Modelling Fine Particulate Matter Concentrations inside the Homes of Pregnant Women in Ulaanbaatar, Mongolia

By

Weiran Yuchi

Bachelor of Arts (Health Sciences), Simon Fraser University, 2015

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Master of Science Program
Faculty of Health Sciences

SIMON FRASER UNIVERSITY
Summer 2017

Approval

Name:	**Weiran Yuchi**
Degree:	**Master of Science**
Title:	**Modeling Fine Particulate Matter Concentrations inside the Homes of Pregnant Women in Ulaanbaatar, Mongolia**
Examining Committee:	**Chair: Bruce Lanphear** Professor
Ryan Allen Senior Supervisor Associate Professor	_____
Scott Venners Supervisor Associate Professor	_____
Sarah Henderson Supervisor Assistant Professor School of Population and Public Health University of British Columbia	_____
Lawrence McCandless External Examiner Associate Professor	_____

Date Defended/Approved: June 30, 2017

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Fine particulate matter ($PM_{2.5}$) is a leading public health risk factor globally. Indoor concentrations are an important determinant of exposure because people spend the majority of time indoors. I developed models for predicting $PM_{2.5}$ concentrations inside the homes of pregnant women in Ulaanbaatar, Mongolia. The work was part of a randomized controlled trial of portable air cleaner use during pregnancy, fetal growth, and early childhood development. Multiple linear regression (MLR) and random forest regression (RFR) were used to model indoor $PM_{2.5}$ concentrations using 7-day indoor $PM_{2.5}$ measurements and potential predictors obtained from outdoor monitoring data, questionnaires, home assessments, and geographic data sets. The MLR (R^2 = 50.5%) and RFR (R^2 = 47.8%) models explained a moderate amount of variation in log-transformed indoor $PM_{2.5}$. Model predictions can be used to evaluate associations between indoor $PM_{2.5}$ concentrations during pregnancy and development in early life.

Keywords: intervention, prediction, air pollution, HEPA, RCT

Table of Contents

List of Figures

List of Tables

Chapter 1.

Introduction

1.1 Background

Exposure assessment plays a crucial role in air pollution epidemiology and risk assessment. Because people spend an average of 80 to 90% of their time indoors, pollution concentrations indoors are an important determinant of total personal exposure (Gauvin et al., 2002; Götschi et al., 2002). Indoor concentrations are the result of both indoor-generated pollution and outdoor-generated pollution that infiltrates indoors (Allen et al., 2012). While researchers may wish to quantify the exposure of every study participant at all times and in all locations, direct measurements of indoor concentrations and/or personal exposures are generally not feasible among large cohorts or over long periods (Cohen et al., 2009; Gerharz, Krüger, & Klemm, 2009; Shilpa & Lokesh, 2013). Thus, modelling techniques have been used to overcome the financial and logistic constraints of indoor and personal measurements, enabling researchers to predict concentrations in circumstances where direct measurements are unavailable (Shilpa & Lokesh, 2013).

1.2 Modeling techniques

One common modeling technique involves the use of mass balance conservation, predicting indoor air pollution concentrations as a function of relevant input parameters. These parameters include, but are not limited to, air exchange rate, indoor volume, and pollution emission and deposition rates (Byun, 1999; Ott, 1999). However, mass balance-based models have some limitations. Results produced by the models cannot be generalized to other locations or populations because the input parameters are unique to defined study sites. While the models may predict concentrations at finer spatial resolution compared with statistical models (Milner, Vardoulakis, Chalabi, & Wilkinson, 2011), they have limited applicability in multi-room settings, because each micro-environment (e.g. living room, kitchen, office, etc.) requires detailed information on each of the input parameters (Keil, 2000).

Other types of modeling techniques include Computational Fluid Dynamics (CFD) and Artificial Neural Networks (ANN). They are far more complex than the mass balance models. CFD

uses applied mathematics, physics, and computational software to predict and visualize how a gas flows. It applies momentum and energy equations in addition to the numerical solution of the mass balance (Anderson & Wendt, 1995; Ferziger & Peric, 2012). Input data requirements for CFD are extensive. In the context of indoor air pollution, CFD has been used to evaluate hypothetical scenarios, such as simulating the dispersion of volatile organic compounds in an apartment (Panagopoulos, Karayannis, Kassomenos, & Aravossis, 2011). However, this technique is not commonly used for modelling indoor exposures (Emmerich, 1997; Gan, 1995; P. Jones & Whittle, 1992; Nielsen, 2004; Sørensen & Nielsen, 2003).

In comparison, ANNs are a type of machine learning computational model, replicating the simple function of a biological network, such as the brain, where neurons exchange information. These models use many equations to convert input variables ("input neurons") into output. Input variables include, but are not limited to, outdoor air pollutant concentrations, temperature, relative humidity, volume, day of the week, and season. During the conversion, interconnections between neurons also involve a complex weighting system (Challoner, Pilla, & Gill, 2015; Kindangen, 1996; Sun & Hoff, 2009). Challoner et al (2015) applied ANN coupled with the personal-exposure activity location model (PALM) to model indoor nitrogen dioxide (NO_2) and particulate matter 2.5 ($PM_{2.5}$) concentrations in three inner city commercial buildings in Dublin (Challoner et al., 2015). PALM provided outdoor air concentrations at a particular building, and the ANN technique modeled the indoor/outdoor concentration relationship based on PALM and input variables mentioned above, yielding indoor air pollutant concentrations. However, extensive experience and knowledge is required for determining the weights put on interconnections, and the functions used to appropriately convert the input variables.

Recent indoor air quality modeling studies have made use of statistical regression methods, such as multiple linear regression (MLR). These empirical approaches are relatively easy to apply, because information about most variables that affect indoor pollutant concentrations can be easily collected through questionnaires or existing databases. This, in turn, may allow researchers to generalize results to larger settings and populations. However, MLR can be sensitive to outliers and involves a number of assumptions.

An alternative approach to MLR models is to apply random forest regression (RFR). RFR is an ensemble learning approach based on classification or regression techniques (depending on the class of dependent variables), involving decision tree applications (Chan & Paelinckx, 2008). A decision tree is a hierarchical, tree-like representation of decisions. It is an objective segmentation technique to iteratively split input data (called the node) into two or more data

samples (called the child nodes). This recursive partitioning of input data continues until certain specified conditions are met, such as when the best found split does not yield noticeable improvement or variation in a child node becomes too small (Liaw & Wiener, 2002; P. F. Smith, Ganesh, & Liu, 2013). In RFR, each node is split into two or more child nodes using the best in a subset of predictors that are randomly chosen at that node. The data in each child node are used to predict values of the dependent variable in that node. Results from all child nodes are then combined to produce final predictions (Figure 1) (Breiman, 2001; Liaw & Wiener, 2002; P. F. Smith et al., 2013).

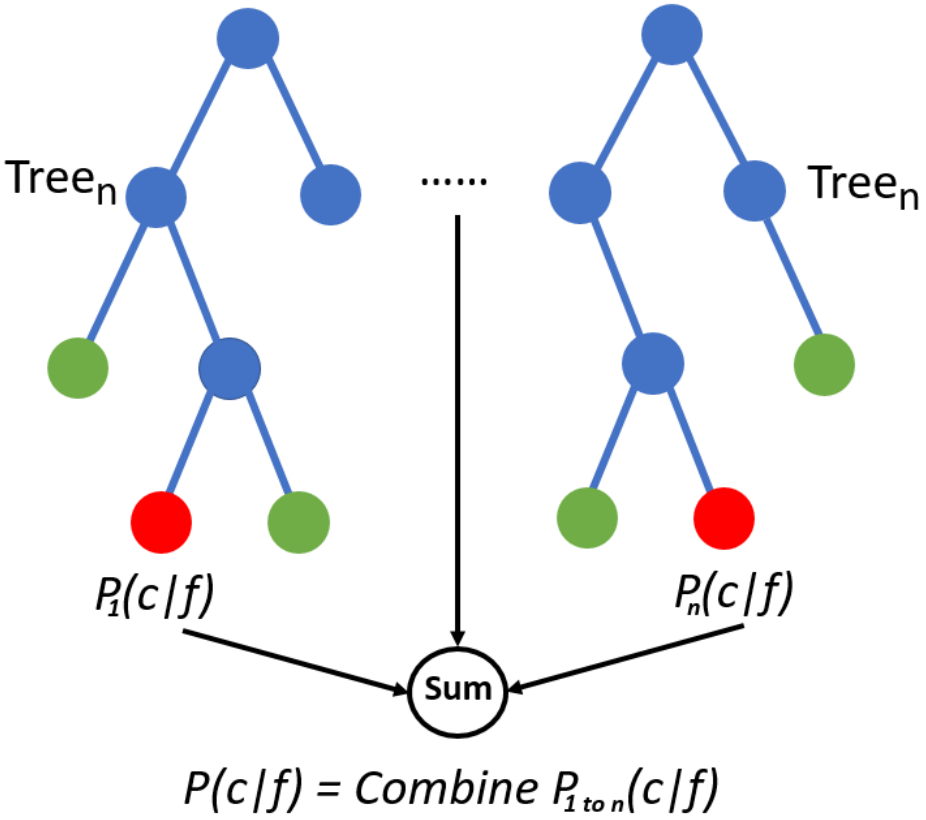

Figure 1: Illustration of random forest model. Blue dot: initial node, Green dot: split node, Red dot: terminal node. P represents predictions. c represents split condition. f represents split function. n represents number of tree (Jiang, 2016).

Because modeled results from each child node are averaged to produce predictions, RFR can minimize the effects of overfitting. In addition, unlike MLR, no data distribution assumptions are needed for RFR, and RFR is capable of detecting interactions and collinearities between predictors during the modelling process without the modeller selecting interaction terms. However, RFR has been criticized for being a “black box” when it comes to interpreting the model, because

of the large number of decision trees generated. In addition, for categorical variables with different numbers of levels, variable importance scores from RFR are not reliable because RFR is biased in favor of those with more levels (Cutler et al., 2007; Z. Jones & Linder, 2015; Prasad, Iverson, & Liaw, 2006; P. F. Smith et al., 2013).

In summary, mass balance models, CFD and ANN all require numerous input parameters and the quality of some parameters remains unknown due to uncertainty introduced during estimation and/or measurement. These limitations undermine their practical applicability in comparison with the use of MLR and RFR for the purposes of modelling indoor air pollution.

1.3 The UGAAR study

1.3.1 Overview

This thesis research made use of data collected as part of the Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study, a randomized controlled trial of high efficiency particulate air (HEPA) filter air cleaners and fetal growth. Ulaanbaatar, the capital city of Mongolia, has some of the highest air pollution concentrations in the world (Allen et al., 2013). The primary source of air pollution in Ulaanbaatar in wintertime is coal smoke from home heating stoves in the >100,000 gers (traditional Mongolian yurts) surrounding the city centre (Yuchi et al., 2016). Each ger stove burns an average of 5 tons of coal per year (Guttikunda, 2008). Spatial variation of ambient air pollutants within the city is generally driven by ger density and, to a lesser extent, local traffic patterns (Allen et al., 2013). Allen et al (2013) conservatively estimated that nearly 10% of the mortality in Ulaanbaatar can be attributed to ambient air pollution.

The primary aims of the UGAAR study were to: (1) understand the impact of HEPA filter air cleaners on indoor $PM_{2.5}$ concentrations; (2) assess the effect of $PM_{2.5}$ exposures on fetal growth; and (3) evaluate the efficacy of filter air cleaners to reduce the adverse effects of $PM_{2.5}$ on fetal growth. $PM_{2.5}$ is the pollutant of interest in UGAAR, both because it is the pollutant most consistently linked with impaired fetal growth (Ritz & Wilhelm, 2008; Shah & Balkhair, 2011) and because the HEPA filter air cleaners reduce particle concentrations but do not affect gaseous co-pollutants.

1.3.2 Study population

Initially, recruitment of participants was done in coordination with the reproductive health clinic at the Sukhbaatar District Health Center in Ulaanbaatar. This city district was targeted due to its large population living in apartments and its proximity to the ger area north of the city centre. To increase participant recruitment, we established a second study office in September 2014 at the 1st branch of the Sukhbaatar District Health Center. Women living in gers were excluded due to concerns about the reliability of electricity in ger neighbourhoods and the possibility that higher air exchange rates in gers would make portable HEPA filter air cleaners ineffective. Only women who were non-smokers, over 18 years of age, lived in an apartment, and in the first trimester of a single gestation pregnancy were eligible to participate. Women who were using an air cleaner at the time of enrollment or planned to give birth outside a hospital or clinic in the city were excluded.

UGAAR participants randomized to the control group received no HEPA filter air cleaners. Previous air cleaner studies (Sublett, 2011; Vijayan, Paramesh, Salvi, & Dalal, 2015) have used sham filtration to blind participants to their intervention status, but instead of purchasing sham air cleaners we chose to allocate resources to recruit a larger number of participants and deploy two air cleaners in larger apartments. Participants randomized into the intervention group received one or two Coway AP-1009CH HEPA filter air cleaners to use from enrollment to the birth of their child. The Coway AP-1009CH has a clean air delivery rate for tobacco smoke of 149 cubic feet per minute, which is appropriate for use in rooms up to approximately 22 m^2. The number of HEPA filter air cleaners deployed in each home was based on home size. All intervention group participants had an air cleaner placed in the main living area of the home, and a second unit was placed in the in the bedroom of the participant if the total home area was greater than 36 m^2.

The commercially available AP-1009CH has an internal PM sensor and “mood light” that changes colour based on the PM concentration, but this feature was disabled to avoid biasing the behaviour of UGAAR participants. The units used in UGAAR were also modified to operate on the second highest fan setting (out of four) and with an internal timer that counted total hours of use. Unfortunately, the internal timers proved to have limited value because initiating the timer required the air cleaner to be turned on while also pressing specific buttons. Participants were given instructions on the procedure, but if a participant turned on the air cleaner (e.g., after the unit was turned off, unplugged, or in the event of a power failure) without initiating the timer then all subsequent air cleaner usage was not logged. Thus, these timer usage data were not used. Air cleaner use in intervention homes was also quantified using information provided on the questionnaire administered at approximately 32 weeks gestation.

A total of 540 eligible women (269 in intervention group and 271 in control group) were recruited and gave informed consent. The number of participants enrolled into the study each season was similar between the control and intervention groups and generally similar across seasons, with slightly lower enrollment in summer. The control group and 1-HEPA air cleaner group had highest enrollment during fall and winter, while the 2-HEPA air cleaners group had slightly higher enrollment during winter and spring (Table 1). Twenty-seven participants (8 in intervention group, 19 in control group) were lost to follow up and 46 participants (18 in intervention group, 28 in control group) experienced miscarriage or stillbirth. Thus, 467 women (243 in the intervention group and 224 in the control group) were followed to the end of a successful pregnancy. The study protocol was approved by the Simon Fraser University Office of Research Ethics (2013s0016) and the Mongolian Ministry of Health Medical Ethics Approval Committee (No.7).

Table 1: Enrollment by season and number of air cleaners deployed of UGAAR participants included in this analysis

Groups/Seasons of enrollment	Number (%) of participants				
	Winter	Spring	Summer	Fall	Total
Control	61 (27.2)	53 (23.7)	52 (23.2)	58 (25.9)	224
Received 1 air cleaner	20 (27.0)	17 (23.0)	18 (24.3)	19 (25.7)	74
Received 2 air cleaners	45 (26.6)	48 (28.4)	36 (21.3)	40 (23.7)	169

1.3.3 UGAAR study data collection

Data collection took place from January 2014 to December 2015. Two home visits were made at approximately at 10 and 32 weeks of pregnancy for each participant. At each home visit, field technicians deployed air pollution monitoring equipment. Particle number concentrations were measured in the homes during two 7-day sampling periods using Dylos laser particle counters (Dylos DC 1700) (described in detail in section 2.1). The technicians also completed an assessment of the home and obtained the GPS location of the home during each home visit. The HEPA filter air cleaner(s) was/were deployed at the first home visit among the intervention group and continued to operate until participants gave birth. At approximately the same time as the home visits, study staff administered questionnaires to obtain information on housing characteristics, indoor pollution sources, and resident behaviours.

1.3.4 Thesis study objectives

Statistical regression methods were used to model indoor residential $PM_{2.5}$ concentrations in the homes of UGAAR participants. The objectives of this study were to: (1) build MLR models using outdoor $PM_{2.5}$ concentration measurements, meteorological data, questionnaires, geographic data, and predictions from previously developed land use regression models of outdoor SO_2 and NO_2 concentrations; (2) build a RFR model; (3) predict 7-day average indoor $PM_{2.5}$ concentrations in the homes of all UGAAR study participants over the course of pregnancy (for use in future epidemiologic analyses); and (4) assess and compare the model performance of MLR models with a RFR model.

