Impact of PM2.5 Exposure on Low Birth Weight in Ulaanbaatar from 2016–2025

Ahmadi, Barua, Bhuyan, Karayel

2025-04-30

Contents

List of Figures

List of Tables

1 Introduction

1.1 Rationale and Research Questions

Air pollution, particularly PM2.5, remains a significant public health concern in Ulaanbaatar. This project examines whether exposure to elevated PM2.5 concentrations during pregnancy is associated with an increased risk of low birth weight (LBW) among newborns between 2016 and 2025.

1.1.1 Research Questions

- What are the seasonal and long-term trends in PM2.5 concentrations in Ulaanbaatar?
- Is there a measurable association between PM2.5 exposure levels and low birth weight rates?
- Does the timing of PM2.5 exposure during pregnancy (examined through lagged exposure models) influence birth outcomes?

2 Dataset Information

2.1 Data Sources

Birth data consisted of monthly counts of live births and low birth weight (LBW) births for each district in Ulaanbaatar. Air pollution data included daily and monthly averages of PM2.5 concentrations, aggregated to monthly values. These datasets were merged by month and district for analysis.

2.2 Data Structure

- **PM2.5 data**:
  - Combined annual CSV files across multiple years.
  - Replaced invalid `-999` entries with `NA`.
  - Aggregated hourly readings into daily and monthly averages to align with birth record reporting.
- **Birth outcomes data**:
  - Reshaped from wide format to long format for time series analysis.
  - Merged live births and low birth weight counts into a unified dataset.
- **Final merged dataset**:
  - Monthly PM2.5 exposure data was linked with corresponding monthly birth outcome data by date.

Dataset	Key Variables	Notes
PM2.5 Pollution	DateTime, Raw Concentration, AQI	Aggregated to monthly averages
Birth Outcomes	Date, Aimag, Low Birth Weight count, Live Births	Cleaned and merged records

3 Data Wrangling and Cleaning

3.1 PM2.5 Data Analysis

```
#No we load and prepare the raw PM2.5 and birth outcome datasets as described previously.

# Defining file paths for PM2.5 data
years <- 2015:2025
pm25_files <- paste0("Data/Raw/Ulaanbaatar_PM2.5_", years, "_YTD.csv")
names(pm25_files) <- years

# Reading and binding all PM2.5 files
pm25_all <- map_dfr(pm25_files, ~ read_csv(.x, show_col_types = FALSE))

# Cleaning and converting -999 to NA
pm25_all <- pm25_all %>%
  mutate(across(where(is.numeric), ~ na_if(., -999))) %>%
  clean_names() %>%
  rename(DateTime = date_lt) %>%
  mutate(
    DateTime = parse_date_time(DateTime, orders = "ymd IMp"),
    Date     = as_date(DateTime)
  )

#We start by analyzing the structure, quality, and time-related aspects of the air pollution data.

# Daily average
pm25_daily <- pm25_all %>%
  group_by(Date) %>%
  summarize(
    raw_conc_daily = mean(raw_conc, na.rm = TRUE),
    aqi_daily = mean(aqi, na.rm = TRUE),
    .groups = "drop"
  )

# Monthly average
pm25_monthly <- pm25_daily %>%
  mutate(Month = floor_date(Date, "month")) %>%
  group_by(Month) %>%
  summarize(
    raw_conc_monthly = mean(raw_conc_daily, na.rm = TRUE),
    aqi_monthly = mean(aqi_daily, na.rm = TRUE),
    .groups = "drop"
  )
```

3.2 Birth Outcome Data Analysis

```
# Load birth outcome data
birth_weight_low <- read_csv("Data/Raw/BIRTH WEIGTH LOWER THAN 2500 GRAMS.csv")
live_births      <- read_csv("Data/Raw/LIVE BIRTHS.csv")

# Clean numeric columns in live births (in case of commas)
for (col in names(live_births)[-1]) {
  live_births[[col]] <- as.numeric(gsub(",", "",
                                        live_births[[col]]))
}

# Reshape both datasets to long format
birth_weight_long <- birth_weight_low %>%
  pivot_longer(-Aimag,
               names_to = "Month",
               values_to = "Low_Birth_Weight") %>%
  mutate(Month = gsub("^X", "", Month))

live_births_long <- live_births %>%
  pivot_longer(-Aimag,
               names_to = "Month",
               values_to = "Live_Births") %>%
  mutate(Month = gsub("^X", "", Month))

# Merge and finalize
births_merged <- left_join(birth_weight_long,
                           live_births_long,
                           by = c("Aimag", "Month")) %>%
  mutate(Date = ym(Month)) %>%
  select(Aimag,
         Date,
         Low_Birth_Weight,
         Live_Births)
```

