Modelling Infiltration of Ambient $PM_{2.5}$ in Higher Education Buildings: An Institution Building Stock Indoor Air Quality Model for Assessing Exposure Risks

Tha'er Abdalla

BArch, M.Sc Arch Eng

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

The University of Sheffield

Faculty of Social Sciences / School of Architecture

Submission Date

July 2023

Abstract

Air pollution has been identified as one of the leading causes of morbidity and mortality worldwide. The current trend is predicted to continue until 2040 by the International Energy Agency (IEA) forecasts. It is estimated that ambient fine particles ($PM_{2.5}$) caused 103.1 million disability-adjusted life-years (DALYs) in 2015. As indoor pollutant concentrations, including $PM_{2.5}$, can be even higher than those outdoors, the indoor environments of homes and workplaces may significantly impact population exposure. This doctoral thesis presents a study of indoor air quality in higher education institution (HEI) buildings. In the UK, most universities are located in high-density urban built areas, and air pollution from urban traffic and other sources is the most significant contributor to poor indoor air quality (IAQ). Since people spend long hours indoors working in HEI buildings, there is a concern about chronic exposure to indoor air pollutants such as $PM_{2.5}$.

The main challenge addressed in this research is the high level of heterogeneous characteristics observed in HEI buildings that require many input parameters in developing building stock IAQ models to inform planning and design for better air quality. Robust HEI building stock IAQ models are required for estimating university population exposure to indoor $PM_{2.5}$ from outdoor sources throughout the year. This thesis shows how such estimations can be achieved reliably by a reduced set of input parameters at an HEI building stock level. The IAQ modelling focuses on the annual heating season (November-April) when higher outdoor $PM_{2.5}$ levels often appear during winter in the UK. Based on the outputs of infiltrated $PM_{2.5}$ concentrations, the HEI stock IAQ model is applied to evaluate the impact of increasing the building envelope airtightness (Q_{50}) measure on population exposure.

Five buildings from the University of Sheffield (UoS) were selected and modelled in CONTAM and EnergyPlus using available data sources, such as the Estates and Facilities Management (EFM) and local building regulations and guidelines. The buildings are modelled with multiple Q_{50} values ranging between 3 – 13 $m^3/h/m^2$ to generate indoor $PM_{2.5}$ concentrations due to infiltration at a zone/room level (N =2,729 zones) during the heating season. An analytical framework employing sensitivity analysis is used to examine correlations, regressions, and sample comparisons to identify the input parameters influential on the concentrations of infiltrated $PM_{2.5}$ during the heating season. The advantage of utilising correlation coefficient tests lies in their ability to assess the significance of input variables through the associated *p*-values. The result of the sensitivity analysis shows the top five input parameters influencing infiltrated $PM_{2.5}$ concentrations: (1) variability in building envelope airtightness Q_{50}, (2) zone infiltration air change rates (ACH_{INF}), (3) indoor-outdoor temperature difference (ΔT), (4) wind speed (v), and (5) the area of exposed façade to zone volume ratio (A_{ef}:V_z).

To allow for rapid assessments of the Q_{50} factor on the concentrations of infiltrated $PM_{2.5}$ of existing or proposed UoS buildings, metamodels for the heating season were further developed. Informed by the latest literature, the five input parameters were examined systematically in three machine learning (ML) regression algorithms: *Generalised Additive Models* (GAM), *Random Forest* (RF), and *Extreme Gradient Decision Trees* (XGB). In terms of the best model performance among the three, the XGB metamodel achieves an R^2 value higher than 0.91 for the heating season concentrations of infiltrated $PM_{2.5}$ on the training (N =1,910), testing (N =819), and evaluation (N=40) datasets with a model prediction accuracy greater than 90%.

As a test case, population exposures to indoor $PM_{2.5}$ in a selected UoS building were estimated by a microenvironment modelling approach to evaluate- the effects of changing the airtightness of the building envelope. To directly compare the indoor concentrations with the World Health Organisation's annual exposure limit of 10 µg/m3, the concentrations of indoor $PM_{2.5}$ predicted by the metamodel due to infiltration are combined with the simulated non-heating season concentrations (May-October). The findings reveal that population exposure to indoor $PM_{2.5}$ originating from outdoor sources experiences an 11% and 32% reduction when the Q_{50} values for the buildings are set at 7 and 3 ($m^3/h/m^2$), respectively.

The thesis contributes to the existing knowledge by: (i) developing a novel modelling framework for assessing indoor air quality (IAQ) of HEI buildings at an institutional level by combining physics-based modelling and ML-based metamodelling; (ii) identifying the most influential input parameters impacting the population's exposure to infiltrated $PM_{2.5}$ in a given HEI context, and (iii) demonstrating how an HEI stock IAQ model can be utilised to inform and evaluate the effects of planning and design interventions (e.g., Q_{50} modifications) on IAQ.

Acknowledgements

Writing a thesis and earning a doctorate are lengthy and arduous tasks that one cannot accomplish single-handedly. Therefore, as a first and foremost thank you, I would like to express my gratitude to Almighty Allah, who has blessed me, given me strength, and arranged the circumstances that have enabled me to complete my PhD.

No words of thanks can sum up the gratitude that I owe to ***Dr Chengzhi Peng***, my mentor and research advisor extraordinaire. It would not have been possible for this research and dissertation to be completed without his advice, guidance, expertise, and encouragement. I would like to express my sincere gratitude to ***Dr Abigail Hathway*** and ***Dr Benjamin Jones*** for their meticulous examination of my PhD thesis and their valuable comments that have significantly improved the quality of my work.

I would like to express my gratitude to my wife and daughter for putting up with an absent husband and father. As I have spent my time and energy pursuing goals that have taken me away from her and the family, ***Jakleen*** has consistently been unfailingly supportive – and has carried the burdens that have fallen upon her shoulders. I attribute much of my ability to work during most of our ten years of marriage to her. There would not have been a possibility for me to achieve my educational goals without my family's support, encouragement, and understanding. I wish there was space on my diploma to include the names of my wife and daughter, ***Jakleen*** and ***Juana***.

I would be remiss if I did not acknowledge and thank my ***parents*** for their unconditional love, support, and encouragement throughout my journey, including their contribution to funding my studies. My ***brother*** and ***sister*** for their continuous words of encouragement.

My gratitude reaches out to all those I've collaborated with internally and externally, ***Craig Wootton***, the Building Services Surveyor from the University of Sheffield's Estates and Facilities Management for providing building data, ***Alistair McLean***, Curator of Natural Sciences from the Weston Park Museum for providing the weather data, ***W. Stuart Dols*** from the National Institute of Standards and Technology for technical support, and my colleague ***Ibrahim Ozdemir*** from the School of Architecture for providing field measurements for this research.

As a final thank you to everyone who helped me on this memorable journey directly or indirectly.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACH	Air Change Rates per Hour
ACH_{INF}	Infiltration Air Change Rates per Hour
ADC	Academic Development Centre
ADF	Approved Document F
AMY	Actual Meteorological Year
ANNs	Artificial Neural Networks
ASHRAE	American Society for Heating, Refrigerating, and Air Conditioning Engineers
ASTM	American Society for Testing and Materials
AT	Arts Tower
BH	Barber House Building
CAD	Computer Aided Design
CaRB	CArbon Reduction Model
CFD	Computational Fluid Dynamics
CHAARM	Chilean Housing Archetype AiR Quality Model
CIBSE	Chartered Institution of Building Services Engineers
COMEAP	Committee on the Medical Effects of Air Pollutants
CSV	Comma Sperated Values File
CV	Cross Validation
DALYs	Daily Adjusted Life Year
DAQI	Daily Air Quality Index
DEFRA	Department for Environment, Food, and Rural Affairs

DHN	District Heating Network
EFM	Estates and Facilities Management
EHS	English Housing Survey
ELA_4	Effective Leakage Area at 4Pa pressure Difference
GAM	Generalised Additive Model
GCM	Generic Contaminant Model
GLM	Generalised Linear Model
HEI	Higher Education Institution
HESA	Higher Education Statistics Agency
HVAC	Heating, Ventilation, and Air Conditioning
I/O	Indoor/Outdoor Pollutant Ratio
IAQ	Indoor Air Quality
ICoSS	Interdisciplinary Centre of Social Sciences
IDF	EnergyPlus Input Data File
LOOCV	Leave-One-Out Cross-Validation
ML	Machine Learning
MLR	Multiple Linear Regression
NDCSM	Non-Domestic Carbon Scenario Model
N-DEEM	Non-domestic Energy and Emissions Model
NICE	National Institute for Health and Care Excellence
NIST	National Institute for Standards and Technology
OECD	The Organisation for Economic Cooperation and Development
PCR	Principle Component Regression
PLS	Partial Least Squares
PM	Particulate Matter
$PM_{2.5}$	Particulate Matter with diameter less than 2.5 mircons (μm)
R	Pearson Correlation Coefficient
R^2	Coefficient of Determination

RC	Regent Court Building
RFR	Random Forest Regression
RMSE	Root Mean Squared Error
SCU	Self-Contained Unit
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TMY	Typical Meteorological Year
UoS	University of Sheffield
VIF	Variance Inflation Factor
VOA	Valuation Office Agency
VOCs	Volatile Organic Compunds
WHO	World Health Organisation
XGB	eXtreme Gradient Boosting

Chapter 1 Introduction

1.1 The '*Ambient'* Air

Several factors affect the composition and concentration of pollutants in the ambient air, such as topography, wind conditions, sources of pollution in the area, the time of day, and the season. According to Chapman (2007), the ambient air is deemed contaminated if a substance is found where it should not be or if its concentration exceeds the background level. On the other hand, outdoor air pollution refers to the presence of one or more "contaminants" in the atmosphere at levels and durations exceeding natural limits and can lead to a negative impact on the environment and health (Chapman, 2007). Particulate Matter (PM), Ozone (O_3), Carbon Monoxides (CO), Sulphur Dioxides (SO_2), and Nitrogen Dioxides (NO_2) are among these pollutants that are used for declaring states of ambient air quality emergency in the UK when their limits are exceeded. In recent years, most studies have focused on the impact of air pollution with PM, especially those with a diameter smaller than 2.5 microns (μm), because of their ability to penetrate lung tissue and trigger local and systemic effects (Nemmar et al., 2013). In light of its varied and severe effects on human health of all ages and genders, air pollution (outdoor and indoor) has been described as one of the "great killers of our age" and as a "major threat to health" (Venkatesan, 2016).

The effect of PM on population health is well documented and is now considered one of the most significant causes of mortality and morbidity globally. One example is the association of lung cancer and respiratory and cardiovascular diseases with exposure to fine particles ($PM_{2.5}$). Several scientific studies have shown that excessive exposure to high $PM_{2.5}$ concentrations reduces the expected lifespan of humans by one to five years (Apte et al., 2018; Cserbik et al., 2020). In 2013, the World Health Organisation (WHO) identified PMs as the leading cause of human cancer (WHO, 2013). Additionally, the health effects of this exposure on the global population were

estimated at 103.1 million Disability-Adjusted Life Years (DALYs) [1] in 2015 (Burnett et al., 2018; Cesaroni et al., 2013; Cohen et al., 2017). In 2012, the WHO estimated that more than 92% of the world's population lives in areas with high pollution levels. Moreover, the WHO revealed that outdoor air pollution caused approximately three million deaths worldwide, with 6.5 million deaths (11.6% of all global deaths) resulting from combined indoor and outdoor air pollution (WHO, 2012).

Although there has been a considerable improvement in outdoor air quality in the UK over the past decade, some areas still exhibit high levels of air pollution. In 2018, 115 of the 317 (36%) local authorities had unsafe levels of $PM_{2.5}$ above the WHO's annual recommended concentration level (DEFRA, 2020). The report adds that 214 local authorities had excessive roadside levels, and 55% of the monitored locations exceeded WHO guidelines. A study by Public Health England found that $PM_{2.5}$ air pollution was responsible for 5% of deaths among those over 30 years of age (Public Health England, 2018). That is 1 in every 20 deaths. The figures are based on data collected between 2010 and 2017. Local areas in some parts of the country are even higher than this, with parts of London reaching as high as 7%. The Committee on the Medical Effects of Air Pollutants estimates that outdoor air pollution contributes to approximately 28,000-36,000 premature deaths annually in the UK (COMEAP, 2009).

1.2 Challenges of Indoor Air Quality in a Higher Education Setting

Worldwide, the energy consumed on Higher Education Institution (HEI) premises has been continually rising due to increased student and staff populations and the expansion of energy-use facilities. Thus, an increase in energy consumption is accompanied by an increase in energy wastage and colossal energy cost burdens on HEI premises management. Furthermore, HEIs are significant institutions that can provide leadership in delivering energy transition to net zero via their decarbonisation strategies and as "living laboratories" to advance clean technologies. Hence, it has been recognised by many HEIs that the need to understand and monitor energy consumption patterns on campuses to ensure improved environmental performance and sustainability (Oyedepo et al., 2021).

[1] Disability-Adjusted Life Year (DALY) is the most commonly used measure for quantifying mortality and morbidity to a given disease or risk factor. It combines the years of life lost due to disability with the years of life lost due to death (A. Chen et al., 2015)

Daily running of HEI premises involves many building users and facilities, making potential environmental degradation caused by intensive energy use a significant concern. Hence, energy consumption in HEIs should be effectively managed to reduce waste and environmental impacts. Therefore, HEIs in the UK must implement energy management strategies and programs to align campus operations with sustainability goals. One example is the University of Sheffield (UoS) Energy Strategy developed to achieve net-zero carbon emissions by 2030 (Arup, 2012). HEI energy strategies often include information on the current energy use patterns and provide information on where building interventions, such as fabric upgrades and increasing the airtightness of building envelopes, are planned and scheduled.

Here is a challenge to address: changes made to an HEI building's envelope to reduce energy demand may negatively impact indoor air quality (IAQ) (Figure 1.1). The emphasis on airtightness is of particular interest and concern since people, on average, spend 85-90% of their time indoors (ECA, 2003; Klepeis, Nelson, Ott, Robinson, Tsang, Switzer, & Behar, 2001; Schweizer et al., 2007). In addition, the building envelope with enhanced airtightness may raise building users' exposure to indoor air pollution (Smith et al., 2016; Vardoulakis, 2009). Therefore, HEI energy strategies must simultaneously evaluate and assess the energy demand and IAQ.

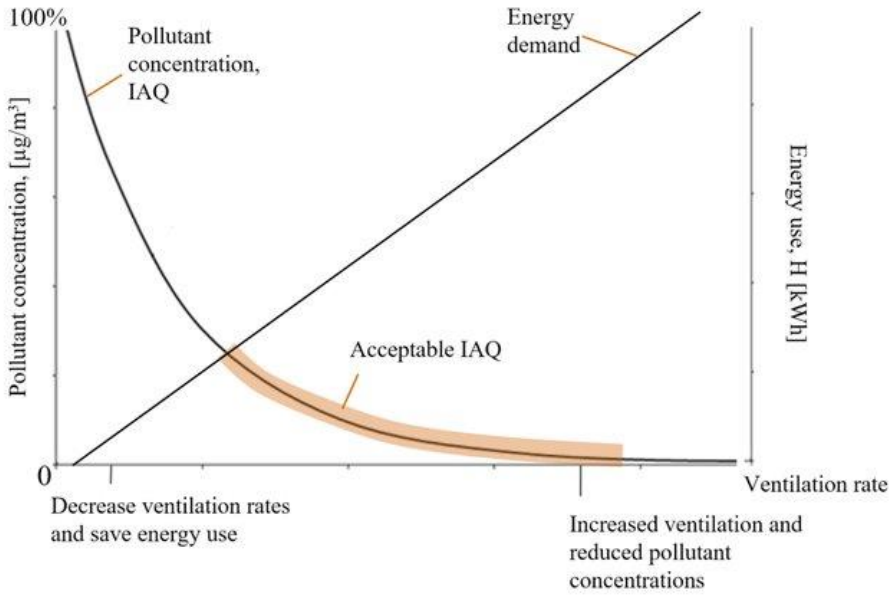


Figure 1.1: The relationship between energy demand, IAQ and ventilation rates in buildings (Molina, 2019)

1.3 Indoor Air Quality in HEI Buildings

Seen as a specific type of institutional building stock, HEI buildings offer complex indoor environments where staff and students work, learn and interact. The total number of HE students in the UK stood at 2.9 million in 2020/21, an increase of 9% from 2019/20 (Bolton, 2022), with 225,000 academic/professional staff. With such a large population, the risk of exposure to indoor air pollution in HEIs should be assessed regularly to inform mitigation strategies and actions. However, there has been relatively little research on indoor air quality (IAQ) in HEI buildings, with most studies focused on factors affecting HEI indoor environmental quality and perception-based measures (Lee et al., 2012; Norbäck et al., 2013; Norbäck & Nordström, 2008; Sarbu & Pacurar, 2015). Moreover, limited studies examined PM exposure in HEI premises and its associated health impacts (Gaidajis & Angelakoglou, 2009; Norbäck et al., 2013). However, these studies were limited in scale and only looked into IAQ in specific room types (e.g., computer rooms). The study by (Elliot et al., 2000) found that even individuals working within the same building will be exposed to varying levels of PM based on the patterns of their daily activities.

Several factors could affect the IAQ in HEI buildings, including the design of the building, its use, location, and the local environmental conditions, making exposure assessment to indoor pollution a complex task. In particular, this is due to (a) the high heterogeneity often observed in HEI building design and construction, (b) the characteristics, composition, and behaviour of building users, and (c) the uncertainties in the environmental parameters that influence building and occupant behaviours. Additionally, these three systems – ambient environment, building, and occupant- constantly interact, and their characteristics and parameters must be examined to account for the interactions. Therefore, more research is needed to better understand the distributions of $PM_{2.5}$ within and between HEI buildings and its impact on population exposure to inform HEI estate planning and design.

1.4 Research Aim and Objectives

This research aims to develop a new modelling capability for estimating population exposure to infiltrated $PM_{2.5}$ at an HEI building stock level throughout the year. The modelling will focus on the heating season (November-April) to determine the concentration of infiltrated $PM_{2.5}$ in indoor environments in HEI building stocks and to evaluate the impact of increasing envelope

airtightness on population exposure. In the UK, outdoor $PM_{2.5}$ concentrations are typically higher during winter due to increased heating systems and transportation use. Additionally, during winter, a phenomenon known as temperature inversion may occur, in which the lower atmosphere is cooler than the upper atmosphere (Gramsch et al., 2014). This leads to the trapping of pollutants, including $PM_{2.5}$, close to the ground by a layer of warm air, resulting in elevated concentrations.

Furthermore, the HEI academic year runs from October to June, with higher occupancy levels in the heating season than in the non-heating season. Previous research suggests that while natural ventilation contributes to increased air exchange rates during the non-heating season (Park et al., 2014), the influence of lower outdoor $PM_{2.5}$ concentrations on IAQ is less pronounced during this season. Therefore, in this study, the efforts to enhance the IAQ using $PM_{2.5}$ as an indicator will prioritise measures that reduce the infiltration of $PM_{2.5,}$ particularly in the heating season. The modelling will focus on the heating season (November-April) to determine the concentration of infiltrated $PM_{2.5}$ in indoor environments in HEI building stocks and to evaluate the impact of increasing envelope airtightness Q_{50} on population exposure.

The research is expected to lay the foundation of novel applications supporting HEI planning and design for better IAQ. In particular, with the heterogeneity and complexity often observed in HEI buildings, variations of indoor $PM_{2.5}$ concentration levels are anticipated from building to building and within each building. To address this challenge, the research tackled the following questions:

1. What key parameters influence the indoor air quality of HEI buildings?
2. Given an existing HEI building stock in its urban context, how can we estimate population exposure to indoor $PM_{2.5}$ outdoor sources at an institutional stock level?
3. How can the building and data science of HEI building stock IAQ modelling inform planning and design for better indoor air quality?

The research questions led to the following six objectives:

Objective (Obj-1): To identify existing data sources that can be used to describe components of the HEI institution stock. Then decompose multiple buildings sampled from an HEI stock into a structured cohort of individual spaces or rooms.

Objective (Obj-2): To develop detailed multi-zone IAQ models of the selected buildings to estimate the concentrations of infiltrated $PM_{2.5}$. (Note: In the absence of indoor air pollution field measurements, the primary information source will be the coupled simulation platform between CONTAM-EnergyPlus.).

As affected by the COVID-19 pandemic restriction, it was not possible to validate the IAQ models developed in this research by following the ASTM guide for validation. The field measurement phase of the research was initially planned to occur between January 2020 and January 2021; however, due to the UK Government's restrictions (i.e., nationwide lockdowns), field measurement data collection was not possible during this period. It was also challenging to gather occupancy-related data through surveys since access to every university building was severely restricted.

Objective (Obj-3): To explore the relationship between indoor $PM_{2.5}$ concentrations and multiple building and environmental variables by applying a sensitivity analysis framework. Here, the outcome of the framework is a reduced set of parameters ranked by their effect on the simulated $PM_{2.5}$ concentrations.

Objective (Obj-4): To develop a data-driven metamodel from a reduced set of input parameters to capture the spatial variation of infiltrated $PM_{2.5}$ concentrations.

Objective (Obj-5): To improve the interpretability of the developed metamodel by applying an interpretation framework (SHAP). Here, the contribution of each explanatory variable to the infiltrated $PM_{2.5}$-concentrations will be calculated.

Objective (Obj-6): To estimate the individual and population exposures to indoor $PM_{2.5}$ according to time-activity groups comparable to IAQ standards or IAQ-related health metrics.

1.5 Contribution to Knowledge

Firstly, a hybrid bottom-up approach is developed to model multiple buildings selected from an HEI building stock as a structured cohort of individual spaces or rooms. The model resolution set at a room level allows for sensitivity analyses of possible relations between measured or simulated infiltrated $PM_{2.5}$ concentrations and the built and environmental characteristics. Second, it identifies the most influential input parameters impacting the population's exposure to infiltrated $PM_{2.5}$ in a given HEI context. Third, it demonstrates how an HEI stock IAQ model can be utilised

to inform and evaluate the effects of planning and design interventions (e.g., Q_{50} modifications) on IAQ. Fourth, the research addresses the need for more detailed studies on developing non-domestic building stock IAQ models. Based on the University of Sheffield estates stock, the thesis presents a UoS Stock IAQ model. Finally, this thesis identifies several areas of data paucity hindering HEI AQ planning and design that requires attention and investment.

1.6 Thesis Outline

This thesis comprises nine chapters reporting the research tasks and results of fulfilling the project objectives.

Chapter 1 Introduction presents the motivation behind this research, the research questions, the aim and objectives, and the contribution to existing knowledge.

Chapter 2, Literature Review, presents a comprehensive review of existing literature on IAQ. The review was dedicated to determining the assumptions, methods and techniques used to develop building stock IAQ models. This review concludes by examining the use of data-driven methods in modelling indoor air quality in different buildings.

Chapter 3, Research Methodology and Data Sources, describes the methodological framework proposed to estimate the infiltrated $PM_{2.5}$ concentrations across selected buildings from an existing HEI stock. Then, existing data sources describing the HEI institution stock components are identified.

Chapter 4, Building Physics-Based Modelling, presents the development of detailed multi-zone models of the sampled UoS buildings in CONTAM and EnergyPlus to estimate infiltrated $PM_{2.5}$ concentrations. In addition, the coupled simulation results are presented in this chapter.

Chapter 5, Predictive Models: Metamodelling Roadmap, presents a framework for developing predictive data-driven models based on the outputs of the coupling simulation. The chapter presents a sensitivity analysis framework to identify the most influential inputs to infiltrated $PM_{2.5}$ concentrations. Then, the roadmap for the development of predictive models is presented.

Chapter 6, Sensitivity Analysis and Model Prediction, presents the results of the sensitivity analysis framework for annual levels of infiltrated $PM_{2.5}$ concentrations. Then, an in-depth analysis of all variables in the dataset is conducted to identify the essential variables associated

with the infiltrated $PM_{2.5}$ concentrations. Finally, the outcome of the SA was used as input to each ML algorithm.

Chapter 7, Microenvironmental Modelling of Population Exposure, presents a microenvironment modelling approach to estimate the average Personal Exposure (E_i) to the infiltrated $PM_{2.5}$ and the average Population-Weighted Exposure (PWE) to the infiltrated $PM_{2.5}$ for different microenvironments across similar time-activity groups.

Chapter 8, Discussion, Limitation, and Future Considerations, discusses the research findings and how these results can be used to inform campus master planning. Then the limitations and areas of improvement in this research are discussed in detail in this chapter.

Chapter 9 Conclusion provides a summary of the key findings of the research and an account of how the six research objectives have been fulfilled, including recommendations for future work.

1.7 Publications Associated with this PhD Thesis

Abdalla, T., & Peng, C. (2021). Evaluation of housing stock indoor air quality models: A review of data requirements and model performance. *Journal of Building Engineering*, *43*(May), 102846. https://doi.org/10.1016/j.jobe.2021.102846

Chapter 2 Literature Review

Declaration

This chapter is based on the literature review I conducted during 2019-21, subsequently published in the *Journal of Building Engineering*. I have duly acknowledged all the sources from which the ideas and extracts have been taken. The publisher retains the Copyright, but I have been permitted to replicate the material in this PhD Thesis.

