


Predicting PM2.5 in Urban and Suburban Beijing: A Comparative Study of Random Forest and Linear Regression Models

Zhuoyang Zhou ^a

Beijing Normal - Hong Kong Baptist University, Zhuhai, 519000, China

Keywords: PM2.5, Regression Model, Random Forest, Prediction Model.

Abstract: This study evaluates the performance of multiple linear regression (MLR) and random forest regression (RFR) models in predicting PM2.5 concentrations across twelve air quality monitoring stations in Beijing, China, using hourly meteorological and pollution data from 2013 to 2017. The analysis reveals that RFR significantly outperforms MLR, with R^2 values improving from 0.11 to 0.22 (MLR) to 0.29–0.41 (RFR), demonstrating better handling of non-linear interactions. However, both models exhibit critical limitations, particularly in predicting extreme pollution events ($PM_{2.5} > 300 \mu g/m^3$), where systematic underprediction occurs. Geographical disparities in model accuracy are evident, with suburban stations (e.g., Dingling, Huairou) exhibiting lower errors than urban-industrial sites (e.g., Dongsì, Aotizhongxin), likely due to the complexity of emission sources and microclimates. Dew point temperature emerges as the most influential predictor, while precipitation shows limited impact. These findings underscore the challenges in air quality forecasting and advocate for localised, hybrid modelling approaches integrating real-time emission data to enhance predictive reliability for public health applications.

1 INTRODUCTION


Air pollution, particularly that 2.5 microns or smaller ($PM_{2.5}$), seriously threatens the environment and people's health worldwide. Air pollution, including $PM_{2.5}$, directly impacts people's health, which may cause heart problems and breathing difficulties (Brook et al., 2010). Additionally, it harms the setting and the essence (Li et al., 2019). To better formulate recommendations for cleaner environments and lessen the damaging effects of air pollution, people must be aware of how these issues impact $PM_{2.5}$.

$PM_{2.5}$ deposition in the air significantly impacts weather conditions, such as temperature, humidity, and wind speed. For instance, higher temperatures can increase $PM_{2.5}$, while higher humidity can make it easier to create tiny antigens (Li et al., 2019; Perrino et al., 2011). The drizzle of pollutants depends on wind speed. $PM_{2.5}$ particles are typically higher because the air doesn't mix well when the wind blows (Tai et al., 2010). $PM_{2.5}$ costs may be accurately predicted because of how connected these elements are. Besides weather conditions, personal activities, especially retreat-related, can drastically change

$PM_{2.5}$. Individuals move more during holidays, corporations may work differently, and more situations take place, which may affect air quality. For example, during major festivals like Chinese New Year, there are fewer factory activities and cars on the road, which makes the air fresh for a short time (Wang et al., 2014; Wang et al., 2017).

On the other hand, trips that involve more travel and tourism may produce more pollutants from transportation and places to stay, leading to higher levels of $PM_{2.5}$ (Zhang et al., 2015). It is crucial to understand how $PM_{2.5}$ costs and air quality change. Using this data, better air quality management strategies can be created during the lively holidays.

Some studies have examined how $PM_{2.5}$ prices, weather conditions, and actions are connected in various locations worldwide. In a study conducted in Beijing, China, the wind's temperature, humidity, and wind speed significantly impacted the amount of $PM_{2.5}$ provided. When the weather and the wind were cooler, $PM_{2.5}$ costs increased (Zhang et al., 2017). A study in the United States found that weather conditions played a key role in modifying $PM_{2.5}$ levels, with temperature and wind speed being

^a <https://orcid.org/0009-0001-6098-6604>

the most important aspects (Perrino et al., 2011). Due to the large number of people using lights and increased traffic on the roads, a study in India found that during the Diwali festival, the levels of harmful PM_{2.5} in the air significantly increased (Guo et al., 2014). When measuring PM₂, these analyses demonstrate that, 5, both weather conditions and specific activities should be taken into account.

