Spatial Max-Stable with Cyclic Generalized Extreme Value Model for Extreme Ground-level Ozone

by

Siti Aisyah Binti Zakaria
(1844012836)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Institute of Engineering Mathematics
UNIVERSITI MALAYSIA PERLIS

2024

ACKNOWLEDGEMENT

All praise be to Allah for providing me with the resources and abilities to complete my studies. I'd like to offer my gratitude to those who have been so encouraging during my research and helped to make it feasible. First and foremost, I want to express my gratitude to my loving supervisors, Dr Nor Azrita Mohd Amin, Dr Noor Fadhilah Ahmad Radi, and Prof Madya Dr Nasrul Hamidin, for their invaluable advice, support, and wisdom throughout my PhD journey. It has been a fantastic pleasure to work with such lovely and beautiful supervisors who never gave up on me and continued to encourage me through thick and thin, allowing me to complete this study.

Special thanks go to my beloved husband (Ibrahim bin Sulong) and all my kids (Anwar Mubarrak, Hidayah Nafisah, Akmal Mushaddiq, Hanis Nazeeha and Hajar Nadeera), parents (Almarhum Hj Zakaria bin Mat and Hjh Hamidah binti Mat), parents in law (Almarhum Hj Sulong bin Hamid and Hjh Senik binti Mohd), brothers and sisters for supporting me spiritually throughout completing this study and my life in general. I would also like to thank all my friends that always support me to complete my journey.

Lastly, I'd like to express my gratitude to University Malaysia Perlis and the Ministry of Higher Education for providing me with the opportunity to pursue this level of study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EVT	Extreme Value Theory
BM	Block Maxima
POT	Peak Over Threshold
GEV	Generalized Extreme Value
GPD	Generalized Pareto Distribution
iid	Independent and identically distributed
GLO	Ground-Level Ozone
ppm	Parts per Million
NO_x	Nitrogen Oxides
VOC_s	Volatile Organic Compound
PCA	Principal Component Analysis
PM10	Particular Matter with the size of less than 10 micron
PM2.5	Particular Matter with the size of less than 2.5 micron
SO2	Sulphur Dioxide
NO_2	Nitrogen Dioxide
CO	Carbon Monoxide
MADA	Lembaga Kemajuan Pertanian Muda
DoE	Department of Environment
MRA	Multiple Regression Analysis
EVA	Extreme Value Analysis
NO	Nitric Oxide
MLE	Maximum Likelihood Estimation
ASMA	Alam Sekitar Malaysia Sdn Bhd
MAAQG	Malaysia Ambient Air Quality Guidelines
ADF	Augmented Dickey Fuller
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
AIC	Akaike's Information Criterion
TIC	Takeuchi Information Criterion
CDF	Cumulative Distribution Function
PDF	Probability Distribution Function
RMSE	Root Mean Square Error
BIC	Bayesian Information Criterion
Q-Q	Quantile-Quantile

KS	Kolmogorov-Smirnov
IDW	Inverse Distance Weighted
ECDF	Empirical Cumulative Distribution Function
MySTIE	Malaysian Science, Technology, Innovation and Economic
AA	Arithmetic Average
NR	Normal Ration Weighted
ECG	Electrocardiogram
HCA	Hierachical Cluster Analysis
G1	Cluster Group 1
G2	Cluster Group 2

LIST OF SYMBOLS

$F(x)$	Cumulative distribution function
x	Random variable
μ	Location parameter
σ	Scale parameter
ξ	Shape parameter
θ	Unknown parameter
t	Time
D	Deviance Statistic
ℓ	Maximize log-likelihood
β_0	Mean value
β_1	Rate of change for linear function
β_1 *and* β_2	Harmonic amplitude for cyclic function
$u(t)$	Non-stationary in the location parameter
y	The number of weeks within the period
8 T	Return period, year
$\tilde{\tau}$	Tau
L	Likelihood function
n	Sample size
N	Total number of samples
p	Number of parameters
Z	Test statistic
x_t	Time series
r_t	Random walk
β_t	Deterministic trend
μ_t	Error time of r_t
ε_t	Stationary error
S_t^2	Partial sums of error
s^2	Consistent estimator of σ^2
p	Lag truncation parameter
x_i	Observed value
x_i	Predicted value at index i

