompokan FCM digabungkan untuk membolehkan elemen data tergolong dalam lebih dari satu kelompok berdasarkan struktur data hujan. Kesimpulannya, hasilnya menunjukkan peningkatan ketara dengan RPCA-FCM berbanding PCA-FCM dari segi purata jumlah kelompok yang diperoleh dan kualiti kelompok. Sebagai implikasinya, pengenalanpastian kelompok corak hujan ruang-masa berguna bagi ahli hidrologi dalam menganalisis model persekitaran dan meningkatkan penilaian terhadap perubahan iklim.



## SPATIOTEMPORAL RAINFALL PATTERNS RECOGNITION USING ROBUST PRINCIPAL COMPONENT ANALYSIS AND FUZZY C-MEANS

### ABSTRACT

The main objective of this study is to identify the spatiotemporal rainfall patterns using Robust Principal Component Analysis and Fuzzy C-means (RPCA-FCM) of torrential rainfall of the East Coast of Peninsular Malaysia. As a methodology, the RPCA-FCM model was proposed to solve issues in identifying torrential rainfall. Generally, most rainfall data were missing for various reasons. The missing data mechanism was identified to choose suitable imputation methods. RF-MLR was chosen as the best imputation method in handling missing rainfall data. Dimension reduction method coupled with clustering approach was applied to reduce the data dimensions and perform the cluster partition. An RPCA-based Tukey's biweight correlation and the optimum breakdown point to extract the number of components in RPCA were proposed. The data used in this study was generated using Monte Carlo simulation to evaluate the performance of the proposed statistical model. The result revealed that a breakdown point of 0.4 at 85% cumulative variance percentage efficiently extracts the number of components to avoid low-frequency variations or insignificant clusters' spatial scale. This study also showed that there is an improvement where the RPCA downweighed the far-from-center outliers and developed the cluster partitions. However, K-Means allows each element to exclusively belong to a cluster. A solution was attained where FCM was combined to allow the data elements to belong to more than one cluster based on the rainfall data structure. In a conclusion, the results show a substantial improvement with the RPCA-FCM than the classical model in terms of the average number of clusters obtained and the cluster quality. As an implication, the identification of spatiotemporal cluster rainfall patterns is useful for hydrologists in analyzing environmental models and improves the assessment of climate change.

## CONTENTS

	Pages
<b>DECLARATION OF ORIGINAL WORK</b>	ii
<b>DECLARATION OF THESIS</b>	iii
<b>ACKNOWLEDGEMENT</b>	ii
<b>ABSTRACT</b>	iv
<b>ABSTRAK</b>	iii
<b>TABLE OF CONTENTS</b>	v
<b>LIST OF TABLES</b>	ix
<b>LIST OF FIGURES</b>	x
<b>LIST OF ABBREVIATIONS</b>	xii
<b>CHAPTER 1 INTRODUCTION</b>	
1.1 Background of the study	1
1.2 Study Area	6
1.3 Database for Torrential Rainfall Data	8
1.4 Problem Statement	10
1.5 Research Objectives	13
1.6 Significance of Study	14



1.7	Notations	14
1.8	Research Methodology	15
1.9	Thesis Organization	16
1.10	Limitations of Study	18

## CHAPTER 2 LITERATURE REVIEW

2.1	Introduction	19
2.2	Missing Data	19
2.3	Imputation Methods for Missing Data	25
2.4	Multiple Linear Regression	30
2.5	Performance Indicator for Imputation Methods	31
2.6	Introduction to Dimension Reduction Methods	32
2.7	Identifying Rainfall Pattern by using Dimension Reduction Method	41
2.8	K-means Clustering	44
2.9	Enhanced Principal Component Analysis	47
2.10	Fuzzy C-means Clustering	52
2.11	Clustering Validity for Statistical Modeling	55

## CHAPTER 3 METHODOLOGY

3.1	Introduction	57
3.2	Imputation Methods	60
3.2.1	Replace by Mean	61
3.2.2	Nearest Neighbor	61

3.2.3	Markov Chain Monte Carlo (MCMC)	63
3.2.4	Nonlinear Iterative Partial Least-Square (NIPALS)	64
3.2.5	Random Forest	65
3.3	Multiple Linear Regression (MLR)	67
3.4	Performance Indicator for Imputation Methods	68
3.4.1	Root Mean Square Error (RMSE)	68
3.4.2	Nash-Sutcliffe Efficiency Coefficient (CE)	69
3.4.3	Mean Absolute Error (MAE)	70
3.5	Classical Principal Component Analysis	70
3.5.1	Method to Select the Number of Principal Component	74
3.6	K-means Cluster Analysis	76
3.6.1	Calinski and Harabasz Index	78
3.7	Disadvantages of Classical PCA combined with K-means Clustering	79
3.8	Statistical Modelling using RPCA-FCM	80
3.8.1	Robust PCA (RPCA)	80
3.8.2	M-estimation	81
3.8.3	Tukey's biweight Function	83
3.8.4	Breakdown Point	87
3.8.5	Fuzzy C-Means (FCM)	88
3.8.6	Proposed Statistical Model of Robust PCA coupled with FCM (RPCA- FCM)	90

