Non-Linear Retrieval of PM2.5 from MODIS AOD: A Comparative Study of Ensemble vs. Linear Architectures in Data-Scarce Urban Environments

Yuval Levental
RIT Center for Imaging Science
54 Lomb Memorial Dr, Rochester, NY 14623
yhl3051@rit.edu

Abstract

Accurate retrieval of Particulate Matter (PM2.5) from satellite Aerosol Optical Depth (AOD) is critical for monitoring air quality in regions lacking ground sensor infrastructure. While recent deep learning approaches achieve state-of-the-art results, they often require massive, clean datasets that are unavailable in developing urban centers. In this study, we evaluate the robustness of ensemble learning architectures (XGBoost, Random Forest) against traditional linear baselines (Lasso, ElasticNet) for retrieving PM2.5 from MODIS Terra/Aqua AOD products. We focus on a comparative analysis between a data-rich environment (Los Angeles, USA) and a data-scarce environment (Addis Ababa, Ethiopia). Our results demonstrate that while linear methods degrade significantly in complex aerosol environments ($R^2 < 0.1$ for linear regression in Los Angeles), ensemble methods maintain robust performance ($R^2 \approx 0.5$, $RMSE < 9.0$) by effectively modeling non-linear interactions between AOD and ground-level particulates. This suggests that lightweight ensemble architectures offer a viable, low-resource alternative to deep learning for establishing air quality baselines in under-monitored regions.

1. Introduction

Air pollution, specifically fine particulate matter with a diameter of less than 2.5 micrometers (PM2.5), represents a significant global health risk. Unlike coarser particulates, PM2.5 can penetrate deep into the respiratory system and enter the bloodstream. A major challenge in global epidemiology is the lack of consistent ground-level monitoring, particularly in developing nations. Satellite remote sensing offers a solution: Aerosol Optical Depth (AOD) products from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) instruments (Terra and Aqua) provide daily global coverage [5].

However, the relationship between column-integrated AOD and ground-level PM2.5 is complex and non-linear, influenced by vertical aerosol distribution, humidity, and surface reflectance. Traditional approaches have relied on linear regression or Chemical Transport Models (CTMs), but these often fail to capture local spatial variability.

Recent advancements in 2024 and 2025 have seen a surge in Deep Learning (DL) applications, such as Convolutional Neural Networks (CNNs) and LSTMs, to model these non-linearities [3]. While powerful, these models are computationally expensive and prone to overfitting when ground-truth training data is sparse [1].

In this work, we argue that Ensemble Learning methods—specifically Gradient Boosting (XGBoost) and Random Forests—provide an optimal balance of accuracy and data efficiency for regions with limited infrastructure. We validate this by benchmarking linear versus ensemble architectures on two distinct datasets: the dense sensor network of Los Angeles, USA, and the sparse, noisy network of Addis Ababa, Ethiopia.

2. Related Work

Satellite-based PM2.5 Retrieval: The use of AOD for PM2.5 estimation is well-established. Early works utilized simple linear correlation, but recent reviews highlight the necessity of non-linear modeling to account for meteorological variables [3].

Machine Learning Approaches: With the availability of open datasets like OpenAQ, data-driven approaches have overtaken physical models. While deep learning models like ResNet-LSTM hybrids show promise for regional forecasting [4], they often require ”Big Data” to converge. Conversely, Bagheri et al. (2024) demonstrated that Deep Ensemble Forests could outperform pure deep learning methods in high-resolution mapping tasks where data density is variable [1].

Ensemble Methods: XGBoost [2] has become a de facto standard for tabular data competitions due to its scalability and handling of missing values—a common issue in satellite AOD retrieval due to cloud cover. Our work extends these findings by explicitly testing the ”transfer of robustness” to developing world contexts.

3. Methodology

3.1. Data Acquisition

We utilized AOD data from NASA’s MODIS Collection 6.1 (Terra and Aqua satellites). Terra crosses the equator at 10:30 AM, and Aqua at 1:30 PM, providing two daily observation windows.

