

Overview of Insight for Wellbeing Task at MediaEval 2021: Cross-Data Analytics for Transboundary Haze Prediction

Asem Kasem¹, Minh-Son Dao², Effa Nabilla Aziz¹, Duc-Tien Dang-Nguyen³, Cathal Gurrin⁴,
Minh-Triet Tran⁵, Thanh-Binh Nguyen⁵, Wida Suhaili¹

¹Universiti Teknologi Brunei

²National Institute of Information and Communications Technology, Japan

³University of Bergen, Norway

⁴Dublin City University, Ireland

⁵University of Science VNU-HCMUS, Vietnam

ABSTRACT

This paper provides an overview of the MediaEval 2021 task on "Insights for Wellbeing: Cross-Data Analytics for Transboundary Haze Prediction". The task targets researchers in multimedia information retrieval, machine learning, data science, environmental and atmospheric sciences. The term "cross-data" refers to approaches that utilize multimodal data across domains, platforms, and prediction models to make predictions. The main objective of this task is to perform accurate multi-day haze prediction in some neighboring countries from the ASEAN region. There are three subtasks: (i) 3-Day Localized Air Pollution Prediction, with emphasis on accurate predictions in each country depending only on its weather and air quality data; (ii) 3-Day Transboundary Air Pollution Prediction, with emphasis on addressing transboundary haze effects through the use of multiple cross-data sources; (iii) Transfer Learning sub-task, focusing on transfer learning techniques to demonstrate that patterns learned from certain regions' data sources help in improving predictions for other regions. The task aims to utilize cross-data sources to provide accurate predictions that can help mitigate the adverse effects of haze air pollution.

1 INTRODUCTION

Haze air pollution describes the pollution consisting of particulate matter of smoke, dust, and other vapors present in the air, which originate from large-scale forest and land fires, factories, and cars. This mixture of air-borne pollutants, when it reaches high levels, causes respiratory health problems [2], and has negative impacts on visibility [3], economic production [6], transportation [4], and tourism [5]. Transboundary haze problem refers to the situation where high levels of haze remain measurable after crossing into another country's air space, resulting a recurrent issue in many regions in the world, especially in Southeast Asia where the sources contributing to haze pollution differ at each country with varying percentages come from localized or transboundary sources. For example, transboundary (haze) pollution episodes are often attributed to the long-range transport of biomass fires from slash-and-burn activities during dry seasons or from forests fires, which travel depending on weather conditions to affect several neighboring countries.

2 RELATED WORKS

Particulate matter concentrations (PM₁₀, PM_{2.5}) are usually used to calculate air quality index measures that describe pollutant severity. Past recorded data of air pollution and meteorological parameters have been used by researchers and practitioners from academia and government agencies to develop air pollution prediction models to forecast changes in air pollution. In [1], the authors aimed to predict the first three hours during transboundary haze events by using three different stepwise Multiple Linear Regression (MLR) models for predicting the PM₁₀ concentration, one for each hour.

In [7], the authors emphasized modeling diverse inter-station relationships for air quality prediction of city-wide stations using an Attentive Temporal Graph Convolutional Network (ATGCN) model. The method could be extended to transboundary haze prediction if considering inter-station relationships in region-wide stations. In [8], the authors generated ordered city clusters by higher-order spectral clustering on pollution-transport networks among cities, then projected those clusters into one-dimensional Euclidean space. The clusters contributed to the partial differential equation (PDE) model for predicting PM_{2.5}. In [9], the authors introduced a hierarchical graph neural network-based air quality forecasting method working on data collected from stations. Here, transboundary haze can be concerned as the diffusion processes of air pollutants between cities and monitoring stations. In [10], authors analyzed transboundary haze from mainland China to Fukuoka, Japan, using atmospheric sensing data. They used a CRNN-LSTM model to predict PM_{2.5}.

