STATISTICAL ANALYSIS OF WIND AND RAINFALL WITH FUNCTIONAL DATA ANALYSIS TECHNIQUE

WAN NORLIYANA BT WAN ISMAIL

A thesis submitted in fulfillment of the
requirement for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

JANUARY 2014

To my beloved mother and father

ACKNOWLEDGEMENT

By the name of Allah, this thesis is finally completed. It is an honor to express my fullest gratitude to the people who have been involved in finishing this thesis. In preparing this research, I have been discussing with my supervisor on the solution to overcome the problem in this thesis. I wish to express my fullest gratitude to my thesis supervisor, Dr. Shariffah Suhaila Syed Jamaludin for the guidance and encouragement that make this thesis finally completed. Without her support, this thesis might be not as appropriate as shown here.

Not to be forgotten, to my beloved parents, Wan Ismail Wan Ali and Seri Sulong as well as my family, I felt so grateful having all of you in my life, thank you for your profound understanding, helping and endless support me for the whole time I spend in this university.

Last but not least, highly thanks also to my friends for supporting me from the beginning of preparing this thesis until it have been completed. There are many motivations and tips that I get from those who support me and I really appreciate it.

ABSTRACT

The speed of wind and rainfall throughout Peninsular Malaysia varies from one region to another, depending on the direction of winds and rainfall that occur because of strong wind and the monsoons. Our data consist of daily mean wind and rainfall data from ten stations and covering 25 years period from 1985 to 2009. The purpose of this study is to convert the wind and rainfall data into a smooth curve by using Functional Data Analysis (FDA) method. The Fourier basis is used in this study in which the wind and rainfall data indicate periodic pattern for ten stations. In this method, to avoid such overfitting data, roughness penalty is added to the least square when constructing functional data object from the observed data. By using small basis functions, the difference is very small between with and without roughness penalty, showing that it is safer to smooth only when required. Meanwhile, with large basis functions and the difference of sum of square is very large, roughness penalty should be added in order to obtain optimal fit data. The graphs with contour plot show the relationship between wind and rainfall data, which illustrate the correlation and cross-correlations functions. Functional linear model also presents the relationship that may exist between wind (functional data) and rainfall (scalar response). Square multiple correlations give strong positive linear correlation, which concludes that the rainfall is influenced by the wind speed.

ABSTRAK

Kelajuan angin dan hujan di seluruh Semenanjung Malaysia berbeza mengikut kedudukan geografi setiap negeri, iaitu bergantung kepada arah angin dan hujan yang berlaku pada musim tengkujuh. Data ini terdiri daripada purata kelajuan angin dan hujan mengikut harian selama 25 tahun dari tahun 1985 hingga tahun 2009 dan juga diperolehi daripada sepuluh stesen. Tujuan kajian ini dijalankan adalah untuk menukarkan data angin dan hujan kepada lengkungan licin dengan menggunakan kaedah Fungsi Analisis Data (FDA). Fourier asas digunakan dalam kajian ini, di mana data angin dan hujan menunjukkan corak berkala selama sepuluh stesen. Dalam kaedah ini, untuk mengelakkan data 'overfitting', penalti kekasaran ditambah kepada kuasa dua terkecil apabila membina fungsi objek daripada data yang diperhatikan. Dengan menggunakan fungsi asas yang kecil, perbezaan adalah kecil di antara penggunaan dan tanpa penggunaan penalti kekasaran, menunjukkan bahawa ia adalah lebih selamat untuk melicinkan lengkungan hanya apabila diperlukan. Sementara itu, dengan menggunakan fungsi asas yang besar dan perbezaan jumlah kuasa dua adalah sangat besar, penalti kekasaran perlu ditambah untuk mendapatkan data yang optimum. Kontur plot menunjukkan hubungan antara data angin dan hujan, menggambarkan fungsi korelasi dan korelasi silang. Fungsi model linear juga menunjukkan hubungan yang wujud antara angin (fungsi data) dan hujan (tindak balas skalar). Korelasi 'square multiple' memberi korelasi linear positif yang kuat, menyimpulkan bahawa hujan adalah dipengaruhi oleh kelajuan angin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