Chapter 2.

Data sources

2.1 Indoor $PM_{2.5}$ measurements

Indoor $PM_{2.5}$ concentrations in the homes of UGAAR participants were measured using Dylos DC1700 laser particle counters (henceforth "Dylos", Figure 2). The Dylos uses light scattering and photodiodes to estimate the mass concentration of indoor $PM_{2.5}$ (Thompson, 2016). The performance of the Dylos has been validated in comparison with multiple conventional gravimetric $PM_{2.5}$ monitoring devices in both urban and rural settings (Brown et al., 2014; Northcross et al., 2013; Semple, Ibrahim, Apsley, Steiner, & Turner, 2013; Steinle et al., 2015) and the agreement has been consistently high, particularly indoors (R^2: 86 – 90%, Table 2). The commercially available Dylos logs particle number concentrations at 1-minute intervals and displays the counts in real time, but the units used in UGAAR were modified to log data at 5-minute intervals (to allow 7 days of data to be logged) and to not display real-time particle counts (to avoid biasing participant behaviour).

Figure 2: Indoor residential monitoring equipment (1: Dylos DC 1700, 2: HPEM, 3: HOBO UX100, 4: and Langan EL-USB-CO). (The HPEM, HOBO UX100, and Langan EL-USB-CO were deployed in a subset of homes. CO and temperature data were not included in this analysis)

Table 2: Summary of previous comparisons between Dylos optical particle counts and $PM_{2.5}$ concentrations

Setting	Measurement duration	$PM_{2.5}$ Monitoring device	R^2 (%)	Reference
Indoors (non-smoking homes)	24 hours	SidePak AM510	86	(Semple et al., 2013; P. F. Smith et al., 2013)
Indoor (smoking/non-smoking households)	7 days	DustTrak	98	(Klepeis et al., 2013)
Indoor (smoking chambers)	46 mins on average	SidePak AM510	90	(Semple, Apsley, & MacCalman, 2012)
Indoor	171 mins on average	SidePak AM510	90	(Dacunto et al., 2015)
Outdoor (rooftop)	24 hours	DustTrak 8520	98	(Northcross et al., 2013)
Rural Urban	24 hours	Tapered element oscillating microbalance (TEOM) and filter dynamics measurement system (FDMS)	90 70	(Steinle et al., 2015)
Outdoor	24 hours	Air monitoring stations at Westport, Connecticut USA	89	(Brown et al., 2014)
Outdoor	24 hours	Grimm Model EDM180	53	(Williams, Kaufman, Hanley, Rice, & Garvey, 2014)
Outdoor	5 mins	Grimm Model EDM180	54	(Williams et al., 2014)

Two 7-day indoor particle measurements were made over the course of pregnancy in the home of each UGAAR participant using the Dylos units. The first measurement was made shortly after enrollment at approximately 10 weeks gestation, on average, and the second measurement was made at approximately 32 weeks gestation. The Dylos units were placed in the main activity room, typically on a table or shelf, away from pollution sources, ventilation systems, and bright light sources.

In total, 911 7-day Dylos measurements (representing over 1.7 million 5-minute averaged concentrations) were made. Several quality control and data cleaning steps were completed prior to data analysis and modeling. First, fifty-seven 7-day measurements were removed if more than 10% of the 5-minute concentrations were 0 particles/cm^3. Next, data from four Dylos co-location experiments conducted by UGAAR field staff were used to identify and remove measurements made with faulty instruments. In these experiments, all 42 Dylos units were operated side-by-side in a room for approximately 24 hours. The data were downloaded and the median concentration across all units was calculated for each 5-minute period. For each Dylos the measured concentrations were then regressed against the median concentrations. Because measurements made with functional Dylos units were expected to agree with the median concentration across all units (i.e., $R^2 \approx 100\%$; slope ≈ 1, and intercept ≈ 0), any unit with co-location regression results that met any of the following criteria was considered faulty: 1) $R^2 < 80\%$; 2) slope < 0.80 or > 1.20; or 3) intercept > 163,830 particles (equivalent of ~ 10 µg/m^3 $PM_{2.5}$). These co-location experiments identified 18 poorly functioning units and all 344 7-day measurements made with those units were removed from the dataset. Finally, 50 measurements where the Dylos recorded a concentration for <50% of the 7-day monitoring period and 5 measurements with incorrect filenames that could not be matched to an UGAAR participant were excluded. These data cleaning steps left 455 7-day measurements for analysis.

Harvard personal environmental monitors (HPEMs) were deployed alongside the Dylos in a randomly selected sub-group of homes to estimate the relationship between Dylos particle counts and $PM_{2.5}$ concentrations. Because the relationship between Dylos particle number counts and $PM_{2.5}$ mass depends on the optical properties of the aerosol, these co-located measurements allow for an empirical estimate of the relationship specific to this setting. The HPEMs were loaded with 37-mm Teflon filters and connected to a pump (BGI 400; BGI, Inc., Waltham, MA) with a mass flow controller operated at 4 L/min. In total, 88 gravimetric filter 7-day measurements were collected. 41 gravimetric measurements were discarded due to pump failures or differences of > 10 % in flow rates at the start and end of the sampling period. The remaining 49 measurements

were matched with Dylos particle count data, resulting in 23 valid paired Dylos-gravimetric $PM_{2.5}$ measurements. Based on these 23 co-located 7-day measurements in UGAAR homes the agreement between the Dylos and $PM_{2.5}$ measured with HPEMs had an R^2 of 94.0% (Figure 3). Dylos particle counts were converted to $PM_{2.5}$ concentrations using the regression relationship (Figure 3).

Relative humidity (RH) was logged at 5-minute intervals using HOBO loggers (ux100-011; Onset Computer Corporation; Bourne, MA, USA) in the homes selected for gravimetric $PM_{2.5}$ monitoring with HPEMs. RH was considered because it could impact the light scattering properties of particles, thereby influencing the relationship between Dylos particle counts and $PM_{2.5}$ mass concentrations. The Dylos has previously been shown to record artificially high particle counts when RH exceeds approximately 90 % (Williams et al., 2014). RH was concluded to not influence Dylos particle counts in UGAAR based on (a) the low RH values measured in homes (all 5-minute RH measurements < 85%), and (b) the weak relationship between RH and Dylos particle counts.

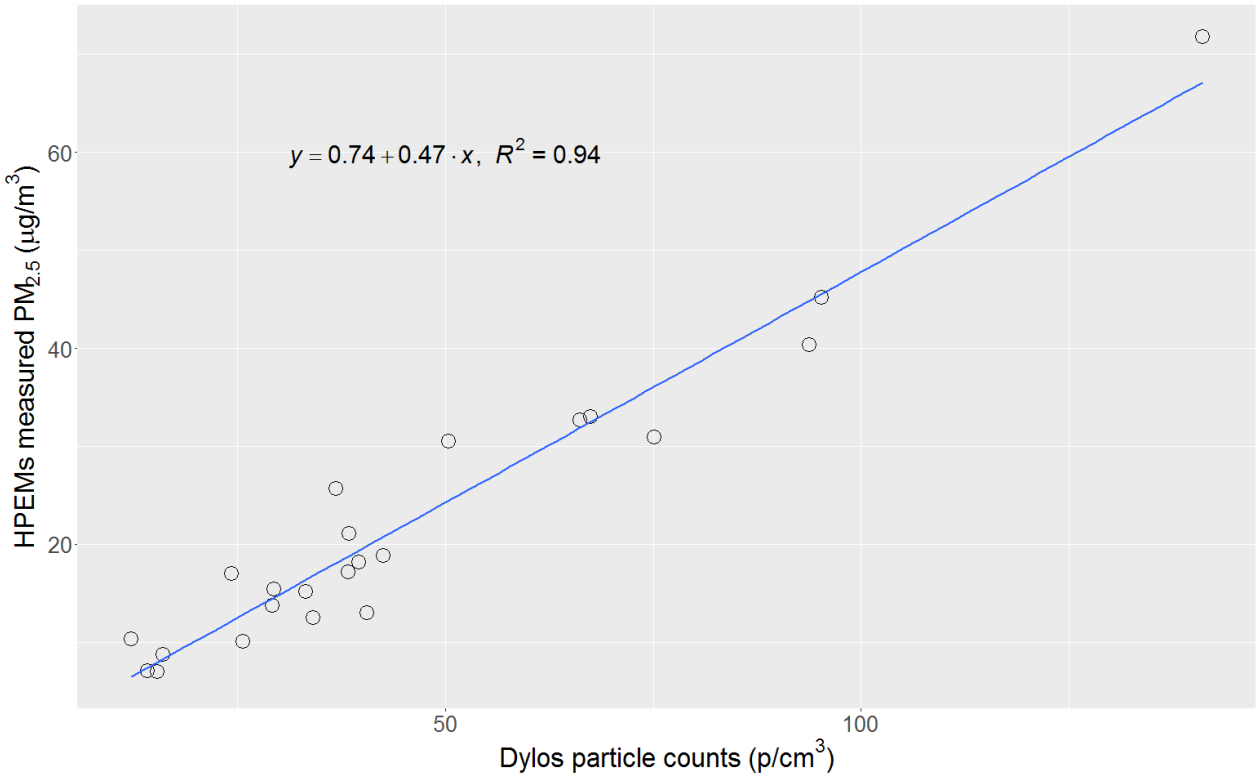

Figure 3: Relationship between 7-day average $PM_{2.5}$ (µg/m3) measured with Harvard personal environmental monitors (HPEMs) and particle counts (p/cm3) measured with Dylos

2.2 Predictors of indoor $PM_{2.5}$

2.2.1 Outdoor $PM_{2.5}$

Hourly outdoor $PM_{2.5}$ measurements from tapered element oscillating microbalances (TEOMs) between June 1, 2013 and December 31, 2015 were obtained from two government-run measurement stations. Outdoor $PM_{2.5}$ concentrations measured at centrally located government monitoring stations have been shown to capture city-wide temporal variations in $PM_{2.5}$ in several previous studies (Kioumourtzoglou et al., 2014). The two stations were selected because they were closest to UGAAR participants' home locations and had the most complete $PM_{2.5}$ data over the period of interest. Hourly data from both stations were aggregated to obtain 7-day average outdoor $PM_{2.5}$ concentration corresponding to the 7-day Dylos measurements for each participant at both home visits. In addition, I also calculated the distance between each UGAAR home and both of the stations, then assigned outdoor $PM_{2.5}$ at the nearest station to each residence. Thus, a total of three outdoor $PM_{2.5}$ variables were considered as possible predictors of indoor $PM_{2.5}$ concentrations (Table 3).

2.2.2 Meteorological data

Because previous studies have shown that meteorological factors can be correlated with outdoor $PM_{2.5}$ concentrations and the infiltration of $PM_{2.5}$ into residences (Allen et al., 2012; Chithra & Nagendra, 2014; Wang et al., 2015), temperature and wind speed measured at the two government-run pollution monitoring sites were also included as potential predictors of indoor $PM_{2.5}$. While temperature data were available from both stations A and B, I only used data from station B because the data from both stations were highly correlated (Pearson's correlation, r = 0.98) and there were some data gaps at station A. Previous literature has demonstrated that increased concentrations of outdoor air pollutants occur during periods of low wind speeds (Norris et al., 2000). Therefore, I followed the same approach used by Norris et al (2000) to create a stagnation index, which was defined as the number of hours in the 7-day Dylos measurement period in which the wind speed was less than the 50th percentile of the hourly wind speed over the period of January 1, 2014 to December 31, 2015. Additionally, I calculated the number of hours in each 7-day Dylos monitoring period that both stations were below their station-specific median, to potentially better indicate periods of stagnation across the entire city.

The Dylos measurements converted to $PM_{2.5}$ indicated substantial seasonal variation in indoor $PM_{2.5}$ concentrations in UGAAR homes, with higher concentrations during the winter season when emissions from ger stoves are highest (Figure 4). Therefore, season was also considered as a potential predictor of indoor $PM_{2.5}$ concentrations (winter = December, January and February; spring = March, April and May; summer = June, July and August; and fall = September, October and November) (Table 3).

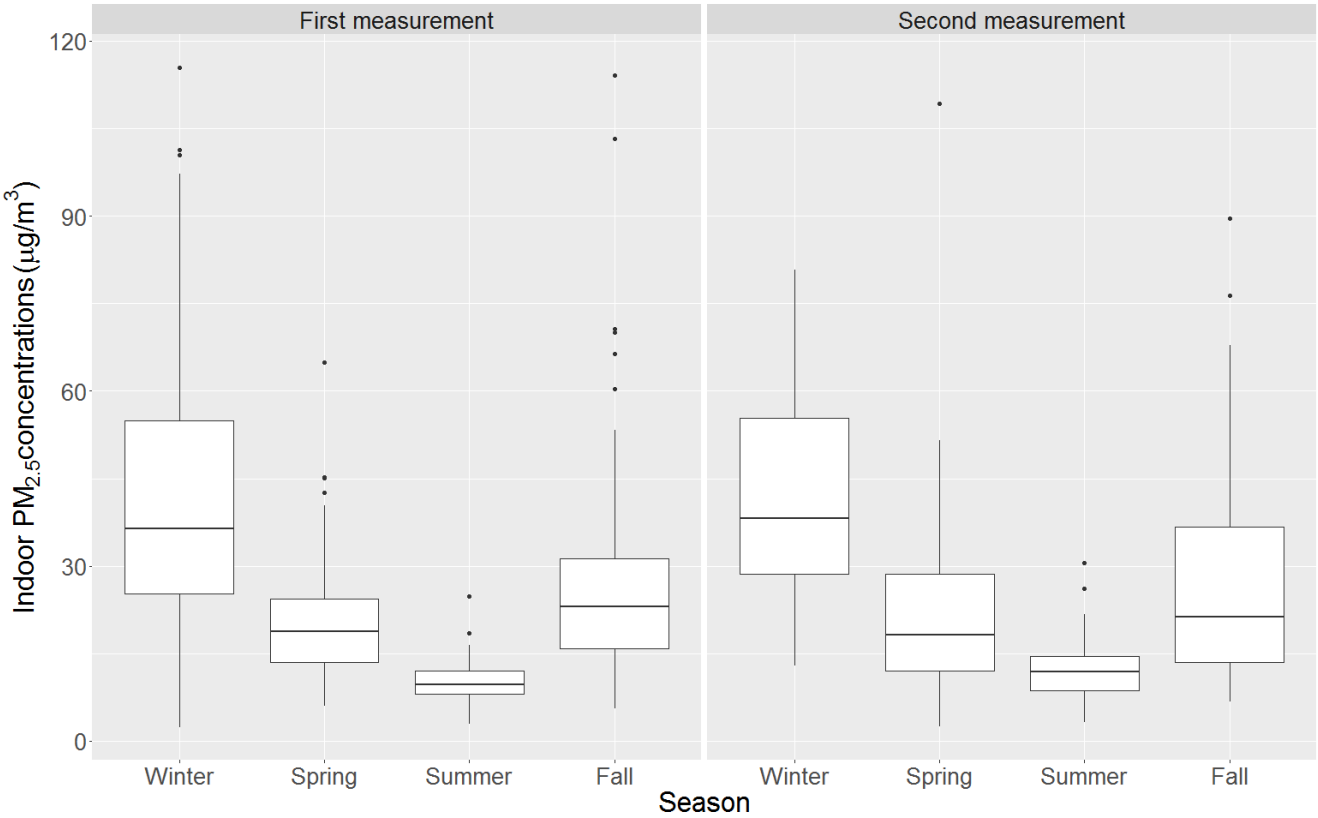

Figure 4: Seven-day indoor $PM_{2.5}$ concentrations (converted from Dylos measurements) by season and measurement. The first measurement was made shortly after enrollment at approximately 10 weeks gestation, on average. The second measurement was made at approximately 32 weeks gestation. Boxes represent the 25th and 75th percentiles. Lines inside boxes represent the median. Whiskers represent the 5th and 95th percentiles

2.2.3 Intervention

Three participants in the control group were mistakenly given air cleaners and two participants in the intervention group were not given air cleaners. Because the primary goal of this research was to build a prediction model, I modeled indoor $PM_{2.5}$ based on what the participants actually received. In total, 242 participants received air cleaners and 225 participants did not. In addition to intervention status, a new variable was created based on the number of air cleaners deployed for each participant: 74 and 168 participants in the intervention group received one and two air cleaners, respectively. The density of air cleaners (the number of air cleaners per square metre of home area) was also calculated for participants in the intervention group (Table 3). These two air cleaner variables aimed to capture and provide more air cleaner information than binary intervention status alone.

2.2.4 Home assessment data

The technicians measured the dimensions (width, height, length) of and counted the number of windows in every room in each home. Total home area and home volume were calculated based on the dimensions.

