ESTIMATING INDOOR AIR QUALITY: DEVELOPING A PREDICTIVE MODEL FOR 7-DAY AVERAGE $PM_{2.5}$ CONCENTRATIONS IN ULAANBAATAR HOMES DURING PREGNANCY

INTRODUCTION

The UGAAR (Ulaanbaatar Gestation and Air Pollution Research) study is a significant research initiative designed to assess the impact of air pollution on health outcomes among pregnant women in Ulaanbaatar, Mongolia. Ulaanbaatar is known for its severe air pollution issues, particularly during the winter months when coal and wood burning significantly increases to meet heating needs. This city is characterized by high concentrations of airborne particulate matter ($PM_{2.5}$), which poses serious health risks (1,2).

The UGAAR study specifically focuses on non-smoking pregnant women, a group particularly vulnerable to air pollution. The research examines the effectiveness of portable HEPA (High-Efficiency Particulate Air) filters in reducing indoor levels of $PM_{2.5}$ and exposure to second-hand tobacco smoke, which are prevalent due to the common practices of indoor heating and smoking in enclosed spaces (2).

This setting provides a unique opportunity to study the direct impacts of air pollution control measures on health outcomes in a population exposed to exceptionally high levels of air pollutants. Ulaanbaatar's extreme climate and urban structure, coupled with its reliance on stoves for heating in densely populated ger districts (traditional Mongolian yurts), make it an ideal location for such studies (3).

Objective

The primary objective of this analysis is to develop a statistical prediction model capable of estimating the 7-day average $PM_{2.5}$ concentrations within the homes of pregnant women in Ulaanbaatar, Mongolia, throughout their pregnancy. By leveraging limited direct $PM_{2.5}$ measurements and supplementing them with data from questionnaires and government monitoring, the model will predict weekly indoor air quality. The analysis will also compute average $PM_{2.5}$ exposure for each trimester and the entire pregnancy period, integrating both environmental and household-specific variables to enhance predictive accuracy.

METHODS

Data Collection Methods

The data collection for this study involved a combination of direct measurements and secondary data derived from both remote sensing and local sources (3). The primary data were obtained from a cohort of pregnant women in Ulaanbaatar, Mongolia, as part of the UGAAR study. Indoor air quality was directly measured using instruments placed in participants' homes to record the 7-day average $PM_{2.5}$ concentration (2). Concurrently, outdoor $PM_{2.5}$ concentrations were recorded at local government monitoring sites. Additional environmental parameters, such as temperature and sulphur dioxide (SO_2) levels, were also measured to provide context to the $PM_{2.5}$ data (1, 3).

Remote sensing data were utilized to assess land use characteristics around participants' homes, which included measures of land cover brightness, greenness, and wetness, along with the density of roads and traditional Mongolian gers (yurts) within specific radii (1, 3). These measures help in understanding the spatial variables that could influence indoor air quality.

The data also included several variables from questionnaires administered to the study participants, capturing lifestyle and household characteristics such as the presence of smokers in the home, the number of meals cooked indoors, and the specific use of HEPA filters provided as part of the intervention study (2).

Dataset Overview

The dataset encompasses several key variables:

- **Indoor_$PM_{2.5}$ and Outdoor_$PM_{2.5}$**: Continuous variables representing the 7-day averaged $PM_{2.5}$ concentration inside the participants' homes and outdoors, respectively.
- **Environmental Factors**: These include wind stagnation index (WindS), outdoor temperature (T4), and wintertime average sulphur dioxide concentration (SO_2).
- **Land Use Characteristics**: Derived from remote sensing, these variables (e.g., $b5000^3$, $wet5000^3$, $Green100^3$) describe the immediate environment around the participants' homes, such as brightness, wetness, and greenness levels, as well as the density of gers and roads at various radii.
- **Household and Behavioral Data**: Data collected through questionnaires that include the number of HEPA filters deployed (Filters_deployed), the intervention status (Intervention), and the number of smokers in the home (Smokers_at_Home).

