TIME SERIES SUPPORT VECTOR REGRESSION MODELS WITH MISSING DATA TREATMENTS FOR WATER LEVEL PREDICTION

NORAINI BINTI IBRAHIM

A Thesis submitted in fulfillment of the
requirements for the award for the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MAY 2014

This thesis is dedicated to:

My husband Mohamad Faizul bin Mat Ali

My father Ibrahim bin Awang and mother Darisah binti Yatim

My brothers and sisters

Whose love and support

For all my friends, especially UA2 friends

Who always accompany me in my happiness and sadness

Without all of you, it is difficult for me to endure all of these adversities

ACKNOWLEDGEMNT

In the name of Allah the Most Gracious and the Most Merciful, I would like to express my sincere gratitude to my thesis supervisor Dr. Antoni Wibowo for the continuous support of my master study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all completing this work.

Besides my advisor, my sincere thanks also goes to UTM and RMC for supporting this study, researchers and academicians for their contribution towards my research including the understanding and thoughts. I thank my fellow labmates in OBI group for the stimulating discussions, for the sleepless nights we were working together and for all the fun we had in the last two years.

Last but not least, I would like to thank my family especially my parents for supporting me spiritually throughout my life.

ABSTRACT

Rise in water level is an important issue because it can be used as an indicator for flood alert. The water level of a river is dependent upon variables such as the month, volume of rainfall, temperature, relative humidity and surface wind. The main purpose of this research is to find a suitable method to predict the water level of Galas River in Kelantan to anticipate flood. In this research, secondary data on water level of Galas River was collected from the Department of Irrigation and Drainage Malaysia and Malaysian Meteorological Department. Some of the data were missing in certain months, thus these data were replaced by the use of means and linear regression based on the related months in other years as treatments of these missing data. Both these treatments were included in the methods to analyse data. Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), Support Vector Regression (SVR) and SVR-based time series regression were used to analyse the data. Using the MLR analysis, multicollinearity was detected and addressed by applying PLSR. However, this technique which is a linear based model may not be appropriate in a nonlinear case such as the Galas River case. In this study, a nonlinear method, SVR, was applied. Besides that, SVR-based time series regression was proposed to cater for the time-based water level data, and to overcome the issue of linearity and multicollinearity. The result shows that linear regression is a better data treatment in SVR and SVR-based time series regressions. In addition, using Gaussian kernel, the results showed that these regressions have lower mean squared error of cross-validation as compared to MLR and PLSR. The major finding from this study is that both SVR and SVR-based time series regression used to anticipate flood by predicting the water level is significantly better than MLR and PLSR.

ABSTRAK

Kenaikan paras air adalah satu isu penting kerana ia boleh digunakan sebagai petunjuk untuk amaran banjir. Paras air sungai adalah bergantung kepada pembolehubah seperti bulan, jumlah hujan, suhu, kelembapan relative dan permukaan angin. Tujuan utama kajian ini adalah untuk mencari kaedah yang sesuai untuk meramalkan paras air di Sungai Galas, Kelantan bagi menjangka banjir. Dalam kajian ini, data sekunder untuk paras air di Sungai Galas telah dikumpul dari Jabatan Pengairan dan Saliran Malaysia dan Jabatan Meteorologi Malaysia. Beberapa data telah hilang pada bulan-bulan tertentu, oleh itu data ini telah digantikan dengan penggunaan purata dan regresi linear berdasarkan bulan-bulan yang berkaitan dalam tahun-tahun yang lain sebagai rawatan kepada data yang hilang ini. Kedua-dua rawatan data ini telah dimasukkan ke dalam kaedah untuk menganalisis data.Regresi Linear Berganda (MLR), Regresi Kuasa Dua Terkecil Separa (PLSR), Regresi Vektor Sokongan (SVR) dan SVR berasaskan regresi siri masa telah digunakan untuk menganalisis data. Dengan menggunakan analisis MLR tersebut, multikolinearan dikesan dan telah ditangani dengan menggunakan PLSR.Walau bagaimanapun, teknik ini merupakan model berasaskan linear yang mungkin tidak sesuai dalam kes tak linear seperti kesdi Sungai Galas.Dalam kajian ini, kaedah tidak linear, SVR, telah digunakan.Selain itu, SVR berasaskan regresi siri masa telah dicadangkan untuk mengatasi data paras air berasaskan masa, dan untuk mengatasi isu kelinearan dan multikolinearan. Hasil menunjukkan bahawa regresi linear adalah rawatan data yang lebih baik dalam SVR dan juga SVR berasaskan regresi siri masa. Selain itu, dengan menggunakan kernel Gaussian, keputusan menunjukkan bahawa regresi-regresi ini mempunyai puratar alat kuasa dua yang lebih rendah dengan menggunakan pengesahan-silang berbanding MLR dan PLSR. Penemuan utama daripada kajian ini ialah bahawa kedua-dua SVR dan SVR berasaskan regresi siri masa yang digunakan untuk menjangka banjir dengan meramalkan paras air adalah jauh lebih baik daripada MLR dan PLSR.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter is divided into five main sections which will wholly summarize the problems and motivations underlying the research. The first section describes the background of problems. The next two sections explain the problem statement and the objectives of the study, followed by the scope, significance of the study and thesis organisation.