Despite numerous studies, People still aren't aware of the relationship between PM_{2.5} levels, temperature, and holidays. Most studies focused on PM₂'s impact on a single cultural problem (Pope and Dockery, 2006). Not many studies have examined how various weather conditions interact with one another to alter it (Pope & Dockery, 2006). Moreover, people don't understand how weather conditions and breaks affect PM_{2.5} (Wang et al., 2017).

To remove these deficiencies, this study will examine how PM_{2.5} charges relate to weather conditions and falls in a particular area. The assessment uses information on air quality and weather conditions to observe how PM_{2.5} (a type of air pollution) rates change over weather changes. To improve air quality management and develop better rules and regulations, the research may utilise multiple linear regression analysis to examine the effects of weather conditions.

Complex components, like the environment and the lives of women, are affected by PM_{2.5} exposure. Knowing how these factors affect PM_{2.5} rates is crucial to making effective air quality management strategies. This research hopes to raise the consciousness by understanding how PM_{2.5} rates, temperature, and holidays relate to a specific area. Policymakers and professionals can comprehend these issues because of this (Tai et al., 2010; Zhang et al., 2015).

2 METHODOLOGY

2.1 Data Source and Description

The dataset used in this study was obtained from Kaggle, comprising hourly air quality measurements from twelve monitoring stations in Beijing, China, spanning from March 1st, 2013, to February 28th, 2017.

2.2 Indicator Selection and Description

Meteorological variables-TEMP, DEWP, PRES, and RAIN-were chosen as independent variables for their

established influence on PM_{2.5} dispersion and formation (as showing in Table 1). For instance, PM_{2.5} concentrations exhibit considerable variability, with hourly readings ranging from 3 to 500 $\mu\text{g}/\text{m}^3$, highlighting the severity of pollution episodes. Temperature and dew point display seasonal trends, while precipitation events are sporadic but critical for pollutant scavenging. Table 1 lists all the variable names and their descriptions, and ranges.

Table 1: Descriptions and ranges of variables

Variable	Description	Range
PM2.5	Fine particulate matter concentration ($\mu\text{g}/\text{m}^3$)	2.0 to 999.0
TEMP	Temperature ($^{\circ}\text{C}$)	-19.9 to 42.6
DEWP	Dew point temperature ($^{\circ}\text{C}$)	-43.3 to 29.1
PRES	Atmospheric pressure (hPa)	982.4 to 1042.8
RAIN	Precipitation (mm)	0.0 to 72.5

2.3 Methodology

The analysis employs two regression techniques: multiple linear regression (MLR) and random forest regression (RFR).

Multiple Linear Regression (MLR): MLR establishes baseline relationships between PM_{2.5} and meteorological factors, providing interpretable coefficients for each predictor. The model is formulated as: $\text{PM}_{2.5} = \beta_0 + \beta_1 \text{TEMP} + \beta_2 \text{DEWP} + \beta_3 \text{PRES} + \beta_4 \text{RAIN} + \epsilon$, where β_0 is the intercept, β_1 to β_4 are coefficients, and ϵ is the error term.

Random Forest Regression (RFR): RFR, a machine learning approach, captures non-linear interactions and improves predictive accuracy. The model uses bootstrap aggregation and random feature selection, with hyperparameters (e.g., $\text{n_tree} = 500$) tuned via cross-validation to prevent overfitting. Key advantages include handling non-linearity and robustness to outliers. Normalisation was included to address scale differences and remove missing values (na.omit).

The analytical workflow began with a stratified data approach. Each station's dataset was divided into training (70%) and testing (30%) subsets. Both MLR and RFR models were then trained on the training subsets. This study used three evaluation indices (the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE)) to compare the performance of MLR and RFR models in predicting PM_{2.5} concentrations across Beijing's monitoring stations.

3 RESULTS AND DISCUSSION

3.1 Multiple Linear Regression

By using the R code, this paper constructs the multiple linear regression model of the 12 stations. The regression coefficients are shown in Table 2.

