Abstract of thesis presented to the Senate of University Malaysia Terengganu in fulfilment of the requirements for the degree of Master of Science

TERENGGANU'S RAINFALL TIME SERIES ANALYSIS USING HIERARCHICAL AGGLOMERATIVE CLUSTERING BASED ON DYNAMIC TIME WARPING AND PERSISTENT HOMOLOGY

KIRTHANA DEVI A/P SELVARAJH

JULY 2023

Main Supervisor : **Gobithaasan Rudrusamy, Ph.D**

Co- Supervisor : **Mohd Sofiyan Sulaiman, Ph.D**

Faculty : **Faculty of Ocean Engineering Technology and Informatics**

Time series analysis is a method of studying time series data in order to derive significant characteristics of the changing data. As time series data gets larger, analysing, discovering, and interpreting patterns becomes more difficult. Topological data analysis (TDA) is an emerging data analysis approach that can help us extract insights from time series data. In this study, we carried out classification and clustering on a monthly rainfall dataset collected from 66 rainfall stations located in Terengganu, Malaysia. First, we ran multiple statistical tests and used the results to classify the station's reliability. Following that, we applied Hierarchical Agglomerative Clustering (HAC) using Dynamic Time Warping (DTW) to generate a dendrogram and observe the pattern of the stations clustered together. Next, we employed TDA tool, Persistent Homology (PH) to generate Persistence Diagram (PD) in order to extract the kdimensional holes (H_k) where zero and one dimensional holes (H_0 and H_1) represents it as Persistence Lifespan Curve (LC) to form lifespan feature vectors ($\widetilde{LC_1}, \widetilde{LC_2}, \ldots, \widetilde{LC_{66}}$) whereas one-dimensional holes (H_1) for Persistence Landscape (PL) to form feature vectors ($\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$).The computation of LC lifespan feature vector as well as average PL of all feature vectors is used to compute L^2 norms and represent as distance matrix of each station for the generation of the dendrogram. The classification of rainfall stations was the first outcome of this study, with 33 stations labelled as "useful" and the remaining 33 stations labelled as "doubtful". The second result is a Ward linkage dendrogram with three clusters based on the HAC of rainfall

stations using DTW. Finally, the clusters generated by PH's topological features resulted in a Ward linkage dendrogram with four clusters. The clusters of DTW-HAC are discovered to be clustered based on their comparable statistical characteristics in terms of elevation, maximum monthly rainfall, average, median, missing data, and altitudes. However, the clusters of PH-HAC (LC and PL) clusters based on extreme values (all the highest and lowest mean, median, maximum monthly rainfall, missing values, and elevation). Both approaches have 65%-67% similarity in the way each station is clustered. When the clusters are compared to gridded data from Climate Hazards group Infrared Precipitation with Stations (CHIRPS), the difference is clear; DTW clusters are similar to mean values of spatiotemporal rainfall plot, whereas PH based clusters are similar to median/maximum based on spatiotemporal rain-fall plot.

Abstrak tesis yang dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan untuk Ijazah Sarjana Sains

ANALISIS SIRI MASA HUJAN TERENGGANU MENGGUNAKAN ANALISIS KELOMPOK AGGLOMERATIF BERHIERARKI BERDASARKAN MASA SEPADAN BERDINAMIK DAN HOMOLOGI GIGIH

KIRTHANA DEVI A/P SELVARAJH

JULY 2023

Penyelia Utama : Gobithaasan Rudrusamy, Ph.D

Penyelia Bersama : Mohd Sofiyan Sulaiman, Ph.D

Fakulti : Fakulti Teknologi Kejuruteraan Kelautan dan Informatik

Analisis siri masa ialah kaedah mengkaji data siri masa untuk mendapatkan ciri-ciri penting bagi data yang berubah. Apabila data siri masa semakin besar, menganalisis, menemui dan mentafsir corak menjadi lebih sukar. Analisis data bertopologi (TDA) ialah pendekatan analisis data baru yang boleh membantu kami mengekstrak cerapan daripada data siri masa. Dalam kajian ini, kami menjalankan pengelasan dan pengelompokan pada dataset hujan bulanan yang dikumpul daripada 66 stesen hujan yang terletak di Terengganu, Malaysia. Pertama, kami jalankan pelbagai ujian statistik dan menggunakan keputusan untuk mengklasifikasikan kebolehpercayaan stesen. Berikutan itu, kami menggunakan teknik kelompok agglomeratif berhierarki (HAC) menggunakan masa sepadan berdinamik (DTW) untuk menjana dendrogram dan memerhati corak pengelompokan stesen. Seterusnya, kami menggunakan satu teknik TDA, Homologi Gigih (PH) untuk menghasilkan Rajah Kegigihan untuk mengekstrak lubang k-dimensi (H_k) yang mana lubang sifar dan satu dimensi (H_0 and H_1) mewakilinya sebagai Keluk Jangka Hayat Kegigihan (LC) kepada membentuk ciri jangka hayat ($\widetilde{LC_1}, \widetilde{LC_2}, \dots, \widetilde{LC_{66}}$) manakala satu dimensi (H_1) untuk Landskap Kegigihan (PL) untuk membentuk vektor ciri ($\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$). Pengiraan vektor ciri jangka hayat LC serta purata semua vektor ciri digunakan untuk mengira norma L^2 dan diwakili sebagai matriks jarak setiap stesen untuk penjanaan dendrogram. Pengelasan stesen hujan adalah hasil pertama kajian ini, dengan 33 stesen dilabel sebagai "berguna" dan baki 33 stesen dilabelkan sebagai "meragukan".

Keputusan kedua ialah dendrogram kaitan Wad dengan tiga kelompok berdasarkan HAC stesen hujan menggunakan DTW. Akhirnya, kluster yang dijana oleh ciri topologi PH menghasilkan dendrogram kaitan Wad dengan empat kluster. Kelompok DTW-HAC didapati dikelompokkan berdasarkan ciri statistik setandingnya dari segi ketinggian, hujan bulanan maksimum, purata, median, data yang hilang dan ketinggian. Walau bagaimanapun, kelompok gugusan PH-HAC (LC dan PL) berdasarkan nilai ekstrem (semua min tertinggi dan terendah, median, hujan bulanan maksimum, data yang hilang dan ketinggian). Kedua-dua pendekatan mempunyai 65%-67% persamaan dalam cara setiap stesen dikelompokkan. Apabila kluster dibandingkan dengan data grid daripada *Climate Hazards group Infrared Precipitation with Station* (CHIRPS), perbezaannya adalah jelas; Kelompok DTW adalah serupa dengan nilai min plot hujan spatiotemporal, manakala kelompok berasaskan PH adalah serupa dengan plot hujan spatiotemporal berdasarkan median/maksimum.