1.2 Background of Problem

Floods are common phenomena which can be defined as the presence of excess water in a place beyond its normal limits. Floods are often cited as being the most lethal of all natural disasters (French *et al*., 1989; Alexander, 1993; Jonkman *et al.*, 2005). The flooding of Malaysian rivers is mainly due to the high amount of rainfall in river basins because of the climate of the country which is greatly influenced by the monsoon winds. Peninsular Malaysia is located in South East Asia and in the equatorial latitudes. Within these latitudes, it receives more than 2,500 mm of rain annually and the average temperature is 27 degrees Celsius. Peninsular Malaysia faces two monsoon winds seasons which are the Southwest Monsoon, from late May to September, and the

Northeast Monsoon, from November to March. The Northeast Monsoon brings in more rainfall compared to the Southwest Monsoon. It is noted that the rainfall over different states in Peninsular Malaysia vary from each other.

The worst flood in Malaysia was recorded in 1926 which has been described as having caused the most extensive damage to the natural environment. Subsequent major floods were recorded in 1931, 1947, 1954, 1957, 1967, and 1971. Kelantan, Terengganu, Pahang, Johor, and Kedah are among the states in Malaysia that suffer the most from floods during monsoon season. Kelantan is a state in the east coast of Peninsular Malaysia that has never missed a flooding event, which occurs between October and March every year during the northeast monsoon period. The citizens in Kelantan always suffer from the floods since the water overflows to the areas adjoining the rivers, lakes, or dams.

A flood affects many of the engineering structures such as bridges, embankments, reservoirs, and significantly disrupts or interferes with human and societal activities. Kuala Krai is one of the districts in Kelantan that is always affected by floods. Factors causing floods in the Kuala Krai district of Kelantan are a combination of physical factors such as elevation and its close proximity to the sea, apart from the heavy rainfall experienced during monsoon period. The severe floods all over Kelantan result from heavy rainfall during the north east monsoon season especially in November and December.

Some of the flood characteristics have been listed by the Department of Irrigation and Drainage Malaysia (DID), which are the water level, area inundation, peak discharge, volume of flow, and duration (Gasim *et al*., 2007). Moreover, there are three categories of critical level stages of water level: alert, warning, and danger, which were also identified by the Department of Irrigation and Drainage Malaysia. The river stage is the variable that is considered when a flood warning is issued. Five factors have been identified to be the variables of this research and the factors that might lead to the rising of water level of Galas River: (1) Months from January to December for 11 years starting from 2001 to 2011, (2) Monthly mean of rainfall, (3) Monthly mean of temperature, (4)

Monthly mean of relative humidity, and (5) Monthly mean of wind speed. These predictor variables were collected from meteorological department and we want to find the most related predictor variables towards response variable which is water level that was collected from hydrological department.

This study attempts to find an appropriate model and dominant variables for predicting water level in Galas River. In order to obtain the model, the study used four different methods which are Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), and SVR for time series regression. It was discovered that the original hydrological data had missing values. Due to the occurrence of the missing values, the four methods could be inappropriate to be used directly for water level prediction; therefore, data cleaning is needed to be performed on the hydrological data beforehand. We performed two types of data cleaning which are type I and type II data cleaning using mean of the corresponding months and linear regression with single predictor.

