ESTIMATING AIRBORNE PARTICULATE MATTER IN THE NATIONAL CAPITAL REGION, PHILIPPINES USING MULTIPLE LINEAR REGRESSION AND GRADIENT BOOSTING ALGORITHM ON MODIS MAIAC AEROSOL OPTICAL DEPTH

R. A. B. Torres [1] *, R. V. Ramos [1, 2], B. A. B. Recto [1], A. M. Tamondong [1, 2], B. J. D. Jiao [1], M. G. Cayetano [3]

[1] Training Center for Applied Geodesy and Photogrammetry, University of the Philippines, Diliman, Quezon City, Philippines - rbtorres@alum.up.edu.ph
[2] Department of Geodetic Engineering, University of the Philippines, Diliman, Quezon City, Philippines
[3] Institute of Environmental Science and Meteorology, University of the Philippines, Diliman, Quezon City, Philippines

Commission IV, WG IV/3

KEY WORDS: Air Quality, MODIS, PM, AOD, NCR, Multiple Linear Regression, Gradient Boosting.

ABSTRACT:

The generation of air quality concentration data is imperative for the health and environment of highly urbanized regions. Through remote sensing, air pollutant concentrations can be obtained over large areas for a long time. In this study, particulate matter ($PM_{2.5}$ and PM_{10}) concentrations were estimated using satellite-derived Moderate Resolution Imaging Spectroradiometer (MODIS) Multi-Angle Implementation of Atmospheric Correction (MAIAC) Aerosol Optical Depth (AOD) values observed in the National Capital Region (NCR), Philippines. Models were generated using multiple linear regression (MLR) and gradient boosting regression to determine the best models for the whole data from 2017 to 2020, dry season, and wet season with a 70-30 split for the train-test sets. Initial models resulted with the best coefficient of determination R^2 values of 2.6% and 1.2% using MLR and 2.0% using gradient boosting regression. The results for $PM_{2.5}$ and PM_{10} showed the lowest Root Mean Square Error (RMSE) values of 8.79 $\mu g/m^3$ and 18.99 $\mu g/m^3$ using MLR and 8.08 $\mu g/m^3$ and 16.85 $\mu g/m^3$ using gradient boosting, respectively. The preliminary results indicate the relatively poor performance of models in estimating particulate matter using satellite-derived AOD images. Improvements in the models will include the integration of more in-situ data from air quality monitoring stations and the addition of additional variables and features such as meteorological parameters and geographical layers.

1. INTRODUCTION

Air quality monitoring in the Philippines is conducted through ground monitoring station data regulated by the Department of Environment and Natural Resources (DENR). Through air quality monitoring, air pollutant concentration data are gathered to assess if concentration levels are good, unhealthy for sensitive groups, or at emergency levels. However, this assessment solely depends on the position of the monitoring stations placed by the regulatory department; therefore, limiting the scale of assessment based on the location of the air quality monitoring stations (Krunpick et al., 2003; Aniceto et al., 2021). Moreover, the generation of regulatory-grade air quality monitoring stations is costly and would take a huge amount of time for planning and budgeting before being operational. Additionally, there are times when insufficient data were recorded for specific stations. In this sense, air quality sensors were developed by researchers and groups to generate more data, however, the coverage is still limited depending on the target area of the respective owners (Galarpe, 2017; Cruz et al., 2019). On the other hand, air quality monitoring is also conducted nowadays by utilizing geospatial and remote sensing techniques to gather air quality data over large areas. Air quality data are either directly obtained from sensors in various satellite platforms or derived from satellite images using regression analysis or dispersion models. Engel-Cox et al. (2012) provided recommendations on the use of satellite remote sensing data for urban air quality assessment. The researchers showed the integration of ground-based and satellite data for air quality monitoring and the difficulties encountered in terms of collaboration, access, resources, and scope of analysis. Landsat, Multi-angle Imaging Spectroradiometer (MISR), Moderate Resolution Imaging Spectroradiometer (MODIS), TOMS, SeaWiFS, and SPOT were a few of the satellite systems discussed that are ready to be used for air quality applications. Duncan et al. (2014) of NASA Goddard Space Flight Center reviewed and discussed the use of satellite data for air quality applications. Satellite data was determined to be useful in tracking pollutant plumes, air quality forecasting, exceptional events demonstration, and data for air quality models. Li et al. (2020) proposed the integration of estimated particulate matter (PM) concentrations from satellite Aerosol Optical Depth (AOD) products and ground-based data from a network of low-cost PM sensors. Specifically, 75 monitoring stations and 2,363 AirBox low-cost sensor data together with MODIS Terra remote sensing data were used to generate surface PM estimates. Moreover, Allen et al. (1997), Chung et al. (2001), Hauck et al. (2004), and Takahashi et al. (2008) determined that traditional methods for PM measuring such as the gravimetric method, β-ray attenuation monitoring, or the use of tapered element oscillating microbalance (TEOM) were costly and require regular maintenance. Therefore, other methods of measuring PM measurements were introduced, especially low-

