Spatial-temporal Pattern Analysis and Prediction of Air Quality in Taiwan

Ping-Wei Soh
Institute of Computer and Communication Engineering, National Cheng Kung University
Tainan, Taiwan
pingweisoh@gmail.com

Kai-Hsiang Chen
PhD Program for Multimedia Systems and Intelligent Computing, National Cheng Kung University and Academia Sinica, Tainan, Taiwan
kevin761015@gmail.com

Jen-Wei Huang
Department of Electrical Engineering, National Cheng Kung University
Tainan, Taiwan
jwhuang@mail.ncku.edu.tw

Hone-Jay Chu
Department of Geomatics, National Cheng Kung University
Tainan, Taiwan
honejay@mail.ncku.edu.tw

***Abstract*—This study explores the spatial-temporal patterns of particulate matter (PM) in Taiwan. Probability map of PM and daily patterns are discussed in this study. Data mining provides more detailed spatial-temporal information for PM variations and trends. The proposed model will show that data mining provides a relatively high goodness of fit and sufficient space-time explanatory power, particularly air pollution frequency and affect areas. In the proposed model, a method using Dynamic Time Warping is proposed to analyse temporal similarity between stations. The proposed model can eliminate global effect on a single station through the performance of multiple stations. The proposed model will further be used for prediction of PM2.5. The prediction results will discuss the spatial-temporal relations between stations. This study will investigate the distribution of PM and its cyclicality.**

***Keywords*—PM2.5; spatial-temporal; data mining; Dynamic Time Warping.**

1. Introduction

With the rising of PM2.5 in various parts of Taiwan, we believe that analysing the distribution of particulate matter in the relation between time and space can effectively control the pollution and issue early warning to reduce the impact on health. PM's spatial changes in air pollution are not only meteorological conditions but also affected by human activities such as industrial development and vehicle emissions [1]. Meteorological conditions cause a significant air pollution changes [2] [3], such as long-distance transport or local air pollution accumulation. Wind direction and wind speed are also key factors [4] [5]. The wind of Taiwan's summer is blowing from southwest of Taiwan, because the west and south of Taiwan are the windward side, air pollution is easy to be blown away after the occurrence. Therefore, the summer has the best air quality of the year. While the PM concentration has a high seasonal variation, especially in the northern and southwestern of Taiwan, because of the monsoon [6] [7]. After October, the northeast monsoon began to prevail. If the winds were not strong enough, the monsoon was susceptible to the central mountain range, especially in the southern area. The air pollutants were more likely to accumulate while the wind speed was reduced by the central mountain range, making the local circulation dominate the Kaohsiung air pollution model, and industrial activity also causes high PM concentrations [8].

Recent studies have indicated that considering time and space in analysing air quality are inevitable [9] [10] [11]. PM has a high cyclicality and is easily affected by space, which may stagnate or diffuse to pollute surrounding environment. If we only analyse PM in time domain, we may neglect the impact and relations between multiple regions. While only considering space in analysis, we may loss the information of PM diffusion from place to place. Therefore, only by considering both time and space relations, we can accurately explain the diffusion behaviour of PM.

Data mining provides a new way to analyse air quality in the absence of physical model [12] [13] [14]. The advantages of using data mining technique is that we may discover hidden information in collected data. In addition, once a model is trained, time consumption for predictions is far more less than traditional physical model.

This study focuses on two main approaches: (i) Spatial-temporal analysis of PM2.5 in Taiwan, (ii) Relations and influences between stations.

2. Preliminary

2.1. Spatial temporal analysis

The probability of PM2.5 was estimated based on a threshold approach. The threshold, i.e., $35\mu g/m^3$, was the standard of PM2.5 control made by the government. After the analysis, the spatial probability map can be generated. Moreover, the daily variation of PM2.5 can be found from a temporal pattern analysis.

