A HYBRID MODEL OF SARIMA AND ANFIS FOR MACAU AIR POLLUTION INDEX FORECASTING

THESIS DISSERTATION

By

Eason, Lei Kin Seng (M-A7-6560-7)

Supervisor: Dr. Wan Feng

In Fulfillment of Requirements for the Degree Of

Master of Science in Electrical and Electronics Engineering

June, 2013

Faculty of Science and Technology
University of Macau
A HYBRID MODEL OF SARIMA AND ANFIS FOR MACAU AIR POLLUTION INDEX FORECASTING

by

Eason, LEI KIN SENG

A thesis submitted in partial fulfillment of the requirements for the degree of

Electrical and Electronics Engineering

Faculty of Science and Technology
University of Macau

June, 2013

Approved by __________________________________________________

Supervisor

__________________________________________________

__________________________________________________

__________________________________________________

Date __________________________________________________________
In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Macau, I agree that the Library and the Faculty of Science and Technology shall make its copies freely available for inspection. However, reproduction of this thesis for any purposes or by any means shall not be allowed without my written permission. Authorization is sought by contacting the author at

Address: 31 AV DA CONCORDIA EDF. VANG HOI BL.4 7-ANDAR-E MACAU

Telephone: +853-66605700
Fax: N/A
E-mail: easonlei@hotmail.com

Signature ______________________
Date __________________________
Air pollution is an increasing problem arising from the rapid population growth and economic expansion in Macau over the past decade while serious harmful to human health such as, asthma and bronchitis are caused in parallel. As a result, more and more public awareness has placed on it and an effective system for supervising and forecasting the future Air Pollution Index (API) becomes obviously important. How to build up an accurate and dependable model to predict the future API is the goal of this research.

In this study, two different kinds of information can be obtained from Macau Meteorological and Geophysical Bureau (SMG) and they are: historical information (The past daily API records logged from January of 2000 to January of 2008); and meteorological information (Five daily pollutants recorded at the same period such as PM$_{10}$, SO$_2$, NO$_2$, CO, O$_3$ and five essential weather elements in daily based including temperature, relative humidity, wind speed, solar radiation and pressure).

To construct a proper model to describe the API system, we may reasonable think that all the related information should be measured as more information we apply to the model, the better performance it should have. Previous studies show that both Box-Jenkins models and Artificial Neuro-Fuzzy Inference System (ANFIS) models have been widely applied in API forecasting but none of them can be concluded as a universal model in different circumstances because of their common drawback – information singularity. Precisely speaking, through the analyses of historical observations, Box-Jenkins models can use to predict the future API without taking any meteorological information into account. With regard to ANFIS, it is not
subject to any historical information, instead, it simply employs the collected meteorological data sets and the actual API values as the input / output pairs and a suitable model can thereby be built for future forecasting after sufficient training. No doubt about it, by applying either model may give inadequate results. Therefore, the hybrid model is developed using the combination of Box-Jerkins model and ANFIS model in order to compensate the shortage of each other. The adopted hybrid model can consider with all the information so that to extend the prediction coverage and improve the forecasting ability.

In addition to hybrid approach, we also address the importance of data pre-processing. More specifically, there are over 30,000 observations stored in our historical and meteorological information; missing parts of data seem to be usual. To neglect the missing parts is not recommended since we may sacrifice some information stored behind and consequently, lead to inefficient analyses and bias the results. On the other hand, ten meteorological variables are found in this research and an excessive number of inputs not only impair the transparency of the underlying model, but also increasing the computation complexity. So, try to find out the missing values and figure out the most meaningful parts from all the observations are obviously required. Through different cases analyses, we verify that both missing data handling and input selection are significant and benefitted to the system performance.

To demonstrate the utility of the proposed scheme, the hybrid model with data pre-process techniques is used to forecast the daily API values of Macau city in January of 2008. The individual Box-Jerkins model and ANFIS model are also applied in order to assess the performance of the hybrid model. By examining the performance index - root mean square error (RMSE) and mean average percentage error (MAPE), the combined model is proved that it can be an effective way to enhance the forecasting accuracy compared with either the models used separately.
## TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ iv  
LIST OF TABLES ........................................................................................................... v  
LIST OF ABBREVIATIONS ............................................................................................. vi  

### CHAPTER 1: INTRODUCTION .............................................................................. 1  
**ABOUT THIS CHAPTER** .......................................................................................... 1  
1.1 Background ............................................................................................................. 1  
1.2 Data Acquisition ...................................................................................................... 3  
1.3 Literature Review ..................................................................................................... 4  
1.4 Objective .................................................................................................................. 9  
1.5 Challenges .............................................................................................................. 10  
1.6 Contribution and Thesis Organization ................................................................... 11