Abdalla, T. and Peng, C. (2021). Evaluation of housing stock indoor air quality models: A review of data requirements and model performance. *Journal of Building Engineering*, *43*(May), 102846. https://doi.org/10.1016/j.jobe.2021.102846

2.1 Introduction

The IAQ of a building is one of the most significant concerns regarding occupant health and comfort. Unfortunately, improvements relating to IAQ are not prioritised due to a lack of information and knowledge regarding indoor air pollution, particularly in complex non-domestic buildings. This chapter aims to critically review the existing literature on indoor air quality in buildings and methods for assessing indoor air quality at a building stock level. This review will reveal gaps in the IAQ modelling of individual buildings and building stocks. This chapter is organised into seven sections. In Section 2.2, outdoor air pollution is discussed as an environmental and health hazard and an active research area. Section 2.3 discusses indoor air's relevant characteristics and several factors that may affect air quality. Section 2.4 presents the methods for modelling indoor air in buildings, focusing on multi-zone modelling methods. Section 2.5 presents the methods for modelling IAQ at a building stock level. Section 2.6 presents existing methods for modelling a non-domestic building stock. Section 2.7 reviews the current data-driven methods in modelling the IAQ of buildings. Finally, Section 2.8 concludes this chapter by presenting the research gaps in building stock IAQ models.

2.2 Outdoor Air Pollution

Polluted ambient air contains a complex mixture of compounds with varying concentrations and toxicity levels and has been identified as a leading cause of mortality and morbidity. Nearly 99% of the volume of the air we breathe consists of nitrogen and oxygen. The remaining 1% corresponds to other gases and particles. Some are common constituents (argon, carbon dioxide, or neon) and harmless in average concentrations, while others (called contaminants) may damage the environment and human health even in relatively low concentrations. In the UK, outdoor air pollution contributes to about 40,000 premature deaths yearly (Holgate, 2017). Globally, the number of premature deaths from outdoor air pollution is estimated to increase from 3 million in 2010 to 6 to 9 million in 2060, with the highest increase in non-OECD[2] economies (OECD, 2016).

Several pollutants are used to assess outdoor air quality (also referred to as *criteria pollutants*): particulate matter, also known as particles, ground-level ozone (O_3), carbon monoxide (CO), sulphur dioxide (SO_2) and nitrogen dioxide (NO_2). Particles, a pollutant rather than a single compound, are typically reported according to their mass concentration. For example, using the mass fraction of particles with aerodynamic diameters smaller than 2.5 or 10 µm, as $PM_{2.5}$ or PM_{10}, respectively. The effect of particles on population health is well documented and is now considered one of the significant causes of mortality and morbidity globally. One example is the association of lung cancer and respiratory and cardiovascular diseases with exposure to $PM_{2.5}$. Scientific reports indicate that excessive exposure to high $PM_{2.5}$ concentrations reduces the expected human lifespan by 1–5.5 years (Apte et al., 2018; Cserbik et al., 2020). In 2013, the WHO identified PMs as the leading cause of human cancer (WHO, 2013). Additionally, the health effects of this exposure on the global population were estimated at 103.1 million DALYs [3] in 2015 (Burnett et al., 2018; Cesaroni et al., 2013; Cohen et al., 2017).

A nationwide outdoor air pollution monitoring network is now well established in the UK, providing long-term evidence for setting national outdoor air quality regulations. Much effort has been expended in analysing the impacts of exposure to outdoor contaminants on the population, and some integrated indices have been developed. In most cases, they convert the time-series air

[2] The Organization for Economic Cooperation and Development (OECD): A unique forum for the cooperation of governments from 37 democratic countries with market-based economies, the Organization for Economic Co-operation and Development (OECD), develops guidelines for policies to promote sustainable economic growth.

[3] Disability-Adjusted Life Year (DALY) is the most commonly used measure for quantifying mortality and morbidity to a given disease or risk factor. It combines the years of life lost due to disability with the years of life lost due to death).

pollution concentrations into a single standard index and their associated health outcomes. One example is the Daily Air Quality Index (DAQI) developed by the UK Department for Environment, Food & Rural Affairs (DEFRA). The DAQI of a specific urban site or a region is based on the highest concentration of five pollutants; $PM_{2.5}$, PM_{10}, O_3, NO_2 and SO_2, which is converted into a standard and dimensionless value between 0 and 10 (Connolly et al., 2013).

Although the DAQI index can indicate when ambient air quality has reached short-term unhealthy levels, it depends on information and the location of stationary monitoring stations, which may not reflect personal exposure conditions (Bravo-Linares et al., 2016). Moreover, reporting a single value by converting the outcomes of only five pollutants into a standard index of short-term health effects does not provide an in-depth analysis of the effects of pollution, nor does it provide information on chronic health consequences. Furthermore, PMs can comprise hundreds of components; typical components of ambient PM include Sulphates, Nitrates, Ammoniums, Sodium, Chlorides, Organic Carbons (black carbons and VOCs) , Minerals, and Water. Finally, it is essential to note that ambient particle fractions differ depending on the source, the location, and the meteorological conditions.

In the UK, domestic combustion is a significant source of PM emissions, accounting for 27% and 44% of PM_{10} and $PM_{2.5}$, respectively. PM concentrations are exceptionally high in winter and autumn, where most emissions come from burning wood in closed stoves and open fires. Domestic combustion of coal was the primary source of PM emissions in the 1970s and 1980s; coal now accounts for only a small portion of the emissions from domestic combustion (Birchby et al., 2019). As of 2018, using wood in residential combustion was the primary source of $PM_{2.5}$ emissions (38%) (Chakraborty et al., 2020). The $PM_{2.5}$ emissions generated by residential wood burning doubled between 2003 and 2018 (from 20 to 41 thousand tonnes) and increased by 6.8% between 2017 and 2018. Despite this, PM concentrations are not negligible in spring and summer due to other factors such as higher temperatures, road traffic, industrial combustion, and agricultural practices (Harrison & Yin, 2004).

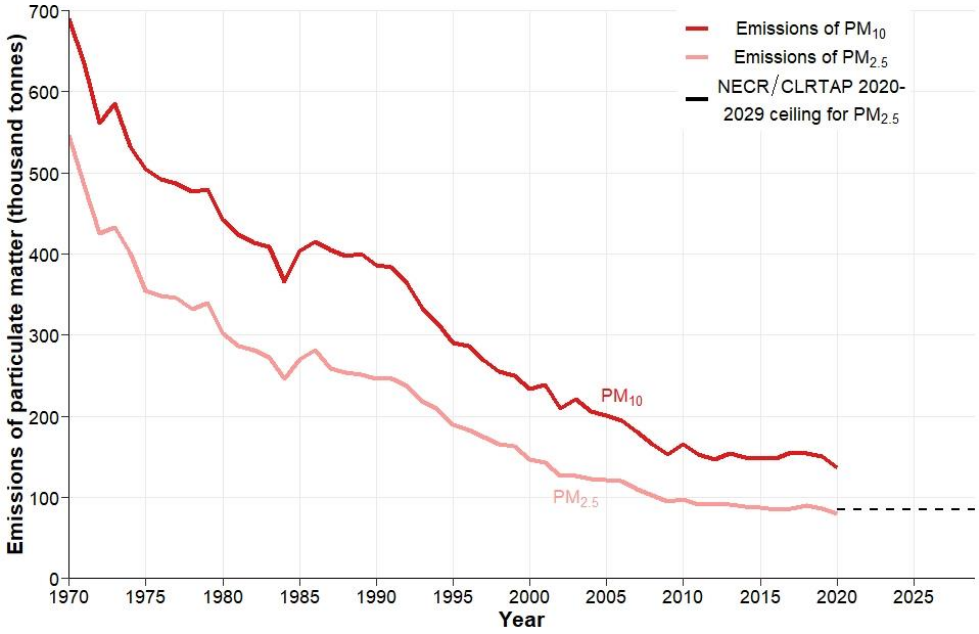


Figure 2.1: Annual emissions of PM_{10} and $PM_{2.5}$ in the UK: 1970-2020 (Brookes et al., 2021)

Currently, road transport accounts for 11.5% of PM_{10} and $PM_{2.5}$ emissions. Exhaust emissions have decreased significantly since 1996 due to stricter pollution control standards (83% for PM_{10} and $PM_{2.5}$ combined) (DEFRA, 2012). However, the increase in traffic activity has partially offset the increase in non-exhaust emissions (such as brake, tyre, and road wear). Another primary PM source is industrial combustion and processes, contributing 43% to PM_{10} and 29 % to $PM_{2.5}$ in 2018 (Birchby et al., 2019). PM emissions from industrial sources have declined over the long term as the demand for chemicals and steel have declined and emission controls have been improved. However, recent industrial combustion increases (especially biomass increases) have partially offset this. With levels of PM and population exposure close to roadsides being much higher than those in background locations, regulations are being implemented to reduce the transport of ambient pollutants to indoor environments. However, a significant proportion of the building stock in the UK was constructed before these regulations or is not eligible for subsidised weatherisation programs.

2.3 Indoor Air Quality

IAQ refers to the air quality inside buildings or other enclosed spaces as it relates to the health and comfort of the occupants. The concentration and type of air pollutants in the indoor environment determine indoor air quality. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) has developed two standards for determining IAQ: ASHRAE Standard 62.1 and ASHRAE Standard 62.2 (ASHRAE, 2019). Both standards provide guidelines for designing, constructing, and operating heating, ventilation, and air conditioning (HVAC) systems to ensure adequate ventilation and IAQ in buildings. ASHRAE Standards 62.1 and 62.2 define IAQ in a binary way, where the air quality is either "acceptable" or "unacceptable". This standard establishes the minimum ventilation rates and other design requirements for residential, commercial and institutional buildings. In addition, the standard considers the amount of outdoor air required to dilute and remove pollutants in indoor air to ensure occupant health and comfort. The standard also considers the building occupancy and use to determine the minimum outdoor air requirements (occupancy densities and outdoor air supply [l/s per person]).

Exposure to air pollutants can occur through inhalation, ingestion, or skin contact. The dose of pollutants an individual is exposed to is determined by the concentration of pollutants in the air and the duration of exposure. Health responses to exposure to air pollutants can vary depending on the type and concentration of the pollutant, as well as the susceptibility of the individual. Health responses can range from mild irritation to severe respiratory illnesses, cardiovascular disease, and cancer (Van Tran et al., 2020). Although people spend 80-90% of their lives in increasingly air-tight buildings, indoor air pollution has received less attention than outdoor air pollution. However, according to the WHO, more than 5 million premature deaths occur yearly due to illnesses caused by poor IAQ (Pai et al., 2022), resulting in multimillion-dollar losses due to reduced employee productivity and increased health system expenses. Several outdoor and indoor factors affect the concentration of indoor air pollutants, including PM, biological pollutants, and over 400 different chemical, organic and inorganic compounds. According to the most recent studies on human exposure to indoor pollution, it was found that indoor environments may be twice as polluted as outdoor environments (González-Martín et al., 2021).

IAQ is a complex issue with potentially acute and chronic health effects. Various metrics measure IAQ, including smell, rating systems, exposure limit values, thresholds, and harm. In the UK, the

National Institute for Health and Care Excellence (NICE) guides IAQ management and emphasizes the importance of addressing IAQ in reducing disease burden. Odour or smell is a subjective measure of IAQ, as it depends on individual perceptions and preferences. While it is not a direct measure of pollutant concentration, it can indicate potential problems with IAQ, such as poor ventilation or chemical or biological contaminants (Dales et al., 1997).

Regulatory agencies establish exposure limit values and thresholds to set safe levels of exposure to various pollutants. These limits are based on scientific studies investigating the relationship between exposure to pollutants and adverse health effects. Exposure limit values and thresholds can vary depending on the pollutant, the duration and frequency of exposure, and the individual's susceptibility. Exposure to pollutants above the recommended exposure limit values and thresholds can lead to acute health effects, such as irritation of the eyes, nose, and throat irritation, or chronic health effects, such as respiratory diseases or cancer.

The World Health Organization (WHO) has established exposure limits to indoor air pollutants based on scientific evidence of their adverse health effects. These limits protect public health by minimizing exposure to harmful pollutants and promoting good indoor air quality. The WHO exposure limits are expressed as concentrations of specific pollutants over a defined period. For example, the WHO recommends a maximum exposure of 15 $\mu g/m^3$ to $PM_{2.5}$ over 24 hours and a maximum exposure of 5 $\mu g/m^3$ over a year (World Health Organization (WHO), 2021a). Other pollutants with WHO exposure limits include NO_2, CO, ozone O_3, and sulfur dioxide SO_2. These limits are essential because indoor air pollution is a significant public health issue. According to the WHO, exposure to indoor air pollution is responsible for an estimated 4.3 million deaths worldwide each year, most of which occur in low- and middle-income countries. In addition, indoor air pollution is associated with various acute and chronic health effects, including respiratory infections, asthma, lung cancer, and cardiovascular disease.

Indoor air pollution levels in the UK are regulated by the Air Quality Standards Regulations 2010, the Air Quality Standards (Wales) Regulations 2010, the Air Quality Standards (Northern Ireland) 2010 and the Air Quality Standards (Scotland) Regulations 2010. These Regulations seek to control exposure to air pollution to protect human health by requiring concentrations within specified limits. The current primary standards and limits are summarised in Table 2.1.

Table 2.1: UK Standard and the WHO recommendations for outdoor concentrations of the DAQI common pollutants.

Pollutant	UK Limit	International Recommendation (World Health Organization (WHO), 2021b)
Particles ($PM_{2.5}$)	-	25 μg/m^3 (24-h mean)
	25 μg/m^3 (annual mean)	10 μg/m^3 (annual mean)
Particles (PM_{10})	50 μg/m^3 (24-h mean) not to be exceeded more than 35 times a year	50 μg/m^3 (24-h mean)
	40 μg/m^3 (annual mean)	20 μg/m^3 (annual mean)
NO_2	200 μg/m^3 (1-h mean) not to be exceeded more than 18 times a year	200 μg/m^3 (1-h mean)
	40 μg/m^3 (annual mean)	40 μg/m^3 (annual mean)
SO_2	350 μg/m^3 (1-h mean) not to be exceeded more than 24 times a year	20 μg/m^3 (24-h mean)
	125 μg/m^3 (24-h mean) not to be exceeded more than 3 times a year	
O_3	100 μg/m^3 (8-h mean) not to be exceeded more than ten times a year	100 μg/m^3 (8-h mean)
		60 μg/m^3 (Peak Season*)

*Average of daily maximum 8-hour mean O_3 concentration in the six consecutive months with the highest six-month running-average O_3 concentration.

Buildings may significantly alter exposure to air pollutants originating both indoors and outdoors (J. Taylor et al., 2014a). Indoor air pollution can result from outdoor pollution due to the infiltration of pollutants caused by human activities, including vehicular traffic (COMEAP, 2018), as well as by natural sources, including radioactive decay in the ground (Turk et al., 1990). Pollution that passively enters a building is influenced by the airtightness of the building, the number of external walls and their exposure to the wind, and how frequently the occupants open their windows (J. Taylor et al., 2014a). There is also the possibility of active infiltration via mechanical ventilation, with the infiltration rate dependent upon the building's ventilation rate and filtration systems. Additionally, indoor activities that can cause air pollution include cooking, using solid fuels for heating, and smoking (Chakraborty et al., 2020). The ventilation rate of a building, internal deposition, and air purification systems acting as air pollutant sinks can improve IAQ. Building characteristics such as airtightness, purpose-installed ventilation systems, internal and external building geometry, and occupant behaviours (e.g., opening windows for ventilation) can impact the infiltration level, and they may vary significantly between different building types.

2.4 Assessing Indoor Air Quality

At the most basic level, there are two approaches to measuring or quantifying air pollution inside a building: direct or indirect. Direct or *in-situ* approaches involve deploying stationary or mobile sensors and data processing facilities. If resourced adequately, direct methods can accurately report personal exposures and indoor contaminant concentrations given known instrument and measurement limitations. In addition, although time-consuming and potentially costly, direct methods can obtain specific IAQ measurements of extant species and magnitudes (extremes) and the sources of pollutants and emission rates (Abdalla & Peng, 2021). Indirect approaches, on the other hand, utilise computational modelling and statistical methods to estimate/predict indoor contaminant concentrations and personal exposures.

The direct approaches seldom capture the complex dynamic interactions of air particles and transient behaviours within a building or a group of buildings. This is due to the limitations imposed by either the instrumental factors (e.g. device selection, calibration and reliability) or the sampling methods (e.g. measurement location, sampling frequency and time-averaging period) (Coleman & Meggers, 2018; B Jones et al., 2018). Moreover, there can be uncertainties associated with an individual or a network of IAQ sensors, resulting in measurements that may be potentially misleading (O'Leary, de Kluizenaar, et al., 2019). Sharing similar purposes of quantifying indoor air environments, indirect approaches attempt to model such complexity computationally, which can be guided by iterative data-based calibration and hypotheses testing. The accuracy and robustness of computational IAQ models can be evaluated and improved through field measurements. However, depending on the methods employed, computational modelling may oversimplify the spatial-temporal dynamics in which the physical-chemical processes of air particles or gases (e.g., $PM_{2.5}$, PM_{10}, O_3, NO_x, CO, SO_2) take place. Nevertheless, one of the desirable benefits of the indirect approaches is the applicability of computational IAQ models to evaluate the likely effects of interventions proposed for improving IAQ at scale.

2.4.1 Mass Balance Models

Due to the complexity of obtaining field measurements of indoor air pollution in buildings or groups of buildings, mass balance equations are commonly used to determine the relationship between indoor particle concentrations and specific variables. Derived from understanding the underlying physical factors and processes that govern the transfers and transformations of

pollutants in indoor environments, mass balance models provide a relatively simple means of estimating changes in average concentrations of indoor pollutants spatially (e.g., in a room or group of rooms) and temporally. They are often applied assuming well-mixed air volumes can characterise a room or a building. In its basic form, a mass balance model that describes the indoor concentration of air pollutants under specified emissions or removal processes can be expressed by an ordinary differential equation:

$$\frac{dC_i}{dt} = (C_{Sources} - C_{Sinks})/V \quad (2.1)$$

Where C_i Is the concentration at time *t*, $C_{Sources}$ is the sum of concentration gain from all sources (μg/m^3), and C_{Sinks} is the sum of concentration loss from all sinks (μg/m^3).

Because the air in a controlled indoor environment is intrinsically complex, no single mass balance model is well suited for addressing all pollutants or issues under investigation, even though all mass balance models are based on the same fundamental principle of mass conservation. Therefore, following (Nazaroff, 2004a), with the only assumption that the indoor particles attributes are uniform (well-mixed) throughout the interior space, Eq. (2.1) can be expanded to represent a range of factors that determine the indoor concentrations of the particle attribute in a single well-mixed zone:

$$\frac{d(C_iV)}{dt} = E + C_o[Q_s(1-\eta_s) + Q_N + Q_LP] - C_i\,[Q_F\eta_F + \beta V + (Q_s + Q_N + Q_L)] \quad (2.2)$$

Where V is room volume (m^3), C_o Is the concentration of particles in outdoor air (μg/m^3), Q_s is the mechanical supply flow rate (m^3/h), Q_N is the natural ventilation flow rate (m^3/h), Q_L is leakage (infiltration) flow rate (m^3/h), η_s Is a filter with single-pass removal efficiency (—), P is the penetration fraction of particles (—), Q_F Is indoor air particle control flow rate (m^3/h), η_F is a filter with single-pass removal efficiency (—), E is an emission source operating at (μg/h), and β is the loss of particles from indoor air by deposition represented by a first-order loss-rate coefficient (h^{-1}). (Nazaroff, 2004a) illustrated the mass balance approach, which systematically depicts the processes represented in Eq. (2.2).

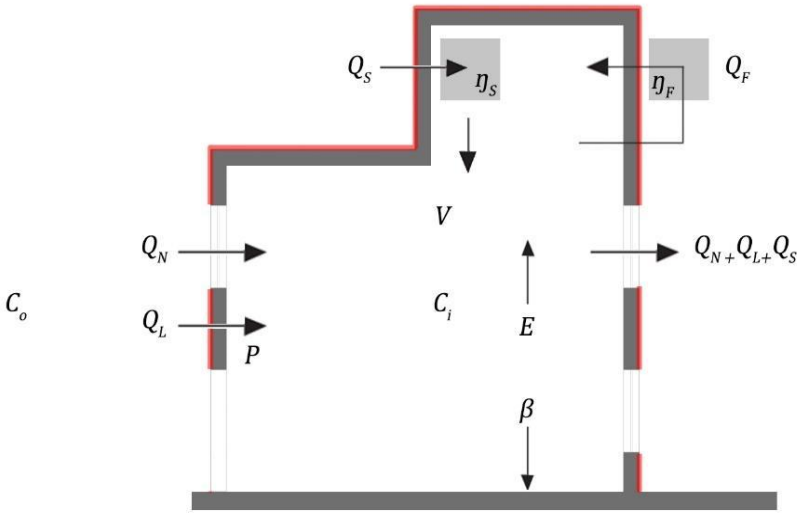


Figure 2.2**:** Schematic representation of the physical processes affecting indoor particle concentration levels (Nazaroff, 2004a)

It is worth noting that Eq. (2.2) may be extended to include different processes when the indoor environment under study is represented as multiple well-mixed zones. This includes terms that account for the supply and loss of particle attributes by inter-zone and infiltration airflows (Miller & Nazaroff, 2001).

In addition to the differences in air pollutants properties, there are myriad variations in how different indoor environments are operated, which renders it difficult nor practical for a single mass balance model to cover all circumstances (Nazaroff, 2004a). Previous studies have tried to illustrate the processes involved in different formulas of the mass balance models (Dimitroulopoulou et al., 2006; Jamieson, 2008; Schneider et al., 2004). However, these studies have primarily been small in scale and applied only for short periods over several locations (Jamieson, 2008). The factors and terms in these models are subject to variability and uncertainty in the relationships between the physical environmental phenomena, building characteristics, and dynamic composition of pollutants (see Figure 2.3 for a summary). To achieve reliable predictions of IAQ at multiple spatial and temporal resolutions, simulation tools need to be built with indoor particle dynamics mathematical models that capture the complex physical and environmental phenomena as accurately as possible.

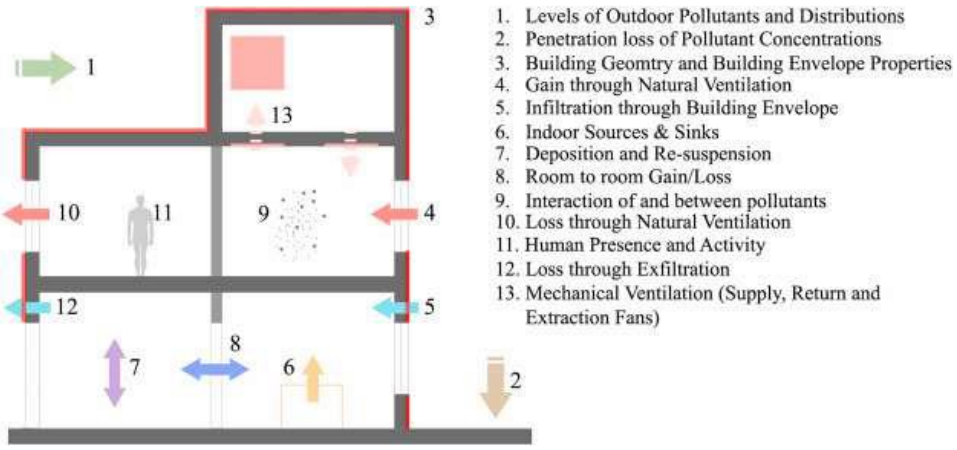


Figure 2.3: Summary of the main factors and processes affecting indoor concentrations of pollutants (red lines indicate the boundary of the building envelope) (IEHIAS)

2.4.2 Computational Modelling of Indoor Air Quality in Buildings

Complex analyses must identify relevant pollutants from each source under actual conditions. Indirect methods simplify any model by relying on assumptions and must consider input parameters and uncertainty about model accuracy. Numerous approaches have been used to address these issues, including methods that complement each other so that measured data better informs model inputs and results can be validated and calibrated. Consequently, the accuracy of the model will be determined by the availability and quality of input data. The manner in which indoor airspace is partitioned to provide a required level of spatial detail can be categorised into three methods based on how the airspace is partitioned. From simple models with (i) a single and well-mixed zone, continuing with (ii) multi-zone models, to (iii) more complex models, using a non–uniform distribution of the pollutants with Computational Fluid Dynamics (CFD) simulation tools.

As summarised in Figure 2.4, single-zone and multi-zone models adopt different principles, strategies and solvers that generate different outputs. It is difficult to determine which model is the best because of the different modelling and simulation requirements, such as building complexity, the parameters investigated, the expected results, and the degree of accuracy required (Yu et al., 2019). A wide range of input parameters is required to perform IAQ simulations, including climate data, building fabric and geometry, building systems, and occupancy schedules (A. K. Persily & Ivy, 2001). As buildings have become more complex, a conceptual understanding

of the fundamental principles of ventilation and building systems, including HVAC, must be coupled with computational modelling to accurately predict contaminant behaviour and its impact on human health.