- **AOD Data:** We extracted 3km resolution AOD values (Optical_Depth_047) and associated quality flags. We focused on ”AOD1” (closest 10km grid) and ”AOD3” (3×3 average) products.
- **Ground Truth:** PM2.5 concentrations were sourced from the OpenAQ project, covering the period 2016–2019.

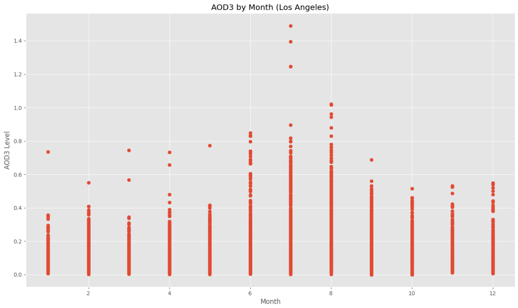

Figure 1. AOD3 levels by month in Los Angeles. As can be seen, aerosol optical depth levels are substantially higher in summer months, correlating with seasonal atmospheric changes.

3.2. Preprocessing & Alignment

AOD and PM2.5 data were temporally aligned (± 1 hour) and spatially matched using nearest-neighbor rounding to 3 decimal places. Missing values (flagged as -999) were removed. As shown in Figure 2, PM2.5 levels exhibit seasonal trends similar to AOD, though the relationship is non-linear.

3.3. Model Architectures

We evaluated two classes of algorithms using the `scikit-learn` framework [6]:

1. Linear Baselines:

- **Ordinary Least Squares (OLS):** Minimizes $\min ||Xw - y||^2$.
- **Lasso/ElasticNet:** Introduces L_1 and L_2 regularization to prevent overfitting on noisy AOD signals.

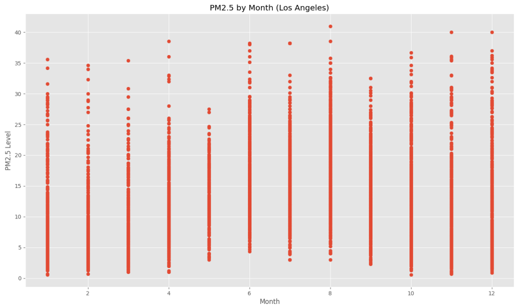

Figure 2. PM2.5 levels by month in Los Angeles. Levels peak in summer months, mirroring the AOD trends observed in Figure 1.

2. Ensemble Methods:

- **Random Forest:** Constructs a multitude of decision trees at training time to reduce variance.
- **XGBoost (Extreme Gradient Boosting):** Uses a gradient descent algorithm to minimize the loss function by adding weak learners sequentially [2].

4. Results

4.1. Addis Ababa (Data Scarce)

In the developing region of Addis Ababa, linear methods largely failed to find a signal, likely due to high noise levels and unmodeled dust events. However, ensemble methods successfully retrieved meaningful correlations. As shown in Figure 3, the correlation improves at higher AOD values, which ensemble methods capture effectively.

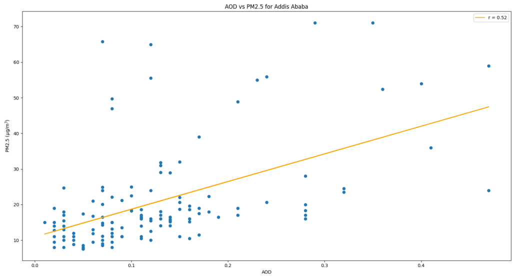

Figure 3. AOD vs PM2.5 for Addis Ababa. The correlation is stronger for higher AOD values, a non-linearity that linear models fail to capture.

Table 1 presents the quantitative results. The Decision Forest achieved an R^2 of 0.52, drastically outperforming the Neural Network baseline ($R^2 = 0.008$) which likely failed to converge due to insufficient data.

Algorithm	**Score (R^2)**	**RMSE**
Linear Regression	0.38	10.09
Neural Network (MLP)	0.008	12.77
Lasso Regression	0.045	12.52
Decision Forest	**0.52**	**8.86**
Extra Trees	0.51	8.97
XGBoost	0.47	9.33

Table 1. Performance metrics for Addis Ababa. Ensemble methods significantly outperform linear and simple neural baselines.