3 TASK DESCRIPTION

The task is organized into three subtasks sharing a common objective, but differing in the approach and data sources utilized to achieve it. The objective for all subtasks is to perform 3-day prediction of air pollution, as measured by PM₁₀, in three different ASEAN countries, namely: Brunei, Singapore, and Thailand. There are training and testing timeseries data for each country, with daily or hourly readings of weather and air pollution parameters. The training data covers mostly 2010-2017 period (Singapore is only for 2016-2017), and the testing data covers mostly 2018-2019 period.

Participants need to predict, as accurately as possible, the PM₁₀ values for certain temporal gaps (of consecutive days) in the testing data. The same objective will be revisited in each subtask, but with differences in terms of the data sources accessible and the prediction approach, as explained below.

3.1 3-Day Localized Air Pollution Prediction

In this subtask, the objective is to predict the PM10 value for the first 3 days in each gap in the testing files, at the location of each air-quality station in a country. To do that, participants are required to develop 3 different predictive models based only on the localized training data from each country. This subtask will explore the accuracy of predicting air pollution for 3 days ahead, and it is designed to evaluate how well this objective can be achieved if each country depends only on its own weather and air pollution data.

The main challenges in this subtask are expected to be on addressing missing reading values, capturing the timeseries changes across multiple stations in a country, and finding insights that may improve prediction quality for each country separately. For example, in some countries, only certain stations will have wind speed and direction readings, or certain stations might be faulty for a long duration. To encourage multi-disciplinary research, we believe that spatial interpolation methods used in environmental studies and GIS-based solutions, or insights that consider the type of parameter, and/or geographical distance information, can be helpful in filling missing parameter readings, or in increasing the spatial granularity of readings in each country.

3.2 3-Day Transboundary Air Pollution Prediction

The objective in this subtask is the same of the previous subtask with respect to the values to predict, but participants are required to consider other data sources available from the same country or neighboring countries (including training data of other countries, remote sensing, terrain information, information, social media streams, news reports, etc.). This subtask attempts to address transboundary haze effects by observing the improvements of prediction accuracy once the haze and weather situation (e.g. wind/fire information) in neighboring countries is taken into account, or through other insights and conclusions that the participants may find.

There will be additional weather and air pollution data from stations in Indonesia, containing daily weather and weekly PM10 readings, which participants are encouraged to utilize in modeling or pre-processing steps. A challenge in this subtask could be the synchronization of multiple data sources, since each country has a different subset of parameters, readings frequency, distribution of stations, and missing periods of recordings. One approach to synchronize the reading frequency differences could be to either summarize the more frequent readings, or apply some form of temporal interpolation to the less frequent readings. Participants can consider additional data sources (cross-data) to improve prediction and/or find insights based on environmental factors, satellite remote sensing, social/news data, etc.

3.3 Use of Transfer Learning

In this subtask, participants will re-visit either of subtask 1 or 2 above, by considering transfer learning techniques in their solutions. For example, participants can reattempt subtask-1 to answer the question: can the use of pre-trained models from one country improve the 3-day prediction of another country?

The application of transfer learning can demonstrate that patterns learnt from certain regions' data sources (e.g., via access to

larger datasets) help to improve predictions in other regions (e.g. where data is scarce or less granular).

4 DATASETS AND EVALUATION

All datasets are provided in CSV format, but with different structure per country, as the granularity, weather and pollution parameters, and number of stations is different across countries. For each country, the testing datasets will have similar structure to the training datasets, but containing multiple temporal gaps (of 8 consecutive days) where only the dates and stations IDs are given. Each gap will be preceded by consecutive periods (of about 10 days) where all available readings are provided (with some potential missing values). Participants need to predict the PM10 values of the first 3 days in each temporal gap. Predicting more than 3 days is also welcomed, but will not be used for ranking the submitted results.

For evaluation, predicted PM10 values at each station in a country will be compared with ground truth values using MAE (Mean Absolute Error) function. Submissions will be ranked (based on MAE) for each country's testing data separately, and points will be assigned based on least average MAE across all stations (there are roughly equal number of required predictions for each station in a country). For each subtask, total points for all countries will be summed and used to determine the overall ranking in that subtask. Furthermore, based on the submitted working notes, the approach used will be evaluated by the task organizers in terms of innovativeness, motivation of used methods, and gained insights. In any subtask, the data and methods used to make predictions should adhere to the given subtask's descriptions and conditions.