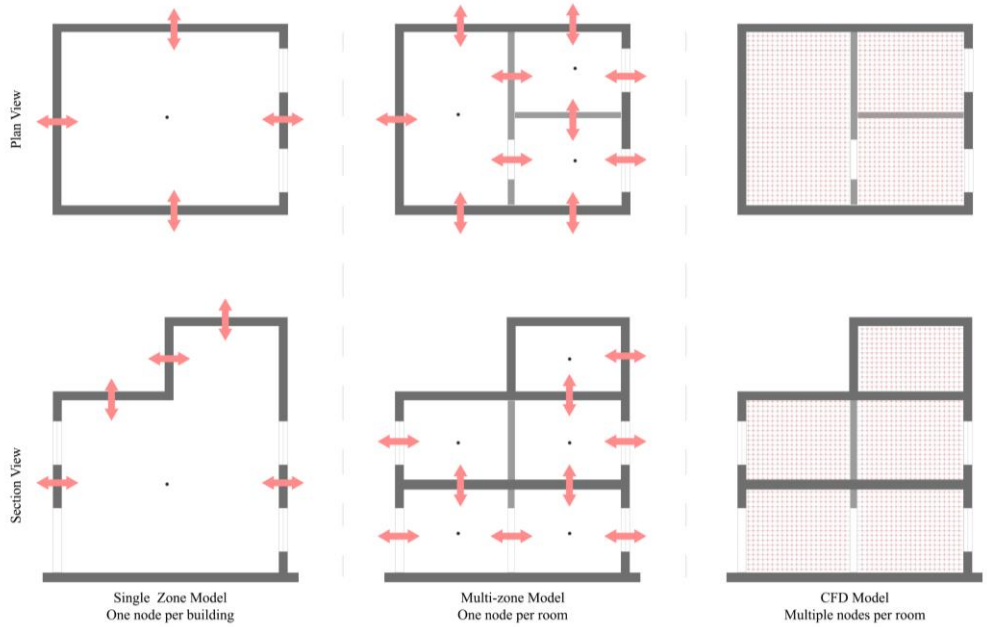


Figure 2.4: Approaches to modelling IAQ in a building. Left: Single zone models; Middle: Multi-zone models; Right: CFD Models. Each node represents a well-mixed volume. (Based on (J. Axley, 2007)).

Assuming homogeneous physical properties of air (i.e. uniform temperature, air pressure, and contaminant concentrations), well-mixed single-zone models typically take a *macroscopic* view of air within one volume represented by a node. Meanwhile, multi-zone models define multiple nodes (or zones), each representing a room or a group of rooms connected by several airflow paths. In both models, the airflows between each zone and the outdoor air are calculated iteratively using mass balance equations until the pressure relationships are solved at each time step. Hence, in the face of multiple challenges, such as the stochastic nature of weather, occupant behaviours, building components, and uncertainties in simulation input parameters, the model choice could have significant implications for estimating indoor contaminant concentrations.

Computational fluid dynamics (CFD) modelling takes a *microscopic* view of airflow in a zone or a group of zones within a building (Yu et al., 2019). CFD models are particularly relevant where uniform mixing within a zone or zones cannot be assumed reasonably to represent the airflow

conditions under investigation (Shimada et al., 1996). CFD-based models can compute fine-grained indoor contaminants concentrations and personal exposures, and they have been widely used to simulate contaminants infiltration from outdoor generated sources and contaminants transport between zones within a building (Panagopoulos et al., 2011; Yang et al., 2014).

As mentioned, the ideal method for IAQ assessment of existing buildings is through large-scale data collection campaigns. However, computational IAQ modelling has become preferable due to cost and time constraints, especially when evaluating intervention proposals. State-of-the-art IAQ models include multi-zone or *airflow networks* and CFD models. These models can calculate indoor air properties such as indoor air temperatures, airflow rates, and indoor contaminant concentrations. In predicting a building's IAQ, airflow network and CFD models perform differently in complexity, reliability and accuracy. CFD-based models are computationally expensive as they often resolve airflow dynamics at high spatial and temporal resolutions. Hensen and Lamberts pointed out that there appeared to be a widespread misconception that using CFD will reduce uncertainties and increase the accuracy of IAQ predictions (Hensen & Lamberts, 2011). Deviating from the ideal case to higher or lower complexity can induce risks of simulation errors. Therefore, the selection of appropriate computation methods should be guided by the purpose of the simulation (e.g., airflow network methods for bulk airflow analysis or CFD to study trends (sensitivity of flow patterns to small changes).

Robinson (2008) stated that all simulation models simplify reality and are based on abstract representations of real-world phenomena. In this regard, it is necessary to make explicit assumptions about the computational methods employed in modelling the IAQ of building stock. Next, the assumptions made in IAQ simulation tools: computational unit, abstraction of building components and systems, and input variables and parameters will be discussed.

Both single-zone and multi-zone models are based on the assumption of perfectly homogeneous or *well-mixed* conditions (i.e., each zone has an average air pollutant concentration value). In single-zone models, a building is simplified to be represented by a single zone or *node* without considering its interior partitions (Megri & Haghighat, 2007). Consequently, the physical details of heat and mass transfer between rooms within a building caused by temperature and pressure variations are ignored (Yu et al., 2019). Figure 2.5 illustrates the assumptions, showing the air temperature in a single-zone model represented by an average value of T_{in} (°C) (Megri &

Haghighat, 2007). A steady state model (see Eq. 2.3 & 2.4.) stipulates that the mass flow rate $\dot{m}_{in}$ (kg/s) should be equal to the outlet mass flow rate $\dot{m}_{out}$ (kg/s) when infiltration is neglected, and the energy is conserved between $\dot{q}_{in}$ (rate of heat energy supplied into room/building (Watts)), $\dot{q}_{out}$ (rate of heat energy removed from room/building (Watts)) , and $\dot{q}_{loss/gain}$ (rate of heat energy transferred through room/building structures (Watts)).

$$\frac{dM_{space}}{dt} = \dot{m}_{in} - \dot{m}_{out} = 0 \tag{2.3}$$

$$\frac{dQ_{space}}{dt} = \dot{q}_{in} + \dot{q}_{out} + \dot{q}_{loss/gain} = 0 \tag{2.4}$$

Additionally, a single node represents the outdoor climate, and the physical parameters of this node are assigned from weather conditions. Notwithstanding, single-zone well-mixed models are relatively easy to implement and fast to compute. They are suitable for estimating bulk airflow properties when the domain of interest can be treated as a single zone or node.

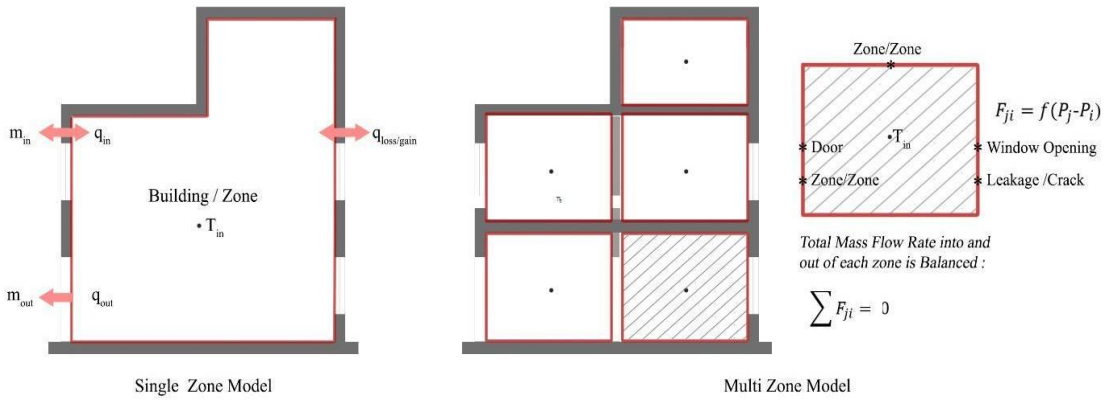


Figure 2.5: A summary of IAQ simulation assumptions of single-zone steady-state and multi-zone models. (Red lines delineate the inner volume of a zone based on (Yu et al., 2019)).

Multi-zone models use *rooms* as the minimum computational unit. They calculate the airflow and contaminant transport inside a building within minutes or seconds. However, shorter computing times can be achieved by assuming homogeneity in each zone; that is, the distributions of air pressure, air temperature, and contaminant concentration in each room are assumed uniform and

leave out the air momentum effect from an inflow opening (J. Axley, 2007). This is not always the case because a vertical temperature gradient exists in rooms filled with stratified flows driven by displacement ventilation or water heating systems (Wang & Chen, 2007). In addition, the well-mixing assumptions could be problematic for simulations of poorly mixed air and contaminants. In an earlier review of airflow and infiltration models, (F. Haghighat, 1989) stated that a multi-zone airflow model should be able to fully account for the driving forces that cause air to flow from outdoor to indoor and between indoor zones, including the stack effect, the wind pressure effect on building envelope, and the effect of HVAC systems on airflow.

In general, multi-zone airflow models are based on constructing a matrix of equations representing all airflow paths connecting zones (nodes) within a building. A mathematical equation describing each airflow path (i.e. door, window, crack, etc.) is used to numerically solve the resulting matrix, typically by the Newton-Raphson method (Conte & de Boor, 1972). All equations are solved iteratively to ensure reaching the convergence state when the sum of all mass flow rates through all flow paths approaches zero, as illustrated in Eq. 2.5.

$$\sum F_{ji} = 0 \tag{2.5}$$

where F_{ji} is the mass airflow rate from zone *j* to zone *i* (kg/s).

In a multi-zone model, the mass airflow rate at each airflow path is some function of the flow pressure drop along the flow path, P_j-P_i (W.S. Dols, 2007), and is expressed as:

$$F_{ji} = f(P_j\text{-}P_i) \tag{2.6}$$

The mass of air, m_i (kg), in zone *i* is given by the ideal gas law:

$$m_i = \rho V_i = \frac{P_i V_i}{R T_i} \tag{2.7}$$

Where ρ is the air density, V_i Is the zone volume (m^3), P_i is the zone pressure (Pa), T_i The zone temperature (K), and *R* is the gas constant for air = 287.055 (J/kg.K).

Although multi-zone models of individual buildings can provide spatial average estimates of pollutant concentrations with a reasonable simplification of indoor physical phenomena, it is possible to describe the building's attributes (e.g., contaminant sources, airflow paths, occupancy schedules, and building service systems) with a high level of resolution. However, to achieve prediction accuracy in building stock IAQ modelling at a reasonable computation cost,

consideration of input parameter variability and uncertainty is required to select appropriate computational modelling methods and tools without risking oversimplification.

2.4.3 Simulation Tools Used in Indoor Air Quality Models

Over the past three decades, several IAQ simulation tools have been developed, such as CONTAM and COMIS (Feustel, 1999; W.S. Dols, 2007). These tools have been used primarily in modelling the IAQ of individual buildings. More recently, CONTAM and EnergyPlus were used to model IAQ of archetypes in building stock studies (see Section 2.4.4). CONTAM is a multi-zone airflow and contaminant transport simulation tool developed and maintained by NIST (W.S. Dols & Polidoro, 2015), which has been validated in many studies in various building types and locations (Emmerich & Hirnikel, 2001a; Fariborz Haghighat, 1996). CONTAM has been built with an updated version of the AirNet model (G.N. Walton, 1989) and provides a graphical user interface for intuitive inputs of building zones and construction, airflow paths and other building elements (McDowell et al., 2003).

More specifically, CONTAM allows users to model airflow rates, including infiltration, exfiltration, zone-to-zone airflows driven by mechanical ventilation systems, wind pressures on the building envelope and buoyancy effects. CONTAM's contaminant dispersal model is an implementation of Axley methods (J. . Axley, 1988; J. W. Axley, 1987) and has been widely used in many studies to predict contaminant concentrations in buildings under multiple designs and retrofitting scenarios (García-Tobar, 2019; L. J. Underhill et al., 2018). However, as a standalone package, CONTAM does not modify zonal air density in response to environmental changes due to building interactions and occupant behaviours. Therefore, CONTAM does not have the capability of performing dynamic thermal simulations on its own.

On the other hand, as one of the widely used whole building energy simulation engines, EnergyPlus (Office of Energy Efficient and Renewable Energy, n.d.) can simulate airflows in buildings using the multi-zone Airflow Network Tool, an airflow model based on the early versions of COMIS and AirNet. The Airflow Network Tool can simulate infiltration and exfiltration rates driven by indoor/outdoor pressure differences, ventilation mechanisms, building envelope permeability, and zone-to-zone airflows. From the perspective of IAQ modelling, CONTAM and EnergyPlus have advantages in respect of each other. For example, CONTAM simulates complex airflow networks in a building and enables users to model absolute airflow

paths and multiple contaminant species. On the other hand, EnergyPlus performs dynamic thermal simulations and accounts for pressure differences between multiple zones in a building. However, Interzone airflows and infiltrations in EnergyPlus are user-specified and not pressure-dependent as in CONTAM. Moreover, EnergyPlus does not require Interzone airflows to balance with system airflow rates (W. Stuart Dols et al., 2016).

Using EnergyPlus to model contaminant transport, Taylor et al. have developed the Generic Contaminant Model (GCM) tool, allowing users to model the behaviour of one specific pollutant within a building. GCM enables the modelling of dynamic thermal behaviour and single pollutant transport within one simulation package (Jonathon Taylor et al., 2014). Polluto, another in-house tool developed at the University College London (UCL), also offers multiple contaminants transport modelling with EnergyPlus. Table 2.2 presents a comparison between CONTAM and the UCL in-house IAQ tools.

Table 2.2: Comparison of IAQ simulation tools used in housing stock IAQ modelling.

	Simulation Tools		
	CONTAM	EnergyPlus GCM	EnergyPlus Polluto
Main Usage	Airflow rates, contaminant transport through airflow, and building occupant exposure.	Energy analysis, thermal load simulation, airflow, and contaminant transport.	Energy analysis, thermal load simulation, airflow, and contaminant transport.
User Interface	Simple	Complex	Complex
Thermal Behaviour	Static [Dynamic if coupled with a thermal engine]	Dynamic	Dynamic
Contaminant Behaviour	Yes (A rich set of sources and sinks, including deposition and re-suspension)	No	No
Changes in Occupant Behaviour Consideration	Yes	Yes	Yes
Modelling of Pollutants	Multiple Pollutants	Single Pollutant	Multiple Pollutants
Air Leakage Points	Multiple Airflow Leakage Points	A one-to-one correspondence between heat transfer and air leakage	A one-to-one correspondence between heat transfer and air leakage
Mechanical Systems Modelling	Complex & Multiple Systems	One System	One System
Warm-up Days	No	Yes, to ensure any thermal capacitance values are representative of the zone.	Yes, to ensure any thermal capacitance values are representative of the zone.
The capability of building control operations	Yes	Yes, indoor concentrations as flags for ventilation system operation	No
Non-trace contaminants	Yes, already included in air density calculations.	Yes, if coupled with the Heat and Moisture Transport (HAMT) model.	Yes, if coupled with the Heat and Moisture Transport (HAMT) model.

Lately, attempts have been made to couple multi-zone airflow models with dynamic multi-zone thermal models to perform dynamic IAQ-Energy co-simulation. The coupling of building energy and airflow modelling has been discussed previously (Grot, 1985; G.N. Walton, 1989). They presented the quasi-dynamic and dynamic coupling methods used in the coupling between EnergyPlus and CONTAM. (W. Stuart Dols et al., 2016) addressed the mathematical description of the energy balance equations of EnergyPlus and the mass balance of airflows in CONTAM. New and existing components and tools have been developed and modified to facilitate synchronising building geometric representations and dynamic data exchange between CONTAM and EnergyPlus. Figure 2.6 illustrates the relationship between the components utilised during a CONTAM/EnergyPlus co-simulation.

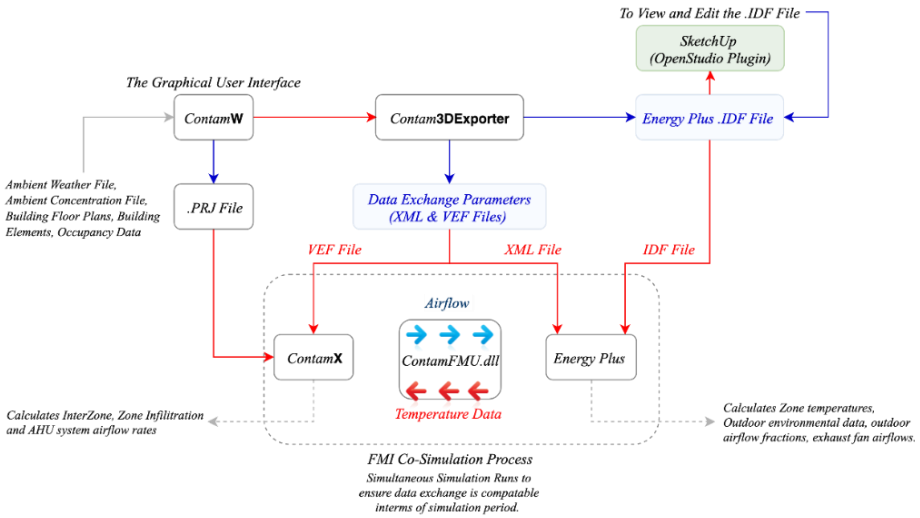


Figure 2.6: Schematic relationship between CONTAM/EnergyPlus co-simulation components (W. Stuart Dols et al., 2016)

CONTAM's graphical user interface, ContamW, allows creating project files (*.prj) representing scaled geometries of building floor plans. Contam3Dexporter tool creates an EnergyPlus input data file (IDF file) and files containing data exchange parameters (VEF and XML files). The IDF file can be edited and exported again using the SketchUp software plugin OpenStudio. Contam3Dexporter tool exports a Windows dynamic link library (ContamFMU.dll) based on the FMI Co-Simulation specification version 1.0 (Blochwitz et al., 2011). ContamFMU.dll manages data exchange and the execution of ContamX during the co-simulation. At present, EnergyPlus transfers zone temperatures, ventilation systems airflows, outdoor airflow fractions, output variables, and outdoor environment data to ContamX. On the other hand, ContamX transfers zone Infiltration rates, Inter-zone airflows and Control values. Previous studies have validated and

verified this process (W. Stuart Dols et al., 2021; Emmerich et al., 2019). More recently, the application of the coupled simulation approach has been part of several studies to estimate indoor $PM_{2.5}$ exposure profiles (Milando et al., 2022a) and to perform a whole building analysis (IAQ, Energy, and Ventilation) (W. Stuart Dols et al., 2021).

2.5 Modelling of a Building Stock

Policymakers in many countries have actively engaged in establishing regulations and guidelines for improving and maintaining urban air quality (DEFRA (Department for Environment Food and Rural Affairs) & DfT (Department for Transport) the UK, 2017; DEFRA (Department for Environment Food and Rural Affairs) the UK, 2018; Department for Environment Food and Rural Affairs, 2007). However, accurate information and the definition of target indicators are required to understand and evaluate building performance on a large scale. In addition, over time, the number, composition, and characteristics of a building stock constantly evolve, which requires models that allow a stock to be outlined and evaluated according to specific indicators and the benefits, drawbacks and trade-offs of potential interventions being assessed before implementation.

Building stock IAQ modelling is an attempt at quantifying and predicting the IAQ of building types statistically representative of a building stock on a city, regional or national scale (Abdalla & Peng, 2021). A building stock located in a geographical domain is the total account of the building subject to planned or organic changes over time. Changes in the composition and characteristics of building stock can be attributed to factors such as climate and environmental changes, socioeconomic and demographic changes, or retrofitting measures applied to improve energy efficiency (G. Sousa et al., 2017b).

A recent review of the UK's housing stock energy models (G. Sousa et al., 2017b) showed that the models could be generally categorised as 'top-down' and 'bottom-up' and developed to work in an aggregated or disaggregated manner. Top-down models primarily predict the macroeconomic performance of a building stock based on the statistical relationships between historical aggregated data and socioeconomic determinants such as the gross domestic product, population, climate conditions and fuel prices (Swan & Ugursal, 2009). Since the top-down models rely on historical data, they are less capable of testing the performance and impact of new technologies and policies. Bottom-up models use empirical data from a hierarchical level less

than the sector/stock as a whole. The bottom-up approaches account for the performance of individual end-uses, individual buildings, or groups of buildings and extrapolate the sector/stock performance with weightings of each modelled dwelling or group of dwellings based on their representations of the sector/stock (Swan & Ugursal, 2009).

The bottom-up methods can be statistical, engineering (physical) based, or combining both. For example, most of the UK's housing stock energy models are bottom-up models developed from simplified steady-state representations of physical phenomena (G. Sousa et al., 2017b). High-resolution housing stock data was used to characterise the stock constitution in terms of building geometry and construction, environmental systems (ventilation, sources and flow paths), and occupancy (patterns of presence and behaviour) to overcome simplification and achieve satisfactory prediction accuracy and consistency. Furthermore, advanced statistical methods such as probabilistic sampling, Gaussian processes and sensitivity analysis are increasingly applied to quantify uncertainties encountered in stock modelling.

According to (Swan & Ugursal, 2009) and (G. Sousa et al., 2017b), bottom-up building stock modelling follows an inductive path of consolidating microscopic measures such as building properties, internal conditions, usage schedules, and building services systems. Bottom-up models thus require extensive empirical data from surveys, field measurements, and assumptions (in the absence of data) to describe each component required of an engineering (physical) approach (ASHRAE, 2019). Based on building physics, several researchers have applied bottom-up modelling techniques to develop representative buildings (archetypes) and used them to calibrate and predict building stock energy performance (e.g.(Ghiassi & Mahdavi, 2017; A. Persily et al., 2006; Sokol et al., 2017), Data entries sharing similar or equal categories were grouped or clustered to classify the dwelling types. After the classification, each archetype was characterised with attributes representing a proportion of the housing stock. So the larger the number of archetypes developed, the more representative of the stock they become and the more widespread the conclusions derived from the modelling results (Molina, 2019).

A recent review by (Abdalla & Peng, 2021) showed that current building stock IAQ models are aimed at the housing stock and are categorised as 'bottom-up' stock models developed to work in a disaggregated manner. Their review showed that housing stock IAQ modelling typically involves (a) classification and characterisation of the dwelling types (archetypes) representative of the housing stock under modelling and (b) utilisation of modelling tools to evaluate the IAQ performance of the archetypes (Shrubsole et al., 2012). The outputs for all archetypes are then

extrapolated to a whole stock of dwellings using weighting factors. However, deterministic bottom-up models produce only one output for one building with given inputs. Hence, the deterministic bottom-up engineering methods can be less applicable to many buildings with different sizes, types, ages, functions, and operating conditions.

As housing stocks are complex dynamic entities that undergo constant evolution, the scope of targeted performance indicators and potential interventions (e.g. likely parameters of dwelling retrofitting) should be considered before stock model implementation (Molina, Kent, et al., 2020). Finally, regular update and calibration processes should be carried out to minimise errors between the predicted and observed values. Table 2.3 summarises the existing housing stock IAQ modelling approaches, sampling methods, and parameter types.

Table 2.3: Existing bottom-up housing stock IAQ models: a stock sampling approach, stock formulation, and parameter selection.

Sampling Approach	**Housing Stock Model Formulation Approach**	**Parameter Selection**	**Parameter Types**	**Variability**
Deterministic	Archetype Approach A	Classification	Deterministic	
		Characterisation	Deterministic Parameters from Literature or Building Data	No
Hybrid	Archetype Approach B	Classification	Deterministic	
		Clustering	Key Descriptive Factors Aggregated into Clusters or *Cells* Utilising Factorial Design	Reflects Variability Between Groups
Probabilistic	*Metamodel (Utilising Machine Learning)	Latin Hypercube / Monte-Carlo	Variable Probability Distribution Functions to represent Uncertainty / Variability	Reflects Variability in the Descriptive Parameters within Groups

* Simplified algebraic or statistical model as a surrogate of the more detailed engineering model, which allows for lower computational requirements (Sokol et al., 2017)

Since the early 2010s, several housing stock IAQ models have been developed to assess the IAQ of housing stock on a city or national scale (Abdalla & Peng, 2021). Based on various datasets and computational IAQ simulation tools as described in the previous section, these models were built to perform mainly simulations of mass transfer processes in sampled representative dwellings. Table 2.4 summarises the housing stock IAQ models developed between 2012 and 2020.

Table 2.4: The housing stock IAQ models developed and published during 2012-2020.

Nation	Model	Date	Stock Scale	IAQ Performance Measure	Simulation Engine	Modelling Approach	Source
US	REIAQM	2018	National	Indoor $PM_{2.5}$ Concentration and HVAC Runtimes	EnergyPlus	Physical Deterministic Approach	(Fazli & Stephens, 2018)
UK	LNDN-A	2012	City	Indoor $PM_{2.5}$ Concentration and Personal Exposure	CONTAM	Physical Deterministic Approach	(Shrubsole et al., 2012)
	LNDN-B	2014	City	Indoor $PM_{2.5}$ Concentration and Mapped I/O	EnergyPlus	Physical Deterministic Approach	(J. Taylor et al., 2014b)
	ENG-A	2014	National	Indoor $PM_{2.5}$ Concentration	CONTAM	Meta-modelling Probabilistic Approach	(Das et al., 2014)
	ENG-B	2016	National	$PM_{2.5}$ I/O Ratio, RH, EUI, & Overheating Metric	EnergyPlus	Meta-modelling Deterministic Approach	(Symonds et al., 2016)
	GBM	2016	Regional	Mapping $PM_{2.5}$ I/O and Overheating	EnergyPlus	Physical Deterministic Approach	(Jonathon Taylor et al., 2019)
	ENGW	2019	Regional	$PM_{2.5}$ and NO_2 I/O and Indoor Concentrations	EnergyPlus	Meta-modelling Deterministic Approach	(Jonathon Taylor et al., 2019)
Chile	CHAARM	2020	National	$PM_{2.5}$ Indoor Concentration, Ventilation and Infiltration Rates	CONTAM	Probabilistic Approach	(Molina, Jones, et al., 2020)

The common approach in most studies is using representative buildings, i.e., archetypes, in modelling the IAQ of housing stock. For example, the Residential Energy-IAQ Model (REIAQM) (Fazli & Stephens, 2018) was developed to model and predict the annual energy use for space conditioning and indoor concentrations of various pollutants across the residential building sector in the US. REIAQM utilised the geometries and housing characteristics of 209 housing archetypes developed previously by (A. Persily et al., 2006), representing 80% of the US residential stock. The Domestic Stock $PM_{2.5}$ Model for London **(**LNDN-A**)** was based on a deterministic physical approach to model and predict the indoor exposure to $PM_{2.5}$ in London's domestic building stock (Shrubsole et al., 2012). London Housing Stock $PM_{2.5}$ Model (LNDN-B) aimed to determine the indoor $PM_{2.5}$ concentrations from outdoor sources for different housing typologies across London (J. Taylor et al., 2014b). The study was based on a deterministic physical approach, utilising the Airflow Network Model and the EnergyPlus GCM Model to simulate the infiltration of $PM_{2.5}$ through the envelope of 15 dwelling archetypes developed previously by (Oikonomou et al., 2012). The Chilean Housing Archetype AiR quality Model (CHAARM) was developed to predict uncertainties in indoor pollutant concentrations, ventilation, infiltration rates and associated

energy demand in the heating season, including the sensitivity of the model outputs to the inputs (Molina, Jones, et al., 2020). The model was based on the previously identified archetypes representing the national Chilean housing stock (Molina, Kent, et al., 2020).