### CHAPTER 2: STOCHASTIC MODELS ................................................................. 13  
**ABOUT THIS CHAPTER** ........................................................................................ 13  
2.1 Auto Regressive Moving Average (ARMA) .......................................................... 13  
2.2 Seasonal Auto Regressive Integrated Moving Average (SARIMA) ...................... 14  
2.3 Conclusion ............................................................................................................. 16

### CHAPTER 3: ARTIFICIAL INTELLIGENT MODELS ............................................ 17  
**ABOUT THIS CHAPTER** ....................................................................................... 17  
3.1 Artificial Neural Networks (ANNs) ....................................................................... 17  
3.2 Fuzzy Inference Systems (FIS) .............................................................................. 18  
3.3 Adaptive Neuro-Fuzzy Inference Systems (ANFIS) ............................................. 20  
3.4 Conclusion ............................................................................................................. 23

### CHAPTER 4: PRE-PROCESSING FOR RAW DATA ............................................ 24  
**ABOUT THIS CHAPTER** ....................................................................................... 24  
4.1 Type of Missing Data ............................................................................................. 24  
4.1.1 Missing Completely at Random (MCAR) ....................................................... 25  
4.1.2 Missing at Random (MAR) ............................................................................. 25  
4.1.3 Missing not at Random (MNAR) .................................................................... 26
4.2 Missing Data Handling ..........................................................................................26
  4.2.1 Listwise Deletion ..........................................................................................26
  4.2.2 Mean Substitution .........................................................................................27
  4.2.3 Multiple Imputation ......................................................................................28
4.3 Input Selection .......................................................................................................35
4.4 Normalization ........................................................................................................37
4.5 Conclusion .............................................................................................................38

CHAPTER 5: DEVELOPMENT AND IMPLEMENTATION OF HYBRID
MODEL ................................................................................................................39
ABOUT THIS CHAPTER ...........................................................................................39
  5.1 Design of Hybrid Model ........................................................................................39
  5.2 Implementation of Hybrid model...........................................................................41
    5.2.1 Design Scheme for Box-Jerkins Model (MODEL 1) ...................................41
    5.2.2 Design Scheme for ANFIS Model (MODEL 2) ...........................................43
    5.2.3 Design Scheme for Hybrid Model ................................................................45

CHAPTER 6: HYBRID MODEL APPLICATION AND RESULTS
VALIDATION .....................................................................................................46
ABOUT THIS CHAPTER...........................................................................................46
  6.1 Simulation Softwares .............................................................................................46
  6.2 Time Series Analysis (MODEL 1) ........................................................................46
  6.3 Meteorological Data Analysis................................................................................51
    6.3.1 Histogram ...................................................................................................51
    6.3.2 Missing Data Analysis ..................................................................................54
    6.3.3 EMB-MI Analysis .........................................................................................57
  6.4 Input Selection .......................................................................................................61
  6.5 ANFIS (MODEL 2) ...............................................................................................62
    6.5.1 Structure Identification .................................................................................62
      6.5.1.1 Selection of Training and Testing Data ...............................................63
      6.5.1.2 Generation of Initial FIS ......................................................................63
    6.5.2 Parameter Identification ................................................................................63
  6.6 Hybrid Model .........................................................................................................64
LIST OF FIGURES

Figure 1-1 Design scheme for this research.................................................................10
Figure 3-1 Typical multi-layer back-propagation (BP) network ...............................18
Figure 3-2 First-order Sugeno fuzzy model..............................................................20
Figure 3-3 General structure for ANFIS.................................................................21
Figure 4-1 Idea of listwise deletion .........................................................................27
Figure 4-2 Concept of mean substitution.................................................................28
Figure 4-3 Matrix of multivariate data with missing values ..................................31
Figure 4-4 Schematic of EMB-MI algorithm ...........................................................35
Figure 4-5 Idea of input selection ............................................................................36
Figure 5-1 Design schematic of Box-Jenkins model (MODEL 1)...............................42
Figure 5-2 Design schematic of ANFIS (MODEL 2)................................................44
Figure 5-3 Design schematic of Hybrid Model .......................................................45
Figure 6-1 Series plot of the historical API from year 2000 to 2007 in daily base ....46
Figure 6-2 Series plot after de-trend and de-seasoning ..........................................48
Figure 6-3 ACF plot of the differenced series up to lags at 2s (s=365) ......................49
Figure 6-4 PACF plot of the differenced series up to lags at 2s (s=365) ....................49
Figure 6-5 ACF plot of the differenced series up to lags at 20 ..................................50
Figure 6-6 PACF plot of the differenced series up to lags at 20 ..................................50
Figure 6-7 Histograms of the 11 variables in our research ....................................53
Figure 6-8 Overall summary of missing values with 11 variables ............................55
Figure 6-9 Missing patterns (11 nos variables) .....................................................56
Figure 6-10 Distribution plot of observed and imputed values ...............................56
Figure 6-11 ANFIS structure for this research .......................................................60
Figure 6-12 The actual API vs predicted API in different cases ............................67
Figure 6-13 Scatter diagrams of predicted API (Hybrid model) and actual API ....68
LIST OF TABLES

