FINE-SCALE MAPPING OF PARTICULATE MATTER USING LANDSAT IMAGERY AND LOW-COST SENSOR DATA FROM PURPLEAIR: A CASE STUDY OF LOS ANGELES

Shengjie Kris Liu, Siqin Wang

Spatial Sciences Institute, Dornsife College of Letters, Arts and Sciences,
University of Southern California, Los Angeles, CA, USA

ABSTRACT

Fine-scale monitoring of particulate matter (PM) is essential to assess the patterns and sources of air pollution. However, to this date, its usage is still limited. In this study, we explore the usage of Landsat imagery for PM mapping at 30 m spatial resolution, where air pollutants at the street level can be distinguished. With in-situ data from the low-cost sensors of the PurpleAir network, we developed a multitask neural network that can simultaneously estimate all three metrics of PM, i.e., PM1, PM2.5, and PM10. With 936 monitoring stations from 25 dates, a total of 15,250 valid observations were used to construct the model. With external validation, results show that the model can achieve R^2 of 0.767, 0.803, and 0.799 in estimating PM1, PM2.5, and PM10, respectively, at the street level. Using one multitask neural network to predict PM1, PM2.5, and PM10 improves the efficiency of real-life air pollutant monitoring.

***Index Terms*—** Air pollution, particulate matter, PurpleAir, multitask learning, low-cost sensor

1. INTRODUCTION

Particulate matter (PM) is a family of major air pollutants. Depending on the diameter, PM can be categorized as PM1, PM2.5, or PM10. PM less than 10 micrometers can penetrate into our lungs and even bloodstream, imposing a great risk to human health [1].

Estimation and prediction of PM in space and time can help us understand the patterns and sources of air pollution. Predicting PM in time is important to provide information for people to adjust their daily behaviors. Tremendous efforts have been devoted to predicting PM. For example, the city of Los Angeles hosted a project named "Predicting What We Breathe" to predict future air quality within 10 days for the 24 monitoring stations in Los Angeles County. They used a graph convolutional network with a Long Short-Term Memory (LSTM) component and meteorological properties to achieve the prediction, with 91% accuracy for the first frame prediction [2]. Fine-scale mapping of PM in space locates highly polluted areas, estimating people's exposure to PM, allowing for the adoption of policies, and advising people to avoid contaminated areas. However, due to the sparse distribution of monitoring stations (e.g., 24 stations in Los Angeles County), prediction based on monitoring stations provides little knowledge in terms of the heterogeneity of PM concentration and exposure.

Several methods have been developed to enhance the spatial monitoring of air pollutants. For example, ordinary and universal kriging have been used to map 1-km NO2, PM10, and O3 concentrations over Europe, with an accuracy of R^2 between 0.45 and 0.70 [3]. Another family of algorithms is the land use regression model, in which air pollution concentration is estimated through predictors such as distance to the road networks [4]. Recently, land use regression models were enhanced with machine learning. Wong et al. (2021) [5] used XGBoost to construct a land use regression model for estimating PM2.5, explaining 58%-94% of the seasonal variation with in-situ validation. While kriging and land use regression models are popular, they are static in the sense that land use is consistent over a period of time, and the models cannot produce spatiotemporal estimates without adjustments.

For continuous spatiotemporal monitoring, some studies proposed to integrate land use regression models with long-term satellite aerosol optical depth (AOD) products [6, 7, 8]. AOD is a combined measure of the amount of aerosol present in the atmosphere that is known to produce three types of scattering (Rayleigh, Mie, and Nonselective). The occurrence of three types of scattering depends on the size of the particles. Mie scattering is caused by particles of the same size or larger than the wavelength of light. For PM1, PM2.5, and PM10, their wavelengths are covered by the common wavelength of satellite data. Integrating Mie scattering into models has helped facilitate real-time PM2.5 monitoring in several places [9, 10, 11]. While AOD can be used to predict PM, a more direct way is to estimate PM from top-of-atmosphere (TOA) reflectance [12].

To make full usage of the satellite data, their combined use with in-situ monitoring data is critical, but it has been limited due to the sparse distribution of monitoring stations. The development of low-cost sensors and the establishment of

E-mail: liusheng@usc.edu (SKL). This work was supported by the Southern California Environmental Health Sciences Center (Grant No. P30ES007048).