5 DISCUSSION AND OUTLOOK

Prediction of haze air pollution is an important task that can guide policy and decision making in many countries that are affected by it. However, its usefulness depends on accurate prediction for a sufficient duration of time, and hence the 3-day ahead requirement of the task. Given the transboundary haze effect, it is expected that neighbouring countries can benefit from each other's data sources to improve their own prediction. In practice however, agencies in different countries will have data sources that are different in the parameters they record, and the (spatial and temporal) granularity of recordings. Besides, other non-traditional data sources such as satellite images or social media streams may help the overall objective of accurate prediction. The subtasks were organized to address the practical challenges related to transboundary haze pollution, and we hope to motivate researchers to present innovative approaches in tackling them. Details on the methods and results of each participating team can be found in the working note papers of the MediaEval 2021 workshop proceedings.

ACKNOWLEDGMENTS

We would like to thank Brunei's Meteorological Department (BDMD) and Department of Environment, Parks and Recreation (JASTRE); Singapore's Meteorological Service Singapore (MSS) and National Environment Agency (NEA); Thailand's Meteorological Department (TMD) and Pollution Control Department (PCD); and Indonesia's Meteorology, Climatology, and Geophysical Agency (BMKG), for providing the meteorological and air quality data.

REFERENCES

- [1] Samsuri Abdullah, Nur Nazmi Liyana Mohd Napi, Ali Najah Ahmed, Wan Nurdiana Wan Mansor, Amalina Abu Mansor, Marzuki Ismail, Ahmad Makmom Abdullah, and Zamzam Tuah Ahmad Ramly. 2020. Development of Multiple Linear Regression for Particulate Matter (PM10) Forecasting during Episodic Transboundary Haze Event in Malaysia. *Atmosphere* 11, 3 (2020).
- [2] Kang Hao et. al. Cheong. 2019. Acute Health Impacts of the Southeast Asian Transboundary Haze Problem: A Review. *Int. J. Environ Res Public Health* 16, 18 (2019).
- [3] Tim Forsyth. 2014. Public concerns about transboundary haze: A comparison of Indonesia, Singapore, and Malaysia. *Global Environmental Change* 25 (2014), 76–86.
- [4] Ding A. Wang Z. et al. Huang, X. 2021. Amplified transboundary transport of haze by aerosol–boundary layer interaction in China. *Nat. Geosci.* 13 (2021), 428–434.
- [5] M. Mostafanezhad and O. Evrard. 2018. Geopolitical Ecologies of Tourism and the Transboundary Haze Disaster in Thailand, Laos and Myanmar. *The Tourism–Disaster–Conflict Nexus (Community, Environment and Disaster Risk Management)* 19 (2018), 53–66.
- [6] Euston Quah, Wai-Mun Chia, and Tsiat-Siong Tan. 2021. Economic impact of 2015 transboundary haze on Singapore. *Journal of Asian Economics* 75 (2021), 101329.
- [7] Chunyang Wang, Yanmin Zhu, Tianzi Zang, Haobing Liu, and Jiadi Yu. 2021. Modeling Inter-Station Relationships with Attentive Temporal Graph Convolutional Network for Air Quality Prediction. In *WSDM '21 (WSDM '21)*. Association for Computing Machinery, 616–634.
- [8] Yufang Wang, Haiyan Wang, and Shuhua Zhang. 2020. Quantifying prediction and intervention measures for PM2.5 by a PDE model. *Journal of Cleaner Production* 268 (2020), 122–131.
- [9] Jiahui Xu, Ling Chen, Mingqi Lv, Chaoqun Zhan, Sanjian Chen, and Jian Chang. 2021. HighAir: A Hierarchical Graph Neural Network-Based Air Quality Forecasting Method. (2021). arXiv:cs.LG/2101.04264
- [10] P.J. Zhao and K. Zettsu. 2018. Convolution recurrent neural networks for short-term prediction of atmospheric sensing data. In *IEEE CPSCoM-SmartData*.