NAZARBAYEV UNIVERSITY

Department of Civil and Environmental Engineering

MCEE 602 MSc Thesis II

Quantifying PM Infiltration in Kazakh Homes under Extreme Weather Conditions

Aziz Talapov ID: 201964449

Date of submission: 28.03.2025

Astana, 2025

Abstract

Since people spend most of their time indoors, indoor air quality has emerged as a serious concern. Particulate matter with a diameter of less than 2.5 micrometers ($PM_{2.5}$) is widely studied due to its serious health impacts. The $PM_{2.5}$ concentration inside a building depends on indoor pollution sources and outdoor concentration. This study focuses on $PM_{2.5}$ infiltration modeling in Astana city households.

Eight buildings across Astana were analyzed. Data collection involved indoor and outdoor air monitoring in each household. The effect of indoor pollution sources was minimized by filtering the indoor concentration data. An increase in $PM_{2.5}$ concentration indoors should appear following an increase in outdoor concentration. Multiple linear regression was performed to identify the relationship between indoor $PM_{2.5}$ concentration and both indoor and outdoor parameters.

As a result, the model accuracy for two households reached almost 90%, while the lowest value was 47%. The results could be improved with better separation of $PM_{2.5}$ data based on its origin. More efficient removal of indoor-related $PM_{2.5}$ concentration changes could result in better model performance. Additionally, the relationships between indoor $PM_{2.5}$ and other parameters, such as temperature and relative humidity, are discussed, as well as the seasonal characteristics of infiltration.

1. Introduction

1.1 Overview

This is the draft of the master's thesis. It includes the research topic, literature review, experiment results and future work plan.

1.2. Thesis Statement

This study aims to collect empirical data on atmospheric fine particles ($PM_{2.5}$) concentrations both indoors and outdoors within household environments. Furthermore, it conducts a detailed analysis of air quality and infiltration mechanisms by aiming to develop a multiple regression based model which explains the infiltration mechanisms. It provides a detailed methodology on modeling principles and the used parameters, provides the first infiltration model for Kazakh houses. The independent factors such as extreme weather conditions also make this research a globally noval challenge since the developed model is one of its kind in the literature.

1.3. Objectives

The main objectives of the work are:

1. Performing several experiments in household environments for monitoring $PM_{2.5}$ concentrations.
2. Explaining the infiltration mechanisms by developing a theoretical background.
3. Understanding the main mechanisms play critical roles in the modeling of the PM infiltration.
4. Performing data analyses and developing a mathematical model of the houses to predict PM concentrations related to outdoor levels.

1.4 Literature review

Indoor air pollution has emerged as a critical concern globally, with adverse impacts on human health and quality of life. With increasing urban population and growing cities people spend more time indoors. The indoor environment has a lower degree of dispersion, chemical transformation and dilution. Thus, the indoor exposure per unit mass of particulate matter (PM) is two to three times higher than outdoors (Scibor and Anita, 2020). The outdoor environment is not dependent on the number of people, however, the indoor environment is very sensitive to the number of occupants. High number of occupants has significant effects on the indoor environment than the outdoor. Different toxicological and physicochemical properties of particles generated outdoor and indoor also emphasize the need to differentiate between exposure to indoor- and outdoor-generated particles (Diapouli et al., 2013).

Indoor air PM primarily comprises salts such as ammonium sulfate, ammonium nitrate, sodium chloride, and potassium chloride, along with soot (elemental carbon, EC), minerals like silicon, aluminum, calcium, iron, manganese, titanium, zinc, among others, organic compounds (particulate organic matter or organic carbon, OC), and biological materials such as bacteria, fungi, dander, pollen, and fragments of plants and insects. All mentioned components have different particle size distributions (Arvanitis, 2010). In the past, air pollution research was mostly focused on the outdoor environment due to the significant contribution of industrial activities. However, despite their smaller scale, indoor air pollution sources also have adverse impacts on human health and quality of life. Fine fraction of the particulate matter ($PM_{2.5}$) is among the primary pollutants of concern, originating from various sources such as burning processes, building materials, and human activities. Nevertheless, there remains a gap in understanding the specific dynamics of indoor air quality in residential settings, particularly in regions with harsh climatic conditions such as Astana.

When outdoor pollutants concentration is higher than indoor pollutants concentration, the indoor environment is highly influenced and the indoor pollution levels are dependent on the outdoor environment (Kalimeri et al., 2019). Chen and Zhao showed that cities can have the same increase in the concentration of outdoor PM, however the increases in indoor concentrations are different from place to place (Chen and Zhao, 2012). They focused on 18 cities in the USA and

defined the annual average infiltration rate for each city. They used an overall air change rate ($\lambda_{overall}$) which considers both open and closed windows. The condition with closed windows is marked as λ_{infilt} and it was assumed windows were open only in residences without air conditioning. Air conditioning was considered as an alternative to window opening. There were different types of windows but one important parameter which describes the windows is their airtightness.

Wan et al. (2015) performed experiments in two office buildings in Beijing. Logically, the windows with low airtightness allowed more air to infiltrate into and leave the building while windows with high airtightness limited the amount of air transfer. Additionally, it was found that the I/O ratio was increasing with higher wind speed and increasing humidity of the outdoor air decreased the I/O ratio.

The older buildings usually have more leakages through exterior walls. Air conditioning is designed for cooling the indoor air and not to provide "fresh" outdoor air. As a result, households with air conditioning systems are less exposed to infiltrating pollutants. However, using air conditioning may cause a buildup of CO_2. Wong and Huang' study in Singapore where air conditioners were widely used showed that in all monitored houses with different air conditioner models, the CO_2 concentrations were higher than 1000 ppm (Wong and Huang, 2004). Their study compared bedrooms where the air conditioner was used during night sleep with naturally ventilated bedrooms. The former bedrooms had higher CO_2 concentrations. It should be also noted that reducing the ventilation rate has a potential to decrease pollutants' concentration that infiltrate from the outside into the household, however, it may increase pollutants concentration of indoor origin (Chen and Zhao, 2012).

Low indoor air quality may lead to health issues and even mortality of vulnerable populations. In Kazakhstan, outdoor air quality is monitored, and some preliminary research in this domain exists. However, indoor air quality is not well explored with only a limited number of recent ongoing investigations. For instance, the sources of indoor pollution were investigated, along with their effects on particulate matter (PM), volatile organic compounds (VOCs), carbon dioxide (CO_2), and other pollutant concentrations (Jones, 1999). However, there isn't any research into the infiltration and air exchange mechanisms. Outdoor pollution sources are

diverse, meteorological conditions are extreme in winter, and the infiltration effect of outdoor pollutants in residential buildings in Kazakhstan, specifically in Astana, has yet to be thoroughly studied.

The harmful health impacts of outdoor $PM_{2.5}$ extensively evidenced in numerous studies, predominantly signify the health consequences due to PM exposures indoor, originating from outdoor sources. Over the past two decades, extensive multi-city studies, cohort analyses, and time-series investigations have consistently revealed connections between environmental exposure to PM and elevated rates of cardiovascular and respiratory hospitalizations as well as mortality (Arvanitis, 2010). $PM_{2.5}$ is often linked with a high risk of non-accidental mortality. Multiple studies conclude that $PM_{2.5}$ leads to lung cancer, ischemic heart disease, and diabetes (Bekierski, 2021). The size of $PM_{2.5}$ is smaller than inhalable particles (PM_{10}), which allows them to penetrate deeper inside the respiratory system. PM_{10} is mostly trapped by nasal hair and mucosa, which filter the inhaling air. However, $PM_{2.5}$ can reach the lungs' alveoli where gas exchange occurs and where it can be absorbed into the bloodstream through the bronchioles.

In Europe, typical indoor $PM_{2.5}$ concentrations fall within the range of 10 to 65 $\mu g/m^3$ (Arvanitis, 2010). The regions with arid or semi-arid climate are strongly influenced by dust storms (Krasnov et al., 2015). Many people in Asia and Africa are exposed to strong PM concentrations because of proximity to dust sources and dust storms. Walker et al. measured indoor and outdoor (I&O) $PM_{2.5}$ concentrations during a wildfire season. Measurements were taken from July to October in 2022 in Western Montana (Walker et al., 2023). A low cost $PM_{2.5}$ sensors were set up in 20 residences. The median outdoor $PM_{2.5}$ concentration was 3.7 $\mu g/m^3$ for the whole four months study period and median 29.0 $\mu g/m^3$ for a two weeks wildfire period. Corresponding indoor $PM_{2.5}$ concentrations were 2.5 $\mu g/m^3$ for four months and 10.4 $\mu g/m^3$ during the wildfire smoke influence. The indoor PM concentrations in most cases remain significantly lower than the outdoor concentrations.

Scibor and Anita measured I&O concentrations of $PM_{2.5}$ and PM_{10} for 23 days during cold periods between 2013 and 2015. The study was conducted during good and very poor weather conditions in Krakow, Poland. The authors aimed to study the effect of wind on $PM_{2.5}$ and PM_{10}

infiltration. The poor weather conditions is assigned to days with wind velocity less than 2 m/s according to three measuring stations in the Krakow rural area. The good weather conditions appear when the wind velocity is higher than 3 m/s. The measurement results were divided into sub-series: good weather conditions with open or closed windows and the same for bad weather conditions. Both I&O concentrations for $PM_{2.5}$ and PM_{10} are higher during the poor weather conditions while the same variables are three to four times lower during good weather conditions. The wind is necessary for the dilution and diffusion of the pollutants in the atmosphere. However its impact on the indoor environment is controversial. Addressing these variations, especially in human behavior, such as opening the window, is necessary for better epidemiology studies and estimation of personal exposures (Orch et al., 2014). Some studies also used questionnaires to identify what activities people do indoors, what is their lifestyle, how do they feel themself indoors (Bai et al., 2019; Wong and Huang, 2004).

Experiments by Thatcher and Layton concluded that PM greater than 5 micrometers can be resuspended, PM less than 5 micrometers is not easy to resuspend, while PM less than 1 micrometer almost do not resuspend (Thatcher, 1995). Once particles enter the home, they undergo deposition onto indoor surfaces, a process influenced by their size. The rate of deposition varies among homes because of architecture and interior. The amount and size of the furniture can have an effect on velocity of airflows. The ratio of interior surface area to volume, and temperature differentials between the air and surfaces are also important (Bennett and Koutrakis, 2006).

Assuming there are no indoor sources of pollution, the main indoor pollution enrichment mechanism will be infiltration. The indoor and outdoor indicator can be implemented to describe the temporary steady state of the airborne particles (Equation 1).

$$\frac{I}{O} ratio = \frac{C_{in}}{C_{out}} \quad (1)$$

The indoor to outdoor ratio (I/O) allows us to assess the difference between I&O concentrations. It is considered good when the I/O ratio is less than 1, indicating minimal indoor pollution sources. In a study conducted in Krakow, the I/O ratio for good weather conditions was 0.92 while windows are open and 0.79 for closed windows (Scibor and Anita, 2020). The results for

bad weather conditions were 0.46 and 0.47, respectively. Precipitation can settle down coarse particles and wash them away, but it has little effect on fine particle concentrations. Weather conditions are important for indoor PM concentration, for example, increase in outdoor temperature, horizontal wind velocity and precipitation during winter showed decrease in indoor PM_1 concentration (Scibor and Anita, 2020).

The investigation by Chan (2002) explored the I/O ratio for different meteorological conditions. The PM I/O ratio increases during higher temperatures, humidity, and wind speed. Morawska and Gilbert studied particle deposition rates in houses. Deposition rates were found by fitting a size resolved particle number and $PM_{2.5}$ concentration decay curve. Their model estimated for small particles (<0.5 μm) were different from the experimental results (Morawska and Gilbert, 2005). The study of infiltration efficiency estimation in Beijing in winter season showed that the rise in relative humidity (RH) from 19.20 to 51.30 percent led to decrease of improved ventilation (F_{inf}) (Ma et al., 2023).

Nadali et al. (2020) constructed a symmetrical correlation matrix to examine possible correlations between concentrations of PM_1, $PM_{2.5}$, and PM_{10}. The correlation between indoor concentrations and factors such as the date of construction, ventilation, number/size of windows, and indoor smoking showed less significance. Diapouli (2013) also mentioned different methods of estimation of particle infiltration and gives categorization into 4 groups as summarized in Table 1.

Steady-State Assumption	**Dynamic Solution of the Mass Balance Equation**	**Experimental Studies**	**Infiltration Surrogates**
This approach utilizes the mass balance equation in its steady-state form to estimate infiltration	These models offer a more advanced and realistic depiction of PM behavior by considering the	These studies involve conducting experiments under controlled conditions to estimate	This methodology uses a surrogate particulate matter constituent with the absence of indoor

factors. It assumes constant values for all parameters and requires long-term data to ensure the steady-state assumption holds.	dynamic nature of all parameters. They allow for the study of spatiotemporal variations in infiltration parameters and can provide short-term estimates, making them suitable for analyzing specific site and day data.	infiltration parameters. They may simplify model calculations by reducing the number of unknowns but can be challenging to conduct across multiple sites and over extended periods.	sources to estimate the infiltration of outdoor particles into indoor environments. It requires chemical composition data and offers a simpler approach when suitable surrogates are available.

Table 1. Infiltration estimation methods categories.

Each methodology has its advantages and limitations, such as the need for long-term data, difficulties in conducting experiments, and challenges in selecting suitable surrogates for infiltration. Despite challenges, methodologies using dynamic models and infiltration surrogates showed promise for estimating infiltration factors, especially when chemical speciation data were available. Over extended time frames, such as several hours, when air exchange rates and outdoor concentrations remain relatively stable and there were no indoor sources, the infiltration factor could be calculated using a steady-state model. This factor represents the I/O ratio. Various studies measured infiltration ratios when indoor contributions are minimal, such as nighttime. These ratios were derived for different particle size fractions using regression techniques under the assumption of steady-state conditions. Huang et al. in 2015 studied $PM_{2.5}$ infiltration efficiency during the non-heating season. It was estimated by solving mass balance model with steady-state and recursive cases (Huang et al., 2015). Park et al. in 2022 worked on a dynamic mass-balance prediction model for indoor particle concentrations (Park et al., 2022). Their study included the effect of air purifiers on indoor $PM_{2.5}$ concentrations.

Since machine learning (ML) models have been developing fast, their use in indoor air pollution modeling is also noteworthy. A prediction accuracy of $R^2 = 0.94$ was achieved using the machine learning models combined with mechanistic approach (Kim et al., 2024). The advantage of ML modelling is that it can be used for general conditions while conventional mechanistic models are based on the performed experiments and measurements with controlled conditions. Ott et al. (2000) introduced the random-component superposition (RCS) model for estimation of the F_{inf} of ambient PM_{10}. This model employs linear regression analysis to examine ambient and indoor $PM_{2.5}$ concentration data and is commonly utilized to provide estimates of the mean F_{inf} within a specified period. Meng et al. (2007) applied the RCS model to assess F_{inf} of ambient $PM_{2.5}$. The model was used in Houston, Los Angeles, and Elizabeth, across 279 households. They reported F_{inf} for $PM_{2.5}$ equals 0.51, 0.78, and 0.04. This PM originates from combustion, secondary formation, and mechanical generation, respectively (Meng et al., 2007). While the RCS model can estimate F_{inf} for a single family, it is crucial to note that the potential error is minimal only when the data collection duration exceeds 24 hours (Diapouli et al., 2013; Sun et al., 2019). If sampling duration is shorter, a recursive model may be a more suitable alternative. Some authors propose that, over extended durations, the average infiltration rate remains consistent across different homes (Bennett and Koutrakis, 2006).

The research aims to address the issue of indoor air pollution in household environments in Astana, particularly focusing on the infiltration of $PM_{2.5}$. The outdoor air quality measurements in Kazakhstan allow for the assessment of the environment in Kazakhstan cities and the definition of public health risks. Additionally, awareness of the detrimental effects of indoor air pollutants on residents is growing, and this specific area is being studied. Infiltration and air exchange mechanisms remain relatively underexplored in Kazakhstan, particularly in the context of Astana. The research obtained simultaneous measurements of the indoor and outdoor PM concentrations, described infiltration, and air exchange mechanisms, and provided multiple regression based modeling results. The results are useful in creating a healthier living environment if the research catches the attention of policymakers, designers, and builders.

2. Study area description

Astana is the capital city of Kazakhstan. It had an official population of 1,350,228 in 2022. It is the second largest city in Kazakhstan and is located in the north-central part of the country. Since becoming the capital in 1997, the city's growth rates have been increasing. Astana experiences an extreme continental climate characterized by warm summers with occasional brief rain showers and long, very cold, dry winters. Summer temperatures can reach as high as 35°C, while temperatures as low as -30 to -35°C are common between the middle of December and early March (our study period). The city's Ishim river typically freezes over from the second week of November until the beginning of April. The city is known for its frequent high winds, which have a significant impact, especially on the rapidly expanding left bank of the city because of wide streets, huge walking areas and low building density. Astana falls under the continental climate classification according to the Köppen scheme (Dfb). The city experiences an average annual temperature of 3.9°C. January stands out as the coldest month, with an average temperature of -14.5°C , with a record low set during the January 1893 cold wave, plummeting to -51.6°C. On the other hand, July is the warmest month, with an average temperature of 20.6°C. The heating of the city is centralized and powered by three thermal power plants and district boiler houses. The main source of energy is coal. Coal stands out as the biggest pollution contributor (Tursumbayeva et al., 2023). Currently, two thermal power plants are fully operational, providing the city with heating and hot water, while a third plant serves solely for providing hot water. The main air pollutants generated by thermal power plants are: PM, CO, NO, SO_2 and SO_2. The emitted amount of the mentioned pollutants in tons by thermal power plant №1 in 2017 were: 94, 1354, 2558 and 1043 respectively (Ecokarta, 2017). All three thermal power plants are located on the right bank of the Astana. Additionally, there are rural areas that were once villages but are now almost seamlessly integrated into the city. These areas consist mainly of one or two-story houses that are not connected to centralized heating and water supply systems. Instead, they rely on individual coal usage, which stands out as a significant contributor to local air pollution.

The air quality of Astana is not healthy because pollutant concentrations are higher than WHO limit. For instance, the $PM_{2.5}$ annual concentration was 4.5 times higher than the WHO limit in

2021 (Mukhtarov et al., 2023). The air quality in Astana is monitored by seven stationary stations within the city, primarily situated in the Saryarka district, with only one stationary station in the Yessil and Almaty districts. Additionally, mobile stations monitor air quality at various locations such as the Khan-Shatyr district, Sport complex Alau, city hospital №2, National museum, "Alatay" sport complex, Children's city hospital, Children's palace, and high school (see the details in Figure 1).

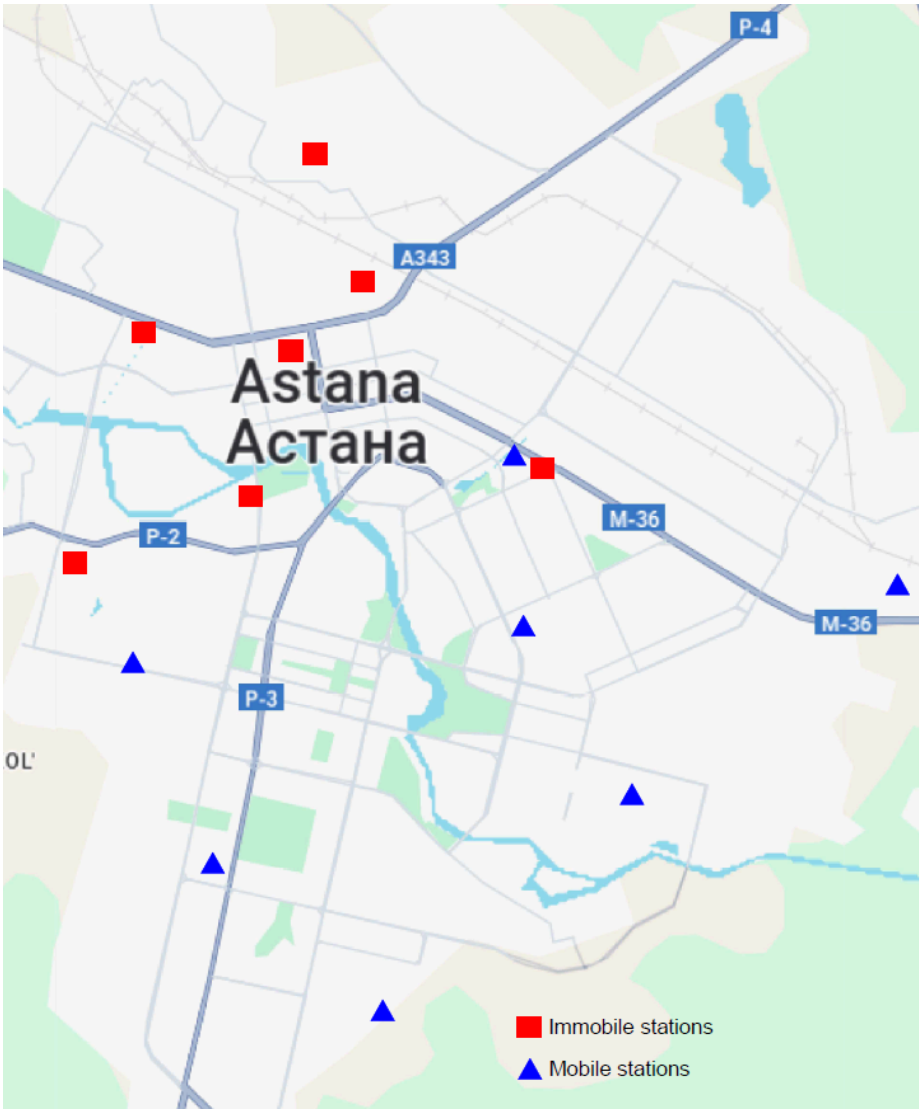

Figure 1. Astana city air quality monitoring stations.

These mobile stations conduct manual air quality measurements three times a day, focusing on pollutants like total suspended particles (TSP), SO_4^{2}, CO, HF, NO_2, SO_2, H_2S, and others (Kerimray et al., 2018). Multiple sources contribute to air pollution in Astana, including

transportation, electricity and heat generation, in addition, individual combustion units. Despite a public transportation system which includes zero-emission buses, the high number of individual drivers exacerbates pollution. Fuel quality remains a challenge due to delayed refinery modernization, leading to the continued use of lower-grade fuels. Astana heavily relies on coal for electricity and heat, with a significant portion of households using coal for heating, often in inefficient stoves without pollutant controls. The city's rapid construction and hosting of events like the 2017 EXPO-2017 International Specialized Exhibition contribute to increased pollution levels. With millions of expected visitors, air pollution is anticipated to rise considerably in 2017 (Kerimray et al., 2018).

The concentrations of air pollutants, as reported by the National Hydrometeorological Service of the Republic of Kazakhstan, have shown alarming levels, with PM_{10}, NO_2, SO_2, and total suspended particles (TSP) exceeding the annual limit values set by WHO, EU, and Kazakhstan standards. Notably, the average annual $PM_{2.5}$ concentration in 2016 was twice the WHO limit value, with PM_{10} levels reaching five times the annual limit value by WHO. PM concentrations tend to peak during the heating season, likely due to increased fuel consumption in winter (Kerimray et al., 2018). The seasonal pattern of the $PM_{2.5}$ was observed by (Mukhtarov et al., 2023). Summer had the lowest average $PM_{2.5}$ concentration of $12.6\ \mu g m^{-3}$ and maximum of $35.5\ \mu g m^{-3}$ was observed in winter.

3. Methodology

3.1 Measurements

In general, there are two types of instruments for measuring PM. The first type is gravimetric method instruments. These instruments can be used if they satisfy the requirements of PN-EN 12341 (2014). The gravimetric method instruments provide average concentrations over the sampling period. Different types of filters (such as glass fiber filter, quartz fiber filter, polytetrafluoroethylene, etc.) are installed inside the equipment, and the filters must be weighed before and after collection.

The second type is optical instruments based on light scattering. The occultation or absorption of light by particles allows for the measurement of real-time concentrations. The instrument emits a beam of light which is interrupted by particles in the air. The sensor contains a photodetector that detects the scattered light. The concentration of the particles is proportional to the intensity of the scattered light. The sensor converts the intensity of the scattered light into a measurement of PM concentration using calibration factors and algorithms. Calibration is typically done by comparing the sensor readings with measurements from reference instruments, such as gravimetric samplers or other high-precision PM monitors. Light scattering sensors offer several advantages for PM measurement, including real-time monitoring capability, low cost, compact size, and ease of use. However, they also have limitations, such as sensitivity to particle composition and size distribution, as well as potential interference from environmental factors like humidity and temperature.

In this study, to monitor I&O $PM_{2.5}$ concentrations simultaneously, indoor activities such as cooking or smoking, which produce PM, are minimized to get a better observation of infiltration. The household environments to be tested have different architecture, location, and elevation levels. The devices used in our measurements are from the TSI company: AirAssure and BlueSky. The AirAssure IAQ Model 8144 is used to measure PM_1, $PM_{2.5}$, PM_4, PM_{10}, Number-based measurements count (NC), CO, CO_2, Barometric Pressure, SO_2, Temperature, and RH (see the details in Table 2). It is installed indoors on the table and connected to a power supply. The AirAssure monitor records variables to a microSD card every minute. There is an option to connect the monitor to Wi-Fi and send results to the TSI account, where online

monitoring is available. To access the results, it is necessary to eject the microSD card and connect it to the computer or download the results from the TSI account.