3.9	Validity Indices for Clustering	91
3.10	Validity Indices for RPCA-FCM Statistical Model	94
3.11	Evaluating Performance of RPCA-FCM Statistical Model	95
3.11.1	Data Generation	95
3.11.2	Simulation	96

## CHAPTER 4 RESULTS AND DISCUSSION

4.1	Introduction	99
4.2	Handling Missing Rainfall Data	100
4.3	Choice of Breakdown Point for RPCA	110
4.4	Evaluating Performance of Classical PCA against RPCA	113
4.5	Validity Measures of Clusters	116
4.6	Fuzzy C-means Clustering	117
4.7	Evaluating Performance of Proposed Statistical Model based on Simulated Data	118
4.8	Description of Clustering Rainfall Patterns	121

## CHAPTER 5 CONCLUSION

5.1	Introduction	126
5.2	Summary	126
5.3	Future Research	130

REFERENCES	132
------------	-----

## LIST OF TABLES

Table No.		Pages
1.1	Geographical coordinates of 48 rainfall stations in the East Coast of Peninsular Malaysia	8
1.2	Geographical coordinates of observed 30 rainfall stations in East-Coast of Peninsular Malaysia	10
2.1	Missing data rules according to (Widaman, 2006)	22
4.1	Average of RMSE and CE values for five imputation methods	106
4.2	The Results for MLR coupled with Imputation Methods	107
4.3	Number of components, clusters and validations for chosen breakdown point, r a) $r=0.2$ b) $r=0.4$ c) $r=0.6$ d) $r=0.8$	112
4.4	Number of components obtained based on classical PCA and RPCA from rainfall data of the East Coast Peninsular Malaysia	114
4.5	Indices to measure the quality of clustering results for torrential rainfall data	117
4.6	Validity measures of statistical model for original data	118
4.7	Number of components and clusters obtained for classical PCA and RPCA based on simulated data	120
4.8	Indices to measure the quality of clustering results of simulation data	120
4.9	Validity measures of statistical model for simulation data	121

## LIST OF FIGURES

Figure No.		Pages
1.1	The locations of 48 rainfall stations in East Coast Peninsular Malaysia	7
1.2	The locations of the observed 30 rainfall stations in the East-Coast of Peninsular Malaysia	9
1.3	Flow Chart of Research Methodology	16
2.1	Operational framework in handling high dimensional data based on dimension reduction	34
3.1	Procedure of classical PCA	74
3.2	Procedure of K-means algorithm	78
3.3	The derivative function, $\psi$ -function for the biweight M-estimator	84
3.4	The weight function for the biweight M-estimator	86
3.5	Procedure of RPCA with unsupervised clustering	87
3.6	Procedure of FCM	90
3.7	Procedure of RPCA-FCM	91
4.1	Missing percentage of dataset	100

4.2	Correlation visualization of missing and non-missing data for the East Coast of Peninsular Malaysia	102
4.3	Correlation of rainfall days and rainfall amount for 3 chosen stations	104
4.4	Data imputation results of 175 missing rainfall data for the MCMC, NIPALS, Nearest Neighbour and Replace by Mean models.	108
4.5	Number of clusters obtained based on classical PCA against RPCA from rainfall data of the East Coast of Peninsular Malaysia	115
4.6	Torrential rainfall composites in the East Coast of Peninsular Malaysia	123



## LIST OF ABBREVIATIONS

CE	Nash-Sutchliffe Coefficient Efficiency
DID	Department of Irrigation and Drainage
FCM	Fuzzy C-means
IDW	Inverse Distance Weighting
MAE	Mean Absolute Error
MAR	Missing at Random
MCAR	Missing Commonly at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MLR	Multiple Linear Regression
MNAR	Missing Not at Random
NIPALS	Nonlinear Iterative Partial Least Squares
NN	Nearest Neighbor
NR	Normal Ratio
PCA	Principal Component Analysis
RF	Random Forest
RMSE	Root Mean Square Error
RPCA	Robust Principal Component Analysis





05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

xiii

SI                      Single Imputation

RP                     Rainfall Pattern



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi



## CHAPTER 1

### INTRODUCTION

#### 1.1 Background of the study

Climate change is becoming more acknowledged as an insidious world crisis. It comes with an increasing intensity of many extreme events such as thunderstorms, tornadoes, severe dust storms and intense rainfall. These climatic issues receive considerable attention from various researchers and the rainfall activities analysis has become progressively significant in various fields, predominantly in water industry like water resources management, hydrology and agronomy. Rainfall is undoubtedly one of the most significant natural phenomena and it highly crucial in the natural environment. However, excessive rainfall can trigger natural hazards that threaten countless human's population. Malaysia, which is classified as having equatorial climate, experiences warm and humid weather all year round. The episodes of the torrential rain character occurrences, which are conceivably disastrous, are a principally pertinent characteristic of Malaysia's rainfall regime.