* Corresponding author

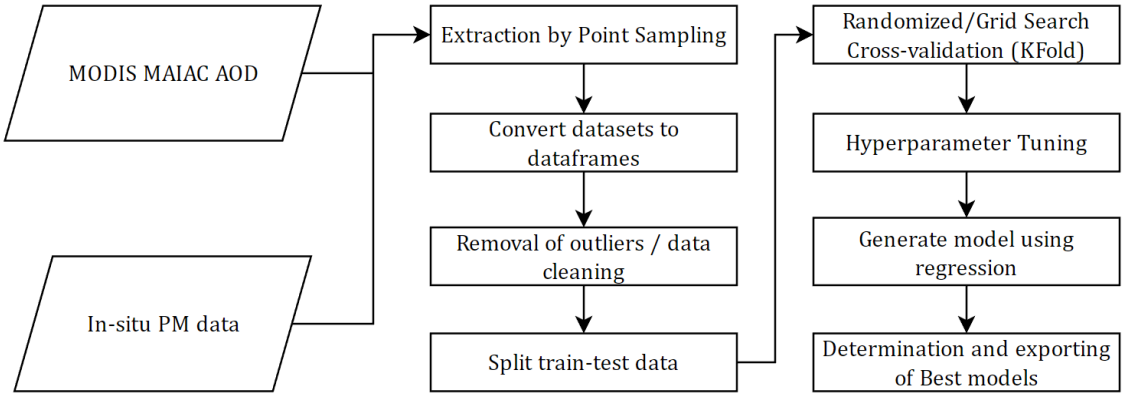

Figure 1. PM modelling workflow using MLR and gradient boosting algorithm.

cost PM sensors and satellite-derived PM concentrations. Satellite-derived AOD is used in several studies for the estimation of ground-based particulate matter with particles of less than 2.5 µm in diameter ($PM_{2.5}$). Various models such as linear regression, multiple linear regression (MLR), artificial neural networks, and geographically weighted regression were tested out by studies as a means of improving the relationship between AOD and $PM_{2.5}$. One of the fundamental requirements in conducting the said models is the availability of satellite data. Gogikar, et al. (2020) conducted a study for the estimation of $PM_{2.5}$ values from satellite-derived AOD using different kinds of regression models, namely, simple linear regression, MLR, log-linear regression, and conditional-based MLR. The study area is in Agra and Rourkela region, India with inclusive years ranging from 2009 to 2015. It was found that the coefficient of determination (R) was statistically significant from the use of Model II or MLR. Moreover, Othman et al. (2010) generated PM_{10} estimates using Landsat 7 ETM+ satellite images and in-situ measurements from the DustTrak aerosol monitor 8520. Multiple linear regression analysis was used to generate the PM_{10} estimation model from the RGB bands of Landsat 7 ETM+. The coefficient of determination resulted in 0.888 and validation with in-situ data gave an accuracy greater than 0.8 for the R coefficient. Machine learning algorithms, specifically gradient boosting, were also used by several researchers in estimating air quality concentration from satellite-derived images. Gradient boosting is a machine learning algorithm used to produce prediction models from an ensemble of weak predictive models for regression and classification. Gradient boosting optimizes a loss function depending on the type of problem needed to be solved, with squared error for regression as an example. The algorithm uses weak learners in the form of decision trees to make predictions. The output of the final model is improved and corrected by adding the output of the new tree to the existing sequence of trees until the loss reaches an acceptable level where no improvement can be observed (Natekin and Knoll, 2013). Gradient boosting is similar to random forests in terms of combining decision trees in the algorithm, however, random forests combine them at the end while gradient boosting combines the trees along the way In general, gradient boosting results in better performance than random forests if parameters were tuned. It would even work with unbalanced data, and reduce the chances of overfitting (Cai et al., 2020). Extreme Gradient Boosting (XGBoost) is an optimized gradient boosting algorithm that applies level-wise tree growth designed to be highly efficient, flexible, and portable. Chen et al. (2019) improved the estimation of ground $PM_{2.5}$ estimation using satellite-derived AOD and extreme gradient boosting (XGBoost) to reduce limits and biases in estimating $PM_{2.5}$. Additionally, a two-step method was used to interpolate missing values in AOD, reducing the missing value rate of daily AOD data to 13.83% from 87.91%. Using XGBoost regression with a non-linear exposure-lag-response model (NELRM) resulted in a cross-validation coefficient of determination of 0.86, a Root Mean Square Error (RMSE) of 14.98, and a Mean Absolute Percentage Error (MAPE) of 23.72%. Another study by Fan et al. (2020) introduced a development in estimating $PM_{2.5}$ using satellite-derived AOD using spatially local extreme gradient boosting (SL-XGB) to obtain accurate results in unsampled spatial areas while also filling gaps in satellite-derived AOD. This study shows the initial models generated for $PM_{2.5}$ and PM_{10} estimation for seasonal and yearly models using MODIS MAIAC AOD and ground monitoring station PM as training-test data for the regression analysis.

2. METHODOLOGY

The study is divided into three major components as shown in Figure 1: (1) Gathering of input parameters, (2) pre-processing of satellite images and monitoring station data, and (3) regression modeling.

2.1 Study Area

Figure 2. The National Capital Region of the Philippines.

