STATS 531 Midterm Project

Abstract

We investigate the association between meteorological variables such as temperature, relative humidity, air pressure and wind speed in relation to fine and coarse PM in an industrial region in South Korea. We fit a linear regression model with ARMA(3,1) errors to account for temporal autocorrelation in the residuals. We further compare it to the regression model with SARIMA(3,1)(1,0,0,7) to explicitly account for weekly trends, performing a nested hypothesis test to select the better model. We conclude our model with ARMA(3,1) errors as our final model after failing to reject the null and interpret its result. Most notable of our findings is a positive association between temperature, and relative humidity, and a negative association between wind speed and fine PM with these associations growing weaker for coarse PM. We also note that relative humidity flips and shows a negative association for coarse PM.

1 Introduction

Particulate matter (PM) is a critical indicator of air pollution, harmful to the human body (Fuzzi et al. (2015)). It is not a single pollutant, but rather a complex mixture, ranging from extremely small particles and liquid droplets containing acids, organic chemicals, and metals, to soil and dust particles (Yoon et al. (2022)). Typically categorized into PM_{10} (coarse PM), which consist of particles with diameters less than 10 μm and $PM_{2.5}$ (fine PM), which consist of particles 2.5 μm or less (Yoon et al. (2022)), these particles harm the human body upon inhalation, causing diseases such as stroke and lung cancer (Yoon et al. (2022), Zhou, Chen, and Tian (2018)) with prolonged exposure even at low concentrations leading to reduced life expectancy (Pope III and Dockery (2006), Fuzzi et al. (2015)). Given these public health implications, studying and understanding the behavior of $\mathrm{PM}_{2.5}$ and PM_{10} is a pivotal step toward effective air quality management and public health protection.

We are interested in investigating whether there is a significant association between various meteorological variables and the concentration of $\mathrm{PM}_{2.5}$ and PM_{10}. Our initial suspicion is that there may be a positive association for temperature, relative humidity, and air pressure, and a negative association for wind speed. We further hypothesize that the associations may be weaker for coarse PM as fine PM tends to remain suspended in the air longer and travel farther (Yoon et al. (2022)). We base our initial hypothesis on scientific literature and intuition, further explored in the Supplementary (Section 4).

We analyze a dataset from the Asian Initiative for Clean Air Networks (AICAN), which maintains multiple air quality stations across South Korea (AICAN (2026)). In particular, we focus on the Sihwa Industrial Complex station, located near the coast in Gyeonggi Province. This station is at an industrially active region where both anthropogenic emissions and meteorological conditions are expected to significantly influence particulate matter concentrations. The dataset provides hourly measurements of PM$_{10}$($\mu g/m^3$) and PM$_{2.5}$($\mu g/m^3$), as well as meteorological variables such as Temperature ($^\circ C$), Relative Humidity (%), Air Pressure (hPa), and Wind Speed (m/s), spanning December 12th, 2019, to April 1st, 2024. The preprocessing steps are detailed in the Supplementary (Section 4).

2 Methods

2.1 Exploratory Data Analysis

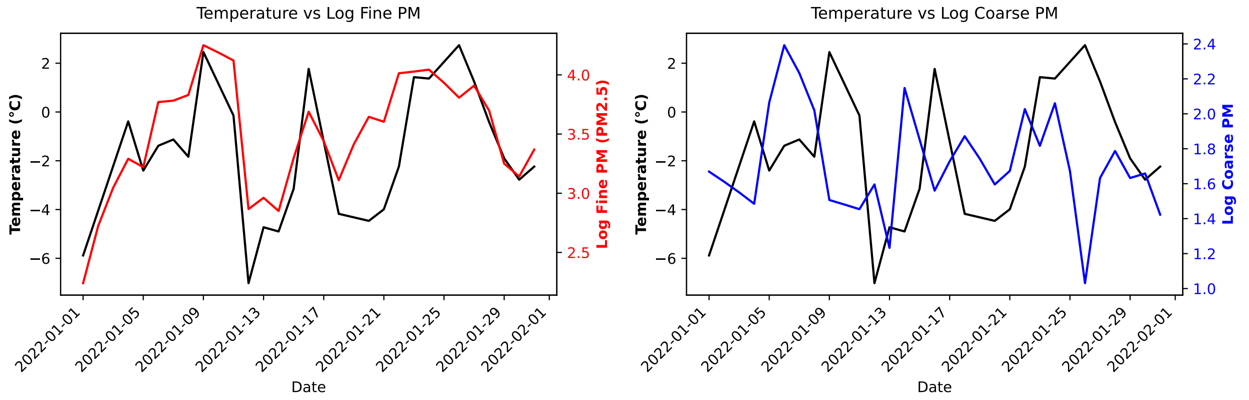


Figure 1: Comparison of Temperature's effect on Fine PM and Coarse PM for January 2022.

We first examined the relationship between temperature and wind speed in relation to PM by zooming in on a particular month, visualizing their daily fluctuations. Figure 1 and Figure 2 are for the window of January in 2022 and they show a fairly intuitive pattern that seems to suggest some signficant associations.