We use MLR since it is simple regression, however, we should pay attention the presence of multicollinearity in the DID and MMD data. We can apply PLSR if the multicollinearity present, however, PLSR has more complicated procedures compared to MLR. It noticed that MLR and PLSR are linear methods, and we do not know that relation between water level and the five predictor variables are either linear or nonlinear. We apply SVR if the relation between those variables is nonlinear. Another advantage of SVR is SVR is not influenced by the present of multicollinearity. In many cases, however, time series based methods is more powerful compared to non-time series methods. Therefore, we consider to modify SVR model with DID and MMD data in the form of SVR-based time series regression.

1.2.1 Geology of Kelantan River

The Kelantan state consists of more than 25 rivers and seven major river basins, namely Galas, Kelantan, Golok, Semerak, Pengkalan Chepa, Pengkalan Datu, and Kemasin. Out of the seven, Kelantan River Basin is the largest in Kelantan (Rohasliney, 2010). It drains a catchment area of about 12,000 km^2 in north-east Malaysia, including part of the National Park, and flows northwards into the South China Sea (Rohasliney, 2010). The Kelantan River is about 248 km long and occupies more than 85 percent of the State of Kelantan. It separates into the Galas and Lebir Rivers near Kuala Krai, about 100 km from the river mouth which means that Kelantan River is the main river while Galas and Lebir Rivers are the tributary rivers. For this research, the focus is on forecasting the water level of Galas River in Kuala Krai district.

1.2.2 Geomorphology of Galas River

For this research, we focused on two main tributary rivers which are Galas and Lebir Rivers in Kuala Krai, Kelantan. **Table 1.1** presents the characteristics of the river and the main tributaries and **Figure 1.1** shows the location of the study area which is Galas River in Kuala Krai and labeled as Galas R. in the map.

Table 1.1: The Characteristics of the Main River and the Main Tributaries

No.	Name of river	Length [km] Catchment area [km²]	Highest peak [m] Lowest point [m]
1	Kelantan River	248 11,900	Mt. Korbu (2,183 m)
2	Galas River	178 7,770	Mt. Setong (1,422 m)
3	Lebir River	91 2,430	Cintawasa Hill (1,185 m)

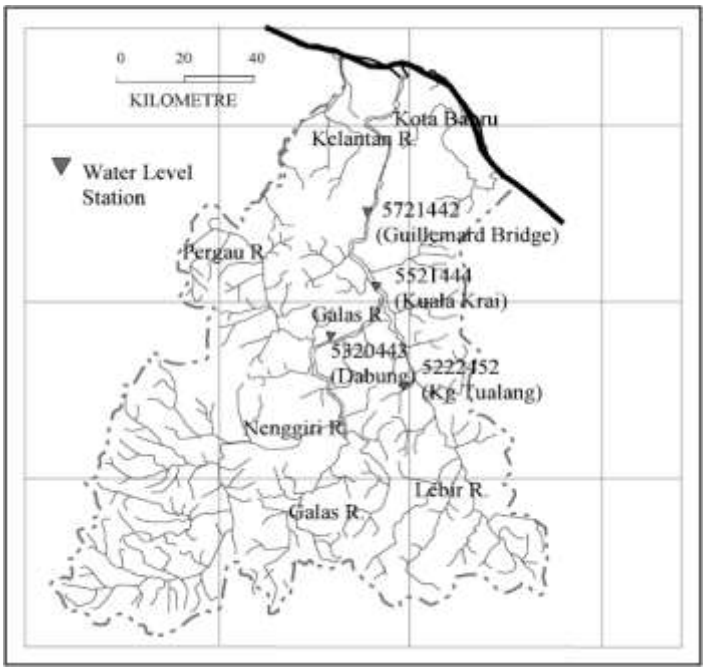

Figure 1.1: Location of Galas River

As mentioned before, five factors have been identified as related to the increasing of the water level of Galas River which leads to the rising of river stages in Galas River, Kelantan. To observe and monitor the water level in Kuala Krai area, eight hydrological stations were set up by the Water Resources Management and Hydrology Division.