Figure 2 shows the National Capital Region, a metropolitan area composed of four districts divided into 16 cities and 1 municipality with a total land area of 619.57 km^2, which serves as the seat of the government, population, and economy of the Philippines (Chua, et al., 2021). The region is affected by two seasons based on temperature and rainfall, specifically the wet season from June to November and the dry season from

December to May. Tomacruz (2018) determined that the Southeast Asian region averages 21 μg/m^3 annually in $PM_{2.5}$ concentrations which is over twice the recommended value. With the region being the center of population and economy where 2.5 million annual average daily traffic was recorded, several risk factors that may cause hazardous effects are present in the region due to significant air pollution, especially in urban areas (Villas-Alvaren, 2016). Specifically, in terms of PM2.5, the World Health Organization (WHO) discovered the region's average concentration of 30.44 μg/m^3 from 2016 to 2018 far exceeded the annual $PM_{2.5}$ standard recommended value of 20 μg/m^3. Particularly, the WHO report for 2016 showed that the Philippines' $PM_{2.5}$ concentration value of 18.4 μg/m^3 is approximately 80% higher than the indicated safe levels, with Metro Manila registering a value of 55 μg/m^3 (Ambag, 2019). These high levels of ambient $PM_{2.5}$ cause prolonged outdoor exposure to be hazardous, especially for those with occupations that are required to do so (Estoque, 2020).

2.2 Input Parameters for Regression Analysis

In this study, MODIS MAIAC satellite-derived AOD was used as the independent parameter for the regression analysis. MODIS is an instrument carried by NASA's Terra and Aqua satellites launched on December 18, 1999, and May 4, 2002, respectively. One of the algorithms used to retrieve AOD from the MODIS spectral reflectance is the Multi-angle Implementation of Atmospheric Correction (MAIAC). . Two bands are available for analysis namely: Optical_Depth_47 (AODB) and Optical Depth 55 (AODG). MAIAC AOD data (MCD19A2.006) which reports daily AOD in 1-km spatial resolution were directly downloaded using Google Earth Engine from 2017 to 2020. Datasets were clipped using the regional boundary shapefile and images with very high cloud cover were removed. In-situ PM data that were used as predicted variable data were obtained from the 15 ground monitoring stations of DENR placed all over NCR as shown in Figure 3. From these ground monitoring stations, PM observations every hour were extracted from 2017 to 2020. Hourly in-situ PM data were then aggregated to daily averages to match the temporal frequency of the satellite-derived AOD images.

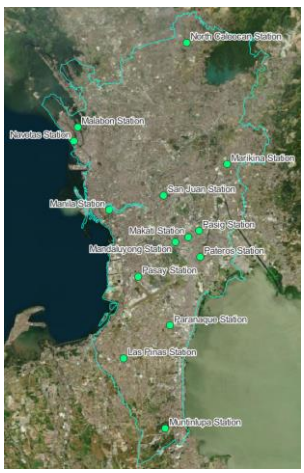

Figure 3. Air quality monitoring stations in the NCR.

2.3 Pre-processing of satellite images and monitoring station data

In-situ PM data were converted to point shapefiles to combine them with the respective AOD concentration data from the satellite images. Concentration values were extracted to comma-separated values files using point sampling. Obtained data from point sampling were combined into one datasheet containing all the data from 2017 to 2020 (yearly), all data for the dry season, and all data for the wet season. Datasets were then converted to a two-dimensional data structure composed of rows and columns called dataframes using a script to prepare them as machine-ready datasets for regression modeling. Negative values and null data values, including random string values, were removed from the data frame. Moreover, outliers were removed through outlier detection techniques. Lastly, data were split into train-test data with 70-30 split percentages.

2.4 Regression Analysis

Models were generated using MLR and Extreme Gradient Boosting (XGBoost). MLR is a regression model that involves two or more independent variables in estimating its relationship with a dependent variable using a straight line (Tranmer et al., 2020). On the other hand, XGBoost is an optimized gradient boosting algorithm that applies level-wise tree growth designed to be highly efficient, flexible, and portable (Wade, 2020). Hyperparameters were tuned using Randomized Search Cross Validation with 10 k-folds, iterating different combinations of hyperparameters to determine the best models, yearly and each season, to avoid overfitting and optimize processing time. The best-performing model was then saved using Joblib. Generated model was then tested using the test data, and the feature importance was determined to evaluate which features affected the models the most.

3. RESULTS AND DISCUSSION

3.1 Data Distribution Plots

Data distribution was observed by checking the histogram and box plots of the input parameters for the PM modeling shown in Figure 4, Figure 5, Figure 6, and Figure 7, respectively. $PM_{2.5}$ values range from 0 to 50 μg/m^3 with the first quantile value of 15 and 3rd quantile value of 30, while 0 to 110 μg/m^3 with the first quantile value of 38 and 3rd quantile value of 62 for PM_{10}. On the other hand, AODB values range from 0 to 500 while AODB ranges from 0 to 350. These observations imply that the values concentrate more in areas below the average value of the respective input parameters.

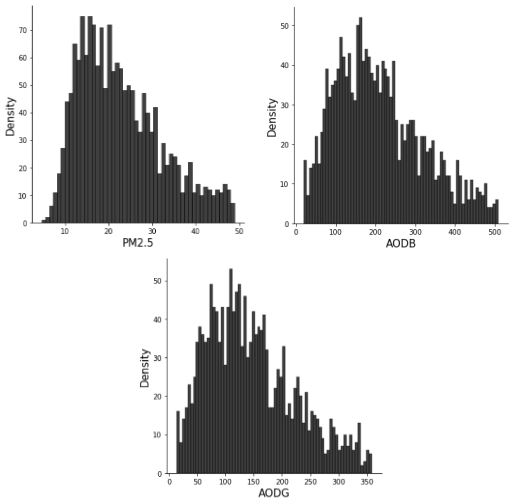

Figure 4. Histogram of the input parameters for $PM_{2.5}$ models.