2.2.5 Questionnaire data

Questionnaires were administered to each participant at approximately the same time as the two home visits to collect information on housing characteristics, indoor pollution sources, and resident behaviours. Some participants moved during the study period, such that their residence was different for the 10-week measurements than for the 32-week measurements. Regardless of whether participants moved, values of the time-varying predictors (e.g. number of meals cooked inside the home per week) were matched to the Dyloes measurement made closest to the time that the questionnaire was administered. In addition, when either questionnaire had missing values for variables that were not expected to vary over time, the non-missing value was matched to both Dylos measurements. For temporally fixed variables in the situation where a participant changed address, I used the responses that corresponded to the home where the pollution measurements were made. For example, a participant who moved at the 25th week of pregnancy reported living on the 3rd floor on the first questionnaire and on the 7th floor on the second questionnaire. In this case, the first Dylos measurement was matched to living on the 3rd floor, and the second Dylos measurement was matched to living on the 7th floor.

In addition to outdoor $PM_{2.5}$ concentrations, meteorology, and intervention-related variables, the questionnaires provided another 25 predictor variables for consideration. Of the 25 potential predictors, 10 were removed from further analysis. Five were removed because of large number of missing values (>75%): frequency of air conditioning use in summer; possession of air conditioning; double/"eagle" door; oven type; and, stove or range type. Four were removed because all participants provided the same responses: types of tobacco smoked; primary source of heat in home; secondary heat source in home; and type of air cleaner. The percentage of HEPA air cleaners running time was also removed because it focused on overall use, not use specifically during the Dylos measurement periods, and had large number of missing values (77%).

Thus, a total of 15 potential predictor variables from questionnaires (Table 3) were considered in the model building process. These variables are listed below in three categories: housing characteristics, indoor combustion, and the use of air cleaners not provided as the intervention in

UGAAR (i.e., some participants in the control group may have chosen to operate their own air cleaners) (Table 3).

i. *Housing Characteristics* variables include: (1) age of home; (2) floor level; (3) typical window opening behavior in July; (4) typical window opening behavior in January; (5) window types.
ii. *Indoor Combustion* variables include: (1) number of people living in the home; (2) living with smokers; (3) number of smokers living in the home; (4) smoking inside home; (5) number of meals cooked inside the home per week; (6) number of meals cooked by frying inside the home per week; (7) presence of exhaust fan in the kitchen area; (8) frequency of candle or incense burning in the previous month.
iii. *Air Cleaner* variables include: (1) possession of air cleaner or not provided as part of the UGAAR study; (2) frequency of air cleaner use in the past month.

Questionnaire data were recorded on paper forms by trained interviewers. Research assistants entered these data into a Microsoft Access database. Some data were missing due to lack of participant knowledge, refusal to answer questions, or errors on the part of the interviewers. I double-checked blank fields in the database against the original paper forms and filled in missing data where possible.

2.2.6 Geographic data

A total of 58 potential spatial predictors were considered in the modelling (Table 3). We recently generated spatial data on ger locations, road locations, and land cover variables (brightness, greenness, wetness) using satellite images, high resolution aerial photographs and Google Maps images (Yuchi et al., 2016). These data were used as proxies for air pollution emissions in circular buffers surrounding the residence of each participant. The radii of these buffers ranged from 100 to 5000m to account for different dispersion patterns from different sources in the city. Brightness measures the overall surface reflectance. Low brightness values indicate dark surfaces, including water, dark soils, and objects in shadows. High brightness values indicate bright surfaces, including bright soil, impervious surfaces, and most human-made materials. Greenness measures the presence and density of green vegetation and wetness is primarily associated with soil moisture content (Yuchi et al., 2016). In addition, we used elevation from a digital elevation model (Environmental Systems Research Institute, 2012) as a potential predictor. Finally, wintertime NO_2 (a marker of traffic emissions) and SO_2 (a marker of coal

combustion) concentrations from land use regression models developed using measurements from 2010 were taken into consideration. The modeled SO_2 and NO_2 concentrations, as well as the brightness, greenness, wetness, road length and ger density variables, can provide information on spatial variation in outdoor pollution emissions and concentrations. Unless participants moved, I used latitude and longitude measured by field staff at the first home visit to extract values of the geographic data at the home location for each participant.

2.3 Final dataset

The final dataset had a total of 445 observations from 323 homes, one dependent variable (indoor 7-day $PM_{2.5}$ concentrations from Dylos measurements) and 87 potential predictor variables (3 for outdoor $PM_{2.5}$, 3 for intervention status, 3 from housing assessment data, 15 from questionnaire data, 5 from meteorological data, and 58 from geographic data) available for model building (Table 3).

Table 3: Summary of dependent variable and potential predictors used for model building procedures

Data	Categories	Predictors names	Class/Levels	Unit
Indoor $PM_{2.5}$ measurements	$PM_{2.5}$	7-day average indoor $PM_{2.5}$ concentrations	continuous	μg/m^3
Outdoor $PM_{2.5}$ measurements	$PM_{2.5}$ station A $PM_{2.5}$ station B $PM_{2.5}$ nearest station	7-day average outdoor $PM_{2.5}$ concentrations (aggregated from hourly measurements)	continuous	μg/m^3
Meteorological variables	Temperature station B		continuous	℃
	Wind stagnation station A Wind stagnation station B Wind stagnation both station			hours
	Season		categorical, Winter/Spring/Summer/Fall	
Intervention	Intervention status		categorical, Control/Filter	
	Number of filter deployed		categorical, 0/1/2	
	Air cleaner density		continuous	number of air cleaner/m^3
Housing assessment data	Housing characteristics	numbers of windows	categorical, One/Two/Three/More than three	
		total area	continuous	m^2
		total volume	continuous	m^3
Questionnaire data	Housing characteristics	age of home	continuous	years
		floor level	categorical, Low (1-3)/Medium(4-10)/High (11+)	
		typical window opening behaviors in July	categorical, Never/A few days in the month/More than half of the days, but not every day/Every day or almost every day	
		typical window opening behaviors in January	categorical, Never/A few days in the month/More than half of the days, but not every day/Every day or almost every day	
		window types	categorical, Wooden/Vacuum/Both/Other	

Data	Categories	Predictors names	Class/Levels	Unit
	Indoor combustion	number of people living in home	continuous	
		living with smokers or not	binary, Yes/No	
		number of smokers living in home	continuous	
		smoking inside home or not	binary, Yes/No	
		number of meals cooked inside the home per week	continuous	meals per week
		number of meals cooked by frying inside the home per week	continuous	meals per week
		presence of exhaust fan in kitchen area	categorical, Yes/No/Don't know	
		use frequency of burned candles or incense in the previous month	categorical, Less than once per week/1-2 time per week/3-4 time per week/5-6 time per week/Every day	
	Air cleaner	use frequency of air cleaner or air cleaner in the past month	categorical, Never/A few days in the month/More than half of the days, but not every day/Every day or almost every day	
		possession of air cleaner or air cleaner	binary, Yes/No	
Geographic variables	Brightness	100, 200, 300, 400, 500, 750, 1000, 1500,2000,2500, 5000 buffer radii	continuous	
	Greenness			
	Wetness			
	Ger density			gers/hectare
	Road length			km in circular
	Elevation	From a digital elevation model		buffer m
	SO_2	From previous developed land use		$\mu g/m^3$
	NO_2	regression model in Ulaanbaatar, Mongolia		

Chapter 3.

Methods

Because the 7-day average indoor $PM_{2.5}$ measurements were heavily right-skewed, a natural log transformation was applied to the data to achieve an approximately normal distribution. Dummy variables were created for categorical variables with more than two levels. Some continuous variables were converted into categorical variables (e.g. floor level was converted into an ordinal variable with categories of 1-3, 4-10, and ≥11) in cases where a non-linear relationship between predictor and indoor $PM_{2.5}$ concentrations was hypothesized. I developed four models in total: a primary MLR model, two secondary MLR models with alternative variable selection procedures, and a RFR model.

3.1 Primary MLR model building procedure

I adapted a model building procedure that has been used previously in the development of land use regression models of outdoor pollution concentrations (Henderson, Beckerman, Jerrett, & Brauer, 2007; Kalkbrenner et al., 2010).

(a) First, I identified variables that I expected *a priori* might modify the relationship between pollution emissions (outdoors or indoors) and indoor $PM_{2.5}$ concentrations. These included intervention status, outdoor temperature, and typical window opening behavior.

(b) I regressed log-transformed indoor $PM_{2.5}$ on each of the 87 predictor variables without any stratification (overall) and with stratification based on the variables identified in step (a). (control/air cleaner, temperature ≤ or > 15 ℃ ≤ or > 12 ℃, typically opening windows ”often” in July, typically opening windows “often” in January). The temperature cut offs were assumed to correspond with changes in window opening behaviors and the use of home heating.

(c) P-values were then extracted for each variable following step (b). Variables that were not significantly associated with log-transformed indoor $PM_{2.5}$ (p-value > 0.05) overall or in any of the strata were removed from further analysis.

(d) The remaining predictors were classified into 15 sub-categories: 1) Intervention, 2) Season, 3) Window opening behaviors, 4) Air cleaner use, 5) Indoor combustion, 6) Outdoor $PM_{2.5}$,

7) Land use model predictions, 8) Temperature, 9) Wind stagnation, 10) Brightness, 11) Greenness, 12) Wetness, 13) Ger density, 14) Road length, and 15) Elevation.

(e) In categories with more than one predictor, the predictor with the highest coefficient of determination (R^2) with log-transformed indoor $PM_{2.5}$ was identified.

(f) Predictors that were highly correlated (r >0.6 between continuous predictors) or associated (p-value <0.05 from Chi-Square test between categorical predictors) with the highest-ranking predictor in the category were removed from further analysis.

(g) The remaining predictors were ranked based on the non-stratified (overall) coefficient of determination (R^2, highest to lowest). Each of the predictors was then entered in a regression model in order. Only predictors that were significant ($p<0.05$), had a partial R^2 greater than 1%, and had a coefficient consistent with *a priori* assumptions (e.g. positive coefficient for outdoor $PM_{2.5}$), were retained.

The UGAAR study was conducted based largely on the assumption that indoor air cleaners can reduce the impact of outdoor $PM_{2.5}$ on indoor $PM_{2.5}$ concentrations. Therefore, an interaction term between outdoor $PM_{2.5}$ and the number of air cleaners deployed was also considered following step (g). Three other possible interaction terms (temperature x typical window opening behavior in January, season x intervention, outdoor $PM_{2.5}$ x typical window opening behavior in January) were also considered (see Appendix). The interaction term was included in the primary model if it was significant ($p<0.05$), did not make other predictors (step g) insignificant, and improved model performance. All models were evaluated using five indicators: (1) Mean Absolute Error (MAE), (2) Root Mean Square Error (RMSE), (3) Normalized Root Mean Square Error (NRMSE), (4) Mean Squared Error (MSE), and (5) Index of Agreement (IOA).

There were some repeated measurements from individual homes in the dataset. Possible dependence between measurements made in the same home was assessed by re-running a mixed model with random home intercept after a primary MLR model was generated, to ensure all the retained predictors behaved in a similar way.

3.2 Regression diagnostics

Multiple linear regression involves the following assumptions: (1) linearity, (2) homoscedasticity, (3) multivariate normality of model errors, and (4) no or little multicollinearity. Each of these assumptions was verified to ensure the proper use of multiple linear regression and robustness of model results.

Linearity means that the conditional means of the dependent variable are a linear function of the predictor variables. Homoscedasticity refers to the assumption of constant variance of residuals. A spread-level (residual-fit spread) plot was used to check both assumptions. The fitted values were plotted on the horizontal axis and absolute studentized residuals were plotted on the vertical axis. If the plot shows a random pattern, it suggests that the assumption of linearity is not violated (Berry, 1993; Tabachnick, Fidell, & Osterlind, 2001). A linear fit line and a loess smooth fit line were generated to check for any non-linearity. Relatively constant width of residuals around fitted values indicates constant variation of residuals, indicating that the assumption of homoscedasticity is not violated (Berry, 1993; Osborne & Waters, 2002).

Multivariate normality (normally distributed model errors) was evaluated using histograms and quantile-quantile (Q-Q) plots. The Q-Q plots were created by plotting two pairs of quantiles against one another. If points on the plot form a line that is approximately straight, it means the normality assumption is not violated.

Multicollinearity occurs when the predictor variables in a regression model are not independent of each other. Variance inflation factor (VIF) was used to evaluate multicollinearity in this study. VIF measures the extent to which the variance of the estimated regression coefficients is “inflated” due to collinearity between predictors (Alin, 2010; Mansfield & Helms, 1982; Tabachnick et al., 2001).

3.3 Sensitivity analysis and secondary MLR model building procedures

The influence of each observation was evaluated using Cook's distance to ensure the MLR model results were not highly influenced by a few extreme observations. The conservative cut-off for influential observations is a Cook’s distance larger than 4/(n-k-1), where n is the sample size and the k is the number of independent variables (Hair, Black, Babin, & Anderson, 2010). To evaluate the role of influential observations in the primary model, influential observations were removed and the model building procedures were repeated to allow for a comparison of models with and without influential observations.

Because the aim was to build a predictive model, I also developed a model using all 22 predictor variables that remained after model building step f, aiming to maximize R^2. In addition, stepwise (backward direction) regression was also used to fit the data and build a model for comparison with my primary MLR model.

3.4 RFR modelling procedure

I used a RFR modeling procedure modified from previous work by Smith et al (2013). Original input data consisted of the variables following step (g) in the primary MLR modelling procedure. I selected 500 as the number of trees to grow using bootstrap samples from the original input data. An unpruned regression tree grew for each of the bootstrap samples. At each node, a number of predictors was randomly sampled and the best split from those variables was selected. Next, new data were predicted by aggregating the predictions. The RFR model was assessed using the same five indicators: (1) MAE), (2) RMSE, (3) NRMSE, (4) MSE, and (5) IOA. R version 3.3.2 and the package "randomForest" were used to conduct the RFR analysis. The detailed R code is provided in the Appendix.

3.5 Model performance evaluation

3.5.1 10-fold cross validation

A 10-fold cross-validation (CV) method was applied (Allen et al., 2012; Trevor, Robert, & Jerome, 2001) to evaluate the prediction errors for both the MLR and RFR models:

(1) The dataset was randomly divided into 10 sub-groups with approximately the same number of observations in each group. Observations from the same home were forced into the same group.

(2) The predictive model was parameterized based on data from 9 of the 10 groups.

(3) The estimated coefficients were used to predict indoor $PM_{2.5}$ concentrations for observations in the excluded group.

(4) Steps (1) – (3) were repeated to obtain predictions for all 10 groups and, therefore, all observations.

(5) 7-day $PM_{2.5}$ concentration predictions and measurements on the untransformed scale were compared and model performance was evaluated based on R^2 and five additional indicators (described in section 3.5.2).

3.5.2 Model performance indicators

The analysis of model prediction performance generally involves the calculation of errors between predicted and observed values. Numerous performance indicators have been used in previous studies. The five indicators used to evaluate MLR and RFR model performance in this

analysis were: (1) Mean Absolute Error (MAE), (2) Root Mean Square Error (RMSE), (3) Normalized Root Mean Square Error (NRMSE), (4) Mean Squared Error (MSE), and (5) Index of Agreement (IOA).

The MAE measures the average magnitude of the errors between predictions and measurements (Equation 1). Values range from zero to positive infinity, with values closer to zero indicating better model performance (Chai & Draxler, 2014).

$$MAE = \frac{\sum_{i=1}^{N} | X_{obs,i} - X_{pre,i} |}{N} \quad (1)$$

Where N is sample size, $X_{obs,i}$ is the observed indoor $PM_{2.5}$, $X_{pre,i}$ is the predicted indoor $PM_{2.5}$ from MLR or RFR.

The RMSE aggregates the difference between predicted and observed values into a single measure of predictive power. RMSE values rang from zero to positive infinity, where lower values indicate better fit (Moriasi et al., 2007).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (X_{obs,i} - X_{pre,i})^2}{N}} \quad (2)$$

where N is sample size, $X_{obs,i}$ is the observed indoor $PM_{2.5}$, $X_{pre,i}$ is the predicted indoor $PM_{2.5}$ from MLR or RFR.

The NRMSE is a non-dimensional form of the RMSE, which measures the residuals between predicted and observed values. NRMSE values ranged from zero to positive infinity, where zero is a perfect score indicating no error (Nethery, Leckie, Teschke, & Brauer, 2008).

$$NRMSE = \frac{RMSE}{X_{obs,max} - X_{obs,min}} \quad (3)$$

where $X_{obs,max}$ is the maximum observed indoor $PM_{2.5}$, $X_{obs,min}$ is the minimum observed indoor $PM_{2.5}$.