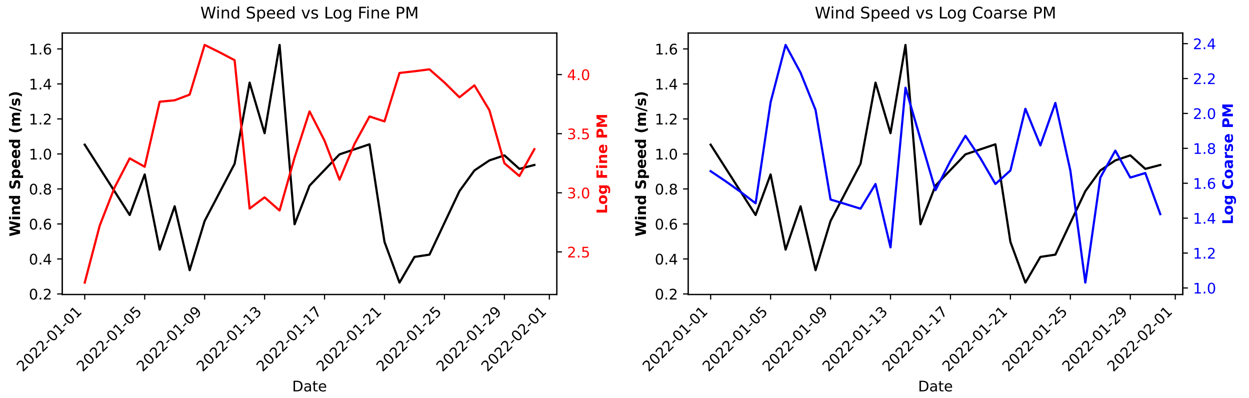


Figure 2: Comparison of Wind Speed's effect on Fine PM and Coarse PM for January 2022.

We also checked the correlation across different lags based on the method provided in class (Ionides (2026)).

First-differencing was applied prior to CCF analysis to account for autocorrelation that may be caused by shared seasonal trends(Rizzo, Scheff, and Ramakrishnan (2002)). These would inflate cross-correlations across all lags, obscuring genuine lag-specific relationships. The differening was applied specifically for this CCF analysis.

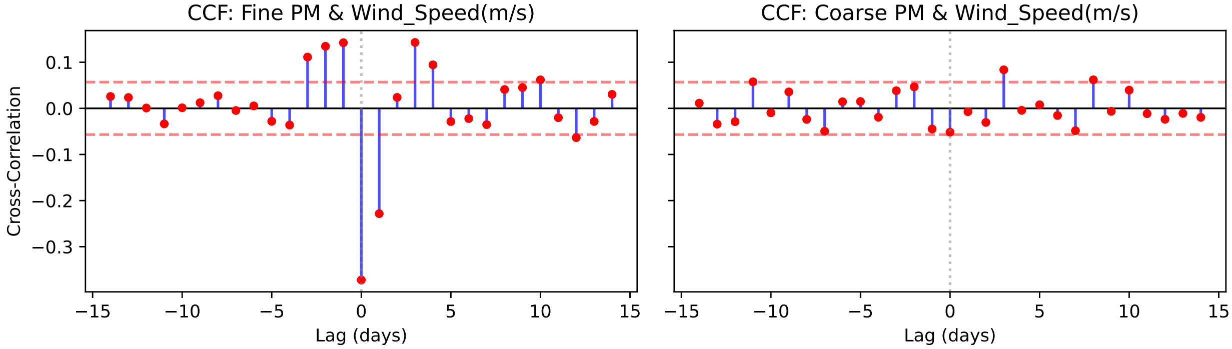


Figure 3: Cross-Correlation Functions for Fine and Coarse PM against Wind Speed.

Figure 3 reveals that cross correlation across various lags are stronger for fine PM than coarse PM. These results are encouraging because this is in line with our intuition, where fine PM is more sensitive to wind speed since it is lighter and stays in the air longer(Yoon et al. (2022)).

Other variables show similar results and are included in Section 4, although we do note that the pattern is not as clear for air pressure, noting strong cross correlation for coarse PM as well. However, the overall pattern is consistent across the other variables.

These preliminary studies do look promising, however, and we were encouraged to further investigate these associations.

2.2 Linear Regression

We initialize our experiment by fitting standard linear regression models to both fine and coarse PM by taking the meteorological variables as covariates. The model specification is straightforward and detailed in (Section 4).

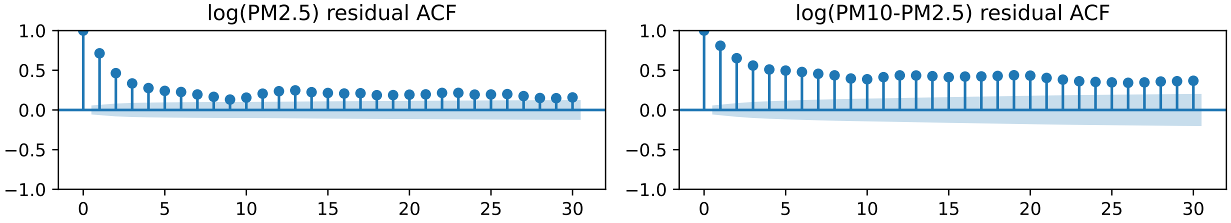


Figure 4: ACF of OLS residuals for fine and coarse PM

Residual diagnostics revealed significant autocorrelation in the OLS residuals for both models, as shown in Figure 5. This seems to indicate that the residual process exhibits temporal dependence that is not fully accounted for by the regression model. We conclude that our model assumptions are violated and the OLS estimates are unreliable even though the p-values appear significant.

