# PREDICTIVE MODELLING FROM MONTHLY RAINFALL PATTERNS USING IMPUTATION APPROACHES COMBINED WITH MULTIVARIATE ANALYSIS

MUHAMAD AFDAL BIN AHMAD BASRI



SULTAN IDRIS EDUCATION UNIVERSITY

2024

# PREDICTIVE MODELLING FROM MONTHLY RAINFALL PATTERNS USING IMPUTATION APPROACHES COMBINED WITH MULTIVARIATE ANALYSIS

MUHAMAD AFDAL BIN AHMAD BASRI



DISERTASI DIKEMUKAKAN BAGI MEMENUHI SYARAT UNTUK MEMPEROLEHI IJAZAH SARJANA SAINS
(MOD PENYELIDIKAN)

FAKULTI SAINS DAN MATERMATIK
UNIVERSITI PENDIDIKAN SULTAN IDRIS

2024

UPSI/IPS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Sila tanda (√)

| | |
|---|---|
| Kertas Projek | |
| Sarjana Penyelidikan | ✓ |
| Sarjana Penyelidikan dan Kerja Kursus | |
| Doktor Falsafah | |

**INSTITUT PENGAJIAN SISWAZAH**

**PERAKUAN KEASLIAN PENULISAN**

Perakuan ini telah dibuat pada ...14...(hari bulan)......05....... (bulan) 20...24...

**i. Perakuan pelajar :**

Saya, MUHAMAD AFDAL BIN AHMAD BASRI, M20201000214, SAINS DAN MATEMATIK (SILA NYATAKAN NAMA PELAJAR, NO. MATRIK DAN FAKULTI) dengan ini mengaku bahawa disertasi/tesis yang bertajuk PREDICTIVE MODELLING FROM MONTHLY RAINFALL PATTERNS USING IMPUTATIONS APPROACHES COMBINED WITH MULTIVARIATE ANALYSIS.

adalah hasil kerja saya sendiri. Saya tidak memplagiat dan apa-apa penggunaan mana-mana hasil kerja yang mengandungi hak cipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hak cipta telah dinyatakan dengan sejelasnya dan secukupnya

afdal
Tandatangan pelajar

**ii. Perakuan Penyelia:**

Saya, DR SHAZLYN MILLEANA SHAHARUDIN (NAMA PENYELIA) dengan ini mengesahkan bahawa hasil kerja pelajar yang bertajuk PREDICTIVE MODELLING FROM MONTHLY RAINFALL PATTERNS USING IMPUTATIONS APPROACHES COMBINED WITH MULTIVARIATE.

(TAJUK) dihasilkan oleh pelajar seperti nama di atas, dan telah diserahkan kepada Institut Pengajian Siswazah bagi memenuhi sebahagian/sepenuhnya syarat untuk memperoleh Ijazah ______________________ (SLA NYATAKAN NAMA IJAZAH).

14/5/2024
Tarikh

Tandatangan Penyelia

UPSI/IPS-3/BO 31
Pind.: 01 m/s:1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

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

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

Tajuk / *Title*: PREDICTIVE MODELLING FROM MONTHLY RAINFALL PATTERNS USING IMPUTATIONS APPROACHES COMBINED WITH MULTIVARIATE ANALYSIS.

No. Matrik /*Matric's No.*: M20201000214

Saya / *I* : MUHAMAD AFDAL BIN AHMAD BASRI

(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*** | |

afdal

(Tandatangan Pelajar/ Signature)

Dr. SHAZLYN MILLEANA SHAHARUDIN
Senior Lecturer
Department of Mathematics
Faculty of Science and Mathematic
Universiti Pendidikan Sultan Idris
35900 Tanjung Malim, Perak

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

Tarikh: 14/5/2024

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 CONFIDENTAL 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 and my father, Fatimah Abustan and Basri Daud for always faithfully listening to the ups and downs in finishing this journey. Not forgetting also for my wife and friends who are endlessly giving me words of encouragement and praying for this success. I believe I could not have reached this point without the prayed and support from all of them. I would also like to thank the Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris for the chance and facilities provided, and also a special thanks to the Meteorological, Environment, and Geophysical Agency, Indonesia for the data and information provided for this study.