Table 1.2 shows the categories of the water level for Galas River in Kuala Krai that were introduced by the Department of Irrigation and Drainage Malaysia. The four stages are normal level, alert level, warning level, and danger level.

Table 1.2: The Categories of the Water Level Stages for Galas River

Station	Station Name	River Basin	Normal Level (metre)	Alert Level (metre)	Warning Level (metre)	Danger Level (metre)
5320443	Galas River at Dabong	Kelantan River	28	32	35	38

1.3 Problem Statement

Mostly people will suffer losses due to flooding event caused by water level rising. An investigation is needed to overcome this problem or as an effort to manage it the future. For this research, the data were taken from the Water Resources Management and Hydrology Division (WRMHD) in Kuala Lumpur and Malaysian Meteorological Department (MMD) in Selangor.

However, it was found that the raw data consisted of missing values and the existing methods, such as MLR, PLSR and SVR might be inappropriate to be used to develop the prediction models using these data. Hence, it was necessary to perform data pre-processing to replace the missing values. There are two types of data pre-processing that have been executed, namely type I data pre-processing (based on the mean) and type II data pre-processing (based on linear regression with single predictor). The data have been analysed using linear and nonlinear methods which are MLR, PLSR, SVR, and SVR-based time series regression.

The MLR is the simplest model in regression analysis and widely used in many fields in engineering and social sciences. MLR can be used to predict the water level but it can be unsuitable when multicollinearity exists. Multicollinearity in DID and MMD data can bring harmful effects to the regression model when correlations are present among these predictors. The negative effects of multicollinearity can be overcome by applying PLSR which is often used when multicollinearity is present. PLSR has been paid an increasing attention nowadays as an important measure of each explanatory variable or predictor (Chong and Jun, 2005). However, it has limitations in its application since PLSR is a linear model.

The SVR and SVR-based time series regression can be used to overcome the issue of linearity and multicollinearity. However, the suitable parameters and lag are need to be determined in both methods. Furthermore, model selection is also needed to obtain the best prediction model and dominant variables among the four methods.

1.4 Objective of Study

The objectives of the study are:

i. To deal with the issue of missing data on rainfall and water level by performing data cleaning.
ii. To develop water level prediction models for Galas River using PLSR, SVR, and SVR for time series regression with considering the existence of multicollinearity.
iii. To select the appropriate dominant variables and prediction models for water level using *k-fold* cross-validation to obtain the best model.

1.5 Scope of Study

i. The study focused on developing the algorithm for water level prediction on Galas River using the methods of MLR, PLSR, SVR, and SVR-based time series regression.
ii. The data were collected from the Water Resources Management and Hydrology Division, Department of Irrigation and Drainage Malaysia (DID), and Malaysian Meteorological Department in Selangor (MMD).
iii. The data consist of six variables, namely month, rainfall, temperature, relative humidity, surface wind, and water level.
iv. The data covers the record on Galas River from 2001 to 2011.
v. The monthly average of rainfall and water level from 11 years are used in order to predict the water level of Galas River.
vi. Matlab software is utilised for most of the analysis in this research.

1.5 Significant of Study

For this research, we studied the appropriate method to predict the water level based on the existing methods of Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), and SVR-based time series regression. However, these methods cannot be directly used when there exists the missing values in the original data. Hence, the appropriate measures to overcome the issue of missing values is needed by performing data cleaning using mean of the corresponding months and linear regression with single predictor. The information from these linear and nonlinear methods is obtained for the research to help the people in Kuala Krai, Kelantan to be prepared for during the monsoon seasons in the future. Furthermore, this research would help the government to analyse the water level patterns and find ways to reduce floods in Kuala Krai. The development of the models using nonlinear methods will give higher accuracy in predicting the water level of the river.

1.6 Thesis Organization

The following chapters will present an approach to the development of the water level models. Chapter 2 contains the literature review of previous researches while the framework of the present study and the theoretical methods of MLR, PLSR, SVR, and SVR-based time series regression are discussed in Chapter 3. The preliminary results are described in Chapter 4 and finally, the conclusion and future works are imparted in Chapter 5.

REFERENCES

Abudu, S., King, J.P., P.E.& Pagano, T.C., 2010. Application of partial least squares regression in seasonal streamflow forecasting, *Journal of Hydrologic Engineering*, 8, pp.612-623.