The BlueSky Model 8143 measures PM_1, $PM_{2.5}$, PM_4, PM_{10}, Temperature and Relative Humidity (see the details in Table 2). It is installed outdoors, at an equivalent elevation level. Similar to the indoor monitor, BlueSky records results on a microSD card or sends results to the TSI account. The measurement duration should be several days to capture enough information in a varying environment. The fluctuations in wind, temperature and other parameters influence the results of the measurements. The studies performed by other researchers often include several days of measurements (Huang et al., 2015; Hossain et al., 2021).

Measuring device	**Measuring variable**	**Units**	**Range**	**Resolution**	**Accuracy**
BlueSky Air Quality Monitor 8143	PM_1 $PM_{2.5}$ PM_4 PM_{10}	$\mu g/m^3$	0 to 1000	1	±10 % (PM 2.5)
	Barometric Pressure	inHg (hPa)	8 to 35 (270 to 1185)	-	± 0.12 (± 4)
	Temperature	°C	-40 to 50	-	±0.5
	Carbon Dioxide	ppm	0 - 10,000	1 ppm	+/- 30 ppm + 3% of reading (Typical)
	Carbon Monoxide	ppm	0 - 20	0.001 ppm	+/- 0.150 ppm (Typical)
	Humidity	RH	0 to 100% RH	-	±3% RH

AirAssure IAQ Model 8144	PM_1 $PM_{2.5}$ PM_4 PM_{10}	$\mu g/m^3$	0 to 1000	1	±10 % (PM 2.5)
	Barometric Pressure	inHg (hPa)	7.7 to 37.2 (270 to 1185)	0.01 (1)	± 0.12 (± 4)
	Temperature	°C	0 - 60	0.1	± 0.5
	Carbon Dioxide	ppm	400 - 10,000	1 ppm	+/- 30 ppm + 3% of reading (Typical)
	Carbon Monoxide	ppm	0 - 1000	100 ppb	±15% of reading or ±150 ppb
	Humidity	RH	0 to 100% RH	1 % RH	±3% RH

Table 2. Specification of the measuring devices (TSI Incorporated BlueSky, 2023), (TSI Incorporated AirAssure, 2023).

Good-quality instruments and measuring methods cost from $10,000 to $100,000. In crowdsourced networks and public air quality initiatives, low-cost sensors are often used. Many low-cost sensors use the aerosol light scattering principle. Manufacturers often omit the precision of their instruments. In general, low-cost instruments show good measurements of PM_1 and $PM_{2.5}$. However, the PM_{10} results are unreliable (Molina Rueda et al., 2023). During our measurements, PM_4 and PM_{10} values are often the same and equal to the $PM_{2.5}$ value as illustrated in Figure 2.

Timestamp	Timestamp (Local)	PM 1.0	PM 2.5	Applied PM 2.5 Custom Calibration Setting - Multiplication Factor	Applied PM 2.5 Custom Calibration Setting - Offset	PM 4.0	PM 10
UTC	UTC+06:00	ug/m3	ug/m3			ug/m3	ug/m3
02/25/2024 00:00:34	02/25/2024 06:00:34	10	11	1.03		11	11

Figure 2. The obtained data example.

Molina Rueda et al., (2023) studied three low-cost sensors from a manufacturer other than TSI and concluded that the $PM_{2.5-10}$ size fraction exhibits significant uncertainty. One reason for such measurements is the inertial losses during the transmission and aspiration of particles from ambient air to sensing zones. This research only focuses on $PM_{2.5}$ and the others size fractions are not used as experimental measurements may have shown unclear results.

The building standards of Kazakhstan require proper ventilation of buildings. The type of ventilation system depends on the purpose of the building and its rooms. The code requirement for apartments where people live is natural ventilation. Natural ventilation is designed for kitchens and bathrooms, and it works efficiently when the air pressure inside the household is high, and typically achieved by opening windows. Apartments do not have air supply except through opening windows, while other building types, such as offices, may have forced air supply. This study started the data collection in a single room office where the window is always closed during the measurement. Then, we proceeded in the other households, where residents open the windows when they need to without limitations.

3.2 Analyses, Mathematical Model

Infiltration modelling requires understanding building and meteorological parameters. Argyropoulos et al., (2020) addressed a model namely; CONTAM, which is used to study building infiltration and indoor air quality. CONTAM is a widely used a multi-zone IAQ model that determines airflow, pressures, concentrations of contaminants, and personal exposure under different conditions. The wind pressure is an important factor for modeling the infiltration. It depends on wind direction and speed, urban airshed, terrain effects, and building heights. The CONTAM libraries can be used to obtain correlations of wind pressure profiles on the parameters listed above (Dols and Polidoro, 2015). Authors mention that these libraries express

the average wind pressure profiles for a simplified building of a cubic form. To calculate the concentrations of pollutants the CONTAM uses the following terms given in Table 3.

Term	**Input data**	**Data source**
Emission	Pollution generation rate	Input by user
Deposition	Pollution removal rate Zone concentrations	Input by user Output ny CONTAM
Interzone air mixing and resultant transfer	Air flow rate between zones Concentrations in other zones	Air exchange module Output ny CONTAM
Infiltration and ventilation transfer	Outdoor concentration Air flow due to infiltration and ventilation	Infiltration and ventilation module
Transfer due to HVAC	Air mass supply rate Concentration in HVAC	HVAC simulations (EnergyPlus) Output ny CONTAM

Table 3. CONTAM model terms and descriptions (Jose and Perez-Camanyo,2023).

The drawback is that CONTAM does not consider exact building geometry. This drawback led to the employment of another model based on more advanced modeling methods of computational fluid dynamics (CFD): QUIC. The purpose of this model is to compute the wind pressure coefficients for the path of airflow which represents the point (crack or opening) on the exterior walls of a building. The points' coordinates were provided by CONTAM via creating a Path Location Data file (PLD). The QUIC calculates pressure coefficient hourly at each inserted point and generates a Wind Pressure and Contaminant file (WPC). The described combination is applicable for large buildings to obtain accurate and realistic predictions of airflow particles movement.

The equation for the model, a transient partial differential equation, used to analyze airflow (Argyropoulos et al, 2020) can be written as:

$$\frac{\partial m_i}{\partial t} = \rho_i \frac{\partial V_i}{\partial t} + V_i \frac{\partial \rho_i}{\partial t} = \sum_j F_{ji} + F_i \qquad (2)$$

The infiltration is further modeled using Bernoulli's equation in the CONTAM model (Walton and Dols, 2005).

$$\Delta P = (P_1 + \frac{\rho V_1^2}{2}) - (P_2 + \frac{\rho V_2^2}{2}) + \rho g(z_1 - z_2) \quad (3)$$

Here P_1, V_1, z_1 stand for static pressure, speed and elevation of the inlet while P_2, V_2, z_2 represent the same variables for outlet. The main characteristics of the model: the architecture and geometrical measurements of the household environment can be obtained in two ways: through real measurements or technical drawings. PM is assumed not to affect the density of the air and is divided into three size fractions ($PM_{2.5\text{-}10}$, $PM_{1\text{-}2.5}$, PM_1), for example. It is necessary to adopt the model to exclude high-size fractions because, as mentioned before, PM_4 and PM_{10} measurements are not clear due to their larger sizes which do not go well with the infiltration mechanism. Jose and Perez-Camanyo (2023) used the EnergyPlus model, supported by the CONTAM for analysis of contaminant transport. Prior studies typically relied on outdoor weather and pollution data from nearby monitoring sites. However, this study introduces a novel approach by generating specific meteorological and air pollution data tailored to the location of the building using an atmospheric and chemical simulation model. This model provides boundary conditions for the building with hourly frequency, considering local microclimatic conditions. Furthermore, the outdoor model is integrated with an indoor air pollution and energy model, incorporating factors such as air fluxes, ventilation dynamics, decay processes, and indoor emission sources. By combining methods for modeling energy and airflow, the indoor model can simultaneously simulate the building's energy demand and airflow between zones. Unlike previous methods that relied on a linear relationship between outdoor and indoor concentrations, this study's tool enables direct simulation of indoor air quality, energy demand, and room temperature using a single integrated model. Additionally, obtaining outdoor data from another simulation ensures accuracy and prevents reliance on measurements from distant locations (Jose and Perez-Camanyo, 2023). The authors conclude that infiltration has a more pronounced effect on gases (i.e., NO_2) levels compared to particles (i.e., $PM_{2.5}$), except during the winter season when outdoor $PM_{2.5}$ concentrations surpass those of NO_2. Therefore, outdoor concentrations play a critical role in determining the influx of pollutants indoors.

The process of PM infiltration is a dynamic process, and the form of the mass balance equation was investigated by Morawska (2005) and Diapouli et al., (2013). The temporal concentration of PM depends on the particle decay curve or particle rebound curve. The equation below describes this process.

$$\frac{dC_{in(t)}}{dt} = a * P * C_{out(t)} - (a + k) * C_{in(t)} + \frac{Q_{is}}{V} \quad (4)$$

$C_{out}(t)$ and $C_{in}(t)$ represent the concentrations of particles outside and inside at time, measured in mg/m^3. a is a factor of multiplicity of air change in h^{-1}. The particle penetration coefficient - P is dimensionless, it represents penetration efficiency. A particle deposition rate is k in h^{-1}. Volume of the room is V in m^3. Q_{is} stands for a rate of particle generation by indoor sources in mg/h. The equation assumes perfect conditions for indoor air mixing. Particle mass gains and losses are ignored as well as temperature and relative humidity changes. The P and k coefficients are dependent on building characteristics, particle composition, electric charge and size. Estimating infiltration parameters such as P and k is challenging because of significant variability and interrelatedness. Different methodologies yield varying results, particularly for deposition rate, which shows wide ranges between studies and particle size fractions (Diapouli, 2013) . The P factor is defined by dividing the mass fraction of particles in the infiltrating air by total mass fraction.

$$P = \frac{N_{escape}}{N_{total}} \quad (5)$$

The P factor is useful for describing particle balance of the building by describing penetration mechanism through leaks and cracks. The I/O ratio is influenced by the sizes of the cracks in the building envelope (Chen et al., 2011). It is accurately calculated using dynamic models. However, the literature where deposition rate k is investigated, showing a wide range of results depending on size fraction. Researchers have computed P and k values independently, sometimes under controlled environmental settings, such as manipulating ventilation conditions and particle levels. For instance, in a study conducted in Hong Kong, penetration and deposition rates were

assessed in six homes by deliberately increasing indoor particle concentrations through the opening of windows and doors. Following this, the windows and doors were shut, and the decay of indoor particles was monitored (Bennett and Koutrakis, 2006). Thatcher et al., (2003) employed a dynamic model to ascertain P and k values in two experimental homes located in California. In these experiments, particle concentrations were uniformly elevated throughout the homes and subsequently allowed to decline to calculate k. It's noteworthy that they ensured well-mixed conditions before determining k, as reductions in particle concentration resulting from mixing throughout the home cannot be distinguished mathematically from reductions due to particle deposition. Additionally, the researchers performed an experiments to identify penetration efficiency P. The pressurized and filtered air was pumped into the indoor environment to reduce the indoor concentrations. Then indoor PM concentration increased naturally due to building's ventilation, which is given as F_{inf}.

Previous studies estimated the outdoor infiltration factor F_{inf} considering the whole monitoring period, encompassing both occupancy and non-occupancy time. This approach was necessitated by using time-integrated filter-based gravimetric methods that couldn't differentiate between these periods and as the actual occupancy time was not recorded constantly. However, including non-occupancy time in F_{inf}estimation may introduce bias due to differences in indoor pollutant concentrations between occupancy and non-occupancy periods. With advancements in sensor technology, presence of people at households can now be detected from continuous monitoring of indoor carbon dioxide concentrations. Such monitoring allows to assess the impact of residents on indoor environment, including non-occupancy or occupancy time in F_{inf} estimation and gain insights into potential biases (Hossain et al., 2021). F_{inf} depends on the building type. A study in Daqing city in China focused on identifying F_{inf} for different rooms and buildings: classroom, office, urban residential and rural residential (Lv et al., 2017). Indoor and outdoor particle concentrations' regression analysis showed they mostly have linear relationships. The calculated F_{inf} for urban residential and office were 0.7499 and 0.7214. In rural residential and classroom F_{inf} is 0.9019 and 0.9217. Authors explain it with higher air exchange in classroom and rural residential as windows in those buildings are usually open.

3.3 Occupancy time

Hossain et al. (2021) studied factors affecting infiltration variability of gaseous pollutants and ambient particles into households in urban environments of Hong Kong. In their study the occupancy time within the household is identified by the presence of at least one individual while air quality measurements are running. While occupancy times were recorded for 10 households, they were not documented for the remaining 45 due to participant burden. In these instances, presence within the home was inferred based on three criteria. Firstly, it was assumed that at least one person was present in the house between 21:00 and 8:00, considering that individuals in Hong Kong usually spend approximately 14.9 hours per day at home. Secondly, information provided in the daily questionnaire, such as changes in ventilation or air conditioning operations and indoor activities (such as preparing food, air humidifier usage, cleaning or smoking), were considered to deduce the presence of at least one occupant. For instance, if cooking activities were recorded between 13:00 and 14:00 on a particular day, it demonstrated the presence of at least one individual at home during that time frame. Lastly, the concentration of carbon dioxide (CO_2) within occupied homes is higher due to human metabolic processes. By analyzing backward differences in hourly average indoor CO_2 concentrations over two-hour intervals, estimated CO_2 differences falling within the range of - 60 ppm/h to +65 ppm/h were considered indicative of the departure or arrival of occupants between 8:00 and 21:00.

4. Experiment monitoring setup and procedure

The experimental monitoring setup and procedures are detailed in the first monitoring example as follow. The first monitoring experiment took place in Astana, at Tauelsizdik 3 building. The room measures 2.5x5 meters and has one window and one door. The experiment started on 02.26.2024 and ended on 03.12.2024. The room plan is provided in Figure 3. The plan was created in AutoCAD after measuring the room. The room is located on the fifth floor and serves as an office. The window faces North-West, while the door leads to the corridor with one neighboring room. The room is situated at the corner of the building, with two walls facing the

outside environment - one with the window and one blank. The architectural plan is shown in Figure 3, the single office room is small and usually occupied by three people. The room has stationary computers and no primary indoor pollution sources.

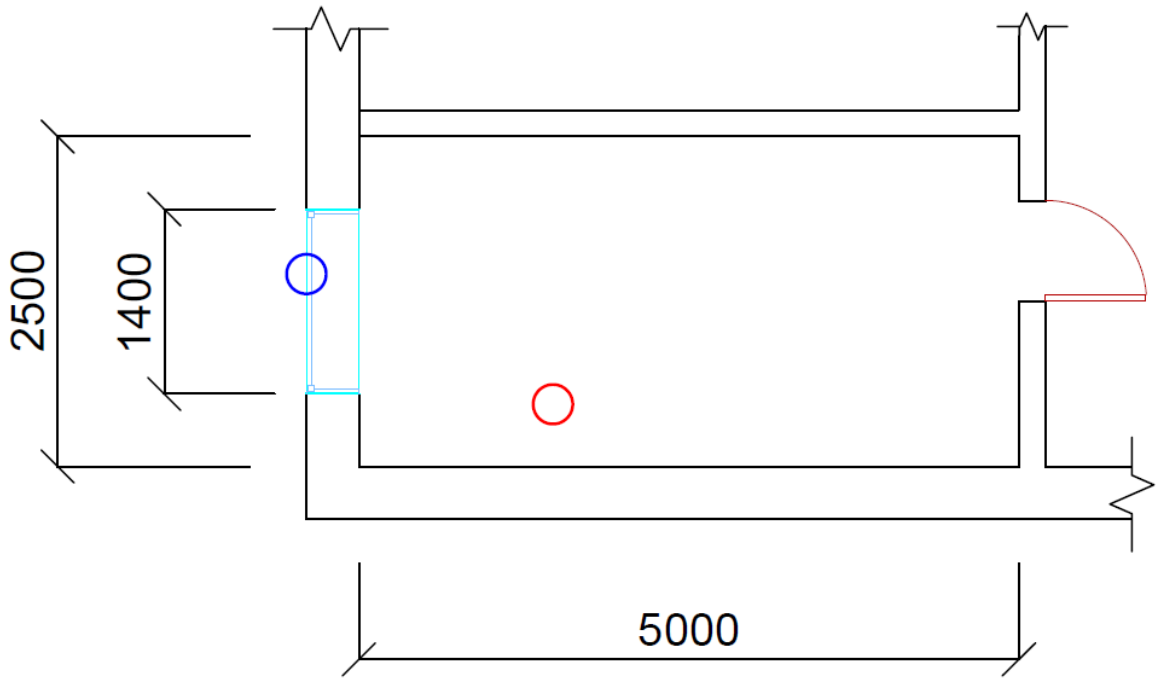

Figure 3. The Office floor plan (units are in mm)

Since there are no indoor pollution sources, infiltration of particles can be observed without any disturbance. The window measures 1.4 meters in length and 0.6 meters in height. A red circle depicts the AirAssure indoor air quality monitoring device (Figure 4), which is installed on the table at the shown location. The elevations of the outdoor and indoor devices are almost similar. A white color of the AirAssure means that the device is working and blue color means that the device is connected to the internet. It is sending the data to the TSI account.

Figure 4. AirAssure setup and active measuring moment.

A blue circle represents the Bluesky outdoor air quality monitoring device, which is installed on the outer side of the window. It has one led diode in the bottom which shows the status of the device. The device on Figure 5 had one blinking of the diode which means correct operation of the device and obtained data availability online on the TSI account.

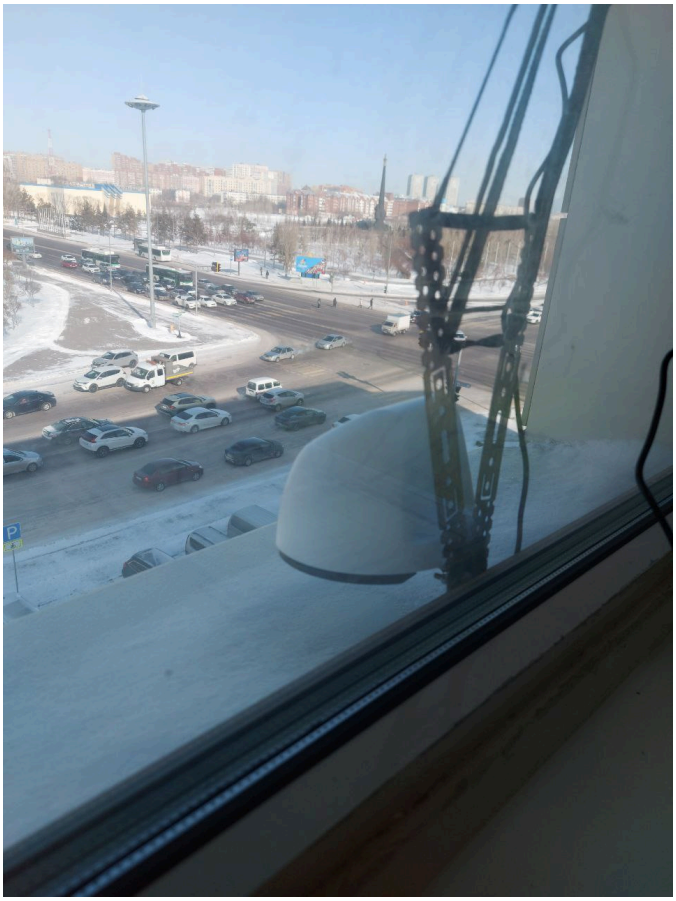

Figure 5. BlueSky setup and active measuring moment.

The monitoring devices were active throughout the experiment, recording results every minute. During the measurement, the window remained closed. Table 4 shows the daily average $PM_{2.5}$ concentrations from February 26 to March 11. Results for March 12 were omitted because the readings were not complete for a full 24 hours. Although the experiment began on the evening of February 26 and the reading for that day is not complete, it was included because it represents the starting day and helps observe how concentrations stabilize on subsequent days.

Day	**Outdoor $PM_{2.5}$ $\mu g/m^3$**	**Indoor $PM_{2.5}$ $\mu g/m^3$**	**I/O**
26.02.2024	27.06	6.20	0.23
27.02.2024	18.63	6.64	0.36

28.02.2024	29.37	12.46	0.42
29.02.2024	38.26	18.05	0.47
01.03.2024	21.75	9.33	0.43
02.03.2024	22.46	7.53	0.34
03.03.2024	9.98	4.90	0.49
04.03.2024	11.33	5.32	0.47
05.03.2024	11.58	6.28	0.54
06.03.2024	11.38	4.58	0.40
07.03.2024	16.50	6.36	0.39
08.03.2024	52.50	18.56	0.35
09.03.2024	30.90	12.47	0.40
10.03.2024	13.43	5.31	0.40
11.03.2024	11.79	5.64	0.48

Table 4. Daily averages and Indoor/Outdoor ratio

The first observation is that the outdoor concentration of $PM_{2.5}$ is higher than the indoor concentration. This holds true for every day, although in minute readings, the values for outdoor and indoor concentration rarely match. Secondly, the concentration values in the table are averages of minute readings throughout the day. As mentioned before, the I/O ratio necessary for analyzing infiltration. The I/O ratio for February 26 is the lowest at 0.23. Starting from the next day, the I/O ratio never dropped below 0.3. The highest I/O ratio recorded was 0.54 on March 5, possibly due to weather conditions. Generally, the I/O ratio fluctuated between 0.34 and 0.54. I/O ratios tend to overstate the proportion of outdoor air pollutants that enter indoor spaces, as noted by Bennett and Koutrakis (2006). A more accurate method involves estimating the infiltration factor - the fraction of outdoor particles that penetrate indoors and remain suspended. Thirdly, the correlation between outdoor and indoor $PM_{2.5}$ concentrations was calculated in Microsoft Excel, which showed a value of 0.93 for Table 4. This strong positive linear

correlation suggests that when outdoor PM concentration rises, indoor concentration also rises due to infiltration. It also confirms that the indoor environment has no pollution sources. Wichmann's study of indoor–outdoor relationships in Stockholm (Wichmann, 2010) showed that despite the similarity between indoor and outdoor $PM_{2.5}$ levels, indoor sources compensate for the lower infiltration of outdoor $PM_{2.5}$, resulting in comparable indoor levels.

Figure 6 illustrates I&O concentration for $PM_{2.5}$. The graph contains more than six thousand points representing minute measurements. The orange line represents outdoor concentration, which is usually higher than the indoor blue line. Occasionally, the blue line crosses the orange line to go higher. The infiltration process requires time (here after called as "lag" time), as evident from the graph. The lag time is clearly observed when outdoor concentration peaks while indoor concentration remains constant. The maximum indoor concentration is observed some time after the outdoor concentration peak. The pattern of these two graphs is similar, with clear upward and downward trends and some time lag. The increase in outdoor concentration is much higher than indoor concentration, with the maximum outdoor $PM_{2.5}$ concentration reaching almost 180 μg/m^3 while for indoor it is around 60 μg/m^3.

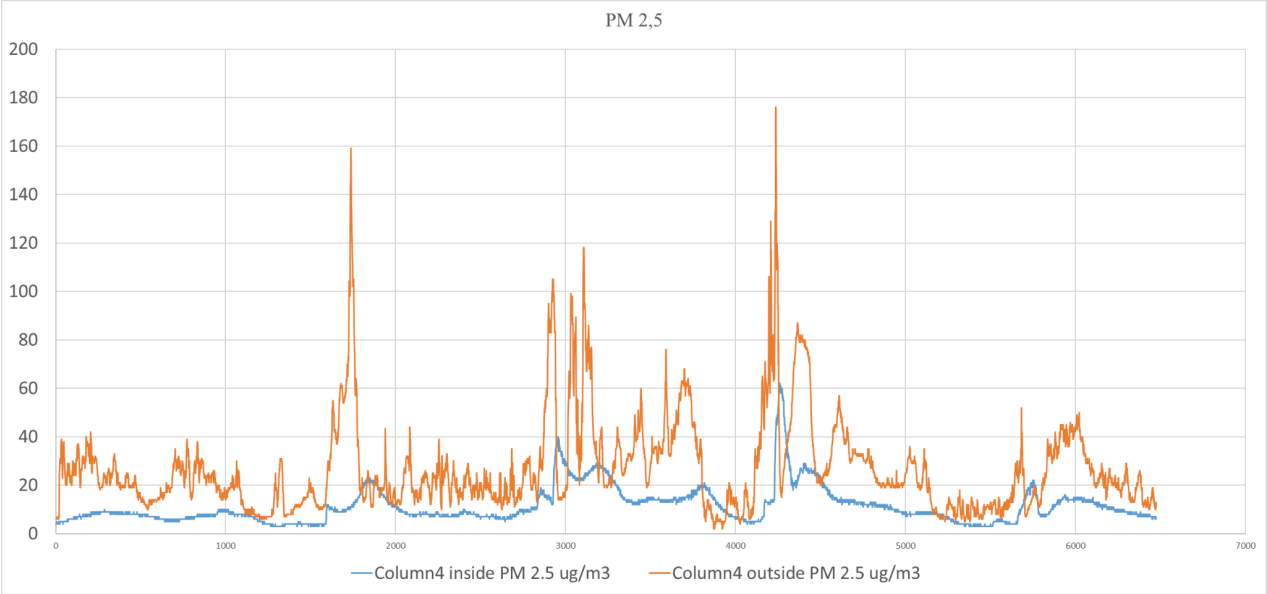

Figure 6. PM 2.5 minute readings for indoor(blue) and outdoor(orange) for one week.

