Personal Air Quality Index Prediction Using Inverse Distance Weighting Method

Trung-Quan Nguyen[1], Dang-Hieu Nguyen[2], Loc Tai Tan Nguyen[3]
[1, 2, 3] University of Information Technology, Ho Chi Minh City, Vietnam
[1, 2, 3] Vietnam National University, Ho Chi Minh City, Vietnam
quannt.13@grad.uit.edu.vn,hieund.12@grad.uit.edu.vn,locntt.12@grad.uit.edu.vn

ABSTRACT

In this paper, we propose a method to predict the personal air quality index in an area by only using the levels of the following pollutants: PM2.5, NO2, O3. All of them are measured from the nearby weather stations of that area. Our approach uses one of the most well-known interpolation methods in spatial analysis, the Inverse Distance Weighted (IDW) technique, to estimate the missing air pollutant levels. After that, we can use those levels to calculate the Air Quality Index (AQI). The results show that the proposed method is suitable for the prediction of those air pollutant levels.

1 INTRODUCTION

The need to know the personal air pollution data is vital because it is better to provide each individual with regional air quality data, which seems to be more accurate than the global data measured from far away weather stations. The problem is finding a suitable method to predict air quality data in a local area from the global data. This paper reports our solution to tackle this challenge.

To know more about this challenge and the dataset that we will use, you can refer to the overview paper of MediaEval 2020 - Insight for Wellbeing: Multimodal personal health lifelog data analysis [1].

2 RELATED WORK

The inverse distance weighting method [4] is used commonly in spatial interpolation [3]. This paper will apply the basic form of IDW without any modification.

3 APPROACH

Due to the limited time available for experimenting with algorithms requiring more time to train data, such as neural network-related algorithms, we choose the IDW. Moreover, because there are no statistical assumptions involved [2], it is simpler than Kriging or other statistical interpolation methods. The way it works is easy to understand. Based on the assumption that closer points will have similar values than further points, it will use the measured values surrounding the unknown point to predict the value. By giving each known point a weight, the predicted value will be the average of those points.

The weight w_i for a known point i is the inverse of the distance d from that point to the unknown point x, which is computed as:

$$w_i = \frac{1}{d(x, x_i)^p} \tag{1}$$

with p is the power value that is used to control the value of the weight. It should be noticed that the Haversine method is used to calculate the distance between the two coordinates.

The value y of an unknown point x is calculated as:

$$y(x) = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

with w_i is the weight, X_i is the value of the known point i_{th}.

3.1 Prediction

At first, all possible time frame in hour-interval is listed by grouping the training data. Then, we start to loop through the training data per time frame.

In each loop, we get the coordinates of all unknown points that need to be predicted. After that, we get the values of the known points and their respective coordinates from the public air pollution data provided by 26 weather stations surrounding the Tokyo area also in that time frame.

With all the necessary data gathered, we can use the IDW formula to make the prediction. Please note that the initial power value p of the IDW formula is 2.

After repeating those steps for each air pollutant data (PM2.5, NO2, O3), we have the final results.

3.2 Optimization

To have the best performance, we could find the optimal value of power value p by trying different values of p until the IDW produces acceptable values of SMAPE/RMSE/MAE.

After evaluating the p-value ranges from 0 to 5, we find that the best power values for PM2.5, NO2, and O3 are 1.5, 3.5, and 0, respectively.

4 RESULTS AND ANALYSIS

The evaluation of PM2.5, NO2, O3, and AQI prediction provided by MediaEval task organizers are shown in Table 1, Table 2, Table 3, and Table 4, respectively.

In general, PM2.5 prediction is acceptable, but there is a big gap in NO2 and O3 prediction results. It is mainly because the IDW formula does not have any offset parameters to compensate for the big difference between weather stations' public weather data and the one carried out by personal equipment used by volunteers. This could be because of some differences in methods and devices of those two data providers.

MediaEval'20, December 14-15 2020, Online

Table 1: Evaluation of the PM2.5 prediction

Sensor	MAE	RMSE	SMAPE
100001	5.190319201	8.732748788	0.45931373
100002	3.720370835	5.511739014	0.406428735
100003	1.619832154	2.095919331	0.133032135
100005	2.874009812	4.055352722	0.35371517
100006	3.233921439	4.341966928	0.468214919
100007	1.695290448	1.707219278	0.625317245
200003	6.465190052	9.724716828	0.444137991
200004	4.815504659	7.436923815	0.400557289

Table 2: Evaluation of the NO2 prediction

Sensor	MAE	RMSE	SMAPE
100001	30.15104	34.62797	0.729989
100002	13.80071	18.2614	0.399087
100003	18.85267	20.40416	1.218212
100005	12.69285	16.3694	0.411915
100006	11.92978	14.12164	0.452494
100007	14.99076	15.85102	0.562354
200003	12.27167	15.1809	0.364154
200004	7.664357	9.571268	0.257642

Table 3: Evaluation of the O3 prediction

Sensor	MAE	RMSE	SMAPE
100001	11.14697072	16.74763774	0.474838877
‘ 100002	13.71316126	18.17918429	0.595873229
100003	12.15603603	14.13207772	0.554840783
100005	12.91552723	15.99672071	0.53328839
100006	15.72452576	19.40818331	0.728461886
100007	30.3013034	31.07255621	1.600495059
200003	14.62686484	18.79131409	0.490170718
200004	22.0919231	31.69232972	0.58440423

Table 4: Evaluation of the AQI prediction

Sensor	MAE	RMSE	SMAPE
100001	18.21506046	34.20371647	0.496721967
100002	18.10474466	38.8695944	0.49921946
100003	30.32401094	78.4465017	0.311432437
100005	10.79848535	19.6665506	0.389208159
100006	14.29939129	34.48844262	0.44466795
100007	23.5094483	60.19537217	0.521219253
200003	16.31585216	22.42326978	0.4097449
200004	12.93598111	19.188617	0.378573048

5 DISCUSSION AND OUTLOOK

We intend to explore more advanced algorithms in our future work, such as the advanced form of IDW [4], the combination of IDW with multiple regression. Also, we plan to utilize more weather data, such as wind direction, wind speed, temperature, to improve accuracy.

REFERENCES

[1] Zhao P. J Nguyen N.T. Nguyen T.B. Dang-Nguyen D. T. Gurrin C. Dao, M. S. 2020. Overview of MediaEval 2020: Insights for Wellbeing Task - Multimodal Personal Health Lifelog Data Analysis. In *MediaEval Benchmarking Initiative for Multimedia Evaluation, CEUR Workshop Proceedings*.

[2] Leonardo Ramos Emmendorfer and Graçaliz Pereira Dimuro. 2020. A Novel Formulation for Inverse Distance Weighting from Weighted Linear Regression. In *Computational Science – ICCS 2020*, Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira (Eds.). Springer International Publishing, Cham, 576–589.

[3] Jin Li and Andrew D. Heap. 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics* 6, 3 (2011), 228 – 241. https://doi.org/10.1016/j.ecoinf.2010.12.003

[4] Donald Shepard. 1968. A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. In *Proceedings of the 1968 23rd ACM National Conference (ACM '68)*. Association for Computing Machinery, New York, NY, USA, 517–524. https://doi.org/10.1145/800186.810616