APPENDIX **TITLE**

(6.1) Temerloh Station for Wind Data

(6.2) Cameron Highlands Station for Rainfall Data

(7) Determine the Descriptions of Functional Data

(7.1) Functional Mean and Variance with Correlation Functions for Wind Data

(7.2) Functional Mean and Variance with Correlation Functions for Rainfall Data

(7.3) Cross-Covariance and Cross-Correlation Functions

(8) Functional Linear Models with Scalar Responses

(8.1) Functional Linear Regression and Correlation

CHAPTER 1

INTRODUCTION

1.1 Introduction

Functional data analysis is a branch of statistics. Functional Data Analysis (FDA) develops fast in statistics area with the aim of estimating a set of related functions or curves rather than focusing on a single entity, like estimating a point, as in classical statistics. Some methods are simply the extension of existing techniques in conventional statistics while others need more than exchanging the summation, which is used in discrete observation, to an integration, which is a continuum. Data sets in FDA are in the form of curves, surfaces or anything else varying over a continuum compare to multivariate statistics, where data are considered as vectors (finite sets of values).

Recently, the application of statistical modeling to medicine, biomedicine, public health, biological sciences, biomechanics, environmental science geology, psychology, and economics has largely and increasingly being driven by the need for better data to assist in government policy making and planning processes for any services required. As a matter of fact, discrete data points collected over a continuum can be looked upon as a function or curve that is presumed to be a reasonably smooth to those discrete smoothed points. This continuum is not necessarily a physical time point, but rather attributes such as age, spatial location, seasons, or temperature. Importantly, such models will only be useful in the long term if they are accurate, based on good quality data, and generated through the application of robust appropriate statistical methods.

1.2 Research Background

In conventional statistical practice, observation is usually a number or a vector. But in many real-life situations, observed values are continuous curves, vectors of curves, images, or vectors of images. The characteristics about functional data are the possibility of using information on the rates of change or derivatives of the curves. We use the slopes, curvatures, and other characteristics made available because these curves are intrinsically smooth, and can use this information in many useful ways. Not only that, FDA gives a strong link with the multivariate statistical paradigm and for the regularization. The stable link with multivariate statistics comes from the fact that methods, such as principal component analysis, multivariate linear modeling, functional linear model, functional ANOVA, canonical correlation analysis, etc. can be applied within the functional data analysis framework.

In this study, pattern characteristics of daily average wind-rainfall in Peninsular Malaysia for the period 1985 to 2009 are investigated for ten stations. In FDA, a set of observations is transformed into a functional object, and statistical analysis is then performed on this continuous function, rather than on the original discrete data points. These make it possible to extract information from the temporal process as a whole, instead of merely point-by-point. In order to describe the characteristics of wind-rainfall data, FDA can represent the data in the form of smooth functions or curves that give meaningful information.

The wind in Malaysia is generally light and varies; however, some uniform periodic changes in the wind flow patterns can also influence rainfall distribution. A strong wind is expected to bring heavy rainfall at the location. We can say that when there is strong wind, this can influence the rainfall. In other words, the heavy rain can happen because of the wind speed is higher.

1.2 Problem Statement

In this analysis, two climate variables that often change throughout the year are daily mean wind speed and rainfall. The seasonal wind flow patterns paired with the local topographic features determine the rainfall distribution patterns over the area. By using FDA, the discrete data are transformed into curves. If the discrete values are assumed to be errorless, the process is an interpolation. But if they have some observational errors that need to be removed, the transformation from discrete data to the functions may involve smoothing. We use FDA to interpret wind-rainfall data with certain smooth functions that are assumed to underlie and to generate the observations. FDA methods are not necessarily based on the assumption for a single subject but values observed at different times. For calculating the coefficients of measurement error and to compute the coefficients to obtain an optimal fit to wind-rainfall data, powerful basis expansion is used. However to avoid overfitting the data, a penalty on the “roughness” of the function is imposed. To examine the relationship between wind and rainfall, having a functional linear model is suitable for continuous time process.

1.3 Research Objectives

The objectives of the study are:

- To determine the number of basis functions for each weather station.
- To summarize the pattern of wind-rainfall data using the functional descriptive statistics.
- To establish the relationship between wind and rainfall data using the concept of functional linear model.

1.4 Scope of the Study

This study focuses on profiling ten stations for wind and rainfall data throughout Peninsular Malaysia. Kuala Krai, Batu Embun, Temerloh, Muadzam Shah, Mersing, Senai, Bayan Lepas, Cameron Highlands, Ipoh and Subang are among the ten selected stations in this study. In this analysis, we used daily wind-rainfall data for the period of 25 years which from the year 1985 until 2009. Data were obtained from Malaysian Meteorological Service (MMS).