Prediction Modelling

To develop a predictive model for indoor $PM_{2.5}$ concentrations, a forward selection statistical method was employed. This method begins with an empty model, systematically adding variables one at a time. Each predictor is selected based on its statistical significance, adhering to a significance level entry criterion of 0.05. The process continues until no further variables meet the criterion for inclusion. The variables ultimately included in the final model were Season, Number of air cleaners deployed (comprising filter1 and filter2), Outdoor_$PM_{2.5}$, Number of gers in 750m radius circular buffer (ger750), Average land cover greenness in a circular buffer 100m (Green100), Number of gers in 5000m radius circular buffer (ger5000). These predictors were incorporated based on their ability to explain the variance observed in the log-transformed indoor $PM_{2.5}$ concentrations, ensuring that the final model is comprised only of those variables that most significantly enhance the model's predictive accuracy. This methodological approach guarantees that the model includes only the most statistically relevant variables, optimizing its effectiveness in predicting indoor $PM_{2.5}$ levels.

RESULTS

The results from the analysis illustrate a significant variation in $PM_{2.5}$ concentrations across different conditions and interventions within Ulaanbaatar. Key findings demonstrate the influence of seasonal changes, household smoking habits, the effectiveness of air cleaners, and external environmental factors on indoor air quality. The provided regression model further quantifies the impacts of these variables.

Table 1: 7-day average $PM_{2.5}$ concentrations in participant's homes in Ulaanbaatar, Mongolia

Stratum	N	Geometric Mean (μg/m³)	Geometric Standard Deviation (μg/m³)	Median (μg/m³)	25th Percentile (μg/m³)	75th Percentile (μg/m³)
Overall	382	20.5	2.0	19.3	12.4	34.1
Season of Measurement						
Spring	100	36.6	1.9	38.5	27.1	56.8
Summer	91	17.8	1.7	18.1	13.5	24.5
Fall	84	11.0	1.4	11.1	8.5	14.4
Winter	107	22.2	1.8	22.0	14.6	31.8
Live with smokers						
Yes	165	21.7	2.0	20.3	13.9	35.9
No	217	19.7	2.0	18.9	11.7	32.8
Intervention Status						
Control	187	24.5	2.0	24.5	14.3	45.1
Filter	195	17.3	1.9	17.1	11.3	25.8
Number of air cleaners deployed						
0	187	24.5	2.0	24.5	14.3	45.2
1	54	19.1	2.0	18.7	14.9	33.5
2	141	16.6	1.9	16.9	10.4	25.3

Table 1 offers a comprehensive summary of the 7-day average $PM_{2.5}$ concentrations across different categories. It quantifies how $PM_{2.5}$ concentrations differ across seasons, with spring showing the highest average $PM_{2.5}$ levels of 36.6μg/m³, which could be linked to specific seasonal activities or changes in external air quality that influence indoor conditions (4). The table shows the impact of smoking within homes, showing higher average $PM_{2.5}$ levels in households with smokers compared to those without. Furthermore, homes with air filters (intervention group) showed notably lower $PM_{2.5}$ levels compared to control homes, emphasizing the effectiveness of HEPA filters in reducing indoor particulate matter.

Table 2: Summary of Correlation coefficient and p-value between continuous variables

Variables	log_Indoor_$PM_{2.5}$ ($\mu g/m^3$)	Outdoor_$PM_{2.5}$ ($\mu g/m^3$)	Sulphur dioxide SO_2 (ppb)	Temperature Measured (^0C)	# of gers in buffer 5000m radius (gers/hectare)
log_Indoor_$PM_{2.5}$ ($\mu g/m^3$)	1.000	0.425 (<.0001)	0.130 (0.0110)	-0.414 (<.0001)	0.166 (0.0011)
Outdoor_$PM_{2.5}$ ($\mu g/m^3$)	0.425 (<.0001)	1.000	-0.043 (0.4021)	-0.824 (<.0001)	0.008 (0.8717)
Sulphur dioxide SO_2 (ppb)	0.130 (0.0110)	-0.043 (0.4021)	1.000	0.047 (0.3582)	0.902 (<.0001)
Temperature Measured (^0C)	-0.414 (<.0001)	-0.824 (<.0001)	0.047 (0.3582)	1.000	0.002 (0.9617)
# of gers in buffer 5000m radius (gers/hectare)	0.166 (0.0011)	0.008 (0.8717)	0.902 (<.0001)	0.002 (0.9617)	1.000

Table 2 shows the relationships between various key continuous variables affecting indoor air quality through correlation coefficients. A strong positive correlation is observed between indoor and outdoor $PM_{2.5}$ concentrations, indicating the influence of external air quality on indoor environments. Negative correlations between temperature and both indoor and outdoor $PM_{2.5}$ indicate lower pollution levels at higher temperatures, which may be due to differences in heating requirements.