Aladeniyi, O. B., Olowofeso, O.E. &Fasoranbaku, O.A., 2009.On new techniques for testing incomplete data using regression models, *The Pacific Journal of Science and Technology*, 10, pp.149-159.

Alexander, D., 1993. *Natural disasters*. UCL Press, London.

Alvisi, S., Mascellani, G., Franchini, M. &Bardossy, A., 2006. Water level forecasting through fuzzy logic and artificial neural network approaches. *Hydrology and Earth System Sciences*, 10, pp.1-17.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L. &Ridella, S., 2012. The 'K' in K-fold cross validation, ESANN 2012 proceedings, *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 25-27 April 2012.

Atkinson, A.C., 1982. Regression diagnostics, transformations and contructed variables, *Journal of the Royal Statistical Society, Series B*, 44, pp.1-36.

Batista, G. E. A. P. A. &Monard, M. C., 2010. An analysis of four missing data treatment methods for supervised learning, *Applied Artificial Intelligence*: An International Journal.

Belsley, D. A., Kuh, E. &Welsch, R. E., 2004. *Regression Diagnostics*: Identifying Influential Data and Sources of Collinearity.

Bergant, K. &Bogataj, L. K., 2005. N-PLS regression as empirical downscaling tool in climate change studies, *Theoritical and Applied Climatology*. 81, pp.11-23.

Bessaih, N., Bustami, R. &Saad, M., 2004.Water level estimation for Sarawak River.*Proceeding of Rivers '04*, USM , Penang.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C., 1984. *Classification and regression trees*, Wadsworth: Belmont, CA.

Bush, B. L. &Nachbar, Jr R. B.,1993. Sample-distance partial least squares : PLS optimized for many variables, with application to CoMFA, *J. Comput.-Aided Mol. Design*, 7, pp.587-619.

Bustami, R., Bessaih, N., Bong C., &Suhaili S., 2007. Artificial neural network for precipitation and water level predictions of Bedup River, *IAENG International Journal of Computer Sciences*, pp.34:2.

Cheesemen, P. & J. Stutz., 1996. Bayesian classification (AutoClass): theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthunsamy (Eds.), Advances in knowledge discovery and Data Mining, *AAAI Press/MIT Press*.

Chen, G. Y. &Xie, W. F., 2007.Pattern recognition with SVM and dual-tree complewavelets, *Image and Vision Computing*, 25, pp.960–966.

Chen, X., Li, Y., Harrison, R., & Zhang, Y.Q., 2008.Type-2 fuzzy logic-based classifier fusion for support vector machines.*Applied Soft Computing*, 8, pp.1222–1231.

Chong, I. K. & Jun, C. H., 2005. Performance of some variable selection methods when multicollinearity is present, *Chemometricsand Intelligent Laboratory Systems*, 78, pp.103-112.

Coulibaly, P., Anctil, F. &Bobee, B., 2001. Multivariate reservoirinflow forecasting using temporal neural networks, *J. Hydrologic Eng*, 6, pp.367–376.

Davison, A. C. &Hinkley, D. V., 1997. *Bootstrap methods and their application*: Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Erguven, M., 2012.Comparison of the efficiency of principal component analysis and multiple linear regression to determine students' academic achievement, *IEEE*, pp.1-5.

Fletcher, R., 1987. *Practical methods of optimization*.Second ed.Wiley, New York.

Franchini, M. &Lamberti, P., 1994. A flood routing Muskingum type simulation and forecasting model based on level data alone. *Water Resources Research*,30, pp.2183–2196.

French, J.G. & Holt, K.W., 1989. '*Floods*', In M.B. Gregg (ed.) The public health consequences of disasters, US Department of Health and Human Services, Public Health Service, CDC, Atlanta, GA, pp. 69–78.

Gasim, M. B., Adam, J. H., Toriman, M. E. H., Rahim S. A. &Juahir, H. H., 2007. Coastal flood phenomenon in Terengganu, Malaysia: special reference to Dungun, *Research Journal of Environmental Sciences*, 1(3), pp. 102-109.

Geisser, S., 1975.The predictive sample reuse method with applications, *J. Am. Stat. Asso.*, 70, pp.320-328.