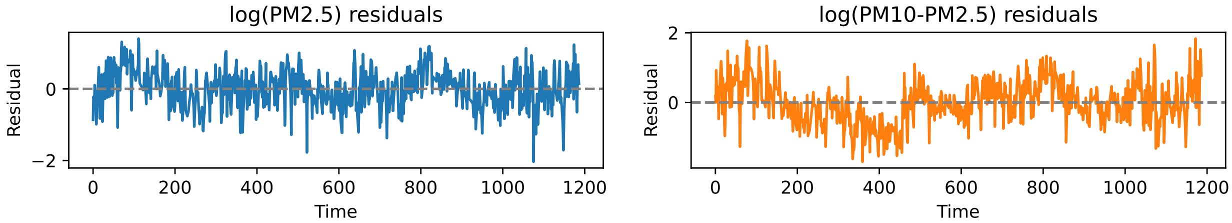


Figure 5: OLS residual time series for fine and coarse PM

2.3 Linear Regression with ARMA errors

We now fit the model by specifying the error term as an ARMA process(via SARIMAX with exogeneous regressors)(Ionides (2026)). We hope to see an improvement in residuals by doing so, capturing the temporal dependence.

To select ARMA orders, we conducted an AIC-based grid search over $(p, q) \in \{0, 1, 2, 3\} \times \{0, 1, 2, 3\}$ (Ionides (2026)). We restricted the search to small orders to avoid unnecessary complexity and overfitting. Table 3 and Table 4 summarize AIC values among converged fits and have been appended to the Supplementary section. We note that $(p, q) = (3, 1)$ seems a reasonable choice with low AIC.

We proceed with an ARMA(3,1) error structure for both the fine and coarse PM regression models.

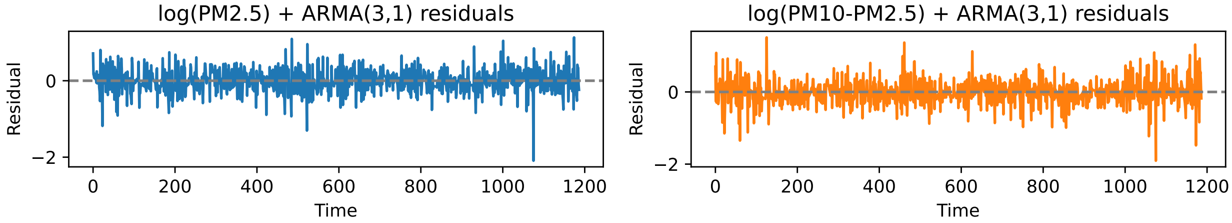


Figure 6: Residual time series from SARIMAX models (Fine: ARMA(3,1), Coarse: ARMA(3,1))

An examination of the residual time series shows that, for both models, the residuals fluctuate around zero with no clear trend or persistent pattern. The dispersion also appears broadly stable over time.

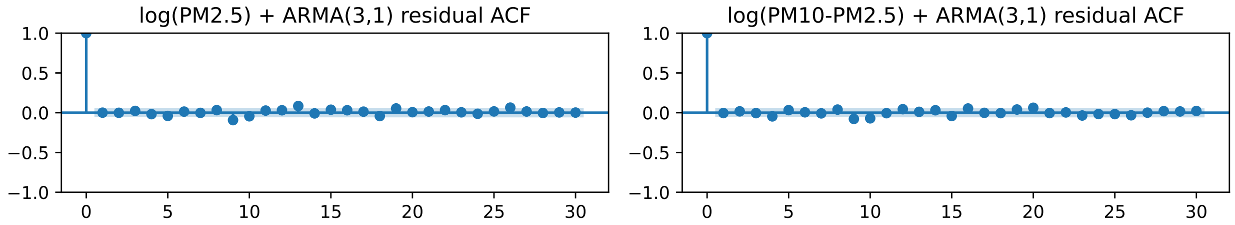


Figure 7: ACF of SARIMAX residuals (Fine: ARMA(3,1), Coarse: ARMA(3,1))

The residual ACF plot shows no statistically significant autocorrelation as well backing our visual inspection of Figure 6. We surmise that the ARMA(3,1) was effectively enough to absorb the remaining temporal dependence, yielding residuals as white noise.

We acknowledge that the ARMA(3,1) model may not be perfectly stable, however. It has a AR root near the unit circle boundary, too close for comfort. Assessing the roots for all combinations of ARMA orders that had decent AIC, we found that models seemed to share the same issue across the board. This suggests that an ARIMA model might be needed to difference the data. However, this is where we make our compromise. By using an ARIMA model, that would make interpreting our linear regression estimates difficult and much less intuitive to study the association between variables.

The roots are as follows:

```
AR roots: [1.00569075 2.60167148 2.60167148]
MA roots: [1.15351829]
AR roots: [1.00716259 2.12369157 4.138346  ]
MA roots: [1.09491249]
```

Acknowledging this concern, we check the estimates given by our linear regression ARMA error models.