Mostly, in the Malaysian context, it has become the utmost prominent cataclysm with regards to the impacted populations, recurrence, zone coverage, flood length and socio-economical impairment. With 150 river systems in Peninsular Malaysia and an additional of 50 river systems in East Malaysia, the ecosystem and human mutually necessitate these rivers as well as the river corridors of floodplains satisfy various purposes. Fundamentally, rivers take on a huge duty for natural, social, and economical systems regardless of their settings. Since the beginning of civilization, a nation's development and cultures are heavily influenced by the rivers. According to the Department of Irrigation and Drainage (2015) data, Malaysia has been blessed with reasonably plentiful annual rainfall of 3000 mm (which is approximately 990 million,  $m^3$ ), where its surface run-off is about 57%, as it is located in the tropical regions. Seasonal distribution and variation, both temporal and spatial, however, caused certain regions to become water-stressed at times. 60% of the rain comes annually from November to January. Generally, the major east Malaysia's wet season, for instance the event of heavy rains, occurs from November to February. On the other hand, in the west coast, the rainiest time interval would be in August. East Malaysia has heavy rains from November to February. The average rainfall for Peninsular Malaysia is 2500 mm, while in East Malaysia is 5080 mm.

Since 1920, the country had experienced major floods which led to tremendous life and property losses as well as environmental degradation. On January 1971, a devastating flood had struck several Malaysian states and the city of Kuala Lumpur, causing over RM 200 million and 61 casualties. Among the most recent events the Johor flood that hit the state by the end of 2006 and early January 2007. The estimated cumulative loss of these flood events was RM 1.5 billion due to a couple of abnormally





heavy rainfall occurrences triggering the major flooding. It was then considered the most destructive flood occurrence in the history of Malaysia. The cost of disruption to the utilities, bridges, highways, farmland as well as residential, private and commercial properties was amplified by current urbanization. An evacuation mission was carried out for about 110 thousand residents who were housed in the disaster relief zones at the peak of the Johor flood, and the death toll was 18 people. A more recent disaster between December 2014 and January 2015 in Kelantan was recorded as the worst one ever occurred in Peninsular Malaysia (Azlee, 2015). Over 200 thousand residents were at the flood relief centers during the worst flood point, and 24 people were killed during the worst flood struck in the world. Other than that, on the 11<sup>th</sup> November 2018, flash floods hit parts of the Kuala Lumpur city following two hours of heavy downpour (Kumar, 2018). The latest flash flood occurred in December 2019, forcing many Malaysians to flee to the relief centers due to continuous heavy downpour in Terengganu and Kelantan. In Southeast Asia that is known for its humid tropics' climate, generally, the torrential rain mostly occurs between 2 and 4 hours. Hence, factors such as seasonal monsoon wind flow, relative humidity, topography, and distance from the sea could heavily influence the local heterogeneity. The east coast states of Malaysia face issues of annual flooding largely due to the monsoon rains from the year end to the start of year. This critical situation causes overflowing rivers, triggering the flood event.

In addition to flood, torrents of rain will endanger public health, overwhelm sewage treatment plants and rise microbial water pollutants. Heavy rain can also cause landslides, agricultural destruction, buildings and bridges to collapse, destroy houses, and havoc on roads and transport, with enormous financial losses. Due to that,



meteorologists have studied the rainfall trend in Malaysia with an emphasis on extreme rainfall events. This has become an alarming problem; thus, this study's findings would be able to steer the directions for climatologists or hydrologists in suggesting further measures to mitigate the destructions caused by the flooding hazard as well as carrying out essential preventive actions for any possible future recurrence.

However, Malaysian hydrologist suffered a lack of complete procedure in analyzing rainfall patterns specifically the long time series analysis of torrential rainfall of Peninsular Malaysia due to the shortage of detailed quantitative research. This was partially because of the lack of long-recorded stations and the issue of missing records (Wong et al., 2009). The number of rain gauge stations with full records in Malaysia is very limited due to the missing data that usually occurred for different reasons such as rainfall station relocation, change in environment, defective tools and network reorganization (Kamaruzaman et al., 2017). To obtain accurate results, the missing rainfall data needs to be handled properly for an accurate inference about the rainfall patterns.

In this study, 'spatial' can be described as being related to or consuming space, whereas 'temporal' indicates being related to time. Hence, when the data is collected in both space and time, the combination of both terms into 'spatiotemporal' also known as 'spatial temporal' is employed in the data processing. Specifically, it can be associated with phenomena at specific locations and times. Researches focusing on clustering-based approach in distinguishing spatiotemporal rainfall patterns were conducted to enumerate the features of an observation set where it placed the data into the same groups. This implied that the rainfall patterns were comparably highly-structured (Mitra et al., 2018; Priyan, 2015; Wickramagamage, 2010).