Table 1. Acquisition time of Landsat imagery, number of available stations, and median PM1, PM2.5, and PM10 ($\mu g/m^3$). Local time differs due to daylight saving time.

Date	Local Time	# station	PM1	PM2.5	PM10
2021-02-28	10:28:19	537	0.21	0.88	1.42
2021-03-16	11:28:10	562	2.56	3.77	4.30
2021-05-03	11:27:50	581	6.74	10.65	12.98
2021-06-04	11:28:06	594	12.95	21.47	23.40
2021-06-20	11:28:11	592	12.47	19.85	21.88
2021-07-06	11:28:14	611	10.57	16.77	18.17
2021-07-22	11:28:18	621	7.34	12.22	14.39
2021-08-07	11:28:26	627	14.53	24.47	26.70
2021-08-23	11:28:31	636	13.50	23.84	26.43
2021-09-08	11:28:35	651	18.90	35.14	39.32
2021-09-24	11:28:38	647	15.30	23.78	26.14
2021-10-10	11:28:44	635	3.71	5.47	6.36
2021-10-26	11:28:46	640	2.59	4.26	5.12
2021-11-11	10:28:42	635	0.76	1.36	1.67
2021-11-23	10:24:37	504	4.48	6.48	7.12
2021-11-27	10:28:41	641	1.79	2.96	3.43
2021-12-05	10:28:43	647	31.86	63.33	76.19
2022-01-06	10:28:36	662	15.44	23.29	24.82
2022-01-14	10:28:33	689	11.20	16.88	18.37
2022-01-22	10:28:36	667	0.00	0.00	0.05
2022-01-30	10:28:28	697	6.58	9.97	10.82
2022-02-07	10:28:35	688	0.36	0.86	1.15
2022-03-11	10:28:21	715	0.61	1.23	1.54
2022-04-04	11:28:05	709	14.4	25.50	28.42
2022-04-12	11:28:12	701	0.00	0.22	0.45

low-cost monitoring network (e.g., PurpleAir) provide an unprecedented opportunity to estimate air pollutants at the street level [13, 14, 15]. There were over 20,000 stations within the North America PurpleAir network. In this study, we aim to develop a model based on a large amount of PurpleAir PM observations for fine-scale mapping of PM. Due to long sunshine periods (fewer clouds) and prevalence of PurpleAir stations, we selected Los Angeles as a case study and used Landsat with the low-cost sensor data to predict PM.

2. DATA AND METHOD

2.1. Landsat imagery

A total of 25 Landsat 8/9 scenes between 2021-03-14 and 2022-04-12 across Los Angeles were collected, as shown in Table 1. These scenes were with a cloud coverage of less than 20%. The data were first converted to TOA reflectance with adjustment of the sun elevation. We used eight multispectral bands (B1, B2, B3, B4, B5, B6, B7, B9) and two thermal bands (B10, B11) to build the model. A diameter of 30 pixels (900 m) was used to extract image patches corresponding to each monitoring station. In model training, a TOA reflectance threshold of 0.4 of the B1 band was used to exclude clouds and anomalies, resulting to 15,250 sample points.

2.2. Low-cost sensor data from PurpleAir

PurpleAir is a network of low-cost stations equipped with two laser scattering particle sensors and a pressure-temperature-humidity sensor [16]. It measures PM1, PM2.5, and PM10. In this study, we downloaded data using the PurpleAir API[1], which wraps historical PurpleAir data from the ThingSpeak API. A total of 16,738 valid outdoor stations were available in North America as of 2022-04-26. A total of 936 stations were on record in Los Angeles. The sensor takes data every two minutes. Some data had extra-large values that seemed to be reading errors (a constant value of about 5000 $\mu g/m^3$ for all PM readings) and were discarded in the analysis.

The number of stations with valid records of each date is shown in Table 1 along with the Landsat imagery. Of the 25 dates, the highest median station PM2.5 was 63.33 $\mu g/m^3$ on 2021-12-05, and the lowest median was 0.00 $\mu g/m^3$ on 2022-01-22. These are also the dates for the highest and lowest median station PM10. The value of PM10 is, in general, larger than PM2.5, and PM2.5 larger than PM1.