978-1-5386-2761-7

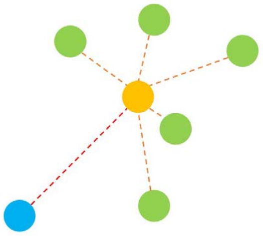


Figure 1. kNN-ED example (k=5)

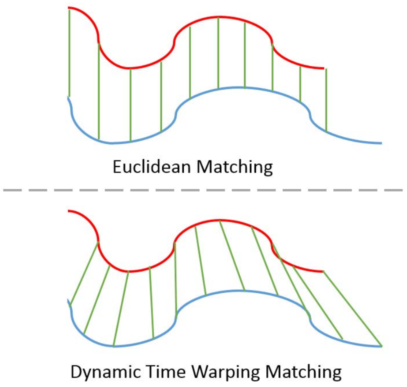


Figure 2. Euclidean Matching vs DTW Matching

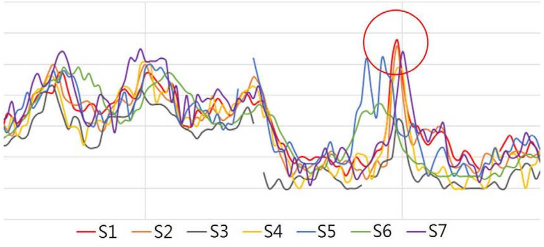


Figure 3. Time series of 7 stations' PM2.5 data

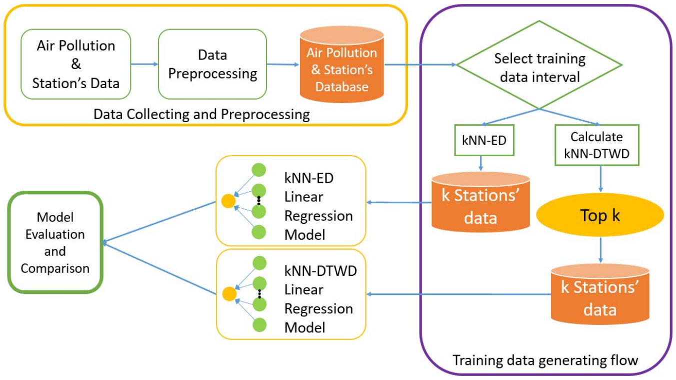

Air Pollution & Station's Data
Data Preprocessing
Air Pollution & Station's Database
Data Collecting and Preprocessing
Select training data interval
kNN-ED
Calculate kNN-DTWD
k Stations' data
Top k
k Stations' data
Training data generating flow
kNN-ED Linear Regression Model
kNN-DTWD Linear Regression Model
Model Evaluation and Comparison

Figure 4. System flow of the prediction model

2.2. Prediction

k-Nearest Neighbour [15] and k-Most-Similar time series using Dynamic Time Warping [16] [17] [18] are used for prediction and analysis of PM2.5 trends.

2.2.1. k-Nearest Neighbour by Euclidean Distance (kNN-ED).

We calculate Euclidean distances between stations by using their geographical coordinates, and for each station's model we choose top-k nearest stations as its candidates for the prediction of its PM2.5 sequence. In Fig.1, we calculate the distances between the target stations, orange station, and its neighbour stations. The stations with the top-5 shortest distances, green stations, are selected as candidates.

2.2.2. k-Nearest Neighbour by DTW Distance (kNN-DTWD).

DTW algorithm compare two sequences to calculate their distances and the errors on shifting and scaling between sequences is minimized. In Fig.2, red sequence is compared with blue sequence. These two sequences are not similar when using Euclidean distance. In contrast, DTW can restore sequence distortion that mapping the data points to corresponding intervals.

Target station's candidates are generated by using DTW to calculate distance between other stations. The similarities are sorted ascendingly and we select the top-k most similar stations as its candidates for prediction of its PM2.5 sequence. We refer to this method as kNN-DTWD. In Fig.3, time series of PM2.5 have shifting and scaling. S7, the purple time series, has delays comparing to S1, the red one, and S2, the orange one, marked at red circle. DTW can reduce the error in similarity in this case.