Other methods used to model the IAQ of a housing stock include using metamodels to represent a combination of various inputs in a mathematical form. For instance, two types of artificial neural networks (cascade-forward and feed-forward) were used to predict the winter indoor concentrations of $PM_{2.5}$ from external and internal sources in the English housing stock using a simplified single zone model ENG-A (Das et al., 2014). Furthermore, The ENG-B was developed to predict the indoor overheating and air pollution risk in England's domestic stock (Symonds et al., 2016). This model used two metamodeling methods (neural networks and radial basis function) to reproduce non-linear non-monotonic relations between model inputs and simulation outputs. Finally, the English and Welsh Housing Stock IAQ Metamodel (ENGW) (Jonathon Taylor et al., 2019) was developed as an updated version of the ENG-B model.

To improve the quality of predictions in housing stock IAQ modelling, the issue of uncertainty needs to be addressed. Some uncertainties are related to the mathematical models used to represent the physical phenomena, some to the heterogeneity of the housing stock under investigation, and some to unknown or random variations of the input's values (*epistemic* or *aleatoric*). The methods used to quantify these uncertainties include clustering techniques, which reflect the variability between groups, and Monte-Carlo sampling methods, which account for variability in group descriptive parameters (Parag Wate et al., 2019). Recent methods include Gaussian process emulators for uncertainty quantification and sensitivity analyses for complex stochastic building performance modelling (Lim & Zhai, 2017; P. Wate et al., 2020).

2.6 Modelling Non-Domestic Building Stocks

Non-residential building stocks are characterised by heterogeneous and complex stock morphologies, which make it difficult to model their energy demand and indoor air quality. Therefore, more progress has been made in developing databases for domestic stocks. Modelling non-domestic stocks depends on data availability for each building within the stock. (Pout, 2000) Introduced a methodology to estimate the UK's Non-domestic stock current energy consumption utilising the Non-domestic Energy and Emissions Model (N-DEEM) developed by (Steadman et

al., 2000). Built-up floor areas, classified by non-domestic building occupiers' activity types, are used and grossed up to cover the whole UK. Energy end-users are estimated for each classified activity and grouped by the fuel type used. Energy consumption for each activity is sourced from the Sheffield Hallam University energy surveys (Mortimer et al., 2000). However, this survey did not provide sufficient data for all activity types and, therefore, plagued all attempts to model energy consumption on a large scale.

The Carbon Reduction in Buildings (CaRB) project developed a preliminary energy use model for the non-domestic stock in England and Wales (Bruhns et al., 2006). It used a bottom-up approach to building energy utilising data on individual premises sourced from the UK's Valuation Office Agency (VOA) and using floor areas for individual premises rather than building areas. The Non-Domestic Carbon Scenario Model (NDCSM) estimates the non-domestic stock's energy consumption and carbon emissions (Hinnells & Shea, 2008). These scenarios are based on interventions and climate change projections. Data for developing this model is once again sourced from the SHU surveys and the VOA. S. Taylor et al., 2014 Introduced a new definition of space in a built environment called the Self-contained Unit (SCU). This unit represents the 3D geometrical and spatial relationship between premises and buildings and the activities on different floors within a building. It facilitates two national data sets in the UK, Ordnance Survey (OS) and the VOA, to automatically assemble the geometrical and geographical structure of the nNon-domestic stock. After that, electricity and gas consumption data are attached to the premises within each SCU; see Figure 2.7. The SCUs are then aggregated to model the energy consumption of the non-domestic stock. This concept was used later to develop a new three-dimensional model of the British Building Stock, called '3DStock' (Evans et al., 2017).

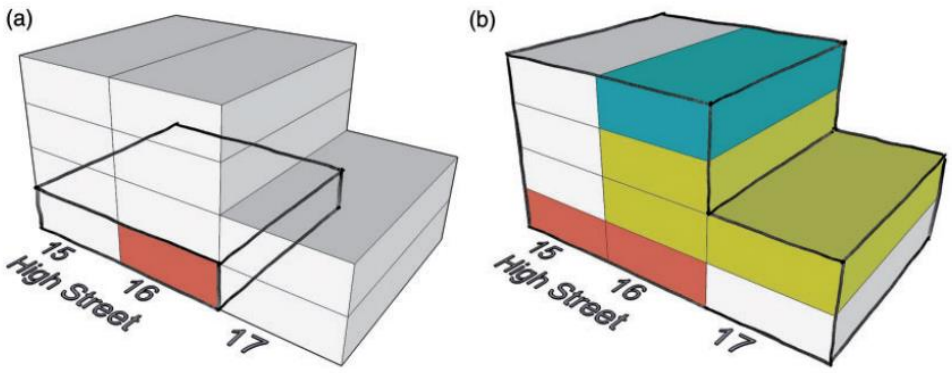


Figure 2.7: Self-contained Units (SCUs) (Evans et al., 2017)

As of this writing, there have been no studies on the IAQ of the non-domestic building stock. This could be due to the complexity and heterogeneity often observed in building design and the morphological characteristics of non-domestic buildings. Additionally, the complex behaviour of airflow and thermal dynamics in non-domestic buildings could not be captured easily without advanced computational methods, which can be time-consuming and expensive. As such, modelling the non-domestic stock from an IAQ point of view cannot be achieved using the methods already used for modelling the IAQ in a housing stock using archetypes or the non-domestic energy stock models using the SCUs; thus, it requires alternative methods.

2.7 Machine Learning and Statistical IAQ Prediction Models

Machine Learning (ML) can estimate indoor air quality as an alternative method. For example, ML can be used to relate the measured concentrations of pollutants to questionnaires to estimate indoor air concentrations in indoor environments. Furthermore, ML makes it possible to predict a pollutant concentration based on other indoor and/or outdoor parameters in a building. ML is very useful, even though existing computational models (CONTAM) may appear more reliable, especially when a specific mechanism or its dynamic variation is not well established and when large data sets are available (Wei et al., 2019). The majority of ML algorithms are based on supervised or unsupervised learning. Supervised learning uses labelled examples as training data and predicts all unknown points (Mohri et al., 2012). In unsupervised learning, unlabelled training data is used to reduce, summarise, and synthesise information. While unsupervised learning cannot provide predictions for unknown data, it provides valuable insight into the structure of the data, which facilitates the selection of an appropriate supervised model, see Figure 2.8. As this thesis aims to develop a model to predict the indoor concentrations of $PM_{2.5}$, this review will focus on supervised ML models used in the domain of IAQ modelling.

An example of a supervised learning model is regression models (such as multiple linear regression MLR, generalised linear regression GLM, regularised regression, partial least squares PLS, and principle component regression PCR), decision tree models (such as gradient boosting trees and random forests), classifiers (such as Bayes classifications, KNN classifications, and support vector machines SVM), and some artificial neural networks ANNs (such as feed-forward back-propagation networks and cascade correlations) (Wei et al., 2019). Generally, these models can be divided into continuous variables (such as pollutant concentrations) and categorical

variables (such as air quality indices). Various models can be used for continuous variables, including MLR, regularised regression, PLS, and PCR (Mohri et al., 2012). Several models for categorical variables exist, including Bayes classifiers, SVM, and k-NNs. In addition, several models can address both variable types, such as GLM, Generalised Additive Models GAM, decision trees, gradient boosting trees, random forests and ANNS. Depending on the variable type, these models may be divided into linear and non-linear models.

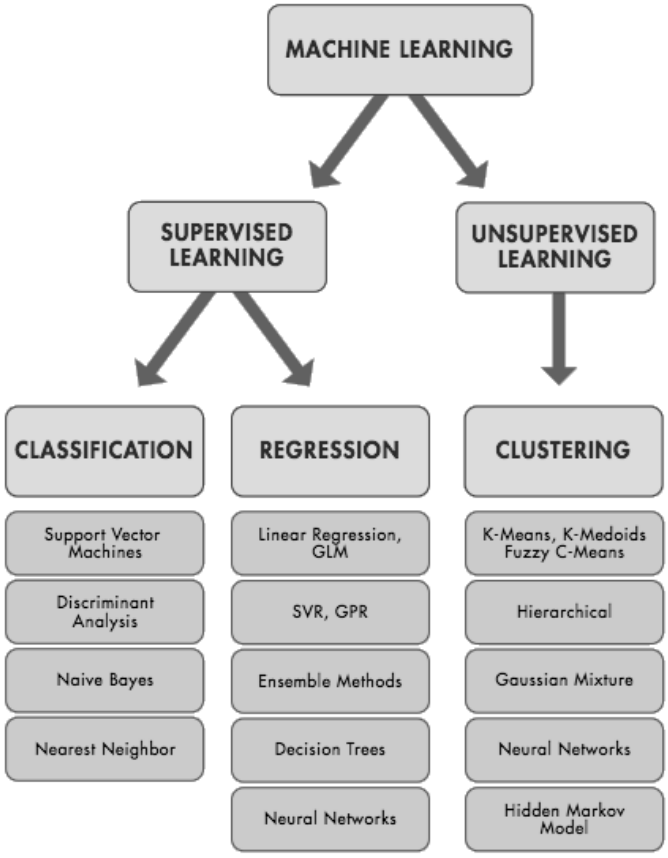


Figure 2.8: Supervised and Unsupervised Machine Learning Algorithms

Linear regression models can be used as suitable predictions when the response and prediction variables are linearly related or when the two variables are transformed into linear relationships. Normalisation, log transformation, and rank transformation can be employed when the scale of multiple variables differs significantly (Mohri et al., 2012). Due to its simplicity and clear display of the best predictors of the variable of interest, MLR is considered to be the classical regression

model. Nonetheless, it cannot handle missing data and requires more observations than variables. It is possible to minimise these negative aspects of MLR by using PLS or PCR, which groups individual explanatory variables into components to reduce their regression effects. In the prediction context, linear models are easy to develop and apply and are often considered the first measure. In cases where the response and predictive variables are unlikely linearly related, other non-linear models may be more suitable, regardless of the data structure. Non-linear modelling is discussed in detail in Chapter 5, Section 5.3.

The dataset is usually divided into three sections to develop an ML model to predict IAQ: training, validation, and hold-out. A typical training set uses about 70-80% of the data, and the model is trained using various algorithms to learn from the patterns and relationships present in the data. The training aims to develop a model that can generalise to new, unseen data and make accurate predictions. The remaining 20-30% is used for model validation and testing (Elbayoumi et al., 2015). The model is tested against the validation data, and the results are used to adjust the model's parameters, such as hyperparameters or learning rate, and to optimise its performance. In addition, the validation data helps to prevent overfitting, which occurs when a model becomes too complex and performs well on the training data but poorly on new data. In most cases, the leave-one-out cross-validation algorithm (LOOCV) is used to validate the data. (Isiugo et al., 2019).

Model testing may also be conducted using new datasets derived from different cases than those in the original study (Fernando et al., 2020). Finally, hold-out data is a separate, independent dataset used to evaluate the final performance of the trained model. It is not used during training or validation but only after the model has been fully trained. Hold-out data is critical to understanding how well the model can generalise to new, *unseen* data. The performance of the model on the hold-out data is the ultimate metric of success, as it indicates how well the model can perform in real-world scenarios.This is discussed in detail in Chapter 5, Section 5.4.

There have been several applications of ML and statistical modelling in outdoor environments to predict the concentrations of ambient pollutants (Ausati & Amanollahi, 2016; Niu et al., 2016; S. Sousa et al., 2007) and in indoor environments to predict thermal comfort (Patil et al., 2008; Soleimani-mohseni, 2007) and energy efficiency (Edwards et al., 2012; Seyedzadeh et al., 2020; Tsanas & Xifara, 2012). Among the models used in these studies are regression models, partial

least squares (PLS), decision trees (classification and regression trees), Bayesian hierarchical models, generalised boosting models, support vector machines, random forests, generalised linear models (GLM), and artificial neural networks (ANN) (Bellinger et al., 2017; Zhao & Magoulès, 2012).

The application of ML to predict the concentrations of indoor air pollutants, for example, PM, NO_2, VOCs, and CO_2, is much less advanced than applications to determine the concentrations of outdoor air pollutants. However, ML can be used to predict indoor air quality in an existing building when using secondary sources of data (questionnaires and/or measurements) (Elbayoumi et al., 2015) or when the primary source of data is simulation results (Das et al., 2014). Among the models used in the domain of IAQ modelling are the ANNs, Regression models, and Decision trees. Several regression models, including MLR and stepwise regression models, have been used to investigate indoor PM and NO2 levels in schools and dwellings (Elbayoumi et al., 2015; Jafta et al., 2017; Kropat et al., 2015; Yuchi et al., 2019). Other regression models have been developed to predict indoor radon concentration at large scales in Switzerland and Italy, such as kernel regression and Bayesian spatial quantile regression (Kropat et al., 2015; Sarra et al., 2016).

In a given environment, the regression model's performance is primarily determined by the selection of inputs. For instance, an MLR model with outdoor $PM_{2.5}$ concentration and indoor relative humidity as inputs predicted $PM_{2.5}$ in a school for three consecutive days during school hours with an R^2 value of 0.58 (Elbayoumi et al., 2015). However, when ventilation rate, wind speed, and indoor temperature were included as inputs, the R^2 value reached 0.69. This is because the ventilation rate can strongly modify PM's indoor/outdoor transport. A regression study of indoor PM2.5 concentrations compared MLR, LASSO, and stepwise regression (Yuchi et al., 2019). During the training period, the LASSO and stepwise regressions performed better, as indicated by R^2 and root mean square errors (RMSE), than MLR, resulting from better variable selection procedures. It was observed, however, that the regression models performed similarly during the validation period. As part of the prediction of indoor PM concentrations, two studies developed both MLR and ANN models (Kim et al., 2009). As a result of the non-linear relationship between inputs and outputs for MLR during the training period, the RMSE values for MLR are generally more significant than those for ANN during the training period. Nevertheless, both MLR and ANN have substantially higher RMSEs during the validation and testing phases and are similar.

A series of studies were conducted in dwellings, a university building, and a hospital to predict $PM_{2.5}$ concentrations, CO_2, radon, and viruses. $PM_{2.5}$ predictions were based on outdoor PM concentrations (Yuchi et al., 2019), and virus predictions were based on fine dust particles (Choi et al., 2017). The R^2 values ranged from 0.74 to 0.94 for training and 0.33 to 0.49 for validation. For the studied cases, the decline in R^2 values for validation indicates that the trained model may require improvement to provide more accurate predictions. In addition, the performance metrics RMSE and index of the agreement also exhibit similar gaps between the training and validation sets.

When comparing two random forest regression (RFR) models of $PM_{2.5}$ concentrations in dwellings with MLR (Yuchi et al., 2019), the RMSE values for the RFR models were smaller than those for the MLR for both prediction methods. Moreover, an Extreme Gradient Descent Boosting model (XGBoost) was used to predict CO_2 concentrations in a university building (Martínez-Comesaña et al., 2022). In the study, the results demonstrated that the built XGBoost model is capable of estimating the 1-min CO_2 concentrations of a building efficiently. The average CV(RMSE) yielded was below 10% for CO2 levels.

There have been multiple studies on using ANNs for the prediction of IAQ in buildings, including dwellings, schools, and offices, to address several IAQ variables using various techniques (Challoner et al., 2015; Elbayoumi et al., 2015; Liu et al., 2017; Skön et al., 2012). In ANN applications indoors, PM concentration was the most studied parameter. The models have shown good performances (R^2 ranging from 0.62 to 0.79, a normalised absolute error between 0.01 and 0.19, and an index of agreement between 0.89 and 0.95). Figure 2.9 shows a list of inputs used in the development of models. Most prediction models used regional environmental variables, such as outdoor PM concentrations, temperatures, and wind speeds, as inputs. The outdoor variables can reasonably explain indoor air PM originating from the outdoors.

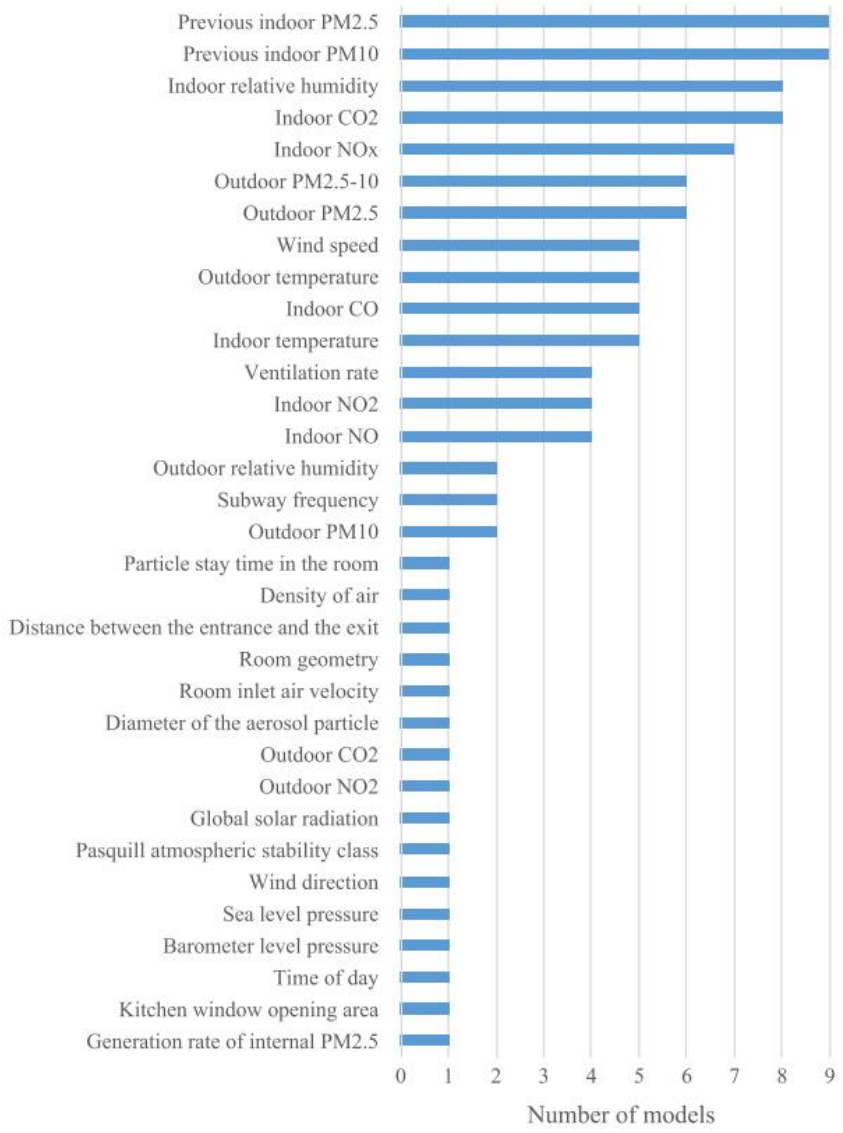


Figure 2.9: List of input variables used in previous models to predict indoor PM concentrations (Wei et al., 2019)

An investigation of $PM_{2.5}$ and PM_{10} concentrations in twelve naturally ventilated schools in Palestine found an R^2 value between 0.65 and 0.79 (Elbayoumi et al., 2015). A limited number of models are available for other indoor environments, such as dwellings, where indoor sources are negligible. To address the issue of indoor sources, (Das et al., 2014) proposed a few indoor variables, such as the rate at which internal $PM_{2.5}$ is produced. However, the importance of this kind of input cannot be overstated since it is not easy to provide and is critical to the prediction model. Hence, further studies are required to determine the inputs and performance of the models when strong indoor sources are present.

2.7.1 Model Transparency

The transparency of models becomes a concern when computational and ML methods are used to make predictions. Previous research has suggested classifying model transparency into *White-Box Models* (Physics-Based Models), *Black-Box Models* (Data-Driven Models), and *Grey-Box Models* (Hybrid Models) (X. Li & Wen, 2014). Although data-driven *(Black-Box)* IAQ modelling for individual buildings has gained increased attention in the last decade (Wei et al., 2019), current building stock *IAQ* models rely on building physics models in their simplest form for data collection and validation. In comparison, historical IAQ data measured at the stock level are scarce. Based on this, the past and current building stock IAQ models identified in this thesis are *White-Box* or *Grey Box* models.

Accessibility or transparency is a pre-condition to achieving model reproducibility, representing the minimum attainable standard compared to replicability (Morrison, 2018). It has been suggested previously that black-box data-driven models are potentially not reproducible. As a result, they are constrained by limited applicability specific to the range of datasets used in developing the models (J. Li et al., 2020). For example, a model that was trained to predict the IAQ by learning from limited datasets (e.g., data collected from a small group of buildings) may not perform well outside of the training data (e.g. different physical properties, occupant behaviour, climate context, future interventions, chaotic events, etc.) (Tardioli et al., 2015). Thus, for non-expert users, the purpose of prediction should be made clear, and guidance on whether the models apply in a new context should be provided.

White-Box models offer higher transparency by releasing and maintaining the core calculation algorithms as open-source programs (e.g., CONTAM and EnergyPlus). The high transparency offered makes white-box models highly reproducible and versatile. However, there are foreseen issues surrounding the deployment of such models (O'Leary, Jones, et al., 2019): (1) these models can be oversimplified when the spatial resolution is increased, i.e., a specific level of abstraction or spatial resolution, therefore, outputs could be erroneous; (2) expert knowledge is required when model assumptions are made or when prediction outputs need interpretation, and (3) assumptions pertaining to the input variables of these models are prone to all kinds of uncertainties.

Open-source IAQ simulation engines allow user interaction or integration of scripting tools such as the EnergyPlus Generator 2 Tool (EPG2) developed in Python (Rossum & Drake, 2022a). For example, EPG2 was used in the ENG-B model (Symonds et al., 2016) for batch-processing building input files configured with user-defined variables. In addition, the REIAQM project (Fazli & Stephens, 2018) used multiple Python scripts to solve mass balance equations and automate most simulation processes. This can allow flexibility, input variability, automation, and increased computation versatility (X. Li & Wen, 2014). However, *White-Box* housing stock IAQ models tend to be static and deterministic, often assuming linear relationships exist between multiple variables in an ideal system without uncertainty (Oladokun & Odesola, 2015).

Alternatively, developing *Grey-Box (hybrid)* models integrating physics-based models and multiple statistical analyses can account for uncertainty assessment and quantification. This has been achieved by deploying sampling methods such as Latin Hypercube for Monte Carlo integration. Although the near-random samples generated could be pretty significant, applying a statistical significance test can reduce the number of samples required to represent the entire building stock with reduced model resolution (e.g. CHAARM (Molina, Jones, et al., 2020)). Furthermore, hybrid models can account for linear and non-linear systems by constructing metamodels (e.g. Artificial Neural Networks) and using them in predictions. This is particularly interesting as non-linear and irregular behaviours best characterise air pollutant concentrations outdoors and indoors due to behavioural, social, and chaotic events (Chelani & Devotta, 2006). However, in contrast to *White-Box* models, some hybrid models (for instance, the ENG-A (Das et al., 2014) and ENG-B models (Symonds et al., 2016)) suffer from reduced transparency and accessibility, particularly in the metamodel construction phase, whereby multiple hidden layers and neurons generate outputs that are extremely difficult to replicate.

In contrast to black-box models, grey-box models can be *scalable* and *versatile*. For instance, the ENG-A and ENG-B models are based on multiple metamodels constructed individually for each housing stock's physical properties, locations, epochs and occupancy profiles. This makes the models *scalable* to include additional information without reconstructing the entire model from scratch and *versatile* in comparing the results of different what-if scenarios (e.g. seasonal variation, future technological interventions, chaotic events, etc.). On the other hand, despite the aforementioned advantages, grey-box models may incur higher computation costs. Both white- and grey-box models can be computationally expensive when many *archetypes* or *metamodels* are involved. However, some of these models managed computing efficiency by (1) using only

single-zone models to represent the entire building stock, such as REIAQM (Fazli & Stephens, 2018) (but with reduced prediction accuracy for both airflow and contaminant concentrations); (2) running the simulations on a high-performance parallel computing platform as in ENG-A (Das et al., 2014) and ENG-B models (Symonds et al., 2016); and (3) reducing the number of archetypes to a statistically acceptable level while acknowledging the loss in model resolution as in the CHAARM (Molina, Jones, et al., 2020).

2.8 Conclusions

In this chapter, an investigation has been undertaken through a comprehensive review of existing literature on the field of IAQ. The review was based on a systematic selection of relevant journal articles using the keywords "IAQ" AND "prediction" AND "stock modelling" AND ("building stock" OR "housing" OR "domestic" OR "deterministic" OR "probabilistic" OR "metamodelling" OR "sensitivity analysis" OR "building simulation" OR "multi-zone model" OR "machine learning" OR "statistical models"). The review was dedicated to determining the methods and techniques used to develop building stock IAQ models. At first, an introduction to outdoor air pollution and its risk to population health was presented with a particular focus on $PM_{2.5}$. Then, the review examined indoor air pollution, its causes, and methods to assess IAQ in buildings. It is clear from the review that using computational methods (e.g., simulation models) to simulate the concentrations of indoor air pollutants in building stock is considered a preferable alternative to the direct methods.