2.3. Network design

The multitask network (Fig. 1) consists of multi-scale input filters, residual blocks, and the Squeeze-and-Excitation module [17, 18]. The proposed network here did not directly adopt the referred components but integrated the ideas of multi-scale filtering [19], residual learning [20], and Squeeze-and-Excitation [21], all winners of ImageNet and now basic components in modern neural networks.

Specifically, the multi-scale input filters enlarge the network's perceptual fields before going through the nonlinear residual blocks. We assumed the first input filters could capture some useful low-level information from the imagery before going through the nonlinear transformations. The residual blocks were adopted to facilitate the training of the deep network and introduce nonlinearity. Within the residual blocks, batch normalization and ReLU are used. Batch normalization is known to accelerate network training and introduce variations that can prevent overfitting, whereas ReLU is the standard activation function for nonlinearity. The Squeeze-and-Excitation Block serves as what can be referred to as an attention mechanism to enhance the extracted channel information that the network regards as important in training (self-adapted). We added the MaxPooling layer between each residual block to reduce feature dimension and capture high-frequent information. After feature extraction, we use a global average pooling to produce a 256-dimension feature array with extracted features useful for PM estimation. Finally, we use a simple fully connected layer for regression with the ReLU function (to ensure non-negative outputs) to co-estimate three PM metrics. The network contains 422,543 parameters (421,772 in feature extractors; 771 in regressors).

[1] https://zenodo.org/badge/latestdoi/163108208

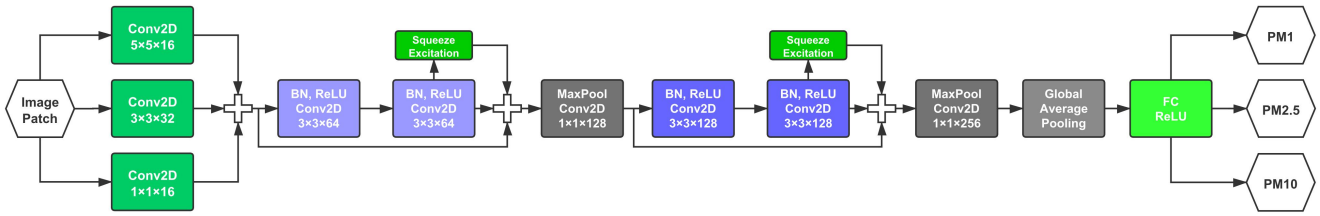


Fig. 1. Multitask network in this study.

3. RESULTS AND DISCUSSION

Fig. 2 shows the scatter plots between observed and estimated PM1, PM2.5, and PM10 values on the testing set. The model achieved R^2 of 0.767, 0.803, and 0.799 in predicting PM1, PM2.5, and PM10, respectively.

Table 2 shows the performance in terms of R^2 and mse of the multitask network, in comparison with a single-task network. The proposed multitask network achieved better results in predicting PM1 and competitive results on predicting PM2.5 and PM10. The multitask network improved the performance of PM1 prediction where a single-task network performed the worst. By training the multitask network with all samples of PM, the network generalized well on tasks that single-task networks failed.

Table 2. Performance of single-task and multitask networks.

	Network	PM1	PM2.5	PM10
R^2	Single-task	0.7640	0.8061	0.8169
	Multitask	0.7667	0.8027	0.7987
mse	Single-task	2.8064	4.4950	5.0612
	Multitask	2.7403	4.5039	5.1277

Fig. 3 shows the mapping of PM2.5 over eight dates in Los Angeles. It is clear to see the heterogeneity of spatiotemporal PM2.5 over time. As for spatial heterogeneity, three places (South Central Los Angeles, freeway 605, and LAX airport) appeared to have higher PM2.5 concentration compared to other areas. Specifically, on 2021-10-10, LAX appear to have higher PM2.5 concentration than other areas. On 2022-01-30, Culver City appeared to have higher PM2.5 concentration in areas with gas plants. The model successfully captured the spatiotemporal heterogeneity of PM2.5 concentration over the eight dates in Los Angeles.