Table 2: Multiple Linear Regression Coefficients by Station

Station	Intercept	PRES	TEMP	DEWP	RAIN
Aotizhongxin	1407.088	-1.243	-5.819	4.005	-3.686
Changping	1039.030	-0.903	-4.621	3.260	-4.126
Dingling	1016.191	-0.885	-4.611	3.342	-3.747
Dongsi	2048.266	-1.858	-6.647	4.300	-6.038
Guanyuan	1335.387	-1.172	-5.770	4.047	-5.724
Gucheng	1373.740	-1.208	-5.801	3.843	-6.324
Huairou	490.354	-0.379	-3.600	2.885	-3.180
Nongzhanguan	1886.982	-1.698	-6.840	4.300	-6.392
Shunyi	1258.115	-1.100	-5.488	3.907	-4.653
Tiantan	1935.161	-1.755	-6.278	3.960	-5.063
Wanliu	1341.457	-1.182	-5.534	3.722	-2.901
Wanshouxigong	1889.810	-1.705	-6.570	3.939	-6.647

From the results provided by the R code, this paper can summarise the coefficient ranges and offer some possible explanations for these results. The summary will be shown in Table 3 below.

As table 3 shows, RAIN (mm) -6.65 to -3.18 Rainfall significantly reduces PM2.5, with urban stations (e.g., Wanshouxigong) showing stronger effects. This is possibly due to rain efficiently depositing PM2.5.

Urban stations (e.g., Dongsi, Nongzhanguan) exhibit larger coefficients for PRES, TEMP, and RAIN, suggesting that meteorological factors play a more pronounced role in PM2.5 variability in densely populated areas. Suburban stations (e.g., Huairou, Dingling) show weaker relationships, possibly due to fewer local emissions and greater influence of

regional transport. Huairou Station has the smallest coefficients (e.g., PRES: -0.38, TEMP: -3.60). This is possibly because of its location in a rural, mountainous area, which reduces the sensitivity of PM2.5 to local weather.

Table 3. MLR Coefficient Ranges and Interpretations

Variables	Range	Explanation
Intercept	490.35 to 2048.27	Intercepts of PM2.5 are significantly different from station to station. Urban sites (e.g., Dongsi, Nongzhanguan) have higher intercepts, likely due to more substantial local emissions.
PRES (hPa)	-1.86 to -0.38	Higher atmospheric pressure will lead to lower PM2.5. This is possible because stable weather conditions suppress vertical spread. But the effect is weaker in suburban stations.
TEMP (°C)	-6.84 to -3.60	Temperature has a negative effect. This is possible because warmer conditions enhance atmospheric mixing, and seasons have higher temperatures, like summer reduce coal heating emissions.
DEWP (°C)	2.89 to 4.30	Higher dew point strongly increases PM2.5. This is possibly caused by moisture-enhanced secondary aerosol formation and stagnant air masses.
RAIN (mm)	-6.65 to -3.18	Rainfall significantly reduces PM2.5, with urban stations (e.g., Wanshouxigong) showing stronger effects. This is possibly due to rain efficiently depositing PM2.5.

Dongsi Station shows the strongest negative effect of TEMP (-6.647), potentially linked to its central urban setting, where temperature inversions trap pollutants. The negative RAIN coefficients align with Beijing's observed "post-rain blue sky" phenomenon, where precipitation removes particulate matter. The stronger effect at urban stations (e.g., coefficient of -6.65 at Wanshouxigong) may reflect higher initial PM2.5 concentrations available for wet deposition.

3.2 Random Forest Regression

This study also uses R code to construct a random forest regression model (RFR) and calculate the importance score (IncMSE) of these variables.