The clustering-based approach is commonly known as a statistical tool that efficiently manages the duty of regions grouping and determines the time intervals according to the grouping findings, which indicates the rainfall episodes. Nonetheless, for tropical climate, the use of classical clustering methods in determining the spatial and temporal rainfall patterns is practically inappropriate for various qualities of the rainfall data. Firstly, the regions in the tropics basically have been experiencing rainfall all year round. Regardless of the tropical regions having dry and wet seasons, the overall rainfall has no significant varieties in contrast to regions with four seasons (Shaharudin et al., 2018). With large differences in regional geographic and atmospheric characteristics, the variability of rainfall is different across geographical regions. Due to this, discerning specific cluster patterns can be a challenge for such rainfall data.



Due to that, a lengthy time series of the analyzed rainfall data tends to generate a data set with high dimensions. High dimensional data generally contains many irrelevant features that will affect the accuracy of the results. Therefore, feature selection is considered as an essential procedure in the high dimensional data processing (Zhang & Cao, 2019). High dimensional data also complexifies the data when extracting a substantial fact since such level of dimensions would encompass greatly inappropriate or unnecessary input that degrades the analysis findings.

Furthermore, classical clustering requires certain assumptions which contradict the characteristics of precipitation. For instance, the clustering techniques operate on a full-dimensional space. In clustering two-dimensional daily rainfall database, clustering a full-dimensional space corresponds to clustering the days while assuming that all rainfall variables behave in a similar manner. However, since variations are bound to



exist between the different rainfall variables, this assumption is not practical. Additionally, classical clustering methods normally split the rainfall patterns database of rainfall patterns into some clusters; this is by considering that each data is categorized into a certain cluster. Suggesting the exclusiveness of every data to a single cluster even though, in practice, this is not accurate.

On account of the above-mentioned rainfall matters, this research is constructed for the purpose of handling the missing data of the East Coast of Peninsular Malaysia as the pre-processing step. The dimensions are then reduced to define the cluster pattern of torrential rainfall using several method and approaches. The core of this study is the East Coast of Peninsular Malaysia as it is the most affected states from the monsoon rains occurrence, especially from the year end to the beginning of the following year.

This study would be constructive in countering the problems for the affected states if they recur in the future.

## 1.2 Study Area

The East-Coast regions of Peninsular Malaysia which is the focal area of this study, specifically, Kelantan, Terengganu and Pahang with the latitude from 3.5° N to 6.5° N and the longitudes from 102° E to 104° E. In Peninsular Malaysia, the temperature is generally between 21°C to 32°C with all-year warm and humid weather, which ideally describes a humid tropical climate. The resource for the high-dimensional datasets employed in this research is the Department of Irrigation and Drainage Malaysia (DID), in which the 32-year interval daily rainfall data covered the necessary information from

1987 to 2018. As stated by Chin et al. (2016), with the use of the tipping bucket rain gauge, the raw rainfall data was measured with a sensitivity of 0.5 mm per tip, while the tipping bucket measurement used was 0.2 mm per tip. The tipping bucket used contains a funnel that gathers the rain and then was channeled into a narrow seesaw-like jar. After an amount of rainfall equals to 0.22 mm falls, the lever tips dumped the collected water and sent an electrical signal to document the rainfall. The rainfall dataset employed for the purpose of this study was gathered from 48 stations and 11,680 days, which means there are adequate data to be able to identify the rainfall pattern. The rainfall stations' locations were scattered on various geographical coordinates around the East Coast of Peninsular Malaysia. The selection of these areas was focused on the duration, dependability and consistency of the daily rainfall data spanning over 3 decades. Figure 1.1 and Table 1.1 illustrates the locations of all 48 rainfall stations sites all over Peninsular Malaysia's East Coast.