ℓ	Maximized log-likehood
p	Number of parameters
D_n	Maximum value between CDF of the observed and generated data
$F_n(x)$	CDF of the observed data
$F(x)$	Generated data
h	Euclidean distance
T	Return period
z_p	Return level
$\gamma(h)$	Semivariogram function

Kestabilan Maksimum Ruang dengan Model Kitaran Nilai Ekstrim Umum untuk Ozon Aras Tanah yang Esktrim

ABSTRAK

Tahap kepekatan ozon aras tanah (GLO) yang tinggi mempunyai kesan buruk yang serius terhadap masalah kesihatan dan mempengaruhi alam sekitar. Kajian ini mengintegrasikan proses kestabilan maksimum ruang dengan model nilai ekstrim umum (GEV) kitaran untuk menganalisis dan meramalkan tahap GLO yang ekstrim. Ruang ekstrim menyediakan rangka kerja untuk menganalisis dan memodelkan tingkah laku peristiwa yang jarang berlaku dengan mengambil kira corak data yang ekstrim dan ciri-ciri beberapa stesen. Salah satu komponen utama dalam pendekatan ini adalah memilih taburan marginal GEV yang sesuai berdasarkan struktur data setiap stesen pemantauan. Pilihan pengedaran marginal bergantung kepada sama ada terdapat trend yang jelas dalam siri data yang ekstrim. Dalam kes di mana data menunjukkan variasi bermusim yang kuat, model kepegunan mungkin tidak sesuai. Kajian ini mengakui variasi bermusim dalam data GLO di pelbagai stesen pemantauan, dipengaruhi oleh pertukaran monsun. Oleh itu, model variasi bermusim dianggap sebagai taburan marginal untuk model ruang ekstrim. Di samping itu, kajian ini memperluaskan model bukan kepegunan dari kes univariat ke model ruang ekstrim. Model ini menggabungkan corak kitaran dalam parameter lokasi untuk melengkapkan taburan GEV sebagai taburan marginal yang baru dalam proses kestabilan maksimum, model pergantungan standard untuk ruang ekstrim yang diperhatikan di lokasi yang berbeza. Proses kelompok menggunakan analisis kelompok hierarki (HCA) mendapati kesemua stesen boleh dikumpulkan kepada dua kelopok bergantung kepada ciri data mingguan maksimum yang sama. Pengesahan model yang dibangunkan adalah penting untuk ramalan yang tepat. Data sintetik yang menghampiri ciri-ciri data sebenar dihasilkan untuk mengesahkan model dan memudahkan ramalan kes-kes ekstrim masa depan berdasarkan nilai pulangan untuk tempoh pulangan tertentu. Tahap pulangan, yang menunjukkan jumlah purata peristiwa ekstrim dalam tempoh pulangan yang ditentukan, digunakan untuk meramalkan tahap kepekatan GLO di lokasi yang berbeza, meningkatkan pemahaman corak kepekatan GLO berdasarkan kategori lokasi. Penyampaian tahap pulangan menghasilkan pemetaan tahap pulangan lebih membantu dalam memvisualisasikan dan menafsirkan ramalan untuk semua stesen pemantauan. Penemuan utama kajian ini menunjukkan tahap pulangan kepekatan GLO meningkat apabila tempoh pulangan meningkat. Hasil menunjukkan hamper semua tahap pulangan melebihi standard MAAQG 8 jam 0.06 ppm. Terutama, stesen di Kota Bahru (CA22) menunjukkan tahap pulangan terendah, sementara stesen Shah Alam (CA25) menunjukkan nilai anggaran tertinggi. Kepekatan GLO yang tinggi di Shah Alam ini disebabkan oleh lokasi bandarannya, yang ditandai dengan kepadatan lalu lintas yang tinggi, operasi perindustrian, dan pelbagai pengaruh meteorologi. Kesimpulannya, kajian ini sangat penting kerana ia menawarkan maklumat berharga yang boleh digunakan dalam bidang alam sekitar dan klimatologi, khususnya mengenai GLO di Semenanjung Malaysia. Metodologi yang terperinci dalam kajian ini boleh disesuaikan untuk analisis set data yang ekstrim yang lain.