3. Proposed Method

We apply the technique of spatial-temporal analysis to explore the sequences' delay and the interactions between stations through the information of the past temporal pattern. Next, we analyse the trend of the station measurement for potential factors. We consider the adjacent stations or stations with similar temporal patterns since they may have high correlations with the target station.

In addition, we propose a model that predicts target station's PM2.5 concentrations using information of adjacent stations or stations with similar temporal patterns. The system flow is shown as Fig.4. First, we collect data from Taiwan Environmental Protection Administration (TWEPA) [19] and pre-process them. After importing processed data into database, we determine the most related stations to the target station according to kNN-ED and kNN-DTWD. We generate training datasets and testing datasets from top-k stations for the model. Lastly, we train the model and compare the performance of predictions.

3.1. kNN-ED geographical relation

Our stations information is retrieved from open data [20], the distance between two stations is calculated from

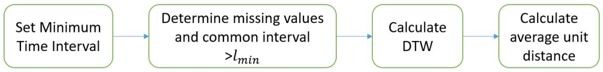

Set Minimum Time Interval
Determine missing values and common interval >l_{min}
Calculate DTW
Calculate average unit distance

Figure 5. Procedure to generate kNN-DTWD similarity

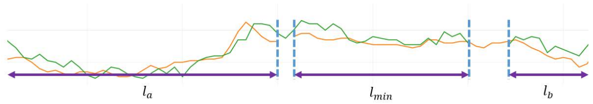


Figure 6. Shortest interval threshold diagram

their coordinates. The distance will be sorted after the calculation. Top 20 of the nearest geographical distance of each target station will be recorded and used for training the prediction model.

3.2. kNN-DTWD temporal similarity

The procedure of generating kNN-DTWD is shown in Fig.5. After the two sequences are selected, both sequences are cut into a plurality of common interval sequences. As the short sequence is not meaningful, we set the shortest interval threshold l_{min} to filter the common interval sequence. The filtered sequences are subjected to DTW calculations and all calculated values are converted into unit similarity through the conversion formula. Since the conventional DTW computes the similarity between time series, coherent data are used in order to eliminate the problem of time shift and scaling between time series. However, the original DTW cannot calculate the similarity between time series with missing data. We tackle this problem by: (i) define the upper bound of the shortest time interval threshold, (ii) average unit distance.

When two intervals of time series have non-missing values simultaneously and the common interval is greater than l_{min}, this interval will be considered to calculate the similarity by DTW. As shown in Fig.6, common interval l_a is larger than l_{min}, and it will be included in the calculation. In contrast, common interval l_b is less than l_{min}, and it will be ignored in the calculation. Choosing different l_{min} will affect the continuity of the time series, and the results of similarity sorting. The ignored interval may induce some loss of information, but can effectively remove most of the errors caused by the noise.

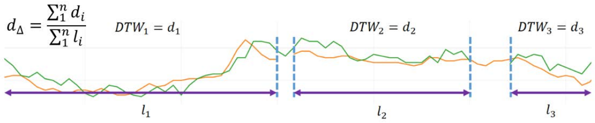


Figure 7. Example for calculating average unit distance

The average unit distance is defined as

$$d_\Delta = \frac{\sum_1^n d_i}{\sum_1^n l_i} \tag{1}$$

In Eq.(1), d_i is the distance in a common interval and l_i is the length of that interval. The equation sums all the distances of common intervals and divide it with the total sum of lengths of the intervals, which averages the distances to a unit length. The average unit distance can combine multiple fragment sequences in order to facilitate the identification of the final similarity. In Fig.7, three interval distances d_1, d_2, d_3 are calculated by DTW and common interval lengths l_1, l_2, l_3 are recorded for the calculation of the average unit distance. In kNN-DTWD model, each station will record the top 20 stations that are most similar to it.