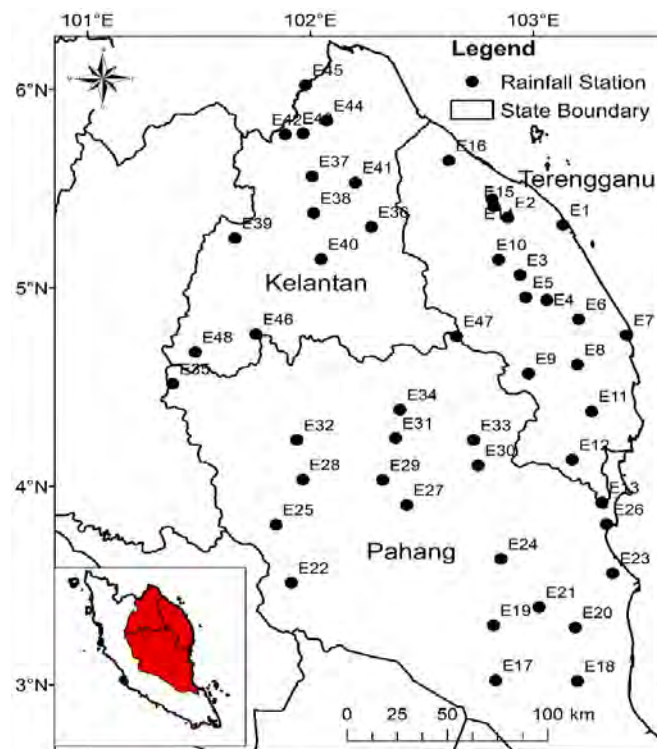


Figure 1.1. The locations of 48 rainfall stations in East Coast Peninsular Malaysia

Table 1.1

*Geographical coordinates of 48 rainfall stations in the East Coast of Peninsular Malaysia*

Stations	Code	Stations	Code
Setor JPS KT	E1	Stor JPS Raub	E25
Kg. Sg. Tong	E2	Pejabat JPS Pahang	E26
Kg. Dura	E3	Rumah Pam Paya Kangsar	E27
Kg. Menerong	E4	JKR Benta	E28
Kg. Embong Sekayu	E5	Kg. Sg. Yap	E29
Jambatan Jerangau	E6	Kawasan B Ulu Tekai	E30
SM Sultan Omar	E7	Kg. Merting	E31
Al-Muktafi	E8	Bkt. Betong	E32
Rumah Pam Paya Kempian	E9	Ulu Tekai (A)	E33
Sg. Gawi	E10	Kuala Tahan	E34
Jambatan Tebak	E11	Gunong Brinchang	E35
Kg. Ban Ho	E12	Kg. Laloh	E36
Hulu Jabor	E13	Ulu Sekor	E37
Kg. Batu Hampar	E14	Dabong	E38
Klinik Chalok Barat	E15	Gob	E39
Inst. Pertanian Besut	E16	Balai Polis Bertam	E40
Sg. Kepasing	E17	Sek. Men. Teknik Kuala Krai	E41
Temeris	E18	Air Lanas	E42
Sg. Cabang Kanan	E19	Kg. Durian Daun	E43
Kg. Unchang	E20	Bendang Nyior	E44
Kg. Batu Gong	E21	Rumah Kastam	E45
Kuala Marong	E22	Blau	E46
Rumah Pam Pahang Tua	E23	Gunung Gagau	E47
Pintu Kawalan Pulau Kertam	E24	Brook	E48

### 1.3 Database for Torrential Rainfall Data

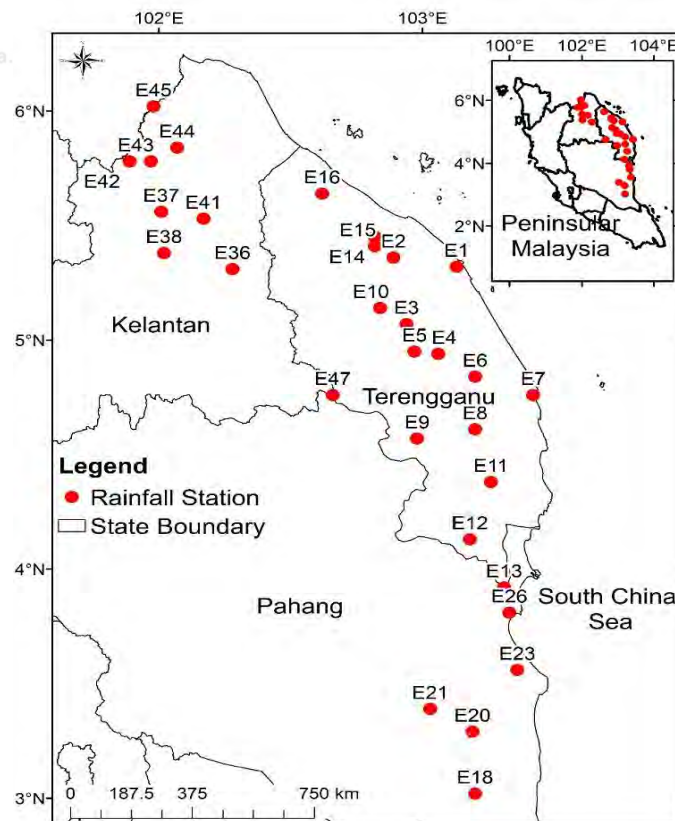
The frequency of heavy rainfall defined as torrential rainfall is the core of this study.