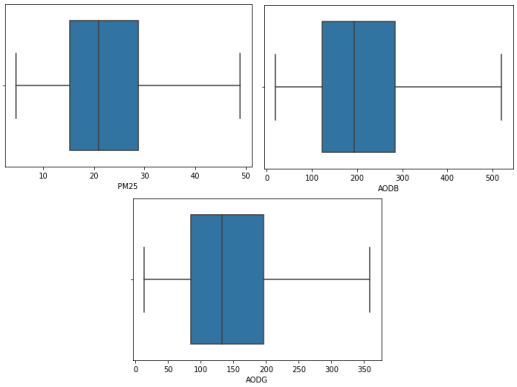

Figure 5. Box plots of the input parameters for $PM_{2.5}$ models.

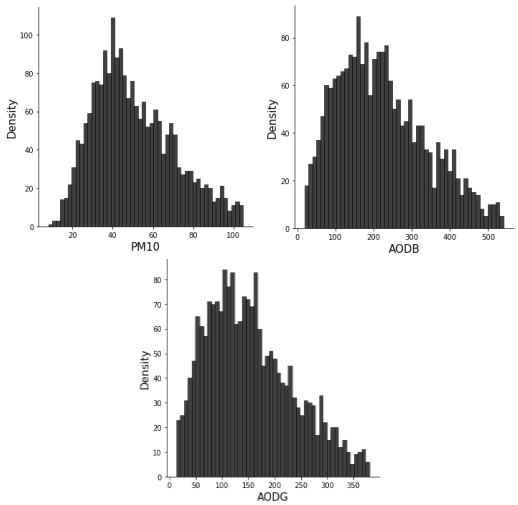

Figure 6. Histogram of the input parameters for PM_{10} models.

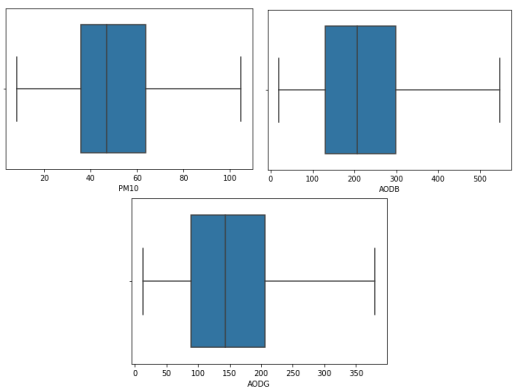

Figure 7. Box plots of the input parameters for PM_{10} models.

3.2 Correlation Analysis

Correlation between variables was first determined before regression modeling. Correlation analysis was performed to determine the strength of the relationship between the defined variables. However, it is necessary to remember that correlation analysis does not define causation between the variables. A positive correlation only denotes a direct relationship between the variables. If there is an increase/decrease in one variable, an increase/decrease can also be observed in the other variable. On the other hand, a negative correlation signifies an increase in one variable while the other variable decreases, and vice versa. Table 1 shows the correlation coefficient between AOD and PM. $PM_{2.5}$ in general shows a greater correlation to AODB and AODG than PM_{10} to the independent variables.

Correlation	$PM_{2.5}$			PM_{10}		
	Yearly	Dry	Wet	Yearly	Dry	Wet
AODB	0.15	0.12	0.14	0.086	0.11	0.05
AODG	0.15	0.12	0.14	0.086	0.11	0.05

Table 1. Correlation coefficient (Pearson) of AODB and AODG with $PM_{2.5}$ and PM_{10}.

Models containing all the data from 2017 to 2020 (Yearly model) resulted in the highest correlation to $PM_{2.5}$ while the dry season model showed the highest correlation between the variables for PM_{10}. Correlation coefficients for all the models between AODB and AODG were very similar and almost equal. This indicates that AODB and AODG were highly correlated with each other and might be showing the same observations over the target areas. The similarity between AODB and AODG might be due to their nature of being AOD with the only difference being one was derived from 0.47 μm and 0.55 μm, respectively, having minute differences for most areas over land. With this result, future models can remove one of the said parameters for a more optimal modeling process and avoidance of multicollinearity between variables.

3.3 Multiple Linear Regression Analysis

After correlation analysis, models were generated using multiple linear regression both for all of 2017 to 2020 and dry and wet seasons. Table 2 shows the coefficient values of the resulting linear regression models for the yearly, dry, and wet seasons.

Coefficient	$PM_{2.5}$			PM_{10}		
	Yearly	Dry	Wet	Yearly	Dry	Wet
Intercept	20.86	19.74	22.5	48.67	48.83	52.33
AODB	-0.21	-0.11	-0.12	-0.39	-0.92	-0.17
AODG	0.32	0.16	0.18	0.58	1.32	0.25

Table 2. Coefficient values of the generated PM models using multiple linear regression.

The intercept value in a regression model represents the mean value of PM when the AOD is zero, though, a rare occasion for this parameter. On the other hand, the coefficient values for AOD showed negative values for AODB across all the models while positive values for AODG. This denotes that an increase in AODB causes a decrease in the mean PM and vice versa, while an increase/decrease in AODG causes an increase/decrease in the mean PM. Moreover, AODG resulted in higher magnitudes than AODB, implying that AODG affects the values of the PM more than AODB in the generated models. Model parameters including error measurements were also checked to determine the performance of the models. Table 3 shows error measurements of the generated models, specifically Mean Absolute Error (MAE), Mean Squared Error (MSE), and RMSE. The lower the error values for all these model parameters implies a better performance of the models.