4 Exploratory Analysis

4.1 PM2.5

4.1.1 Seasonality

```
ggplot(pm25_daily, aes(x = Date,
                       y = raw_conc_daily)) +
  geom_line(color = "steelblue") +
  labs(
    title = "",
    x = "Year",
    y = "Daily Mean PM2.5 (µg/m3) "
  ) +
  theme_minimal()
```

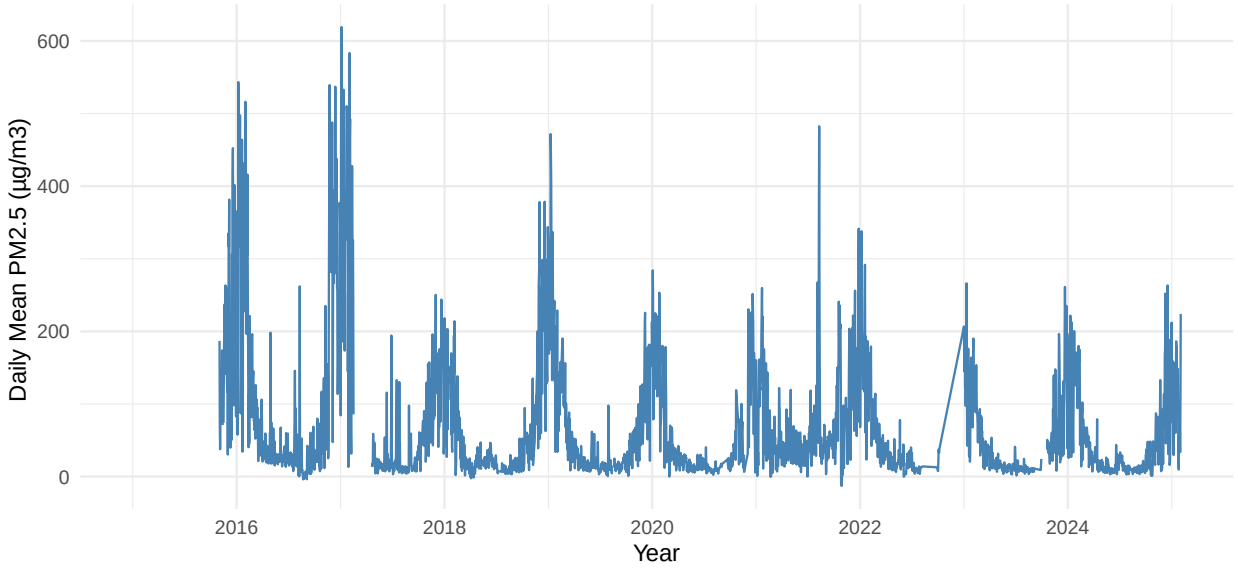

Figure 1: Seasonal Pattern of PM2.5 Concentation in Ulaanbaatar

The figure 1 shows a time series of daily mean PM2.5 levels from 2015 through early 2025. The data reveals a recurring pattern of sharp spikes in pollution levels each year, with the highest concentrations consistently occurring during the colder months. These peaks reach over 600 microgram per cubic meter in some years, notably in 2016 and 2017, suggesting severe air quality events. Although the magnitude of these spikes varies year to year, the pattern of elevated concentrations in specific periods remains consistent throughout the decade, indicating persistent seasonal pollution events.

4.1.2 Missing Data

```
#We evaluate missingness by calculating how many months each year had incomplete PM2.5 daily records.

pm25_yearly_missing <- pm25_monthly %>%
  mutate(Year = year(Month)) %>%
  group_by(Year) %>%
  summarize(
    months_with_missing = sum(is.na(raw_conc_monthly)),
    .groups = "drop"
  )

ggplot(pm25_yearly_missing, aes(x = Year,
                                y = months_with_missing)) +
  geom_col(fill = "tomato") +
  labs(
    title = "",
    x = "Year",
    y = "Months with Missing PM2.5"
  ) +
  theme_minimal()
```