To allow for a reasonable discrepancy of factors that signifies a day of torrential rainfall,

it is thus important to select certain standards that would contribute to the setting of a



threshold. Representing tropical climates, the greatest recorded threshold was 60 mm/day based on the categorization of rainfall intensity by the Department of Irrigation and Drainage Malaysia (DID). The unit for rainfall calculation is presented as mm/day or millimeters per day, precisely, indicating the total rainwater depth (mm) in 24 hours (day). After filtering the total days from the data, the days with rainfall amount of over 60 mm/day were accepted for research purposes (Shaharudin et al., 2018; Tan et al., 2015). After the 11 680-day dataset of over 48 stations was filtered accordingly to the threshold standards for at least 1.5% of the overall stations, the new dataset obtained was 175 days and 30 stations, which were adequate for the representation of the major torrential rainfall centers. The new dataset of torrential rainfall for the East Coast of Peninsular Malaysia is as follows:



*Figure 1.2.* The locations of the observed 30 rainfall stations in the East-Coast of Peninsular Malaysia

Table 1.2

*Geographical coordinates of observed 30 rainfall stations in East-Coast of Peninsular Malaysia*

Stations	Code	Stations	Code
Setor JPS KT	E1	Inst. Pertanian Besut	E16
Kg. Sg. Tong	E2	Temeris	E18
Kg. Dura	E3	Kg. Unchang	E20
Kg. Menerong	E4	Kg. Batu Gong	E21
Kg. Embong Sekayu	E5	Rumah Pam Pahang Tua	E23
Jambatan Jerangau	E6	Pejabat JPS Pahang	E26
SM Sultan Omar	E7	Kg. Laloh	E36
Al-Muktafi	E8	Ulu Sekor	E37
Rumah Pam Paya Kempian	E9	Dabong	E38
Sg. Gawi	E10	Sek. Men. Teknik Kuala Krai	E41
Jambatan Tebak	E11	Air Lanas	E42
Kg. Ban Ho	E12	Kg. Durian Daun	E43
Hulu Jabor	E13	Bendang Nyior	E44
Kg. Batu Hampar	E14	Rumah Kastam	E45
Klinik Chalok Barat	E15	Gunung Gagau	E47

#### 1.4 Problem Statement

Rainfall pattern in both spatial and temporal scales are the most likely evident of the changes occurring in the earth's climate system (Ayugi et al., 2016). To estimate the collective risk in flood insurance, infrastructure networks and water resource management applications, the required rainfall data should be large-scale (Serinaldi & Kilsby, 2014). Rainfall pattern identification matter has become a very significant issue and the results of related studies' findings were applicable as an outline for hydrologists to recommend precautions to minimize the loss. However, for some reasons, identifying daily spatiotemporal torrential rainfall patterns may be a challenge.

Firstly, the number of rain gauge stations with complete records is very limited in Malaysia due to the missing data which normally occurs for various reasons (Kamaruzaman et al., 2017). Since this study focused on torrential rainfall pattern



identification, it is significant to analyze a complete dataset. The missing part may carry important information of the study. Analyzing incomplete dataset may lead to inaccurate rainfall patterns identification, imprecise statistical inference and even incorrect conclusions. Hence, the missing rainfall data needs to be handled properly for an accurate inference about the rainfall patterns. To obtain the most efficient imputation methods, the missing mechanism needs to be analyzed. Due to the stated issue, a proper procedure in handling the missing rainfall data especially for tropic regions is identified to obtain precise valuation of rainfall.

Other than that, tropical rain has a type of tropical climate where the region experiences rainfall throughout the year. Therefore, the rainfall count has no significant discrepancy for each day as compared to that of the four-season regions' rainfall amount. Therefore, it is tough discerning a specific pattern of cluster for this rainfall data form. Due to that, a long time series of monitored rainfall data is needed to obtain a better cluster partition of the torrential rainfall that tends to produce a highly-dimensional data set. This study focused on 32 years of daily torrential rainfall data for 48 rainfall stations over East Coast Peninsular Malaysia. It was equivalent to hundred thousands of data. In high dimensional data set, it can be difficult to classify rainfall patterns as it can involve a high degree of unnecessary and redundant information that greatly degrades the performance of a further study such as forecast or even modelling. The data quality can be improved by removing these redundant, irrelevant, and noisy data. High dimensional dataset is challenging to visualize and interpret. Some algorithms do not perform well on the large number of dimensions taken (Malavika & Selvam, 2015). Thus, reducing the dimensionality helps an algorithm to work efficiently and improves the accuracy of the analysis. Nonetheless, there were variables





with certain significance and should be recognized to deliver noteworthy facts to understand the underlying process. A statistical approach, hence, is needed for dimensionality reduction that could simultaneously maintain the significant information of the data. Dimensionality reduction indicates the process to condense the feature set through the step of selecting an appropriate subset of the original features and eliminating the redundant and irrelevant features of the datasets (Reddy & Reddy, 2010).

Besides that, the amount of rainfall in Peninsular Malaysia (tropical regions) has no significant variations in contrast to subtropical climate regions having annual four seasons. Because of this, it is a challenging feat of discerning a certain cluster pattern for this rainfall data form (Shaharudin et al., 2018). For a study in classifying day-to-day spatiotemporal torrential rainfall patterns, a bigger number of clusters would be required to explain the distinctive rainfall pattern forms in the regions. Every pattern displays identifiable features. Therefore, a highly robust procedure in improving the cluster partition of the tropic rainfall data is mandatory.