4. Results and Discussions

TWEPA monitors the air quality throughout Taiwan regularly. TWEPA records PM2.5 data from 2008. Our data were collected from TWEPA stations with hourly sampling frequency, i.e., total 17520 instances within 2014 and 2015. For prediction, this study uses the data of 2014 as the training set and the data of 2015 is used as the testing set. We use the data of 2014 for finding the top k candidates by kNN-DTWD, and l_{min} is set as 6 hours.

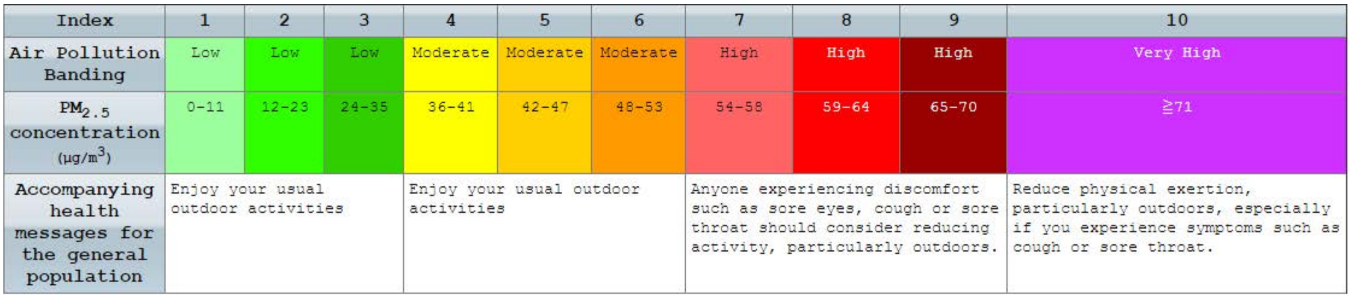
4.1. Probability Map of PM2.5

Spatial pattern of air quality was various in Taiwan. Fig.8 shows the spatial probability map when the daily PM2.5 is over $35\mu g/m^3$. In average, the probability in north Taiwan is less than 0.25. The PM2.5 in southern Taiwan is more serious than northern Taiwan. Fig.9 shows the temporal patterns of average daily PM2.5 at three stations in Taipei. The PM2.5 at SinDian and Linkou is higher than SanChong. In local scale, the PM2.5 concentrations are low at 10PM to 6AM (night time) and the PM2.5 concentrations are high at 8AM to 2PM (day time). The PM2.5 changes seasonally. In winter, the concentration of PM2.5 is more serious than it is in summer, especially in January, based on the observation. Several factors such as wind direction and the location of pollution affect the spatial-temporal patterns of PM2.5 concentrations.

4.2. Linear Regression Model with Proposed Methods

In Fig.10, lime time series is actual measured values, red time series is the prediction result of kNN-DTWD, and the grey time series is the prediction result of kNN-ED. The prediction results of two methods are similar to the ground truth and grasp the global trend. Overall, kNN-DTWD is better than kNN-ED.

As we can see in Fig.11 and Table 1, the accuracy of the current model, whether using kNN-ED and kNN-DTWD, can accurately predict the PM2.5 classification



Index	1	2	3	4	5	6	7	8	9	10
Air Pollution Banding	Low	Low	Low	Moderate	Moderate	Moderate	High	High	High	Very High
$PM_{2.5}$ concentration ($\mu g/m^3$)	0-11	12-23	24-35	36-41	42-47	48-53	54-58	59-64	65-70	≧71
Accompanying health messages for the general population	Enjoy your usual outdoor activities			Enjoy your usual outdoor activities			Anyone experiencing discomfort such as sore eyes, cough or sore throat should consider reducing activity, particularly outdoors.			Reduce physical exertion, particularly outdoors, especially if you experience symptoms such as cough or sore throat.