The daily correlation coefficient between I&O $PM_{2.5}$ is 0.93. However, the minute-by-minute correlation showed a more accurate but lower value of 0.448 for PM_1 and 0.438 for $PM_{2.5}$. Minute readings provide the shortest available reading time. The concentration of outdoor PM is unstable due to weather conditions and human activities. Throughout the experiment, there were windy and rainy days, as well as road repair works and traffic jams, particularly during the evening hours. The indoor concentration is influenced by the number of occupants and the position of the door (open or closed). There is one heating radiator in the room, which was operational during the experiment in March, and there was no ventilation inside the room. Since minute readings fluctuate, average values were calculated. The 10-minute and 1-hour averages selected, and correlations are computed in Excel and presented in Table 5.

Correlation between I&O PM					
PM_1 1 minute	PM_1 10 min avg	PM_1 1 hour avg	$PM_{2.5}$ 1 minute	$PM_{2.5}$ 10 min avg	$PM_{2.5}$ 1 hour avg
0.448	0.458	0.534	0.438	0.447	0.523

Table 5. Correlations between I&O PM concentrations in different time scales.

From Table 5, it is evident that correlation increases with a higher average scale. The average value can be more useful as it exhibits fewer fluctuations. As indoor PM concentration is dependent on outdoor concentration, the indoor readings lag behind the outdoor readings. One option to determine the lag time is by adding some time and tracking the correlation value. For this building, the "lag" time for $PM_{2.5}$ indoor concentration is increasing by 1-minute each time. Resultant correlation between I&O $PM_{2.5}$ values is tabulated in Table 6.

Lag time	**Correlation**	**Lag time**	**Correlation**	**Lag time**	**Correlation**	**Lag time**	**Correlation**
1 min	0.44501	21 min	0.57209	41 min	0.66654	61 min	0.69981
2 min	0.45131	22 min	0.57823	42 min	0.66966	62 min	0.69999

3 min	0.45726	23 min	0.58429	43 min	0.67263	63 min	0.69991
4 min	0.46339	24 min	0.59050	44 min	0.67490	64 min	0.69987
5 min	0.46954	25 min	0.59635	45 min	0.67694	65 min	0.69958
6 min	0.47542	26 min	0.60214	46 min	0.67884	66 min	0.69925
7 min	0.48131	27 min	0.60782	47 min	0.68064	67 min	0.69904
8 min	0.48764	28 min	0.61300	48 min	0.68213	68 min	0.69872
9 min	0.49404	29 min	0.61812	49 min	0.68388	69 min	0.69825
10 min	0.50029	30 min	0.62298	50 min	0.68553	70 min	0.69766
11 min	0.50676	31 min	0.62768	51 min	0.68700	71 min	0.69731
12 min	0.51345	32 min	0.63232	52 min	0.68839	72 min	0.69711
13 min	0.52001	33 min	0.63687	53 min	0.69026	73 min	0.69686
14 min	0.52657	34 min	0.64124	54 min	0.69178	74 min	0.69625
15 min	0.53322	35 min	0.64553	55 min	0.69312	75 min	0.69571
16 min	0.53996	36 min	0.64944	56 min	0.69496		
17 min	0.54666	37 min	0.65321	57 min	0.69660		
18 min	0.55307	38 min	0.65674	58 min	0.69781		

19 min	0.55939	39 min	0.66017	59 min	0.69876
20 min	0.56577	40 min	0.66334	60 min	0.69948

Table 6. Correlations between I/O concentrations with different time lags.

The correlation increases when indoor concentration values are considered with a "lag" time. For example, the correlation value with indoor concentration 1 minute earlier is 0.445, while the correlation value without any adjustments is 0.438. The correlation between I/O $PM_{2.5}$ minute readings increases with increasing lag time and reaches its maximum value at 62 minutes. The possible conclusion is that it takes one hour for particles to infiltrate. It is necessary to compare this lag time with the lag time on the graph (the time between the peaks of I/O concentrations). After 62 minutes, the correlation value decreases, indicating that increasing the "lag" time does not yield useful results and may lead the correlation in the wrong direction. Krasnov et al. (2015) experiments showed variable lag time between I/O concentrations during the dust events. According to the intensity of the dust storm the time lag in their study varied from several minutes to one hour. The I/O ratio in their experiment was always less than one for all measured houses. During the low level storms the mentioned ratio was 0.79 while in case of the strongest storms it was 0.58. The daily $PM_{2.5}$ concentrations for each day are shown on Figures 7 and 8. Each line represents a single day on a 24 hour scale. Figure 7 is for outdoor concentration and Figure 8 is for indoor concentration. The 15 minute average is taken. The figures 7 and 8 are similar in terms of peaks. Generally, the outdoor daily concentration is less than 10, however some days show a sharp increase.

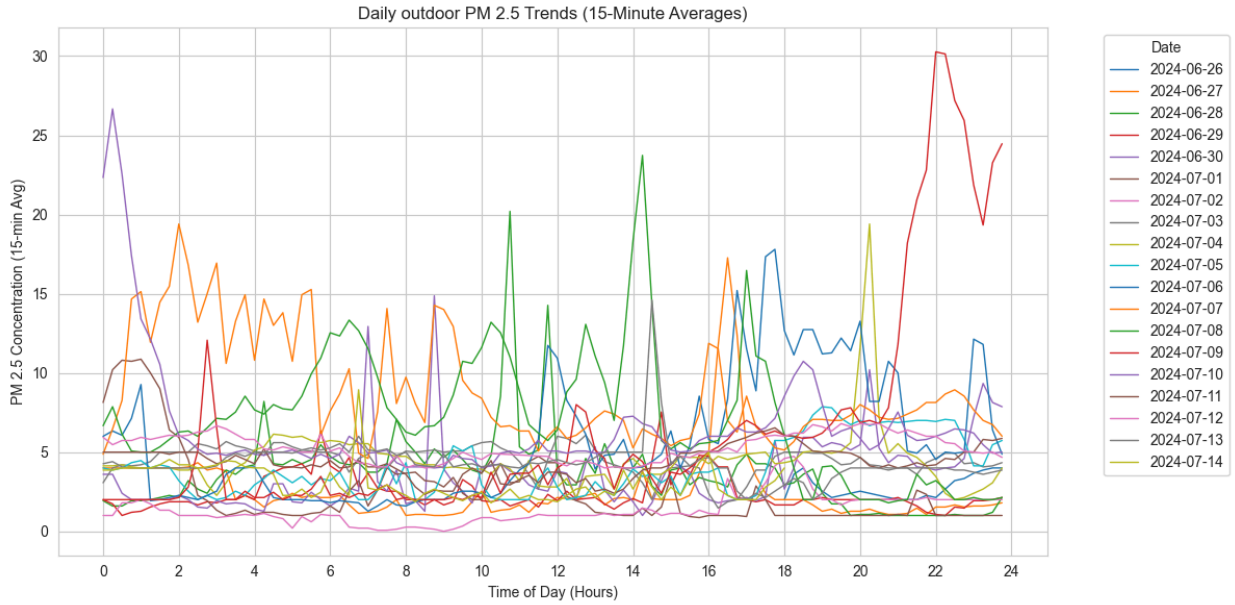

Figure 7. Daily $PM_{2.5}$ 15 minute average concentrations for outdoor environment.

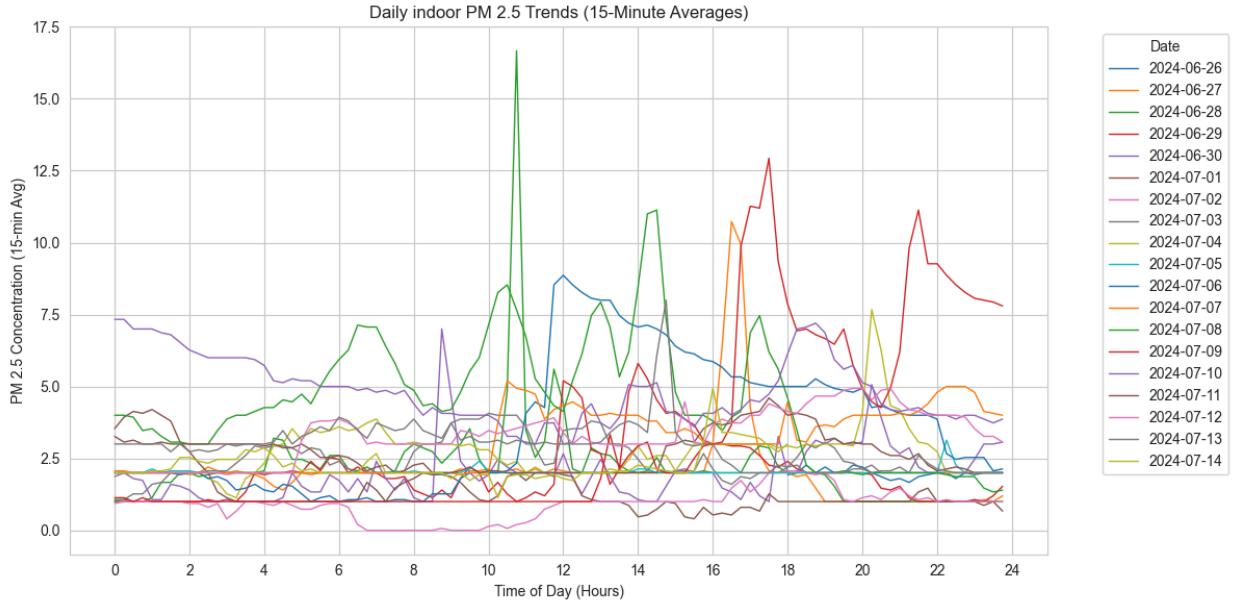

Figure 8. Daily $PM_{2.5}$ 15 minute average concentrations for indoor environment.

5. Data collection places

The data is collected in many households of Astana. Table 7 provides information about all the sampling households.

	Dates	ID	Address	Latitude	Longitude	Apartment type	Floor	Room volume, m^3
1	26.02.2024 11.03.2024	Office	Tauelsizdik Ave 3	51.150500	71.455824	Office room	5	44.10
2	05.05.2024 24.05.2024	E.B.	Qabanbay Batyr Ave 4/2	51.146698	71.426211	1 room apartment	8	47.46
3	28.05.2024 20.06.2024	A.B.	Ul. Shamshi Kaldayakova 33	50.966285	71.361427	Individual house	1	73.50
4	26.06.2024 14.07.2024	A.K.	Uly Dala Avenue 27	51.101260	71.390305	2 room apartment	3	68.06
5	25.07.2024 10.08.2024	D.K.	A-98 10/1	51.121734	71.496752	1 room apartment	5	46.33
6	18.08.2024 31.08.2024	L.A.	Turan Ave 59	51.101583	71.394379	3 room apartment	2	33.03
7	27.10.2024 08.11.2024	N.K	Anet Baba 13	51.139698	71.385229	1 room apartment	2	50.00
8	09.11.2024 23.11.2024	A.S	Kenzhebek Kumisbekov St 2	51.16263	71.401829	1 room apartment	6	54.00

Table 7. Locations of data collection.

The data collection campaigns are performed during an extended period of time, between November 2024 and February 2025. The heating period of 2024 started on September 24. As coal burn for heating needs is the main source of air pollution in Astana, the concentration of $PM_{2.5}$ is expected to increase during October and November. Figure 9 shows all 8 data collection places. One of them is an office room and one is a single storey individual building. The other 6 households are apartments, 4 of them are single room apartments, one has two rooms and another has three rooms. The red circle represents the position of an indoor monitoring device and the blue circle is an outdoor monitoring device.

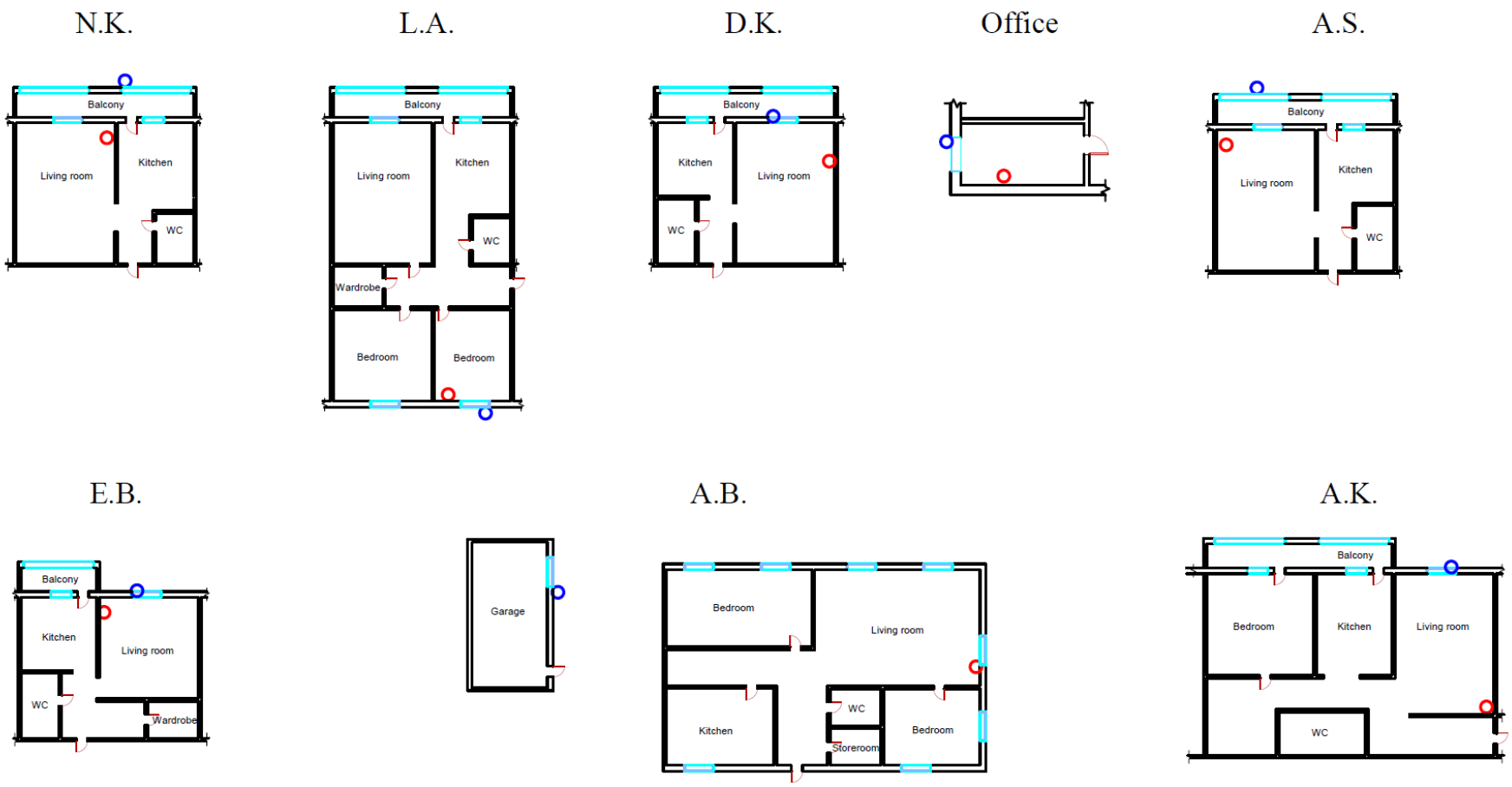

Figure 9. Floor plans of households.

The measurements in N.K. household cover the end of October and beginning of November. The people who live there smoke cigarettes on the balcony of their home. Thus, the indoor $PM_{2.5}$ concentrations were observed to be higher than the outdoor concentation. Considering the negative effects of indoor smoking, Table 8 depicts average I&O concentrations.

Date	Indoor $PM_{2.5}$ concentration	Outdoor $PM_{2.5}$ concentration
28.10.2024	5.77	5.81
29.10.2024	7.62	7.45
30.10.2024	4.50	5.61
31.10.2024	2.36	5.00
01.11.2024	5.47	4.85
02.11.2024	2.51	3.45
03.11.2024	5.60	7.17
04.11.2024	5.10	7.56
05.11.2024	4.34	8.57
06.11.2024	2.14	2.18
07.11.2024	0.98	1.78

Table 8. $PM_{2.5}$ concentrations for N.K. household where people smoke.

6. Data preparation

Removal of first and last hours of the data. During the device's transportation, some amount of dust and other particles accumulate on it. The devices start recording data immediately after turning on. However, it may have a different temperature from the environment it is meant to read, because: the previous household had a different environment, dust accumulation and temperature change during the transportation. In addition, The outdoor BlueSky device is turned

on inside the household before being hung outside. Considering this active environment changes, it was decided to remove at least the first 5 hours of the measurements. It is necessary because after devices get used to a new environment, they read more reliable data. The last hour of data is also removed because the outside device is taken inside before turning it off.

Data alignment. The data collection is performed using two devices and each of them creates a csv file containing the data it collected. Both devices have one minute average data collection settings. The installation of the devices is not simultaneous, it is easy to install the indoor device and start its operation, however, to install the outdoor device outside the apartment takes time. The outdoor device hangs out of the window, as it was shown on the photos, and attached by a rope. Thus, consequent setup procedure makes it impossible to start simultaneous data collection and one of the devices starts operating several minutes later. The original data is shown in Figure 10.

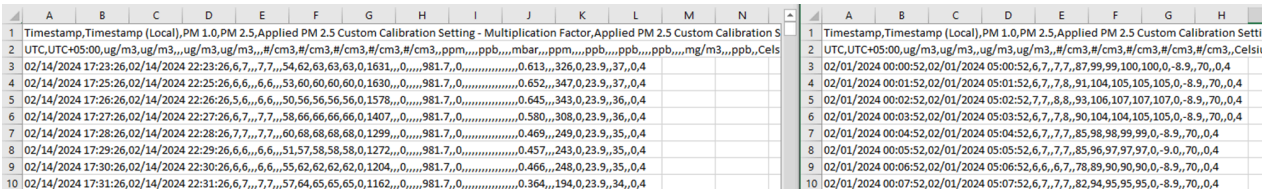

Figure 10. Original CSV files example.

In addition, sometimes devices read twice during a one minute interval. For example, one device may read at 17:00:00 and 17:00:59 while the second device has data for 17:00:20 (hour, minute, second format). Considering the two week duration of data collection these deficiencies result in unequal amounts of data. If the first CSV file for an indoor device has 26 thousand records(rows) the outdoor CSV file may have 21 thousand. Python is used to make perfect minute to minute data alignment, while removing extra rows. In the result, the number of rows is equal. The modified data is shown in Figure 11.

	B	D	Q	Y	AW	AY	BD	BF	BR	BT
1	Timestamp (Local)_A8	PM 2.5_A8	CO2	Barometric Pressure	Temperature_A8	Relative Humidity_A8	Timestamp (Local)_B8	PM 2.5_B8	Temperature_B8	Relative Humidity_B8
2	2024-02-27 00:00:31	8	587	996,3	20,4	31	2024-02-27 00:00:42	25	-5,2	65
3	2024-02-27 00:01:31	9	585	996,3	20,4	31	2024-02-27 00:01:42	30	-5,1	64
4	2024-02-27 00:02:31	8	585	996,3	20,4	31	2024-02-27 00:02:42	31	-5,1	65
5	2024-02-27 00:03:31	8	582	996,3	20,4	31	2024-02-27 00:03:42	31	-5,1	64
6	2024-02-27 00:04:31	8	579	996,3	20,4	31	2024-02-27 00:04:42	32	-5,1	64
7	2024-02-27 00:05:31	8	574	996,3	20,4	31	2024-02-27 00:05:42	31	-5,1	64
8	2024-02-27 00:06:31	8	572	996,3	20,4	31	2024-02-27 00:06:42	30	-5,1	64
9	2024-02-27 00:07:31	8	570	996,3	20,4	31	2024-02-27 00:07:42	31	-5,1	64
10	2024-02-27 00:08:31	8	569	996,3	20,4	31	2024-02-27 00:08:42	31	-5,1	64
11	2024-02-27 00:09:31	8	567	996,3	20,4	31	2024-02-27 00:09:42	31	-5,1	64
12	2024-02-27 00:10:31	9	563	996,3	20,4	31	2024-02-27 00:10:42	30	-5,1	64

Figure 11. Single file example containing aligned data from both devices.

In the resultant file, readings of devices are aligned minute to minute, different reading seconds does not have influence on further calculations. The Python code used for this operation is provided in the Appendix.

Multiple linear regression model.

The multiple regression is performed to predict the indoor $PM_{2.5}$ concentration based on several parameters. The data collected includes both I&O $PM_{2.5}$ concentrations. The indoor concentration serves as the Y parameter to be predicted, while the input parameters are in Table 9.

X1	PM 2.5 concentration outside	ug/m^3
X2	Temperature difference = temperate inside - temperature outside	C^o
X3	Barometric Pressure	$mbar$
X4	Relative humidity inside	%
X5	Relative humidity outside	%
Y	PM 2.5 concentration inside	ug/m^3

Table 9. Parameters for multiple regression with 5 variables.

The first multiple regression analysis is performed for the office room where data is shifted based on infiltration "lag" time of 62 minutes. The calculated coefficients are shown in Table 10.

	Coefficients
Y-intersection	-221.60
X1_PM 2.5_B8	0.235
X2_Tin-Tout	1.13
X3_Barometric Pressure	0.21
X4_Relative Humidity_A8	-0.08
X5_Relative Humidity_B8	-0.18

Table 10. Coefficients for multiple regression for Office with 5 variables.

The indoor $PM_{2.5}$ concentration now can be calculated using the coefficients from Table 10. The Figure 12 plots measured and predicted $PM_{2.5}$ concentration. The graphs are similar and there is positive correlation.

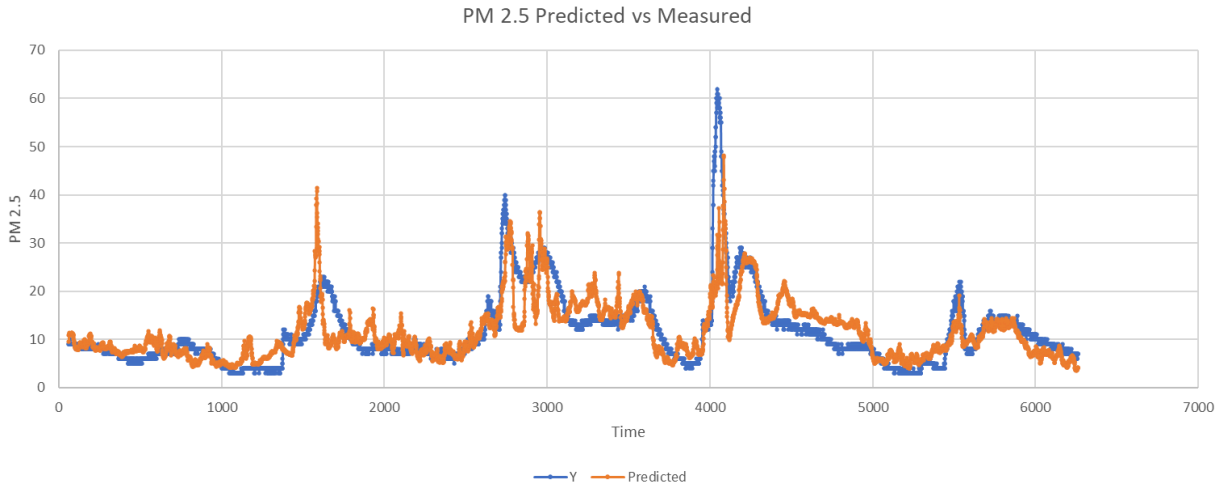

Figure 12. The Office indoor $PM_{2.5}$ measured and predicted concentrations.

The results of multiple regression are provided in Table 11.

Regression Statistics	
Multiple R	0.765425787
R Square	0.585876635
Adjusted R Square	0.585475483
Standard Error	4.901316028
Observations	6201

Table 11. Results for the Office multiple regression with 5 variables.

The second household is - E.B. household. The multiple linear regression analysis is performed similar to the Office. The calculated coefficients and regression results are given in Table 12 and Table 13.