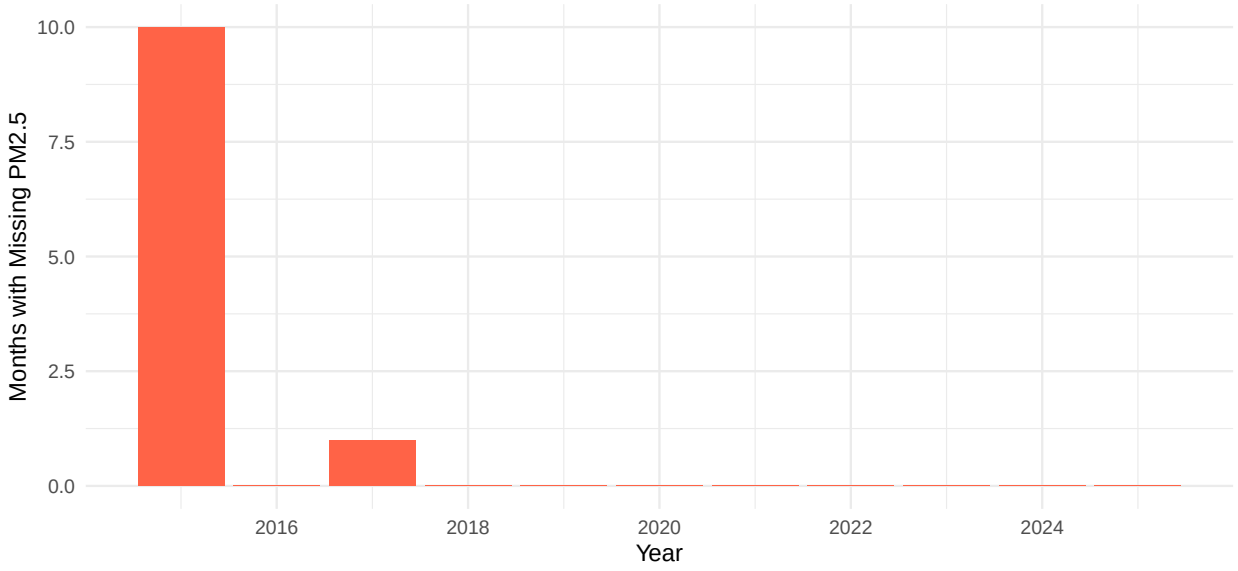

Figure 2: Number of Months with Missing PM2.5 Data

Our analysis in Figure 2 reveals that data gaps were mostly depicted in 2016, with 10 months lacking complete PM2.5 measurements. Additionally, one month in 2017 also shows missing data. From 2018 onward, there have been no recorded months with missing data. Therefore, the team had to analyse months with missing PM2.5 data.

4.1.3 Outlier Detection

```
#We also explore the variability of monthly PM2.5 to check for seasonal peaks and outliers.

ggplot(pm25_monthly, aes(y = raw_conc_monthly)) +
  geom_boxplot(outlier.color = "red") +
  labs(
    title = "Distribution of Monthly PM2.5 Concentrations",
    y = "Monthly Mean PM2.5"
  ) +
  theme_minimal()
```

The boxplot summarizes the variation in monthly mean PM2.5 levels in Ulaanbaatar. The plot illustrates that the median monthly PM2.5 concentration is just above 50 microgram per cubic meter, with the interquartile range (IQR)—representing the middle 50 percent of values—extending from approximately 20 to 110 microgram per cubic meter. Several outliers are displayed above the upper whisker, with some months reporting concentrations exceeding 300 microgram per cubic meter. These outliers likely indicate extreme pollution events occurring in particular months, aligning with the sharp peaks in Figure 1. The extended upper whisker signifies a right-skewed distribution, implying that while high PM2.5 values are less common, they can be significantly elevated when they do occur. In summary, this boxplot demonstrates that most monthly averages remain substantially above the WHO 24-hour guideline of 15 microgram per cubic meter, with a few months experiencing extremely high levels, underscoring the seriousness and fluctuations of air pollution in Ulaanbaatar.