The MSE is a measure of closeness of a fitted line to actual data points. MSE values are bounded between zero to infinity. The smaller the MSE, the closer the fit is to the data. A MSE of zero indicates perfect accuracy (Shyur, 2003).

$$MSE = \frac{\sum_{i=1}^{N} (X_{obs,i} - X_{pre,i})^2}{N} \quad (4)$$

where N is sample size, $X_{obs,i}$ is the observed indoor $PM_{2.5}$, $X_{pre,i}$ is the predicted indoor $PM_{2.5}$ from MLR or RFR.

The IOA is a standardized measure of the degree of model prediction error. Values of IOA range from zero to one. IOA values closer to one indicate a better match between predicted and observed data, while values closer to zero indicate poor agreement (Willmott, 1981).

$$IOA = 1 - \left[\frac{\sum_{i=1} N(M_{pre} - X_{obs,i})^2}{\sum_{i=1} N(|X_{pre,i} - M_{obs}| + |X_{obs,i} - M_{obs}|^2}\right] \quad (5)$$

where N is sample size, M_{pre} is the mean of predicted indoor $PM_{2.5}$, M_{obs} is the mean of observed indoor $PM_{2.5}$, $X_{obs,i}$ is the observed indoor $PM_{2.5}$, $X_{pre,i}$ is the predicted indoor $PM_{2.5}$ from MLR or RFR.

Chapter 4.

Results

Of the 87 predictor variables, 55 variables were significantly correlated with indoor $PM_{2.5}$ in one or more strata of control/air cleaner, temperature (≤ or > 15 ℃ ≤ or > 12 ℃), typically opening windows ”often” in July, typically opening windows “often” in January. Of these, 38 variables were significantly associated across the full unstratified dataset. A total of 22 variables were retained after removing variables that were highly correlated (r >0.6 between continuous variables) or associated (p-value <0.05 from Chi-Square test between categorical predictors) with the highest-ranking predictor in the categories with more than one variable. The R^2 of the 22 variables ranged from 0.2% to 38%, and 16 variables had R^2 greater than 1.0%. The three strongest individual predictors of indoor $PM_{2.5}$ were season (R^2 = 38.0%), outdoor $PM_{2.5}$ (R^2 = 18.0%), and outdoor temperature (R^2 = 17.0%). The number of air cleaners deployed had an R^2 of 6%. Several predictors were highly correlated. For example, outdoor temperature was highly correlated with outdoor $PM_{2.5}$ (Pearson’s r = - 0.8).

4.1 Descriptive statistics

Summary statistics for the 22 variables considered for building the MLR models are summarized in Table 4 and Table 5. In brief, 7-day outdoor $PM_{2.5}$ had median of 36.6 µg/m^3 with some extreme values of over 100 µg/m^3 and an interquartile range of 43.0 µg/m^3. Temporal 7-day outdoor $PM_{2.5}$ patterns differed between stations A and B, with consistently higher concentrations measured at station B (Figure 5). The correlation between 7-day $PM_{2.5}$ measurements at stations A and B was 0.93 (Figure 6). Dylos measurements were approximately equally distributed across four seasons (Table 5). The majority of participants (169 out of 242) in the intervention group received two air cleaners. The density of air cleaners in participants who received one air cleaner (median = 0.007 air cleaners/m^3) was slightly lower than those who received two air cleaners (median = 0.009 air cleaners/m^3).

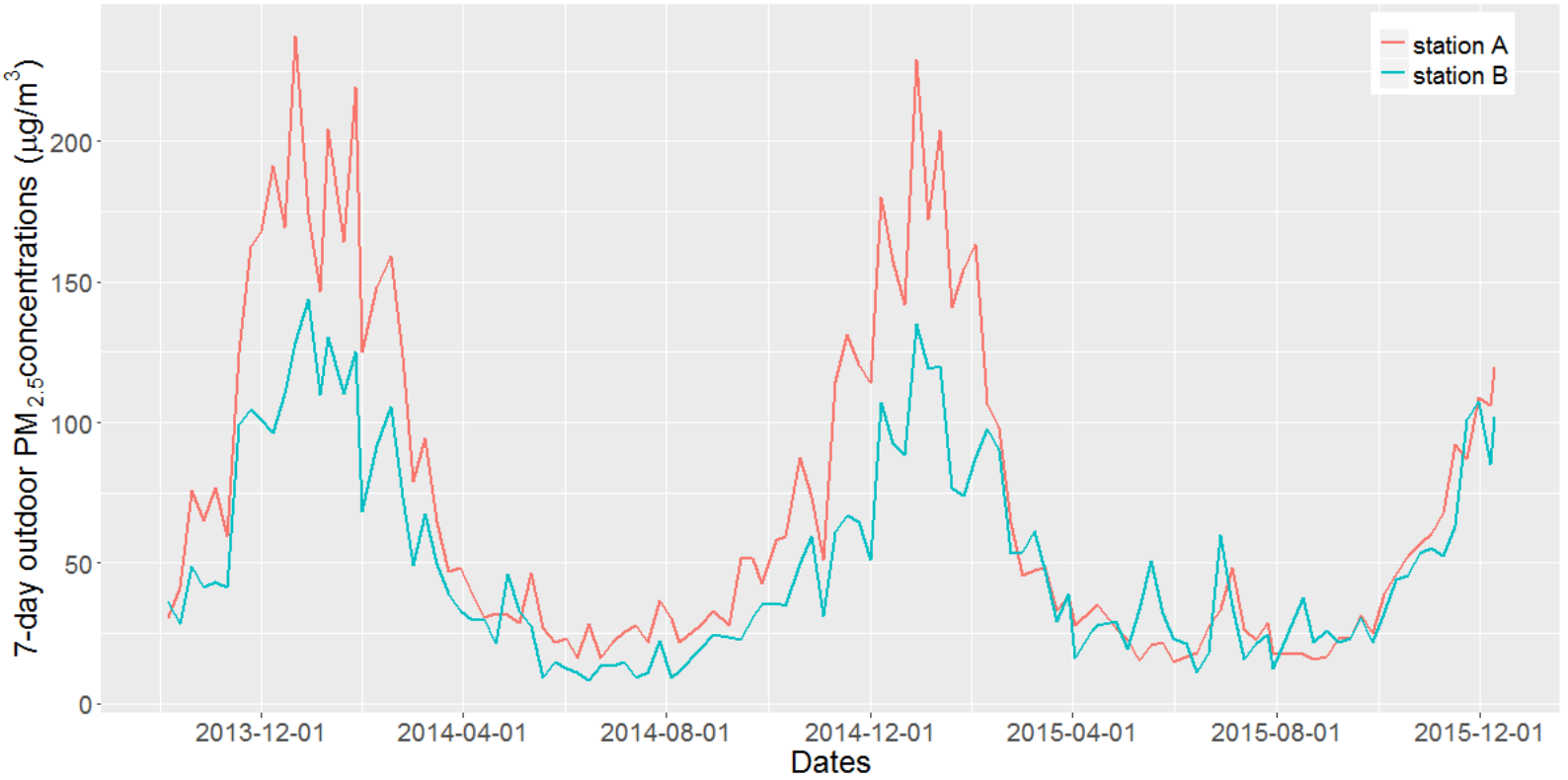

Figure 5: Temporal patterns of 7-day outdoor $PM_{2.5}$ from stations A and B during the period of UGAAR pregnancies

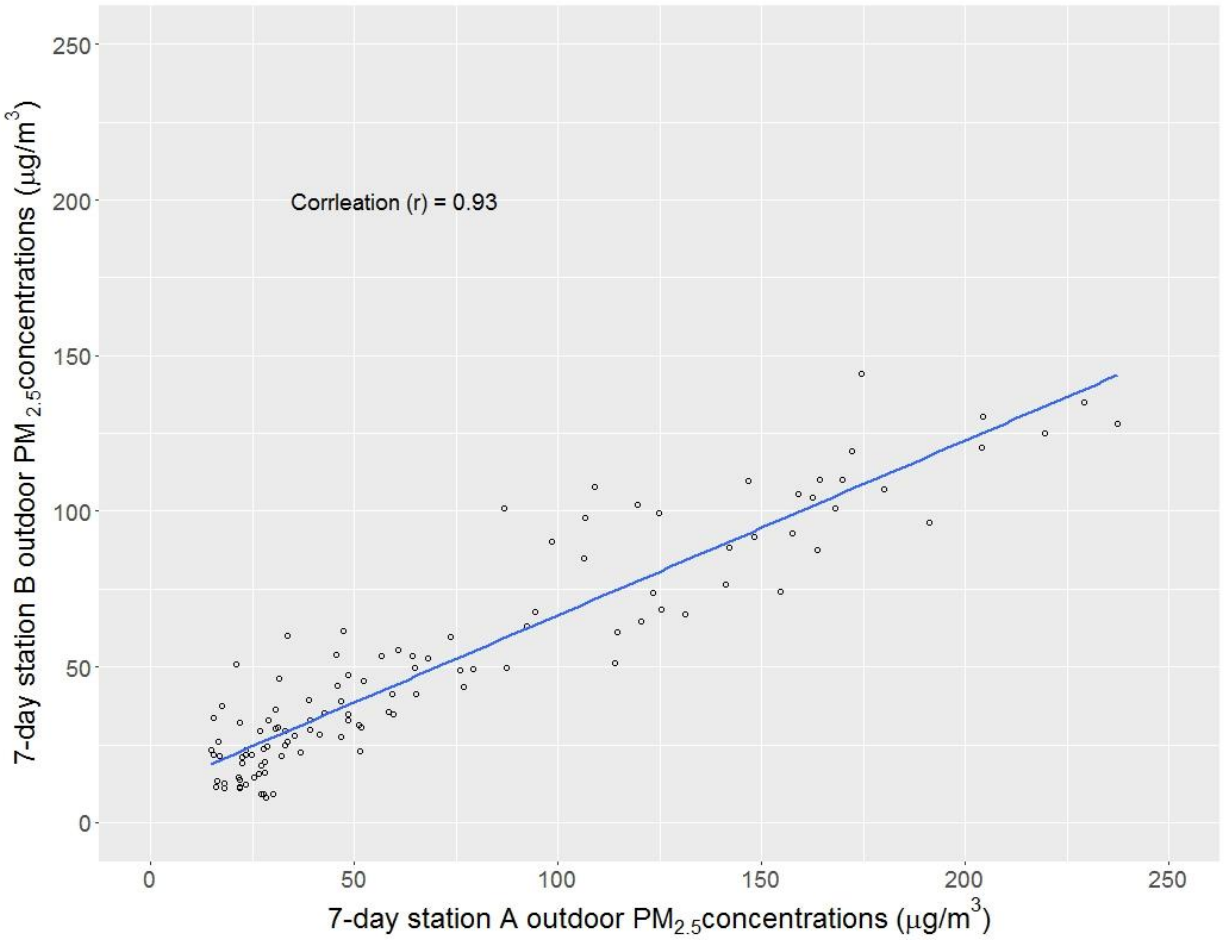

Figure 6: Scatterplot showing the relationship between 7-day outdoor $PM_{2.5}$ concentrations measured at stations A and B during the period of UGAAR pregnancies

Table 4: Descriptive statistics for indoor $PM_{2.5}$ and 17 continuous predictor variables considered in the MLR model building procedure

Continous variables	Unit	Minimum	25th percentile	Median	75th percentile	Maximum	Missing values
7-day average indoor $PM_{2.5}$	µg/m^3	2.3	12.1	19.2	33.4	115.4	0
Log-transformed 7-day indoor $PM_{2.5}$		0.8	2.5	2.9	3.5	4.7	0
7-day average outdoor $PM_{2.5}$ Station B*	µg/m^3	1.0	25.6	36.6	68.6	146.5	7
Air cleaner density (homes received air cleaners only)	Air cleaners/ m^3	0.003	0.005	0.009	0.017	0.081	0
7-day average outdoor temperature Station B	°C	-21.1	-12.4	-0.2	14.0	28.7	21
Wind stagnation both stations	hours	0.0	1.0	4.0	7.0	28.0	0
Ger density in 750m radius	gers/hectare	0.0	0.0	0.4	2.5	12.5	0
Ger density in 5000m radius*	gers/hectare	0.2	3.3	4.0	4.8	5.3	0
Brightness in 5000m radius	unitless	137.1	156.6	159.6	162.7	165.8	0
Greenness in 5000m radius	unitless	-20.0	-19.6	-19.2	-18.7	-7.5	0
Wetness in 5000m radius	unitless	-21.5	-14.0	-12.7	-11.1	-6.0	0
Road length in 100m radius	km/100 metre buffer	0.0	0.1	0.2	0.4	0.8	0
Raod length in 500 radius	km/500 metre buffer	0.0	7.2	8.1	9.7	18.0	0
Road length in 2500m radius	km/2500 metre buffer	17.5	196.7	231.2	272.3	301.8	0
Elevation	m	1279	1292	1301	1308	1388	0
SO_2	µg/m^3	0.0	12.2	18.0	26.4	37.2	0
Number of meals cooked inside the home per week	unitless	0.0	7.0	14.0	14.0	22.0	8

Continous variables	Unit	Minimum	25th percentile	Median	75th percentile	Maximum	Missing values
Number of meals cooked by frying inside the home per week		0.0	3.0	7.0	7.0	15.0	8
Number of smokers living in home		0.0	0.0	0.0	1.0	2.0	8

*Variable retained in the primary MLR model

Table 5: Descriptive statistics for the 5 categorical predictor variables considered in the MLR model building procedure

Categorical variables	Number of levels	Levels	Observations per level	Missing values
Season*	4	Winter	114	
		Spring	116	
		Summer	104	
		Fall	111	
Number of air cleaners deployed*	3	Zero	215	0
		One air cleaner	61	
		Two air cleaners	169	
Intervention status	2	Control	215	
		Air cleaner(s)	230	
Typical window opening behaviors in January	2	Not often	159	98
		Often	188	
Possession of air cleaner**	2	No	431	0
		Yes	14	

*Variable retained in the primary MLR model

**Air cleaners used by the participants in addition to the ones they received from the UGAAR study

4.2 MLR models

After following the model building procedures described in section 3.1, the primary MLR model included the following variables as predictors: season; outdoor $PM_{2.5}$; number of air cleaners deployed; ger density; and an interaction between outdoor $PM_{2.5}$ and the number of air cleaners deployed (Table 6). Season was the most important predictor of indoor $PM_{2.5}$ concentrations (partial R^2 = 31.0%).

This primary MLR model explained 52.5% of the variation in log-transformed indoor $PM_{2.5}$ with a cross-validation (CV) R^2 of 50.5% (Figure 7). After removing 23 highly influential observations with Cook's distance > 0.009, a similar model was produced with the same four predictors and interaction term (outdoor $PM_{2.5}$ x number of air cleaners deployed) and had similar coefficients. This model had an R^2 of 59.6% and a CV R^2 of 58.2%. This suggests that the primary MLR model is not highly sensitive to influential observations and the model results are relatively robust.

Regression diagnostics indicated that the assumptions of MLR were met. The Spread-Level plot showed a random pattern and a linear fit line across the full range on the x-axis, indicating that the homoscedasticity and linearity assumptions were also met. The normal Q-Q plot showed that the model residuals were approximately normally distributed. The multicollinearity assumption was also fulfilled because the VIF for model predictors were all less than 2.5. When

home was modelled with a random intercept, the results showed similar coefficients and standard errors when compared with the primary MLR model. This suggested that measurements made in the same home can be considered as independent.

Table 6: Primary MLR Model. Significant p-values are in bold

Predictor variables	Coefficients	Standard Error	P-value	Partial R^2 (%)
Intercept	1.52	0.13	**<0.001**	
Summer (reference)	-	-	-	
Fall	0.58	0.06	**<0.001**	31
Spring	0.74	0.06	**<0.001**	
Winter	1.07	0.08	**<0.001**	
Outdoor $PM_{2.5}$ from station B ($\mu g/m^3$)	0.005	0.001	**<0.001**	10
Control (reference)	-	-	-	
1 air cleaner deployed	-0.09	0.12	0.46	8
2 air cleaners deployed	-0.14	0.09	0.11	
Ger density in 5000m radius (gers/hectare)	0.19	0.02	**<0.001**	3
Outdoor $PM_{2.5}$ from station B * 1 air cleaner deployed	-0.001	0.002	0.33	0.5
Outdoor $PM_{2.5}$ from station B * 2 air cleaners deployed	-0.005	0.001	**<0.001**	

"/": categorical level; "|": interaction term

A MLR model with all 22 considered variables had an R^2 of 56.2%. However, this model had poor performance (MAE = 0.66, NRMSE = 92.3, RMSE = 0.78, MSE = 0. 46, IOA = 0.67, Table 7) compared with the MLR model (MAE = 0.37, NRMSE = 69.7, RMSE = 0.49, MSE = 0.24, IOA = 0.80, Table 7). In addition, some estimated coefficients from this model violated *a priori* assumptions. For example, greenness in 5000m radius had a positive coefficient. Moreover, this model also had slightly weaker model performance in comparison with primary MLR by cross validation (CV R^2: 48.2% vs. 50.5%). A model based on the automated stepwise regression method had similar predictive performance (R^2 = 56.0%) as the model based on the 22 variables (R^2 = 56.2%), but used only 9 variables. This model also performed poorly (MAE = 0.57, NRMSE = 80.8, RMSE = 0.67, MSE = 0.42, IOA = 0.68, Table 7) compared with the primary MLR model, and also had coefficients that violated *a priori* assumptions. The stepwise model also had a slightly lower CV R^2 compared with the primary MLR model (CV R^2: 49.1% vs. 50.5%).