Model	$PM_{2.5}$			PM_{10}		
	Yearly	Dry	Wet	Yearly	Dry	Wet
MAE	7.76	7.05	7.91	16.51	15.68	18.54
MSE	90.62	77.29	92.18	409.97	360.68	515.31
RMSE	9.52	8.79	9.61	20.25	18.99	22.7

Table 3. Model error measurements of the generated PM models using MLR.

Specifically for RMSE, the resulting models showed relatively low error values with the dependent variable, PM, with minimum and maximum values of 0 to 50 for $PM_{2.5}$ and 0 to 100 for PM_{10}. The coefficient of determination (R^2) for almost all models resulted in positive values significant from zero with the highest percentage of around 2.6%. $PM_{2.5}$ models showed a better fit than PM_{10} to AOD, however, underfitting was present for both models of $PM_{2.5}$ and PM_{10}. This might imply that the variability of PM was not fully taken into account by the independent variables, AOD. Therefore, for the improvement of the model, other variables were added to the modeling process such as meteorological parameters. Moreover, the number of observations used to model PM might not be enough to properly calculate its variability as also seen later on with the gradient boosting models.

Figure 8 and Figure 9 display the scatter plot of in-situ PM vs predicted PM from MLR. Scatter plots showed the concentration of values below the average values which is more predominant in the PM_{10} models. The clustering of values over specific predicted values or an unbalance in the y-axis can also be seen as a result of the underfitting of the models.

Figure 10 and Figure 11 show the residual distribution plots of the best models for $PM_{2.5}$ and PM_{10}. Plots showed near-normal distribution curves with a slight skew to the right. Specifically, plots showed residuals having a higher density in the negative area from 0 to -10. This could mean that the model is biased for lower values resulting in a lower fit across all the models.

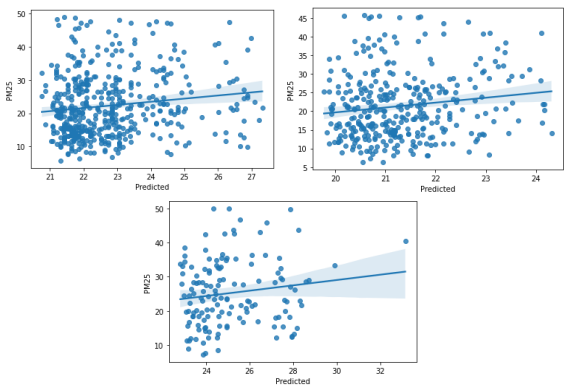

Figure 8. In-situ $PM_{2.5}$ vs. Predicted $PM_{2.5}$ plots for the yearly model (top left), dry model (top right) and wet model (bottom).

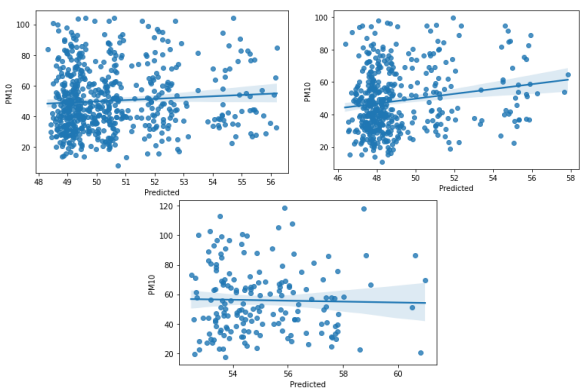

Figure 9. In-situ PM_{10} vs. Predicted PM_{10} plots for the yearly model (top left), dry model (top right) and wet model (bottom).

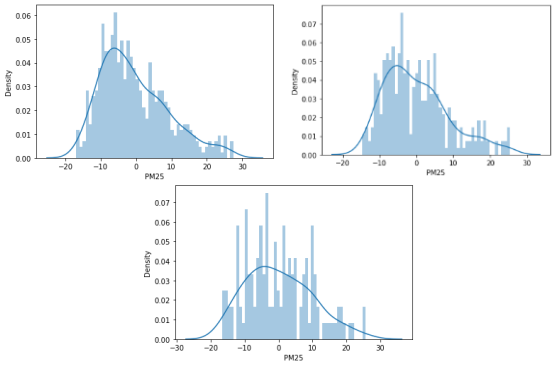

Figure 10. In-situ $PM_{2.5}$ residual plots for the yearly model (top left), dry model (top right) and wet model (bottom).

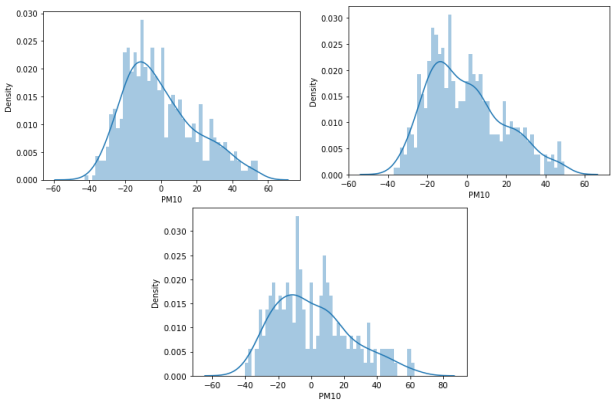

Figure 11. In-situ PM_{10} residual plots for the yearly model (top left), dry model (top right) and wet model (bottom).