Table 1: SARIMAX(3,1) results for log(PM2.5) (left) and log(PM10-PM2.5) (right)

Parameter	Est. (Fine)	SE (Fine)	p (Fine)	Est. (Coarse)	SE (Coarse)	p (Coarse)
Intercept	3.376***	0.508	3e-11	2.215***	0.263	3.7e-17
Temperature	0.473***	0.044	3.3e-27	0.223***	0.044	4.3e-07
Rel. Humidity	0.228***	0.015	6.8e-52	-0.178***	0.015	1.1e-30
Air Pressure	0.071**	0.031	0.02	0.054*	0.029	0.062
Wind Speed	-0.110***	0.011	2e-25	-0.039***	0.012	0.00077
AR(1)	1.737***	0.037	0	1.705***	0.028	0
AR(2)	-0.886***	0.050	1.3e-70	-0.821***	0.044	1.3e-76

Table 1: SARIMAX(3,1) results for log(PM2.5) (left) and log(PM10-PM2.5) (right)

Parameter	Est. (Fine)	SE (Fine)	p (Fine)	Est. (Coarse)	SE (Coarse)	p (Coarse)
AR(3)	0.147***	0.031	2.8e-06	0.113***	0.028	6.6e-05
MA(1)	-0.867***	0.031	2.1e-173	-0.913***	0.021	0
Sigma^2	0.084***	0.002	4e-297	0.104***	0.003	7.5e-288
N	1187			1187		
AIC	457.0			704.6		
BIC	507.8			755.4		

Looking at results in Table 1, we see some interesting results. We see that for temperature, results align with our initial hypothesis, with a coefficient of 0.4728 and a significant p-value less than 0.05. Recall predictors are standardized as stated in the preprocessing section of (Section 4), so one SD corresponds to one unit increase in the standardized scale. A one standard deviation increase in temperature is associated with a $(e^{0.473} - 1) \times 100\% \approx 60\%$ increase in raw fine PM, holding other variables fixed. Relative humidity displays significant positive association as well with fine PM, where a one standard deviation increase is associated with approximately a 26% increase in fine PM. Air pressure displays the weakest signal, with a one standard deviation increase related to about 7% increase in fine PM. Wind speed shows a negative association with a one standard deviation increase associated with approximately a 10% decrease. We find that all results seem to align well with our initial assumptions. The AR(1), AR(2), AR(3), and MA(1) terms are significant as well, indicating a reasonable temporal dependence structure in fine PM.

For coarse PM, we also find a positive and significant association with temperature where a one standard deviation increase is associated with a 25% increase in coarse PM. To our surprise, we find that relative humidity has flipped signs with a one sd increase associated with approximately a 17% decrease. Evidence suggests that coarse particles are already relatively heavy and when the relative humidity is high, water vapor condenses on these larger particles, dragging them down to the ground (Csavina et al. (2014)). Air pressure does not seem significant at the 5% level, which is reasonable since heavier particles would be less sensitive to pressure changes. Wind speed is still significant, although with only a 4% decrease associated with a one sd increase. The AR and MA components remain highly significant, suggesting that substantial serial dependence is present in the coarse particulate matter series as well.

We find that it is also true that the signals are generally weaker for coarse PM, which is consistent with our initial hypothesis that fine PM would be more responsive to meteorological conditions due to its smaller size and a longer airborn period.

It is strange, however, that our linear regression model with ARMA errors does so well. We initially suspected some seasonal or yearly trends and was expecting we would have to explicitly account for them in our model since it does not explicitly decompose seasonal or weekly components. However, the ARMA(3,1) error structure seems to have done a good job in absorbing the temporal dependence. We investigate deeper into why this may be so by looking at coherence and phase plots.

2.4 Coherence and phase analysis

We plot the coherence and phase between the PM and each meteorological variables. Codes are based on (Ionides (2026)). We list our most significant findings here and include the rest in the Supplementary (Section 4).

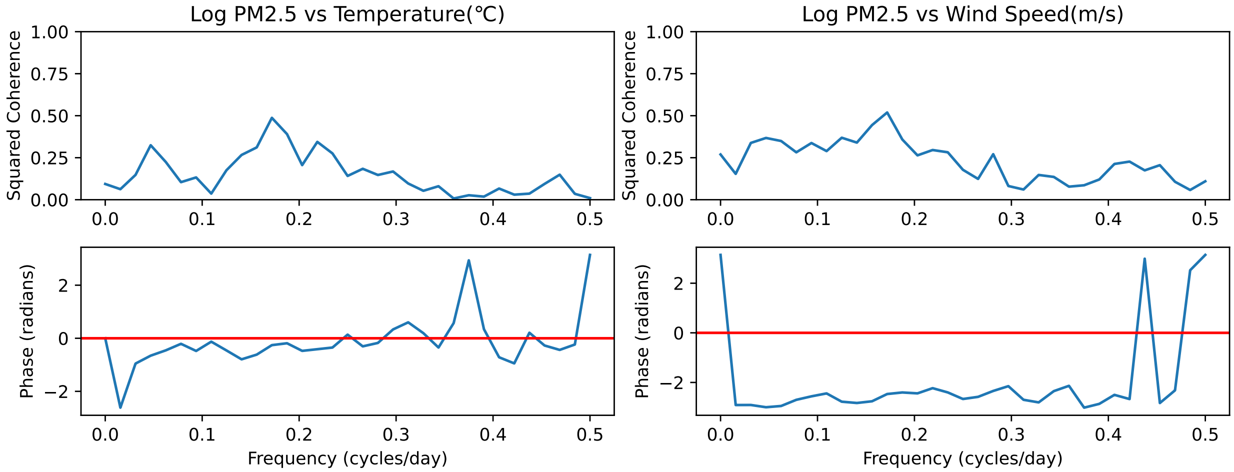


Figure 8: Coherence and phase plots for log(PM2.5) vs Temperature (left) and Wind Speed (right)

We note notable peaks in Figure 8. The strongest peak for the temperature plot is at around 0.15 corresponding to roughly a weekly cycle ($1/0.15 \approx 6.7$ days). This suggests that temperature and fine PM co-move most strongly at a weekly cycle.