Glen, W. G., Dun III W. J. & Scott D. R., 1989. Principal component analysis and partial least squares regression, *Tetrahedron Comput. Methodol*, 2, pp.349-376.

Glen, W. G., Dun III W. J., Sarker M. & Scott D. R., 1989. UNIPALS: Software for principle components analysis and partial least squares regression, *Tetrahedron Comput. Methodol*, 2, pp.377-396.

Golberg, M. A. & Cho, H. A., 2004.Introduction to Regression Analysis, Wit Press, UK.

Goyal, M. K. &Ojha, C. S. P., 2010.Application of PLS-regression tool for Pichola Lake Basin in India, *International Journal of Geosciences*, 1, pp.51-57.

Grzymala-Busse, J., Pawlak, Z., R. Slowinski, and W. Ziarko, 1999.*Rough sets. Communications of the ACM*, pp.11: 38.

Gunn, S. R., 1998. *Support vector machines for classification and regression,Technical Report*, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science.

Hanke, J. E. &Wichern, D., 2009. Business Forecasting, Pearson International Edition.

Hassanien, A.E., 2003. Classification and feature selection of breast cancer data based on decision tree algorithm, *Intl. J. Studies in Information and Control J.*, 1, pp.33-39.

Helland, I. S., 1988. On the structure of partial least squares regression, Communications in Statistics Elements of Simulation and Computation, 17, pp.581–607.

Hulland, J.S., 1999. Use of partial least squares (PLS) in strategic management research: A review of four recent studies, *Strategic Management Journal*, 20, pp.195–204.

Jong, S. D., 1993. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*,Elsevier Science Publisher B.V., Amsterdam, pp.251-263.

Jonkman, S. N. &Kelman, I., 2005. An analysis of the causes and circumstances of flood disaster deaths, *Disasters*, 29, pp.75−97.

Joshi, H., Kulkarni H. & Deshpande S., 2012.Multicollinearity Diagnostics in Statistical Modeling & Remedies to deal with it using SAS, *PhuSE 2012*.

Khan, M. S. &Coulibaly, P., 2006.Application of support vector machine in lake water level prediction, *Journal of Hydrologic Engineering* , 3(199).

Kim, K. J., 2003. Financial time series forecasting using support vector machines. *Neurocomputig*, 55, pp.307-319.

Kim, K.,Lee J. M. & Lee I. B., 2005. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems*, 79, pp.22-30.

Larson S., 1931. The shrinkage of the coefficient of multiple correlation, *J. Educat. Phychol*., 22, 45-55.

Legates, McCabe, D.R., Jr. & G.J., 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, *Water Resour.Res.* 35, pp.233–241.

Li, B., Hu, J. &Hirasawa, K., 2008.Financial time series prediction using a support vector regression network, *International Joint Conference on Neural Networks.*

Liang, Y., Xu, Q. S., Li, H. D.& Cao D. S., 2011. Support Vector Machines and Their Application in Chemistry and Biotechnology, *Taylor & Francis Group, LLC*.

Lindgren, F. &Rannar, S., 1998. Alternative Partial Least-Squares (PLS) Algorithms, *Perspective in Drug Discovery and Design*, pp.105-113.

Lin, G. F., Chou Y. C. & Wu, M. C., 2013.Typhoon flood forecasting using integrated two-stage support vector machine approach, *Journal of Hydrology*, 486, pp. 334-342.

Liong, S. Y. &Sivapragasam, C., 2002. Flood stage forecasting with support vector machines, *Journal of the American Water Resources Association,* 38, pp.173-196.

Little, R.J.A. & Rubin, D.B., 1987.Statistical Analysis with Missing Data, *J. Wiley & Sons*: New York, NY.

Lobaugh, N.J., West R. & McIntosh A.R., 2001.Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares, *Psychophysiology*, 38, pp.517–530.

Longley, J.W., 1967. An appraisal of least squares programs for the electronic computer from the point view of the user, *Journal of the American Statistical Association*, 62, pp.819-841.

Maatta, S., 2011. Predicting groundwater levels using linear regression and neural networks. CS 229 Final Project.

Mahabir, C., Hicks, F. E., Robichaud, C. &Fayek, A. R., 2006. Forecasting breakup water levels at Fort McMurray, Alberta, using multiple linear regression, *Can. J. Civ. Eng*, 33, pp.1227-1238.