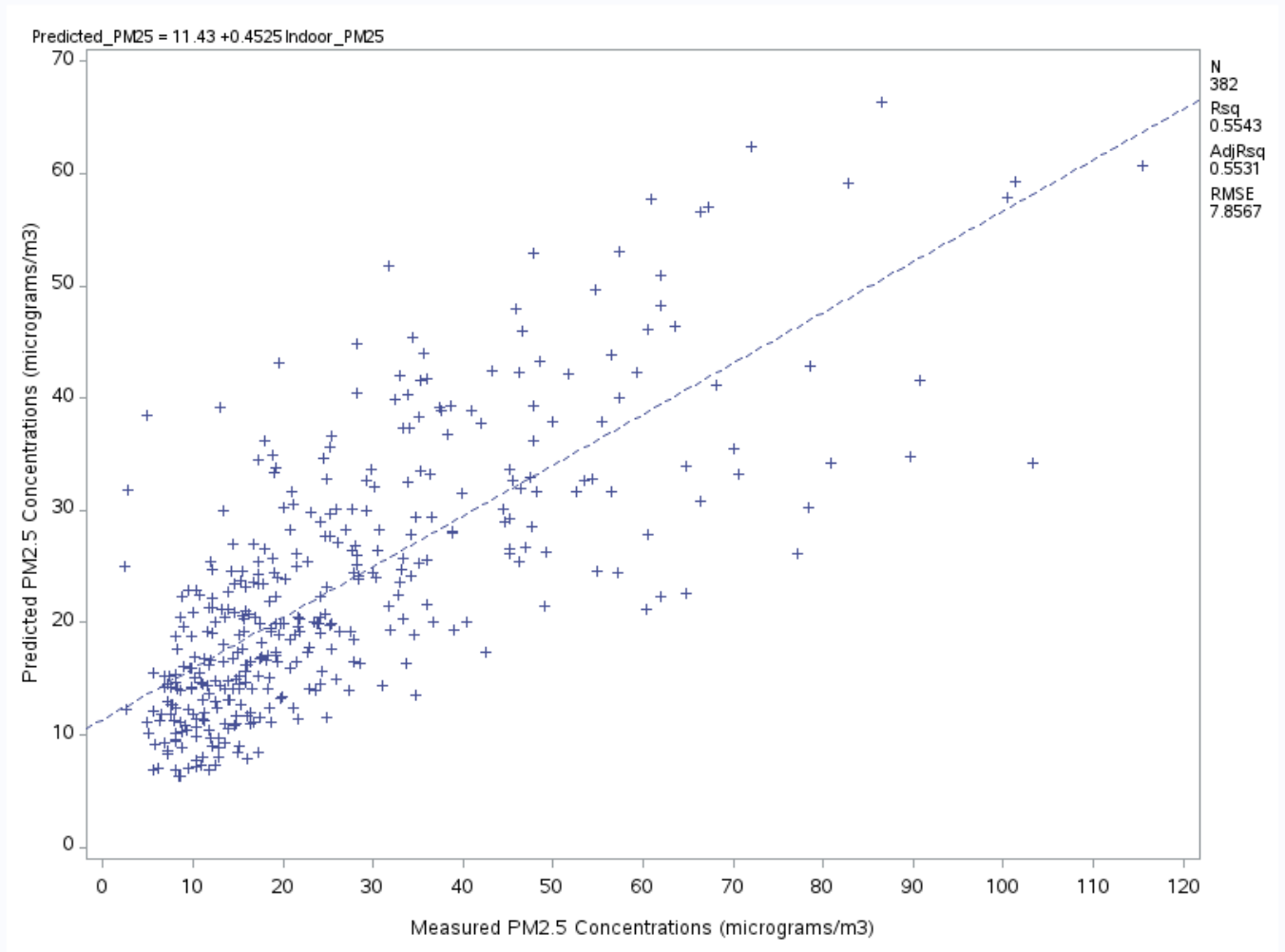
Table 3: Regression Coefficients and 95% Confidence Intervals for Predictors

