UNIVERSITI TEKNOLOGI MARA

AN ENHANCED BOOSTED REGRESSION TREE MODEL FOR THE PREDICTION OF PM_{10} CONCENTRATION LEVEL USING SVM_BRT WITH QR LOSS FUNCTION COUPLING APPROACH

WAN NUR SHAZIAYANI
BINTI WAN MOHD ROSLY

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)

Faculty of Computer and Mathematical Sciences

June 2022

ABSTRACT

Malaysia experiences transboundary haze episodes in which the air contains particulate matter (PM) that is harmful to human health and the environment. Therefore, the main prediction model used in this study is Boosted Regression Trees (BRT) to predict three days ahead of PM_{10} concentration. However, the main problem with the common BRT technique is that it is not suitable for use in predicting extreme values of PM_{10} concentration levels. Besides, the problem with BRT is that over-fitting can occur if the number of trees is not suitable and also because of the complexity of the model, which is caused by the unsuitable number of predictor variables used in the model. Therefore, the aim of this study is to enhance the BRT model with Quantile Regression (QR) and Support Vector Machine (SVM) weight. This study used maximum daily monitoring records from 2002 to 2017 in Alor Setar, Klang, and Kuching which were analysed using four models: a boosted regression tree (BRT) model, a BRT with QR loss function model and a hybrid model between SVM and BRT with and without QR loss function. In order to get the best prediction model and to avoid over-fitting, the number of trees (nt) was optimized by using independent test set (TEST), cross validation (CV) and out of bag estimation (OOB). Then, to solve the extreme value issue in BRT, this study used the QR loss function rather than the Ordinary Least Square (OLS) loss function, since QR is more resistant to outliers. Meanwhile, the model then evaluated and the best method for predicting PM_{10} concentration was selected based on the lowest error and highest accuracy values. The findings revealed that the TEST and CV were the best methods to be used in BRT model while TEST and OOB were the best method in BRT with QR loss function model. In general, hybrid models (SVM-BRT) performed better than the single models with the values of RMSE (14.76, 34.56), NAE (0.15, 0.33), PA (0.58, 0.85), R^2 (0.33, 0.73) and IA (0.67, 0.92) for the first and second days ahead of prediction. The final comparison revealed that the BRT with QR loss function was significantly better at predicting future PM_{10} concentration than common BRT used by other researchers (BRT with OLS loss function). Finally, since the proposed model can accurately predict high air pollution levels, it can be used as a tool for early warning system in giving air quality information to local authorities in order to formulate air quality improvement strategies.

ACKNOWLEDGEMENT

Firstly, I wish to thank God for giving me the opportunity to embark on my PhD and for completing this long and challenging journey successfully. My gratitude and thanks go to my supervisor Assoc Prof Ts. Dr. Ahmad Zia Ul-Saufie Bin Mohamad Japeri who has provided me a good guidance with the most sincerity and similarly to my co-supervisor, Dr. Hasfazilah Binti Ahmat who has contributed good ideas in the process of completing my research.

My appreciation goes to my husband, Syarul Azan Bin Md Said for always being there by my side to support me. Special thanks to my parents, siblings and friends for given me motivations in completing my research.

Finally, I would like to express my greatest appreciation to Universiti Teknologi Mara for giving me study leave and special appreciation to the Department of Environmental, Malaysia, which has provided me with the secondary data on air pollution.

Alhamdulillah.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 Background of Study

Pollution is a general phenomenon which destroys useful substances within air or water. It is not a physical but a chemical situation that may have the ability to move throughout the atmosphere and water. Pollution means the surroundings will be drastically damaged, so it will become a significant threat to living things, either on land or in the aquatic environment (Pulipati, 2018). Air pollution has become a serious environmental problem in the developing Southeast Asian countries. Malaysia is ranked as the 98th worst country among 180 nations worldwide in terms of air quality (EPI, 2020). According to Yahaya (2019), the major sources of air pollutants in Malaysia are motor vehicle emissions and industry, particulate matter from stacks and exhaust, dust from quarrying activities, construction projects, and open biomass-burning aerosols from wildfires in Indonesia, which are also transported over Malaysia during the dry season and southwest monsoon. Therefore, ambient air quality readings in several Malaysian cities are exceeding the national ambient air quality standard. Indeed, air pollution has been shown to have a significant impact on human health, agriculture, and the ecosystem. Therefore, air pollution has been identified as a major cause of respiratory and cardiovascular diseases.

Particulate matters (PM) are notable pollutants within the air and it has a greater effect on human beings compared to other pollutants. The major components of PM are black carbon, ammonia, mineral dust, sodium chloride, nitrates, water and sulphate. It is a complex mixture suspended in the air that consists of liquid and solid particles of inorganic and organic substances. Particulate matter with aerodynamic less than 10 μm (PM_{10}) is one of the major air pollutants monitored by the Malaysian government, and it is included in the Air Pollution Index (API), a measure of air quality in Malaysia. Besides that, it can penetrate and lodge deep inside the lungs because it is a small particle with a diameter of 10 microns or less, ($\leq PM_{10}$). According to WHO (2005), air quality standards for PM_{10} are 50 μgm^{-3} for a 24-hourly concentration limit and 20 μgm^{-3} as an annual mean limit. Furthermore, PM_{10} is an air contaminant that has been implicated in a variety of health problems. Hassan