This was followed by a review of the existing literature on the methods and techniques for modelling a building stock. Based on the review, it is evident that current building stock IAQ models adhere to the same methods used to model building stock's energy demand. It should also be noted that existing non-domestic building stock models were limited to energy use in commercial buildings. This review concludes by examining the use of data-driven methods in modelling indoor air quality in different buildings. This review identified MLR, RFR, XGboost, and ANNs as the most commonly used algorithms in predicting indoor pollutant concentrations. With the current development of ML algorithms and the increasing amount of data collected in buildings, it is clear that ML can provide researchers with methods for modelling indoor air quality that is computationally inexpensive. However, there are concerns regarding

model transparency and reproducibility, as some ML algorithms may be considered black boxes. Below is the summary of the key findings from this literature review:

Gap (1): The current attempts to model indoor air quality in a building stock are confined to residential buildings, with no studies conducted for non-domestic buildings. Considering how much time people spend in offices, schools, and other working environments, it is evident that this issue needs to be addressed since people are likely to be exposed to indoor air pollutants of different magnitudes in these environments. Nevertheless, further research is required to determine the data sources and methods required to perform IAQ models for non-domestic building stocks.

Gap (2): Existing housing stock IAQ models can be oversimplified when the spatial resolution is increased, i.e., a specific level of abstraction or spatial resolution (single-zone models) or when reducing the number of archetypes used to represent a building stock, therefore, outputs could be erroneous.

Gap (3): The complex and dynamic nature of modelling the IAQ does not consider the dynamic variations of indoor air temperatures. Current building stock IAQ models use airflow models only and do not solve the heat balance equations.

Gap (4): The IAQ models developed using machine learning algorithms for residential buildings are limited to Artificial Neural Networks (ANNs). Although they have been demonstrated to be accurate predictors of indoor pollutant concentrations, they are still black-box models regarding model transparency and interpretability.

Chapter 3 Research Methodology and Data Sources

3.1 Introduction

Improving the environmental performance of HEI building stocks is becoming increasingly important in achieving higher energy efficiency and better indoor environment quality, including IAQ, and identifying the prime targets for reducing carbon emissions and the health risks associated with poor IAQ. Due to their volumes and energy use intensities, HEI buildings in the UK bear a significant portion of a city's overall carbon emissions and are subject to social, economic, demographic and technological changes. In addition, there are constant demands for retrofitting and constructing new buildings to accommodate such changes. Decision-making needs reliable models to inform strategies to maximise HEI building stock performance while minimising negative environmental and health impacts. Having actionable policies/guidance to reduce or remove pollutant emissions at their sources effectively is preferable, but there may be limits to what HEIs can achieve independently. This chapter presents a case for pursuing new research to investigate sensitivity-based data-driven modelling to predict infiltrated $PM_{2.5}$ concentrations starting from a zonal level. Seeing an HEI as the key stakeholder, the proposed new modelling framework and capability aims to enable fine-grained estimation of population exposures to $PM_{2.5}$ as the basis for evidence-based decision-making.

As reviewed systematically in Chapter 2, different methods can be used to study the IAQ in a single building and the building stock. They can be classified into two groups: direct and indirect. Direct methods include field measurements using mobile or stationary equipment. Alternatively, indirect methods comprise computational modelling, simulation, and statistical techniques. Simulation models can offer several advantages over field measurements in certain situations, such as cost-effectiveness, flexibility when assessing different scenarios and interventions, and time efficiency. However, it is crucial to acknowledge that simulation models rely on simplifications and presuppositions, and their precision is contingent on both the quality of the

input data and the intricacy of the model itself. Consequently, field measurements are indispensable in validating, calibrating, and refining simulation models.

Moreover, when sufficient high-quality input data is available, both modelling and sensitivity analysis methods can calibrate/validate model predictions and inform uncertainties. The model outputs can be used as data/evidence to inform interested parties about the status quo or probable intervention scenarios. Adopting a bottom-up approach, a modeller can evaluate various factors contributing to IAQ by conducting a sensitivity analysis of the outputs of each input factor. This method has proven beneficial, particularly in countries with limited data. However, to the author's knowledge, HEI stock IAQ models are currently non-existent.

3.2 A Methodological Framework

This research investigates how population exposure to infiltrated $PM_{2.5}$ at an HEI building stock level can be estimated to support planning and design for better IAQ. With the heterogeneity and complexity often observed in HEI buildings, variations of indoor $PM_{2.5}$ concentrations from building to building are anticipated. Therefore, combining building physics and statistical modelling techniques, a hybrid bottom-up approach is proposed to decompose multiple buildings selected from an HEI stock into a structured cohort of individual spaces or rooms. The model resolution at a room level allows for sensitivity analyses of modelled indoor $PM_{2.5}$ concentrations to the built and environmental characteristics. As obtaining field measurements of $PM_{2.5}$ at a large scale can be time-consuming and prohibitively expensive, this research utilises an IAQ and heat transfer coupled simulation platform[4] (CONTAM-EnergyPlus co-sim) developed at NIST, USA.

A wide range of inputs must be considered to accurately simulate mass transfer in buildings. However, it is essential to note that some information may not be available, so assumptions must be made. Additionally, when data cannot be tracked, the collation and processing of data can be time-consuming, resulting in systematic errors. Therefore, any tool used to model a stock must be informed by the most reliable sources of information. As shown in Figure 3.1, the proposed overall methodological framework begins with five buildings at the University of Sheffield (UoS) selected for this research, mainly due to their distinctive morphological characteristics. Unlike domestic building stocks, where input data to model the IAQ can be informed by national housing

[4] https://www.nist.gov/services-resources/software/contam

surveys (e.g., the English Housing Survey), sufficient information and high-quality data on HEI building stocks are often not readily available. However, data from existing surveys and previous research on domestic building stocks may still be referred to inform building and environmental data for some of the UoS buildings as case studies. This is discussed in detail in Section 3.3.

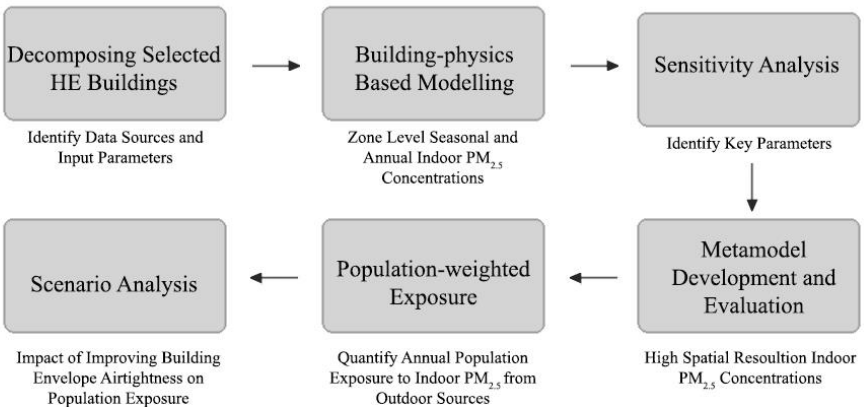


Figure 3.1: A workflow diagram showing the processes involved in carrying out deterministic bottom-up HEI building stock IAQ modelling

Two open-source building-physics simulation packages (CONTAM and EnergyPlus) were used to develop detailed multizone IAQ models of the selected UoS buildings to estimate indoor $PM_{2.5}$ concentrations (see previously in Section 2.4.3). Here, unlike the housing stock IAQ models, the rooms/spaces in each UoS building were modelled as individual zones rather than one zone per building/floor. This ensures that detailed airflow networks reflecting the complexity of the UoS buildings were included to account for the spatial variations of indoor $PM_{2.5}$ concentrations. As such, by adding the airflow paths, building elements, and indoor sinks of $PM_{2.5}$, the multizone modelling results in two types of model files: CONTAM (.prj) and EnergyPlus (.idf). In the model development, some key model parameters were varied across the five buildings while others were held constant, which is discussed in more detail in Chapter 4.

The resultant concentrations of infiltrated $PM_{2.5}$ were checked zone by zone and were used to perform sensitivity analyses (Das et al., 2014). This is to assess the sensitivities of the input variables to the output of indoor $PM_{2.5}$ concentrations. The sensitivity analysis framework is described in detail in Chapter 5, and the results are presented in Chapter 6. Predictive metamodels were developed through machine learning (ML) using the reduced set of input variables. The metamodel development process involved an investigation of three regression-based ML

algorithms chosen for the study. The algorithm with the highest prediction accuracy was selected to estimate the population exposure to infiltrated $PM_{2.5}$. The framework of the ML model development is presented in Chapter 5, and the results of metamodel development are presented in Chapter 6. Existing studies suggest that estimating the population exposure to indoor air pollutants can be achieved through a *microenvironmental modelling* approach using indoor concentrations of $PM_{2.5}$ and the number of occupants as weights. Microenvironmental modelling of population exposure to $PM_{2.5}$ is discussed in detail in Chapter 7. Figure 3.2 presents a more detailed research methodological framework consisting of six key stages:

Stage (1): Decomposition of an HEI Building Stock

At this stage, it is necessary to identify the sources of building stock data for HEIs that can be used to model IAQ. Previous research on IAQ and a literature review are used throughout this stage to guide the data collection process. The objective is to decompose the selected buildings into a cohort of individual rooms. As a result of the proposed model resolution, it is expected that it will be possible to estimate infiltrated $PM_{2.5}$ concentrations in indoor environments considering spatial variability.

Stage (2): Building Physics-Based Modelling

Generation of coupled multizone models (CONTAM and EnergyPlus) using the data identified in stage (1). Here, the simulations will run for the whole year using a time step of 15-min intervals. The outputs of this stage include hourly and seasonal values of indoor $PM_{2.5}$ concentrations, indoor temperature, and infiltration (ACH_{INF}) for each room within the selected buildings. This stage is discussed in detail in Chapter 4.

Stage (3): Sensitivity Analyses

The sensitivity analysis framework will determine the relationships between the inputs and the outputs. The scatter plot of inputs versus the output illustrates the relationships between the individual inputs and the output for visual inspection. Here, the output from this stage will help identify the more critical and related inputs for developing the metamodels. This stage is discussed in detail in Chapters 5 and 6.

Stage (4): Metamodel Development, Tuning and Evaluation

This stage involves following a metamodeling framework to rapidly estimate the spatial variations of infiltrated $PM_{2.5}$ concentrations in an HEI building stock from a set of key explanatory variables identified in stage (3). Here, the selection of candidate ML

algorithms, metamodels development, tuning, and evaluation will be carried out to ensure that the metamodels fit the selected variables.

Stage (5): Metamodel Interpretation through SHAP (SHapley Additive exPlanations)

An essential contribution of this research will be the application of SHAP values to increase the transparency and interpretability of the developed metamodels and statistically quantify the contribution of each input variable to the predicted indoor $PM_{2.5}$ concentration. The significance of this stage will be highlighted by answering research questions (1) and (3) of this thesis.

Stage (6): Microenvironmental Modelling for Exposure Assessment

Based on the metamodel's estimations for various zones, this stage proposes a microenvironment modelling approach to estimate the average Personal Exposure (E_i) to infiltrated $PM_{2.5}$ and the average Population-Weighted Exposure (PWE) to infiltrated $PM_{2.5}$ for different microenvironments.

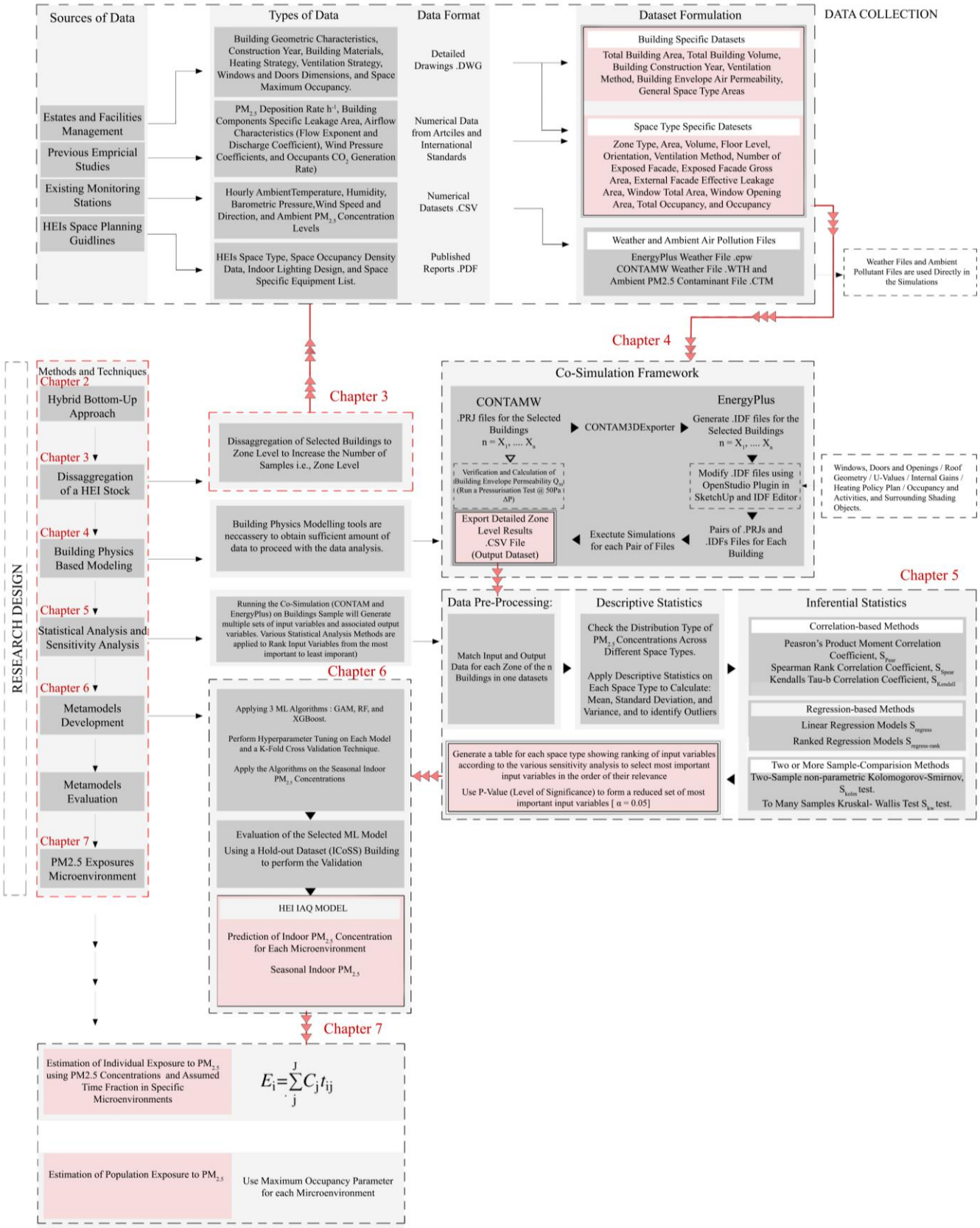


Figure 3.2: A detailed research methodological framework for developing a HEI stock IAQ model to predict the heating season infiltrated $PM_{2.5}$ concentrations and annual population exposures

3.3 The Data Sources

Compared to residential building stocks, collating a representative database of buildings for an HEI building stock can be challenging and technically complex in different manners. This is due to the high heterogeneity in HEI buildings regarding their sizes, functions, designs, constructions, and building uses. As such, this study was designed with *selected* buildings rather than *representative* buildings in mind. How to develop statistically representative archetypes of an HEI building stock is beyond the scope of this research. Here, the idea was to work with an initial selection of buildings while collecting as much data as possible. In what follows, this section introduces the data sources held by the University of Sheffield Estates and Facilities Management (EFM) department and relevant governmental and non–governmental organisations. These sources include general building information and layouts, HVAC systems in use and operational details, heating policies, lighting and appliances, building envelope construction details, U-values, occupancy schedules, and weather and air pollutants monitoring stations.

The data from five UoS buildings were collected for experimental IAQ model development: The Arts Tower (AT), Regent Court Building (RC), Academic Development Centre (ADC), Barber House (BH), and the Interdisciplinary Centre of Social Sciences (ICoSS). These buildings were selected to reflect the UoS building stock of different ages (the 1920s-2000s). Additionally, the buildings differ in size, geometry, construction methods and materials; thus, different building-related input parameters to model the IAQ were required. Moreover, HEI buildings tend to be composed of purpose-built spatial volumes (e.g., classrooms, student-led learning spaces, laboratories, staff offices etc.) connected by circulation routes, often resulting in large built areas exposed to external thermal and air flows.

Figure 3.3: The five buildings selected from the University of Sheffield (UoS) building stock

3.3.1 Building Design and Characteristics

The UoS EFM department maintains records of all university sites, buildings and facilities. It is the primary source of information regarding building design and operation. The EFM maintains various sorts of building information and data, such as computer-aided design (CAD) drawings, facility management databases, occupancies etc. There is also a wide range of data recording the room areas and usage types. Moreover, as part of the UoS Energy Strategy to reduce overall carbon footprint, the EFM team keeps energy use-related data, including HVAC systems, thermostat setpoints, heating policy, indoor lighting design, and appliances. During this research, the Space Management team was beneficial in providing copies of the site and building plans on request. Table 3.1 summarises the main characteristics of the selected buildings.

Table 3.1: Summary of the features of the five selected UoS buildings

	Barber House	**ADC**	**The Arts Tower**	**Regent Court**	**ICoSS**
Construction Period	1920s	1940s	1960s	1990s	2000s
Function	Offices, Seminar Rooms, and Meeting Rooms	Offices, Seminar Rooms, Meeting Rooms, and Open-office Style Study Space	Offices, Seminar Rooms, Meeting Rooms and Studios	Large Computer Rooms, Open-office Style Study Space, Lecture Theatres, Cellular Staff-Offices, and Meeting Rooms	Labs and Seminar Rooms
Distinctive Feature	Linear Circulation System	Compact Floor Layout with Compound Circulation	High-rise Central Core with Radial Circulation	Courtyard with Linear Circulation	Atrium Building
Refurbished	Yes	Yes	Yes	No	No
Height Classification	Low Rise	Low Rise	High Rise	Low Rise	Mid Rise
Total Number of Floors Above Ground	2	2	20	3	5
Building Built-Up Area (m^2)	874.48	1,351.25	16,402.36	9,057.09	1,947.80
Ventilation Method	Natural Ventilation W/ Exhaust System	Natural Ventilation W/ Exhaust System	Natural Ventilation W/ Exhaust System	Natural Ventilation W/ Exhaust System	Natural Ventilation W/ Exhaust System
Heating Method	Gas Fired Wet Heating System	Central Heating via Wall Mounted Radiators	Central Heating via Wall Mounted Radiators	Central Heating via Wall Mounted Radiators	Central Heating via Wall Mounted Radiators
External Walls	Solid Wall	Solid Wall	Double Glazed Curtain Wall System	Cavity Wall	Cavity Wall and Pre-painted Copper Sheets
External Walls U-Values ($W/m^2.K$)	1.80	1.80	2.20	0.60	0.45

Categorising energy use activities according to standard space-type functions seems practical in HEI. Therefore, a total of 15 space types were identified according to the UoS Energy Strategy (Arup, 2012), which form the basis of the building elements of the Energy Strategy (Table 3.2).

Table 3.2: The space types identified in the UoS Energy Strategy 2012 (Arup, 2012)

Room/Space Type	
1. Lecture Theatres	2. High Energy Usage Laboratory
3. General Offices	4. Low Energy Usage Laboratory
5. Classrooms/Seminars Rooms	6. Clean Room Laboratory
7. ICT Suite	8. Circulation/Lobby Spaces
9. Retail and Leisure	10. Back of House
11. Kitchen	12. Accommodation
13. Cold Rooms	14. Toilets and Changing Areas
15. Library	

In this study, these room/space types form the basis for developing an IAQ-specific space-type categorisation to account for the time spent in each space type in the $PM_{2.5}$ exposure estimation (see Chapter 7). However, some short time spent in some spaces can be considered negligible compared to the total time spent in a higher education building. As such, a revised space-type categorisation of 14 types is proposed in this thesis: 1. Academic Offices, 2. Administration Offices, 3. Post-Graduate Research Offices, 4. Lecture Theatres, 5. Laboratories, 6. Library, 7. Seminar Rooms, 8. Educational Facilities (teaching spaces were combined here), 9. Kitchens, 10. Shared Facilities (Toilets and Changing Areas were combined here), 11. Cold Rooms, 12. ICT Suite, 13. Services (including Back of House), and 14. Circulation (corridor, staircase, lobby).

3.3.2 Airtightness

According to the American Society for Testing and Materials (ASTM) (ASTM, 1999), a building's envelope permeability is conventionally determined using a fan pressurisation test. This test systematically and artificially increases the difference between the external and internal air pressures (Pa) to measure the airflow rate through adventitious openings within the envelope $\dot{V}$ (m^3/h). These parameters are related by a power law (W. Stuart Dols & Polidoro, 2020):

$$\dot{V} = C\{\Delta P\}^n \tag{3.1}$$

where C ($m^3/h\ Pa^{-1}$) is a flow coefficient, n is a flow exponent, $\dot{V}$ is commonly reported at 50 Pa, interpolated from measurements, when it is known as an air leakage rate, $\dot{V}_{50}$ (m^3/h). Comparing

the air leakage rates of different buildings is achieved by normalising the $\dot{V}_{50}$ by a common parameter, such as the external envelope area when it becomes an Air Permeability, Q_{50} ($m^3/h/m^2$), or the building's total volume to give N_{50} (h^{-1}). However, comparing buildings with substantially different forms and volumes is best achieved using Q_{50}.

According to (Etheridge, 2011), the usual operational pressure differences are about 4 Pa, roughly one-tenth of 50 Pa. Etheridge also suggests that an appropriate value for Q_4 can be inferred from the acceptable value of Q_{50} if n is known, as noted by (Benjamin Jones et al., 2015). Measuring the air leakage rate is achieved through the standard method, which involves determining C and n values by measuring $\dot{V}$ at intervals between $10 \leq \Delta P \leq 100$ Pa, as specified by ASTM and BSI (ASTM, 1999; BSI, 2001). This helps to minimise the effect of noise from naturally occurring wind and buoyancy on the measurements of $\dot{V}$. However, systematic uncertainty is associated with the measurement because the shape of the leakage characteristic is unknown when $0 < \Delta P <$ 10 Pa, and the equation used may not be valid during this range, as noted by (Cooper et al., 2007).

A previous publication by Hurel and Leprince (2021), which included an in-depth review of relevant literature, explored the impact of wind on the uncertainty associated with airtightness testing. They found that the model error due to wind on estimated airflow rate was relatively insignificant at high-pressure points (up to 12% for wind speeds of up to 10 m/s at 50 Pa). However, it could be significant at low-pressure points (up to 60% at 10 Pa). Therefore, when estimating airtightness at 4 Pa, wind could introduce substantial errors, sometimes exceeding 35% (Bailly, Leprince, Guyot, Carrié, & Mankibi, 2012). In addition, the stack effect can also contribute to non-uniform pressure differences along the envelope of tall buildings, leading to inaccuracies in airtightness testing results (Carrié & Leprince, 2016).

Alternatively, Eq (3.2) describes a quadratic equation that can be used in preference to the power law model because it is believed to depict the flow behaviour of adventitious openings accurately (Cooper et al., 2007).

$$\Delta P = aQ^2 + bQ \tag{3.2}$$

The first term aQ^2 accounts for the momentum change observed in openings with variable geometry. The second term (bQ) corresponds to surface friction and is observed in long gaps with a fixed geometry (Zheng et al., 2020; Zheng & Wood, 2020). However, the issue with determining the coefficients a and b in the quadratic equation used to model infiltration is that they are not uniquely identifiable based solely on the measured infiltration data (Zheng & Wood, 2020). This

is because the quadratic equation has two coefficients, but infiltration data only provides information about the overall shape of the curve, not about the specific values of the coefficients. Therefore, additional information, such as measurements at different pressure differences or consideration of the building's envelope characteristics, is needed to determine the values of a and b accurately. Without such information, the resulting values of a and b could be inaccurate, leading to unreliable infiltration predictions. To simplify the modeling of infiltration in this thesis, Eq (3.1) will be used, while acknowledging the potential uncertainties associated with this approach.

Airtightness values are often used to estimate the rate at which unconditioned ambient air *infiltrates* a building through adventitious openings. However, as Q_{50} values might not be available for non-domestic building stocks in the UK, data from the CIBSE TM23 (CIBSE, 2022) standard alongside the construction year for buildings were used to estimate the Q_{50}, see Table 3.3.

Table 3.3: CIBSE TM23 UK Standard for Allowable Airtightness in Buildings (CIBSE, 2022)

Building Type	**Building Airtightness Q_{50} ($m^3/h/m^2$ @ 50Pa)**	
	Best Practice	**Normal Practice**
Offices (Naturally Ventilated)	3.0	7.0
Offices (Mixed Mode Ventilation)	2.5	5.0
Offices (Air Conditioned)	2.0	5.0
Schools	3.0	9.0

According to the study conducted on the Arts Tower (AT), the Q_{50} before retrofitting was approximately 23 $m^3/h/m^2$ (Everett, 2013), which was reduced to approximately 10 $m^3/h/m^2$ after the retrofitting specified by the HLM Architects in 2009 (Mara, 2010). Based on the data from the EFM, the Barber House (BH) and the Academic Development Centre (ADC) buildings, dating back to the 1920s and 1940s, were assumed to share similar construction materials with several other dwellings built during that period. As such, the Q_{50} for the buildings were assumed to be 13 $m^3/h/m^2$. According to CIBSE TM23 (CIBSE, 2022), all buildings over 1,000m^2 should have a maximum air tightness of 10 $m^3/h/m^2$ after 2002. The Regent Court (RC) building dates back to 1995; however, it was assumed to have a Q_{50} of the maximum allowable Q_{50} of 10 $m^3/h/m^2$. Furthermore, the ICoSS building was assumed to have the best Q_{50} value of 7 $m^3/h/m^2$, representing an environmentally conscious green building design at the time of construction (Arup, 2012).

