Multiple PM Low-Cost Sensors, Multiple Seasons' Data, and Multiple Calibration Models

ABSTRACT

In this study, we combined state-of-the-art data modelling techniques (machine learning [ML] methods) and data from state-of-the-art low-cost particulate matter (PM) sensors (LCSs) to improve the accuracy of LCS-measured $PM_{2.5}$ (PM with aerodynamic diameter less than 2.5 microns) mass concentrations. We collocated nine LCSs and a reference $PM_{2.5}$ instrument for 9 months, covering all local seasons, in Bengaluru, India. Using the collocation data, we evaluated the performance of the LCSs and trained around 170 ML models to reduce the observed bias in the LCS-measured $PM_{2.5}$. The ML models included (i) Decision Tree, (ii) Random Forest (RF), (iii) eXtreme Gradient Boosting, and (iv) Support Vector Regression (SVR). A hold-out validation was performed to assess the model performance. Model performance metrics included (i) coefficient of determination (R^2), (ii) root mean square error (RMSE), (iii) normalised RMSE, and (iv) mean absolute error. We found that the bias in the LCS $PM_{2.5}$ measurements varied across different LCS types (RMSE = 8–29 μg m^{-3}) and that SVR models performed best in correcting the LCS $PM_{2.5}$ measurements. Hyperparameter tuning improved the performance of the ML models (except for RF). The performance of ML models trained with significant predictors (fewer in number than the number of all predictors, chosen based on recursive feature elimination algorithm) was comparable to that of the 'all predictors' trained models (except for RF). The performance of most ML models was better than that of the linear models. Finally, as a research objective, we introduced the collocated black carbon mass concentration measurements into the ML models but found no significant improvement in the model performance.