Towards Time Series Forecasting of Cross-Data Analytics for Haze Prediction

Ali Akbar, Muhammad Atif Tahir, Muhammad Rafi
National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
{k201306,atif.tahir,muhammad.rafi}@nu.edu.pk

ABSTRACT

Atmospheric pollution and a thin/thick layer of dust/smoke (Haze) has become one of the major issues all over the world. It obscures the visibility of the sky[1]. The paper aimed to first explore the dataset with the help of visualizations and ACF and PACF plots and analyzing the trends and seasonality components. Thereafter, Time series methodologies were applied to predict PM10 values, given the same countries data and the data from other neighbouring countries as well. Models, including ARIMA and SARIMA, were applied and tuned along with training methodologies including Grid Search and Walk-Forward validation. The paper also employed Vector Auto-Regression (VAR) methodology to capture the cross data relationship between one country and the other. The implementation and model produced the best (across the two sub-tasks) SMAPE scores of 44.96, 29.07 and 27.74 on the Brunei, Singapore and Thailand datasets, respectively.

1 INTRODUCTION

Environmental pollution is a major concerned. The basic idea of the research in Haze Prediction is presented in [1]. Our main motivation to participate in the challenge, is to be the part of the research aimed at the existing unsolved problem of Haze. It is perceived as a major problem for countries around the world and is becoming a serious health concern across the globe.

The paper aimed to focus on the first two subtasks i.e 1) predicting the PM10 values using the data from the same country, and, 2) predicting the PM10 values from the same country and other country's data as well. The dataset was first visualised, structured and arranged in different files and also tackled one of the challenges of the missing values as the parameters, along with PM10 data, were not available for all the weather stations and at all times. Thereafter several Time Series and Machine Learning models were applied and evaluated based on the results of the train and test datasets.

2 RELATED WORK

The paper in [2] discussed methodology to forecast the haze occurrences in Southeast Asia. The paper utilized a Convolutional Neural Network (CNN) based framework, known as HazeNet which has a 16-layered architecture and had been trained on 18 hydrological and meteorological features and time-sequence maps of about 35 years and achieved about 95.2% accuracy in the validation set, however, neglected the impact and importance of transboundary haze effects.

The authors in paper [3] identified two major research gaps that persisted in predicting PM10 concentrations in Brunei, Darussalam: 1) The recent research did not take the use of CNN and Recurrent Neural Networks (RNN) into account for haze prediction use-case, and 2) The majority of other researchers used the data from 1997 to 1998 period, which was considered a disastrous period in that region, which made the outcomes of research biased and not widely applicable. The authors attempted multiple Time Series and Deep Learning techniques including Moving Averages (average of PM10 values with shifting average-window), Linear Regression, and RNN (with 1-D Convolutional layer), of which CRNN proved to be the best performing model throughout. However, this paper also does not cater transboundary effects for Haze Prediction and additionally explicitly mentions and proposes this approach as a future work possible for this paper.

The authors in paper [4] studies the effect of transboundary haze events in Malaysia by using Multiple Linear Regression (MLR) for estimating PM10. The paper used stepwise MLR with a 95% confidence interval and the dataset was divided into 70% for training and 30% for testing. Along with using normalization, the authors used different tests including Durbin–Watson (DW) test and R-squared test for correlation. The authors concluded by showing different test results and standard deviations and dispersions graphs, with the PM10(t+1) model giving accuracy of 0.668. The paper did incorporate the use of transboundary haze prediction but did not incorporate, analyse and compare the localized version of haze in a region.

The authors in paper [5] used the Convolutional RNN to determine the transboundary based haze levels in island cities and introduced a Dynamic C-RNN that combined a CNN and RNN and can model the interactions spatially and temporally. The paper used Spatial Transformation and techniques such as Inverse Distance Weighting and transformed the data, keeping the constraint and assumption of the island in view. The paper finally concluded by stating results that proved that D-CRNN proved to be successful when compared with existing state-of-the-art algorithms. While this paper makes an attempt to cater the transboundary and local effects, many assumptions and methodology involved were keeping in view that the training and testing is to be done for island cities, especially while transforming data. Therefore, the results, methodology and data transformation techniques may or may not be consistently supportive with non-island cities/countries.

3 APPROACH

The first step in the study was to understand and visualize the data to have the basic understanding of the values. In addition, there were some null values in the dataset which had to be taken care of since the models implementation required the complete

MediaEval'21, December 13-15 2021, Online

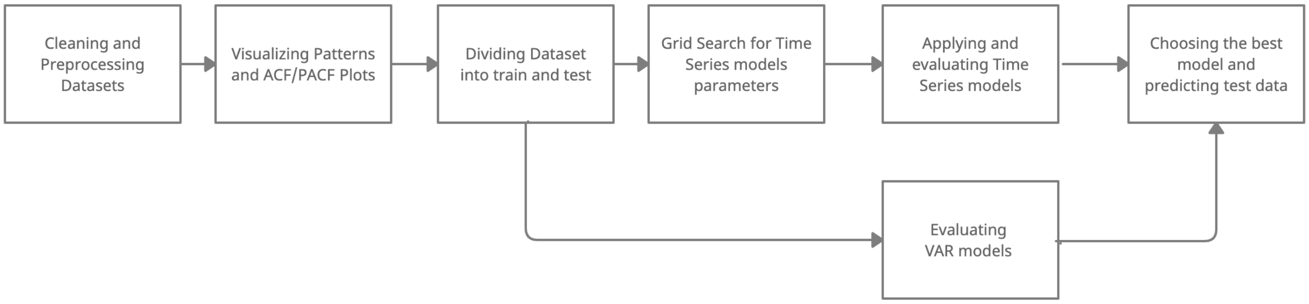
Cleaning and Preprocessing Datasets
Visualizing Patterns and ACF/PACF Plots
Dividing Dataset into train and test
Grid Search for Time Series models parameters
Applying and evaluating Time Series models
Choosing the best model and predicting test data
Evaluating VAR models

Figure 1: Architecture diagram of the proposed approach

dataset without any null values. This was done by trying different methods on different datasets, included using Last observation carried forward and linear imputation. These methods were tried and tested several times and its impact on accuracies were observed. It was finally concluded that the Last observation carried forward method produced the best SMAPE scores and hence was applied to impute the missing values.