We also see that wind speed and fine PM co-move across a wide range of frequencies with a notable peak at 0.2, which corresponds to a weekly cycle ($1/0.2 \approx 5$ days). The phase plot is consistently around $-\pi$ across almost all frequences where coherence is high. This further strengthens our finding that when wind speed is high, fine PM is low, and vice versa.

We suspect that some of the periodic structures are being captured by the temperature and wind_speed variables, which resulted in clean residuals without having explicitly controlled for them. However, adding explicit weekly SARIMA errors could more cleanly separate the weekly autocorrelation from the meteorological variables. We fit a new linear regression model and perform a likelihood ratio test between ARMA(3,1) and SARIMA(3,1)(1,0,0,7) to formally test whether the weekly component adds significant explanatory power.

2.5 Likelihood Ratio Test for Weekly Seasonal Components

To formally evaluate this hypothesis, we compare the baseline ARMA(3,1) specification with the SARIMAX model that incorporates an explicit weekly seasonal component (s = 7) implementing a likelihood ratio test, using Wilk's Approximation (Ionides (2026)).

This test compares two nested models: a nested model $H^{(0)}$ and the more general, nesting model $H^{(1)}$, where the parameter space of the restricted model is a subset of the general one, $\Theta^{(0)} \subset \Theta^{(1)}$, with dimensions $D^{(0)} < D^{(1)} \leq D$.

The maximum log-likelihood over each parameter space is notated as such.

$$\ell^{(0)} = \sup_{\theta \in \Theta^{(0)}} \ell(\theta), \qquad \ell^{(1)} = \sup_{\theta \in \Theta^{(1)}} \ell(\theta).$$

The difference in maximized log-likelihoods follows approximately a scaled chi-squared distribution:

$$\ell^{(1)} - \ell^{(0)} \approx \tfrac{1}{2}\chi^2_{D^{(1)}-D^{(0)}},$$

where the degrees of freedom equals 1 in this particular case.

Table 2: Likelihood ratio tests and AIC comparison: ARMA(3,1) vs SARIMA(3,1)(1,0,0,7)

	Target	LR Statistic	df	p-value	Significant at 0.05
0	Log_PM2.5	0.013571	1	0.907262	No
1	Log_PM10-PM2.5	0.037431	1	0.846591	No

Using a standard $\alpha = 0.05$ criterion, we reject H_0 only if $p < 0.05$ (equivalently, if $\chi^2_1 > 3.84$). For both fine and coarse PM, the LR test p-values are much larger than 0.05, so we fail to reject the null hypothesis that the simpler ARMA(3,1) error model is sufficient.

We conclude that adding the weekly component does not significantly improve model fit. Since our residual diagnostics of the baseline model Figure 6, also showed no clear patterns, this raises the possibility that meteorological variables already account for much of the weekly variation in fine PM.

3 Conclusions

We find strong evidence of associations between meteorological variables and PM concentrations at the Sihwa Industrial Complex. Temperature, relative humidity, and wind speed showed significant associations with fine PM, while coarse PM showed weaker associations overall. Notably, humidity reversed sign betwen fine and coarse PM. Coherence analysis revealed that these associations may be driven by underlying periodic structures, motivating us to fit a linear regression model with SARIMA errors to account for the weekly cycles that seemed to be present in the data. However, we still find that the linear regression with ARMA(3,1) errors is our best model with our nested hypothesis test failing to reject the null hypothesis that the simpler model is better.

One concern is that our AR roots are close to the unit circle boundary, which raises concerns for the reliability of our results. While differencing the response and fitting a linear regression model with ARIMA errors would address this, it would alter the interpretation of the regression coeffients. It would shift from associations with the level of PM to interpreting associations with their daily changes, which is much less intuitive and difficult to interpret. Since our residual diagnostics show no significant autocorrelation, we proceed with caution with our results. Future work could explore alternate approaches that resolve this non-stationarity, while preserving interpretabiility.

Comparison with previous projects

To our knowledge, there has not been a project that has investigated the relationship between meteorological variables and both fine and coarse PM in the same location. Hoewever, in carrying out this project, we were able to find steps where we could strengthen our project by learning from previous projects.

We reviewed a collection of past midterm projects and their corresponding peer reviews to guide our methodological decisions. A recurring critique in past peer reviews involves the improper handling of missing data. For instance, (Anonymous Authors (2024b)) heavily penalized the author for directly dropping dates with missing records. Reviewers noted that "dropping dates disrupts the continuity of the time series" and damages the discovery of underlying autocorrelation structures and seasonal patterns. This motivated us to search for an alternative method and found inspiration from (Mendoza, Castro, and Rodríguez (2024)).

Past projects fail to highlight the motivation behind fitting their time series model and often struggle to integrate residual diagnostics meaningfully in that context. Reviewers in (Anonymous Authors (2024a)) pointed out that the group wasted space by providing multiple residual plots only to write a generic sentence concluding "Gaussian white noise," lacking deeper diagnostic insight. To improve upon this, we placed a heavier emphasis on using the OLS residuals as the baseline and why this motivates us to fit a linear regression model with ARMA errors.

Acknowledgments

We acknowledge that any and all methods used in this project are based on materials provided by (Ionides (2026)). However, we do acknowledge the use of AI in designing the aesthetic of our plots to make them more visually appealing.

