Asia-Pacific Journal of Atmospheric Sciences
https://doi.org/10.1007/s13143-019-00135-8

Online ISSN 1976-7951
Print ISSN 1976-7633

ORIGINAL ARTICLE

The Effectiveness of a Probabilistic Principal Component Analysis Model and Expectation Maximisation Algorithm in Treating Missing Daily Rainfall Data

Zun Liang Chuan[1] · Sayang Mohd Deni[2] · Soo-Fen Fam[3] · Noriszura Ismail[4]

Received: 21 February 2019 / Revised: 6 May 2019 / Accepted: 24 May 2019

Abstract
The reliability and accuracy of a risk assessment of extreme hydro-meteorological events are highly dependent on the quality of the historical rainfall time series data. However, missing data in a time series such as this could result in lower quality data. Therefore, this paper proposes a multiple-imputation algorithm for treating missing data without requiring information from adjoining monitoring stations. The proposed imputation algorithms are based on the *M*-component probabilistic principal component analysis model and an expectation maximisation algorithm (*M*PPCA-EM). In order to evaluate the effectiveness of the *M*PPCA-EM imputation algorithm, six distinct historical daily rainfall time series data were recorded from six monitoring stations. These stations were located at the coastal and inland regions of the East-Coast Economic Region (ECER) Malaysia. The results of analysis show that, when it comes to treating missing historical daily rainfall time series data recorded from coastal monitoring stations, the 2-component probabilistic principal component analysis model and expectation-maximisation algorithm (2PPCA-EM) were found to be superior to the single- and multiple-imputation algorithms proposed in previous studies. On the contrary, the single-imputation algorithms as proposed in previous studies were superior to the *M*PPCA-EM imputation algorithms when treating missing historical daily rainfall time series data recorded from inland monitoring stations.

Keywords Expectation maximization algorithms · Missing daily rainfall · Probabilistic principal component analysis model · VIKOR technique

1 Introduction

In the past decade, climate change has altered the stationary patterns of historical rainfall time series data, creating a non-stationary component in the time series (Agilan and Umamahesh 2016). In particular, climate phenomena such as the monsoon (Tangang et al. 2012), El Nino-Southern Oscillation (Tangang et al. 2012; Villafuerte and Matsumoto 2015), Indian Ocean Dipole (Cai et al. 2014; Tangang et al. 2012) and Madden-Julian Oscillation (Tangang et al. 2012) have altered the stationary patterns in historical rainfall time series data. Therefore, an appropriate risk assessment model for extreme hydro-meteorological events, which takes into account the non-stationary component, is much needed. Moreover, the reliability and accuracy of risk assessment of extreme hydro-meteorological events are also highly dependent on the quality of the historical time series data. However, the quality of the time series data consistently diminishes due to the presence of missing data. As a consequence, the risk assessment analysis is rendered unreliable and inaccurate.

Technically, missingness in a dataset can be categorised as missing completely at random (MCAR),

Responsible Editor: Edvin Aldrian.

✉ Zun Liang Chuan
chuanzl@ump.edu.my

[1] Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang Kuantan, Pahang DM, Malaysia

[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Malaysia

[3] Faculty of Technology Management and Technopreneurship, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

[4] School of Mathematical Sciences, Faculty Science and Technology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor DE, Malaysia

Published online: 13 June 2019