4.3 RFR model results and comparison with primary MLR model

Compared with the primary MLR model, the RFR model (using the same variables as in the primary MLR model) explained a larger fraction of the variation in log-transformed indoor $PM_{2.5}$ (R^2 = 74.4%). The RFR model also had better model performance, achieving a higher index of agreement (IOA =0.86) compared with the primary MLR model (IOA = 0.82) and had consistently lower scores on error-based model performance indicators (Table 7). However, based on the CV the RFR did not perform as well as the primary MLR model (Table 7, Figure 7). These results indicate that the RFR performed slightly better with the sample data than the primary MLR model, while the primary MLR model performed better in predicting indoor $PM_{2.5}$ outside of the sample. Overall, there was strong agreement between cross-validation predictions from the RFR and primary MLR models (R^2: 90.2%, Figure 8).

Table 7: Primary MLR model, MLR model with 22 predictor variables, MLR model with stepwise variable selection and RFR model performance evaluation

Models	Coeeficient of Determination (R^2, %)	Mean Absolue Error (MAE)	Normalized Root Mean Square Error (NRMSE)	Root Mean Square Error (RMSE)	Mean Sqaured Error (MSE)	Index of Agreement (IOA)
Primary MLR model	52.5	0.36	68.3	0.48	0.23	0.82
MLR model with 22 predictor variables	56.2	0.66	92.3	0.78	0.46	0.67
MLR model with stepwise variable selection	56.0	0.57	80.8	0.67	0.42	0.68
Random Forest Regression (RFR)	74.4	0.30	56.3	0.40	0.16	0.86
10-fold Cross validation (based on untranformed values)						
Primary MLR model	50.5	0.37	69.7	0.49	0.24	0.80
MLR model with 22 predictor variables	48.2	0.69	94.5	0.72	0.47	0.62
MLR model with stepwise variable selection	49.1	0.63	85.3	0.64	0.41	0.64
Random Forest Regression (RFR)	47.8	0.39	73.5	0.52	0.27	0.73

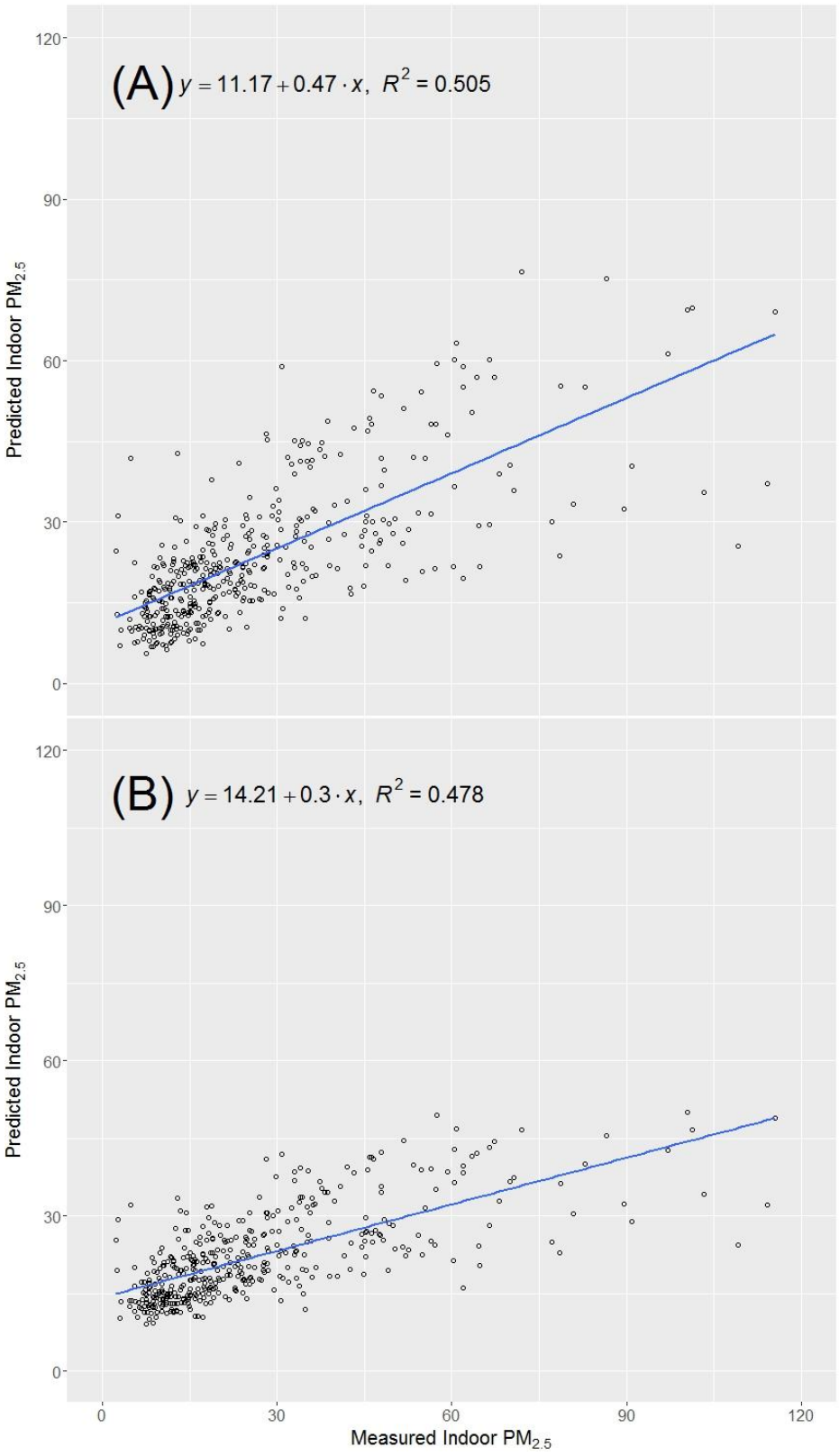

Figure 7: Relationship between predicted and measured indoor $PM_{2.5}$ based on 10-fold cross-validation (untransformed scale) for the (A) primary MLR model and (B) RFR model

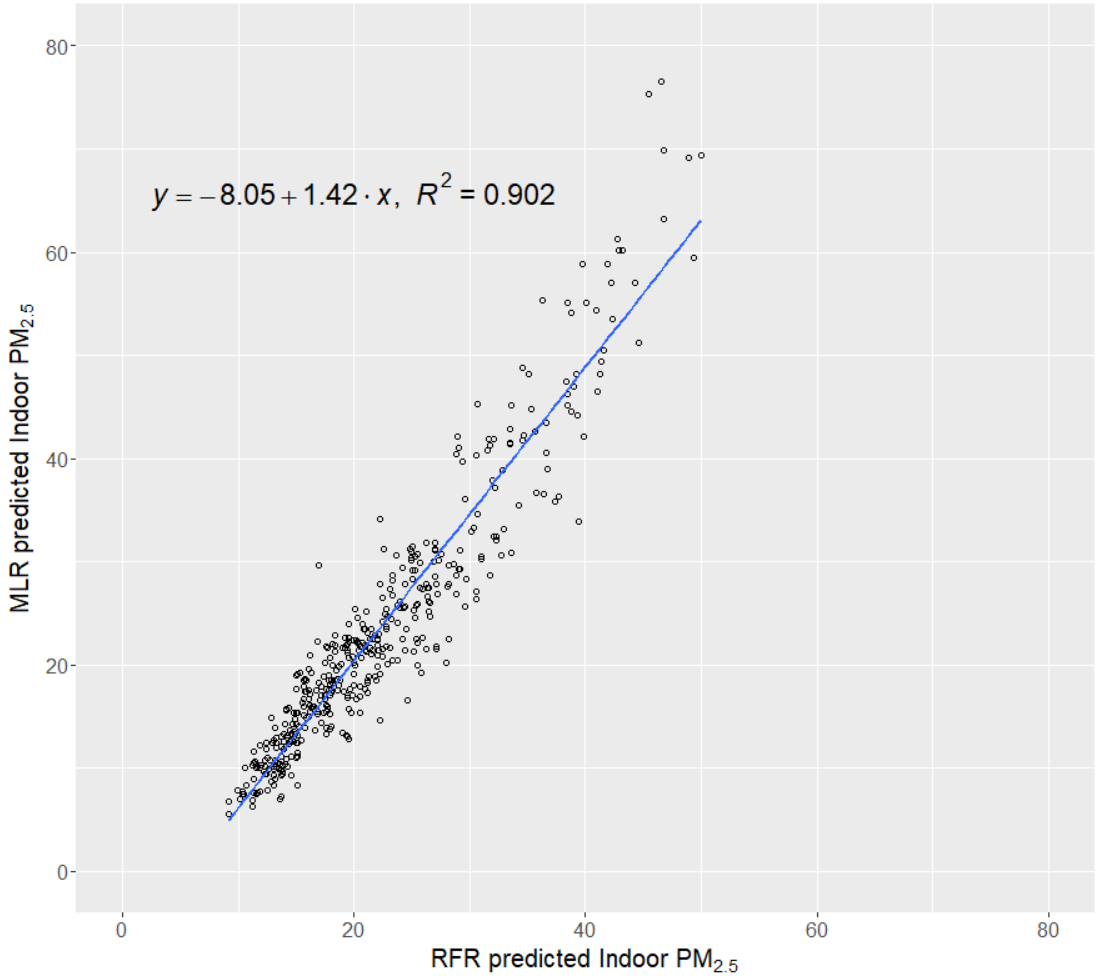

Figure 8: Comparison between indoor $PM_{2.5}$ predicted from the primary MLR and RFR models based on 10-fold cross-validation (untransformed scale)

4.4 Primary MLR model application

The primary MLR model was used to predict 7-day average indoor $PM_{2.5}$ concentrations for each UGAAR participant, and then average concentrations over the course of pregnancy and in each trimester were calculated (Figure 9). Outdoor $PM_{2.5}$ data were missing for 923 hours in the prediction period. As a result, 38 out of 18038 weekly predictions were missing. Thus, in the situation where outdoor $PM_{2.5}$ from station B was not available, I fitted a model using the same four variables using outdoor $PM_{2.5}$ from station A. For participants who changed their addresses, I changed ger density values at the week of their pregnancy corresponding to the time they moved. Because participants in the intervention group received their HEPA filter air cleaners shortly after they enrolled in the UGAAR study, I used the date of the first home visit to each home, when

HEPA air cleaners were deployed, to assign the number of air cleaners deployed. Prior to the first home visit all participants had zero UGAAR study air cleaners deployed at their homes.

Compared with the control group, the full pregnancy median indoor $PM_{2.5}$ concentration was 8.0% lower in homes that received one air cleaner, and 23.1% lower in homes that received two air cleaners (Table 8 and Figure 9). Participants in the control, 1 air cleaner, and 2 air cleaner groups had similar indoor $PM_{2.5}$ concentrations during the first trimester (median = 31.5 $\mu g/m^3$, 28.1 $\mu g/m^3$, 28.5 $\mu g/m^3$, respectively, Table 8 and Figure 9), because the air cleaners were not deployed until 10 weeks gestation, on average. The median indoor $PM_{2.5}$ concentrations for all three groups decreased over the course of pregnancy.

Table 8: Summary statistics of average 7-day prediction of indoor $PM_{2.5}$ concentrations ($\mu g/m^3$) for each trimester and whole pregnancy

Categories	Minimum	Median	Interquartile range	Maximum	Standard deviation
1st trimester	6.8	29.7	21.9	127.0	15.9
Participants who received 0 air cleaners (control)	6.8	31.5	20.9	71.9	15.2
Participants who received 1 air cleaner (intervention)	10.5	28.1	23.6	70.3	16.3
Participants who received 2 air cleaners (intervention)	7.5	28.5	22.9	127.0	16.6
2nd trimester	7.5	21.1	12.4	55.9	10.5
Participants who received 0 air cleaners (control)	9.6	25.7	19.5	55.9	11.7
Participants who received 1 air cleaner (intervention)	9.2	24.8	12.6	44.3	9.3
Participants who received 2 air cleaners (intervention)	7.5	17.7	7.5	27.9	5.0
3rd trimester	5.7	18.0	10.2	61.5	9.6
Participants who received 0 air cleaners (control)	8.3	20.9	12.6	61.5	11.3
Participants who received 1 air cleaner (intervention)	8.3	18.4	7.7	42.7	6.7
Participants who received 2 air cleaners (intervention)	5.7	15.6	8.3	29.7	5.7
Full pregancy	10.7	25.1	8.4	48.6	6.3
Participants who received 0 air cleaners (control)	14.4	27.2	7.4	42.3	5.9
Participants who received 1 air cleaner (intervention)	14.8	25.0	7.8	38.8	5.9
Participants who received 2 air cleaners (intervention)	10.7	20.9	6.3	48.6	5.1

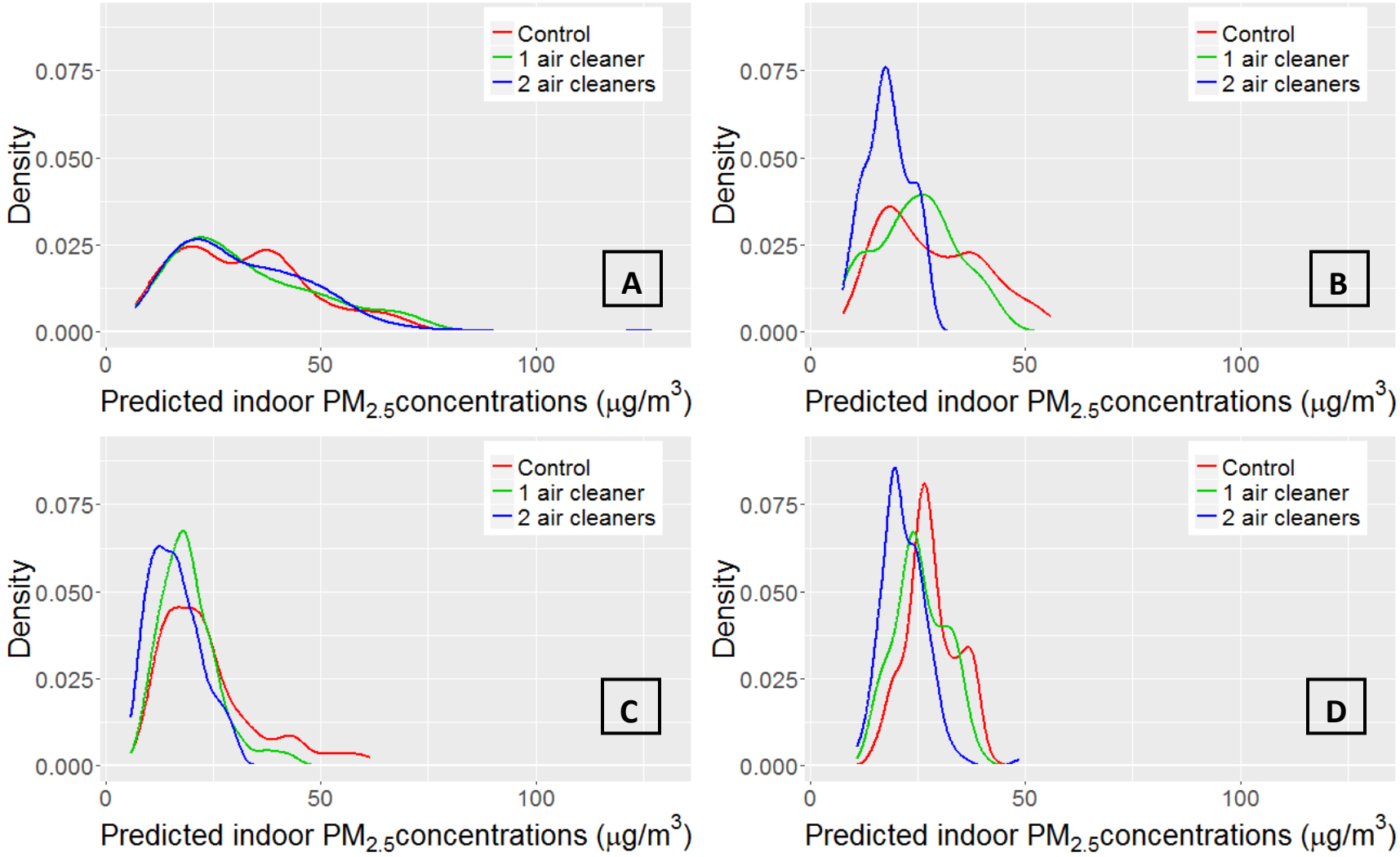

Figure 9: Distribution of predicted 7-day indoor $PM_{2.5}$ concentrations ($\mu g/m^3$)
A: 1st trimester, B: 2nd trimester, C: 3rd trimester, D: full pregnancy

Chapter 5.