Predictor Variable	Regression coefficient ($\mu g/m^3$)	95% confidence intervals
Intercept	3.008	(2.6, 3.4)
Season		
Summer	-0.510	(-0.7, -0.4)
Winter	-0.345	(-0.5, -0.2)
Fall	-1.060	(-1.2, -0.9)
Number of air cleaners deployed		
1-filter	-0.240	(-0.4, -0.1)
2-filters	-0.469	(-0.6, -0.4)
Outdoor_$PM_{2.5}$($\mu g/m^3$)	0.003	(0.001, 0.005)
# of gers in 750m radius circular buffer (ger750) (gers/hectare)	-0.001	(-0.002, -0.0004)
Average land cover greenness in a circular buffer 100m(Green100)	0.012	(0.000002, 0.02)
# of gers in 5000m radius circular buffer (ger5000) (gers/hectare)	0.206	(0.2, 0.3)

Table 3 presents the regression coefficients for each predictor in the model, quantifying their impact on indoor $PM_{2.5}$ levels. Seasonal coefficients indicate lower $PM_{2.5}$ levels in summer compared to other seasons, with fall showing the most significant reduction. Air cleaners (Filter1 and Filter2) demonstrate a substantial decrease in $PM_{2.5}$ levels, confirming their effectiveness. Additionally, the precision of the regression coefficients, indicated by tight confidence intervals, further supports the reliability and predictive quality of the model.

In the predictive model developed, the R-squared (R^2) value was 51.98%.

Figure 1: Scatter Plot Comparing Measured and Predicted Indoor PM2.5 Concentrations in Ulaanbaatar Homes



The scatter plot in Figure 1 compares the measured $PM_{2.5}$ concentrations against those predicted by the model, providing a graphical representation of the model's accuracy. The proximity of data points to the regression line indicates the effectiveness of the predictive model.

DISCUSSION

The predictive model developed to estimate indoor $PM_{2.5}$ concentrations achieved an R-squared value of 51.98%, indicating that approximately 52% of the variance in indoor $PM_{2.5}$ levels can be explained by the predictors included in the model. The significant predictors identified were seasonal variations, the number of air cleaners deployed, outdoor $PM_{2.5}$ concentrations, gers proximity, and land cover greenness. The regression coefficients as illustrated in Table 4, highlight the impact of seasonal variations, the number of air cleaners deployed, and the concentrations of outdoor $PM_{2.5}$ on indoor air quality. Specifically, the negative coefficients for the seasons relative to the reference category of spring imply lower $PM_{2.5}$ concentrations in the summer, fall, and winter months, demonstrating seasonal

improvements in air quality compared to spring. The negative coefficients for the air cleaners also demonstrate their effectiveness in reducing indoor $PM_{2.5}$ levels.

The predictors in the final model are both logical and grounded in empirical evidence, reflecting well-understood sources and dynamics of particulate matter:

- *Seasonal Variations*: The inclusion of seasons (with specific negative coefficients for summer, winter, and fall) as predictors makes sense, given the seasonal fluctuations in air quality due to changes in weather conditions, heating practices, and other seasonal activities (4). The negative regression coefficients suggest that $PM_{2.5}$ concentrations are generally lower in these seasons compared to spring, which may experience higher pollution levels possibly due to specific local practices like increased burning of coal for heating during the colder spring months (1-3).
- *Air Cleaners*: The number of air cleaners deployed (Filter1, Filter2) showing a negative coefficient confirms that the use of HEPA filters significantly reduces indoor $PM_{2.5}$ levels.
- *Outdoor $PM_{2.5}$*: The positive coefficient associated with outdoor $PM_{2.5}$ concentrations indicates that higher outdoor pollution levels lead to higher indoor levels, highlighting the influence of external air quality on indoor environments. This relationship underscores the permeability and infiltration of outdoor air into indoor spaces.
- *Proximity to Gers*: The positive coefficients for gers in the vicinity suggest that closer proximity to these traditional dwellings, which often burn coal and wood, increases indoor $PM_{2.5}$ levels.
- *Land Cover Greenness*: The positive coefficient of 0.012 indicates a direct relationship between the average land cover greenness and indoor $PM_{2.5}$ concentrations. Typically, increased greenness is associated with reduced ambient air pollution due to the capability of plants to absorb pollutants. However, the model has identified a small positive coefficient for greenness, suggesting a potential increase in $PM_{2.5}$ levels with increased land cover greenness. Further research into the types of vegetation, seasonal variations in plant behaviour, and specific local activities within green spaces might be required to fully understand this relationship.