Table 1 EPA PM2.5 index [21]

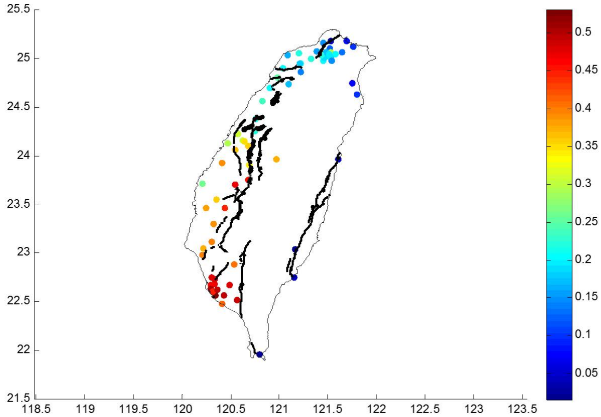


Figure 8. Probability map of the daily PM2.5 over $35\mu g/m^3$

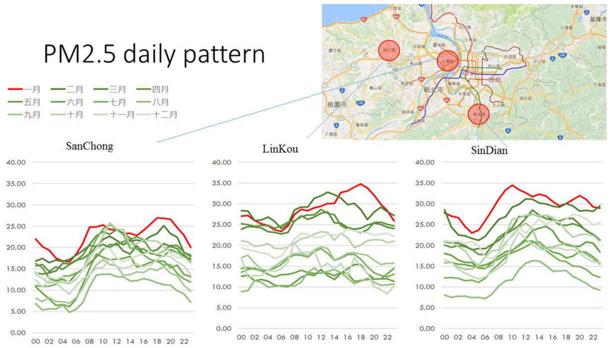


Figure 9. PM2.5 temporal patterns at SanChong, LinKou and SinDian

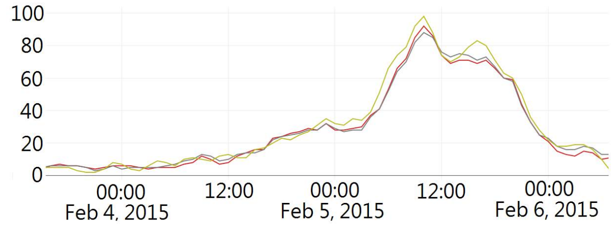


Figure 10. Prediction results of kNN-ED(grey) and kNN-DTWD(red) (units: $\mu g/m^3$)

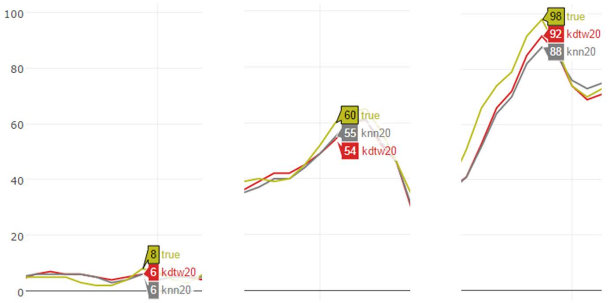


Figure 11. Prediction results of kNN-ED(grey) and kNN-DTWD(red) (units: $\mu g/m^3$) in low, mid, high grade.

section defined by TWEPA. Numerical prediction will be affected by various factors. We will further study the results in accordance with TWEPA provided area partitions, in which the vertical axis unit is $\mu g/m^3$ and horizontal axis is k (in units). As shown in Fig.12 and Fig.15, with addition of more similar stations, the more accurate prediction achieves. kNN-DTWD performs better than kNN-ED, which shows that central area stations have lower impact with adjacent stations. Fig.13 and Fig.16 show the results of northeastern area, kNN-DTWD performs better than kNN-ED. This result is line with expectations. If we use geographical distance, the candidates will cross the Snow Mountain Range to select the stations in northern area. Fig.14 and Fig.17 show the result of eastern area. The prediction results are better when quantity of adjacent stations is low. This phenomenon explains that kNN-ED included southern area stations. In addition, eastern area stations will not be affected by southern area, and the average error and the standard deviation are lower than other area, which indicating that variability is smaller than other area.