4. CONCLUSION

This study used Landsat imagery and low-cost sensor data from PurpleAir for fine-scale mapping of PM1, PM2.5, or PM10., simultaneously, using a multitask network. The multitask network achieved R^2 of 0.767, 0.803, and 0.799 in estimating PM1, PM2.5, and PM10, respectively. The performance is competitive with that of three single-task networks. In the future, an adaptive training strategy should be adopted to fully utilize the potential of multitask networks.

5. REFERENCES

[1] Anderson et al., "Clearing the air: a review of the effects of particulate matter air pollution on human health," *J. Med. Toxicol.*, vol. 8, no. 2, pp. 166–175, 2012.

[2] Pratyush Muthukumar et al., "Predicting pm2. 5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data," *Air Qual. Atmos. Health*, pp. 1–14, 2021.

[3] Rob Beelen et al., "Mapping of background air pollution at a fine spatial scale across the european union," *Sci. Total Environ.*, vol. 407, no. 6, pp. 1852–1867, 2009.

[4] DK Moore et al., "A land use regression model for predicting ambient fine particulate matter across los angeles, ca," *J. Environ. Monit.*, vol. 9, no. 3, pp. 246–252, 2007.

[5] Pei-Yi Wong et al., "Using a land use regression model with machine learning to estimate ground level pm2. 5," *Environ. Pollut.*, vol. 277, pp. 116846, 2021.

[6] Hyung Joo Lee et al., "Enhancing the applicability of satellite remote sensing for pm2. 5 estimation using modis deep blue aod and land use regression in california, united states," *Environ. Sci. Technol.*, vol. 50, no. 12, pp. 6546–6555, 2016.

[7] Chih-Da Wu et al., "Land-use regression with long-term satellite-based greenness index and culture-specific sources to model pm2. 5 spatial-temporal variability," *Environ. Pollut.*, vol. 224, pp. 148–157, 2017.

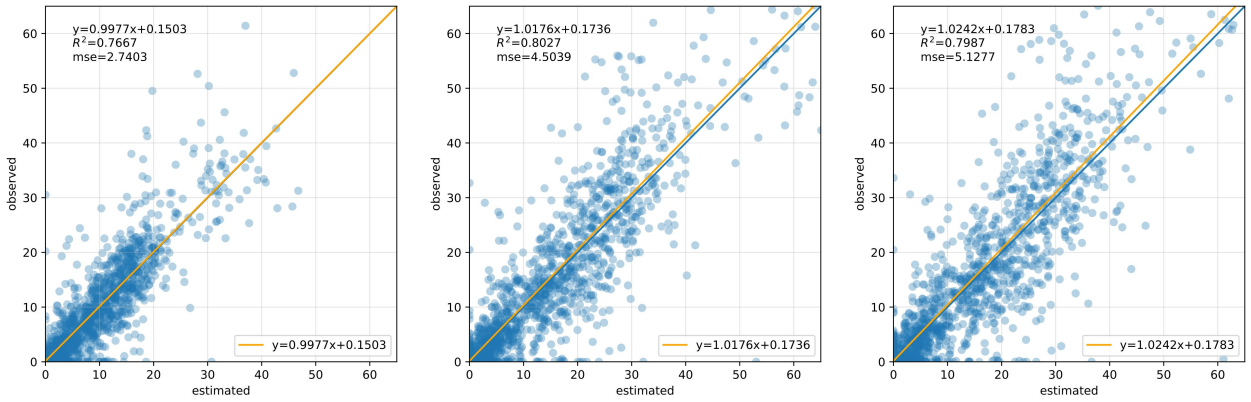
[8] Chih-Da Wu et al., "A hybrid kriging/land-use regression model to assess pm2. 5 spatial-temporal variability," *Sci. Total Environ.*, vol. 645, pp. 1456–1464, 2018.

[9] Constantinos Sioutas et al., "Field evaluation of a modified dataram mie scattering monitor for real-time pm2. 5 mass concentration measurements," *Atmos. Environ.*, vol. 34, no. 28, pp. 4829–4838, 2000.

[10] RS Gao et al., "A light-weight, high-sensitivity particle spectrometer for pm2. 5 aerosol measurements," *Aerosol. Sci. Technol.*, vol. 50, no. 1, pp. 88–99, 2016.