3.4 Gradient Boosting Regression

Table 4 summarizes the train and test RMSE of the generated models using gradient boosting regression. Similar to the results from MLR, the resulting RMSE using gradient boosting regression was relatively low in comparison to the dependent variable's range. Moreover, computed RMSE between the train and test sets showed a minute difference between each other, indicating no case of overfitting for all the generated models. R^2 resulted in the highest training value of 2% but low test scores. Although gradient boosting in general results in better accuracy in modeling than MLR in most cases, machine learning algorithms require an ample amount of training data to properly produce models that were cross-validated with optimized hyperparameters, especially, when there are only two independent variables with 15 observation points.

Model	$PM_{2.5}$			PM_{10}		
	Yearly	Dry	Wet	Yearly	Dry	Wet
Train RMSE	9.01	8.39	8.08	19.13	16.85	20.82
Test RMSE	9.62	8.82	9.64	20.58	20.02	22.79

Table 4. Model error measurements of the generated PM models using gradient boosting regression.

In a similar way of checking the influence of the independent variables in affecting the dependent variable values through the coefficient values, feature importance plots were generated to determine which parameter between AODB and AODG affected the model building the most using gradient boosting regression. Feature importance computes F-scores that represent the importance of each input feature for a given model. The larger

the F-score means the higher the effect of that specific input feature in building the model in predicting a variable or PM in this case.

Figure 12 shows the F-scores of each input feature for all the generated models using gradient boosting regression. Overall, AODG resulted in higher F-score values than AODB, even though the difference between these input features was low. From these results and the ones from MLR, dropping AODB for future modeling processes can be considered for a more optimal modeling process and smaller chances for multicollinearity.

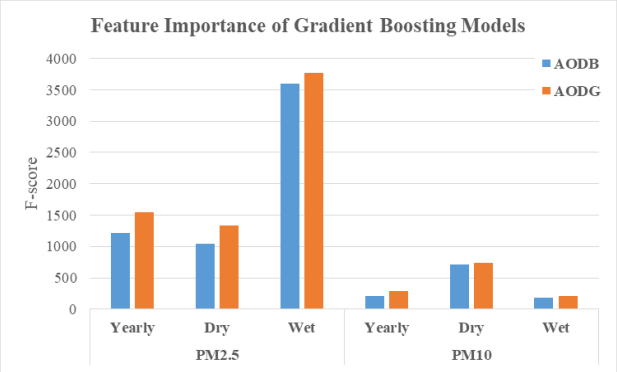

Figure 12. Feature importance of the generated gradient boosting models for PM estimation.

Falk and Miller (1992) recommended coefficient of determination values greater than 0.10 to adequately determine the variability of the endogenous variable. On the other hand, Cohen (1988) recommended endogenous variables with R^2 values of 0.26 to be assessed as substantial, while 0.13 for moderate, and 0.02 for weak.

Some of the models resulted in a weak correlation coefficient significant from zero. This implies that the input variables might not be sufficient to properly take into account the variability of PM, however, this does not mean that the other resulting variables from the correlation analysis and feature importance plots are insignificant. On the other hand, it is imperative that to improve the models, additional input parameters or more in-situ monitoring station data are needed to gather more significant results.

3.5 $PM_{2.5}$ and PM_{10} Maps

Maps were generated using a python script to apply the saved best models with JobLib to the satellite-derived AOD images shown in Figure 13, Figure 14, Figure 15, and Figure 16, respectively.

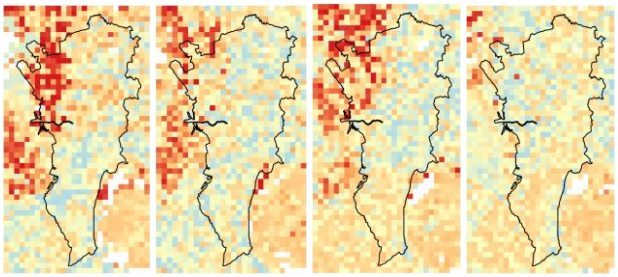

Figure 13. Sample $PM_{2.5}$ estimation maps using the best dry model for March 2017 to 2020.

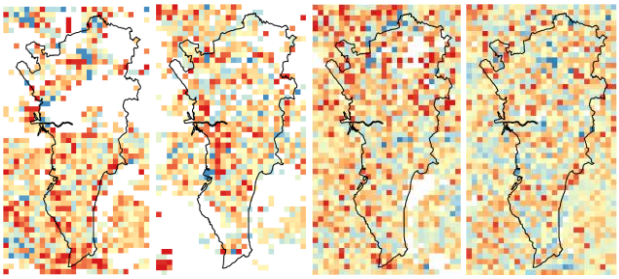

Figure 14. Sample $PM_{2.5}$ estimation maps using the best wet model for July 2017 to 2020.

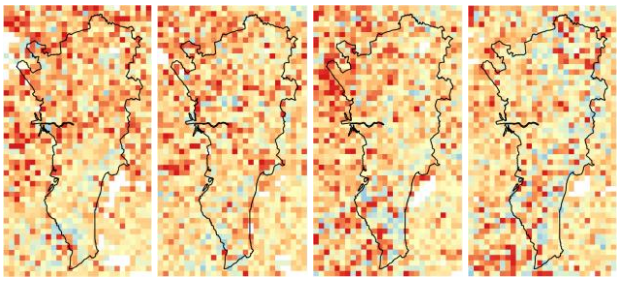

Figure 15. Sample PM_{10} estimation maps using the best dry model for March 2017 to 2020.