Spatial Max-Stable with Cyclic Generalized Extreme Value Model for Extreme Ground-level Ozone

ABSTRACT

The high level of ground-level ozone (GLO) concentration has serious adverse effects on health problems and affects the environment. This study integrates the spatial max-stable processes with the cyclic generalized extreme value (GEV) model to analyze and forecast extreme GLO levels. Spatial extreme provides a framework for analyzing and modelling the behaviour of rare events considering the extreme data pattern and the characteristics of several stations. One key component in this approach is selecting an appropriate GEV marginal distribution based on the data structure of each monitoring station. The choice of marginal distribution depends on whether there is an apparent trend in the extreme data series. In cases where the data exhibit strong seasonal variation, a stationary model may not be appropriate. This study acknowledges the seasonal variation in GLO data at various monitoring stations, influenced by the interchange of monsoons. Therefore, a seasonal variation model is considered as the marginal distribution for the spatial extreme model. Additionally, the study extends the non-stationary model from univariate cases to the spatial extreme model. This model incorporates a cyclic pattern in location parameters to complement the GEV distribution as a new marginal distribution within the max-stable process, a standard dependency model for spatial extreme observed at different locations. The clustering process using hierarchical cluster analysis (HCA) found that all the stations can be grouped into two clusters depending on the same characteristic of the weekly maxima data. The extremal coefficient between 1 to 2 indicates that the stations are dependent on each other's within the cluster. Validation of the developed model is crucial for accurate predictions. Synthetic data approximating real data characteristics are generated to validate the model and facilitate predictions of future extreme cases based on return values for specific return periods. Return levels, indicating the average amount of extreme events within a specified return period, are used to predict GLO concentration levels across different locations, enhancing the understanding of GLO concentration patterns based on location categories. The presentation of return level results in return level mapping further aids in visualizing and interpreting the predictions for all monitoring stations. The main finding of this study indicates that the return level of GLO concentration increased as the return period increased. The results show that most of the return levels exceed the guideline of MAAQG for 8-hour average that is 0.06 ppm. Notably, the station in Kota Bahru (CA22) stands out with the lowest return levels, while the Shah Alam (CA25) station exhibits the highest estimated values. This high GLO concentration in Shah Alam may be attributed to its urban location, marked by high traffic density, industrial operations, and diverse meteorological influences. In conclusion, this study is highly significant as it offers valuable insights that can be applied in the fields of environmental and climatology, specifically regarding GLO in peninsular Malaysia. The methodology detailed in this study can be adapted for the analysis of other extreme datasets.

CHAPTER 1 : INTRODUCTION

1.1 Research Background

Spatial extreme modelling is essential for understanding and predicting rare, high-impact events in various fields such as meteorology, hydrology, and environmental science. These models help in assessing the risk associated with extreme events and their spatial dependencies. The spatial extreme model focuses on the analysis of extreme events that occur over space at multiple locations simultaneously. It aims to capture the spatial dependence among extreme events observed at different locations and estimate the probability of extreme events at each location (Hector & Reich, 2022). Two common approaches for modelling spatial extremes are the max-stable process approach and the copula approach.

The max-stable model is a statistical framework used to model the spatial and temporal dependencies of extreme values. It is based on the theory of extreme value distributions and is designed to describe the behaviour of the maximum values of a dataset across different locations and times. The model assumes that the extreme values follow a max-stable process, which is characterized by its stability condition under the process. Max-stable model is suitable for analyzing and predicting the occurrence of rare and extreme events, such as floods, heatwaves, and other natural disasters (Huser & Davison, 2014). There are three types of max-stable models, namely the Smith, Schlather, and Brown-Resnick models (Smith, 1990). Gaume et al., (2013) demonstrate the expected exceedance probability of a given threshold at different locations using max-stable model. The Smith model focuses on capturing spatial dependence through a spatial process, while the Schlather model uses a random field with a stationary process with a specific correlation function. The Brown-Resnick model, based on Gaussian processes, provides more flexibility in modeling complex dependencies in extreme events.