1.5 Significance of Study

The result of this study will give the advantage which is focused on how the FDA techniques can be used for wind-rainfall data analysis. This study can determine the patterns of each region of wind-rainfall data in Peninsular Malaysia. The FDA techniques can represent both climate data in the form of smooth curves for each region located. Functional data analysis also will produce important basis functions for this research. In statistics, this research will broaden the application of functional data analysis in our daily life.

REFERENCES

Clarkson, D., Fraley, C., Gu, C.C., Ramsay, J.O., 2005. *S+ Functional Data Analysis*. Springer, US of America.

Cuevas, A., Febrero, M., Fraiman, R., 2003. *An Anova Test for Functional Data*. Comput. Stat. Data Anal. 47, 111-122.

Febrero, M., Galeano, P., Manteiga, W.G., 2007. *A Functional Analysis of NOx Levels: Location and Scale Estimation and Outlier Detection.* Comput. Stat. 22, 411-427.

Ferraty, F., Vieu, P., Viguier-Pla, S., 2006. *Factor-based Comparison of Groups of Curve.* Comput. Stat. Data Anal. 51, 4903-4910.

Froslie, K.F., 2012. *Shape Information from Glucose Curves: Functional Data Analysis Compared with Traditional Summary Measures.* BMC. Med. Rese. Method. 13 (6), 1471-2288.

Gao, H.O., Niemeier, D.A., 2008. *Using Functional Data Analysis of Diurnal Ozone and NOx Cycles to Inform Transportation Emissions Control*. Trans. Research. Part D. 13, 221-238.

Guo, M., Zhou, L., Huang, J.Z., Hardle, W.K., 2012. *Functional Data Analysis of Generalized Quantile Regressions*. SBP. Discussion Paper, 001.

Hyndman, R.J., Booth, H., 2008. *Stochastic Population Forecasts using Functional Data Models for Mortality, Fertility and Migration*. Int. J. Forecasting. 24, 323-342.

Laukaitis, A., Rackauskas, A., 2004. *Functional Data Analysis for Clients Segmentation Tasks*. Euro. J. Operational Research. 163, 210-216.

Manteiga, W.G., Vieu, P., 2007. *Statistical for Functional Data*. Comput. Stat. Data Anal. 51, 4788-4792.

Muller, H.G., Stadtmuller, U., 2005. *Generalized Functional Linear Model*. Annals Stat. 33(2), 774-805.

Muniz, C.D., Nieto, P.J.G., Fernandez, J.R.A., Torres, J.M., Taboada, J., 2012. *Detection of Outliers in Water Quality Monitoring Samples using Functional Data Analysis in San Esteban Estuary*. SN. Total Envi. 439, 54-61.

Newell, J., McMillan, K., Grant, S., McCabe, G., 2004. *Using Functional Data Analysis to Summarise and Interpret Lactate Curves.* Comput. Bio. Medi. 36, 262-275.

Nikitovic, V., 2011. *Functional Data analysis in Forecasting Serbian Fertility*. Institute of Social Sciences. 2, 73-89.

Ramsay, J.O., Ramsey, J.B., 2002. *Functional Data Analysis of the Dynamics of the Monthly Index of Nondurable Goods Production*. J. Econometrics. 107, 327-344.

Ramsay, J.O., Hooker, G., Graves, S., 2009. *Functional Data Analysis with R and Matlab*. Springer, New York.

Ramsay, J.O., Silverman, B.W., 2005. *Functional Data Analysis,* second ed.Springer, New York.

Ratcliffe, S.J., Leader, L.R., Heller, G.Z., 2002. *Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression.* John Wiley & Sons, Ltd. 21, 1103-1114.

Song, J.J., Deng, W., Lee, H.J., Kwon, D., 2008. *Optimal Classification for Time-Course Gene Expression Data using Functional Data Analysis*. Comput. Bio. Chemis. 32, 426-432.

Suhaila, J., Jemain, A.A., Hamdan, M.F., Zin, W.Z.W., 2011. *Comparing Rainfall Pattern between regions in Peninsular Malaysia via a Functional Data Analysis*. J. Hydro. 411, 197-206.

Tian, T.S., 2010. *Functional Data Analysis in Brain Imaging Studies.* Front Psychol. 1, 35.

Torres, J.M., Nieto, P.J.G., Alejano, L., Reyes, A.N., 2010. *Detection of Outliers in Gas Emissions from Urban Areas using Functional Data Analysis.* J. Hazard. Material. 186, 144-149.