	Coefficients
Y-intersection	-24.32389019
X1_PM 2.5_B8	0.196863386
X2_Tin-Tout	0.047551686
X3_Barometric Pressure	0.024421166
X4_Relative Humidity_A8	0.057272475
X5_Relative Humidity_B8	-0.028581583

Table 12. Coefficients for multiple regression for E.B. household with 5 variables.

Regression Statistics	
Multiple R	0.719256085
R Square	0.517329315
Adjusted R Square	0.517237255
Standard Error	0.997564757

Observations	26221

Table 13. Results for E.B. household multiple regression with 5 variables.

The regression statistics show similar values for multiple R, R Square and Adjusted R Square. The multiple R for the office is 0.77 and for the E.B. household is 0.72. The calculated coefficients are different. To evaluate the coefficients, Figure 13 is plotted.

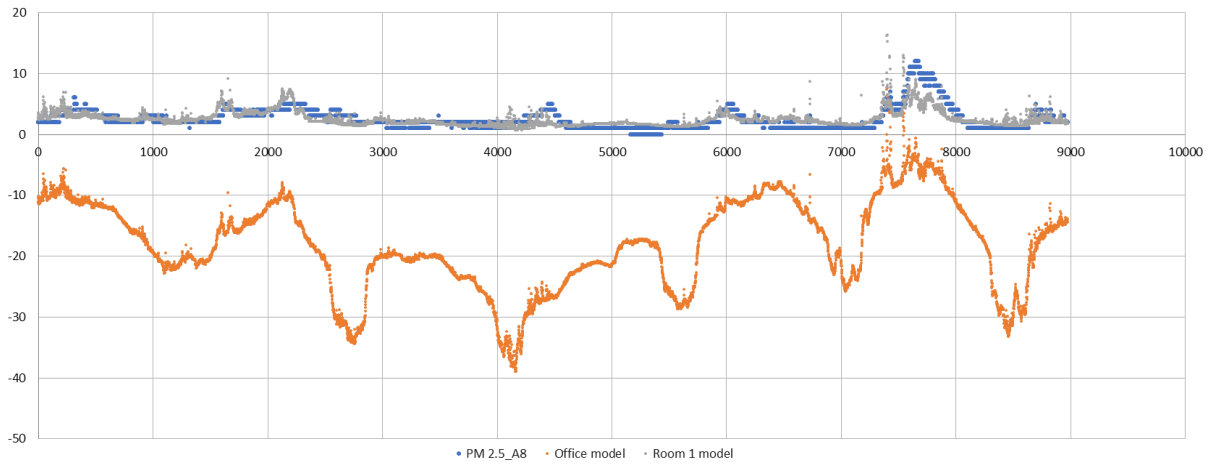

Figure 13. The E.B. household indoor $PM_{2.5}$ measured and predicted concentrations.

The blue color indicates the measured indoor $PM_{2.5}$ concentration in the E.B. household. The gray color is predicted concentration using coefficients of E.B. household and orange is using coefficients of Office. Here it is checked if results of one household can be used in the prediction of $PM_{2.5}$ concentration in another household. Figure 13 showsd that the Office coefficients fail with E.B. household data. However, the orange points correctly show directions of concentration, clearly showing if indoor $PM_{2.5}$ concentration is increasing or decreasing.

About air exchange rate.

The next step is to calculate Air Exchange Rate (AER). AER is important for the estimation of $PM_{2.5}$ infiltration. It describes what amount of air in the room, relative to its volume, is replaced

within some selected time. The AER is calculated using the collected data. The focus is on CO_2 concentration. To calculate the AER, the volume of the room is required. The volume of the whole apartment may be necessary for proper AER calculation, however, considering that only one indoor measurement device is available and CO_2 concentration is nonuniform around the apartment, it was decided to take the room volume. The volume is measured manually with a meter. The CO_2 concentration for E.B. household is shown in Figure 14.

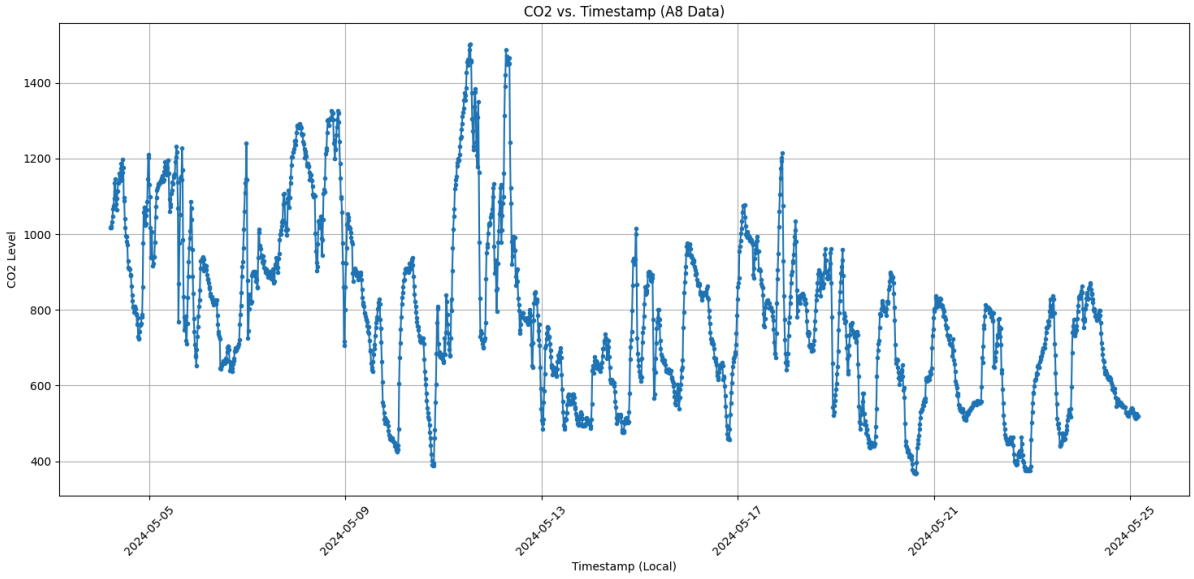

Figure 14. The E.B. household CO_2 concentration.

The CO_2 is measured in ppm and from the figure above it is clear that CO_2 fluctuations represent change in the environment. When people are inside the house they breathe and create CO_2 at some rate depending on their activity. The decay of CO_2 within a room means:

1. Someone opened the window.
2. Windows are closed, but CO_2 is moving to other rooms.
3. People who were the source of CO_2 left the room.

In all three cases, there is air exchange within the room which needs to be calculated. It is included in multiple regression as the 6th input variable. There are 3 cases for AER calculation using indoor CO_2 concentration:

1. Decay
2. Constant
3. Build up

The CO_2 decay was discussed above. The CO_2 build up case means there are people in the room who are increasing CO_2 concentration. Mostly, Astana households have electric stoves which means cooking is not a direct source of CO_2. Thus, people are considered as the main source of CO_2. The amount of CO_2 a person can produce is an individual value and depends firstly on physical activity, secondly on a person's biology (age, weight, gender). For this thesis, it was taken that 1 person produces 20.6 L of CO_2 in an hour during "light office work" conditions. The AER for "constant" and "build up" cases with 1 person in the room is calculated as follows:

$$\frac{(C_1+0{,}343/V*10^6-C_2)}{C_1} \tag{6}$$

Where:

C_1 is CO_2 concentration at time t_1 in ppm.

C_2 is CO_2 concentration at time t_2 in ppm.

V is a volume of the room in liters.

0.343 is CO_2 production rate in liters/min.

It can be used for the "constant" case, because the decay of CO_2 is always present. It means that, if CO_2 concentration is constant, the CO_2 production and decay rates are equal. This formula does not work if CO_2 concentration increases by 8 ppm and more because the value becomes negative in that case. Thus, if there are sharp CO_2 concentration increases the formula is modified by increasing the number of people, by multiplying 0.343 L/min by the number of people. The AER for "decay" case is calculated as follows:

$$\frac{(C_1-C_2)}{C_1} \tag{7}$$

The derivation of the formulas is based on the air volume change in the room. If CO_2 concentration decreases, it means some amount of air, which has CO_2 inside, left the room. Considering the air volume changes in the room with measured volume, the formulas are derived. The AER values for A.S. and L.A. are shown in Figure 15. The derivation is shown in the appendix.

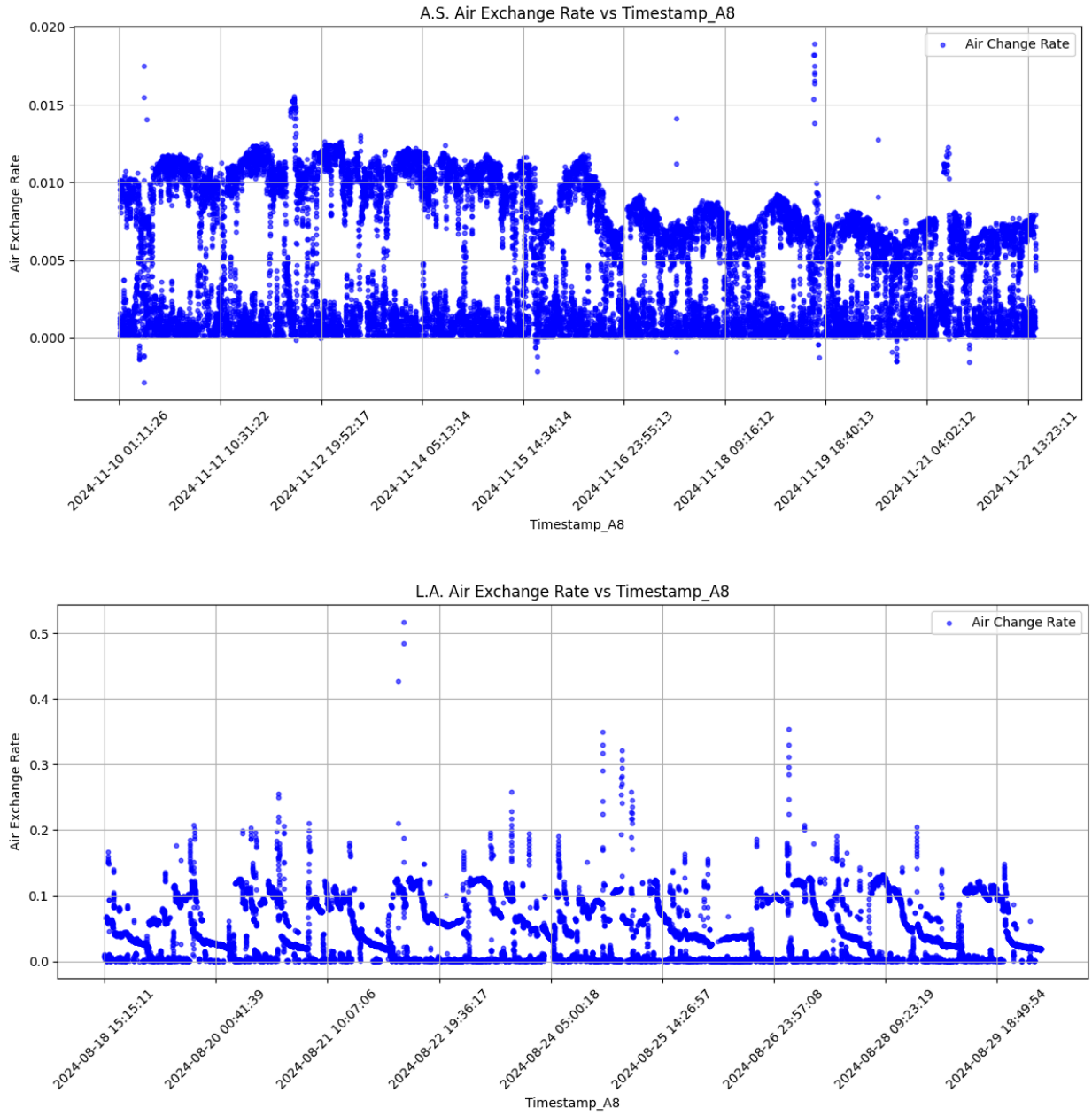

Figure 15. AER for A.S. and L.A. households.

Figure 15 shows that combined formulas for AER calculation work well, showing mostly positive values. The values are reasonable because it is AER per minute. However, there are some points going negative which should be checked and may be removed.

Analyzing E.B. and D.K. households

The multiple regression for the Office and E.B. household work well individually, but failed when predicting indoor $PM_{2.5}$ concentration for E.B. household using the coefficients of the Office (see the details in Figure 12 and Figure 13). It may be because these households are different and in the Office the window was always closed which allowed us to calculate the "lag time". Then, it was decided to analyze households that are similar. The E.B. household and D.K. household are both 1 room apartments where usually one person lives.

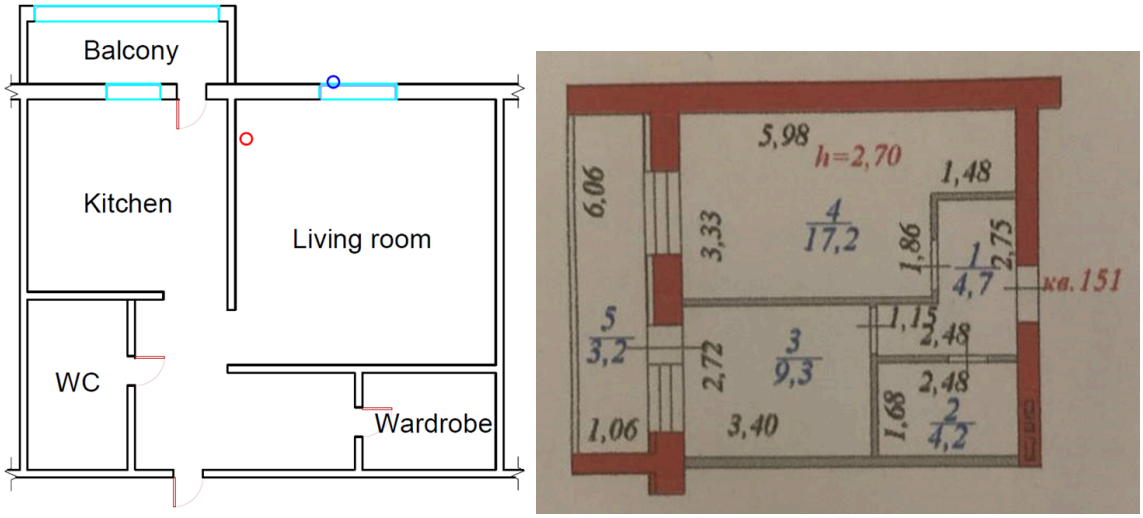

Figure 16. The E.B. household (on the left) and D.K. household (on the right).

Figure 17. The BlueSky installation in E.B. household (left) and D.K. household (right).

These 2 households are analyzed using multiple regression with five input variables as before. The resultant coefficients are given in Table 14.

	E.B.	**D.K.**
Y-intersection	-24.32389019	-17.17358934
X1	0.196863386	0.482751294
X2	0.047551686	-0.06240005
X3	0.024421166	0.019827541
X4	0.057272475	-0.033279869
X5	-0.028581583	0.019016527

Table 14. Coefficients for E.B. household and D.K. household.

E.B. and D.K. households have similar coefficients. To see the performance of 2 models the Figure 18 is plotted.

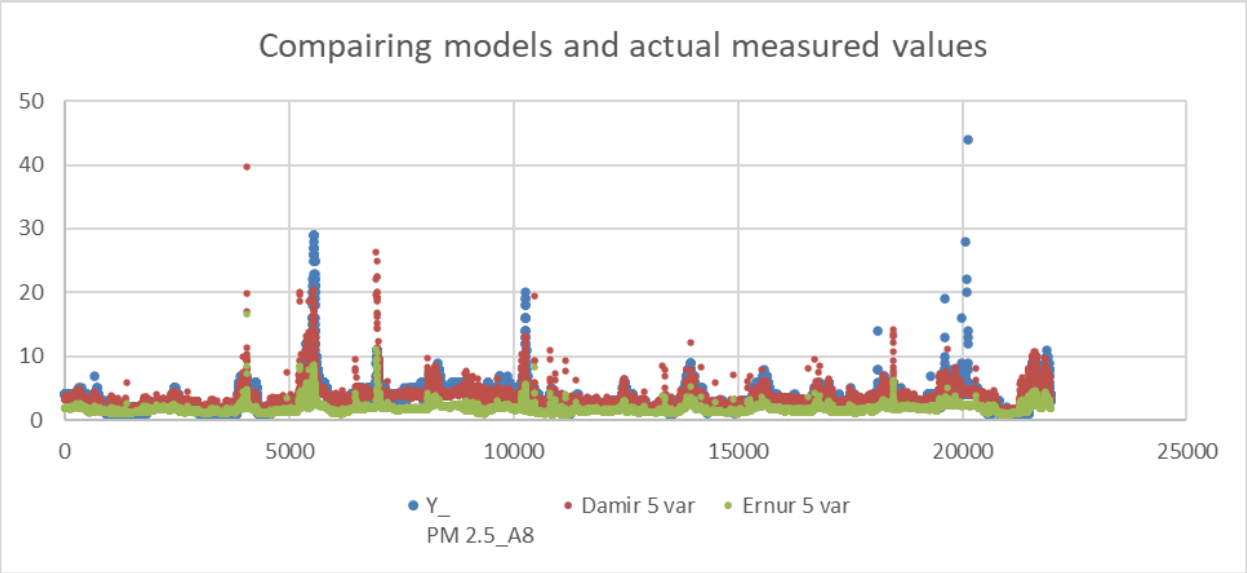

Figure 18. The D.K. household indoor $PM_{2.5}$ measured and predicted concentrations.

The blue color is measured indoor $PM_{2.5}$ concentration for the D.K. household and the red color is predicted using coefficients for the D.K. household. The green color is predicted using coefficients of E.B. household. The red is better in predicting because its coefficients are from this D.K. household data. But what is more important is that coefficients from another E.B. household performed well. The error was calculated using the formula:

$$\frac{(PM\ 2.5_{predicted} - PM\ 2.5_{measured})}{PM\ 2.5_{measured}} * 100\% \quad (8)$$

After calculating the error for each minute, the average error for the D.K. household is found to be 11.89 % for predicting using coefficients of D.K. household and - 37.66 % with coefficients of the E.B. household. The correlation Table 15 below shows that values predicted with D.K. household coefficients are in better correlation with actual measured value. In general, the correlation between two predictions is high (0.96).

	Y_$PM_{2.5}$ indoor	**D.K. model**	**E.B. model**
Y_$PM_{2.5}$ indoor	1		
D.K. model	0.798276515	1	

E.B. model	0.768830283	0.963112742	1

Table 15. Correlation table for measured $PM_{2.5}$ concentration, E.B. and D.K. households.

The next step is to repeat the procedure with AER to check if including the AER leads to better performance. The input variables and predicted variable are shown in Table 16. The calculated coefficients for 2 households are given in Table 17 and regression results are in Table 18.

X1	PM 2.5 concentration outside	ug/m^3
X2	Temperature difference = temperate inside - temperature outside	C^o
X3	Barometric Pressure	$mbar$
X4	Relative humidity inside	%
X5	Relative humidity outside	%
X6	Air Exchange Rate (AER)	-
Y	PM 2.5 concentration inside	ug/m^3

Table 16. Variables in multiple linear regression.

	E.B.	**D.K.**
Y-intersection	-23.35553178	-17.15439133
X1	0.196479872	0.482709279
X2	0.047109022	-0.06230161
X3	0.023536994	0.019794629
X4	0.055551217	-0.,033120357
X5	-0.028237254	0.018923098
X6	-5.623222411	0.88887457

Table 17. Coefficients for E.B. and D.K. households with AER.

Regression Statistics	Without AER	With AER
Multiple R	0.798276515	0.798286844
R Square	0.637245394	0.637261886
Adjusted R Square	0.637162856	0.63716284
Standard Error	1.19546515	1.195465176
Observations	21981	21981

Table 18. D.K. household multiple regression results with and without AER.

Table 18 shows that including AER does not have much effect on the results. The model performed better but the improvement is insignificant. It may mean that AER calculations are not precise enough. The average error is 11.89 % for predicting using coefficients of D.K. household and -38.23 % with coefficients of the E.B. household. Including AER, weakened the performance of E.B. household model. However it is not clear yet if it is right to take the AER coefficient of E.B. household. Because it may be that AER should be taken for original household data.

7. Improved methodology

1. Data alignment is required for the convenience of further work. The data is collected in two files and have different time intervals. The procedure creates one single file containing all the data and ensures they correspond to one time, minute to minute data alignment.
2. Data filtering. The “smoking filter” applied to the data removes unnecessary peaks in the indoor $PM_{2.5}$. It is necessary to analyse only infiltration related data, when indoor $PM_{2.5}$ concentration is influenced only by outdoor concentration. The indoor pollution source related data is removed.
3. Applying moving averages(MA). The data is collected for each minute and observation shows sudden peaks. The MA application can provide clear trends in $PM_{2.5}$ concentration.
4. The lag analysis. The outdoor $PM_{2.5}$ requires time to infiltrate indoor and observation shows that outdoor peak appears before the indoor. Identifying time required for outdoor $PM_{2.5}$ to infiltrate inside or “lag” time. It is necessary to understand till what time forward, we can predict the indoor $PM_{2.5}$ concentration.
5. The data normalisation is required to assess the influence of independent variables on indoor concentration. Some variables have great influence on $PM_{2.5}$ concentration while others may be insignificant.
6. Multiple linear regression is used for analysing the households. The indoor $PM_{2.5}$ concentration is a predicted variable while there are several independent variables.

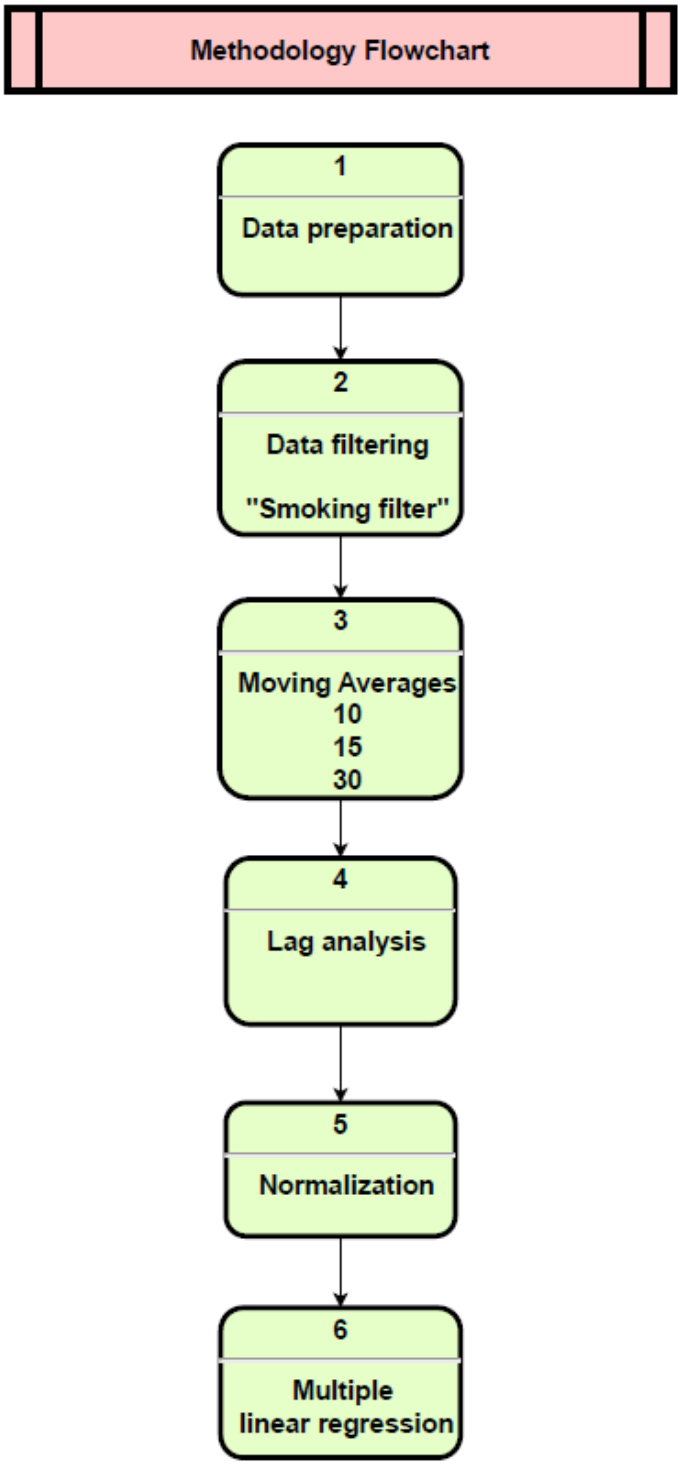

Figure 19. The methodology flowchart.

7.1 About Moving averages.

The Moving average (MA) works as a smoother and allows to eliminate sudden peaks in measurements. The MA is showing better results compared to just average values shown before in experiment 1. MA 10, 20 and 30 min are used to see the effect of different MA time intervals and compare them. The higher interval of MA can increase the Multiple R and identify general trends in data, but it reduces the precision. The MA effect on the accuracy of the model is given in Table 19.