Discussion

My study has demonstrated the ability to model a moderate portion of the variation in indoor $PM_{2.5}$ in the homes of pregnant women in Ulaanbaatar, Mongolia, using both MLR and RFR models. In general, the primary MLR model had slightly better out of sample performance in comparison with RFR model.

The primary MLR model described in this thesis had comparable predictive performance (CV R^2 = 50.5%) to previously published models, which explained 35% - 84% of the variation in indoor $PM_{2.5}$ concentrations (Table 9). Previous MLR models used to predict indoor $PM_{2.5}$ were mainly conducted in developed countries, such as France, Germany and the United States (Cyrys, Pitz, Bischof, Wichmann, & Heinrich, 2004; Gauvin et al., 2002; Lai et al., 2006; Meng, Spector, Colome, & Turpin, 2009). A limited number of studies were conducted in developing regions, such as Palestine and India (Elbayoumi, Ramli, & Yusof, 2015; Elbayoumi et al., 2014; Goyal & Khare, 2011). Most of the previous studies (Elbayoumi et al., 2015; Elbayoumi et al., 2014; Gauvin et al., 2002; Lai et al., 2006; Meng et al., 2009) used stepwise regression to screen potential predictor variables and build their MLR models, while the remaining studies only mentioned that they used MLR but did not specify which variable selection method was applied (Cyrys et al., 2004; Goyal & Khare, 2011). Outdoor $PM_{2.5}$ and meteorological data (e.g., temperature, wind speed etc.) were commonly considered as independent variables in most of the studies (Cyrys et al., 2004; Elbayoumi et al., 2015; Elbayoumi et al., 2014; Goyal & Khare, 2011; Lai et al., 2006; Meng et al., 2009). Only one study used outdoor PM_{10} instead (Gauvin et al., 2002). Other commonly used variables included traffic, heating, cooking, ventilation (window opening), other outdoor air pollutants (e.g. black carbon), while some studies also incorporated information on the interior of the home (e.g. curtains in homes, soft furnishing in homes) (Lai et al., 2006), and another study included data on number of pets (Gauvin et al., 2002). Smoking-related variables were also considered in several studies (Gauvin et al., 2002; Lai et al., 2006; Meng et al., 2009).

Two alternative MLR models - one with 22 considered predictors and one developed using a stepwise variable selection - had higher R^2 compared with the primary MLR model. However, the primary MLR model consistently had better performance across all model indicators than the

two alternative models. These model results demonstrate that the model building procedures adopted from the development of land use regression models of outdoor pollution concentrations have promise for identifying predictors of indoor $PM_{2.5}$ concentrations.

RFR has been used to model outdoor air pollution and its performance has been compared with linear regression models (Pandey, Zhang, & Jian, 2013; Ryan, Brokamp, Fan, & Rao, 2015). To my knowledge, this is the first study to use RFR to model indoor $PM_{2.5}$ concentrations. In the cross-validation the RFR model performed slightly worse than the primary MLR model. Nevertheless, RFR has several advantages over MLR, and RFR is a promising machine-learning approach that should be taken into consideration when modeling indoor pollution concentrations.

Overfitting can be a problem when developing predictive models, and the lower cross-validation R^2 for both the MLR and RFR models may have been due in part to overfitting. One approach to reduce overfitting in MLR models is to set a more stringent significance criterion (p-value) for variables to enter the model. For RFR models, one possible approach would be to test the use of different numbers of decisions trees and compare their results to find an optimal number to minimize overfitting.

Table 9: Summary of previous literature using MLR models to predict indoor $PM_{2.5}$

References	Countries(cities)	Home types(s)	Homes	Measurement duration	CV R^2 (%)	R^2 (%)	Predictors
(Meng et al., 2009)	The United States (Houston, Los Angeles County, Elizabeth)	Non-smoking homes (apartment, single house, town house, mobile house)	240	48 hours	34.9	-	Outdoor $PM_{2.5,}$ environmental tobacco smoke, sweeping, incense, cooking, sander, chainsaw
(Gauvin et al., 2002)	France (Grenoble, Paris, Toulouse, Clermont-Ferrand, Nice)	Classroom, home living-room	434	48 hours	-	36.0	Outdoor PM_{10}, index of traffic, number of smokers, number of persons per room, percentage of rodent, percentage of pets, percentage of gas cooking, percentage of gas heating, percentage of rainfall during 48 hours
(Goyal & Khare, 2011)	India (Delhi)	Classroom	16	8:00am – 2:00pm for 20 weekdays and 8 weekends in winter season, 8:00am – 2:00pm for 17 weekdays and 8 weekends in non-winter season	-	40.0	Outdoor $PM_{2.5}$, outdoor temperature, indoor temperature, relative humidity, building ventilation rate, wind speed
(Elbayoumi et al., 2014)	Palestine (Gaza strip)	Classrooms	36	7:00am – noon during school days between March to May	71	-	Outdoor $PM_{2.5}$, indoor carbon monoxide, outdoor carbon monoxide, ventilation rate, outdoor temperature, indoor temperature, relative humidity, wind speed

References	Countries(cities)	Home types(s)	Homes	Measurement duration	CV R^2 (%)	R^2 (%)	Predictors
(Lai et al., 2006)	Greece (Athens), Switzerland (Basel), Finland (Helsinki), Italy (Milan), The United Kingdom (Oxford), Czech Repubic (Prague)	Any types of homes(non-specify)	413	48 hours	60	-	Total over 40 predictors: weather, smoking, gas stove, heating, cooking, ventilation, traffic, wall, floor, products, interiors, others
(Cyrys et al., 2004)	Germany (Erfurt)	Hospital rooms	111	24 hours	-	84	Outdoor $PM_{2.5}$, ventilation, temperature, relative humidity

Four predictor variables (season, outdoor $PM_{2.5}$, number of air cleaners deployed, and ger density in a 5000m radius buffer) were retained in the primary MLR model. Although the goal of this research was to build a predictive model, and not a causal model for identifying determinants of indoor $PM_{2.5}$, the predictors in the primary MLR model are consistent with literature on sources of $PM_{2.5}$ and its variation in Ulaanbaatar (Allen et al., 2013; Davy et al., 2011; Fisk & Chan, 2017). Previous studies have documented the strong seasonal pattern in $PM_{2.5}$ concentrations in Ulaanbaatar (Allen et al., 2013; Davy et al., 2011). Gers are widely recognized as the most important $PM_{2.5}$ source in Ulaanbaatar (Davy et al., 2011; Ochir & Smith, 2014), and a land use regression models developed for Ulaanbaatar identified gers as a major source of spatial variation in outdoor pollution concentrations (Allen et al., 2013; Davy et al., 2011). Several studies have demonstrated that HEPA filter air cleaners can substantially reduce indoor $PM_{2.5}$ concentrations (Fisk & Chan, 2017). Except season, all of the predictor variables in my primary MLR model are modifiable. Decreasing emissions from ger stoves and increasing the use of portable air cleaners have already been identified as potential strategies to reduce air pollution health risks in Ulaanbaatar. For example, Ochir and Smith (2014) estimated that a shift from coal to cleaner burning home heating fuels could lead to a 60% percent reduction in annual health impacts from air pollution by 2024. A recent UNICEF report identified portable air cleaners as a tool to reduce the impacts of air pollution on children in Ulaanbaatar (United Nations Children's Fund, 2016).

The primary MLR model was used to predict trimester-specific and full pregnancy indoor $PM_{2.5}$ concentratons for all UGAAR participants, and these predictions will be used in future UGAAR epidemiologic studies. The variation explained by the primary MLR model was much higher than that explained by air cleaner deployment alone (R^2 of number of air cleaner deployed = 6%). This is consistent with results from the RESPIRE randomized trial of chimney stoves and pneumonia risk in Guatemala (Smith et al., 2011). McCracken et al (2009) applied mixed models to predict carbon monoxide concentrations in RESPIRE, and found that the predictive power of mixed models that combined individual- and group-level predictors was better than solely using individual (e.g. child mean) or group level (e.g. stove) variables alone (McCracken et al., 2009).

The median predicted indoor $PM_{2.5}$ concentrations for control, 1 air cleaner, and 2 air cleaner groups decreased over the course of pregnancy. While a decrease from the first trimester was expected for the intervention participants, the decrease among control participants was suprising. To evaluate the cause of this decrease I further evaluated the time varying model predictors (season and outdoor $PM_{2.5}$). Participants were more likely to begin their pregnancies during the high pollution seasons, with 64% of UGAAR control participants becoming pregnant in

the fall or winter. Outdoor $PM_{2.5}$ concentrations tended to be higher during participants' first trimesters. For each of the three groups, the 90th percentile of the outdoor $PM_{2.5}$ distribution in the first trimester was ≥ 119 μg/m^3, while for the second and third trimesters the 90th percentiles were all ≤ 101 μg/m^3 (Table 10).

Table 10: 90th percentile of outdoor $PM_{2.5}$ by trimester and intervention group

Categories	90th percentile of outdoor $PM_{2.5}$ (μg/m^3)			
	1st trimester	2nd trimester	3rd trimester	Full pregnancy
Control	119.9	96.6	90.3	95.6
1 air cleaner	126.5	101.0	85.7	100.3
2 air cleaners	119.2	96.5	90.5	97.7

Although the primary analyses in UGAAR will be intention-to-treat, secondary analyses will evaluate relationships between modeled indoor $PM_{2.5}$ and health outcomes. Limitations in statistical power caused by the relatively small size of the UGAAR cohort will be offset somewhat by the relatively large exposure gradients. The interquartile range (IQR) in modeled full-pregnancy indoor $PM_{2.5}$ was 8.4 μg/m^3 and trimester-specific IQRs ranged from 10.2 to 21.9 μg/m^3. Previous research of outdoor $PM_{2.5}$ and birth outcomes was conducted using much larger study populations, but reported much less exposure variation with full-pregnancy and trimester-specific IQRs typically below 3.5 μg/m^3 (Bell et al., 2010; Hyder et al., 2014; Lavigne et al., 2016; Savitz et al., 2014).

Several limitations should be taken into consideration when interpreting these results. First, the particle concentrations were measured with the Dylos, a low cost optical sensor that approximates $PM_{2.5}$ concentrations. Although $PM_{2.5}$ was not measured directly in most homes, the Dylos measurements were highly correlated (R^2: 94%) with indoor $PM_{2.5}$ measured with HPEMs in a subset of homes, consistent with previous results (Table 2) indicating that the Dylos provides a very good approximation of $PM_{2.5}$ concentrations. Second, the extensive data cleaning procedures led to a large loss of Dylos data. Nevertheless, my research made use of a relatively large dataset with ample observations for model building (445 7-day measurements from 323 homes). Third, many predictors were derived from questionnaires, and while questionnaires are easily implemented in a large cohort, information obtained by questionnaire may be inaccurate. This might be the reason that the primary MLR model did not have any predictors related to indoor pollution sources, which were all assessed by questionnaire. Fourth, due to problems with the way the internal air cleaner usage timer was created by the manufacturer, the data were unreliable and I was unable to incorporate this measure of air cleaner use into the models. In addition, I had

limited information on the use of other air cleaners (e.g. number of operating hours during pollution monitoring periods) purchased by participants themselves. Finally, I had access to data from only two outdoor $PM_{2.5}$ monitoring sites. Despite this limitation, the strong correlation between 7-day average $PM_{2.5}$ concentrations measured at the two locations suggested that the temporal patterns in $PM_{2.5}$ are very similar in different areas of the city.

Chapter 6.

Conclusion

The MLR model based on a variable selection procedure used previously in the development of land use regression models generally had slightly better predictive performance than a RFR model and MLR models that used other variable selection methods. The prediction model developed in this thesis explained much more variability than intervention status alone, and the predictions from my model will be used in future epidemiologic analyses in the UGAAR cohort. Based on indoor $PM_{2.5}$ predictions from MLR model in the homes of all UGAAR study participants over the course of pregnancy, deploying one or two air cleaners was effective to reduce indoor $PM_{2.5}$ exposures. In particular, deploying 2 air cleaners led to considerable reductions in indoor $PM_{2.5}$ concentrations over the full pregnancy.

References

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 370-374.

Allen, R. W., Adar, S. D., Avol, E., Cohen, M., Curl, C. L., Larson, T., . . . Kaufman, J. D. (2012). Modeling the Residential Infiltration of Outdoor PM^ sub 2.5^ in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environmental Health Perspectives, 120*(6), 824.

Allen, R. W., Gombojav, E., Barkhasragchaa, B., Byambaa, T., Lkhasuren, O., Amram, O., . . . Janes, C. R. (2013). An assessment of air pollution and its attributable mortality in Ulaanbaatar, Mongolia. *Air Quality, Atmosphere & Health, 6*(1), 137-150.

Anderson, J. D., & Wendt, J. (1995). *Computational fluid dynamics* (Vol. 206): Springer.

Bell, M. L., Belanger, K., Ebisu, K., Gent, J. F., Lee, H. J., Koutrakis, P., & Leaderer, B. P. (2010). Prenatal exposure to fine particulate matter and birth weight: variations by particulate constituents and sources. *Epidemiology (Cambridge, Mass.), 21*(6), 884.

Berry, W. D. (1993). *Understanding regression assumptions* (Vol. 92): Sage Publications.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brown, D. R., Alderman, N., Weinberger, B., Lewis, C., Bradley, J., & Curtis, L. (2014). Outdoor wood furnaces create significant indoor particulate pollution in neighboring homes. *Inhalation toxicology, 26*(10), 628-635.

Byun, D. W. (1999). Dynamically consistent formulations in meteorological and air quality models for multiscale atmospheric studies. Part II: Mass conservation issues. *Journal of the Atmospheric Sciences, 56*(21), 3808-3820.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions, 7*, 1525-1534.

Challoner, A., Pilla, F., & Gill, L. (2015). Prediction of Indoor Air Exposure from Outdoor Air Quality Using an Artificial Neural Network Model for Inner City Commercial Buildings. *International journal of environmental research and public health, 12*(12), 15233-15253.

Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment, 112*(6), 2999-3011.

Chithra, V., & Nagendra, S. S. (2014). Impact of outdoor meteorology on indoor PM 10, PM 2.5 and PM 1 concentrations in a naturally ventilated classroom. *Urban Climate, 10*, 77-91.

Cohen, M. A., Adar, S. D., Allen, R. W., Avol, E., Curl, C. L., Gould, T., . . . Larson, T. V. (2009). Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and air pollution (MESA air). *Environmental science & technology, 43*(13), 4687-4693.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Cyrys, J., Pitz, M., Bischof, W., Wichmann, H.-E., & Heinrich, J. (2004). Relationship between indoor and outdoor levels of fine particle mass, particle number concentrations and black smoke under different ventilation conditions. *Journal of Exposure Science and Environmental Epidemiology, 14*(4), 275-283.

Dacunto, P. J., Klepeis, N. E., Cheng, K.-C., Acevedo-Bolton, V., Jiang, R.-T., Repace, J. L., . . . Hildemann, L. M. (2015). Determining PM 2.5 calibration curves for a low-cost particle monitor: common indoor residential aerosols. *Environmental Science: Processes & Impacts, 17*(11), 1959-1966.

Davy, P. K., Gunchin, G., Markwitz, A., Trompetter, W. J., Barry, B. J., Shagjjamba, D., & Lodoysamba, S. (2011). Air particulate matter pollution in Ulaanbaatar, Mongolia: determination of composition, source contributions and source locations. *Atmospheric Pollution Research, 2*(2), 126-137.

Elbayoumi, M., Ramli, N. A., & Yusof, N. F. F. M. (2015). Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM 2.5–10 and PM 2.5 concentrations in naturally ventilated schools. *Atmospheric Pollution Research, 6*(6), 1013-1023.

Elbayoumi, M., Ramli, N. A., Yusof, N. F. F. M., Yahaya, A. S. B., Al Madhoun, W., & Ul-Saufie, A. Z. (2014). Multivariate methods for indoor PM 10 and PM 2.5 modelling in naturally ventilated schools buildings. *Atmospheric Environment, 94*, 11-21.