The experimental results show that the data of 2014 can be used to predict the values of PM2.5 in 2015 successfully and there are only slight deviations, which are related to the deterioration of greenhouse effect. Our proposed model can predict the measured values of PM2.5 accurately, and we can eliminate the annual growth error caused by the deterioration of the greenhouse effect on a single station through the performance of multiple stations. We have considered the effect

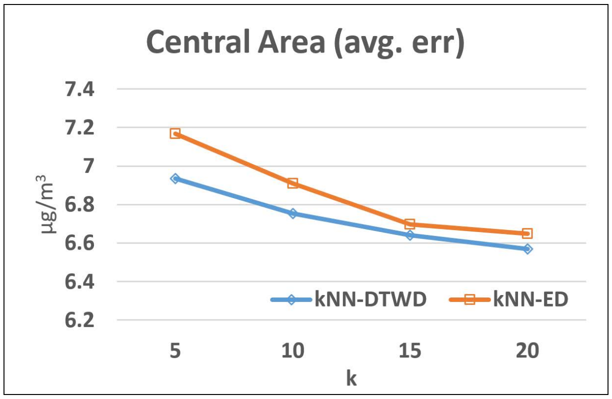


Figure 12. Average error of EPA in Central Area

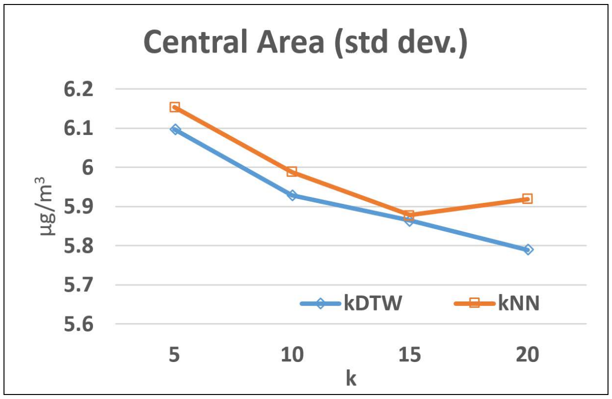


Figure 15. Standard deviation of EPA in Central Area

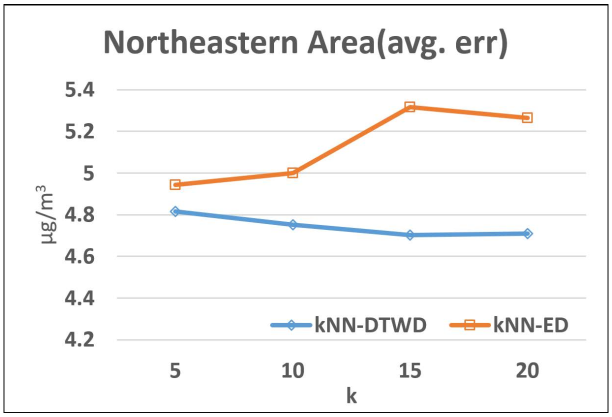


Figure 13. Average error of EPA in Northeastern Area

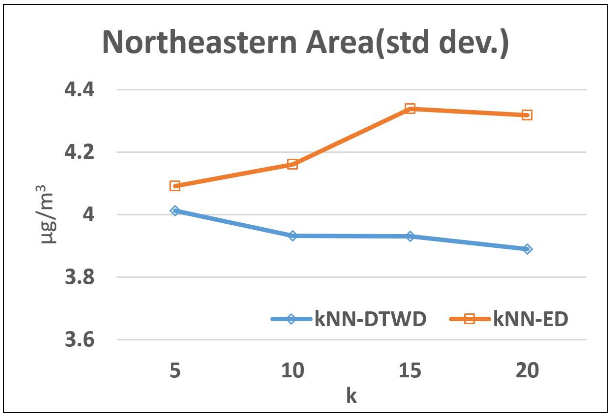


Figure 16. Standard deviation of EPA in Northeastern Area

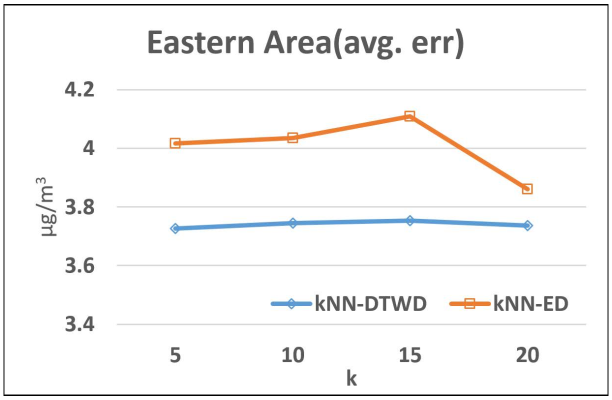


Figure 14. Average error of EPA in Eastern Area

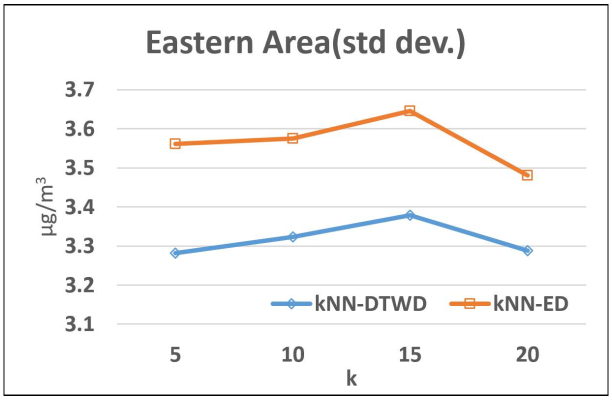


Figure 17. Standard deviation of EPA in Eastern Area

of periodic coherence in the training model. It is confirmed that the station performance is periodic if the average error and the standard deviation are within acceptable limits. With the above deduction and experimental results, the values of PM2.5 is indeed cyclical. The results also show that kNN-DTWD outperforms kNN-ED model. The prediction results are affected by the relevance of stations, which confirms that there are spatial-temporal relations between stations.