Condition	**D.K. model Average error, %**	**E.B. model Average error, %**
No MA	11.89	-38.23
10 min MA	7.44	-38.51
20 min MA	6.11	-38.06
30 min MA	5.34	-37.65

Table 19. Results for D.K. household with different MA.

The MA use has a positive effect on Multiple R. The D.K. model prediction accuracy increases with increasing MA interval (see the details in Table 20). However, MA has a small effect on E.B. household model performance.

Regression Statistics	**Without MA**	**10 min MA**	**20 min MA**	**30 min MA**
Multiple R	0.798286	0.875825	0.899109	0.912154
R Square	0.637261	0.767069	0.808396	0.832025
Adjusted R Square	0.63716	0.767006	0.808344	0.831979
Standard Error	1.195465	0.927703	0.831223	0.769121
Observations	21981	21980	21962	21952

Table 20. The D.K. household multiple regression results with different MA.

The effect of MA is visually shown in Figures 20 and 21. The 10 minute MA is applied to both I&O $PM_{2.5}$.

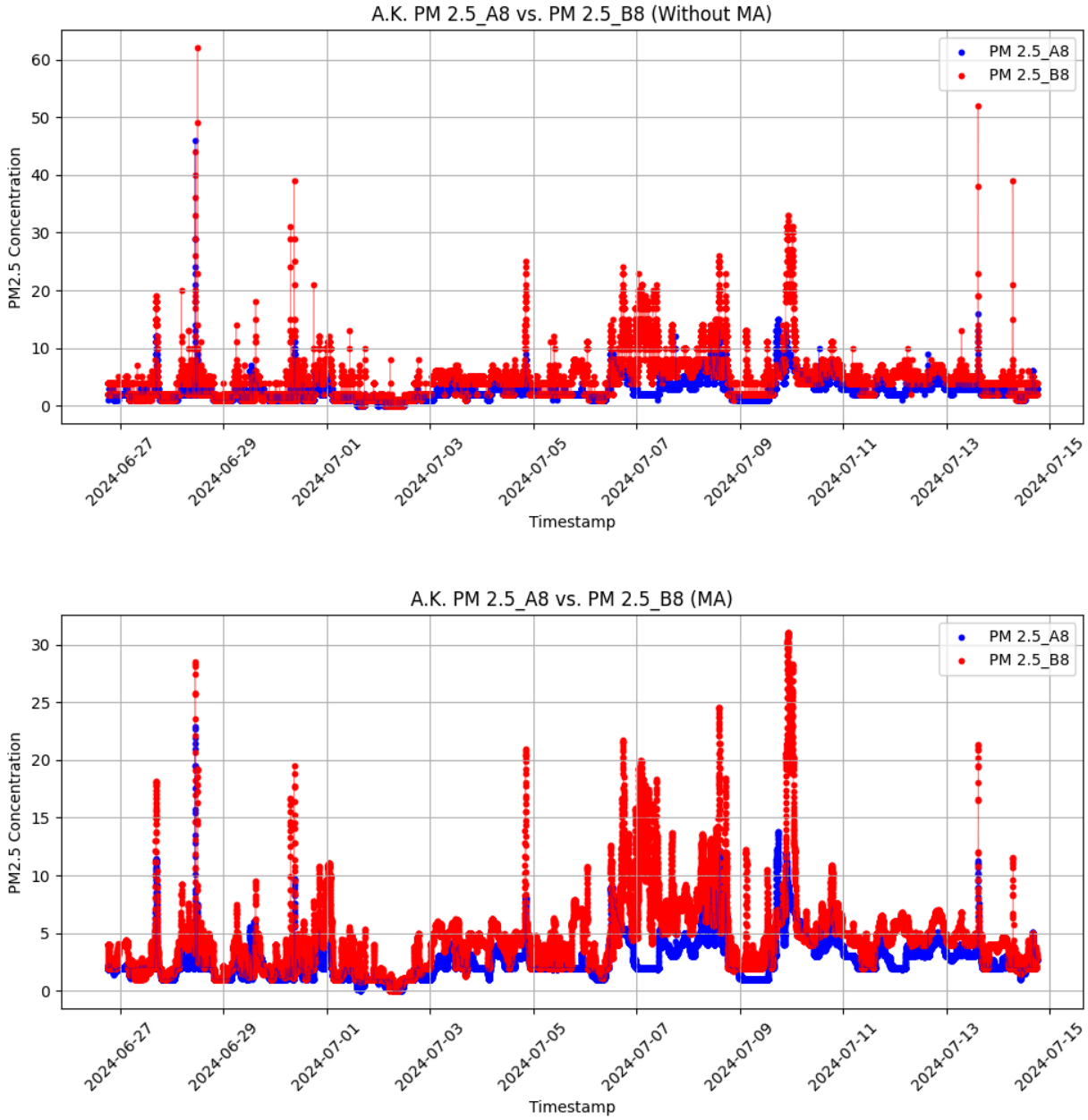

Figure 20. The A.K. household I&O $PM_{2.5}$ concentration before and after MA.

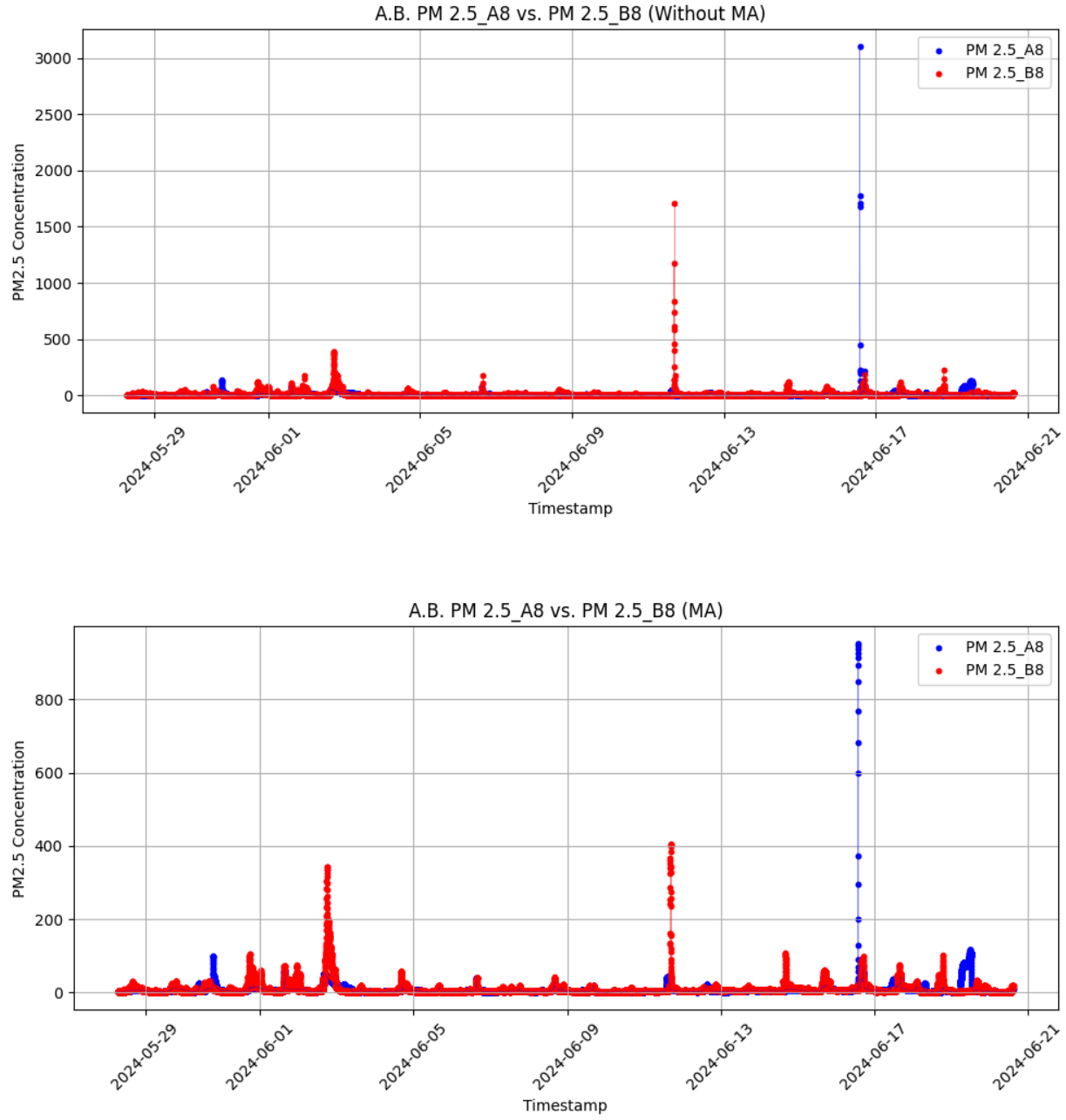

Figure 21. The A.B. household I&O $PM_{2.5}$ concentration before and after MA.

Applying MA removes outliers in the data. In the A.K. household, multiple peaks seen in outdoor concentration in the beginning and in the end of the data collection period (see Figure 20). The peaks like that with short duration(couple minutes) decrease the predictability of indoor concentration. The MA applied to this data removes part of the peaks while others are smoothed.

The A.B. household is highly affected by unknown source of pollution, thus both I&O $PM_{2.5}$ concentrations go to unrealistic values of 1700 and 3000 μg/m^3 (see Figure 21). Both devices may experience some short time influence from human activities which result in such measurements. For example, smoking and vacuuming inside the household, or burning and cooking something outside the house. The A.B. household is an individual building with a yard. The MA decreased the peaks and the new values are 400 and 900 μg/m^3 which is still high.

7.2 Smoking filter

The residents of the households which are considered in this research have usual habits which increase indoor $PM_{2.5}$ concentration. Cooking, indoor smoking and vacuuming are the main polluting activities. The plot of I&O $PM_{2.5}$ concentrations can provide some understanding of indoor pollution sources. For example, Figure 22 shows the peaks in the concentrations during the monitoring period.

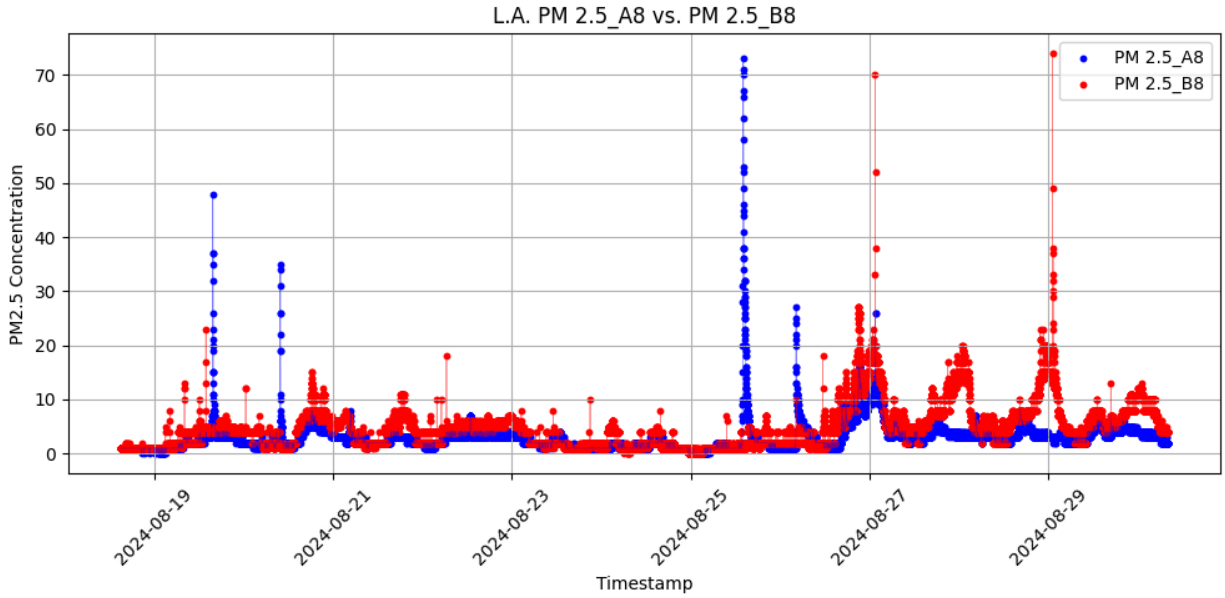

Figure 22. The L.A. household I&O $PM_{2.5}$ concentrations.

The outdoor concentration(red) is ideally the only pollution source since we investigate infiltration. Following this logic, the indoor concentration cannot be higher than outdoor concentration, some time before. The indoor concentration(blue) peak should appear after outdoor concentration(red) peak. However, the figure above shows sudden increases in indoor concentration which appear due to indoor pollution sources and can be smoothed with a "smoking filter".

The "smoking filter" removes sudden peaks in indoor concentration. There are two criterias for detecting the peaks: The subsequent increase by some value two times and single increase by some big value. For example, in the L.A. household the subsequent increase by 4 $\mu g/m^3$ two

times and single increase by 10 $\mu g/m^3$ identifies the unwanted indoor peaks. The peak identification for both criterias is shown in Figure 23.

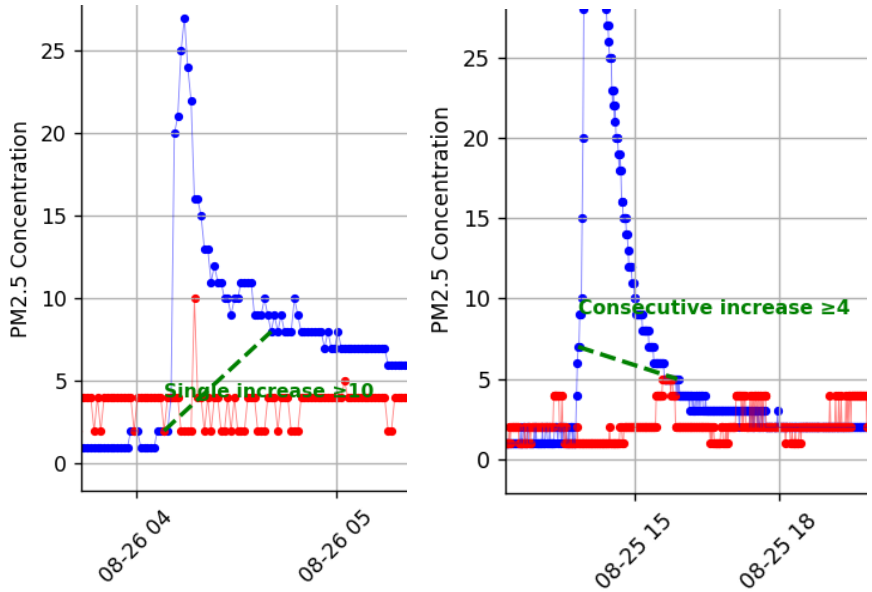

Figure 23. The L.A. household indoor PM 2.5 filtering.

After peak identification, in the first criteria, the line connecting the indoor concentration value some minutes before the peak detection and some time after peak detection is drawn. For example in the L.A. household, in figure 23, the line connects concentrations values 3 minutes before peak detection and length of this line is 30 minutes. In the second criteria, a line starts 2 minutes before the detected peak and its length is 2 hours. The values used in filtering the L.A. household are identified by observing the data and adjusting its parameters. The indoor peaks are of two types: Sudden one minute peak or sharp peak with constant gain. In both cases, the peaks have no outdoor $PM_{2.5}$ concentration peak. The one minute peak is dissolved in 30 minutes for this household which is observed using the data. The peak with constant gain dissolving time is longer and it is 2 hours in this household. The indoor peaks are not the same since indoor activities vary in duration and other factors. The filtering criterias and time durations do not ideally smooth all peaks, however they detect all unwanted peaks and modify the data.

The new data points are created along the green lines. They represent a filtered indoor $PM_{2.5}$ concentration. The peaks are removed and a modified data plot is shown in Figure 24. The green

data points are filtered values which follow the original indoor concentration but do not include unwanted indoor source related peaks. The other households are also filtered and their filtering criterias are similar but values are selected differently to suit the pattern of each household.

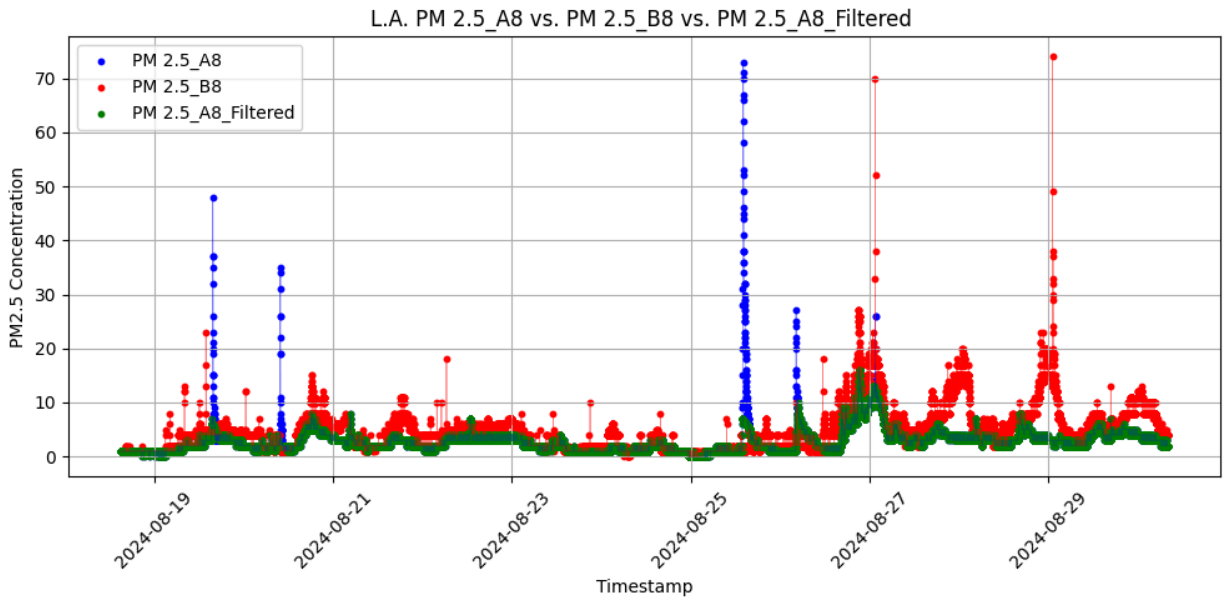

Figure 24. The L.A household indoor, outdoor and filtered indoor $PM_{2.5}$ concentrations.

7.3 Lag analysis

The outdoor $PM_{2.5}$ infiltration requires some time. The peak of outdoor concentration appears first and after some "lag" time we observe indoor peak. To identify best "lag" time the outdoor concentration is shifted to the right and to the left while calculating the correlation between the I&O concentrations. Since the indoor concentration increase can appear after outdoor concentration increase, the right shift of outdoor concentration gives the best result. The example calculation is given in Figure 25.

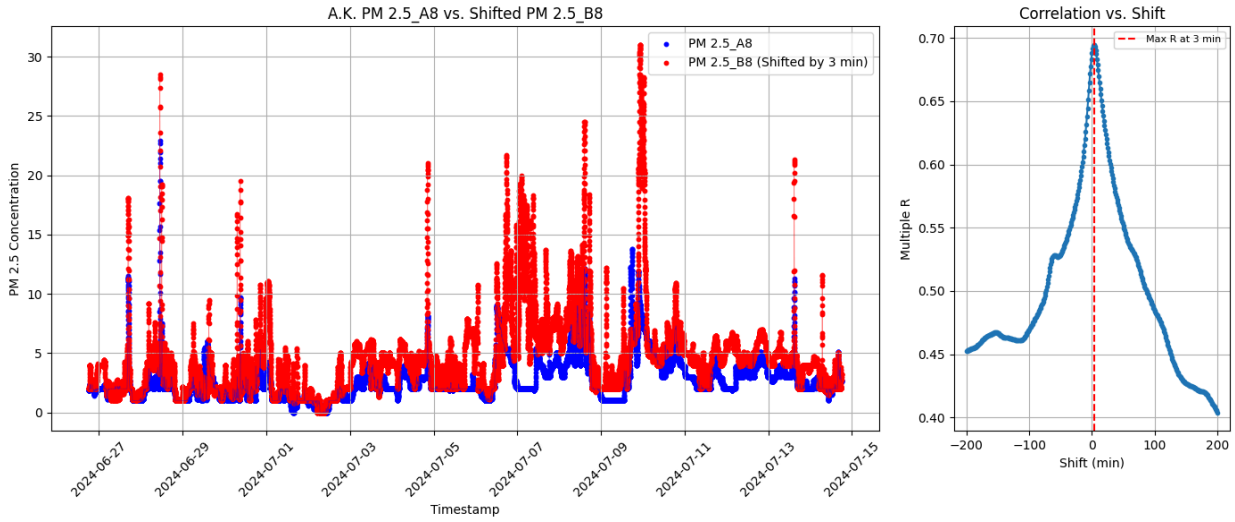

Figure 25. The shortest "lag time" identification.

From the correlation vs. shift part in Figure 25, it is seen that the A.K. household has a 3 min "lag" time when correlation between I&O concentrations reaches its maximum. The left part of Figure 25 shows indoor and shifted outdoor $PM_{2.5}$ concentrations. The Figure 26 is the same procedure performed for another household.

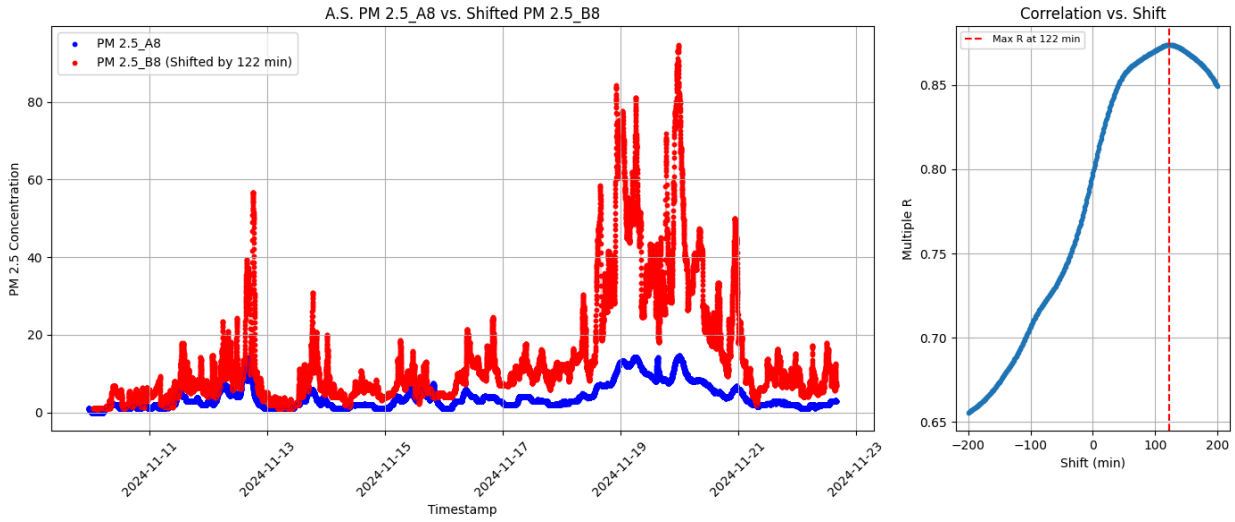

Figure 26. The longest "lag time" identification.

The two households have very different "lag" time of 3 minutes for A.K. and 122 minutes for A.S.. The period of data collection is summer for A.K. and winter for A.S.. The windows in summer are usually open, thus outdoor $PM_{2.5}$ enters the indoor environment quickly. On the other hand, in the winter, windows are usually closed to preserve heat and outdoor $PM_{2.5}$ takes some time to infiltrate indoors. The "lag" time for other households is shown in Table 21.

№	Household	Lag time, minutes	Data collection period
1	Office	63	26.02.2024 - 11.03.2024
2	E.B.	46	05.05.2024 - 24.05.2024
3	A.B.	3	28.05.2024 - 20.06.2024
4	A.K.	3	26.06.2024 - 14.07.2024
5	D.K	7	25.07.2024 - 10.08.2024
6	L.A.	6	18.08.2024 - 31.08.2024
7	N.K.	95	27.10.2024 - 08.11.2024
8	A.S.	122	09.11.2024 - 23.11.2024

Table 21. The "lag time" for all households.

In 2 households where data collection is done in summer, the “lag” time is 3 minutes which is the lowest value. The highest “lag” time is observed in November and it is 122 min. The values are adequate since the Office room showed 63 minutes “lag” time with always closed single window. On the other hand, there are two windows in A.S. and N.K. households, one between outside and balcony, second between balcony and living room (see the details in Figure 9). It means the PM infiltrates the balcony first, then infiltrates the living room where an indoor monitoring device is installed.

7.4 Regression

The multiple linear regression is performed using python. The results include Multiple R, average percentage error and mean squared error (MSE). Firstly, the multiple linear regression is runned for 5 variables without AER. The general results for all 8 households are shown in Table 22. The calculated coefficients for input variables are shown in Table 23.