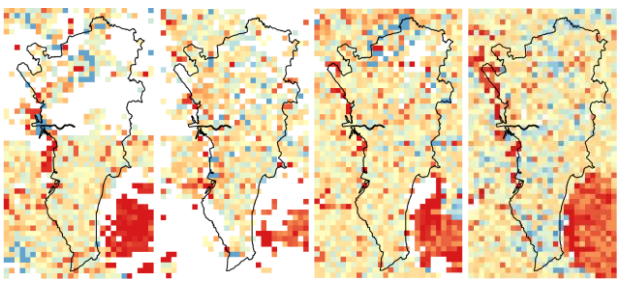

Figure 16. Sample PM_{10} estimation maps using the best wet model for July 2017 to 2020.

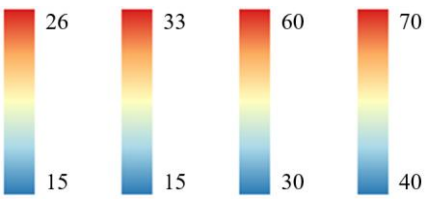

Figure 17. Map legends in $\mu g/m^3$ for $PM_{2.5}$ for March, $PM_{2.5}$ for July, PM_{10} for March, and PM_{10} for July (left to right).

Map gradient of Blue-Yellow-Red was used to show the low to high concentrated areas of PM as shown in Figure 17. White spaces are null values due to cloud cover. Even though it is difficult to determine clusters of PM from the maps alone, $PM_{2.5}$ maps for March 2017 to 2020 show a high concentration of particulate matter in the Northwest – West region of the image with a lower concentration near the central area of NCR.

4. CONCLUSION AND RECOMMENDATIONS

Regression models were generated for the estimation of $PM_{2.5}$ and PM_{10} using MODIS MAIAC AOD satellite images. Models were generated by integrating in-situ monitoring station data with satellite-derived data using MLR and XGBoost.

Results showed low coefficient of determination values and significantly low RMSE values. Cases of underfitting for the

models might be the result of the insufficient number of data points, specifically for machine learning algorithms such as gradient boosting regression due to its integrated cross-validation process when generating the model. This results in models with a coefficient of determination value less than zero. However, PM and AOD showed a correlation, especially for $PM_{2.5}$ across all the models. Variability and fit may be improved when AOD is combined with other input parameters in estimating PM.

$PM_{2.5}$ and PM_{10} concentration estimation maps were generated by applying the best models to the MODIS MAIAC AOD satellite images.

Undergoing improvement of the models is the addition of several input parameters such as meteorological parameters. Specifically, parameters such as wind speed and direction, mean temperature, minimum temperature, maximum temperature, pressure, precipitation, land use/land cover, and other geospatial layers are added to encapsulate the variability of $PM_{2.5}$ and PM_{10}. Additional years of analysis might also be helpful if air quality monitoring station data are available.

ACKNOWLEDGEMENTS

This research was done as part of the Ambient Air Remote Sensing, Modeling, and Visualization Environment (Project AiRMoVE). The Project was implemented by the University of the Philippines Training Center for Applied Geodesy and Photogrammetry (TCAGP), through the support of the Department of Science and Technology (DOST) of the Republic of the Philippines and the Philippine Council for Industry, Energy, and Emerging Research and Development (PCCIERD).

REFERENCES

Allen, T.D., Rush, M.C., 1998. The effects of Organizational Citizenship Behavior on performance judgments: A Field Study and a laboratory experiment. *Journal of Applied Psychology*, 83(2), 247–260. doi.org/10.1037/0021-9010.83.2.247.

Ambag, R., 2019. *How bad is air pollution in the Philippines?* FlipScience. Retrieved May 12, 2022, from flipscience.ph/health/how-bad-air-pollution-philippines/.

Aniceto, K.R., Macam, J.J., Salmorin, E.I., Sison, Z.K., Mission, M.P., Camacho, I.K., Poso, F.D., 2021. Seasonal mapping and air quality evaluation of total suspended particulate concentration using arcgis-based spatial analysis in Metro Manila, Philippines. *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management*. doi.org/10.1109/hnicem54116.2021.9732031.

Cai, J., Xu, K., Zhu, Y., Hu, F., Li, L., 2020. Prediction and analysis of Net Ecosystem Carbon Exchange based on gradient boosting regression and random forest. *Applied Energy*, 262, 114566. doi.org/10.1016/j.apenergy.2020.114566.

Chen, Z.Y., Zhang, T.H., Zhang, R., Zhu, Z.M., Yang, J., Chen, P.Y., Ou, C.Q., Guo, Y., 2019. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. *Atmospheric Environment*, 202, 180–189. doi.org/10.1016/j.atmosenv.2019.01.027.

Chua, P.L., Ng, C.F., Rivera, A.S., Salva, E.P., Salazar, M.A., Huber, V., Hashizume, M., 2021. Association between ambient temperature and severe diarrhea in the National Capital Region, Philippines. *International Journal of Environmental Research and Public Health*, 18(15), 8191. doi.org/10.3390/ijerph18158191.

Chung, A., Chang, D.P.Y., Kleeman, M.J., Perry, K.D., Cahill, T.A., Dutcher, D., McDougall, E.M., Stroud, K., 2001. Comparison of real-time instruments used to monitor Airborne Particulate Matter. *Journal of the Air & Waste Management Association*, 51(1), 109–120. doi.org/10.1080/10473289.2001.10464254.