The results of kNN-DTWD still have room for improvement, we believe that kNN-DTWD may have error due to the missing data, and this can be solved by data filling technique using historical data. In addition, most of the similar stations selected by kNN-DTWD do not fully contain kNN-ED selected stations, mainly because kNN-DTWD is not affected by geographical distance. In future, we will first consider the results of kNN-ED and calculate similarity with them by using kNN-DTWD, which will be more effective to shorten the training time and improve accuracy.

5. Conclusions

This study proposed a spatial-temporal air quality analysis model to monitor and predict PM2.5 values. Our proposed model predicts a stations PM2.5 concentrations using information of adjacent stations or stations with similar temporal patterns and eliminate the annual growth error caused by the deterioration of the greenhouse effect. The proposed method kNN-DTWD mostly outperforms kNN-ED. From the experimental results, we can learn that the spatial distance in the air pollution is a certain impact, but the temporal relations can help to model changes. In addition, the values of PM2.5 is indeed cyclical and southern Taiwan is more serious than northern Taiwan. Eastern Taiwan has the best air quality all year.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research under Contract No. MOST 105-EPA-F-007-004. We are grateful to the Taiwan Environmental Protection Administration (TWEPA) for providing the monitoring data used in this study.

References

[1] H.-J. Chu, C.-Y. Lin, C.-J. Liau, and Y.-M. Kuo, "Identifying controlling factors of ground-level ozone levels over southwestern taiwan using a decision tree," *Atmospheric environment*, vol. 60, pp. 142–152, 2012.

[2] Y.-S. Choi, C.-H. Ho, D. Chen, Y.-H. Noh, and C.-K. Song, "Spectral analysis of weekly variation in pm10 mass concentration and meteorological conditions over china," *Atmospheric Environment*, vol. 42, no. 4, pp. 655–666, 2008.