Bibliography

AICAN. 2026. "Particulate Matter Data." https://www.data.go.kr/data/15080316/fileData.do.

Anonymous Authors. 2024a. "Analysis of Battery-Powered Electric Vehicles Sales in the United States." STATS 531 Winter 2024 Midterm Project, University of Michigan. https://ionides.github.io/531w24/midterm_project/project16/blinded.html.

———. 2024b. "Investigating Trends in Household Electricity Consumption." STATS 531 Winter 2024 Midterm Project, University of Michigan. https://ionides.github.io/531w24/midterm_project/project12/blinded.html.

Csavina, J, J Field, O Felix, A Corral-Avitia, A Saez, and E A Betterton. 2014. "Effect of Wind Speed and Relative Humidity on Atmospheric Dust Concentrations in Semi-Arid Climates." *Science of the Total Environment* 487: 82–90. https://pmc.ncbi.nlm.nih.gov/articles/PMC4072227/.

Fuzzi, S, U Baltensperger, K Carslaw, S Decesari, H Denier van der Gon, M C Facchini, D Fowler, et al. 2015. "Particulate Matter, Air Quality and Climate: Lessons Learned and Future Needs." *Atmospheric Chemistry and Physics* 15: 8217–99. https://doi.org/10.5194/acp-15-8217-2015.

Ionides, Edward. 2026. "Notes for STATS 531, Modeling and Analysis of Time Series Data." https://ionides.github.io/531w26/.

Mendoza, Leonel, Wendy Castro, and José Rodríguez. 2024. "PM2.5 Time Series Imputation with Moving Averages, Smoothing, and Linear Interpolation." *Computers* 13 (12): 312. https://doi.org/10.3390/computers13120312.

Pope III, C A, and D W Dockery. 2006. "Health Effects of Fine Particulate Air Pollution: Lines That Connect." *J Air Waste Manag Assoc* 56: 709–42. http://dx.doi.org/10.1080/10473289.2006.10464485.

Rizzo, Michael, Peter Scheff, and Viswanathan Ramakrishnan. 2002. "Defining the Photochemical Contribution to Particulate Matter in Urban Areas Using Time-Series Analysis." *J Air Waste Manag Assoc* 52 (5): 593–605. https://doi.org/10.1080/10473289.2002.10470808.

Yoon, S, Y Heo, CR Park, and W Kang. 2022. "Effects of Landscape Patterns on the Concentration and Recovery Time of PM2.5 in South Korea." *Land* 11: 2176. https://doi.org/10.3390/land11122176.

Zhou, L, X Chen, and X Tian. 2018. "The Impact of Fine Particulate Matter (PM2.5) on China's Agricultural Production from 2001 to 2010." *Journal of Cleaner Production* 178: 133–41. https://doi.org/10.1016/j.jclepro.2017.12.204.