Mardia, K.V., Kent, J.T. &Bibby, J.M., 1997. Multivariate analysis, *Academic Press*.

Martens, M. & Martens, H., 1986. Partial least squares regression, In J.R. Piggott, editor, *Statistical Procedures in Food Research*, Elsevier Applied Science, London, pp.293–359,

Mevik, B. H. &Cederkvist, H. R., 2004. Mean squared error of prediction (MSEP) estimates for principle component regression (PCR) and partial least squares regression (PLSR), *Journal of Chemometrics*, 18, pp.422-429.

Minsk, M. L. &Papert, S.A., 2009.Perceptrons. Cambridge, MA: *MIT Press*.

Montgomery, D. C., Peck, E. A. & Vining, G. G., 2001. *Introduction to linear regression analysis*, 3rd edition, Wiley, New York.

Morozov, V.A., 1984. Methods for solving incorrectly posed problems, *Springer*.

Nguyen, D.V. &Rocke, D.M., 2002. Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, 18, pp.39–50.

Nilsson, J., Jong, S. D. &Smilde A.K., 1997.Multiway Calibration in 3D QSAR, *Journal of Chemometrics*, 11, pp.511–524.

Pawlak, Z., 1982. Rough Sets, *Intl J. Computer and Information Sci*., 11, pp.341-356.

Pawlak, Z., 1991. Rough Sets-Theoretical Aspect of Reasoning About Data, *Kluwer Academic Publishers*.

Pawlak, Z., Grzymala-Busse, J., Slowinski, R. &Ziarko, W., 1995.Rough sets.Communications of the ACM, 11, pp.89-95.

Peng, C. Y. J., Michael, H., Liou, S. M. &Ehman, L. H., 2003. Advances in missing data methods and implications for educational research, Advance in Missing Data 1.

Quinlan, J.R., 1989. Unknown attribute values in induction. Proc. *Sixth Intl. Workshop on Machine Learning*, pp: 164-168.

Raghunatha, T.E. 2004. "What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data", *Annual Review of Public Health Journal*, 25, pp.99-117.

Refaeilzadeh, P., Tang, L. and Liu, H. (2008), *Cross-validation*, Arizona State University.

Rohasliney, H., 2010. Status of river fisheries in Kelantan,Peninsular Malaysia, Malaysia, *World Academy of Science, Engineering and Technology*, 65.

Rosipal, R. & Kramer, N., 2006.Overview and recent advances in partial least squares, *LNCS*, 3940, pp. 34–51.

Rosipal, R., 2003. Kernel partial least squares for nonlinear regression and discrimination, *Neural Network World*, 13, pp.291–300.

Rosipal, R., Trejo, L.J. & Matthews, B., 2003.Kernel PLS-SVC for linear and nonlinear classification, *Proceedings of the Twentieth International Conference on Machine Learning* (ICML-2003), Washington DC, pp.640–647.

Rubin, D.B. 1987. Multiple Imputation for Non-response in Surveys, *Wiley*: New York.

Sapankevych, N. I. &Sankar, R., 2009. Time series prediction: Using Support vector machines, *IEEE Computational Intelligence Magazine*, pp.24:38.

Schafer, J.L. & Graham, J.W., 2002. "Missing Data: Our View of the State of the Art". *The American Psychological Association Incorporated*. 7, pp.147-177.

Scholkopf. B., Smola. A., & Muller, K. R., 1988.Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10, pp.1299-1319.

See, L. and Openshaw, S. 1999. Applying soft computing approaches to river level forecasting, *Hydrological Sciences Journal*, 44(5), pp.763–778.

Setiono, R., 2000. Generating concise and accurate classification rules for breast cancer diagnosis, *Artif.Intell.Med.*, 3, pp.205-219.

Shalabi, L. A., Najjar, M. &Kayed, A. A., 2006. A framework to deal with missing data in data sets, *Journal of Computer Science*, 2(9), pp.740-745.

Sheridan, R. P., NachbarJr, R. B. & Bush B. L., 1994. Extending the trend vector: The trend matrix and sample based partial least squares, *J. Comput.-Aided Mol. Design*, 8, pp.323-340.