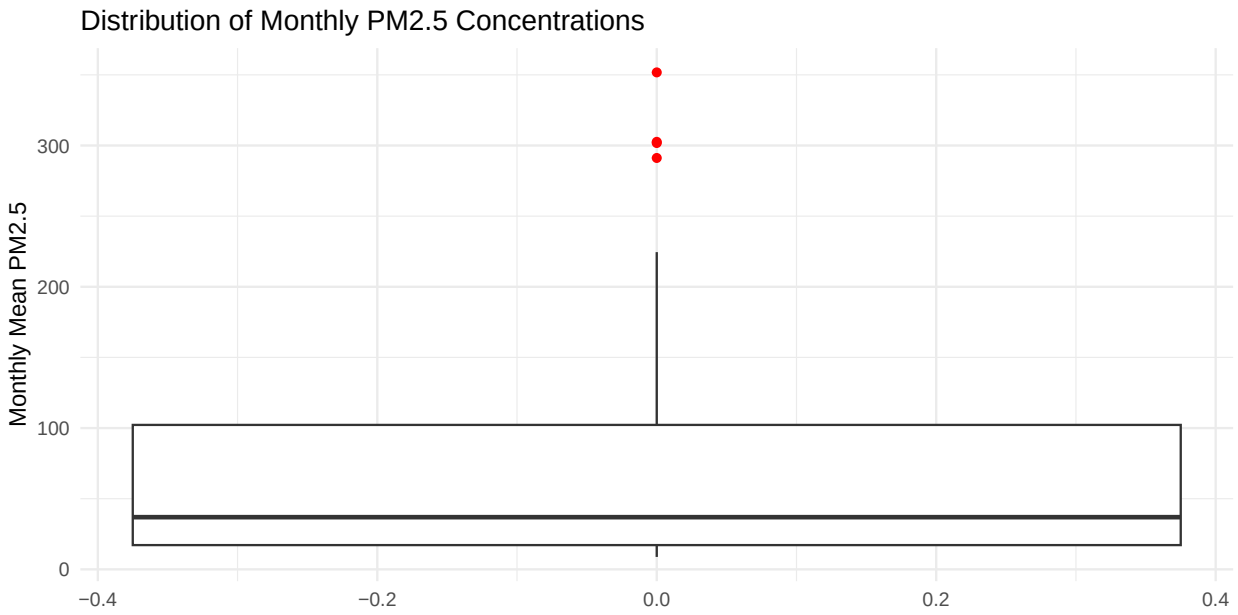

Figure 3: Distribution of Monthly PM2.5 Concentrations

4.2 Birth Outcome

4.2.1 Monthly Number of Birth

```
# Plot live births over time
ggplot(births_merged, aes(x = Date, y = Live_Births)) +
  geom_line(color = "darkgreen") +
  labs(
    title = "",
    x = "Date",
    y = "Number of Live Births"
  ) +
  theme_minimal()
```

```
# Plot live births over time
ggplot(births_merged, aes(x = Date, y = Low_Birth_Weight)) +
  geom_line(color = "purple") +
  labs(
    title = "Monthly Low Birth Weight",
    x = "Date",
    y = "Number of LBW Births"
  ) +
  theme_minimal()
```

The figures show that between 2016 and approximately 2021, the rate of LBW births was notably high and fluctuated significantly, consistently surpassing 160 births per month, occasionally exceeding 200. However, starting in 2022, a clear downward trend is evident, with both the average rate of LBW births and their levels of variability decreasing. By 2024, the number of LBW births often dips below 150 per month, marking some of the lowest figures recorded during the observed timeframe.