4.2. Los Angeles (Data Rich)

In Los Angeles, the complexity of the urban aerosol layer (traffic emissions + marine layer) rendered simple linear models ineffective. Ensemble methods, capable of modeling these non-linear decision boundaries, showed superior performance.

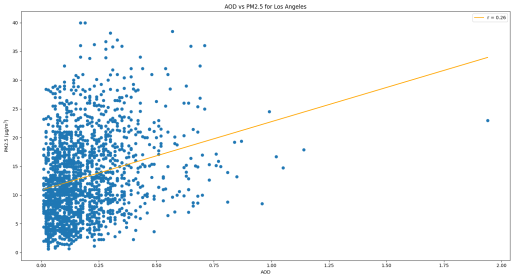

Figure 4. AOD vs PM2.5 for Los Angeles. The data spread indicates complex atmospheric interactions requiring non-linear modeling.

Table 2 highlights that XGBoost achieved the highest accuracy ($RMSE = 4.71$).

Algorithm	**Score (R^2)**	**RMSE**
Linear Regression	0.09	6.35
ElasticNet	0.012	6.61
XGBoost	**0.50**	**4.71**
Decision Forest	0.48	4.79
Extra Trees	0.476	4.81

Table 2. Performance metrics for Los Angeles. XGBoost achieves the highest accuracy, minimizing RMSE to 4.71.

5. Discussion

Our results highlight a critical "robustness gap." While Deep Learning is the current trend [4], simple Neural Networks (MLP in Table 1) failed to converge effectively on the smaller Addis Ababa dataset ($R^2 = 0.008$). In contrast, Tree-based ensembles consistently achieved $R^2 \approx 0.5$.

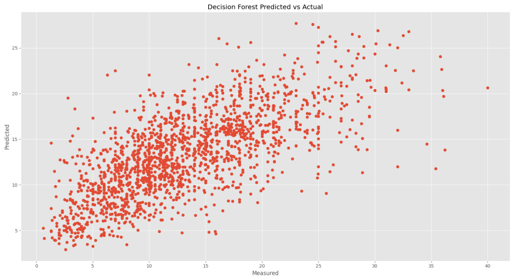

Figure 5. Measured vs predicted PM2.5 values for the Decision Forest model for Los Angeles. The model shows strong predictive capability along the diagonal.

This finding is significant for "Data Equity." It suggests that for many regions in the Global South, where sensor networks are developing, implementing computationally efficient ensemble methods like XGBoost or Extra Trees is a more viable strategy than waiting for datasets large enough to train Transformer-based models.

Future work should integrate meteorological reanalysis data (MERRA-2) to further constrain the AOD-PM2.5 relationship, as humidity and boundary layer height are critical unobserved variables in this study.

6. Conclusion

We compared ensemble and linear methods for predicting PM2.5 from satellite AOD. We found that non-linear ensemble methods (Random Forest, XGBoost) are essential for accurate retrieval in both data-rich and data-scarce urban environments. These methods offer a robust baseline for global air quality monitoring, bridging the gap between satellite observations and ground-level health impacts.

7. Acknowledgements

Many thanks to Dr. John Kerekes of the RIT Center for Imaging Science for guidance on data preparation and analysis, and to Robert Ritz for foundational work on air quality algorithms.

References

[1] Hossein Bagheri. Using deep ensemble forest for high resolution mapping of pm2.5 from modis maiac aod in tehran, iran. *arXiv preprint arXiv:2402.02139*, 2024.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[3] Ahmad Makhdoomi et al. From detection to solution: A review of machine learning in pm 2.5 sensing and sustainable green mitigation approaches (2021–2025). *Processes*, 13(7), 2025.

[4] A. Makhdoomi, M. Sarkhosh, and S. Ziaei. Pm2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Scientific Reports*, 15:8076, 2025.

[5] NASA. Surface-to-air quality mission, 2019. International Space Apps Challenge.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.