Table 4: Variable Importance Score (IncMSE)

Variable	RAIN (%)	DEWP (%)	TEMP (%)	PRES (%)
Aotizhongxin	42.5	24.5	18.6	14.4
Changping	33.9	29.1	19.2	17.7
Dingling	39.9	30.2	16.6	13.3
Dongsi	38.3	24.2	19.2	18.3
Guanyuan	39.6	23.3	17.9	19.2
Gucheng	36.2	25.1	18.7	20
Huairou	32.9	26.5	20.9	19.8
Nongzhanguan	39.1	21.9	18.2	20.7
Shunyi	33.9	24.7	19.1	22.3
Tiantan	31	29.7	22.3	17
Wanliu	44.9	23	17.9	14.2
Wanshouxigong	40.9	24.4	18.5	16.2

The variable importance scores (IncMSE) from Random Forest Regression reveal distinct patterns in how meteorological factors influence PM_{2.5} concentrations across Beijing's monitoring stations (Table 4). The results highlight both consistent trends and notable spatial variations in atmospheric processes affecting air quality. RAIN emerges as the most important predictor at all stations (31.0–44.9% importance), particularly at Wanliu (44.9%) and Aotizhongxin (42.5%). This reflects Beijing's reliance on wet deposition for particulate removal, where precipitation effectively scavenges aerosols from the atmosphere. The stronger effect at urban stations suggests higher initial PM_{2.5} loading available for removal. DEWP shows moderate importance (21.9–30.2%), peaking at Dingling (30.2%) and Tiantan (29.7%). This importance likely represents moisture-enhanced secondary aerosol formation and stagnant conditions during high humidity episodes. TEMP (16.6–22.3%) and PRES (13.3–22.3%) show more variable importance across stations. For example, Tiantan station shows unusually high TEMP importance (22.3%), possibly due to its location near parks where temperature inversions may trap pollutants.

Urban stations (Dongsi, Nongzhanguan) show balanced importance across all variables. Suburban stations (Huairou, Shunyi) display elevated PRES importance (19.8–22.3%), suggesting the greater influence of synoptic weather patterns. Wanliu Station shows extreme RAIN dominance (44.9%) with low DEWP importance (23.0%), possibly due to its location near water bodies enhancing rain effects. Tiantan Station has unusually high TEMP importance (22.3%), potentially reflecting the urban heat island effect in this cultural landmark area. Huairou Station demonstrates the most balanced distribution, consistent with its rural location, where no single factor dominates.

The strong RAIN importance suggests that weather modification (e.g., cloud seeding) could be particularly effective during pollution episodes. High DEWP importance indicates that humidity control measures might help reduce secondary aerosol formation. Urban stations may benefit most from emission controls before forecasted precipitation events. Suburban stations require more attention to pressure systems and temperature variations.

The IncMSE results demonstrate that while rainfall universally dominates PM_{2.5} variability across Beijing, the relative importance of other factors varies substantially by location. This spatial heterogeneity underscores the need for tailored air quality management strategies that account for local meteorological sensitivities. The outlier behaviour at stations like Wanliu and Tiantan suggests that microclimate effects may significantly modify pollution-weather relationships in specific urban contexts. Future work should incorporate finer-scale topographic and land-use data to better explain these station-level differences.

3.3 Model Performance Metrics

Table 5 lists all the R^2 , RMSE and MAE of the 12 stations. The evaluation metrics reveal several key patterns in the performance of MLR and RFR models across Beijing's air quality monitoring stations.

Both models show limited predictive power overall, with test R^2 values ranging from 0.112–0.225 for MLR and 0.287–0.411 for RFR, indicating that meteorological factors alone explain less than half of PM_{2.5} variability. This suggests that additional predictors like wind patterns, emission sources, or temporal factors may be necessary for improved accuracy. The RFR models consistently outperform MLR in training data (R^2 0.425–0.509 vs 0.117–0.220), but this advantage diminishes in test data, revealing moderate overfitting, particularly at stations like Aotizhongxin where the train-test R^2 gap exceeds 0.12. This overfitting likely stems from the RFR's complexity of capturing noise in the training data.