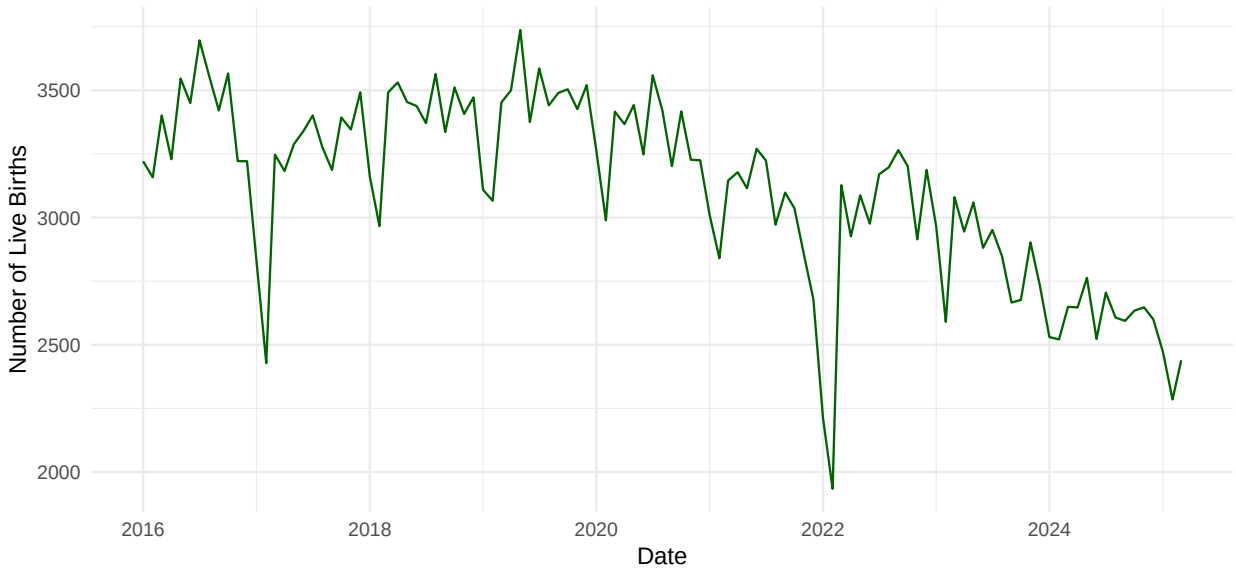

Figure 4: Monthly Live Births in Ulaanbaatar

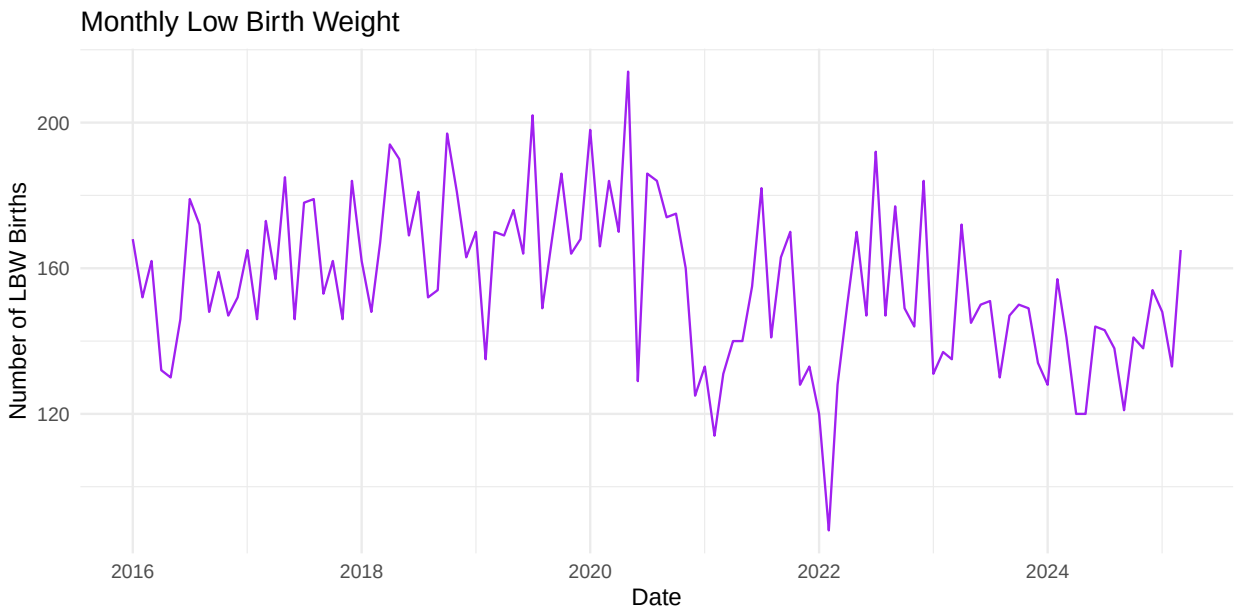

Figure 5: Monthly Low Birth Weight

4.3 Merge Exposure and Outcome Data

We link monthly PM2.5 exposure data with corresponding monthly birth outcomes. Therefore, in this section, we merge exposure and outcome datasets, calculate the low birth weight (LBW) rate, and begin preliminary modeling to investigate associations between PM2.5 exposure and birth outcomes.

```
# Merge datasets by Date
full_data <- births_merged %>%
  left_join(pm25_monthly,
            by = c("Date" = "Month")) %>%
  arrange(Date)
```

To make the comparison and to run the model we have created a new variable called Low Birth Weight Rate