4 Supplementary material

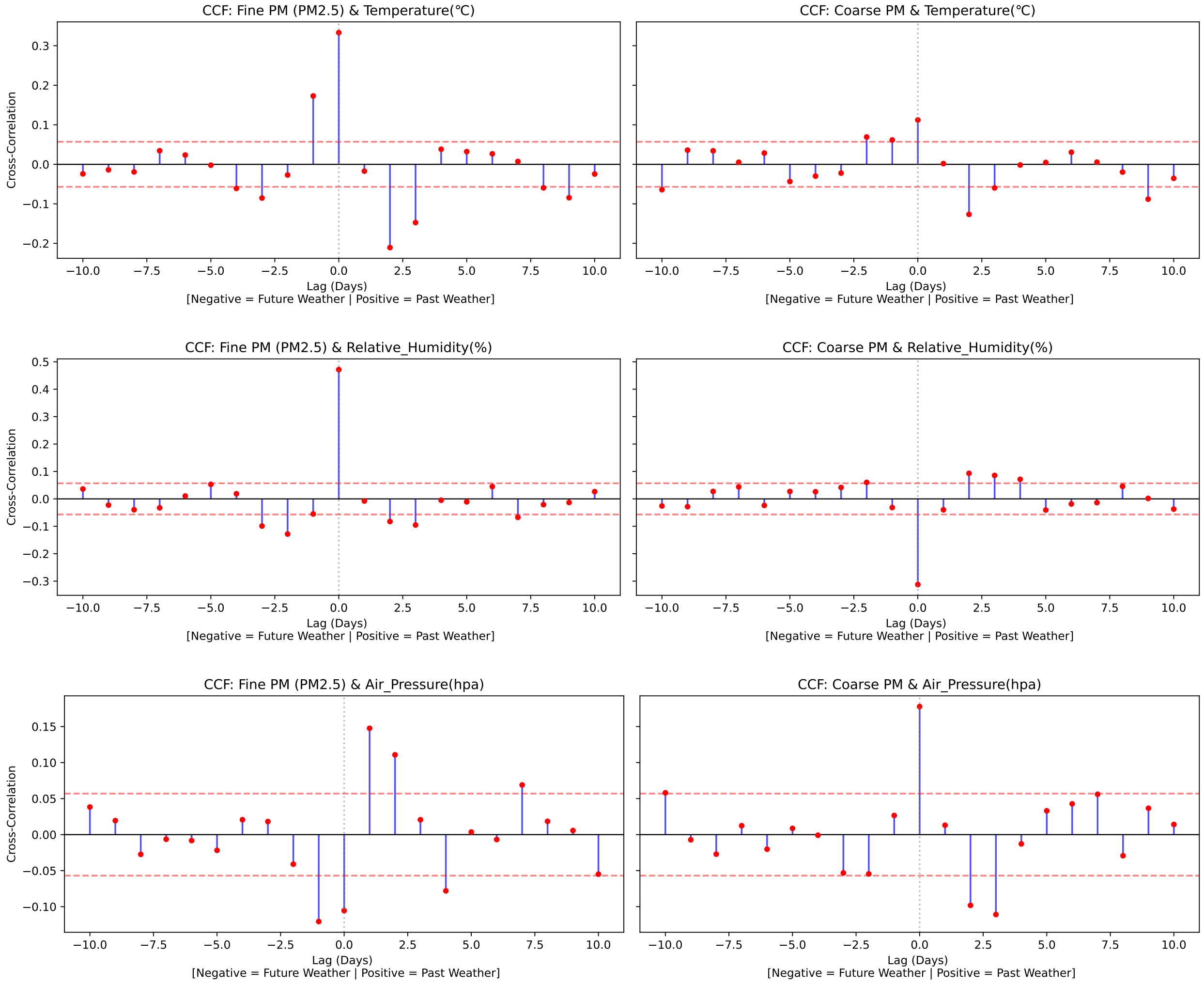
4.1 Preprocessing

We take daily averages of the hourly measurements to reduce noise and take into account weekly and seasonal trends. We also handle missing data, which is common in environmental datasets due to sensor malfunctions or maintenance. In particular, there is a huge missing gap between 2020-10-14 and 2020-12-30. We truncated the dataset to exclude this gap and worked with the dataset spanning 2021-01-01 and onward. The rest of the missing data in that time frame amounted to around 100 points, however, the longest consecutive missing period was only 8 days, which we imputed using linear interpolation, noting the procedures in (Mendoza, Castro, and Rodríguez (2024)). Note that analyzing diurnal patterns with hourly measurements seemed infeasible as the 8 consecutive missing points would blow up to 192 missing points, which would be too much to impute reliably.

We also log-transformed the PM measurements to stabilize the variance and make the data more normally distributed (Ionides (2026)). Since PM_{10} includes all particles with diameter $\leq 10\ \mu m$, and therefore already contains the $\text{PM}_{2.5}$ fraction, any variation in $\text{PM}_{2.5}$ would automatically appear in PM_{10} as well. To get rid of this dependence, we construct a new variable $\text{PM}_{10} - \text{PM}_{2.5}$, which represents the coarse particle fraction with diameter between 2.5 μm and 10 μm. By using $\text{PM}_{10} - \text{PM}_{2.5}$ as our coarse particle measure instead, we obtain two independent response variables, allowing for a cleaner comparison between coarse and fine particles. From henceforth, we will refer to $\text{PM}_{2.5}$ as fine PM and $\text{PM}_{10} - \text{PM}_{2.5}$ as coarse PM for simplicity.

For the meteorological variables, we experienced numerical issues in optimization since there was a significant difference in numbers due to varying units. We standardizd them to have mean 0 and standard deviation 1 to put them on the same scale to resolve this issue.

4.2 CCF plots



4.3 Linear Regression Model

The linear regression model is specified as follows:

$$y_t^{(k)} = \beta_0^{(k)} + \beta_1^{(k)}\,\mathrm{Temp}_t + \beta_2^{(k)}\,\mathrm{RH}_t + \beta_3^{(k)}\,\mathrm{Pressure}_t + \beta_4^{(k)}\,\mathrm{Wind}_t + \varepsilon_t^{(k)}, \quad k \in \{\mathrm{fine}, \mathrm{coarse}\}$$

$$\begin{aligned} y_t^{(\mathrm{fine})} &= \log(\mathrm{PM2.5})_t, \\ y_t^{(\mathrm{coarse})} &= \log(\mathrm{PM10} - \mathrm{PM2.5})_t. \end{aligned}$$

$$\begin{aligned} \mathrm{Temp}_t &: \text{temperature}, \\ \mathrm{RH}_t &: \text{relative humidity}, \\ \mathrm{Pressure}_t &: \text{air pressure}, \\ \mathrm{Wind}_t &: \text{wind speed} \end{aligned}$$

4.4 Model Selection

Table 3: AIC grid search results for log(PM2.5)

q	0	1	2	3
p				
0	1703.417	1055.932	846.168	791.176
1	545.957	547.778	507.963	476.443
2	547.835	547.489	457.835	459.500
3	523.593	456.983	458.818	461.825

Table 4: AIC grid search results for log(PM10-PM2.5)