Table 1-1 API range in Macau and its associated health influence and advice to public.................................................................2
Table 1-2 Sub-index and breakpoint pollutant concentration for Macau-API ........3
Table 1-3 Summary of different approaches for forecasting.............................8
Table 1-4 Summary of some previous methodologies for API forecasting..........8
Table 1-5 Propose solutions for the challenges in this research .......................11
Table 6-1 ADF test for the observed API series .............................................47
Table 6-2 ADF test for the de-trend & de-seasoning series ...............................48
Table 6-3 Coefficient of model SARIMA (2,1,2)(1,1,1)_365 ..........................51
Table 6-4 Univariate statistics of the 11 variables in our research ....................54
Table 6-5 Means & SD of observed data set ..................................................57
Table 6-6 EMB-MI analysis – Means & SD .....................................................58
Table 6-7 EMB-MI analysis – Covariance Matrix ...........................................59
Table 6-8 Correlation coefficient examination ...............................................61
Table 6-9 Performance examination for different cases .................................66
LIST OF ABBREVIATIONS

ACF. Autocorrelation Function

ADF-test. Augmented Dickey-Fuller test

AIC. Akaike’s Information Criterion

ANFIS. Adaptive Neuro-Fuzzy Inference System

ANNs. Artificial Neural Networks

AR. Auto-Regressive

ARMA. Autoregressive Moving Average

ARIMA. Autoregressive Integrated Moving Average

API. Air Pollution Index

AQI. Air Quality Index

BIC. Schwartz Bayesian Information Criterion

BP. Backpropagation

CCA. Complete Case Analysis

CO. Carbon Monoxide

EM. Expectation Maximization algorithm

FCM. Fuzzy C-Mean

FIS. Fuzzy Inference System

FL. Fuzzy Logic

GDP. Gross Domestic Product

LSE. Least Squares Error

MA. Moving Average

MAPE. Mean Average Percentage Error
MAR. Missing at Random
MCAR. Missing Completely at Random
MCMC. Markov Chain Monte Carlo
MNAR. Missing Not at Random
MI. Multiple Imputation
NO₂. Nitrogen Dioxide
O₃. Ozone
PACF. Partial Autocorrelation Function
PM₂.₅. Fine Suspended Particulate
PM₁₀. Respirable Suspended Particulate
PSO. Particle Swarm Optimization
RMSE. Root Mean Square Error
SARIMA. Seasonal Autoregressive Integrated Moving Average
SD. Standard Deviation
SMG. Meteorological and Geophysical Bureau of Macau
SO₂. Sulphur Dioxide
ACKNOWLEDGMENTS

This dissertation would not be possible without the guidance and the assistance of several individual parties for their valuable support and advice during the preparation and completion of this study.

First and foremost, my utmost gratitude to Dr. Wan Feng, I never forget the inspiration and motivation from Dr. Wan as I hurdle with the obstacles in the completion of this research work. He always mentions that we must keep the working manner in calm to handle with the problem. This valuable advice will absolutely help for my prospective career, not only in the study stage.

Mr. Joe Cheang, Control Lab Technician. Thanks and appreciates with his assistance in granting the access to control lab room so that I can prepare this research in a silence place without any interruption.

Mr. LM Tam, Senior Construction Manager (MEP) in Sands Cotai City project. His continuous support is not only providing for my daily working but also my master study.

Mr. Woody Cho, Construction Manager (MEP) in Sands Cotai City project. His technical advice from engineering point of view is critical and useful for this study.

My teammates in Sands Cotai City project. Thanks for their kindness for sharing my workload as they know I was preparing my master thesis in these years.

Lastly, I would like to say thank you to my family: Thank you for all love and support to me.
DEDICATION

I wish to dedicate this thesis to my parents, my wife Miller and my lovely daughter Mavis.