[11] Rohan Jayaratne et al., "Low-cost pm2. 5 sensors: An assessment of their suitability for various applications," *Aerosol and Air Quality Research*, vol. 20, no. 3, pp. 520–532, 2020.

[12] Jianhua Yin, Feiyue Mao, Lin Zang, Jiangping Chen, Xin Lu, and Jia Hong, "Retrieving pm2. 5 with high spatio-temporal



(a) PM1 (b) PM2.5 (c) PM10

Fig. 2. Scatter plots between observed and estimated PM values on the testing set.

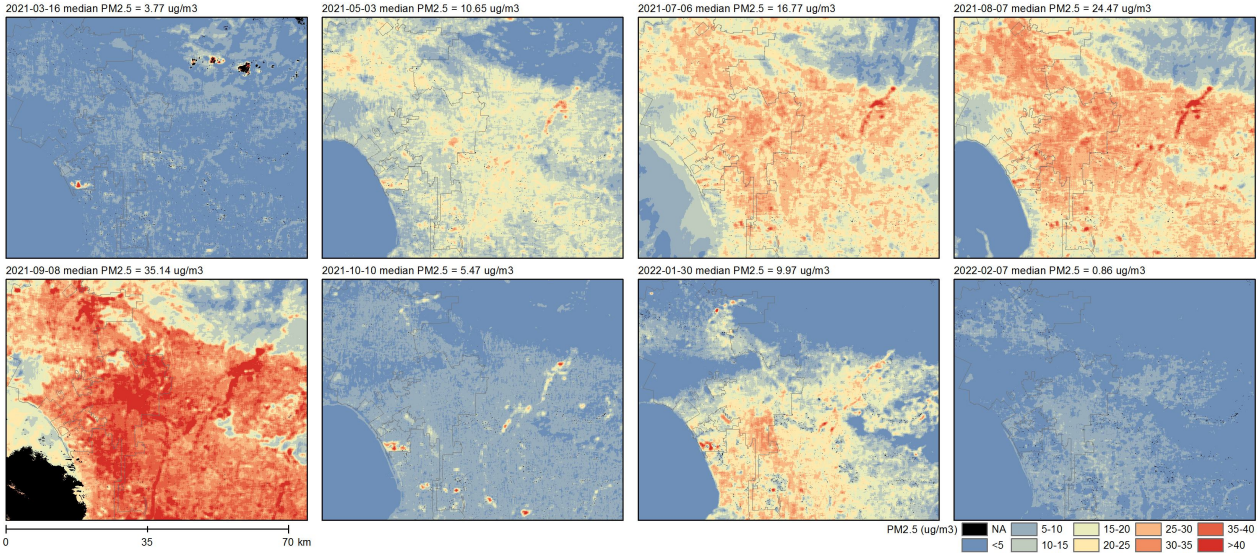


Fig. 3. Mapping of PM2.5 over eight dates in Los Angeles.

coverage by toa reflectance of himawari-8," *Atmospheric Pollution Research*, vol. 12, no. 4, pp. 14–20, 2021.

[13] Lidia Morawska et al., "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?," *Environ. Int.*, vol. 116, pp. 286–299, 2018.

[14] P Gupta et al., "Impact of california fires on local and regional air quality: the role of a low-cost sensor network and satellite observations," *GeoHealth*, vol. 2, no. 6, pp. 172–181, 2018.

[15] Iasonas Stavroulas et al., "Field evaluation of low-cost PM sensors (Purple Air PA-II) under variable urban air quality conditions, in Greece," *Atmosphere*, vol. 11, no. 9, pp. 926, 2020.

[16] Karoline K Barkjohn et al., "Development and application of a united states-wide correction for pm 2.5 data collected with the purpleair sensor," *Atmos. Meas. Tech.*, vol. 14, no. 6, pp. 4617–4637, 2021.

[17] Liu and Shi, "Multitask deep learning with spectral knowledge for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 12, pp. 2110–2114, 2020.

[18] Shengjie Liu, Qian Shi, and Liangpei Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5085–5102, 2021.

[19] Christian Szegedy et al., "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *ECCV 2016*. Springer, pp. 630–645.

[21] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR 2018*, pp. 7132–7141.