LBW_rate = (Low Birth Weight / Live Births) * 100

```
full_data <- full_data %>%
  mutate(LBW_rate = 100 * Low_Birth_Weight / Live_Births)
```

4.4 Summary Statistics

```
summary_stats <- full_data %>%
  summarise(
    Mean_LBW_Rate = mean(LBW_rate, na.rm = TRUE),
    Median_LBW_Rate = median(LBW_rate, na.rm = TRUE),
    Mean_PM25 = mean(raw_conc_monthly, na.rm = TRUE),
    Median_PM25 = median(raw_conc_monthly, na.rm = TRUE)
  )

summary_stats %>%
  kable(caption = "", digits = 2)
```

Mean_LBW_Rate	Median_LBW_Rate	Mean_PM25	Median_PM25
5.03	4.96	67.28	35.22

4.5 Bivariate Visulization

```
ggplot(full_data, aes(x = raw_conc_monthly,
                      y = LBW_rate)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm",
              se = TRUE,
              color = "blue") +
  labs(
    title = "",
    x = "Monthly PM2.5 ",
    y = "Low Birth Weight Rate"
  ) +
  theme_minimal()
```

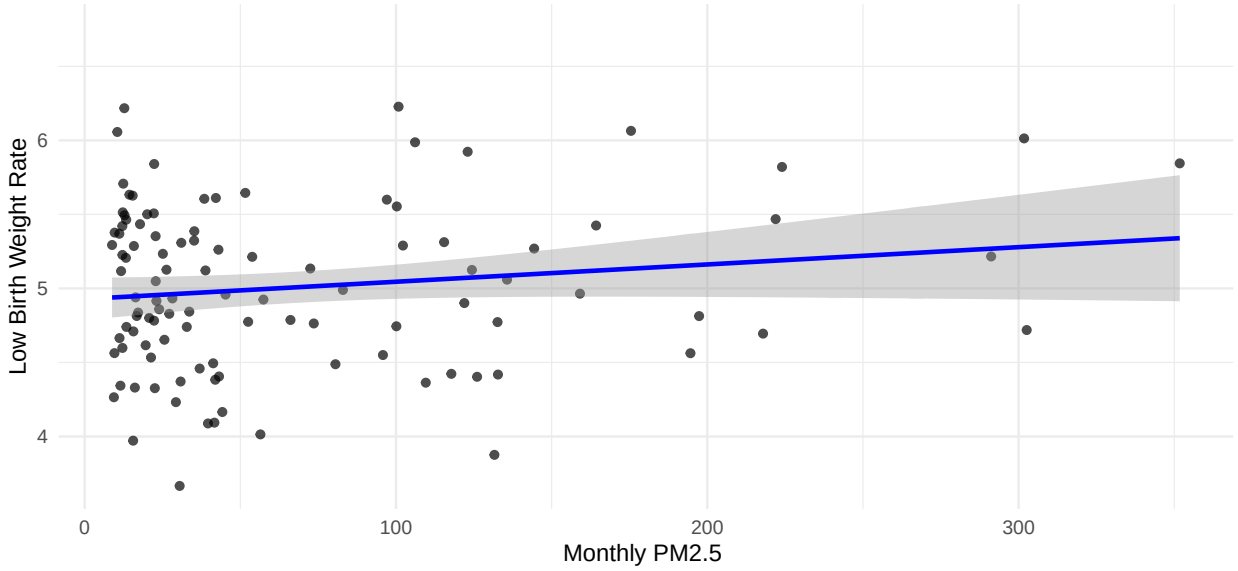

Figure 6: Relationship between Monthly PM2.5 and LBW Rate

The figure – reveals a positive trend, indicating that higher levels of PM2.5 are associated with slightly higher rates of low birth weight. Although the slope of the fitted line is not steep, the upward direction suggests that increases in air pollution may contribute to adverse birth outcomes. The shaded area around the line represents the 95% confidence interval, showing some uncertainty in the estimate, but the trend remains visible.