# ABSTRACT

This study identifies torrential rainfall patterns in Yogyakarta, Indonesia using multivariate and univariate approaches to propose a statistical model for solving associated issues. First, addressing its long-gap missing rainfall data (approximately 52.8%) is crucial. Therefore, classical imputation methods were enhanced by combining them with the bootstrap algorithm. The hybrid of Random Forest (RF) and the bootstrap algorithm, with the lowest Root Mean Square Error (RMSE) of 7.96 and Mean Absolute Error (MAE) of 0.29, is the best statistical method for imputing Yogyakarta rainfall data missing values. Cluster analysis then classified stations into different rainfall regimes; hierarchical clustering analysis (HCA) recognised four distinct, homogenous regions. The multivariate approach, principal component analysis (PCA), homogenised the rainfall series and optimally reduced long-term rainfall data to validate the HCA analysis results. From the 75% cumulative variation, 14 factors for the Dry and the Rainy seasons and 12 factors for the Intermonsoon season were extracted using varimax rotation. Subsequently, the rainfall pattern forecasting was done using a univariate approach, Singular Spectrum Analysis (SSA). Recurrent Forecasting and Vector Forecasting-Singular Spectrum Analysis (RF-SSA and VF-SSA) were proposed by establishing length ($L$) and eigentriple ($ET$) parameters. This forecasting model effectively discriminated noise in a time series trend, producing significant forecasting results. Overall, the best performances are from $L = 6$ and $L = 5$ for the Rainy and Dry seasons and the Intermonsoon season, respectively, based on the lowest MAE and Mean Forecast Error (MFE). For the dry season, $L = 6$ has the lowest MAE (2.7116); the MFE tends to over forecast monthly rainfall data by 0.0126%. For the Intermonsoon season, $L = 5$ has the lowest MAE (3.3940); the MFE tends to under forecast by -0.0621%. All results utilised the RF-SSA algorithm. Concisely, the hybrid of the bootstrap algorithm and SSA improves the forecasting results. Different data classifications may advance climate warning systems.

**PEMODELAN DESKRIPTIF BERDASARKAN CORAK HUJAN BULANAN MENGGUNAKAN GABUNGAN PENDEKATAN IMPUTASI DENGAN ANALISIS MULTIVARIAT**

## ABSTRAK

Kajian ini mengenal pasti corak hujan lebat di Yogyakarta, Indonesia menggunakan pendekatan multivariat dan univariat untuk mencadangkan model statistik bagi menyelesaikan isu berkaitan. Pada asasnya, penting untk menangani isu data hujan dengan jurang yang besar (kira-kira 52.8%). Oleh itu, kaedah imputasi klasik telah dipertingkatkan dengan menggabungkannya dengan algoritma *bootstrap*. Hibrid *Random Forest* (RF) dan algoritma *bootstrap* merupakan kaedah statistik terbaik untuk mengira data hilang dalam data hujan Yogyakarta yang menunjukkan Ralat Punca Min Kuasa Dua (RMSE) terendah iaitu 7.96 dan Min Ralat Mutlak (MAE) terendah iaitu 0.29. Analisis kluster kemudian mengelaskan stesen-stesen hujan ke dalam rejim yang berbeza; analisis pengelompokan hierarki (HCA) mengiktiraf empat kawasan homogen yang berbeza. Pendekatan multivariat, iaitu analisis komponen utama (PCA), menyeragamkan siri hujan dan mengurangkan data hujan jangka panjang secara optimum untuk mengesahkan keputusan analisis HCA. Daripada 75% variasi kumulatif, 14 faktor untuk musim Kering dan Hujan dan 12 faktor untuk musim Antara Monsun telah diekstrak menggunakan putaran varimaks. Tambahan, peramalan corak hujan dilakukan menggunakan pendekatan univariat iaitu Analisis Spektrum Tunggal (SSA). Ramalan Berulang dan Ramalan Vektor-Analisis Spektrum Tunggal (RF-SSA dan VF-SSA) telah dicadangkan dengan mewujudkan parameter panjang (L) dan *eigentriple* (ET). Model ramalan ini mendiskriminasi bunyi secara berkesan dalam aliran siri masa dan menghasilkan hasil ramalan signifikan. Secara keseluruhan, prestasi terbaik adalah daripada $L = 6$ dan $L = 5$ untuk musim Hujan dan Kering dan musim Antara Monsun, masing-masing berdasarkan MAE dan Ralat Ramalan Min (MFE) terendah. Bagi musim kemarau, $L = 6$ mencatatkan MAE terendah (2.7116) manakala MFE cenderung meramal data hujan bulanan secara lebih sebanyak 0.0126%. Bagi musim Antara Monsun, $L = 5$ menunjukkan MAE terendah (3.3940) manakala MFE cenderung meramal secara kurang sebanyak -0.0621%. Semua keputusan menggunakan algoritma RF-SSA. Secara ringkas, gabungan algoritma *bootstrap* dan SSA meningkatkan hasil ramalan. Klasifikasi data yang berbeza dikira berupaya memajukan sistem amaran iklim.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AR | Autoregressive |
| BMKG | Regulation of Head of Meteorology, Climatology, and Geophysics Agency |
| BNPB | The National Agency for Disaster Management |
| CA | Cluster analysis |
| EM | Expectation-maximisation |
| EM-MCMC | Monte Carlo Markov chain method-based expectation-maximization |
| ET | Eigentriple |
| EV | Eigenvectors |
| FA | Factor analysis |
| GDP | Gross domestic product |
| HCA | Hierarchical Clustering Analysis |
| IDWA | Inverse Distance Weighting Algorithm |
| IPCC | The Intergovernmental Panel of Climate Change |
| KNN | K-nearest neighbor |
| LRF | Linear recurrent formula |
| LRRs | Linear recurrence relations |
| MAE | Mean Absolute Error |
| MAR | Missing at Random |
| MCAR | Missing Completely at Random |
| MCMC | Markov Chain Monte Carlo |
| MI | Multiple imputations |
| MNAR | Missing Not at Random |