These values may underestimate or overestimate the air tightness, but they are likely to represent the *typical* Q_{50} values for non-domestic buildings in the UK according to the CIBSE TM23 standard for several types of buildings. However, as the aim of this thesis is to assess the impact of increasing the building's airtightness on infiltrated $PM_{2.5}$ concentrations and population exposure, all buildings are modelled with $3 \leq Q_{50} \leq 13$ $m^3/h/m^2$ so that any changes to indoor $PM_{2.5}$ concentrations as a result of improved airtightness can be quantified.

Table 3.4: Baseline airtightness Q_{50} values for the five selected buildings

Building	Airtightness Q_{50}	Construction Year	Reference
Barber House	13 $m^3/h/m^2$	1920s	-
Academic Development Centre	13 $m^3/h/m^2$	1940s	-
Arts Tower	10 $m^3/h/m^2$	1960s	HLM Architects
Regent Court	10 $m^3/h/m^2$	1990s	CIBSE TM23
ICoSS	7 $m^3/h/m^2$	2000s	CIBSE TM23

3.3.3 Ventilation Assumptions

There are no reported measurements in the literature of window opening behaviour in the UK's HEI building stock. Without this knowledge, the models will consider a multiplier for every window where 1 is fully open (during the non-heating season), and 0 is fully closed (during the heating season). Furthermore, it should be noted that certain areas, such as lecture halls, are equipped with Air Handling Units (AHUs) designed to supply fresh air to indoor environments, given their classification as high occupancy spaces. However, it is important to acknowledge that these AHUs were intentionally excluded from the CONTAM models employed in this project. The primary objective of the research was to investigate the effects of enhancing the airtightness of buildings on the infiltration-induced presence of indoor $PM_{2.5}$ originating from the outdoor environment. The ramifications of this particular assumption are examined and elaborated upon in Chapter 8 of this thesis.

3.3.4 Weston Park Weather Station

In order to model the actual performance of a building or group of buildings, two types of weather data can be used: Actual Metrological Year (AMY) and Typical Metrological Year (TMY) data. TMY data files are created by looking at 15-30 years of hourly data at the site in question and selecting, in series, the most typical January, February, … and so on of all available years based on the weighted average of eleven weather variables. As such, TMY files are likely to miss the extremes and not likely to be local. Thus, they do not reflect the weather conditions at a site. On

the other hand, AMY data files provide actual hourly datasets over a time period in which a study is conducted. Therefore, AMY data is essential if more extreme weather conditions are considered to cross-check design performances under stress. However, AMY files must be as close to the buildings under investigation as possible.

Weston Park Weather Station (Latitude 53.38, Longitude -1.49) is a weather station in Weston Park, Sheffield. It was founded in September 1882 and is managed by the Weston Park Museum. It provides continuous hourly data for air temperature, air pressure, rainfall, wind speed and direction, humidity, sunshine, and solar radiation. Access to this weather data can be arranged by contacting the staff at the Weston Park Museum. It is essential to point out that the criteria for selecting the weather data were by looking at (1) the year with some extreme conditions, (2) the year where there were no missing values in the requested dataset, and (3) the year where no interruptions caused by large-scale events (e.g., the COVID-19 pandemic).

Figure 3.4: Weston Park Weather Station, Sheffield (photographed by the author, 2022)

For this study, the weather data from 2018-2021 was requested from the Weston Park weather station. The year 2019 was the only year meeting the above selection criteria. Figure 3.5 shows that in 2019, Sheffield recorded its highest-ever temperatures, with 18.2 °C in February and 35.6 °C in July. In April, the highest recorded temperature was 26.4 °C; however, in 2019, the temperature was 23.7 °C, which is still above the typical April maximum temperature of 20 °C. In January, the average minimum temperature is usually 2 °C; however, in 2019, it was recorded at -4.70 °C, close to the minimum temperature ever recorded in January 1993 (-6.5 °C). In November 2019, the minimum temperature recorded was -2.45 °C, below the typical minimum temperature of -1.5 °C.

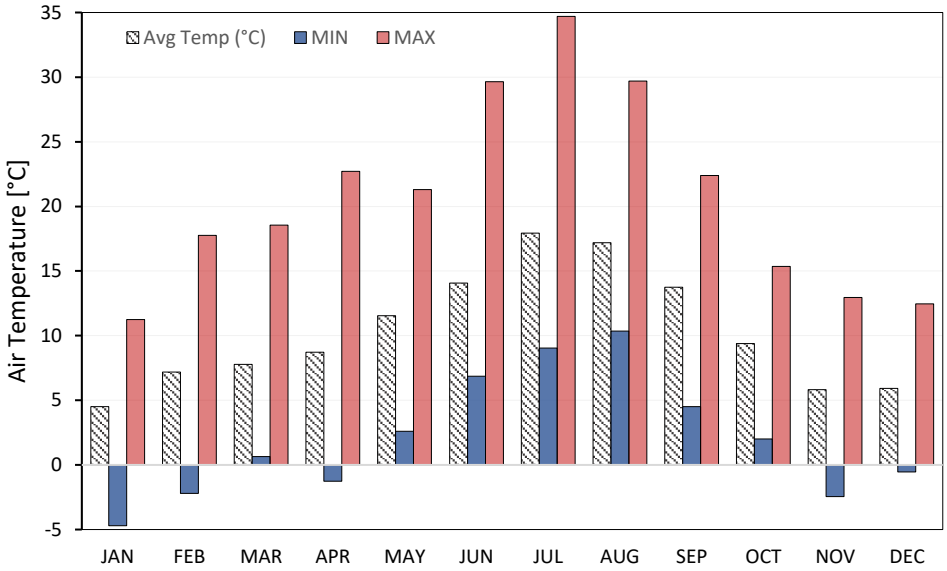


Figure 3.5: Monthly Maximum, Average, and Minimum Temperatures for Sheffield (Weston Park Weather Station, 2019)

3.3.5 DEFRA Air Quality Monitoring Station

The UK's Department for Environment, Food and Rural Affairs (DEFRA) is a ministerial environmental monitoring and protection department. DEFRA provides outdoor air quality information online via its UK Air Information Resource website. It is considered the primary source of information regarding air pollution in the UK nationally and locally via its air pollution monitoring stations. In addition, DEFRA uses a network of automatic monitoring stations across the UK for PMs. The reference method used by DEFRA to measure PM is based on the European Union's (EU) Air Quality Directive. This method is known as the reference equivalent method to measure PMs with a diameter of 10 micrometres or less (PM_{10}) and a diameter of 2.5 micrometres or less ($PM_{2.5}$). It involves two main methods, the gravimetric method and the optical method.

The gravimetric method involves collecting PM on a filter and then measuring the mass of the filter before and after sampling. The difference in mass is used to determine the amount of PM collected. The advantage of this method is that it accurately measures PM's mass concentration. However, it does not provide information about the size or chemical composition of the particles. In the optical method, a device such as a photometer or a nephelometer detects the amount of light scattered or absorbed by the particles in the air. This technique offers an advantage in that it can provide insights into both the size distribution and chemical makeup of the particles, in addition

to measuring their mass concentration. Nevertheless, this method is comparatively less precise when measuring mass concentration than the gravimetric method.

$PM_{2.5}$ data can be downloaded from its archive through the website: http://ukair.defra.gov.uk/data. The Devonshire Green AQ Monitoring Station in Sheffield (UKA00575) was selected as the inner urban monitoring site from which hourly $PM_{2.5}$ was downloaded. It was selected as it is considered the closest AQ monitoring station to the Weston Park Weather Station (Latitude 53.37, Longitude -1.48) and has a distance of less than 1 km (905 meters). This ensures that the weather and ambient $PM_{2.5}$ data used in the study are of close (approximate) geo-locations and urban context, see Figure 3.6.

Figure 3.6: DEFRA's Devonshire Green AQ Monitoring Station (UKA00575) location in approximation to the Weston Park Weather Station (grid cell of 100m x 100m)

The selection of hourly $PM_{2.5}$ data followed a similar approach to selecting the weather data for 2018-2021. It was noticed that there was much-missing data in 2018 due to maintenance of the AQ monitoring station. Years 2020 and 2021 were not considered in this study due to the disruption caused by the COVID-19 pandemic, in which the outdoor levels of $PM_{2.5}$ do not represent the typical $PM_{2.5}$ levels due to traffic and other anthropogenic activities. Moreover, most people were working from home due to the governmental restrictions on travel and building access. Hence, the 12 months of 2019 were selected for this study.

Figure 3.7 shows that during the winter months, the levels of $PM_{2.5}$ peaked in February and April with an average monthly level of 20.49 and 23.11 µg/m^3, respectively. On the other hand, the monthly outdoor $PM_{2.5}$ between May and October ranged from 10.61 to 6.41 µg/m^3. When comparing the seasonal outdoor $PM_{2.5}$, it can be seen in Figure 3.8 that in the heating season (November to April), the average seasonal outdoor $PM_{2.5}$ level is 17.04 µg/m^3, which is higher than the WHO annual average permissible $PM_{2.5}$ levels of 10 µg/m^3. Meanwhile, in the non-heating season (May to October), the average seasonal $PM_{2.5}$ was 6.78 µg/m^3. This suggests the importance of investigating population exposure to indoor $PM_{2.5}$ from outdoor sources exhibiting seasonal fluctuations.

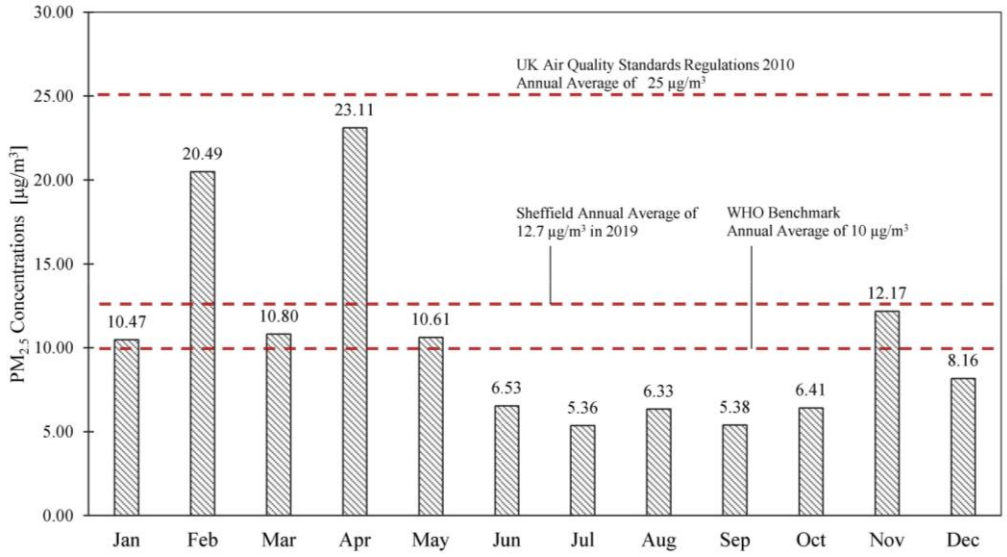


Figure 3.7: Monthly Average Outdoor $PM_{2.5}$ Concentrations in Sheffield in 2019 (DEFRA, 2019)

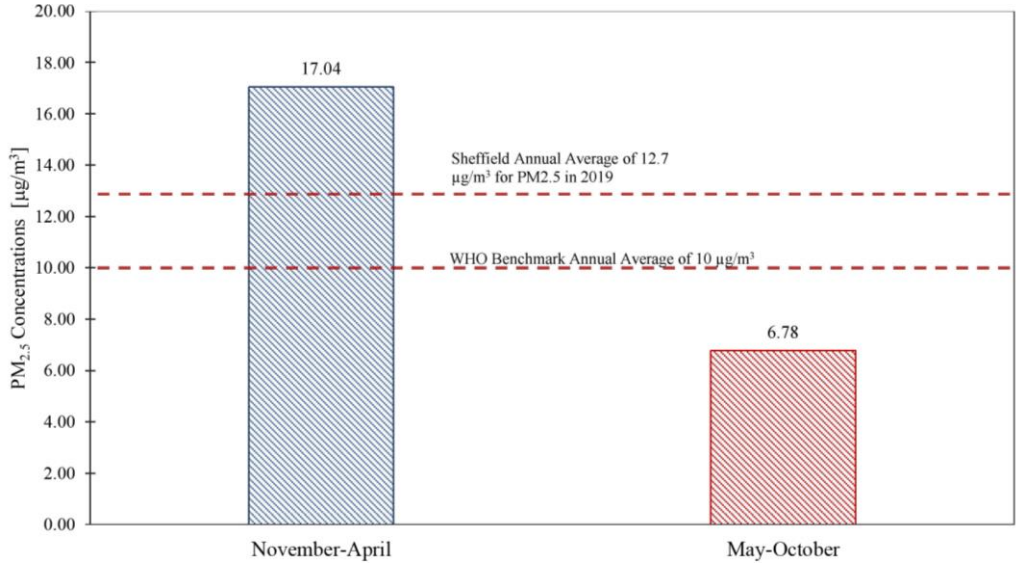


Figure 3.8: Comparison between the Heating season (Nov-Apr) and Cooling (Non-Heating) season (May-Oct) outdoor $PM_{2.5}$ concentrations (DEFRA, 2019)

3.3.6 The UoS Heating Policy

The UoS campus is part of the Veolia District Heating Network (DHN) (Arup, 2012), a unique district heating system whereby heat is mainly distributed in a building using traditional wet heat radiators. A low-carbon energy source is generally produced at a central location, converted into hot water, and distributed via underground pipes to a heat exchanger in buildings of all sizes and types. Figure 3.9 shows that out of the selected buildings in this study, the Barber House building is not part of the Veolia DHN, and a local gas-fired boiler supplies heat with wet heat radiators. However, the operation of its heating system was assumed to follow the UoS Heating Policy.

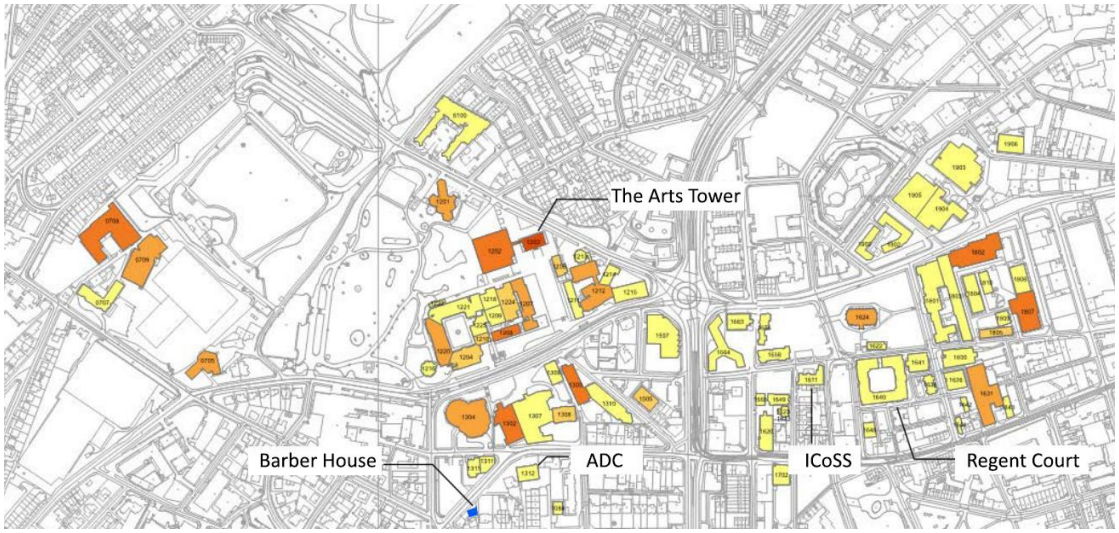


Figure 3.9: Buildings heated through the Veolia District Heating Network (Arup, 2012)

The heating season for the UoS is defined by the EFM and runs between 01 November to 30 April. In this period, the EFM aims to maintain the indoor air temperature between 19 °C and 21 °C during core working hours (9:00 - 17:00). At 9:00 during the heating season, the EFM's target is to have an indoor air temperature of 16 °C. The heating system is inoperable outside the heating season, outside the core working hours, and during weekends, and is kept to 12 °C. The heating is activated by a thermostat which responds to the outside temperature. If the temperature outside is 18 °C or above, the heating system will not operate. When outdoor temperatures plummet to extreme temperatures, an additional 'booster' is added between 12:00 and 14:00. Figures 3.10 and 3.11 show the UoS heating setpoints plotted against hourly outdoor air temperatures for January and April 2019. For January, the heating system operates at total capacity, in contrast to April, where the heating system is inoperative in the third week as the temperature reaches above 18 °C.

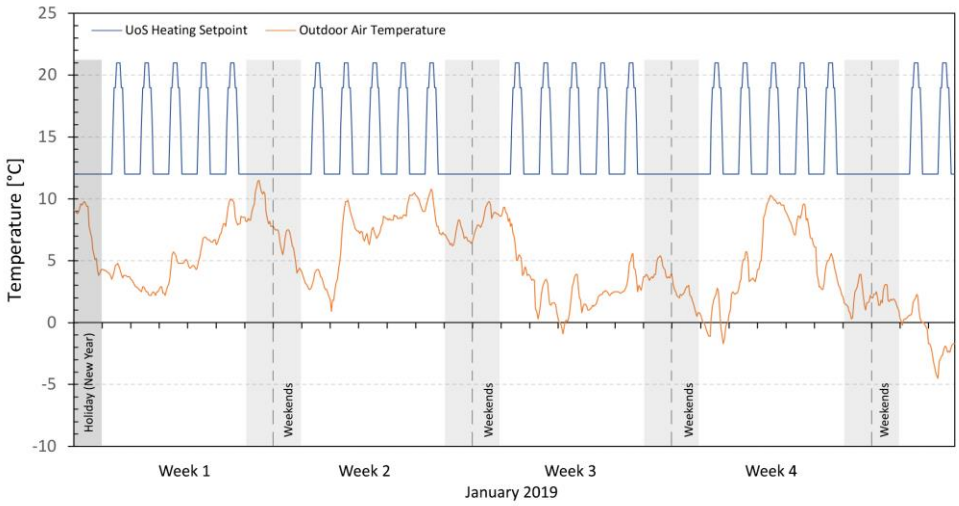


Figure 3.10: UoS Heating Setpoint plotted against January Outdoor Temperature

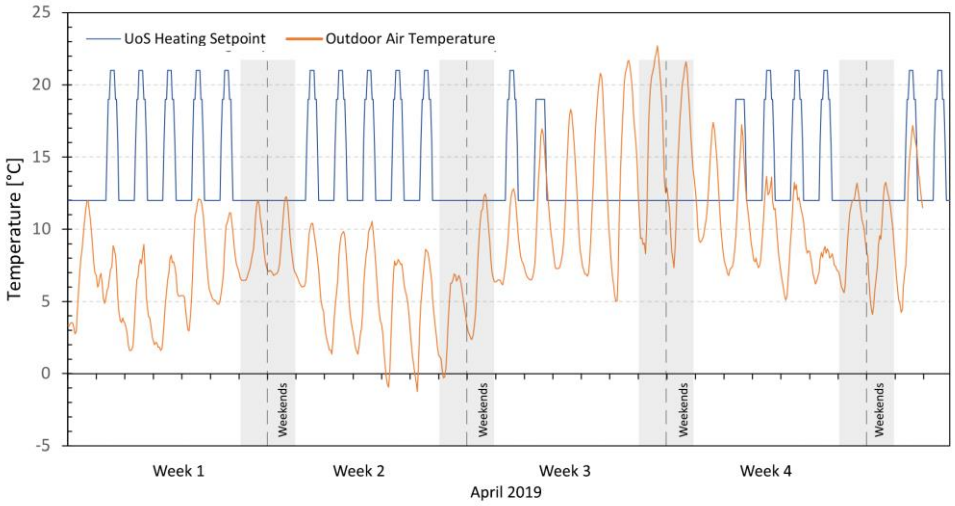


Figure 3.11: UoS heating setpoints plotted against April outdoor temperatures

3.3.7 $PM_{2.5}$ Properties

$PM_{2.5}$ concentrations can vary significantly between different types of buildings, depending on various factors, including PM sources, building design and construction, ventilation, and occupant activities. For example, common $PM_{2.5}$ sources in residential buildings include outdoor air pollution, cooking and smoking. $PM_{2.5}$ sources may vary in other buildings, such as office or education buildings. These buildings may have additional $PM_{2.5}$ sources from office equipment, such as printers and copiers. In addition, the design and construction of buildings can also impact $PM_{2.5}$ concentrations. For example, buildings with inadequate ventilation or with air filtration systems that are not properly maintained may have higher $PM_{2.5}$ concentrations. On the other hand, buildings with high-efficiency air filtration systems may have lower $PM_{2.5}$ concentrations.

For the modelling tool to simulate the transport of $PM_{2.5}$ between the rooms of a building and the local ambient environment, the physical and behavioural characteristics of $PM_{2.5}$ are required. This information is summarised in Table 3.5 and includes Molecular Weight (g/mol), Mean Diameter (μm), Penetration Factor P, and the Deposition Rate *k* (h^{-1}). The removal of particles from the ventilation system, their deposition on the building shell during air infiltration, and their deposition on indoor surfaces can significantly impact the indoor concentration of particles from outside (Riley et al., 2002).

The penetration factor (P) accounts for the filtering effect of $PM_{2.5}$ as it passes through a crack or an opening in the building envelope. It is a non-dimensional parameter with a value ranging between 0 and 1 (Ott et al., 2006). Its value depends on the size distribution of the aerosol and the airflow characteristics through the path (Hering et al., 2007), with P = 1 when the airflow path is through large openings like open windows (in natural ventilation) and P < 1 for other paths (infiltration). Previous studies found that when $PM_{2.5}$ passes through a crack in the building envelope, the value of P can vary between 0.7 and 0.9 with particle size (Ott et al., 2006). For the heating season in this study, $PM_{2.5}$ is considered a homogeneous pollutant and follows a similar uniform distribution between 0.7 and 0.9, and it is modelled with a P of 0.8 (O'Leary, Jones, et al., 2019). Alternatively, $PM_{2.5}$ is modelled during the summer with a P of 1 representing natural ventilation (J. Taylor et al., 2014b).

$PM_{2.5}$ deposition rates refer to the rate at which $PM_{2.5}$ settles or deposits onto indoor surfaces. $PM_{2.5}$ can be deposited through various mechanisms, including gravitational settling, diffusion, interception, and impaction. Gravitational settling occurs when particles settle onto a surface due to the force of gravity (Thatcher et al., 2002). Diffusion happens when particles move from an area of high concentration to a low concentration due to random molecular motion. Interception refers to the process where particles are intercepted by surfaces such as walls or ceilings, and impaction occurs when particles collide with a surface due to their inertia. Various factors, including particle size, air velocity, temperature, relative humidity, and surface characteristics, influence $PM_{2.5}$ deposition rates. For example, higher airflow velocities and warmer temperatures increase $PM_{2.5}$ deposition rates, while the impact of relative humidity on deposition rates is complex and can be influenced by many other factors such as coagulation, hygroscopic growth, airflow patterns, and electrostatic effects (Oezkaynak et al., 1996).

Assessing deposition rates is crucial in evaluating human exposure to indoor air pollution. Understanding the mechanisms and factors affecting $PM_{2.5}$ deposition rates can aid in designing and implementing effective indoor air quality interventions. Additionally, measuring $PM_{2.5}$ deposition rates can aid in estimating the effectiveness of various mitigation strategies for reducing indoor $PM2_{.5}$ concentrations. Several studies have been conducted to measure $PM_{2.5}$ deposition rates in residential buildings (Nazaroff, 2004b; Riley et al., 2002; Schneider et al., 2004); see Figure 3.12. However, such information might not be available in other types of buildings due to the complexity of measurements.

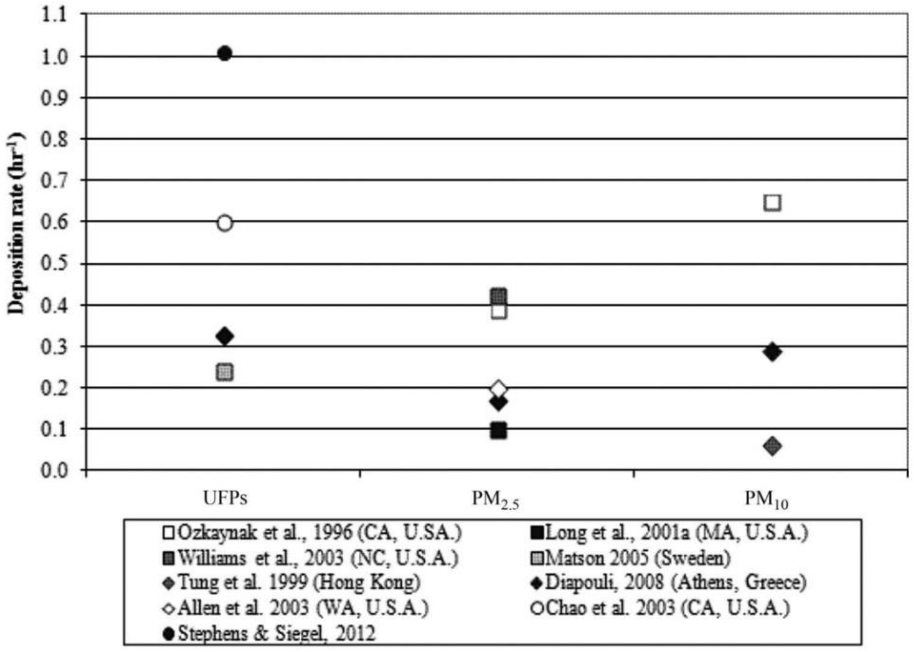


Figure 3.12: Summary of previous results on deposition rates of particles $\boldsymbol{k}$ (h^{-1}) (Diapouli et al., 2013)

Therefore, this is considered an uncertain input, and a probability distribution of $PM_{2.5}$ deposition rates N (0.39, 0.16), reported in the literature for residential settings, is considered (Oezkaynak et al., 1996). Other behavioural characteristics of $PM_{2.5}$ that can significantly impact the concentrations of $PM_{2.5}$ in indoor environments include the emission rates and removal rate by filter efficiency. However, based on the data received from the EFM, there was no information on the $PM_{2.5}$ filters installed across the UoS estates. Additionally, there were no tobacco smoking and cooking activities inside buildings and no use of gas heaters. As such, the indoor emission rate in this study was assumed to be zero.