Figure 1, an illustration of the model performance, clearly shows the relationship between the predicted and actual values. The regression line and the distribution of data points around it provide a straightforward visualization of the model's predictive accuracy. Additionally, it supports the R-squared value reported in the results, providing a visual representation of how much of the variance in indoor $PM_{2.5}$ the model captures.

STRENGTHS

This analysis incorporated data from multiple studies, including direct measurements of indoor $PM_{2.5}$ levels and land use regression (LUR) predictors, which offers a comprehensive view of the factors influencing air quality. Additionally, the model benefits from robust predictors such as season, the number of HEPA filters deployed, the density of gers, and outdoor $PM_{2.5}$ levels, which have a clear and direct impact on indoor air quality, as demonstrated in the UGAAR study (2). By accounting for seasonal variations, the model aligns with the recognized fluctuations in

$PM_{2.5}$ concentrations due to seasonal activities, as identified in the studies by Allen et al. and Barn et al. (1,2). The use of LUR predictors can capture complex spatial interactions and environmental variables that are otherwise difficult to quantify, enhancing the model's ability to predict PM2.5 levels accurately (3).

LIMITATIONS

One of the primary limitations encountered in our study was the absence of a designated dataset for model validation. Typically, the robustness of a predictive model is assessed by its performance on an independent dataset that was not used during the model-building process. In our case, however, the data utilized for developing the model was the full extent of the data available, leaving us without a separate validation set. This presents a challenge in thoroughly assessing the model's accuracy and potential overfitting.

Additionally, an R-squared value of 51.98% indicates that a substantial portion of the variance in $PM_{2.5}$ concentrations remains unexplained, pointing to other factors not captured by the model. Direct measurements of $PM_{2.5}$, even when conducted rigorously, are subject to measurement error, which can propagate through the model and affect the accuracy of predictions. Not all potential confounders, such as specific household behaviours, socioeconomic status, or adherence to using air filters, may have been fully accounted for in the model, despite some of these aspects being highlighted in the studies reviewed (1-3).

CONCLUSION

The model's effectiveness and the logical consistency of the predictor variables and their coefficients suggest that the predictive model is both robust and informative. It provides valuable insights into the factors affecting indoor air quality in Ulaanbaatar, Mongolia. Further analysis on evaluation could be done to enhance the model's accuracy.

REFERENCES

1.
Allen RW, Gombojav E, Barkhasragchaa B, Byambaa T, Lkhasuren O, Amram O, et al. An assessment of air pollution and its attributable mortality in Ulaanbaatar, Mongolia. Air Quality, Atmosphere & Health. 2011 Aug 9;6(1):137–50.
2.
Barn P, Gombojav E, Ochir C, Laagan B, Beejin B, Naidan G, et al. The effect of portable HEPA filter air cleaners on indoor PM2.5 concentrations and second hand tobacco smoke exposure among pregnant women in Ulaanbaatar, Mongolia: The UGAAR randomized controlled trial. Science of The Total Environment [Internet]. 2018 Feb;615:1379–89. Available from: https://www.sciencedirect.com/science/article/pii/S0048969717326426
3.
Yuchi W, Knudby A, Cowper J, Gombojav E, Amram O, Walker BB, et al. A description of methods for deriving air pollution land use regression model predictor variables from remote sensing data in Ulaanbaatar, Mongolia. The Canadian Geographer / Le Géographe canadien. 2016 May 13;60(3):333–45.
4.
Zareba M, Weglinska E, Danek T. Air pollution seasons in urban moderate climate areas through big data analytics. Scientific Reports [Internet]. 2024 Feb 6 [cited 2024 Apr 20];14(1):3058. Available from: https://www.nature.com/articles/s41598-024-52733-w