Cohen, J., 1988: *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associate.

Cruz, L.A.A., Griño, M.T.T., Tungol, T.M.V., Bautista, J.T., 2019. Development of a low-cost air quality data acquisition IOT-based system using Arduino Leonardo. *International Journal of Engineering and Manufacturing, 9(3)*, 1–18. doi.org/10.5815/ijem.2019.03.01.

Duncan, B.N., Prados, A.I., Lamsal, L.N., Liu, Y., Streets, D.G., Gupta, P., Hilsenrath, E., Kahn, R.A., Nielsen, J.E., Beyersdorf, A.J., Burton, S.P., Fiore, A.M., Fishman, J., Henze, D.K., Hostetler, C.A., Krotkov, N.A., Lee, P., Lin, M., Pawson, S., Ziemba, L.D., 2014. Satellite data of atmospheric pollution for U.S. air quality applications: Examples of applications, summary of data end-user resources, answers to faqs, and common mistakes to avoid. *Atmospheric Environment*, 94, 647–662. doi.org/10.1016/j.atmosenv.2014.05.061.

Engel-Cox, J.A., Hoff, R.M., Haymet, A.D.J., 2004. Recommendations on the use of satellite remote-sensing data for Urban Air Quality. *Journal of the Air & Waste Management Association*, 54(11), 1360–1371. doi.org/10.1080/10473289.2004.10471005.

Estoque, R.C., Ooba, M., Seposo, X.T., Togawa, T., Hijioka, Y., Takahashi, K., Nakamura, S., 2020. Heat health risk assessment in Philippine cities using remotely sensed data and social-ecological indicators. *Nature Communications*, 11(1). doi.org/10.1038/s41467-020-15218-8.

Falk, R.F., Miller, N.B., 1992: *A primer for Soft Modeling.* University of Akron Press.

Fan, Z., Zhan, Q., Yang, C., Liu, H., Bilal, M., 2020. Estimating PM2.5 concentrations using spatially local Xgboost based on full-covered Sara AOD at the Urban Scale. *Remote Sensing*, 12(20), 3368. doi.org/10.3390/rs12203368.

Gogikar, P., Tripathy, M.R., Rajagopal, M., Paul, K.K., Tyagi, B., 2020. PM2.5 estimation using multiple linear regression approach over industrial and non-industrial Stations of India. *Journal of Ambient Intelligence and Humanized Computing*, 12(2), 2975–2991. doi.org/10.1007/s12652-020-02457-2.

Hauck, H., Berner, A., Gomiscek, B., Stopper, S., Puxbaum, H., Kundi, M., Preining, O., 2004. On the equivalence of gravimetric PM data with TEOM and beta-attenuation measurements. *Journal of Aerosol Science*, 35(9), 1135–1149. doi.org/10.1016/j.jaerosci.2004.04.004.

Heyasa, B.B., Galarpe, V.R., 2017. Preliminary development and testing of microcontroller-MQ2 GAS SENSORFOR University Air Quality Monitoring. *IOSR Journal of Electrical and Electronics Engineering, 12(03)*, 47–53. doi.org/10.9790/1676-1203024753.

Krupnick, A.J., Morgenstern, R.D., Fischer, C., Rolfe, K., Logarta, J., Rufo, B., 2003. Air Pollution Control Policy Options for Metro Manila. *Discussion Papers 10612, Resources for the Future*.

Li, J., Zhang, H., Chao, C.Y., Chien, C.H., Wu, C.Y., Luo, C.H., Chen, L.J., Biswas, P., 2020. Integrating low-cost air quality sensor networks with fixed and satellite monitoring systems to study ground-level PM2.5. *Atmospheric Environment*, 223, 117293. doi.org/10.1016/j.atmosenv.2020.117293.

Natekin, A., Knoll, A., 2013. Gradient Boosting Machines, a tutorial. Frontiers in Neurorobotics. *Front. Neurorobot.* 7. doi.org/10.3389/fnbot.2013.00021.

Othman, N., Mat Jafri, M.Z., San, L.H., 2010. Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia. *Modern Applied Science*, 4(11). doi.org/10.5539/mas.v4n11p131.

Takahashi, K., Sugi, Y., Hosono, A., Kaminogawa, S., 2009. Epigenetic regulation of TLR4 gene expression in intestinal epithelial cells for the maintenance of intestinal homeostasis. *The Journal of Immunology*, 183(10), 6522–6529. doi.org/10.4049/jimmunol.0901271.

Tomacruz, S., 2018. *Air Pollution Deaths 3rd highest in ph*. RAPPLER. Retrieved May 12, 2022, from rappler.com/nation/208192-air-pollution-deaths-3rd-highest-philippines/.

Tranmer, M., Murphy, J., Elliot, M., Pampaka, M., 2020. Multiple Linear Regression (2nd Edition); *Cathie Marsh Institute Working Paper 2020-01*. hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1-multiple-linear-regression.pdf.

Villas-Alvaren, A.L., 2016. *MMDA faces greatest challenge: managing 2.5M vehicles in MM*. PressReader.com - Digital Newspaper & Magazine subscriptions. Retrieved May 12, 2022, from pressreader.com/philippines/manila-bulletin/20161225/282016146985660.

Wade, C., 2020: *Hands-on gradient boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with python.* Packt Publishing.