Household ID	Multiple R	Average Percentage Error	Mean Squared Error
A.K.	0.7388	15.27	1.3244
A.S.	0.8983	17.45	1.9243
A.B.	0.4703	63.16	67.7240
D.K.	0.8991	6.27	0.6984
E.B.	0.8088	17.12	0.6871
L.A.	0.7747	13.94	1.3361
N.K.	0.6232	33.98	18.4890
Office	0.7751	11.61	22.5268

Table 22. Regression results (without AER).

Household ID	const	PM 2.5_B8 _Shifted	Temperature Difference	Barometric Pressure	Relative Humidity_A8	Relative Humidity_B8
A.K.	-85.5	0.3249	-0.1240	0.0909	-0.0501	0.0305
A.S.	18.3	0.1634	0.0284	-0.0102	-0.0485	-0.0923
A.B.	-312.0	0.1752	0.3194	0.2981	0.7459	-0.4043
D.K.	1.1	0.5735	-0.0512	0.0007	-0.0257	0.0061
E.B.	-17.8	0.2400	0.0315	0.0181	0.0438	-0.0239
L.A.	-49.1	0.3101	-0.4243	0.0489	0.0067	0.0655

N.K.	-41.0	0.6188	0.1491	1.1900	0.2491	-0.0967
Office	-203.9	0.2441	1.1369	0.1938	-0.0702	-0.1971

Table 23. Regression coefficients.

The tables above are results without AER. The results with AER calculation are given in Table 24 and Table 25. The AER provides some improvements. The Multiple R increased, while average percentage error and MSE are decreasing. However the effect of AER is very small.

Household ID	Multiple R	Average Percentage Error	Mean Squared Error
A.K.	0.7393	15.24	1.3224
A.S.	0.8983	17.43	1.9244
A.B.	0.4703	63.16	67.7238
D.K.	0.8992	6.27	0.6982
E.B.	0.8088	17.11	0.6870
L.A.	0.7756	13.83	1.3315
N.K.	0.6281	33.42	18.3042
Office	0.7760	11.51	22.4496

Table 24. Regression results (with AER).

Household ID	const	PM 2.5_B8 _Shifted	Temperature Difference	Barometric Pressure	Relative Humidity_A8	Relative Humidity_B8	Air Exchange Rate
A.K.	-85.4	0.3252	-0.1212	0.0906	-0.0481	0.0294	4.6817
A.S.	18.3	0.1633	0.0283	-0.0102	-0.0489	-0.0923	-1.0350
A.B.	-313	0.1751	0.3220	0.2985	0.7481	-0.4056	3.7168
D.K.	1.15	0.5732	-0.0510	0.0006	-0.0254	0.0060	2.5441

E.B.	-17.6	0.2399	0.0313	0.0178	0.0433	-0.0237	-2.4241
L.A.	-49.9	0.3092	-0.4147	0.0495	0.0102	0.0634	1.6236
N.K.	-40.8	0.6246	0.1593	1.1278	0.2674	-0.0944	33.3194
Office	-208	0.2424	1.1402	0.1988	-0.0825	-0.1983	-12.7449

Table 25. Regression coefficients.

The regression results show that households A.S. and D.K. show almost 90% results and A.B. has the lowest value of Multiple R of 47%. From the regression coefficients we see that outdoor $PM_{2.5}$ concentration has always a positive effect on indoor concentration. The same for indoor barometric pressure with one exception in the A.S. household. To identify the more specific influence of independent variables on indoor $PM_{2.5}$ concentration, the normalisation is performed.

Normalization.

The normalization to the range is performed to see regression coefficients. All input variables and Y variable(indoor $PM_{2.5}$ concentration) are normalized in a range between 0 to 1. Table 26 provides multiple regression results.

Household ID	**Multiple R**	**Average Percentage Error**	**Mean Squared Error**
A.K.	0.7393	15.24	0.0025
A.S.	0.8983	17.43	0.0090
A.B.	0.4703	67.51	0.0064
D.K.	0.8992	10.65	0.0010
E.B.	0.8088	17.11	0.0052
L.A.	0.7756	13.83	0.0055

N.K.	0.6281	33.42	0.0058
Office	0.7760	58.93	0.0069

Table 26. Regression results.

The regression coefficients changed due to normalization (see the details in Table 27). All the independent variables are in the range between 0 and 1, thus we can see which variables are more important in the equation.

Household ID	**const**	**PM 2.5_B8 _Shifted**	**Temperature_ Difference**	**Barometric Pressure**	**Relative Humidity_A8 Inside**	**Relative Humidity_B8 Outside**	**Air Exchange Rate**
A.K.	0.0860	0.4403	-0.1068	0.0803	-0.0659	0.0769	0.0296
A.S.	0.3339	1.0505	0.0364	-0.0272	-0.0442	-0.2889	-0.0015
A.B.	-0.0312	0.6895	0.0771	0.0611	0.2933	-0.2089	0.0150
D.K.	0.0498	0.7770	-0.0531	0.0003	-0.0203	0.0135	0.0152
E.B.	0.0345	0.9200	0.0859	0.0363	0.0821	-0.1361	-0.0055
L.A.	0.1131	0.6821	-0.2882	0.0607	0.0170	0.1680	0.0542
N.K.	-0.0676	0.3732	0.0619	0.0165	0.1138	-0.0668	0.0507
Office	-0.0058	0.5913	0.2119	0.0425	-0.0246	-0.1344	-0.0747

Table 27. Regression coefficients(normalised).

A.S. and L.A. households have big constants compared to others. It means they have some background level of pollution. For example, in the A.S. household, the indoor $PM_{2.5}$ level rarely goes to less than one.

The outdoor $PM_{2.5}$ concentration is the most important variable. It is always positive and increases the indoor concentration. The A.S. household has a constant of 1.0505 which means in this household there are very low indoor pollution sources. On the other hand, the N.K. household where residents smoke on the balcony the constant is the lowest, 0.3732.

Temperature difference is indoor temperature minus outdoor temperature. It is the second most important variable. The negative values in temperature difference correspond to hot weather seasons. It is the time in summer when indoor temperature can be lower than outdoor temperature. In contrast, positive coefficients of temperature difference show cold seasons.

The barometric pressure coefficients are positive with one exception in A.S. household, though this household has Multiple R of 0,8983. It is not related to the height of the building, because E.B. household for example is on the 8th floor while A.S. household is on the 6th floor. It means that indoor pressure changes have a small influence on $PM_{2.5}$ infiltration.

The relative humidity inside the house has mixed effects on the infiltration. While it has negative value for A.S. and D.K. households with the highest Multiple R, it has positive value for other households.

The relative humidity outside the household has positive values during the heavy rain period. A.K., D.K. and L.A. households were measured during the summer of 2024 when there was a huge increase in precipitation and hail.

The AER has a mixed effect on indoor $PM_{2.5}$. The low AER means bad ventilation of the household, however, it means $PM_{2.5}$ infiltration from the outdoors is not intensive. When the household has high AER, it loses heat, it gets “fresh air” more and gets outdoor $PM_{2.5}$ more. The AER requires more thorough investigation as it stands out as a separate research topic. In this work, a simple AER formula is added to see if it has an effect on the infiltration. For now, the effect is very small.

The worst performance model belongs to A.B. household. There are peaks which are not predicted by the model. The performance is shown in Figure 27, The blue points represent true measured concentrations while red points are predictions. The performance is generally poor .

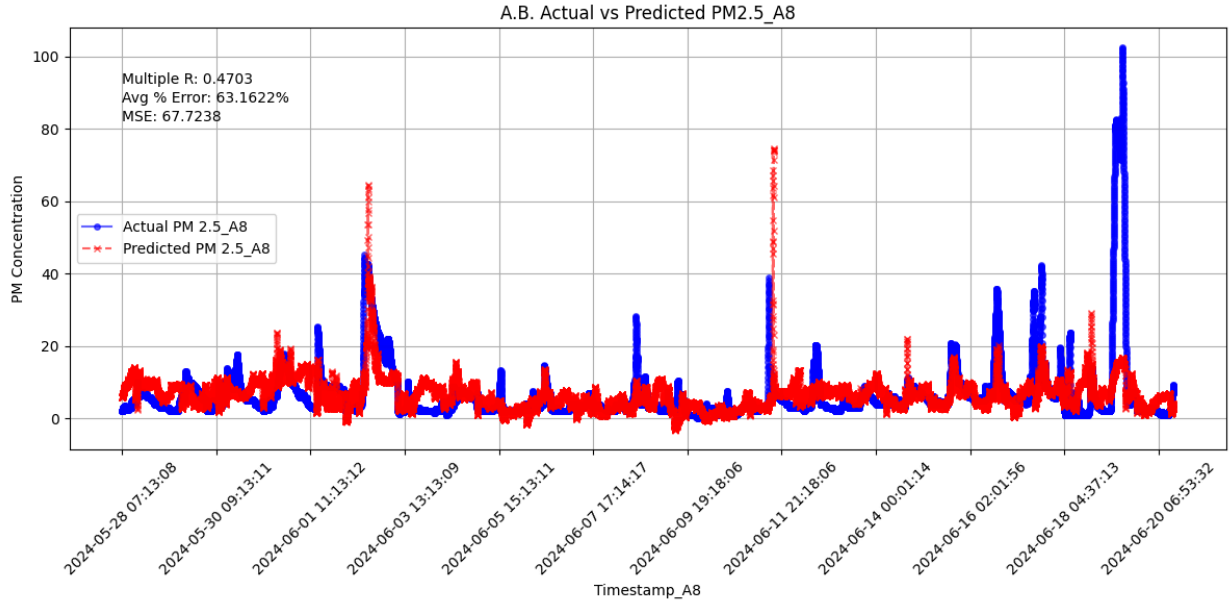

Figure 27. The worst performance model.

The data was modified using a “smoking filter”, and a “lag” time of 3 minutes. Figure 28 shows the wrong “lag” time identification before applying the “smoking filter”. After “smoking filter” the “lag” time is identified correctly (see the details in Figure 29). The resulting Multiple R is 0.47 which is low. However, before the “smoking filter” the Multiple R was 0.29 and “lag” time was -114 minutes which is wrong, because indoor sources in this case can not be a source of outdoor pollution.

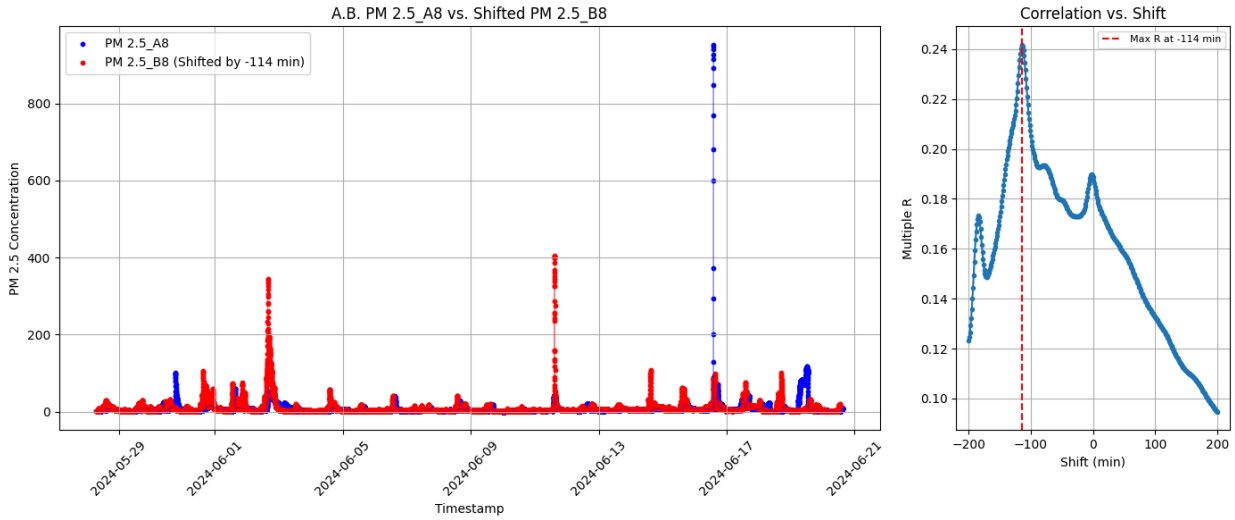

Figure 28. A.B. household before filtering.

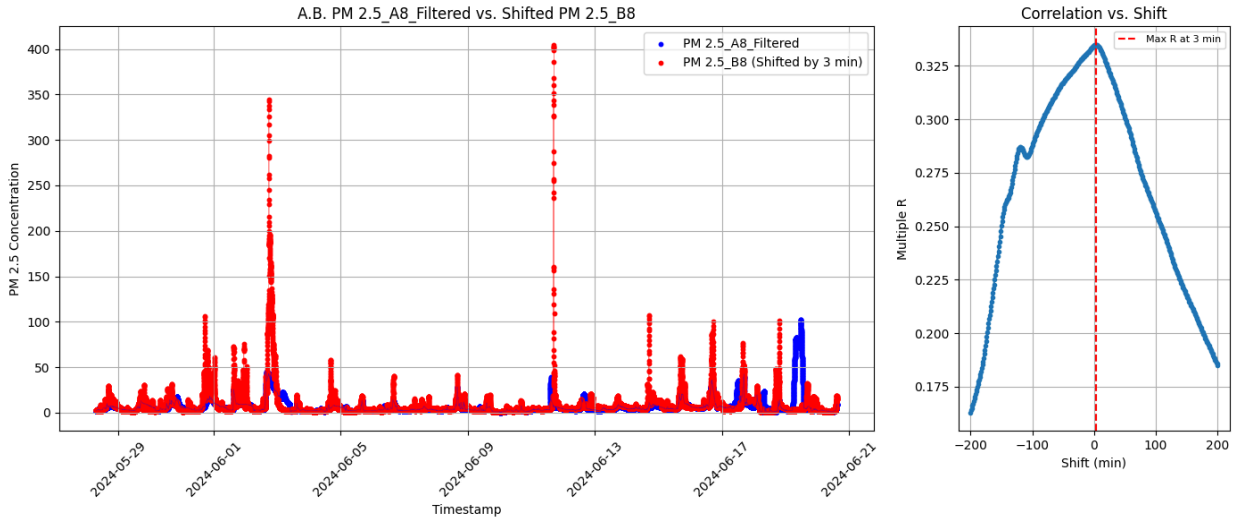

Figure 29. A.B. household after filtering.

The "smoking filter" performance is shown on the figures above. The effectiveness of this filter can be increased by thorough data observation and new peak identification criterias.

The best performance model belongs to the D.K household. It shows good results with very low MSE which is seen in Figure 30. The similar performance is shown in the A.S. household in Figure 31.

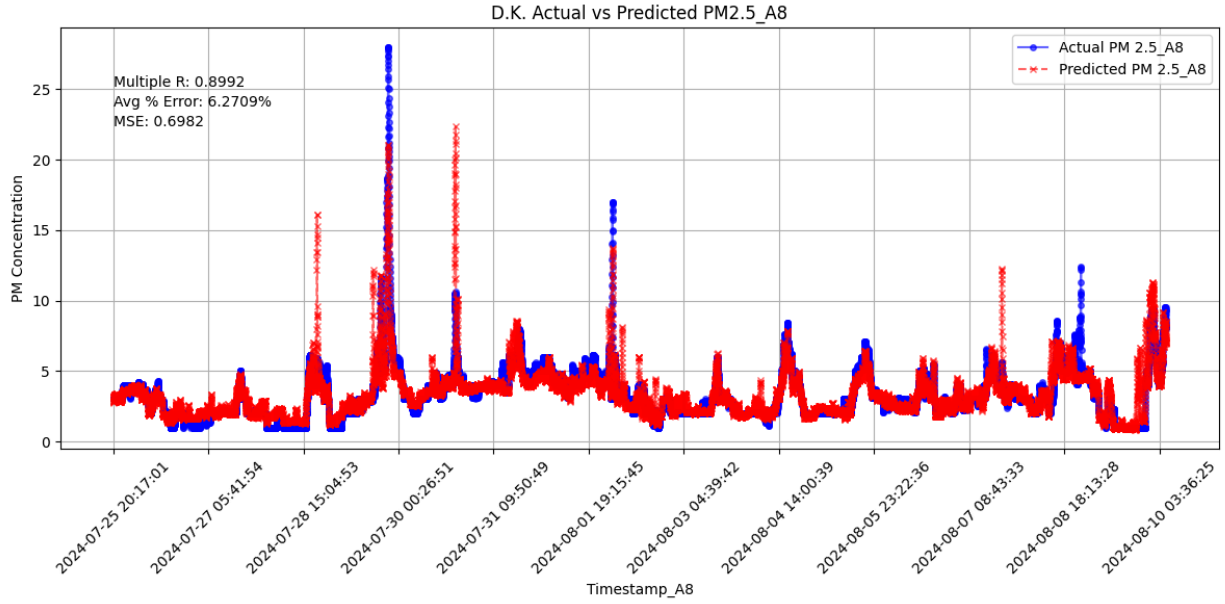

Figure 30. The best performance model.

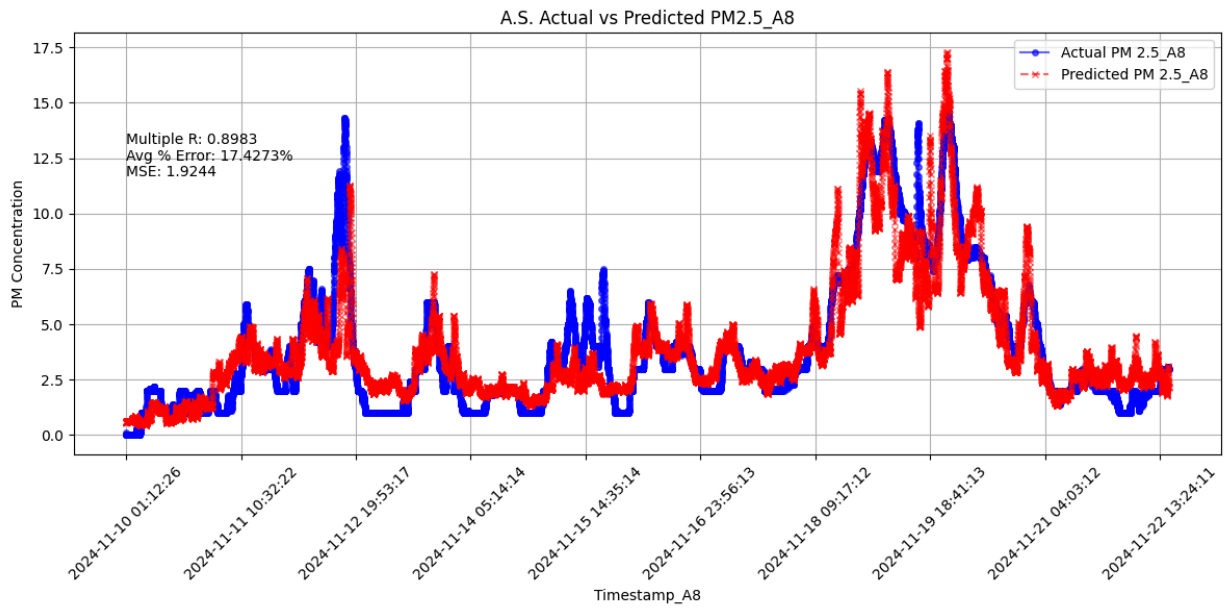

Figure 31. A.S. household model.

These households have similar Multiple R values of almost 90%, but D.K. model outperforms because it has lower values of average percentage error and MSE.

8. Conclusion.

The purpose of the work is to predict indoor $PM_{2.5}$ concentration. The multiple linear regression can be used for predicting indoor $PM_{2.5}$ concentration due to infiltration. The Multiple R is high for two households reaching almost 90%. The lowest value is 0.47 for one household, while others have good predictability of more than 70%. The human factor may be the main reason for low Multiple R because identifying indoor $PM_{2.5}$ contributors and separating them is difficult.

The work provides data collected over the year. The data is collected at 8 different locations of Astana. The attempt taken to model the infiltration of outdoor $PM_{2.5}$ gives useful information for future study. The procedures described in this work can be a good guide on pollutants infiltration study. There are successful decisions made which increase the predictability of indoor concentration while some had no significant impact.

9. Limitations.

Literature review shows that there are many limitations in the study.

1. Absence of some data. The study does not collect data about CO_2 concentration outside. We also miss data about wind speed and direction. However, in the scope of the study, there are enough variables necessary to identify general infiltration mechanisms. The Kazhydromet can provide some data from its measuring stations, but things like wind speed, CO_2 are very specific to each house location. In addition, the measurements are minute based while wind speed and direction are hourly measured.
2. Only one indoor measuring device (Schreck et al., 2024). The indoor environment does not perfectly mix. CO_2 concentration in the kitchen may be higher than in the living room. Thus, the position of the device is important in data collection. I used only one indoor device and if I had 2 or more I could compare them(find average). But I have one. Considering that house architecture and big furniture dictate indoor air flow, there may be differences in $PM_{2.5}$ measurements across one room.
3. Low $PM_{2.5}$ values can be a problem. Some households have very low values for I&O $PM_{2.5}$ concentrations. The concentration outside is 1 ug/m^3 while concentration inside is 0 ug/m^3 . Generally, low $PM_{2.5}$ concentration is good for health and the environment is considered as clean. The purpose of the thesis is to predict indoor $PM_{2.5}$ concentration due to infiltration. If the outdoor concentration is 1 ug/m^3, then the indoor concentration can be 1 or 0 ug/m^3. 1 is unlikely because it means 100% infiltration. It comes out, indoor concentration is very low, between 0 - 1 ug/m^3.
4. In some cases it is easy to see indoor related $PM_{2.5}$ concentration increase. Sharp increases in indoor level without preceding outdoor increase say that some human activity is taking place inside. However, there are indoor related $PM_{2.5}$ contributions that are difficult to detect. And even with removed peaks, when $PM_{2.5}$ concentration drops fast to some level, we cannot be sure to what level it dropped. One possibility is that it drops till outdoor level or peak can be removed based on visual judgement. While the start of the peak can be easily seen, the end is subjective.

10. Suggestions for the future research

It is anticipated that the outcomes across various buildings will vary due to differences in architectural design, construction materials, and other factors. Nonetheless, certain similarities may emerge as a result of shared outdoor environmental conditions, which present an intriguing aspect for observation. It is probable that lag times for different household environments vary and depend on many factors mentioned before. In addition, it is interesting to identify daily patterns in the PM concentrations. The study by Mukhtarov et al., found that $PM_{2.5}$ concentration starts increasing in the morning because people travel to work. The local peak is reached before noon and decreases until the evening when people go home. I also consider using more convenient and advanced software. Nadali et al. (2020) carried out research on I/O PM concentrations in Qom city, which is located in Iran. Most of the statistical analysis, including calculations of mean, maximum, and minimum concentrations, as well as standard deviation, were performed using SPSS software 20.0 (SPSS Inc., NY, USA). Prior to analysis, the normality of the data was assessed using the Shapiro-Wilk test. The regression analysis was carried out to examine the relationship between indoor and outdoor PM concentrations for each size fraction. They introduced (r) - coefficient which represents the correlation between I&O PM concentrations. It stands out as an indicator of the extent to which PM concentration monitored indoors is influenced by outdoor PM concentration and its infiltration.

It is necessary to compare the results with other studies to confirm the results. The study by Nadali et al, (2020) obtained the following results: a negative correlation between PM concentration and wind speed, between PM concentration and temperature. There was no significant correlation between PM concentration and relative humidity. The increasing temperature and wind speed, make the I/O ratios of PM decrease, whereas humidity did not show a clear association with I/O ratios. Higher wind speeds facilitate dispersion and dilution of outdoor PM concentrations, resulting in decreased indoor PM concentrations. The positive correlation between temperature and PM concentration may be attributed to thermal diffusion, where higher outdoor temperatures push PM indoors through openings like windows and doorways, while the opposite occurs with lower ambient temperatures.

From the research methods mentioned above, it is necessary to choose one that is more appropriate for studying household environments in Astana. The various approaches used by researchers for estimating PM, its constituent variables, and modeling the environment with PM sources and transportation have their advantages and disadvantages, which require more detailed study.

Identifying people's presence.

It may be useful to select time when people are not at home. Because CO_2 decreases when people are not home. People contribute to $PM_{2.5}$ production by some activities(cooking, aerosols)and contribute to CO_2 production by breathing. Empty households may provide more accurate data because PM infiltration is not disturbed by human activities. It is necessary to identify the criterias to identify if people are present. The first variable to be used is CO_2. If CO_2 is increasing, it means people are present. If it is decreasing, people leave the household or open the window and CO_2 is diluting. Since there are two options we need more specific criteria. To do so, we can check the rate of CO_2 change. If the window is opened the decrease of CO_2 should be sharp.