Other constrictive aspect of classifying the daily spatiotemporal torrential rainfall data would be that the classical clustering requires certain assumptions that contradict the characteristics of rainfall in which it separates the database of rainfall patterns to generate different clusters with a supposition that each of the weather pattern could only fit a certain cluster. Using classical clustering, each day is respectively assigned to only one cluster (Padilha & Campello, 2017). On the other hand, each day must be assigned to some clusters since practically, it might be a rainy day for day 1 at station A but it also might be a sunny day for day 1 at different stations. These



assumptions may not reflect the reality since a day could take part in several weather for different stations.

Hence, a statistical model of a robust PCA combined with a supervised clustering is proposed to the daily torrential rainfall analysis. Its purpose is countering the issues in the clustering approach to identify the rainfall patterns in the East Coast of Peninsular Malaysia.

## 1.5 Research Objectives

For this study, four main research objectives are highlighted:

1. To propose suitable imputation methods in handling the missing daily rainfall data in the East Coast of Peninsular Malaysia.
2. To identify the spatiotemporal torrential rainfall patterns in the East Coast of Peninsular Malaysia through the use of the Principal Component Analysis combined with cluster analysis.
3. To develop a statistical model to improve the clustering results in identifying the rainfall patterns.
4. To evaluate and compare the proposed statistical model against the classical procedure in identifying the rainfall patterns in the East Coast of Peninsular Malaysia.

## 1.6 Significance of Study

The propositions of this study are:

- a) Discovering the most suitable method in handling the missing daily rainfall data in the East Coast of Peninsular Malaysia.
- b) Establishing the pattern of torrential rainfall in the East Coast of Peninsular Malaysia through the use of the clustering method, in which each cluster has been known to relate to particular topographic features.
- c) Establishing a novel approach in a multivariate technique for specific hydrologic applications particularly in the tropical climate.
- d) Assisting in an established analysis and enhancement of suitable models in predicting the torrential rainfall events in the East Coast of Peninsular Malaysia.

## 1.7 Notations

To alleviate the computation process, the database is constructed to take into the form of a large rectangular  $n$  rows by  $p$  columns matrix  $\mathbf{X}$ , with  $n > p$ . We denote  $x_{ij} \in \mathbf{X}$  to be the rainfall amount for each  $i^{th}$  at each  $j^{th}$  rainfall station where  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ .

Throughout this thesis, the terms ‘rows’ of the data matrix refers to the rainfall observations while ‘columns’ of the data matrix refers to the rainfall station.

## 1.8 Research Methodology

Concisely, this thesis proposes a statistical model to identify the torrential rainfall patterns in the East Coast of Peninsular Malaysia in which it provides a directed cluster rainfall pattern related to the torrential rainfall occurrences respective of the regions. The research methodology is illustrated in Figure 1.3. Handling the missing data is the pre-processing step for this study. The missing percentage as well as the missing data mechanism have been analyzed in order to choose the suitable methods for missing data imputation. The imputation methods are then compared to choose the most appropriate approach for the missing rainfall data of the East Coast of Peninsular Malaysia. Subsequently, the classical Principal Component Analysis (PCA) is employed to find the torrential rainfall patterns. The classical PCA is then found to be unsuitable for the tropical climate because of some problems with the rainfall data in the tropics. To overcome such challenges, a robust approach in PCA is proposed. Robust PCA then coupled with supervised clustering to cluster the partition of rainfall data. Additionally, the effectiveness of the proposed statistical model is then assessed through a simulation.