Emmerich, S. J. (1997). *Use of computational fluid dynamics to analyze indoor air quality issues*: National Institute of Standards and Technology.

Environmental Systems Research Institute. (2012). ArcGIS Release 10.1.

Ferziger, J. H., & Peric, M. (2012). *Computational methods for fluid dynamics*: Springer Science & Business Media.

Fisk, W. J., & Chan, W. R. (2017). Effectiveness and cost of reducing particle-related mortality with particle filtration. *Indoor air*.

Gan, G. (1995). Evaluation of room air distribution systems using computational fluid dynamics. *Energy and Buildings, 23*(2), 83-93.

Gauvin, S., Reungoat, P., Cassadou, S., Dechenaux, J., Momas, I., Just, J., & Zmirou, D. (2002). Contribution of indoor and outdoor environments to PM2. 5 personal exposure of children—VESTA study. *Science of the total environment, 297*(1), 175-181.

Gerharz, L. E., Krüger, A., & Klemm, O. (2009). Applying indoor and outdoor modeling techniques to estimate individual exposure to PM2. 5 from personal GPS profiles and diaries: a pilot study. *Science of the total environment, 407*(18), 5184-5193.

Götschi, T., Oglesby, L., Mathys, P., Monn, C., Manalis, N., Koistinen, K., . . . Künzli, N. (2002). Comparison of black smoke and PM2. 5 levels in indoor and outdoor environments of four European cities. *Environmental science & technology, 36*(6), 1191-1197.

Goyal, R., & Khare, M. (2011). Indoor air quality modeling for PM 10, PM 2.5, and PM 1.0 in naturally ventilated classrooms of an urban Indian school building. *Environmental monitoring and assessment, 176*(1), 501-516.

Guttikunda, S. (2008). Urban air pollution analysis for Ulaanbaatar, Mongolia.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). Advanced diagnostics for multiple regression: A supplement to multivariate data analysis: Upper Saddle River, NJ: Prentice Hall.

Henderson, S. B., Beckerman, B., Jerrett, M., & Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental science & technology, 41*(7), 2422-2428.

Hyder, A., Lee, H. J., Ebisu, K., Koutrakis, P., Belanger, K., & Bell, M. L. (2014). PM2. 5 exposure and birth outcomes: use of satellite-and monitor-based data. *Epidemiology (Cambridge, Mass.), 25*(1), 58.

Jiang, J. (2016). How to Apply Data Science in Commodity Industry Analysis- Part Three. Retrieved from https://www.linkedin.com/pulse/how-apply-data-science-commodity-industry-analysis-part-jeff-jiang-1

Jones, P., & Whittle, G. (1992). Computational fluid dynamics for building air flow prediction—current status and capabilities. *Building and Environment, 27*(3), 321-338.

Jones, Z., & Linder, F. (2015). *Exploratory data analysis using random forests.* Paper presented at the Prepared for the 73rd annual MPSA conference.

Kalkbrenner, A. E., Hornung, R. W., Bernert, J. T., Hammond, S. K., Braun, J. M., & Lanphear, B. P. (2010). Determinants of serum cotinine and hair cotinine as biomarkers of childhood secondhand smoke exposure. *Journal of Exposure Science and Environmental Epidemiology, 20*(7), 615-624.

Keil, C. (2000). A tiered approach to deterministic models for indoor air exposures. *Applied occupational and environmental hygiene, 15*(1), 145-151.

Kindangen, J. I. (1996). Artificial neural networks and naturally ventilated buildings: a method of predicting window size and location with subsequent effect on interior air motion using neural networks. *Building research and information, 24*(4), 203-208.

Kioumourtzoglou, M.-A., Spiegelman, D., Szpiro, A. A., Sheppard, L., Kaufman, J. D., Yanosky, J. D., . . . Suh, H. (2014). Exposure measurement error in PM 2.5 health effects studies: a pooled analysis of eight personal exposure validation studies. *Environmental Health, 13*(1), 2.

Klepeis, N. E., Hughes, S. C., Edwards, R. D., Allen, T., Johnson, M., Chowdhury, Z., . . . Hovell, M. F. (2013). Promoting smoke-free homes: a novel behavioral intervention using real-time audio-visual feedback on airborne particle levels. *PloS one, 8*(8), e73251.

Lai, H., Bayer-Oglesby, L., Colvile, R., Götschi, T., Jantunen, M., Künzli, N., . . . Nieuwenhuijsen, M. (2006). Determinants of indoor air concentrations of PM 2.5, black smoke and NO 2 in six European cities (EXPOLIS study). *Atmospheric Environment, 40*(7), 1299-1313.

Lavigne, E., Ashley-Martin, J., Dodds, L., Arbuckle, T. E., Hystad, P., Johnson, M., . . . Fisher, M. (2016). Air Pollution Exposure During Pregnancy and Fetal Markers of Metabolic Function The MIREC Study. *American journal of epidemiology*, kwv256.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician, 36*(3a), 158-160.

McCracken, J. P., Schwartz, J., Bruce, N., Mittleman, M., Ryan, L. M., & Smith, K. R. (2009). Combining individual-and group-level exposure information: child carbon monoxide in the Guatemala woodstove randomized control trial. *Epidemiology, 20*(1), 127-136.

Meng, Q. Y., Spector, D., Colome, S., & Turpin, B. (2009). Determinants of indoor and personal exposure to PM 2.5 of indoor and outdoor origin during the RIOPA study. *Atmospheric Environment, 43*(36), 5750-5758.

Milner, J., Vardoulakis, S., Chalabi, Z., & Wilkinson, P. (2011). Modelling inhalation exposure to combustion-related air pollutants in residential buildings: application to health impact assessment. *Environment international, 37*(1), 268-279.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. Asabe, 50*(3), 885-900.

Nethery, E., Leckie, S. E., Teschke, K., & Brauer, M. (2008). From measures to models: an evaluation of air pollution exposure assessment for epidemiological studies of pregnant women. *Occupational and environmental medicine, 65*(9), 579-586.

Nielsen, P. V. (2004). Computational fluid dynamics and room air movement. *Indoor air, 14*(s7), 134-143.

Northcross, A. L., Edwards, R. J., Johnson, M. A., Wang, Z.-M., Zhu, K., Allen, T., & Smith, K. R. (2013). A low-cost particle counter as a realtime fine-particle mass monitor. *Environmental Science: Processes & Impacts, 15*(2), 433-439.

Ochir, C., & Smith, K. R. (2014). *Air pollution and health in Ulaanbaatar*. Retrieved from http://ehsdiv.sph.berkeley.edu/krsmith/publications/2014/UB%20Final%20Report%20-%20July%2010.pdf

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research & evaluation, 8*(2), 1-9.

Ott, W. R. (1999). Mathematical models for predicting indoor air quality from smoking activity. *Environmental Health Perspectives, 107*(Suppl 2), 375.

Panagopoulos, I. K., Karayannis, A. N., Kassomenos, P., & Aravossis, K. (2011). A CFD simulation study of VOC and formaldehyde indoor air pollution dispersion in an apartment as part of an indoor pollution management plan. *Aerosol and Air Quality Resarch, 11*(6), 758-762.

Pandey, G., Zhang, B., & Jian, L. (2013). Predicting submicron air pollution indicators: a machine learning approach. *Environmental Science: Processes & Impacts, 15*(5), 996-1005.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199.

Ritz, B., & Wilhelm, M. (2008). Ambient air pollution and adverse birth outcomes: methodologic issues in an emerging field. *Basic & clinical pharmacology & toxicology, 102*(2), 182-190.

Ryan, P. H., Brokamp, C., Fan, Z.-H., & Rao, M. (2015). Analysis of Personal and Home Characteristics Associated with the Elemental Composition of PM2. 5 in Indoor, Outdoor, and Personal Air in the RIOPA Study. *Research report (Health Effects Institute)*(185), 3-40.

Savitz, D. A., Bobb, J. F., Carr, J. L., Clougherty, J. E., Dominici, F., Elston, B., . . . Matte, T. D. (2014). Ambient fine particulate matter, nitrogen dioxide, and term birth weight in New York, New York. *American journal of epidemiology, 179*(4), 457-466.

Semple, S., Apsley, A., & MacCalman, L. (2012). An inexpensive particle monitor for smoker behaviour modification in homes. *Tobacco Control*, tobaccocontrol-2011-050401.

Semple, S., Ibrahim, A. E., Apsley, A., Steiner, M., & Turner, S. (2013). Using a new, low-cost air quality sensor to quantify second-hand smoke (SHS) levels in homes. *Tobacco Control*, tobaccocontrol-2013-051188.

Shah, P. S., & Balkhair, T. (2011). Air pollution and birth outcomes: a systematic review. *Environment international, 37*(2), 498-516.

Shilpa, B., & Lokesh, K. (2013). Models for Indoor Pollution and Health Impact Assessment–An Overview. *International Journal of Emerging Technology & Advanced Engineering (ISSN 2250-2459, ISO 9001: 2008 Certified Journal), 3*(4), 519-525.

Shyur, H.-J. (2003). A stochastic software reliability model with imperfect-debugging and change-point. *Journal of Systems and Software, 66*(2), 135-141.

Smith, McCracken, J. P., Weber, M. W., Hubbard, A., Jenny, A., Thompson, L. M., . . . Bruce, N. (2011). Effect of reduction in household air pollution on childhood pneumonia in Guatemala (RESPIRE): a randomised controlled trial. *The Lancet, 378*(9804), 1717-1726.

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods, 220*(1), 85-91.

Sørensen, D. N., & Nielsen, P. V. (2003). Quality control of computational fluid dynamics in indoor environments. *Indoor air, 13*(1), 2-17.

Steinle, S., Reis, S., Sabel, C. E., Semple, S., Twigg, M. M., Braban, C. F., . . . Lin, C. (2015). Personal exposure monitoring of PM 2.5 in indoor and outdoor microenvironments. *Science of the total environment, 508*, 383-394.

Sublett, J. L. (2011). Effectiveness of air filters and air cleaners in allergic respiratory diseases: a review of the recent literature. *Current allergy and asthma reports, 11*(5), 395-402.

Sun, G., & Hoff, S. J. (2009). *Prediction of indoor climate and long-term air quality using a building thermal transient model, artificial neural networks and typical meteorological year.* Paper presented at the 2009 Reno, Nevada, June 21-June 24, 2009.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). Using multivariate statistics.

Thompson. (2016). Crowd-sourced air quality studies: A review of the literature & portable sensors. *Trends in Environmental Analytical Chemistry, 11*, 23-34.

Trevor, H., Robert, T., & Jerome, F. (2001). The elements of statistical learning: data mining, inference and prediction. *New York: Springer-Verlag, 1*(8), 371-406.

United Nations Children's Fund. (2016). *Understading and addressing the impact of air pollution on children's health*. Retrieved from https://www.unicef.org/environment/files/Understanding_and_addressing_the_impact_of_air_pollution.pdf

Vijayan, V. K., Paramesh, H., Salvi, S. S., & Dalal, A. A. K. (2015). Enhancing indoor air quality–The air filter advantage. *Lung India: official organ of Indian Chest Society, 32*(5), 473.

Wang, Y., Chen, C., Wang, P., Wan, Y., Chen, Z., & Zhao, L. (2015). Experimental Investigation on Indoor/Outdoor PM 2.5 Concentrations of an Office Building Located in Guangzhou. *Procedia Engineering, 121*, 333-340.

Williams, R., Kaufman, A., Hanley, T., Rice, J., & Garvey, S. (2014). *Evaluation of Field-deployed Low Cost PM Sensors*. Retrieved from https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=297517

Willmott, C. J. (1981). On the validation of models. *Physical geography, 2*(2), 184-194.

Yuchi, W., Knudby, A., Cowper, J., Gombojav, E., Amram, O., Walker, B. B., & Allen, R. W. (2016). 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*.

Appendix: Figures and Codes

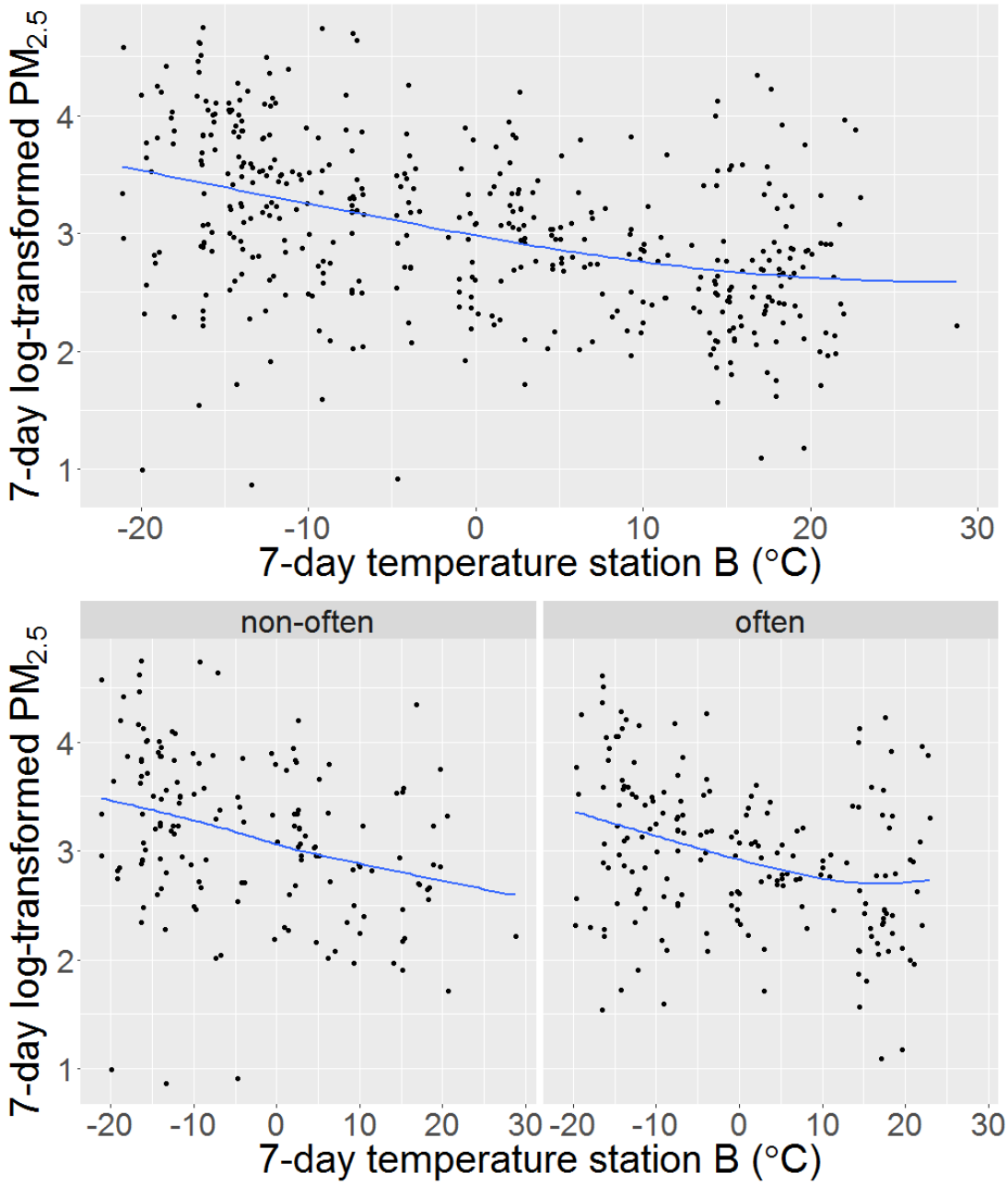

Figure A1a. Relationship between 7-day log-transformed $PM_{2.5}$ and temperature from station B (top panel). Relationship between 7-day log-transformed $PM_{2.5}$ and temperature from station B stratified by typically window opening behavior in January (bottom panel).