Spatial patterns in model performance reflect Beijing's air pollution dynamics. Urban stations (Dongsi, Nongzhanguan) show the highest R^2 values for both models, with Nongzhanguan's RFR achieving the best test performance ($R^2=0.406$). This urban advantage may result from stronger, more consistent relationships between meteorological conditions and local emissions in built-up areas. In contrast, suburban stations like Huairou demonstrate the poorest performance (test $R^2=0.287$ for RFR),

likely because regional transport of pollutants weakens local weather-PM2.5 correlations.

Table 5: Model Performance Metrics (R^2 /RMSE/MAE)

Station	Model	R^2_{Train}	R^2_{Test}	RMSE _{Train}	RMSE _{Test}	MAE _{Train}	MAE _{Test}
Aotizhongxin	MLR	0.188	0.178	74.021	74.537	53.944	53.836
	RFR	0.461	0.337	65.133	69.123	47.468	49.970
Changping	MLR	0.155	0.151	65.941	67.954	48.491	49.303
	RFR	0.457	0.340	56.216	61.523	41.079	44.279
Dingling	MLR	0.153	0.145	67.105	65.478	48.690	48.192
	RFR	0.474	0.335	56.718	59.053	40.956	43.165
Dongsi	MLR	0.213	0.219	76.742	76.722	55.833	55.340
	RFR	0.501	0.411	67.263	70.584	48.959	51.219
Guanyuan	MLR	0.191	0.185	73.075	72.507	52.945	52.631
	RFR	0.463	0.366	64.945	67.030	47.160	49.145
Gucheng	MLR	0.175	0.178	75.272	74.926	53.918	53.742
	RFR	0.454	0.356	65.984	68.739	47.581	49.774
Huairou	MLR	0.117	0.112	67.277	66.298	49.268	48.953
	RFR	0.432	0.287	57.353	60.362	41.664	44.030
Nongzhanguan	MLR	0.220	0.225	76.351	75.519	55.229	55.291
	RFR	0.509	0.406	67.077	69.787	48.395	51.138
Shunyi	MLR	0.173	0.173	74.018	73.608	53.696	53.561
	RFR	0.453	0.350	64.362	67.306	46.436	48.655
Tiantan	MLR	0.210	0.206	72.270	71.382	52.263	52.236
	RFR	0.488	0.374	63.772	66.289	46.457	48.765
Wanliu	MLR	0.166	0.164	75.071	74.280	54.659	54.362
	RFR	0.425	0.334	66.881	68.877	48.628	50.350
Wanshouxigong	MLR	0.201	0.210	76.775	76.697	54.994	55.165

The models' relative performance varies spatially too—at Dingling, RFR reduces test RMSE by 9.8% compared to MLR, while at Wanshouxigong the improvement is just 6.1%.

Notable anomalies include Dongsi station, where MLR unexpectedly matches RFR's test performance ($R^2=0.219$ vs 0.411), suggesting linear relationships may suffice at this urban location. Meanwhile, Wanliu shows unusually poor RFR performance despite its urban setting, possibly due to microclimate effects from nearby water bodies. The consistent MAE values (45–55 $\mu\text{g}/\text{m}^3$ across stations) indicate both models struggle with extreme PM2.5 events, a critical limitation for pollution warning systems. These results underscore that while RFR generally outperforms MLR, its advantages are modest and station-specific, highlighting the need for localised model tuning in Beijing's heterogeneous airshed. The persistent low R^2 values across all stations suggest that effective PM2.5 forecasting requires incorporating non-meteorological predictors like real-time emission data.

4 CONCLUSION

This comprehensive evaluation of MLR and RFR models for PM2.5 prediction across Beijing's monitoring network yields several important insights with significant implications for air quality management. The analysis reveals that while both models demonstrate limited predictive capability using only meteorological variables, the Random Forest approach consistently outperforms traditional linear regression, albeit with notable spatial variations in performance. The urban stations, particularly Nongzhanguan and Dongsi, show relatively better model performance (test R^2 up to 0.41), likely due to a stronger coupling between local emissions and meteorological conditions in densely populated areas. In contrast, suburban stations like Huairou exhibit poorer performance, suggesting that regional pollutant transport and other non-local factors play a more dominant role in these locations. The consistent gap between training and test performance in RFR models (average ΔR^2 of 0.12) indicates moderate overfitting, emphasising the need for more robust regularisation or inclusion of additional relevant predictors.