5 Methods and Models

5.1 Model Fitting

We fit a simple linear regression model at first LBW Rate ~ Monthly PM2.5 Concentration

We started the analysis with a simple linear regression to test the association between monthly PM2.5 concentrations and LBW rates in Ulaanbaatar. This initial model, which regressed LBW rates on same-month PM2.5 levels, produced a weak and statistically non-significant relationship [p = 0.113].

```
model_simple <- lm(LBW_rate ~ raw_conc_monthly,
                   data = full_data)

tidy(model_simple) %>%
  kable(caption = "Simple Linear Regression", digits = 3)
```

Table 3: Simple Linear Regression

term	estimate	std.error	statistic	p.value
(Intercept)	4.928	0.072	68.089	0.000
raw_conc_monthly	0.001	0.001	1.597	0.113

To account for the possibility that air pollution during pregnancy may affect birth outcomes with a delay, a distributed lag model was applied, incorporating PM2.5 exposures lagged by 0–3 months. However, this approach revealed high multicollinearity among lagged variables (variance inflation factors greater than 8 for some lags), which inflated standard errors and made it difficult to isolate the effect of any single month's exposure.

```
# Now We fit a distributed lag linear model:
#LBW Rate ~ PM2.5 exposure in current and previous 3 months

model_data <- full_data %>%
  arrange(Date) %>%
  mutate(
    pm25_lag0 = raw_conc_monthly,
    pm25_lag1 = lag(raw_conc_monthly, 1),
    pm25_lag2 = lag(raw_conc_monthly, 2),
    pm25_lag3 = lag(raw_conc_monthly, 3)
  )
model_lag <- lm(LBW_rate ~ pm25_lag0 + pm25_lag1 + pm25_lag2 + pm25_lag3,
                data = model_data)

tidy(model_lag) %>%
  kable(caption = "Distributed Lag Model", digits = 3)
```

Table 4: Distributed Lag Model

term	estimate	std.error	statistic	p.value
(Intercept)	5.053	0.097	52.155	0.000
pm25_lag0	-0.001	0.002	-0.369	0.713
pm25_lag1	0.002	0.003	0.750	0.455
pm25_lag2	0.001	0.003	0.441	0.660
pm25_lag3	-0.003	0.002	-1.742	0.085

Cumulative exposure models were then constructed by averaging PM2.5 over biologically plausible windows (e.g., 1–2 months and 1–3 months before birth) to capture sustained exposure during critical gestational periods. Despite this, the results remained non-significant, with small coefficients beta = 0.001, high p values [p = 0.158 for 1-2 months, p = 0.390 for 1-3 months]

```
#Finally, we fit simple models of LBW Rate on cumulative exposures:
cum_model_data <- model_data %>%
  mutate(
    pm25_cum12 = (pm25_lag1 + pm25_lag2) / 2,
    pm25_cum123 = (pm25_lag1 + pm25_lag2 + pm25_lag3) / 3
  )

model_cum12 <- lm(LBW_rate ~ pm25_cum12, data = cum_model_data)
model_cum123 <- lm(LBW_rate ~ pm25_cum123, data = cum_model_data)

bind_rows(
  tidy(model_cum12) %>% mutate(Model = "Cumulative Lag 1-2"),
  tidy(model_cum123) %>% mutate(Model = "Cumulative Lag 1-3")
) %>%
  select(Model, term, estimate, std.error, statistic, p.value) %>%
  kable(caption = "LBW Rate on Cumulative PM2.5 Exposures", digits = 3)
```

Table 5: LBW Rate on Cumulative PM2.5 Exposures

Model	term	estimate	std.error	statistic	p.value
Cumulative Lag 1-2	(Intercept)	4.944	0.080	62.183	0.000
Cumulative Lag 1-2	pm25_cum12	0.001	0.001	1.421	0.158
Cumulative Lag 1-3	(Intercept)	4.985	0.085	58.483	0.000
Cumulative Lag 1-3	pm25_cum123	0.001	0.001	0.863	0.390

Both cumulative models show a very small estimated increase in LBW rate (0.001 percentage points) per unit increase in cumulative PM2.5 exposure, but neither result is statistically significant. This suggests that, based on the available data, cumulative PM2.5 exposure in the last 2–3 months of pregnancy is not strongly or consistently associated with changes in LBW rate.

5.2 Result Diagnostics

Now we assess multicollinearity among the lagged PM2.5 variables using VIFs.A VIF greater than 5–10 indicates problematic multicollinearity.

```
# Calculate VIFs
vif_values <- vif(model_lag)

# Display neatly
vif_values %>%
  as.data.frame() %>%
  rownames_to_column("Predictor") %>%
  rename(VIF = ".") %>%
  kable(caption = "Variance Inflation Factors for Lagged PM2.5 Model", digits = 2)
```

Table 6: Variance Inflation Factors for Lagged PM2.5 Model

Predictor	VIF
pm25_lag0	4.06
pm25_lag1	8.89
pm25_lag2	8.08
pm25_lag3	3.44

Analysis:The high VIF values (especially for lags 1 and 2) suggest that the lagged PM2.5 variables are strongly correlated with one another, which can inflate standard errors and make it difficult to detect statistically significant effects in your model.