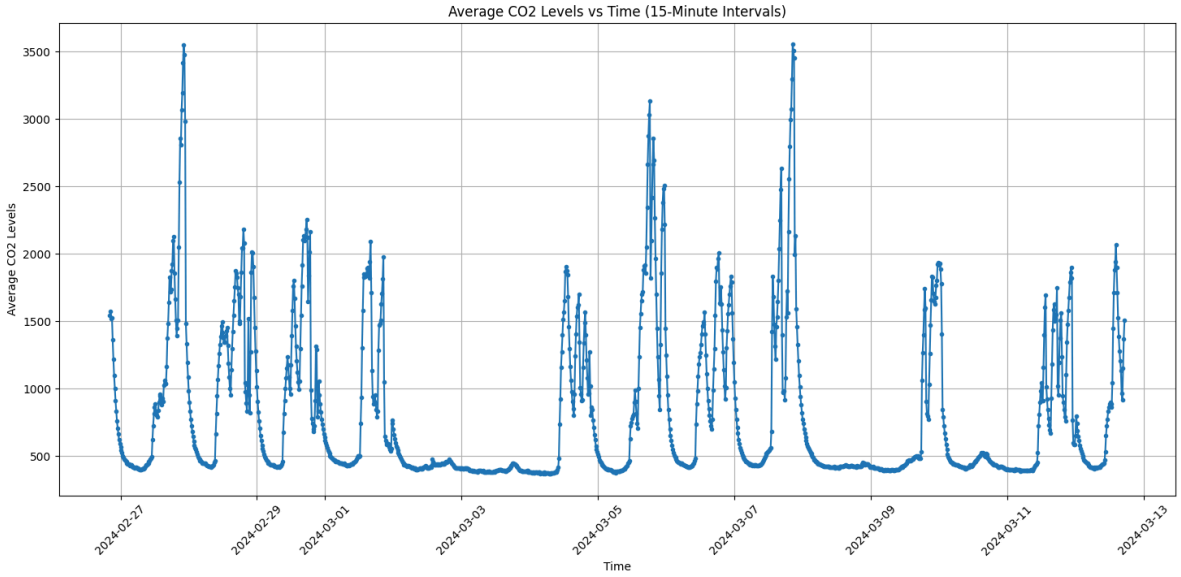

Figure 32. The Office CO_2 concentration.

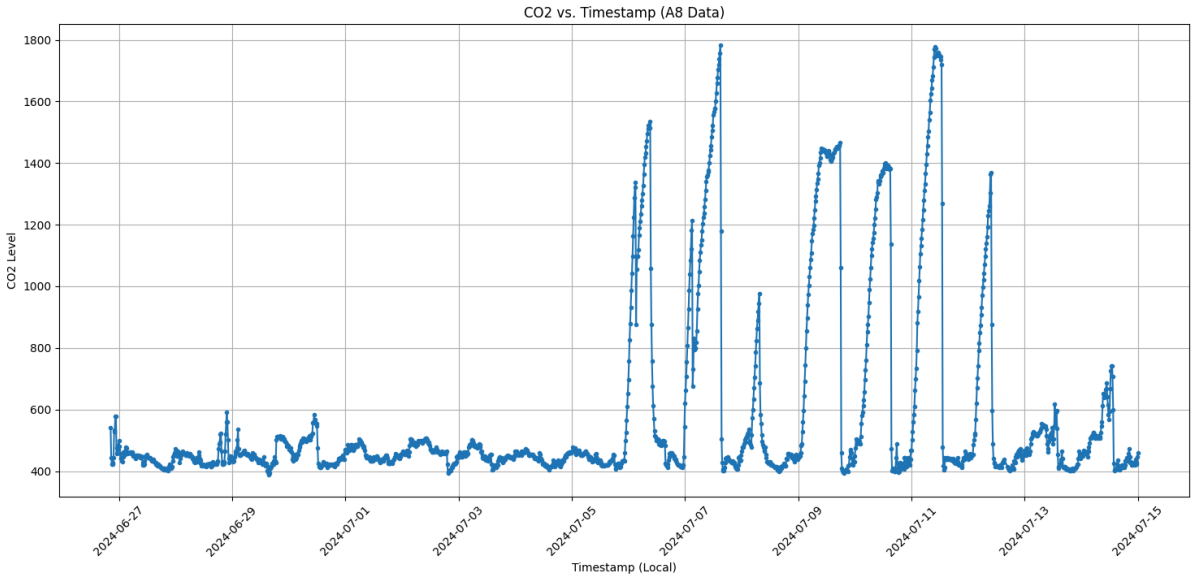

Figure 33. The A.B. household CO_2 concentration.

The CO_2 in the household № 1 could only decay because the window was always closed. In Figure 19 we can notice a clear decrease pattern. The toe of the decreasing part becomes smooth and contains several data points. On the other hand, the CO_2 concentration in Figure 20 drops fast which may only mean the window opening. Knowing that usually people close the windows when leaving the house, we may consider the rate of CO_2 decrease as the first criteria of identifying if people are present. Smooth decay is likely to mean that people left the household.

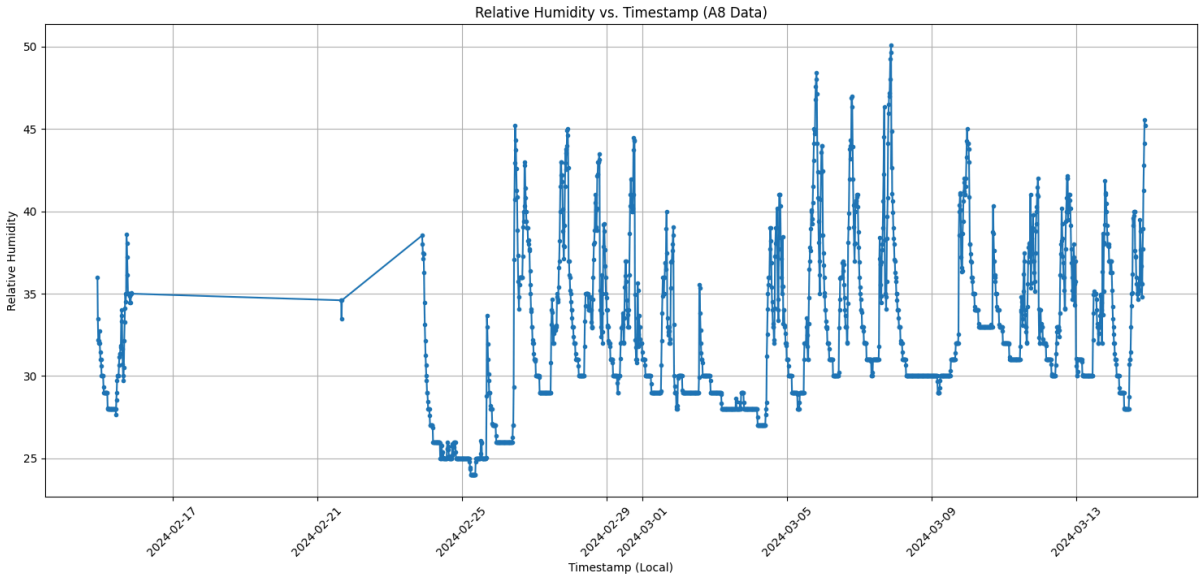

Figure 34. The Office relative humidity inside.

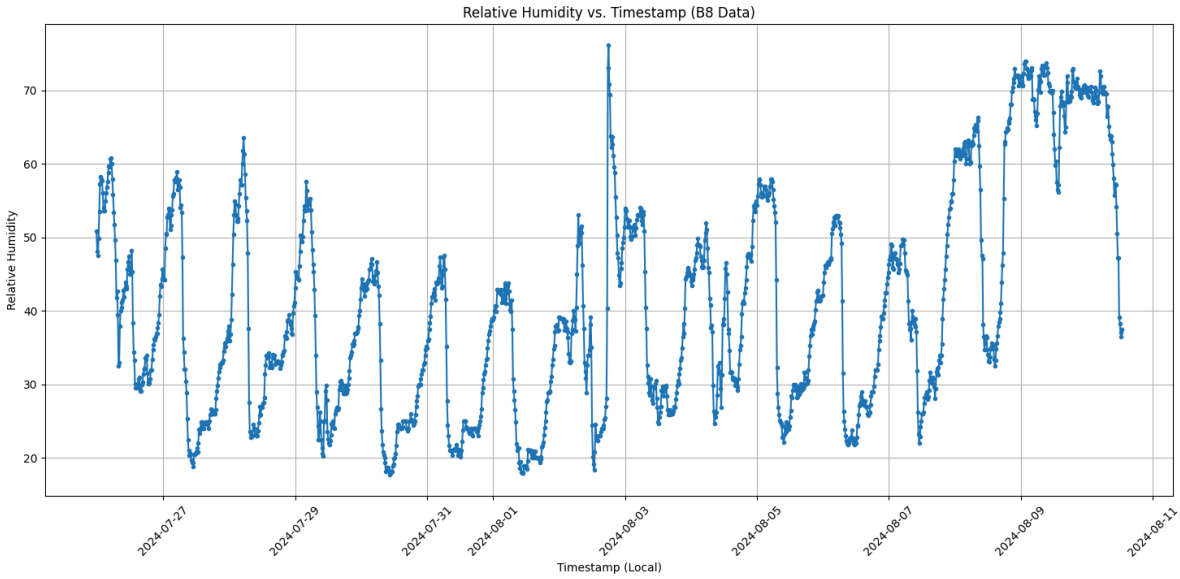

Figure 35. The D.K. household relative humidity outside.

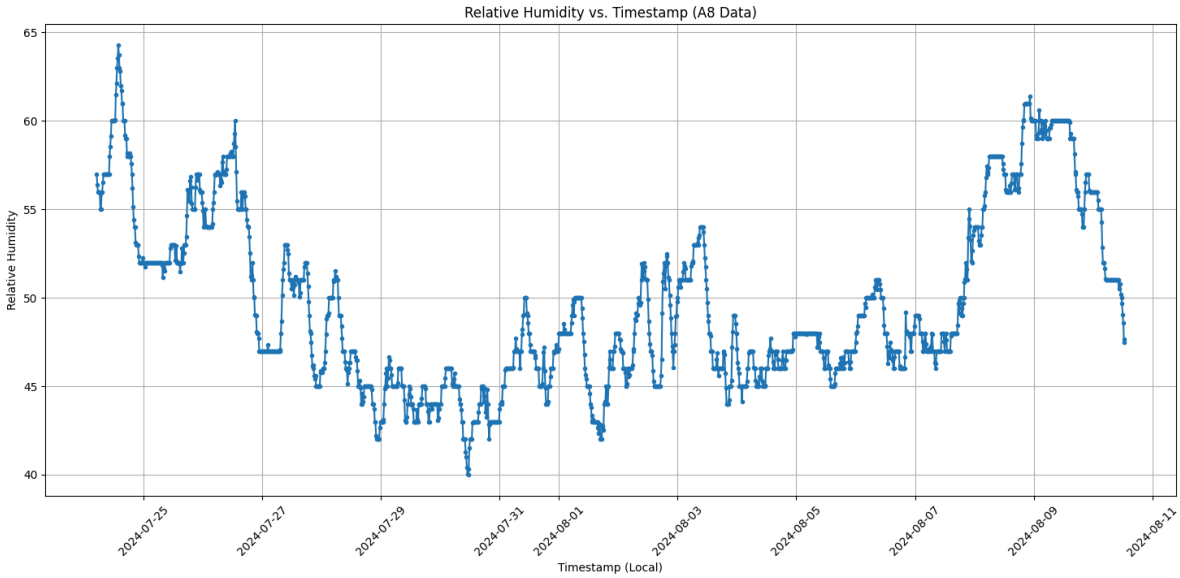

Figure 36. The D.K. household relative humidity inside.

The second criteria can be relative humidity. From Figures 19 and 21 we can see that with a closed window relative humidity and CO_2 concentration are almost identical. The number of peaks can be counted and they are equal. Relative humidity is also a good indicator of people's presence. However as for the CO_2, it decreases slowly when people leave and may decrease sharply when the window is opened. The same rate of change approach may be implemented for relative humidity and it may serve as a second criteria after CO_2. The advantage of the relative humidity indicator is that we have data for both I&O relative humidity, while for the CO_2 we have only indoor concentration. Thus, relative humidity may be the first priority indicator.

The relative humidity is measured in percent. The figures 22 and 23 show relative humidity inside and outside household № 5. The time when data was collected was very humid because there often were rains in the august. The outside relative humidity is often high. The pattern of relative humidity outside is clearly detected while relative humidity inside does not have clear pattern and fluctuates a lot.

The third criteria may be time. Usually people are inside during the night time because they sleep and leave the household in the morning to go to work. There will be exceptions in people’s

behavior, for example staying home during holidays and weekends. They may also arrive late or early. Thus, time may serve as a third priority indicator.

One more controversial criteria is increase in indoor $PM_{2.5}$ concentration without increase of outdoor $PM_{2.5}$ concentration. It may mean that people are doing some of the many activities that produce $PM_{2.5}$. Thus, people are present if indoor concentration is increasing while outdoor concentration is stable or decreasing. However, it may be a “lag time” mentioned earlier.

The Python code will be used for removing data which corresponds to time when people are present. The draft code used now colors red the rows where $CO_2 > 700$ and relative humidity > 40. The code is provided in the appendix.

Appendix

The Python code used for data alignment

```
import pandas as pd

# Read both CSV files
file1 = pd.read_csv('C:\\Users\\aziz\\Desktop\\plan\\Selected data. First & Last hours removed\\D.K A8.csv')
file2 = pd.read_csv('C:\\Users\\aziz\\Desktop\\plan\\Selected data. First & Last hours removed\\D.K B8.csv')

# Convert the 'Timestamp (Local)' column to datetime, specifying the exact format MM/DD/YYYY %H:%M:%S
file1['Timestamp'] = pd.to_datetime(file1['Timestamp'], format='%m/%d/%Y %H:%M:%S', errors='coerce')
file2['Timestamp'] = pd.to_datetime(file2['Timestamp'], format='%m/%d/%Y %H:%M:%S', errors='coerce')

# Drop rows where the timestamp couldn't be parsed
file1 = file1.dropna(subset=['Timestamp'])
file2 = file2.dropna(subset=['Timestamp'])

# Create new columns for year, month, day, hour, and minute (ignore seconds) for both files
file1['Year'] = file1['Timestamp'].dt.year
file1['Month'] = file1['Timestamp'].dt.month
file1['Day'] = file1['Timestamp'].dt.day
file1['Hour'] = file1['Timestamp'].dt.hour
file1['Minute'] = file1['Timestamp'].dt.minute

file2['Year'] = file2['Timestamp'].dt.year
file2['Month'] = file2['Timestamp'].dt.month
file2['Day'] = file2['Timestamp'].dt.day
file2['Hour'] = file2['Timestamp'].dt.hour
file2['Minute'] = file2['Timestamp'].dt.minute

# Merge the two dataframes based on matching Year, Month, Day, Hour, and Minute
merged_data = pd.merge(file1, file2, on=['Year', 'Month', 'Day', 'Hour', 'Minute'], suffixes=('_A8', '_B8'))

# Drop the extra 'Year', 'Month', 'Day', 'Hour', 'Minute' columns from the final output
merged_data = merged_data.drop(columns=['Year', 'Month', 'Day', 'Hour', 'Minute'])
```

```
# Write the combined data to a new Excel file
merged_data.to_csv('D.K Aligned.csv', index=False)

print("D.K data aligned in csv")
```

The Python code used for calculating moving averages.

```
import pandas as pd

# Load the data from the CSV file
file_path = 'C:\\Users\\aziz\\Desktop\\air\\1 print PM 2.5 indoor outdoor\\A.B. Aligned.csv'
data = pd.read_csv(file_path)

# Convert 'Timestamp' column to datetime format if necessary (assuming it's already in datetime format)
data['Timestamp_A8'] = pd.to_datetime(data['Timestamp_A8'], format='%Y-%m-%d %H:%M:%S')

# Set 'Timestamp' as index for easier time-based calculations
data.set_index('Timestamp_A8', inplace=True)

# Select only numeric columns for moving average calculation
numeric_columns = data.select_dtypes(include=['int64', 'float64']).columns

# Calculate moving averages for numeric columns over a 10-minute window, starting at 10th row
moving_averages = data[numeric_columns].rolling('10min').mean()

# Since we want full windows only, we'll drop rows where MA isn't fully calculated (first 9 rows in this context)
fully_calculated_moving_averages = moving_averages.iloc[9:]  # Start from index 9 (10th row)

# Reset index to include timestamp in output DataFrame
fully_calculated_moving_averages.reset_index(inplace=True)

# Save fully calculated moving averages to a new CSV file
fully_calculated_moving_averages.to_csv('A.B. MA.csv', index=False)
print("A.B. data ready!")
```

The Python code used for “smoking filter”.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Load the datA.B.K
file_path = 'C:\\Users\\aziz\\Desktop\\plan\\Aligned\\L.A. Aligned.csv'
data = pd.read_csv(file_path, parse_dates=['Timestamp_A8'])

# Ensure required columns exist
if 'PM 2.5_A8' not in data.columns or 'PM 2.5_B8' not in data.columns:
    raise KeyError("Missing required columns: 'PM 2.5_A8' or 'PM 2.5_B8'")

# Detect sudden increases
threshold = 4  # Increase threshold for consecutive increases
consecutive_count = 2  # Detecting 2 consecutive increases
large_jump_threshold = 10  # Threshold for a large single increase
indexes_to_connect = []
skipped_ranges = []  # List of (start, end) timestamps to prevent overlap
labels = []  # Labels for annotation
filtered_values = data['PM 2.5_A8'].copy()  # New column for filtered values

for i in range(len(data) - consecutive_count):
    if any(start <= data['Timestamp_A8'].iloc[i] <= end for start, end in skipped_ranges):
        continue  # Skip indexes within the ignored range

    # Condition for consecutive increases
    if all(data['PM 2.5_A8'].iloc[j + 1] - data['PM 2.5_A8'].iloc[j] > threshold for j in range(i, i + consecutive_count)):
        start_index = max(0, i - 1)  # Adjust to start 3 minutes earlier if possible
        start_time_adjusted = data['Timestamp_A8'].iloc[start_index] - pd.Timedelta(minutes=3)

        if start_time_adjusted < data['Timestamp_A8'].iloc[0]:
            start_time_adjusted = data['Timestamp_A8'].iloc[0]  # Prevent going before the datA.B.K start

        time_later = data['Timestamp_A8'].iloc[start_index] + pd.Timedelta(hours=2)
        end_index = data['Timestamp_A8'].sub(time_later).abs().idxmin()  # Find closest timestamp 2 hours later
```

```
        indexes_to_connect.append((start_time_adjusted, data['Timestamp_A8'].iloc[end_index]))
        labels.append((data['Timestamp_A8'].iloc[start_index], "Consecutive increase ≥4"))

        # Store the range to prevent overlap
        skipped_ranges.append((start_time_adjusted, data['Timestamp_A8'].iloc[end_index]))

# Detect large single increases (non-overlapping)
for i in range(len(data) - 1):
    if any(start <= data['Timestamp_A8'].iloc[i] <= end for start, end in skipped_ranges):
        continue  # Skip indexes already covered

    if data['PM 2.5_A8'].iloc[i + 1] - data['PM 2.5_A8'].iloc[i] >= large_jump_threshold:
        start_time_adjusted = data['Timestamp_A8'].iloc[i] - pd.Timedelta(minutes=2)

        if start_time_adjusted < data['Timestamp_A8'].iloc[0]:
            start_time_adjusted = data['Timestamp_A8'].iloc[0]  # Prevent going before the datA.B.K start

        time_later = data['Timestamp_A8'].iloc[i] + pd.Timedelta(minutes=30)
        end_index = data['Timestamp_A8'].sub(time_later).abs().idxmin()  # Find closest timestamp 30 min later
        indexes_to_connect.append((start_time_adjusted, data['Timestamp_A8'].iloc[end_index]))
        labels.append((data['Timestamp_A8'].iloc[i], "Single increase ≥10"))

        # Store the range to prevent overlap
        skipped_ranges.append((start_time_adjusted, data['Timestamp_A8'].iloc[end_index]))

# Apply smoothing to the new column
for start_time, end_time in indexes_to_connect:
    start_index = data['Timestamp_A8'].sub(start_time).abs().idxmin()
    end_index = data['Timestamp_A8'].sub(end_time).abs().idxmin()

    new_values = np.linspace(data['PM 2.5_A8'].iloc[start_index], data['PM 2.5_A8'].iloc[end_index], end_index -
start_index + 1)
    filtered_values.iloc[start_index:end_index + 1] = new_values

# Add the new column to the dataframe
data['PM 2.5_A8_Filtered'] = filtered_values
```

```
# Save the new datA.B.K
data.to_csv('L.A._Aligned_Filtered.csv', index=False)

# Create the plot
plt.figure(figsize=(10, 5))
plt.scatter(data['Timestamp_A8'], data['PM 2.5_A8'], label='PM 2.5_A8', color='blue', marker='o', s=10)
plt.plot(data['Timestamp_A8'], data['PM 2.5_A8'], color='blue', linewidth=0.5, alpha=0.5)

plt.scatter(data['Timestamp_A8'], data['PM 2.5_B8'], label='PM 2.5_B8', color='red', marker='o', s=10)
plt.plot(data['Timestamp_A8'], data['PM 2.5_B8'], color='red', linewidth=0.5, alpha=0.5)

# Draw green lines for detected sudden increases
for idx, (start_time, end_time) in enumerate(indexes_to_connect):
    start_index = data['Timestamp_A8'].sub(start_time).abs().idxmin()
    end_index = data['Timestamp_A8'].sub(end_time).abs().idxmin()

    plt.plot(
        [data['Timestamp_A8'].iloc[start_index], data['Timestamp_A8'].iloc[end_index]],
        [data['PM 2.5_A8'].iloc[start_index], data['PM 2.5_A8'].iloc[end_index]],
        color='green', linewidth=2, linestyle='--'
    )

    # Add annotation above the green line
    plt.text(
        data['Timestamp_A8'].iloc[start_index],
        data['PM 2.5_A8'].iloc[start_index] + 2,  # Adjust height slightly above
        labels[idx][1],
        fontsize=9, color='green', fontweight='bold'
    )

plt.title('L.A. PM 2.5_A8 vs. PM 2.5_B8 (Sudden Increase Smoothed)')
plt.xlabel('Timestamp')
plt.ylabel('PM 2.5 Concentration')
plt.legend()
plt.grid()
plt.xticks(rotation=45)
```

```
plt.tight_layout()
plt.show()

print("L.A. data ready!")