Table 3.5: Physical and behavioural properties of $PM_{2.5}$ used in this study.

Pollutant	Molecular weight (g/mol)	Mean Diameter (μm)	Deposition Rate (h^{-1})	Penetration Factor P
Particulate Matter $PM_{2.5}$	1.4	2.50	0.39 (±0.16)	P = 0.8 (Infiltration), P = 1 (Natural Ventilation)

3.3.8 Occupancy Schedules

While assessing a building's energy performance and indoor air quality should ideally draw empirical data from extensive and sufficient field studies (surveys) or occupancy sensors, such data collection procedures are time-consuming and costly. Moreover, measured occupancy diversity factors for use in building energy and IAQ simulations are generally rare and practically non-existent for HEI buildings. Therefore, the uncertainty associated with building users' behaviours in the simulation study of indoor air quality remains relatively unexplored to date (Davis & Nutter, 2010). To the author's knowledge, the UoS does not hold occupancy data across all buildings. If some are available (e.g., Information Commons, The Diamond), it only records the total number of occupants at the building level, not per room/space. Conducting a fine-grained occupancy survey would be ideal for developing typical occupancy profiles for common space types for the sampled buildings. Unfortunately, an occupancy survey did not occur due to time constraints and the COVID-19 pandemic. Therefore, the uncertainty regarding occupancy schedules is accepted and acknowledged. As this thesis focused on modelling population exposure to $PM_{2.5}$ from outdoor sources during the heating season, user interaction with windows and natural ventilation was assumed negligible during the seasonal period. From an IAQ point of view, occupancy profiles can be highly influential when the heating systems are occupant-controlled or where significant indoor emissions of $PM_{2.5}$ exist (e.g., cooking or smoking activities).

University buildings exhibit various space types with specific academic purposes and operational characteristics. For the buildings selected in this study, a general characterisation of the space types identified in Section 3.3.1 may suggest some "theoretical" occupancy profiles to be used in simulation studies. Thus, hourly occupancy factors were specified for the sampled buildings (Table 3.6). These profiles formed the basis for calculating the internal heat gains in different space types of HEI buildings. Combining the theoretical occupancy profiles with the designed maximum occupancy for each space provided by the EFM, the maximum internal heat gain from occupants in each space/room within the sampled buildings can be modelled in EnergyPlus.

Table 3.6: Theoretical Occupancy Profiles for Each Space Type Used in the Co-simulation of the Selected Buildings

Space Type		Theoretical Occupancy Profile
General Offices	Administration and Academic Offices	9:00 – 17:00 (1-hour break between 12:00-13:00)
Educational Facilities	Lecture Theatres / Labs / Workshops / Studios	2-hour occupancy interval between 10:00 – 17:00 (including a 10-min break at each interval and a 1-hour break between 12:00-13:00)
	Seminar Rooms	1-hour occupancy interval between 10:00 – 17:00 (including a 10-min break at each interval and 1-hour break between 12:00-13:00)
	Study / Computer Rooms	10:00 – 17:00 (1-hour break between 12:00-13:00)
Shared Facilities	Common Areas	10:00 – 17:00
	Kitchen/Toilets	N/A
Circulation		N/A
Services		N/A

3.4 Summary

This chapter presents a hybrid bottom-up framework for developing HEI stock IAQ models to predict indoor $PM_{2.5}$ concentration and exposure. Five buildings of the University of Sheffield (UoS) were selected and introduced, briefing the data sources, scopes and granularities relevant to the study. The predicted indoor $PM_{2.5}$ will be used to estimate the population exposure to indoor $PM_{2.5}$ following a microenvironmental modelling approach (Chapter 7). The key steps and workflow constituting the research methodology are described. This chapter examined how the data on the selected buildings and the ambient environments could be used to develop an HEI stock IAQ model to inform future building planning and design. Furthermore, areas of paucity are highlighted, thus informing future surveys and research. Table 3.9 presents the data collected from the UoS EFM and existing governmental and non-governmental organisations that can be used as inputs to CONTAM and EnergyPlus, the application and use of this data in CONTAM and EnergyPlus will be described in detail in Chapter 4.

Table 3.7: A summary of the input parameters of CONTAM(.prj) file and EnergyPlus (.idf) file

Category	Parameter	Symbol [unit]	Variation Type	Data Source	Attribution Level
Building/Zone Variants	Zone Height	H [m]	Variability	CAD Drawings (Architectural Plans, Sections and Elevations from UoS EFM)	Development of CONTAM Project Files [.prj] for each Higher Education Building includes detailed zones geometries, adjacencies and juxtapositions.
	Zone Area	A [m^2]			
	Zone Volume	V [m^3]			
	Zone Orientation	φ [°]			
Building Construction Year	External Building Envelope Effective Leakage Area	ELA [cm^2/m^2]	Uncertainty	NIST[1] Library	Calculation of Building Envelope Air Permeability Values Q_{50} [$m^3/h/m^2$] after performing CONTAM Pressurisation Test at ΔP=50 Pa
	Building Envelope and Building Components Thermal Transmittance Value	U-Value [W/m^2K]	Variability	CAD Drawings (Wall Sections and Material Specifications from UoS EFM)	Development of Building Age Representative Material and Construction .idf files to Perform Dynamic Thermal Simulation in EnergyPlus
Window Parameters	Window Glazing Area	A_{wt} [m^2]	Variability	CAD Drawings (Architectural Plans, Sections and Elevations from UoS EFM)	Calculate the amount of Heat Gain and Heat Loss in EnergyPlus
	Window Opening Area	A_{wo}[m^2]	Design	EFM Heating Policy Plan	Window Opening Schedules to Account for Natural Ventilation in Cooling Season
	Window Leakage Area	A_{wl} [cm^2/m]	Uncertainty	NIST[1] Library	Development of Window Leakage Elements in CONTAM Project Files using a Power Law Model $Q= C(\Delta P)^n$
Building User Characteristics	Maximum Occupancy	Occ_m	Design	UK University Space Planning Guide for Space Standards and University Timetables for Different Space Types.	Space Use / Occupancy Schedules in CONTAM and EnergyPlus to Account for Indoor Heat Gains and CO_2 Generation Rates and the Calculation of $PM_{2.5}$ Exposure Levels.
	Occupancy Density	Occ_d [m^2/person]	Design		
	Occupancy Level	Occ	Uncertainty	Assumed Occupancy Data was Used	

THE TABLE CONTINUES THE FOLLOWING PAGE

Category	Parameter	Symbol [unit]	Variation Type	Data Source	Attribution Level
Pollutant Characteristics	$PM_{2.5}$ Deposition Rate	k [h^{-1}]	Uncertainty	Literature	Identify Sink Elements in CONTAM Project Files to represent the loss of $PM_{2.5}$ Indoors.
Airflow Characteristics	Flow Exponent	n	Uncertainty	CONTAM User Guide	Indicator of the Nature of Airflow (Turbulent or Laminar), Typical Values for Infiltration Airflow between 60 and 70
	Flow Coefficient	C	Uncertainty		Airflow Openings Dynamic Effects, Typical Values C=60 for Small Openings and slightly higher for Larger Openings
	Wind Pressure Coefficient	C_p	Uncertainty	Swami and Chandra Model	Calculate Wind Pressure Coefficients for Different Wind Angles in CONTAM to account for Wind Pressure Effect on Building Façade.
Ambient Weather Characteristics	Outdoor Temperature	T_{amb} [°C]	Uncertainty	Local Weather Stations (Sheffield's Weston Park Weather Station)	Generating EnergyPlus Weather Files (.epw) and CONTAM Weather Files (.WTH) using Actual Meteorological Year (AMY) Data
	Wind Speed	v [$m.s^{-1}$]	Uncertainty		
	Wind Direction	u [°]	Uncertainty		
	Ambient $PM_{2.5}$ Concentrations	C_{amb} [$\mu g/m^3$]	Uncertainty	Local Pollutant Monitoring Station (Sheffield Devonshire Green (UKA00575))	Generating CONTAM Ambient PM2.5 Concentration Levels Files (.CTM) using Hourly Data
Indoor Environment Characteristics	Heating Season Indoor Air Temperature	T_{in} [°C]	Variability	University of Sheffield Indoor Space Heating Policy	Development of EnergyPlus Indoor Space Heating Schedules and Setpoints to Control Indoor Air Temperature for the Co-Simulation

Chapter 4 Building Physics-Based Modelling

4.1 Introduction

In Chapter 3, the potential data sources to model the IAQ of an HEI building stock were demonstrated through the selected buildings from the UoS. The data received from the EFM guided the selection of individual rooms as the model resolution for performing IAQ modelling. The diversity of parameter most directly related to a building's design were used to reduce the primary uncertainty. Others with other sources of uncertainties (aleatory or epistemic) must be determined differently.

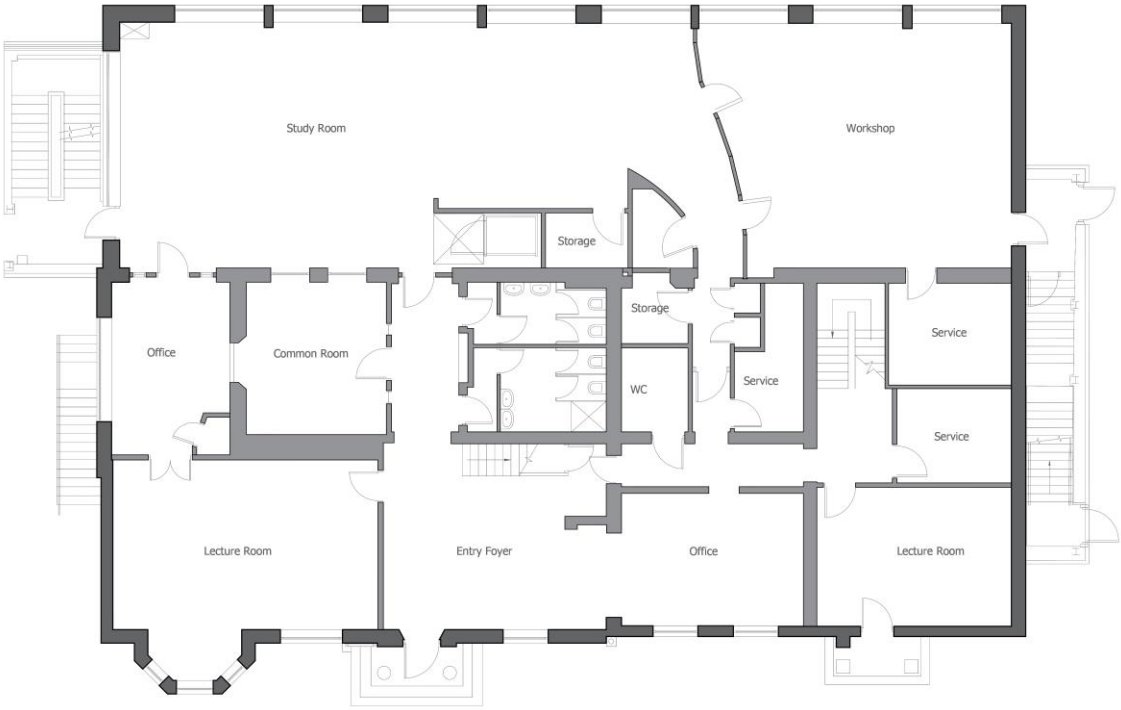
This chapter presents the IAQ modelling methods and software tools for estimating the concentrations of infiltrated $PM_{2.5}$ in the context of an HEI building stock. Here, the specific aspects of CONTAM and EnergyPlus, such as restrictions or boundaries and modelling assumptions, are discussed. Simulations were made under these conditions to analyse some aspects of IAQ across the selected buildings from the UoS building stock. This chapter illustrates how buildings are modelled in CONTAM and EnergyPlus. Then, a detailed analysis of the coupled simulation results is presented. The outputs include infiltrated $PM_{2.5}$ concentrations as a result of infiltration, infiltration ACH (ACH_{INF}), and indoor temperature. The analysis of the results focuses on two temporal scales: the **hourly** time-series outputs and the resampled **annual** outputs. Finally, the chapter concludes by demonstrating how the CONTAM-EnergyPlus co-simulation (CoSIM) results can be validated against field measurements.

4.2 Modelling the UoS Buildings

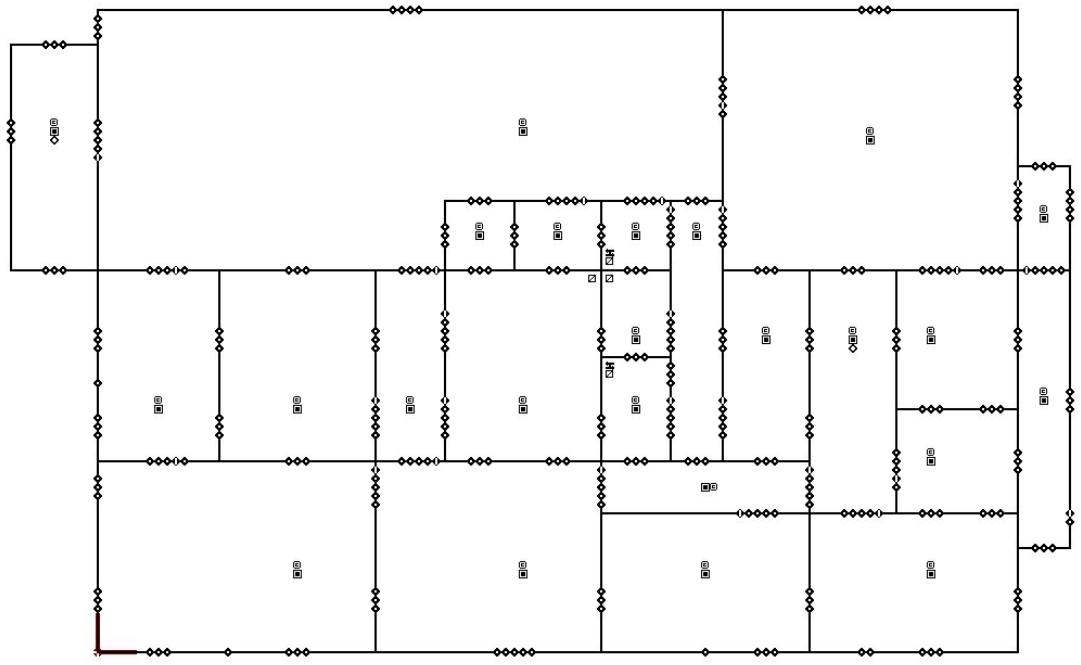
By utilising ContamX V3.4.0.2[5] (the CONTAM simulation engine), a multizone indoor air quality and ventilation model of each selected building was developed. For example, the Academic Development Centre (ADC) model is presented in Figure 4.1(a). The Barber House (BH), Arts Tower (AT), Regent Court (RC), and ICoSS are presented in Appendix A, Figures A.2-A.5. Each room/space (as labelled on the EFM building floor drawings) was considered a single volume to which doors connect all other rooms. After the rooms and doors were drawn in ContamW[6] (the CONTAM graphical user interface), well-mixed zones and airflow paths (represented as diamond-shaped dots) were mapped, as shown in Figure 4.1(b). A fundamental assumption in using ContamW is that the modelling must capture the (i) juxtaposition of zones to account for inter-zone flows, (ii) zone volumes to account for contaminant dilution, and (iii) wind pressure coefficients to account for the effect of wind on the building envelope. These aspects can be defined without exact geometrical representation via the graphical user interface (CONTAM SketchPad). However, as mentioned in Section 3.2, this thesis utilised a coupled modelling approach to account for heat transfer. Therefore, each CONTAM model was drawn using the *pseudo-geometry* option to define the scaling factor for drawing and viewing the wall, zone and duct dimensions on the SketchPad.

ContamX calculates the rate at which air pollutants are transported through airflow paths according to the information entered for every path. Thus, it uses information on the actual building layouts (Figure 4.1a), which the CONTAM-EnergyPlus coupled simulation requires. To determine the wind pressures at a building's location, CONTAM models require inputs of building orientation, or azimuth angle, for each external element. The orientation is assigned according to the actual building orientation. The CAD drawings show that the total floor areas were given and mapped in ContamW to replicate the building geometry. ContamW calculates the zone volumes and wall surface areas given the actual floor height inputs. After the layout is completed, the exterior walls of each Zone are used to determine the total envelope area.

[5]https://www.nist.gov/el/energy-and-environment-division-73200/nist-multizone-modeling/software/contam/download
[6] https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1887r1.pdf



(a)



(b)

Figure 4.1: The Ground Floor of the Academic Development Centre (ADC) – (a) Original CAD drawing and (b) CONTAM model

4.2.1 Airflow Paths

Each Zone has a variable *volume* and *floor area* resulting in variable *airflow paths*. Figure 4.2 shows another example of a CONTAM model built for the ground floor of the ADC building. Airflow paths (in red) and sinks (in green) are identified for each Zone, and exhaust systems (in blue) in kitchens and toilets. Air leakage paths are modelled using a single graphic element to represent potential airflows through walls and windows. Three airflow paths are used to model air leakage paths in each external and internal wall of a zone, which is assumed to be uniformly porous, by locating at its top, midpoint, and bottom following (Jones et al., 2013). An exhaust fan is linked to a central AHU unit for each floor to account for potential airflows in kitchens and toilets. This information can be acquired by accessing the EFM's mechanical CAD drawings for each building. The process of identifying all air leakage paths was similar for all buildings.

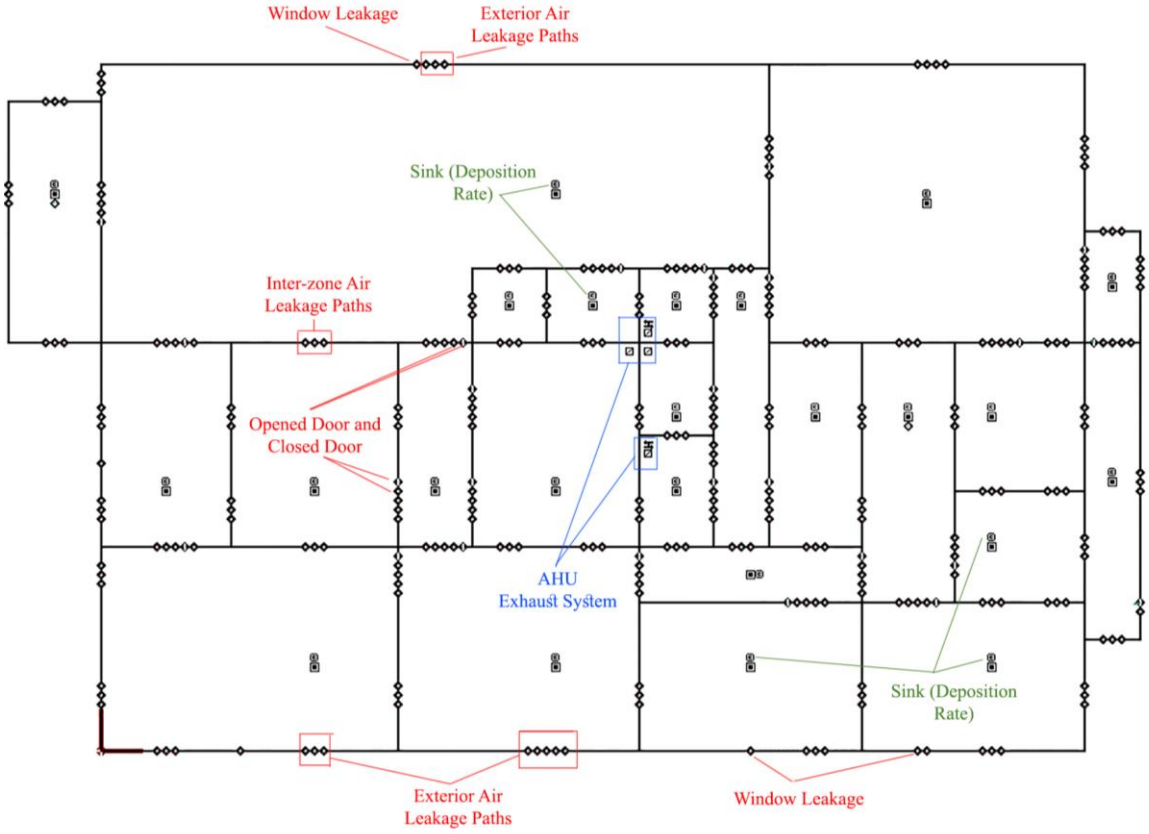


Figure 4.2: CONTAM Elements for the Ground Floor Layout of the ADC Building

The flow elements were adjusted to account for the corresponding surface areas of the elements they represent. CONTAM allows this by adding a *multiplier* to an element. For simplicity, airflows through floors and ceilings were not accounted for in any of the buildings modelled in

this study, and so their air leakage paths were ignored by giving them a multiplier with a value of zero. Wall area (multipliers) are equally distributed into the three paths. Air leakage paths were modelled using the power law model given in Eq. (4.1). To match the calculated airtightness level normalised by thermal envelope area Q_{50} ($m^3/h/m^2$) with the values provided in Section 3.3.2, the effective leakage area at 4Pa ELA_{4pa} (cm^2/m^2) was used as an input in ContamW and is represented by the following:

$$ELA_{4Pa} = \sqrt{\frac{p}{2(4Pa)}} \cdot Q_{50} \cdot \left\{\frac{4Pa}{50Pa}\right\}^{n} \tag{4.1}$$

where, ELA_{4Pa} is the effective leakage area at 4 Pa (cm^2/m^2), n is the dimensionless flow exponent, and ρ is the air density (kg/m^3). The value of the flow exponent n ranges from 0.5 for large and 1 for small openings. Previous studies in the US suggested that n can be sampled from a normal distribution N(0.651, 0.077) (Sherman & Dickerhoff, 1998) and between 0.6 – 0.7 for typical infiltration openings (W. Dols & Polidoro, 2020). With the absence of similar nationwide studies in the UK and for non-domestic buildings, n was assumed in this study to follow the same normal distribution and was given the value 0.65. This value was then entered into CONTAM with a coefficient equivalent to the corresponding section of the envelope area, as shown in Figure 4.3.

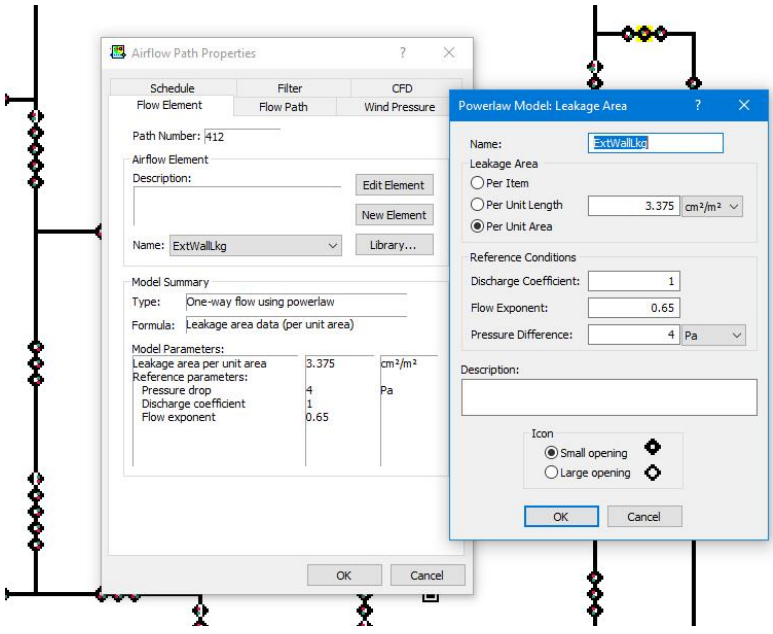


Figure 4.3: Using Eq (4.1) and ContamW interface to model Q_{50} using the values of ELA_{4Pa} (an example drawn from the ADC building)

Windows were all modelled with the assumption that they were permanently closed during the heating season as per the EFM heating policy. However, to account for the air leakage of windows, they were modelled in ContamW using Eq. (4.1) and were assigned a leakage value that represents the total leakage value for an item (cm^2). Information on each window was informed by the CAD drawings provided by the EFM. This includes the window height, width, and number of windows used to assign the ContamW multipliers. The discharge coefficient *Cd* was 0.6 as per the CONTAM User Guide (W. Dols & Polidoro, 2020).