Further, as shown in Figure 1, the data and its structure was arranged as required in different models, for example combining the date, month and year columns to make it a date-time column. Thereafter the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) was studied to get the real insights of the dataset(including observing the trend and seasonality components and thereby using 1 seasonal differencing order in the model fitting and selection step) and to find the appropriate lag variables to be used at the time of fitting a Time Series model. It was identified that lag 1 and 2 were of most value in the datasets and hence this observation was applied in the model fitting stage and was further verified by the grid search method. This step, along with other steps that were dataset specific, were carried out three different times since we had data from three different countries. For the training and model evaluation purposes, the datasets were divided into two parts: 70% for training and 30% for testing, i.e choosing the first 70% of the data as training and the last 30% data as test data and hence preserving the chronological sequence of the time series nature.

The task 1 was to predict the PM10 value at different locations in multiple countries using data only from each country itself, and therefore several Time Series techniques were applied and evaluated for example using Moving Averages, and ARIMA [6]/SARIMA [7] models. Methods such as Grid Search and Walk-Forward validation method were used while evaluating and validating the Time Series models so that the best and accurate models were selected while weighing the values of PM10 with the amount of time being passed. While these methods proved to give good accuracies on Brunei dataset (lower error rate compared to others), the methodology slightly failed on other datasets mainly due to Google Colab(Online Python notebook and execution engine), used to train and test the models, being crashed with bigger amount of data and heavy RAM consumption. This heavily impacted the results since the data then had to be divided into batches and then given for training which impacted the results. However, the best model that could be produced was SARIMA with optimal pdq parameters as (1, 0, 1) and optimal PDQS parameters as (0, 1, 1, 12).

For the task 2, the Vector Auto Regression (VAR) models were implemented to find the relationship of one country's data with other countries data along with other station's provided data. The data from other country, including other weather variables were merged into a data-frame and then fed to the VAR model. Additionally, statistical tests including Cointegration Test were used to find the relationship between multiple features and the lag variables were identified with the help of the least AIC and BIC scores.

4 RESULTS AND ANALYSIS

The models provided considerably fair results for all the three datasets including Brunei, Singapore, and Thailand datasets across the two sub-tasks. However, it is noted that there is no much difference between the two sub-tasks and this may be due to under-fitting of the model and non-effective parameter selection since the models in Time Series vary a lot depending upon the lag variables and the choice of other parameters. The individual scores are reported below:

Task	Brunei		Singapore		Thailand	
	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
1	9.418	44.96	7.437	29.07	7.473	27.85
2	9.419	44.97	7.436	29.07	7.443	27.74

5 DISCUSSION AND OUTLOOK

The proposed model for task 1 achieved encouraging results for all the regions, however, for the task 2, the model's performance was not very satisfactory and did not achieve better results which may be due to under-fitting and non-efficient choice of parameters. Time series models are very much dependent on identifying the key dependency of lag values. Therefore, it is observed that there is a need to reevaluate these values and a modified model may be suggested for the second sub-task. Moreover, the model tuning and hyper-parameters learning can also be used to improve the model and since Task 2 and Task 3 were more challenging and requires more complex model, spending more time and resources on model selection and fine tuning could be very useful. We would finally like to thank MediaEval and the task organizers to provide us with an opportunity to work in this domain and contribute to the society.

REFERENCES

[1] A. Kasem, M.-S. Dao, E. N. Aziz, D.-T. Dang-Nguyen, C. Gurrin, M.-T. Tran, T.-B. Nguyen, and W. Suhaili, "Overview of insight for wellbeing task at mediaeval 2021: Cross-data analytics for transboundary haze prediction," *Proc. of the MediaEval 2021 Workshop, Online*, December 2021.

[2] Wang and Chien, "Exploiting deep learning in forecasting the occurrence of severe haze in southeast asia," *arXiv preprint arXiv:2003.05763*, 2020.

[3] E. N. Aziz, A. Kasem, W. S. H. Suhaili, and P. Zhao, "Convolution recurrent neural network for daily forecast of pm10 concentrations in brunei darussalam," *Chemical Engineering Transactions*, vol. 83, pp. 355–360, 2021.

[4] S. Abdullah, N. N. L. M. Napi, A. N. Ahmed, W. N. W. Mansor, A. A. Mansor, M. Ismail, A. M. Abdullah, and Z. T. A. Ramly, "Development of multiple linear regression for particulate matter (pm10) forecasting during episodic transboundary haze event in malaysia," *Atmosphere*, vol. 11, no. 3, p. 289, 2020.

[5] P. Zhao and K. Zettsu, "Convolution recurrent neural networks based dynamic transboundary air pollution predictiona," pp. 410–413, 2019.

[6] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[7] A. E. Permanasari, I. Hidayah, and I. A. Bustoni, "Sarima (seasonal arima) implementation on time series to forecast the number of malaria incidence," in *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 203–207, IEEE, 2013.