# Load the datA.B.K
file_path = 'C:\\Users\\aziz\\Desktop\\air\\1 print PM 2.5 indoor outdoor\\L.A._Aligned_Filtered.csv'
data = pd.read_csv(file_path, parse_dates=['Timestamp_A8'])



# Ensure required columns exist
if 'PM 2.5_A8' not in data.columns or 'PM 2.5_B8' not in data.columns or 'PM 2.5_A8_Filtered' not in data.columns:
    raise KeyError("Missing required columns: 'PM 2.5_A8' or 'PM 2.5_B8' or 'PM 2.5_A8_Filtered'")

# Create the plot
plt.figure(figsize=(10, 5))
plt.scatter(data['Timestamp_A8'], data['PM 2.5_A8'], label='PM 2.5_A8', color='blue', marker='o', s=10)
plt.plot(data['Timestamp_A8'], data['PM 2.5_A8'], color='blue', linewidth=0.5, alpha=0.5)

plt.scatter(data['Timestamp_A8'], data['PM 2.5_B8'], label='PM 2.5_B8', color='red', marker='o', s=10)
plt.plot(data['Timestamp_A8'], data['PM 2.5_B8'], color='red', linewidth=0.5, alpha=0.5)


plt.scatter(data['Timestamp_A8'], data['PM 2.5_A8_Filtered'], label='PM 2.5_A8_Filtered', color='green',
marker='o', s=10)
plt.plot(data['Timestamp_A8'], data['PM 2.5_A8_Filtered'], color='green', linewidth=0.5, alpha=0.5)



plt.title('L.A. PM 2.5_A8 vs. PM 2.5_B8 vs. PM 2.5_A8_Filtered ')
plt.xlabel('Timestamp')
plt.ylabel('PM 2.5 Concentration')
plt.legend()
plt.grid()
```

```
plt.xticks(rotation=45)

plt.tight_layout()
plt.show()

print("3 columns plot ready!")
```

The Python code used for “lag time” detection.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load the datA.B.K
file_path = 'C:\\Users\\aziz\\Desktop\\air\\2 MA\\N.K. MA.csv'
data = pd.read_csv(file_path, parse_dates=['Timestamp_A8'])

# Ensure required columns exist
if 'PM 2.5_A8_Filtered' not in data.columns or 'PM 2.5_B8' not in data.columns:
    raise KeyError("Missing required columns: 'PM 2.5_A8_Filtered' or 'PM 2.5_B8'")

# Define the range of shifts
shift_range = range(-200, 201)  # Shifting from -50 to +50 minutes
correlations = []

# Compute Multiple R (Pearson correlation) for each shift
for shift in shift_range:
    shifted_data = data[['PM 2.5_A8_Filtered']].copy()
    shifted_data['PM 2.5_B8'] = data['PM 2.5_B8'].shift(shift)
    valid_data = shifted_data.dropna()

    if not valid_data.empty:
        multiple_r = valid_data.corr().iloc[0, 1]  # Pearson correlation
        correlations.append(multiple_r)
    else:
        correlations.append(np.nan)  # Avoid errors when all values are NaN
```

```
# Find the best shift (maximum correlation)
best_shift = shift_range[np.nanargmax(correlations)]
max_r = np.nanmax(correlations)

# Apply the best shift
data['Timestamp_A8_Shifted'] = data['Timestamp_A8'].shift(best_shift)
data['PM 2.5_B8_Shifted'] = data['PM 2.5_B8'].shift(best_shift)
data.to_csv('1 N.K. MA shifted.csv', index=False)

# Create the main figure with two subplots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 6), gridspec_kw={'width_ratios': [3, 1]})

# Main plot (PM 2.5_A8 vs. Shifted PM 2.5_B8)
ax1.scatter(data['Timestamp_A8'], data['PM 2.5_A8_Filtered'], label='PM 2.5_A8_Filtered', color='blue',
marker='o', s=10)
ax1.plot(data['Timestamp_A8'], data['PM 2.5_A8_Filtered'], color='blue', linewidth=0.5, alpha=0.5)

ax1.scatter(data['Timestamp_A8'], data['PM 2.5_B8_Shifted'], label=f'PM 2.5_B8 (Shifted by {best_shift} min)',
        color='red', marker='o', s=10)
ax1.plot(data['Timestamp_A8'], data['PM 2.5_B8_Shifted'], color='red', linewidth=0.5, alpha=0.5)

ax1.set_title('N.K. PM 2.5_A8_Filtered vs. PM 2.5_B8_Shifted')
ax1.set_xlabel('Timestamp')
ax1.set_ylabel('PM 2.5 Concentration')
ax1.legend()
ax1.grid()
ax1.tick_params(axis='x', rotation=45)

# Correlation vs. Shift (placed separately on the right)
ax2.plot(shift_range, correlations, marker='o', linestyle='-', markersize=3)
ax2.axvline(best_shift, color='red', linestyle='--', label=f'Max R at {best_shift} min')
ax2.set_xlabel('Shift (min)', fontsize=10)
ax2.set_ylabel('Multiple R', fontsize=10)
ax2.set_title('Correlation vs. Shift', fontsize=12)
ax2.legend(fontsize=8)
ax2.grid()
```

```
plt.tight_layout()
plt.show()

print("N.K. data ready!")
best_shift, max_r
```

The Python code used for removing empty rows in the data.

```
import pandas as pd

# Define file paths
input_file = r"C:\Users\aziz\Desktop\air\3 Lag Analysis\1 A.B. MA shifted.csv"
output_file = r"C:\Users\aziz\Desktop\air\4 Clean for regression\1 A.B. MA single shifted.csv"

# Load the datA.B.K
data = pd.read_csv(input_file)

# Drop rows where "PM 2.5_B8_Shifted" or "PM 2.5_A8" is empty
cleaned_data = data.dropna(subset=["PM 2.5_B8_Shifted", "PM 2.5_A8"])

# Save the cleaned data to a new CSV file
cleaned_data.to_csv(output_file, index=False)

print("Cleaning complete. New file saved as:", output_file)
```

The Python code used for removing empty rows in the data.

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error

# Load the data from the CSV file
file_path = 'C:\\Users\\aziz\\Desktop\\air\\4 Clean for regression\\1 A.B. MA single shifted.csv'
```

```
data = pd.read_csv(file_path)

# Assuming 'CO2' is the column name for CO2 values
co2_current = data['CO2']  # Current CO2 values
co2_previous = data['CO2'].shift(1)  # Previous CO2 values

# Initialize a list to store air change rates
air_change_rates = []

# Calculate air change rate based on conditions
for index in range(len(data)):
    K2 = co2_previous[index] if index > 0 else None  # Previous value for first row is None
    K3 = co2_current[index]  # Current value

    if K3 is None or K2 is None:
        air_change_rate = None  # Handle cases where K2 or K3 is not available
    elif K3 >= K2 + 64:
        air_change_rate = (K2 + 2*0.343 * 20 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 56:
        air_change_rate = (K2 + 2*0.343 * 8 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 48:
        air_change_rate = (K2 + 2*0.343 * 7 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 40:
        air_change_rate = (K2 + 2*0.343 * 6 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 32:
        air_change_rate = (K2 + 2*0.343 * 5 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 24:
        air_change_rate = (K2 + 2*0.343 * 4 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 16:
        air_change_rate = (K2 + 2*0.343 * 3 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 8:
        air_change_rate = (K2 + 2*0.343 * 2 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2:
        air_change_rate = (K2 + 2*0.343 / 68063 * 1000000 - K3) / K2
    else:
        air_change_rate = -(K3 - K2) / K2
```

```
    # Append the calculated rate to the list
    air_change_rates.append(air_change_rate)

# Add the calculated air change rates to the DataFrame as a new column
data['Air Change Rate'] = air_change_rates

# Select relevant columns for regression
Y = data['PM 2.5_A8']   # Dependent variable

# Independent variables (X)
X1 = data['PM 2.5_B8_Shifted']   # Independent variable X1
X2 = data['Temperature_A8'] - data['Temperature_B8']   # Difference between temperatures (X2)
X3 = data['Barometric Pressure']   # Independent variable X3
X4 = data['Relative Humidity_A8']   # Independent variable X4
X5 = data['Relative Humidity_B8']   # Independent variable X5

# Combine independent variables into a DataFrame, including Air Change Rate as X6
X = pd.DataFrame({
    'PM 2.5_B8_Shifted': X1,
    'Temperature_Difference': X2,
    'Barometric Pressure': X3,
    'Relative Humidity_A8': X4,
    'Relative Humidity_B8': X5,
    'Air Change Rate': data['Air Change Rate']   # Adding Air Change Rate as X6
})

# Check for missing values in Y and X before fitting the model
original_row_count = len(data)
if Y.isnull().any() or X.isnull().any().any():
    print("Warning: Missing values detected in Y or X.")

# Drop rows with missing values in Y or any column of X before regression analysis
data_cleaned = data.dropna(subset=['PM 2.5_A8', 'PM 2.5_B8_Shifted', 'Temperature_A8',
                                   'Temperature_B8', 'Barometric Pressure',
                                   'Relative Humidity_A8', 'Relative Humidity_B8',
```

```
                'Air Change Rate'])

# Calculate how many rows were skipped
skipped_rows_count = original_row_count - len(data_cleaned)
print(f"Number of rows skipped due to missing values: {skipped_rows_count}")

Y_cleaned = data_cleaned['PM 2.5_A8']
X_cleaned = pd.DataFrame({
    'PM 2.5_B8_Shifted': data_cleaned['PM 2.5_B8_Shifted'],
    'Temperature_Difference': data_cleaned['Temperature_A8'] - data_cleaned['Temperature_B8'],
    'Barometric Pressure': data_cleaned['Barometric Pressure'],
    'Relative Humidity_A8': data_cleaned['Relative Humidity_A8'],
    'Relative Humidity_B8': data_cleaned['Relative Humidity_B8'],
    'Air Change Rate': data_cleaned['Air Change Rate']
})

# Add a constant to the model (intercept)
X_cleaned = sm.add_constant(X_cleaned)

# Fit the regression model using cleaned data
model = sm.OLS(Y_cleaned, X_cleaned).fit()

# Get the regression results summary
regression_results = model.summary()

# Calculate Multiple R (square root of R-squared)
multiple_r = model.rsquared ** 0.5

# Make predictions using the model
predicted_values = model.predict(X_cleaned)

# Calculate Mean Squared Error (MSE)
mse = mean_squared_error(Y_cleaned, predicted_values)

# Calculate percentage average error
percentage_errors = ((predicted_values - Y_cleaned) / Y_cleaned) * 100
```

```
average_percentage_error = percentage_errors.mean()

# Print the regression results, Multiple R, and average percentage error
print("A.B. Regression Results Summary:")
print(regression_results)
print("\nMultiple R:", multiple_r)
print("\nAverage Percentage Error of Predicted Values:", average_percentage_error)
print("\nMean Squared Error (MSE):", mse)

# Plotting Actual vs Predicted PM 2.5_A8 values
plt.figure(figsize=(12,6))
plt.plot(data_cleaned.index, Y_cleaned, label='Actual PM 2.5_A8', color='blue', marker='o', markersize=4, alpha=0.6)
plt.plot(data_cleaned.index, predicted_values, label='Predicted PM 2.5_A8', color='red', linestyle='--', marker='x', markersize=4, alpha=0.6)

# Annotate metrics on the plot
plt.title('A.B. Actual vs Predicted PM 2.5_A8')
plt.xlabel('Index')
plt.ylabel('PM Concentration')
plt.legend()

# Display metrics on the plot
plt.text(0.05, max(Y_cleaned)*0.9, f'Multiple R: {multiple_r:.4f}', fontsize=10)
plt.text(0.05, max(Y_cleaned)*0.85, f'Avg % Error: {average_percentage_error:.4f}%', fontsize=10)
plt.text(0.05, max(Y_cleaned)*0.80, f'MSE: {mse:.4f}', fontsize=10)

plt.grid()
plt.tight_layout()
plt.show()



# Create a list to store results for both CSV files
summary_results = []

# Append results for the first CSV file (A.B. 5.csv)
summary_results.append({
    'File Name': 'A.B..csv',
```

```
    'Multiple R': multiple_r,
    'Average Percentage Error': average_percentage_error,
    'Mean Squared Error': mse
})

# Create a list to store coefficients for both CSV files
coefficients_results = []

# Append coefficients for the first CSV file (A.B. 5.csv)
coefficients_results.append({
    'File Name': 'A.B..csv',
    **model.params.to_dict()  # Add regression coefficients as key-value pairs
})
```

The Python code used for normalisation.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler

# Load the datA.B.K
file_path = 'C:\\Users\\aziz\\Desktop\\air\\4 Clean for regression\\1 A.B. MA single shifted.csv'
data = pd.read_csv(file_path)

# Compute Temperature Difference
data['Temperature_Difference'] = data['Temperature_A8'] - data['Temperature_B8']

# Compute Air Change Rate
co2_current = data['CO2']
co2_previous = data['CO2'].shift(1)
air_change_rates = []

for index in range(len(data)):
    K2 = co2_previous[index] if index > 0 else None
    K3 = co2_current[index]
```

```
    if K3 is None or K2 is None:
        air_change_rate = None
    elif K3 >= K2 + 64:
        air_change_rate = (K2 + 2*0.343 * 20 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 56:
        air_change_rate = (K2 + 2*0.343 * 8 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 48:
        air_change_rate = (K2 + 2*0.343 * 7 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 40:
        air_change_rate = (K2 + 2*0.343 * 6 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 32:
        air_change_rate = (K2 + 2*0.343 * 5 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 24:
        air_change_rate = (K2 + 2*0.343 * 4 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 16:
        air_change_rate = (K2 + 2*0.343 * 3 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2 + 8:
        air_change_rate = (K2 + 2*0.343 * 2 / 68063 * 1000000 - K3) / K2
    elif K3 >= K2:
        air_change_rate = (K2 + 2*0.343 / 68063 * 1000000 - K3) / K2
    else:
        air_change_rate = -(K3 - K2) / K2

    air_change_rates.append(air_change_rate)

data['Air Exchange Rate'] = air_change_rates

# Remove first row since it has NaN for Air Exchange Rate
data = data.iloc[1:].reset_index(drop=True)

# Columns to normalize
norm_columns = ['PM 2.5_A8', 'Relative Humidity_A8', 'Barometric Pressure', 'Relative Humidity_B8', 'PM
2.5_B8_Shifted', 'Temperature_Difference', 'Air Exchange Rate']

# Store original data for plotting before normalization
original_data = data[norm_columns].copy()
```

```
# Normalize the data
scaler = MinMaxScaler(feature_range=(0, 1))
data[norm_columns] = scaler.fit_transform(data[norm_columns])

# Function to plot histograms with trendlines
def plot_histograms(original, normalized, columns):
    for col in columns:
        plt.figure(figsize=(12, 5))

        # Before normalization
        plt.subplot(1, 2, 1)
        sns.histplot(original[col], kde=True, bins=30, color='blue')
        plt.title(f'A.B. Before Normalization: {col}')
        plt.xlabel(col)
        plt.ylabel('Frequency')

        # After normalization
        plt.subplot(1, 2, 2)
        sns.histplot(normalized[col], kde=True, bins=30, color='red')
        plt.title(f'A.B. After Normalization: {col}')
        plt.xlabel(col)
        plt.ylabel('Frequency')

        plt.tight_layout()
        plt.show()

# Plot histograms before and after normalization
plot_histograms(original_data, data[norm_columns], norm_columns)

# Select relevant columns
final_columns = ['Timestamp_A8', 'Timestamp_A8_Shifted', 'PM 2.5_A8', 'PM 2.5_B8_Shifted',
'Temperature_Difference', 'Barometric Pressure', 'Relative Humidity_A8', 'Relative Humidity_B8', 'Air Exchange
Rate', 'CO2', 'Temperature_A8', 'Temperature_B8']
data = data[final_columns]

# Save the final datA.B.K
data.to_csv('A.B. norm.csv', index=False)
```

```
print("Processed data saved successfully. Histograms with trendlines generated.")
```

References

Argyropoulos, C. D., Hassan, H., Kumar, P., & Kakosimos, K. E. (2020). Measurements and modeling of Particulate Matter Building Ingress during a severe dust storm event. *Building and Environment*, *167*, 106441. https://doi.org/10.1016/j.buildenv.2019.106441

Arvanitis, A., Kotzias, D., Kephalopoulos, S., Carrer, P., Cavallo, D., Cesaroni, G., Brouwere, K. D., de Oliveira-Fernandes, E., Forastiere, F., & Fossati, S. (2010). The index-pm project: Health risks from exposure to indoor particulate matter. *Fresenius Environmental Bulletin*, *19*, 2458–2471.

Bai, L., He, Z., Li, C., & Chen, Z. (2019). Investigation of yearly indoor/outdoor PM 2.5 levels in the perspectives of health impacts and air pollution control: Case study in Changchun, in the northeast of China. *Sustainable Cities and Society*, *53*, 101871. [https://www.sciencedirect.com/science/article/abs/pii/S2210670719320141?via%3Dihub]

Bekierski, D., & Kostyrko, K. B. (2021). The Influence of Outdoor Particulate Matter PM 2.5 on Indoor Air Quality: The Implementation of a New Assessment Method. *Energies, 14*, 6230. [https://doi.org/10.3390/en14196230]

Bennett, D. H., & Koutrakis, P. (2006). Determining the infiltration of outdoor particles in the indoor environment using a dynamic model. *Journal of Aerosol Science*, *37*(6).[https://www.sciencedirect.com/science/article/abs/pii/S002185020500113]

Chan, A. T. (2002). Indoor–outdoor relationships of particulate matter and nitrogen oxides under different outdoor meteorological conditions. *Atmospheric Environment,* *36*, 1543–1551. [https://www.sciencedirect.com/science/article/abs/pii/S135223100100471X?via%3Dihub]

Chen, C., & Zhao, B. (2011). Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor. *Atmospheric Environment*,

45, 275–288. [https://www.sciencedirect.com/science/article/abs/pii/S1352231010008241?via%3Dihub]

Chen, C., Zhao, B., & Weschler, C. J. (2012). Assessing the Influence of Indoor Exposure to "Outdoor Ozone" on the Relationship between Ozone and Short-term Mortality in U.S. Communities. *Environmental Health Perspectives*, *120*, 235–240. [https://ehp.niehs.nih.gov/doi/10.1289/ehp.1103970]

Diapouli, E., Chaloulakou, A., & Koutrakis, P. (2013). Estimating the concentration of indoor particles of outdoor origin: A review. *Journal of the Air & Waste Management Association*, *63*, 1113–1129. [https://www.tandfonline.com/doi/full/10.1080/10962247.2013.791649]

Dols, W. S., & Polidoro, B. J. (2015). CONTAM user guide and program documentation (Technical Note 1887, Version 3.2). Gaithersburg, MD: National Institute of Standards and Technology.

Ecokarta. Company information: АО "Астана-Энергия" (2017). Retrieved from https://ecokarta.kz/company/show/114

He, C., Morawska, L., & Gilbert, D. (2005). Particle deposition rates in residential houses. *Atmospheric Environment*, *39*, 3891–3899. [https://www.sciencedirect.com/science/article/abs/pii/S1352231005002815]

Hossain, M. S., Che, W., Frey, H. C., & Lau, A. K. H. (2021). Factors affecting variability in infiltration of ambient particle and gaseous pollutants into home at urban environment. *Building and Environment*, *206*, 108351. [https://www.sciencedirect.com/science/article/pii/S0360132321007484]

Huang, L., Hopke, P. K., Zhao, W., & Li, M. (2015). Determinants on ambient PM 2.5 infiltration in non-heating season for urban residences in Beijing: Building characteristics, interior surface coverings and human behavior. *Atmospheric Pollution*

Research, *6*(6), 1046-1054. [https://www.sciencedirect.com/science/article/pii/S1309104215000161]

Jones, A. P. (1999). Indoor air quality and health. *Atmospheric Environment, 33*(30), 4535-4564. https://doi.org/10.1016/S1352-2310(99)00272-1

Kalimeri, K. K., Bartzis, J. G., Sakellaris, I. A., & Fernandes, E. D. O. (2019). Investigation of the PM 2.5, NO2 and O3 I/O ratios for office and school microenvironments. *Environmental Research*, *179*, 108791. [https://www.sciencedirect.com/science/article/abs/pii/S0013935119305882?via%3Dihub]

Kerimray, A., Bakdolotov, A., Sarbassov, Y., Inglezakis, V., & Poulopoulos, S. (2018). Air pollution in Astana: analysis of recent trends and air quality monitoring system. *Materials Today: Proceedings, 5*(11), 22749-22758.

Kim, J., Son, J., & Koo, J. (2024). Hybrid models of machine-learning and mechanistic models for indoor particulate matter concentration prediction. *Journal of Building Engineering*, *86*, 108836. [https://www.sciencedirect.com/science/article/pii/S2352710224004042]

Krasnov, H., Katra, I., & Friger, M. D. (2015). Insights into indoor/outdoor PM concentration ratios due to dust storms in an arid region. *Atmosphere, 6*(7), 879–890. [https://www.mdpi.com/2073-4433/6/7/879]

Lv, Y., Wang, H., Wei, S., Zhang, L., & Zhao, Q. (2017). The correlation between indoor and outdoor particulate matter of different building types in Daqing, China. *Procedia Engineering*, *205*, 360–367. [https://www.sciencedirect.com/science/article/pii/S1877705817345514?via%3Dihub]

Ma, Z., Huang, J., Wang, X., Wei, Y., & Huang, L. (2023). Estimation of infiltration efficiency of ambient PM 2.5 in urban residences of Beijing during winter. *Urban Climate*, *52*, 101677. [https://www.sciencedirect.com/science/article/pii/S2212095523002717?via%3Dihub]

MacNeill, M., Wallace, L., Kearney, J., Allen, R. W., Van Ryswyk, K., Judek, S., Xu, X., & Wheeler, A. (2012). Factors influencing variability in the infiltration of PM 2.5 mass and its components. *Atmospheric Environment*, *61*, 518-532. [https://www.sciencedirect.com/science/article/pii/S1352231012006772]

Meng, Q. Y., Turpin, B. J., Jong, H. L., Polidori, A., Weisel, C. P., Morandi, M., Colome, S., Feng, Z. J., Thomas, S., & Arthur, W. (2007). How does infiltration behavior modify the composition of ambient PM 2.5 in indoor spaces? An analysis of RIOPA data. *Environmental Science & Technology, 41*(21).

Molina Rueda, E., Carter, E., L'Orange, C., Quinn, C., & Volckens, J. (2023). Size-resolved field performance of low-cost sensors for Particulate Matter Air Pollution. *Environmental Science & Technology Letters*, *10*(3), 247–253. https://doi.org/10.1021/acs.estlett.3c00030

Mukhtarov, R., Ibragimova, O. P., Omarova, A., Tursumbayeva, M., Tursun, K., Muratuly, A., Karaca, F., & Baimatova, N. (2023). An episode-based assessment for the adverse effects of air mass trajectories on PM 2.5 levels in Astana and Almaty, Kazakhstan. *Urban Climate*, 49, 101541. https://doi.org/10.1016/j.uclim.2023.101541

Nadali, A., Arfaeinia, H., Asadgol, Z., & Fahiminia, M. (2020). Indoor and outdoor concentration of PM10, PM 2.5 and PM1 in residential building and evaluation of negative air ions (NAIs) in indoor PM removal. *Environmental Pollution and Bioavailability, 32*, 47–55. [https://www.tandfonline.com/doi/full/10.1080/26395940.2020.1728198]

Orch, Z. E., Stephens, B., & Waring, M. S. (2014). Predictions and determinants of size-resolved particle infiltration factors in single-family homes in the U.S. *Building and Environment*, *74*. [https://www.sciencedirect.com/science/article/pii/S0360132314000092]

Ott, W., Wallace, L., & Mage, D. (2000). Predicting particulate (PM10) personal exposure distributions using a random component superposition statistical model. *Journal of the Air & Waste Management Association, 50*(8), 1390-1406.

Park, S. B., Park, J.-H., Jo, Y. M., Song, D., Heo, S., Lee, T. J., Park, S., & Koo, J. (2022). Development and validation of a dynamic mass-balance prediction model for indoor particle concentrations in an office room. *Building and Environment*, *207*(Part A), 108465. [https://www.sciencedirect.com/science/article/pii/S0360132321008611]

PN-EN 12341: 2014-07. (2014). *Atmospheric Air—Standard Gravimetric Measurement Method for Determining the Mass Concentrations of PM10 or PM 2.5 Suspended Dust*. Polski Komitet Normalizacyjny: Warszawa, Poland.

San Jose, R., & Perez-Camanyo, J. L. (2023). Modelling infiltration rate impacts on indoor air quality. *International Journal of Thermofluids*, *17*, 100284. [https://www.sciencedirect.com/science/article/pii/S266620272300006X]

Scibor, M., Bokwa, A., & Balcerzak, B. (2020). Impact of wind speed and apartment ventilation on indoor concentrations of PM10 and PM 2.5 in Kraków, Poland. *Air Quality, Atmosphere & Health*, *13*, 553–562. [https://link.springer.com/article/10.1007/s11869-020-00816-8]

Schreck, C., Rouchier, S., Foucquier, A., Machefert, F., & Wurtz, E. (2024). In situ air change rate estimation from metabolic CO_2 measurement: Summer experimental campaign in a single-family test house. *Building and Environment, 259*, 111646. https://doi.org/10.1016/j.buildenv.2024.111646

Thatcher, T. L., & Layton, D. W. (1995). Deposition, resuspension, and penetration of particles within a residence. *Atmospheric Environment, 29*, 1487–1497. [https://www.sciencedirect.com/science/article/abs/pii/135223109500016R?via%3Dihub]

TSI Incorporated. (2023). AirAssure™ Indoor Air Quality Monitor [Brochure]. Retrieved from

[https://tsi.com/getmedia/43145cbb-d386-40be-b307-3fe48f0cb4b3/AirAssure-IAQ_A4_5002719_RevB_Web?ext=.pdf]

TSI Incorporated. (2023). BlueSky™ Air Quality Monitor Models 8143 and 8145 [Brochure]. Retrieved from [https://tsi.com/getmedia/4c72a030-4585-4df4-a7c8-596aa1994734/BlueSky-Air-Quality-Monitor_A4_5002492_RevB_Web?ext=.pdf]

Tursumbayeva, M., Muratuly, A., Baimatova, N., Karaca, F., & Kerimray, A. (2023). Cities of Central Asia: New hotspots of air pollution in the world. *Atmospheric Environment, 309*, 119901. https://doi.org/10.1016/j.atmosenv.2023.119901

Walker, E. S., Stewart, T., & Jones, D. (2023). Fine particulate matter infiltration at Western Montana residences during wildfire season. *Science of The Total Environment*, *896*, 165238. [https://www.sciencedirect.com/science/article/pii/S0048969723038615]

Walton, G. N., & Dols, W. S. (2005). CONTAM 2.4 User Guide and Program Documentation. Gaithersburg, MD: National Institute of Standards and Technology.

Wan, Y., Chen, C., Wang, P., Wang, Y., Chen, Z., & Zhao, L. (2015). Infiltration Characteristic of Outdoor Fine Particulate Matter (PM 2.5) for the Window Gaps. *Procedia Engineering, 121*, 191-198. https://doi.org/10.1016/j.proeng.2015.08.1050

Wichmann, J., Lind, T., Nilsson, M. A.-M., & Bellander, T. (2010). PM 2.5, soot and NO2 indoor–outdoor relationships at homes, pre-schools and schools in Stockholm, Sweden. *Atmospheric Environment*, *44*(36), 4536-4544. [https://www.sciencedirect.com/science/article/pii/S1352231010006977]

Wong, N. H., & Huang, B. (2004). Comparative study of the indoor air quality of naturally ventilated and air-conditioned bedrooms of residential buildings in Singapore. *Building and Environment*, *39*(9), 1115–1123.