Unlike max-stable models, copulas are not limited to extreme value analysis only. Copula is able to model the dependencies of variables across the entire distribution including both the centre and tails. Copulas work by separating the marginal behaviour of each variable from the dependence structure, allowing for the construction of a joint distribution that can capture complex, non-linear dependencies (Ribatet et al., 2013). Max-stable models offer distinct advantages over copula models to capture asymptotic dependence, providing accurate representations of joint distributions for maximum values. Max-stable models offer theoretical consistency and robustness, making them reliable for extrapolating beyond observed data ranges.

In spatial extreme modelling, the selection of an appropriate marginal distribution is a basic step in max-stable models. Marginal distributions define the statistical properties of extreme values at single location to ensure that the model accurately captures the behaviour of these extremes. Spatial dependence between the extreme values exposing the occurrence of an extreme event at one location associated with extremes at other locations. Combining the marginal distributions with the max-stable process describes both individual characteristics and the dependence structure of spatial extremes (Tawn et al., 2018). Spatial dependence identifies the groups or clusters of locations with similar characteristics by using hierarchical cluster analysis.

In spatial extremes, the marginal distributions are often modelled using the Generalized Extreme Value (GEV) distribution and the Generalized Pareto Distribution (GPD). These univariate models provide the necessary statistical foundation to accurately capture the behaviour of extreme values at individual locations before integrating them into a comprehensive spatial framework (Coles, 2001; Phalitnonkiat et al., 2016; Xu et al., 2023). GEV is the limiting distribution for block maxima (BM) series, which corresponds to the maximum value in each block of a fixed size within a dataset. On the other hand, the GPD can be used to model the distribution of the data exceedances above a chosen threshold (Ferreira, 2010).

Finding an appropriate marginal distribution is the major challenges when dealing with non-stationarity in the data. Environmental conditions and underlying processes can vary significantly across different locations, making it difficult to accurately model and predict extremes. Additionally, limited data availability and quality at certain locations can affect the estimation of marginal distributions. Non-stationarity can be linear over time or have cyclic variations which occur in a repeating pattern due to periodic influences. Cyclic non-stationarity requires the model to account for periodic fluctuations that can significantly impact extreme event occurrences. These cyclic variations can be driven by natural phenomena like monsoons, seasonal temperature changes, or human activities that follow a regular pattern. Combining cyclic non-stationarity with spatial extreme modelling using max-stable model allows for a more comprehensive understanding of extreme events by capturing both temporal and spatial dependencies.

The combination of cyclic pattern with max-stable model provides an appropriate model to managing the non-stationary extreme ground-level ozone (GLO) for several stations in peninsular Malaysia. This will capture both temporal and spatial dependencies to improve prediction and risk management. The combination also allows for the modelling of complex dependence structures between different locations to enhances the model's ability to fit real-world scenarios effectively (Tawn et al., 2018). Hazarika et al. (2019) found that GLO extremes are higher in the monsoon and pre-monsoon seasons and lower in winter. GLO patterns in peninsular Malaysia are influenced by the Southwest monsoon (May-August) and the Northeast monsoon (November-February) (Angelina et al., 2020; Awang, Elbayoumi, et al., 2016). The non-stationary GEV models, which can be linear, quadratic, cyclic, or another suitable function to represent the data trends, are use as the marginal distribution in max-stable model.

The high level of GLO concentration has serious adverse effects on health problems such as asthma, shortness of breath, chest pain, throat irritation, and other respiratory conditions (Liu et al., 2020; Zulkifli et al., 2022). It also causes damage to crop, forests, and other vegetation, reducing yields and decreasing biodiversity. In addition, it contributes to climate change by trapping heat in the atmosphere and worsening global warming (Fleming et al., 2018). The proper monitoring of high-level GLO is essential to satisfy the tenth item from socio-economic driven, which is environment and biodiversity, as stated by 10-10 Malaysian Science, Technology, Innovation and Economic (MySTIE) framework. It aims to preserve and conserve the natural environment and biodiversity of Malaysia, which are vital in harnessing its value for sustainable development guided by Ministry of Science, Technology and Innovation (Malaysia, 2020).