Shu, L., Dong, G., Liu L., Tao, Y. & Wang, M., 2008.Water level variation and prediction of the Pingshan Sinkhole in Guizhou, Southwestern China, *Sinkhole and the Engineering and Environmental Impacts of Karst*.

Sjgstrgm, M., Wold, S., Lindberg, W.,Persson, J.A. & Martens, H., 1983. A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables, *AnalyticaChimicaActa*, 150, pp.61-70.

Smola, A. J. &Scholkopf, B., 2004. A tutorial on support vector regression, *Statistics and Computing*, 14, pp. 199-222.

Stone, M., 1974.Cross-validatory choice and assessment of statistical predictions, *J. Royal Stat. Soc.*, 36(2), pp.111-147.

Strang, G., 1988. *Linear Algebra and its Applications*, 3rd ed., Brooks and Cole. Pacific Grove, California.

Taian, L., Xin, X., Xinying, L. &Huiqi, Z., 2009. Application research of support vector regression in Coal Mine ground water level forecasting, *International Forum on Information Technology and Applications*, pp.507-509.

Thissen, U., Brakel, V., R., Weijer, D., A. P., Melssen, W. J., &Buydens, L. M. C., 2003. Using support vector machines for time series prediction, *Chemom. Intell.Lab. Syst.*, 69, pp.35–49.

Tikhonov, A.N. and Arsenin, V.Y., 1977. Solution of Ill-posed problems, *V. H. Winston and Sons*.

Tranmer, M. & .Elliot, M., 2008. Multiple linear regression, *Cathie Marsh Center for Census and Survey Research*.

Vapnik, V. N., 1982, Estimation of Dependences Based on Empirical Data, *Springer*, Berlin.

Vapnik, V., 1995.The Nature of Statistical Learning Theory, *Springer*, New York.

Walczak, B. &Massart, D. L., 1996. Application of radial basis functions- Partial least squares to non-linear pattern recognition problems: Diagnosis of process faults, *Anal. Chim.Acta*, 331(3), pp.187-193.

Wang, K., Han Z., Cui, S. &Zhong P., 2014. Flood runoff prediction using LS-SVR based on sliding time window, Journal of Information & Computational Science, 11(2), pp. 641-647.

Wei, C. C., 2012. Wavelet kernel support vector machines forecasting techniques: Case study on water level predictions during typhoons, *Expert Systems with Applications*, 39.

Weisberg, S., 2005.Applied linear regression, Wiley series in probability and statistics, *Wiley-Interscience*, A John Wiley & Sons, Inc., New York.

Wold, H., 1982. Soft modeling, The basic design and some extensions, in: K.-G. Jo reskog, H. Wold Eds., *System Under Indirect Observation*, vols. I and II, North-Holland, Amsterdam.

Wold, H., 1985. Partial least squares, In S. Kotz and N.L. Johnson, editors, *Encyclopedia of the Statistical Sciences*, 6, pp.581–591.

Wold, S., Ruhe, H., Wold, H.& Dunn W.J. III., 1984. The collinearity problem in linear regression, The partial least squares (PLS) approach to generalized inverse, *SIAMJournal of Scientific and Statistical Computations*, 5, pp.735–743.

Worsley, K.J., 1997. An overview and some new developments in the statistical analysis of PET and fMRI data, *Human Brain Mapping*, 5, pp.254–258.

Wu, C. H., Tzeng, G. H. & Lin R.H., 2009.A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression, *Expert Systems with Applications*, 36, pp.4275-4735.

Wu, C. L., Chau, K. W. & Li, Y. S., 2008. River stage prediction based on a distributed support vector regression, *Journal of Hydrology*, 358, pp.96-111.

Yeniay, O. &Goktas, A., 2002. A comparison of partial least squares regression with other prediction methods, *Hacettepe Journal of Mathematics and Statistics*, 31, pp.99-111.

Yu, S. P., Chen, S. T. & Chang, I. F., 2006. Support vector regression for real-time flood stage foreasting, *Journal of Hydrology*, 328, pp.704-716.

Zou, Y., Zhou, W. &Zhong, M., 2010. A model on the relation between the rainfall in Poyang Lake Basin and its Water Level, *Bioinformatics and Biomedical Engineering, 4th International Conference.*