These findings contrast with robust associations reported in the literature, where studies using individual-level data and precise gestational timing (such as aligning PM2.5 exposure with the second or third trimester) consistently found significant effects (Amnuaylojaroen & Saokaew, 2024; Zhang et al., 2019). For example, multinational meta-analyses and cohort studies that adjusted for confounders like socioeconomic status and healthcare access have reported a 5–8 pecent increase in LBW risk per 10 microgram per cubic meter rise in PM2.5 (Lee & Holm, 2022). The lack of significance in the current analysis likely stems from methodological limitations, including the use of aggregated monthly data (which cannot pinpoint critical gestational windows), unmeasured confounders (such as maternal age and poverty), and ecological bias.

5.3 Discussion

The present analysis did not identify a statistically significant relationship between monthly PM2.5 concentrations and LBW rates in Ulaanbaatar, regardless of whether exposure was considered in the same month, in previous months, or as a cumulative average. This outcome contrasts with findings from other regions, where higher air pollution during pregnancy has been linked to increased risk of LBW, especially in low-income or high-exposure settings (Amnuaylojaroen & Saokaew, 2024; Lee & Holm, 2022; Zhang et al., 2019). The discrepancy likely reflects differences in data structure and methodology. Specifically, the use of monthly aggregated data in this study limited the ability to align pollution exposure precisely with the most sensitive periods of pregnancy. Additionally, the absence of individual-level data and adjustment for important confounders, such as socioeconomic status and access to healthcare, may have masked subtle or time-specific effects of PM2.5 exposure.

6 Limitations

Several limitations should be considered. First, monthly aggregated data limited the ability to pinpoint exposure during the most critical weeks of pregnancy, which is essential in air pollution and birth outcome research (Amnuaylojaroen & Saokaew, 2024; Zhang et al., 2019). Second, important confounders such as maternal age, income, and access to healthcare were not included, even though these factors are known to influence both pollution exposure and birth outcomes (Lee & Holm, 2022). Third, reliance on group-level data rather than individual pregnancy records may have introduced ecological bias. Fourth, high correlation between PM2.5 levels in adjacent months made it difficult to determine which period of exposure had the strongest effect. Finally, the sample size and time span may not have been sufficient to detect small or modest effects. Future research should use data that allow for more accurate timing of exposure and include additional variables to better understand how air pollution affects birth outcomes in Mongolia.

Key findings - PM2.5 concentrations in Ulaanbaatar showed strong seasonal variation, with extreme peaks in winter months. - The highest pollution levels occurred in 2016 and 2017, exceeding 600 microgram per cubic meter on some days. - Monthly low birth weight (LBW) rates showed a declining trend starting in 2022. - Simple and lagged linear models did not detect a statistically significant association between PM2.5 exposure and LBW rate. - Multicollinearity among lagged exposure variables was high, limiting the interpretability of individual lag effects.

#Conclusion While elevated PM2.5 exposure in Ulaanbaatar is clearly a recurring and serious public health concern, this analysis does not find strong statistical evidence that short-term monthly or lagged PM2.5 exposure is independently associated with changes in LBW rates. Future studies should explore longer exposure windows and consider individual-level birth data to strengthen causal inference as we lacked the birth data here.

7 References

Amnuaylojaroen, T., and Saokaew, S. (2024). Prenatal PM2.5 exposure and its association with low birth weight: A systematic review and meta-analysis. *Toxics, 12*, 446.

Lee, J. R., and Holm, S. M. (2022). The association between ambient PM2.5 and low birth weight in California. *International Journal of Environmental Research and Public Health, 19*, 13554.

Zhang, Y., Wang, J., Chen, L., et al. (2019). Ambient PM2.5 and clinically recognized early pregnancy loss: A case-control study with spatiotemporal exposure predictions. *Environment International, 126*, 422–429.

8 GitHub

All project files, including data and code, are available at the following GitHub repository:
AhmadiBaruaBhuyanKarayel

9 Acknowledgment of AI Assistance

ChatGPT was used for troubleshooting code syntax and LaTex formatting. All models and analysis were executed and validated locally in RStudio by the authors.