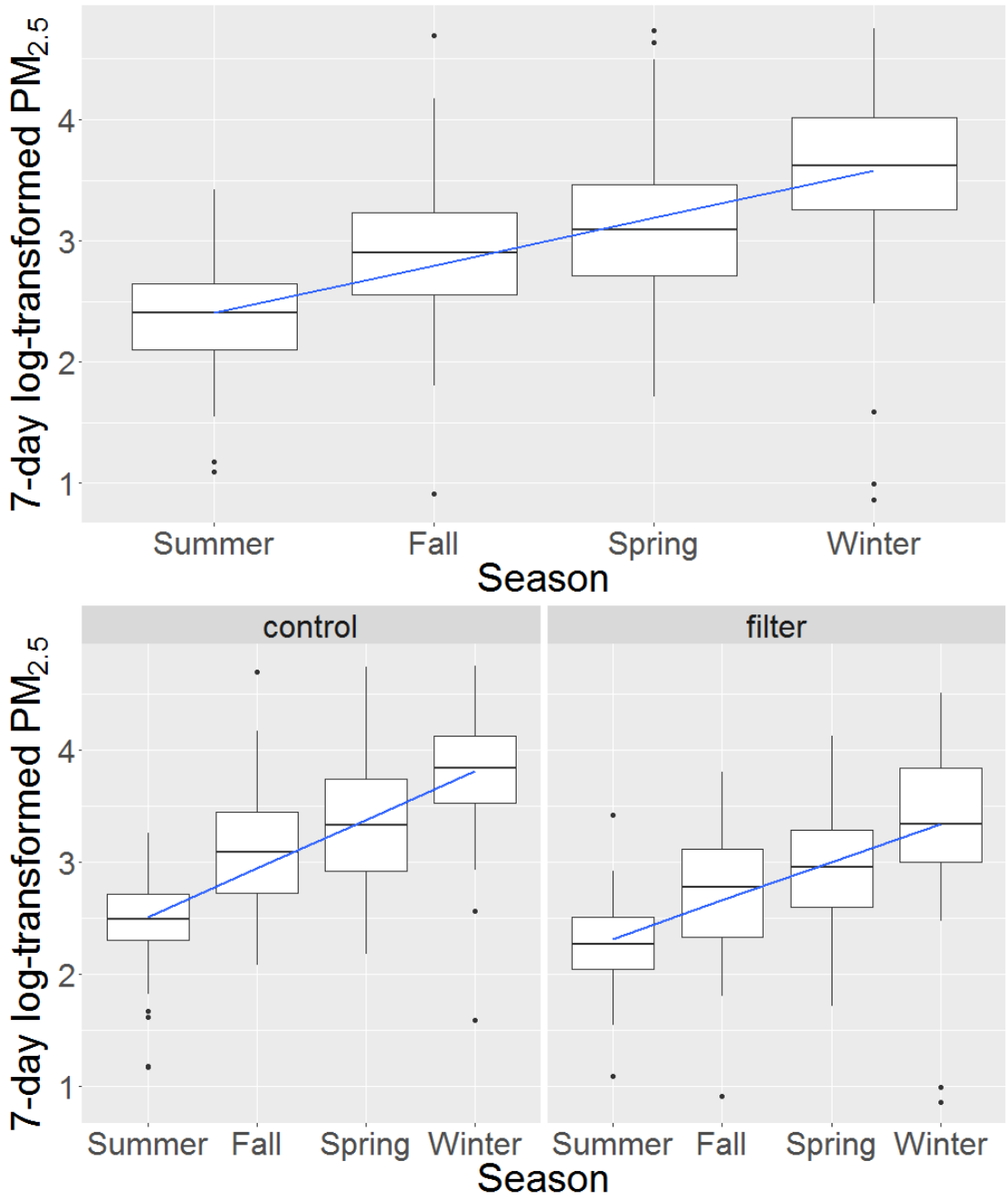

Figure A1b. Relationship between 7-day log-transformed $PM_{2.5}$ and season (top panel). Relationship between 7-day log-transformed $PM_{2.5}$ and season stratified by intervention status (bottom panel).

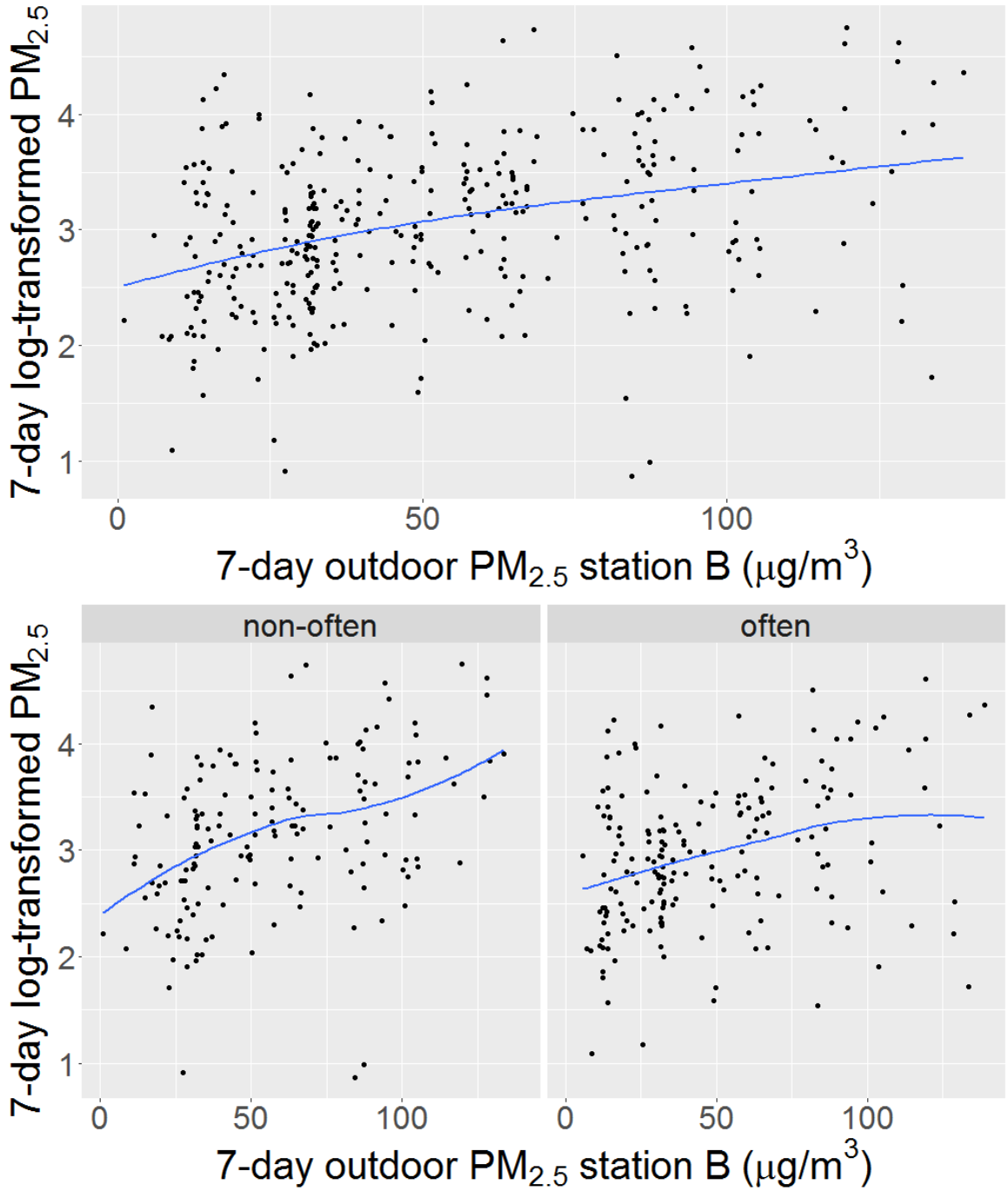

Figure A1c. Relationship between 7-day log-transformed $PM_{2.5}$ and outdoor $PM_{2.5}$ from station B (top panel). Relationship between 7-day log-transformed $PM_{2.5}$ and outdoor $PM_{2.5}$ from station B stratified by typically window opening behavior in January (bottom panel).

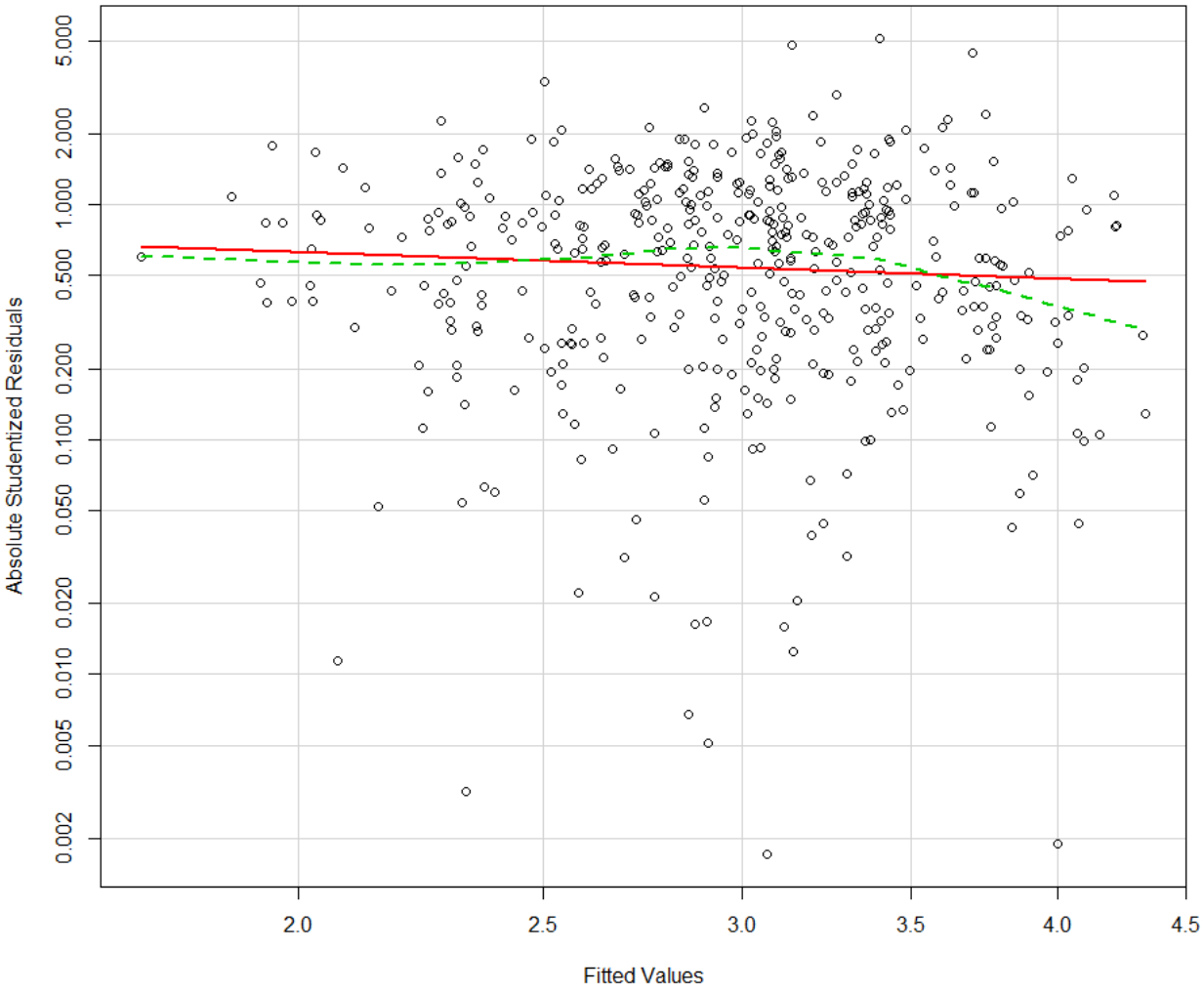

Figure A2. Spread Level plot for the test of linearity and homoscedasticity assumptions

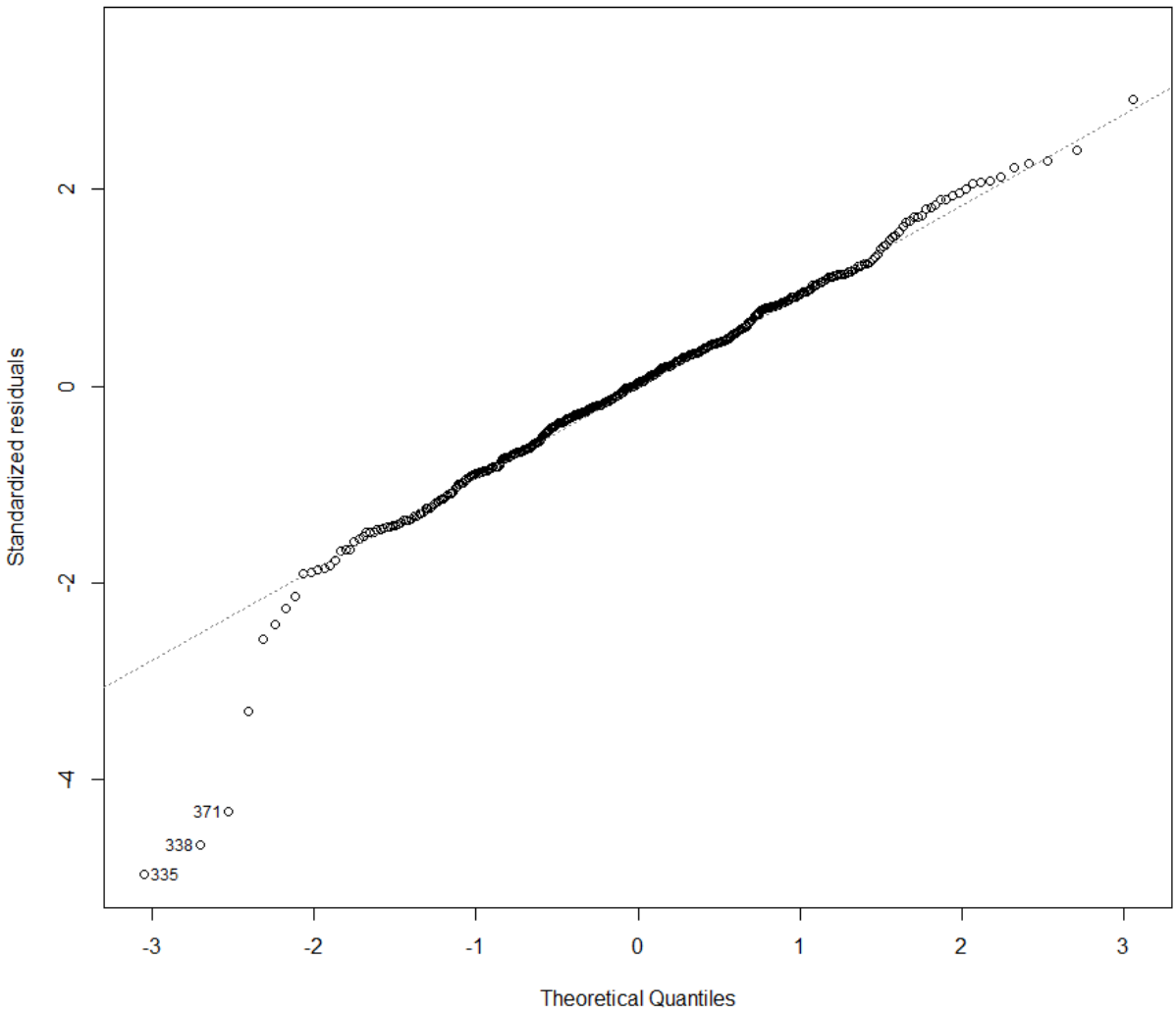

Figure A3. Q-Q plot for the test of normality assumption

Random Forest Regression R codes:

```
library(randomForest)
  set.seed(1)
  ## number in this bracket can be any number. It makes sure that I can reproduce final results

  thesis_rfr = randomForest(Indoor_PM2.5 ~ ., data=dataset, importance=TRUE,
                            na.action=na.omit, ntree=500)

  library(plyr)
  library(dplyr)
  data = dataset

  set.seed(0)
  k = 10
  data$id = sample(1:k, nrow(data), replace =TRUE)
  list = 1:10
  ## randomly create 10 numbers

  prediction = data.frame()
  testsetCopy = data.frame()
  progress.bar = create_progress_bar("text")
  progress.bar$init(k)
  ## create empty dataframes to store predicted values and test dataset

  #function for 10 fold validation
  for(i in 1:k){
    # remove rows with id i from dataframe to create training set
    # select rows with id i to create test set
    trainingset <- subset(data, id %in% list[-i])
    testset <- subset(data, id %in% c(i))

    # run a random forest
    mymodel <- randomForest(trainingset$Indoor_PM2.5 ~ ., data = trainingset, ntree =
500,na.action=na.omit)
    #mymodel                    <-lm(data                    =                    trainingset,
Indoor_PM2.5~Season+OPM2.5_4+filters_deployed+ger_5000+OPM2.5_4*filters_deployed)

    #remove response indoor pm2.5 in column 1
    temp <- as.data.frame(predict(mymodel, testset[,-1]))

    # append this iteration's predictions to the end of the prediction data frame
    prediction <- rbind(prediction, temp)

    # append this iteration's test set to the test set copy data frame
    # keep only the pm2.5 Column
    testsetCopy <- rbind(testsetCopy, as.data.frame(testset[,1]))
    progress.bar$step()
  }

  result <- cbind(prediction, testsetCopy[, 1])
```

```
names(result) <- c("Predicted", "Actual")
result$Difference <- abs(result$Actual - result$Predicted)
```