The spatial patterns in model performance reflect Beijing's complex air pollution dynamics, where urban-scale processes appear more predictable than regional-scale phenomena. The superior performance of RFR models, particularly in urban settings,

suggests that nonlinear relationships between meteorological factors and PM_{2.5} concentrations are important to capture. However, the modest absolute performance levels ($R^2 < 0.5$ even for the best models) strongly indicate that meteorological variables alone are insufficient for accurate PM_{2.5} prediction. This limitation is particularly evident during extreme pollution events, as shown by the consistently high MAE values (45-55 $\mu\text{g}/\text{m}^3$). The station-specific variations in model performance, such as the unexpectedly strong showing of MLR at Dongsi or the poor RFR performance at Wanliu, highlight the importance of localised model development that accounts for microclimate effects and unique station characteristics.

These findings have several important implications for both research and air quality management. First, they underscore the need to incorporate additional predictors beyond basic meteorological variables, particularly emission-related indicators and wind pattern data. Second, they suggest that different modelling approaches may be warranted for different parts of the metropolitan area, with more sophisticated techniques like RFR being prioritised for urban core stations. Finally, the results indicate that current models have limited capability in predicting extreme pollution events, which should be a focus area for future model improvement. Future research directions should include testing more advanced machine learning architectures, incorporating real-time emission data, and developing ensemble approaches that combine the strengths of different modelling paradigms. Ultimately, while meteorological factors provide a useful foundation for PM_{2.5} prediction in Beijing, significant improvements in forecasting accuracy will require a more comprehensive approach that accounts for the full range of physical and chemical processes governing air pollution in this complex urban environment.

REFERENCES

- Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., Kaufman, J. D. 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation*, 121(21), 2331-2378.
- Guo, H., Zhang, Q., Cheng, Y., He, K. 2014. The impact of the Chinese New Year on air quality in China: A review. *Atmospheric Environment*, 98, 647-656.
- Li, J., Zhang, Y., Wang, X., Li, Z. 2019. The impact of the Spring Festival on air quality in China: A review. *Environmental Pollution*, 245, 707-718.
- Perrino, C., Tiwari, S., Catrambone, M., Dalla Torre, S., Rantica, E., Canepari, S. 2011. Chemical characterization of PM_{2.5} during the Diwali festival in Delhi, India. *Atmospheric Environment*, 45(34), 6123-6130.
- Pope, C. A., Dockery, D. W. 2006. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 56(6), 709-742.
- Tai, A. P. K., Mickley, L. J., Jacob, D. J. 2010. 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*, 44(32), 3976-3984.
- Wang, G., Zhang, R., Wang, L., Li, Z. 2014. Impact of meteorological conditions on PM_{2.5} concentrations in Beijing, China. *Atmospheric Chemistry and Physics*, 14(22), 12307-12322.
- Wang, Y., Zhang, Q., Wang, L., Zhang, R., Li, Z. 2017. The impact of meteorological factors on PM_{2.5} concentrations in urban areas: A case study in Beijing, China. *Atmospheric Environment*, 165, 52-63.
- Zhang, Q., Quan, J., Tie, X., Cao, J., Han, S., Wang, Z., Zhao, D., Li, J., Liu, X. 2015. The impact of meteorological conditions on PM_{2.5} concentrations in China: A review. *Atmospheric Environment*, 122, 823-833.
- Zhang, R., Wang, G., Zhang, Q., Li, Z. 2017. The impact of holidays on air quality in China: A review. *Environmental Pollution*, 231, 1234-1245.