| | |
|---|---|
| MVI | Missing value imputation |
| NIPALS | Nonlinear Iterative Partial Least Squares |
| NN | Nearest Neighbor |
| NR | Normal Ratio |
| PC | Principal components |
| PCA | Principal component analysis |
| PD | Pairwise deletion |
| RF | Random Forest |
| RF-B | Random Forest-Bootstrap |
| RF-SSA | Recurrent Forecasting-SSA |
| RSME | Root Mean Square Error |
| SI | Single imputations |
| SSA | Singular Spectrum Analysis |
| SVD | Singular Value Decomposition |
| VF-SSA | Vector Forecasting-SSA |

# LIST OF SYMBOLS

| | |
|---|---|
| $\mathbf{O}$ | Matrix of dummies |
| $\mathbf{X}$ | Data matrix |
| $a$ | Centroid of clusters *A* |
| $b$ | Centroids of clusters *B* |
| $A$ | Cluster with centres *a* |
| $B$ | Cluster with centres *b* |
| $B(t)$ | Number of bootstrap replications |
| $C$ | Number of iterations for convergence. |
| $C_{tt}$ | Correlation matrix |
| $d$ | Variables |
| $d(A,B)$ | Ward's distance between two clusters *A* and *B* |
| $d_q(x_i,x_j)$ | Computation formula of the distance |
| $\mathbf{E}_t$ | Matrix with columns of eigenvectors |
| $g$ | Components |
| $h_{d,g}$ | Each entry of matrix of component loadings |
| $H_c$ | $n \times n$ matrix of component loadings |
| $i$ | Index |
| $j$ | Index |
| $K$ | Number of clusters |
| $l_k$ | Variance of principal component |
| $m$ | Number of components |
| $n$ | Number of rainfall days |

| | |
|---|---|
| $n_A$ | Frequency of clusters *A* |
| $n_B$ | Frequency of clusters *B* |
| *NA* | Number of the missing values in the categorical variables |
| $o_{is}$ | Corresponding matrix of dummies |
| $p$ | Rainfall stations |
| $p_h$ | Principal components |
| $P_i$ | Initial observed rainfall data |
| $P_x$ | Observed rainfall data |
| Q | Diagonal matrix of eigenvalues of correlation matrix |
| $\mathfrak{R}$ | Coefficient vector |
| $t_h$ | Principal components |
| $t_m$ | Cumulative percentage of the variance of $m$ |
| $T$ | Total number of bootstrap replications |
| $u_i$ | Left singular vectors |
| $U$ | $L \times L$ orthogonal matrix |
| $U_j$ | Eigenvector |
| $v$ | Value in the domain of the target feature |
| $v_i$ | Right singular vectors |
| $V$ | $K \times K$ orthogonal matrix |
| $x$ | Rainfall days |
| $x_i$ | Measured value of rainfall |
| $x_{i,j}$ | Elements in the input matrix, i.e. rainfall amount |
| $x_{is}$ | The $i$th observation of the $s$th variable |

| | |
|---|---|
| $X'$ | Spatial T-mode matrix |
| $\mathbf{X}_i$ | Elementary matrices |
| $X_s$ | Arbitrary variable |
| $\tilde{X}_i$ | Reconstructed columns of trajectory matrix |
| $X_{new}^{imp}$ | Newly imputed matrix |
| $X_{old}^{imp}$ | Previously imputed matrix |
| $x_{1,i}^{B(t)}$ | Bootstrap replication data through sampling with replacement |
| $y_i$ | Observed rainfall data |
| $\bar{y}_i$ | Average rainfall data over rainfall stations |
| $\tilde{y}_i$ | Predicted rainfall data |
| $y_{mis}^{(s)}$ | Missing values of variable $X_s$ |
| $y_{obs}^{(s)}$ | Observed values of variable $X_s$ |
| $Y_{miss}$ | Missing values |
| $Y_{obs}$ | Observed data |
| $y_t$ | Reconstructed series |
| $\tilde{\mathbb{Y}}_T$ | Reconstructed time series value |
| Z | Rank of $\mathbf{X}'$ |
| $Z_i$ | Reconstructed time series value |
| λ | Eigenvalue |
| $\vec{e}_t$ | Elements of eigenvector matrix |
| λ | Eigenvalue that obtained from correlation matrix |
| $\theta$ | Parameter of interest |
| Σ | Diagonal matrix |

| | |
|---|---|
| $\Sigma_{ii}$ | Nonnegative real diagonal entries |
| $\sigma_i$ | Singular values |
| $\sigma_i, u_i, v_i$ | Eigen triple of singular value decomposition |
| $1(Y_j = v)$ | Indicator function |