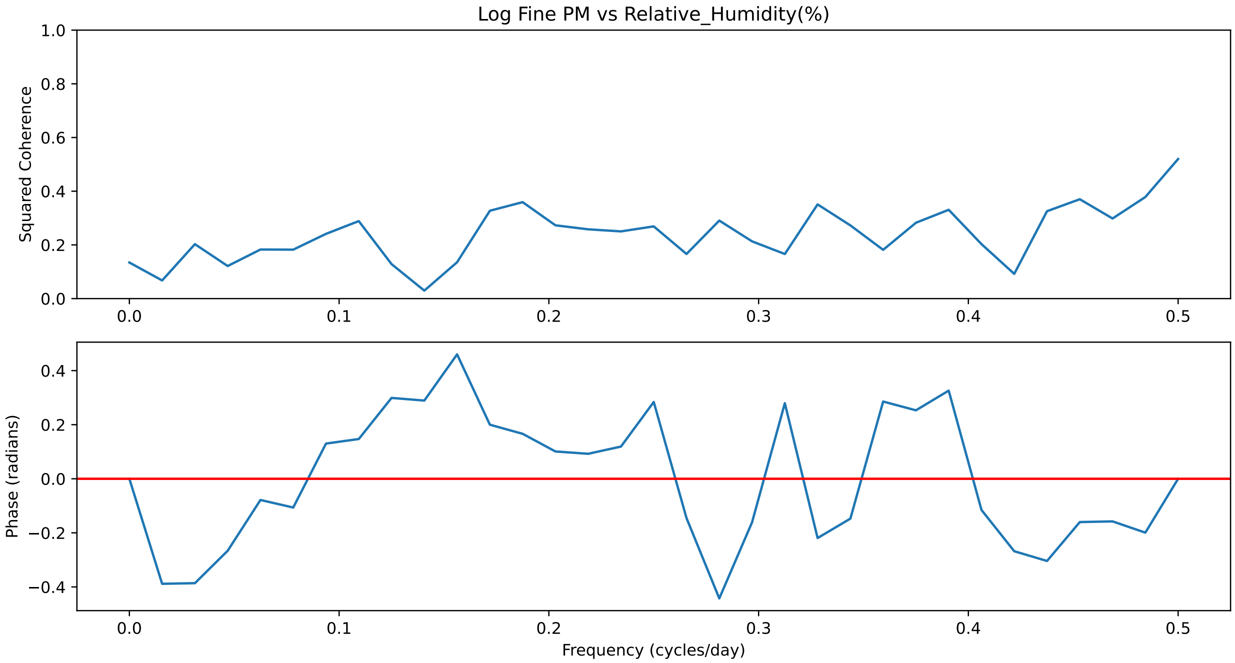
q	0	1	2	3
p				
0	2115.160	1340.225	1075.364	941.611
1	781.054	783.044	759.680	726.577
2	783.046	783.416	705.694	706.546
3	773.027	704.592	717.191	708.917

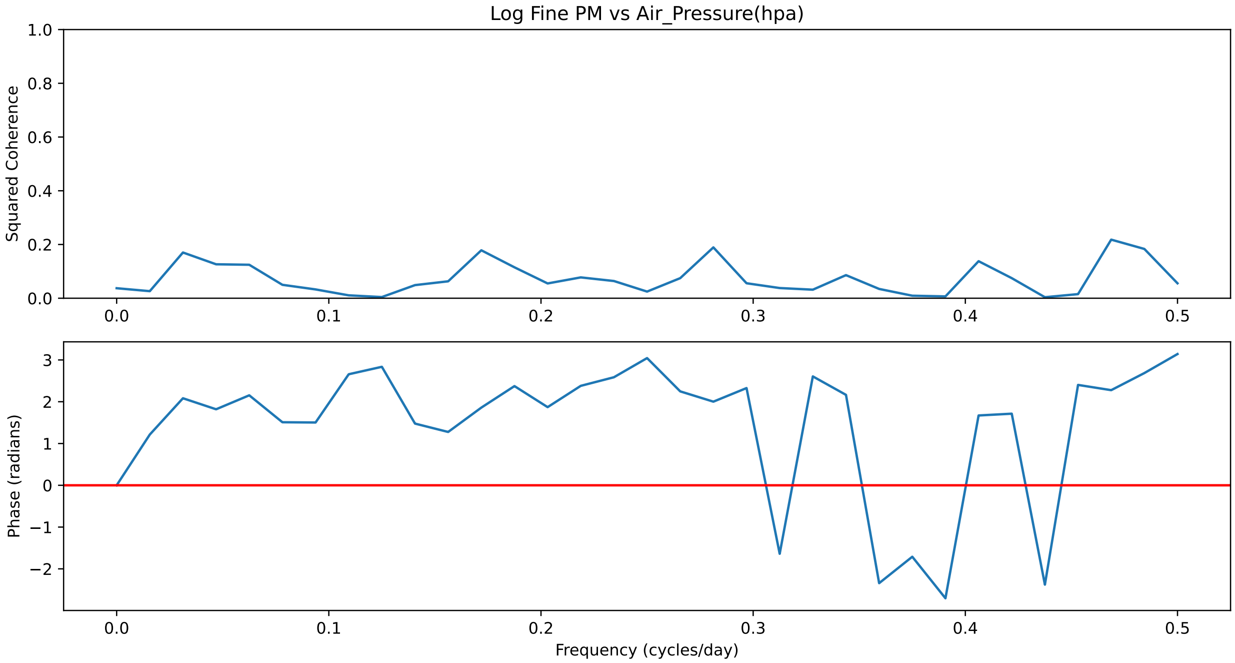
4.5 Roots

Below are a couple of other ARMA models we fitted in the hopes of avoiding the borderline non-causality issue. All models with decent AIC seem to share this issue, however.

```
ARMA(2,1)
AR roots: [1.05006008 1.74143005]
MA roots: [1.63302347]
AR roots: [1.05006008 1.74143005]
MA roots: [1.63302347]
ARMA(2,2)
AR roots: [1.00566251 1.84952728]
MA roots: [1.14197458 4.77802657]
AR roots: [1.00700633 1.80480728]
MA roots: [1.08928259 6.31199278]
```

4.6 Coherence and phase plots


Log Fine PM vs Relative_Humidity(%)


Log Fine PM vs Air_Pressure(hpa)