[3] A. P. Tai, L. J. Mickley, and D. J. Jacob, "Correlations between fine particulate matter (pm 2.5) and meteorological variables in the united states: Implications for the sensitivity of pm 2.5 to climate change," *Atmospheric Environment*, vol. 44, no. 32, pp. 3976–3984, 2010.

[4] C.-Y. Lin, S. C. Liu, C. C.-K. Chou, S.-J. Huang, C.-M. Liu, C.-H. Kuo, and C.-Y. Young, "Long-range transport of aerosols and their impact on the air quality of taiwan," *Atmospheric Environment*, vol. 39, no. 33, pp. 6066–6076, 2005.

[5] C.-Y. Lin, Z. Wang, W.-N. Chen, S.-Y. Chang, C. C. Chou, N. Sugimoto, and X. Zhao, "Long-range transport of asian dust and air pollutants to taiwan: observed evidence and model simulation," *Atmospheric Chemistry and Physics*, vol. 7, no. 2, pp. 423–434, 2007.

[6] C.-M. Liu, C.-Y. Young, and Y.-C. Lee, "Influence of asian dust storms on air quality in taiwan," *Science of the Total Environment*, vol. 368, no. 2, pp. 884–897, 2006.

[7] X.-K. Wang and W.-Z. Lu, "Seasonal variation of air pollution index: Hong kong case study," *Chemosphere*, vol. 63, no. 8, pp. 1261–1272, 2006.

[8] C.-S. Horng, C.-A. Huh, K.-H. Chen, P.-R. Huang, K.-H. Hsiung, and H.-L. Lin, "Air pollution history elucidated from anthropogenic spherules and their magnetic signatures in marine sediments offshore of southwestern taiwan," *Journal of Marine Systems*, vol. 76, no. 4, pp. 468–478, 2009.

[9] Y. Hwa-Lung and W. Chih-Hsin, "Retrospective prediction of intraurban spatiotemporal distribution of pm2. 5 in taipei," *Atmospheric Environment*, vol. 44, no. 25, pp. 3053–3065, 2010.

[10] B. S. Beckerman, M. Jerrett, M. Serre, R. V. Martin, S.-J. Lee, A. Van Donkelaar, Z. Ross, J. Su, and R. T. Burnett, "A hybrid approach to estimating national scale spatiotemporal variability of pm2. 5 in the contiguous united states," *Environmental science & technology*, vol. 47, no. 13, pp. 7233–7241, 2013.

[11] H.-J. Chu, H.-L. Yu, and Y.-M. Kuo, "Identifying spatial mixture distributions of pm2.5 and pm10 in taiwan during and after a dust storm," *Atmospheric environment*, vol. 54, pp. 728–737, 2012.

[12] H.-W. Chen, C.-T. Tsai, C.-W. She, Y.-C. Lin, and C.-F. Chiang, "Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach," *Chemosphere*, vol. 81, no. 10, pp. 1358–1367, 2010.

[13] A. Kurt and A. B. Oktay, "Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986–7992, 2010.

[14] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1436–1444.

[15] E. Fix and J. L. Hodges Jr, "Discriminatory analysis-nonparametric discrimination: consistency properties," DTIC Document, Tech. Rep., 1951.

[16] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10115-004-0154-9

[17] A. W.-C. Fu, E. Keogh, L. Y. Lau, C. A. Ratanamahatana, and R. C.-W. Wong, "Scaling and time warping in time series querying," *The VLDB JournalThe International Journal on Very Large Data Bases*, vol. 17, no. 4, pp. 899–921, 2008.

[18] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 262–270.

[19] TWEPA. Air quality index historical data. [Online]. Available: http://taqm.epa.gov.tw/taqm/tw/YearlyDataDownload.aspx

[20] Government open data : Air quality stations info. [Online]. Available: http://data.gov.tw/node/6075

[21] TWEPA. PM2.5 index. [Online]. Available: http://taqm.epa.gov.tw/taqm/en/fpmi.aspx