1.2 Problem Statement

The most common statistical studies performed in the context of GLO concentration prediction are using regression analysis. This regression analysis examines the relationship between GLO concentration and one or more independent variables. However, these methods are not sufficient to represent the rare events that refer to an occurrence or outcome that happens infrequently or has a low probability of happening. This is owing to the fact that it may produce biased estimates, not accurately estimate the probabilities or risks, and may not be sensitive enough to detect sensitive patterns or changes when applied to rare events data (Thompson et al., 2001). Therefore, the conventional standard model is unsuitable for developing precise models or generating reliable predictions. Hence, spatial extreme provides a framework for analysing and modelling the behaviour of rare events considering the extreme data pattern and the characteristics of several locations.

The spatial extreme GLO model required an appropriate GEV marginal distribution depending on the data structure for each station. If there is no apparent trend in the extreme data series, the use of a stationary model will be the best or most accurate. However, the problem arises when the extreme data series have strong seasonal variation. Hence, the stationary model is no longer suitable. Thus, choosing the best marginal distribution depends on the data trend, which will give the best-fitted spatial extreme model for predicting extreme events occurrences.

On the other hand, Rajab et al. (2013) discovered seasonal variation in GLO data at Subang, Kota Bahru, Penang, Johor, and Kuantan stations, which were influenced by the interchange of two monsoons, the Northeast and Southeast Monsoons. It is supported by Tan et al. (2014), who discovered that atmospheric GLO concentrations varied seasonally for the same stations. Hence, this study considers the seasonal variation in the GEV model that will be used as the marginal distribution for the spatial extreme model. Other choices of the marginal distribution, such as linear, quadratic, log-linear, and exponential trends, are commonly used, such as by Hasan et al. (2014) and Panagoulia et al. (2014).

The plot of extreme data extracted using the BM method demonstrates a seasonal trend, which inspired the idea of constructing a new marginal distribution using a non-stationary GEV model. Generally, the GLO concentration level is at a low level during the northeast monsoon (November to February), while during the southwest monsoon (May to August), the whole of Peninsular Malaysia experiences a dry and hot season. Hence, GLO production is very favourable under these conditions. Therefore, this research introduced the newly constructed model by this condition as a cyclic GEV model. The idea is supported by Coles (2001), who has mentioned that strong seasonal effects in the Wooster temperature range fit a single-year cycle period non-stationary model using trigonometric function. Eastoe and Tawn (2009) have used a model with a cyclic pattern in the analysis of surface GLO based on several air monitoring stations in the United Kingdom using the GPD model. The cyclic model has also been applied by Syafrina et al. (2019) to model the seasonal trend of the rainfall model. However, the existing cyclic is limited to univariate extreme modelling.

In this research, the non-stationary model for univariate cases is extended to the spatial extreme model where the cyclic pattern in locations parameter complements the GEV distribution as a new marginal distribution in the max-stable process. The max-stable process is the standard dependency model for spatial extreme observed at two or more different locations (Davis et al., 2013). To date, there has been a lack of research utilising marginal distribution incorporating cyclic trends when applying the max-stable model to GLO data. Therefore, a new procedure for the non-stationary max-stable process is developed by considering the marginal distribution, which depends on the cyclic trend to represent each air monitoring station in peninsular Malaysia.

The validation process is important to assess the developed model for making an accurate prediction. Therefore, the synthetic data are generated to closely resemble the key characteristics of the real data, including distributions and patterns, to validate the developed model and, hence, make a proper prediction of future extreme cases based on the return value of a specific return period. The return level is used to predict once, on average, the extreme event amount within a specified period of the return period. However, a clearer presentation of return level results will enhance the researcher's understanding of GLO concentration level based on the location category. Consequently, the return level prediction for all stations can be conveniently visualised in return level mapping.

1.3 Research Objective

The main objective of this study is,