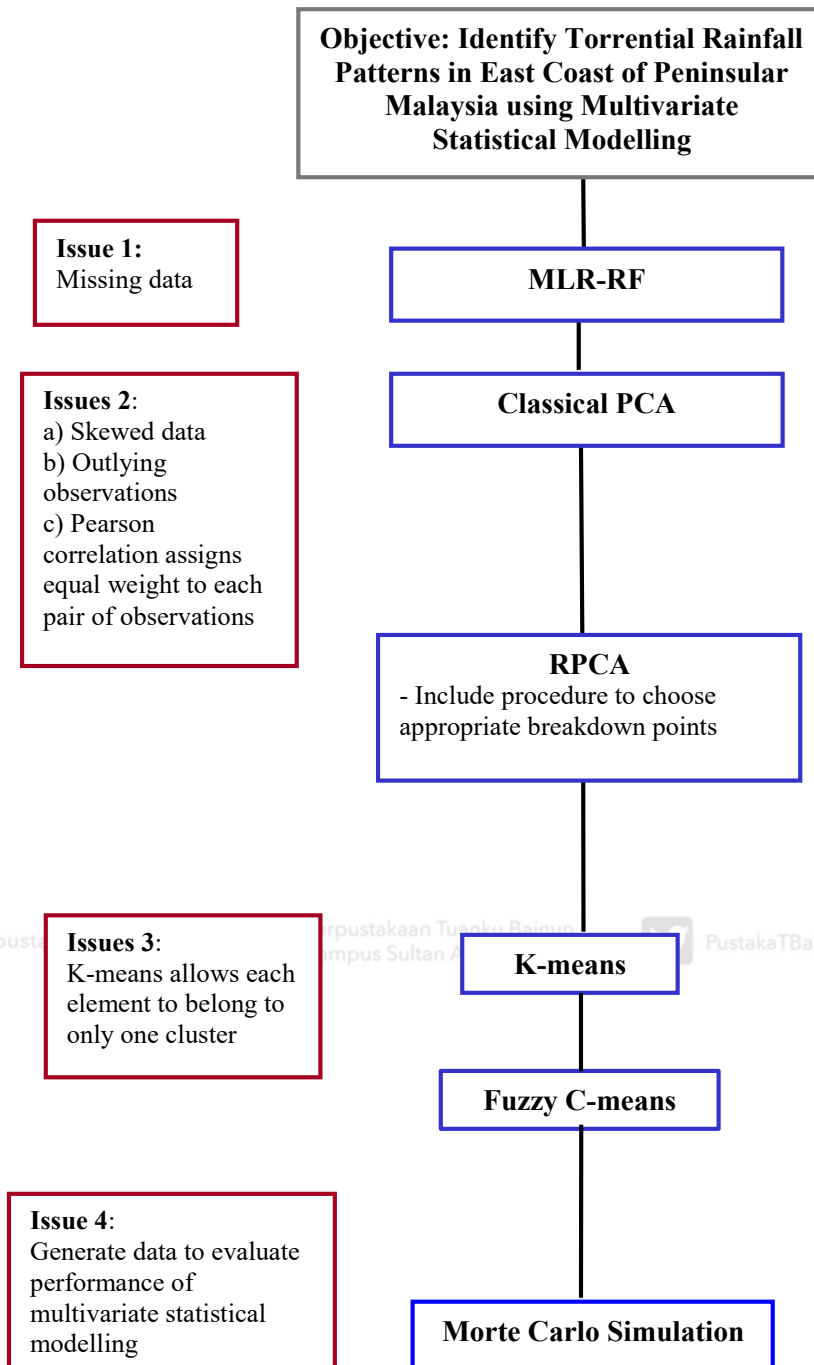


Figure 1.3. Flow Chart of Research Methodology

## 1.9 Thesis Organization

The background of the study and the associated problems in describing the daily torrential rainfall patterns of the East Coast of Peninsular Malaysia are provided in the first chapter of this study. Furthermore, this chapter clarifies the study area and data set



employed including the study objectives, significance of the study, notations and research methodology. In essence, a general outline of the thesis is provided in this section. Additionally, the current research linked to this field of study and the approaches associated with this thesis will be discussed in the following chapter. Chapter 2 starts with ascertaining the rainfall pattern through multivariate analysis which focuses on previous literatures with similar objectives and methods for the data analysis. The next section then centers on the robust methods of verifying the suitable robust procedure for the data set.

Consequently, in Chapter 3, a procedure of identifying the daily torrential rainfall patterns in the East Coast of Peninsular Malaysia is discussed along with the pre-processing phase of the study, which is the handling of the missing data. Problems raised from data with high dimensions are then clarified so that the generic methods applied in the hydrology field could be explained. An enhanced PCA, which is considered more appropriate, is then proposed to the data set. The simulation procedure is further included to assess the performance of Robust PCA. Meanwhile, Chapter 4 will be discussing and presenting the findings of the torrential rainfall pattern identification using this enhanced procedure. This chapter provides a comparison of the results from the classical PCA and Robust PCA. The results of the simulated data also been discussed to assess the performance of the proposed statistical model.

Finally, the findings and discussions of all of the issues discussed in this study will be concluded in Chapter 5. The summary and future study that could be carried out for a better understanding of the specified problems will be addressed clearly in this chapter.

### 1.10 Limitations of Study

In this study, the suggested methodologies are only explained using a series of torrential rainfall in the East Coast of Peninsular Malaysia. Due to that, this research does not focus on all of the rainfall stations in Peninsular Malaysia. The East Coast of Peninsular Malaysia had been identified as the region that is most affected by the monsoon, which comes with torrential rainfall, annually. For the purpose of analyzing a bigger dimension of the data set such as the rainfall data for Peninsular Malaysia, a more powerful and higher speed of computer would be needed. In view of processing the higher dimensional data set, it might be lengthy and time consuming. In addition, the rainfall data set of this study has been chosen without considering the environmental condition of the rainfall stations. Other than that, this study only focuses on the torrential rainfall events seeing how impactful these disastrous occasions are to the human life.