Airflow and the transfer of pollutants and thermal energy can occur between different zones within a building or between inside and outside environments through other large openings like open doorways. These airflows tend to be more intricate, with the possibility of airflows in opposite directions in various parts of the opening. Two models, the *two–way flow one–opening* model and the *two–way flow two–opening model*, can be used to study such airflow in CONTAM (W. Stuart Dols & Polidoro, 2020). The former considers the flow through a single large opening and defines the neutral height, NPL, where the air velocity is zero. The latter model divides an opening vertically and uses two power-law models to estimate the net flow rate in each direction, accounting for the two-way flow due to the stack effect over the height of a tall opening (George N Walton, 1989). The NPL is the height at which the internal pressure equals the external pressure, resulting in no airflow in or out of an opening at that height. Above or below the NPL, the airflow and direction can be determined, with vents positioned below the NPL acting as inlets and those above acting as outlets, or vice versa, see Figure 4.4. The concept of the NPL is helpful in building design and is referenced in design standards such as the CIBSE AM10 guide (CIBSE, 2005).

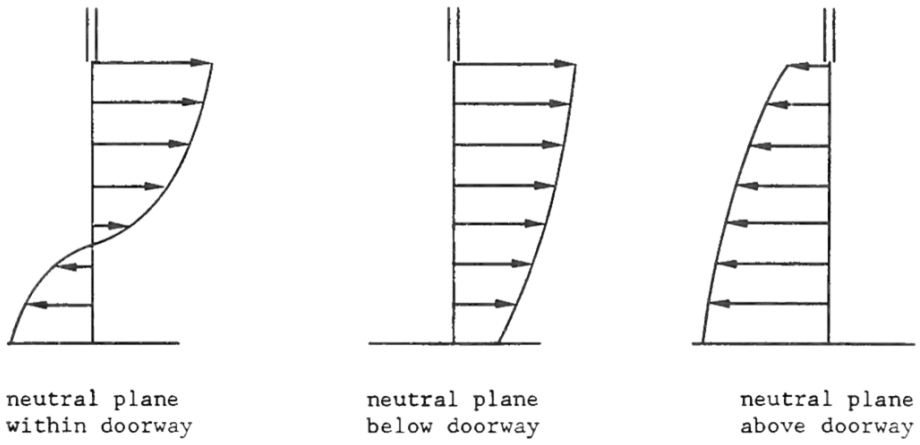


Figure 4.4: Neutral Plane Level (NPL) within doorways (George N Walton, 1989)

For simplification, open internal doors were modelled using the *two–way flow one–opening* model, with a discharge coefficient of 0.78, and its relative elevation is at the bottom of the door. When closed, doors were modelled as leakage elements that represent the door undercut given in (cm^2). Finally, simulations of a blower door test at 50 Pa were run in CONTAM for each building to ensure that the model's external air leakage rate (Q_{50}) was correct, see Table (4.1).

Table 4.1: Summary of the used Effective Leakage Areas ELA_{4Pa} for external and internal walls elements to achieve the airtightness level Q_{50} using CONTAM's blower test at 50Pa

Leakage Level	Airtightness Level Q_{50} ($m^3/h/m^2$)	External Wall and Internal Walls ELA_{4Pa} (cm^2/m^2)
Tight Building Envelope	3	0.925
	5	1.550
	7	2.155
	9	2.775
	10	3.075
	11	3.375
Leaky Building Envelope	13	3.997

4.2.2 Weather and Pollutants Data

To include the local weather conditions in the modelling, the weather data from the Weston Park Weather Station was converted into CONTAM's weather data format (*.wth) (W. Dols & Polidoro, 2020). Data is reported hourly, giving the date and time, ground temperature, atmospheric pressure, wind velocity, wind direction, and absolute humidity. The same weather file was used in the simulations for all selected buildings. To account for the wind effects on each side of the buildings, the wind effects were estimated using a wind pressure profile calculated using wind pressure coefficient (C_P) relationships found in (Swami & Chandra, 1987). Wind pressure profiles for each building are a function of the block aspect ratio (S) and the terrain constants. A variable wind speed modifier corresponding to "urban" terrain and scaled to building height (W. Dols & Polidoro, 2020), was applied to all exterior leakage paths. This parameter was used in CONTAM to account for the effects of local terrain on wind speed variation with height above ground level, see Table 4.2.

Table 4.2: Building Heights, Local Terrain Constant, Velocity Profile Exponent, and Corresponding Wind Speed Modifier Input Data in CONTAM

Building Name	Building Height (m)	Local Terrain Constant	Velocity Profile Exponent	Wind Speed Modifier
Barber House (BH.)	6.0	0.717	0.22	0.410
Academic Development Centre (ADC)	7.5			0.453
The Regent Court (RC.)	13.4			0.581
Arts Tower (AT)	72.0			1.165
ICoSS	21.8			0.718

C_P at different angles were assumed to be between 0◦ to 360◦ and specified for each side of the building. As the *building form* can significantly differ between the selected buildings, it can impact the resultant wind effects. Due to the lack of other resources that may represent such heterogeneous forms, the uncertainty in the resultant wind pressure profiles using the Swami and Chandra model is acknowledged (Figure 4.4), and the implications on the model results are discussed in detail in Chapter 8.

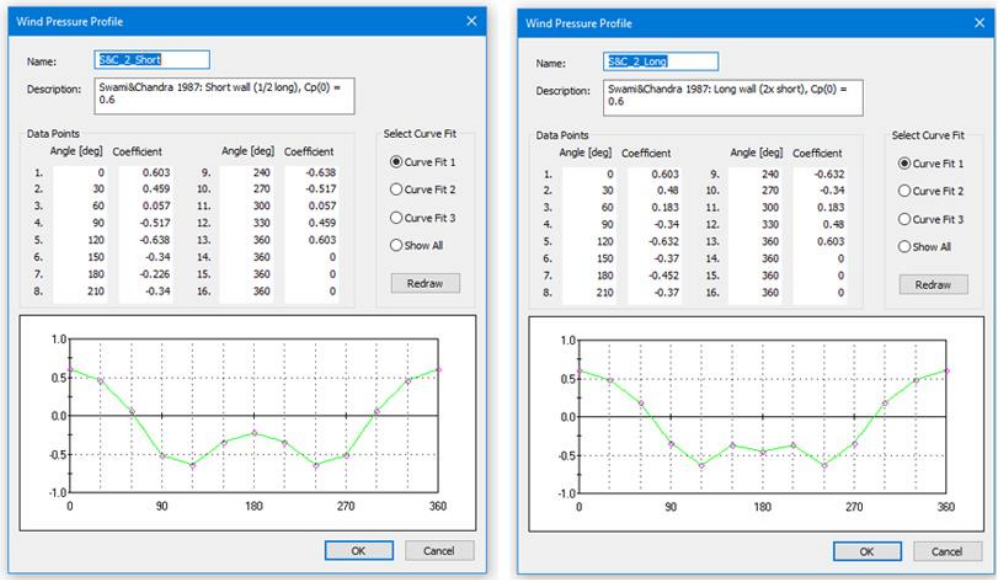


Figure 4.5: Example of a wind pressure profile for the (a) short wall and (b) long wall for all buildings to be used in the CONTAM simulations.

Finally, to include the local ambient $PM_{2.5}$ levels in the models, the DEFRA Devonshire Green Monitoring Station data was converted into the CONTAM's format (*.ctm) (W. Dols & Polidoro, 2020). The ambient AQ data is reported hourly, giving the date, time and ambient $PM_{2.5}$ levels. It was assumed that the ambient $PM_{2.5}$ levels were the same for all buildings, and therefore the same .ctm file was used in all simulations.

4.2.3 Deposition Rates

Deposition rates are essential in determining the removal rates and indoor concentrations of pollutants, especially when ventilation is limited. Here, the deposition of $PM_{2.5}$ was modelled as a deposition rate sink model with a constant value of k =0.39 h^{-1} (see Table 3.5). Nevertheless, this value was considered uncertain (Nazaroff et al., 1993) due to several factors such as room dimension, furniture area, and air velocity. Although all these parameters change by building/zone, this study took a simplified approach to the deposition process and input values.

4.2.4 Indoor Temperatures

As CONTAM is not a thermal model, the internal air temperatures must be specified as constant values for each Zone. However, to account for the dynamic interaction between thermal flow and airflow within a building, CONTAM can be coupled with EnergyPlus following the framework given in (W. Stuart Dols et al., 2016), which was described in detail in Section 2.4.3. In generating the CONTAM project file (*.prj) for each building, a constant indoor temperature of 21 °C was used. This temperature represents the heating season setpoint specified in the EFM heating policy. This allows ContamX to calculate initial infiltration rate values that can be used as dynamic infiltration flows rather than constant values. Then, CONTAM3DExporter[7] was used to export EnergyPlus (IDF) files from COMTAM (PRJ) files. The IDF file contains all data exchange parameters representing the geometry of each building.

The exported IDF files were manually edited for each building to include the thermal-related input parameters. This includes the thermal properties (U-Values) of the building construction, adding windows to account for the heat gain from solar radiation, identifying sources of internal heat gains and occupancy schedules, and the design of the Veolia DHN informed by the EFM, see two such examples in Figure 4.6.

[7]https://www.nist.gov/el/energy-and-environment-division-73200/nist-multizone-modeling/software/contam-3d-exporter

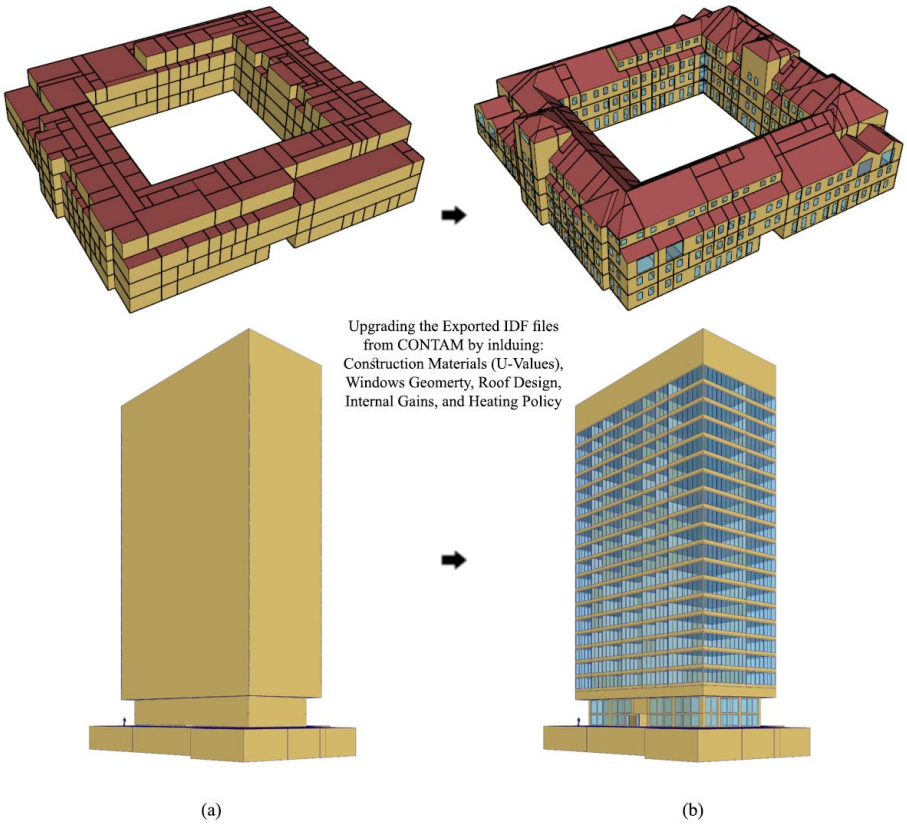


Figure 4.6: The IDF models developed for the Regent Court Building (top) and the Arts Tower (bottom); before editing (a) and after editing (b), using the OpenStudio tool in Sketchup

For IAQ modelling, the main benefit of coupling CONTAM with EnergyPlus is that EnergyPlus enables timestep-level temperature differences for each Zone, responsive to weather, HVAC, air flows, and occupant behaviours. The U-values of the construction materials for each building were obtained from the EFM and followed the values specified in Table 3.1. Information on the sources of internal gains in HEI buildings was informed by the CIBSE Guide A (CIBSE, 2018) and the previous studies conducted on multiple UoS buildings (S Douglas, 2014). The information includes artificial lighting, equipment, and occupants. Based on these studies, the primary source of internal heat gains from equipment in the UoS buildings was the number of computers in use. The number of computers reported in these studies was 0.375 PC/m^2 in computer rooms, 0.1 PC/m^2 in study rooms and shared offices, and 1 PC per cellular office. The benchmark values from CIBSE Guide A (CIBSE, 2018) were used to assign the sensible heat gain from equipment with 20 watts/m^2 in cellular offices, 25 watts/m^2 in shared offices, and 25 watts/m^2 in computer rooms. Sensible heat gain from artificial lighting was given the values of 10 watts/m^2 for rooms where occupants spent some considerable time and 7 watts/m^2 for circulation and service zones (Altan et al., 2009).

The heat released by people is often tabulated in design guides in terms of sensible and latent heat loads. Table 4.3 summarises benchmark allowances of the sensible and latent heat loads for different space types, assuming an indoor temperature of 21 °C and a sedentary occupancy activity level (CIBSE, 2018). Occupancy density for each space type (m^2/person) (Ross, 2019) was used to calculate the total number of occupants and the total occupants' heat gain for each Zone. As the occupancy schedules were assumed to be similar among most UoS space types, the theoretical operational schedules summarised in Table 3.6 were mapped into the IDF file of each selected building.

Table 4.3: Summary of the benchmark allowances for internal heat gain from occupants, artificial lighting, and equipment in different space types (CIBSE, 2018).

Building Type	**Use**	**Floor Area** (m^2/person)	**Sensible Heat Gain** W/m^2			**Latent Heat Gain** W/m^2	
			People	Lighting	Equipment	People	Other
Offices	Cellular Office	9	10	8-12	25	7.5	-
	Shared Office	4.5	20	8-12	20	15	-
	Meeting Rooms	3	27	10-20	5	20	-
Education	Lecture Theatres	1.5	67	12	2	50	-
	Computer Spaces	2.5	53	12	60	40	-
	Seminar Rooms	3	27	12	5	20	-

4.2.5 Heating Policy using HVAC Templates in EnergyPlus

As discussed in Section 3.3.5, the UoS buildings are mainly heated through the Veolia DHN, and heat is distributed to each Zone using traditional wet heat radiators. In order to map this heating system to the IDF files, a group of objects in EnergyPlus was used for the specification of simple zone thermostats and HVAC systems. As the total energy consumption of the sampled buildings was not part of the scope, the sizing and the detailed design of the Veolia DHN were simplified to match the heating policy regardless of the amount of water and energy used. Therefore, the `HVACTemplate` object type was used. Figure 4.7 illustrates the schematic representation of the Veolia DHN in EnergyPlus. For water baseboard heating systems powered by local district heating, the following objects were specified in each IDF file:

(1) `HVACTemplate:Thermostat`,

(2) `HVACTemplate:Zone:BaseboardHeat`,

(3) `HVACTemplate:Plant:HotWaterLoop`,

(4) `HVACTemplate:Plant:Boiler`.

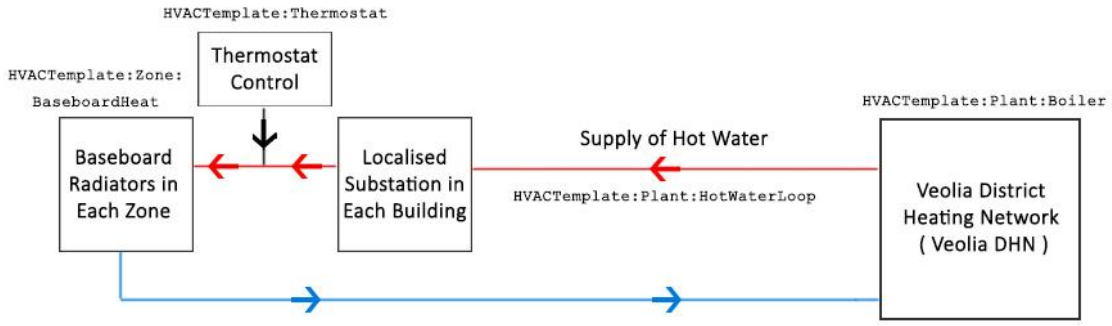


Figure 4.7: Schematic representation of the Veolia DHN mapped in EnergyPlus for each of the sampled buildings

4.3 Processing the Co-Simulation Outputs

This study used deterministic values described by discrete or continuous probability distributions for the input parameters. Based on these input parameters, CONTAM and EnergyPlus can be used to estimate each Zone's annual average infiltrated $PM_{2.5}$ concentrations. Table 3.7 in Section 3.4 summarised the input parameters for co-simulation and their sources and attribution levels. Quantifying the uncertainty in the outputs by systematically varying each set of CONTAM and EnergyPlus inputs and running multiple simulations was not within the scope of the study.

After the co-simulation files were developed for the selected buildings, simulations were run for a whole year. Generally, the shorter the time step, the more accurate the solution is, but at the expense of computational resources and runtime. Due to the sensitivity of the time step selected in the analysis of indoor $PM_{2.5}$ and its behaviour in indoor environments compared to other building-related performance analyses (e.g. annual energy use) (Tabares-velasco, 2013), a time step of 4 per hour (i.e., 15-min interval) was set in both CONTAM and EnergyPlus files. As a result, the airflow calculations, pollutant behaviours, and building envelope thermal responses can be modelled more accurately. The simulation ran using a 15-min interval and performed reasonably well regarding computing time and resources. It also improved the numerical solution of the zone mass balance and heat balance models in CONTAM and EnergyPlus. In CONTAM, concentrations are reported at a moment in time and only according to the "Output" time step, regardless of the "Calculation" time step identified in the simulation settings.

This process was repeated using the six airtightness values specified in Section 3.3.2 and assuming that purpose–provided openings (windows) are closed during the heating season

(November – April) and open during the non-heating season (May – October). In addition, any mechanical ventilation systems were switched off (except extract fans in toilets and kitchens) during the simulation (see Section 3.3.3). With this in mind, the total airflow rate ACH_T (h^{-1}) was assumed to equal the simulated infiltration rate ACH_{INF} (h^{-1}). Each set of outputs calculated the heating season zone-weighted pollutant $PM_{2.5}$ concentrations and added them to the average non-heating season concentrations (see Section 1.4). First, each Zone's average concentration of infiltrated $PM_{2.5}$ over the heating season was calculated over the simulation period. Next, using the building operation period (7 AM-7 PM), the average concentrations in each room were weighted using the 12-hour occupancy time. Then, the time–series data of the infiltration rates ACH_{INF} (h^{-1}) for each Zone, the infiltrated $PM_{2.5}$ concentrations ($\mu g/m^3$) for each of the zones, the outdoor scaled wind speed (m/s), and the indoor temperature T_{in} (°C) were obtained.

Finally, the ACH_{INF} and $PM_{2.5}$ data were extracted from the CONTAM output files, and ΔT was computed using the EnergyPlus time-series indoor temperature (T_{in}) value and the weather data. Table 4.4 shows an example of the metrics recorded and compiled into a single file labelled with the building ID. The results of the co-simulation are presented in four subsections below. Section 4.3.1 describes the predicted time series data from the simulations; while Section 4.3.2 presents the the baseline concentrations of infiltrated $PM_{2.5}$ heating season and the whole year. Section 4.3.3 presents the results of improving the Q_{50} on the concentrations of indoor $PM_{2.5}$. Finally, Section 4.3.4 presents the results of I/O ratio of $PM_{2.5}$ and highlights the importance of other factors such as building height and zone level as a modifier to exposure to indoor $PM_{2.5}$.

4.3.1 Time series data

With a 15-min temporal resolution, the simulated $PM_{2.5}$ concentrations and Air Change Rates (ACH_{INF}) from CONTAM and the Indoor Temperature (T_{in}) from EnergyPlus were resampled to generate hourly averages for data analysis over the heating season. This was achieved using the Pandas Module in Python and the function `resample()`. The data was obtained from each CONTAM simulation file (`.sim`) and read using `simread3.exe`. Additionally, the IDF file outputs the T_{in} time series in a Comma Separated Values (.CSV) file. At the hourly resolution between 01 November and 30 April, the total number of data points for each Zone was 4,344, totalling 1,941,768 data points for the 445 zones simulated.

Table 4.4: Input parameters and output metrics recorded and compiled into a single file labelled with the building and zone ID. (Total Number of Zones N=2729)

	Decriptive_Parameters			Geometrical_Parameters														Simulated_Environmental_Parameters					Computed_Environmental_Parameters				
N	CONTAM_ZONE_ID	Space_Type_Main	Space_Type_Sub	Area (A) (m^2)	Volume (V) (m^3)	Height (H) (m)	Orientation (Θ) (°)	Number of Exposed Facades (N_{ef})	Area of Exposed Facades (A_{ef}) (m^2)	Leakage Area (L_{ef}) (m^2)	Area of Adjacent Walls (A_{aw}) (m^2)	Leakage Area (L_{aw}) (m^2)	Building Airtightness Q_{50} ($m^3/h/m^2$)	Occupancy Density	Maximum Occupancy Occ_{max}	Aef:V	Aef:A	Indoor Temperature T_{in} (°C)	Local Wind Speed V_H (m/s)	Relative Humidity RH (%)	ACH$_{INF}$ (h^{-1})	Indoor $PM_{2.5}$ C_i (μg/m^3)	Temperature Difference ΔT (°C)	Air Permeability Q_A ($m^3/h/m^2$)	Ventilation Rates (litres/s)	Ventilation Rates (litres/s/pp)	Indoor/Outdoor Ratio I/O
1	301_BS_O_001	Offices	Shared Cellular Offices	24	90	-1.9	5	1	23	7.6E-03	56	2.8E-02	3	4.5	5	0.3	0.9	18.3	0.0	49	0.3	5.7	11.7	1.3	7.9	1.5	0.3
2	301_BS_O_002	Offices	Shared Cellular Offices	12	45	-1.9	5	1	11	3.8E-03	41	2.1E-02	3	4.5	3	0.3	0.9	18.7	0.0	51	0.3	5.1	12.0	1.3	3.9	1.5	0.3
3	301_BS_O_003	Offices	Shared Cellular Offices	10	38	-1.9	5	1	9	3.2E-03	39	2.0E-02	3	4.5	2	0.2	0.9	18.6	0.0	49	0.3	4.8	12.0	1.1	3.0	1.3	0.3
4	301_BS_O_004	Offices	Shared Cellular Offices	17	62	-1.9	5	1	11	3.8E-03	53	2.6E-02	3	4.5	4	0.2	0.7	18.9	0.0	51	0.2	4.2	12.2	1.1	3.4	0.9	0.2
5	301_BS_O_005	Offices	Shared Cellular Offices	25	93	-1.9	320	2	21	7.0E-03	54	2.7E-02	3	4.5	6	0.2	0.8	18.3	0.0	51	0.3	4.6	11.6	1.4	7.8	1.4	0.3
6	301_BS_O_006	Offices	Shared Cellular Offices	25	93	-1.9	230	2	38	1.3E-02	38	1.9E-02	3	4.5	6	0.4	1.5	20.2	0.0	46	0.6	4.9	13.6	1.4	14.7	2.7	0.3
7	301_F0_EF_LR01	Educational Facilities	Lecture Rooms	24	90	1.9	140	2	34	1.1E-02	38	1.9E-02	3	1.5	16	0.4	1.4	24.2	3.3	34	1.2	8.7	17.5	3.1	28.9	1.8	0.5
8	301_F0_EF_LR02	Educational Facilities	Lecture Rooms	44	165	1.9	140	2	51	1.7E-02	51	2.5E-02	3	1.5	29	0.3	1.2	24.9	3.3	37	0.6	5.9	18.3	2.1	29.8	1.0	0.3
9	301_F0_EF_SR01	Educational Facilities	Study Rooms	119	446	1.9	320	2	71	2.4E-02	120	6.0E-02	3	2.5	48	0.2	0.6	18.1	3.3	52	0.2	5.4	11.5	1.4	28.1	0.6	0.3
10	301_F0_EF_W01	Educational Facilities	Workshops	64	239	1.9	50	2	49	1.6E-02	71	3.6E-02	3	2.5	26	0.2	0.8	17.8	3.3	50	0.2	6.0	11.2	1.1	15.4	0.6	0.3
11	301_F0_O_001	Offices	Shared Cellular Offices	24	90	1.9	185	1	23	7.6E-03	53	2.6E-02	3	4.5	5	0.3	0.9	28.1	3.3	32	0.4	5.6	21.4	1.4	9.1	1.7	0.3
12	301_F0_O_002	Offices	Shared Cellular Offices	19	72	1.9	275	1	21	7.0E-03	47	2.3E-02	3	4.5	4	0.3	1.1	21.4	3.3	40	0.8	6.0	14.7	2.9	16.4	3.8	0.4
13	301_F1_EF_SR01	Educational Facilities	Study Rooms	99	371	5.6	140	2	53	1.8E-02	120	6.0E-02	3	2.5	40	0.1	0.5	23.0	4.1	47	0.1	4.4	16.3	0.8	11.5	0.3	0.3
14	301_F1_EF_SR02	Educational Facilities	Study Rooms	41	152	5.6	50	2	47	1.6E-02	54	2.7E-02	3	2.5	16	0.3	1.2	17.2	4.1	59	0.2	4.1	10.6	0.8	9.9	0.6	0.2
15	301_F1_EF_SR03	Educational Facilities	Study Rooms	34	127	5.6	5	1	28	9.5E-03	62	3.1E-02	3	2.5	14	0.2	0.8	17.5	4.1	57	0.2	3.7	10.8	0.7	5.4	0.4	0.2
16	301_F1_O_001	Offices	Shared Cellular Offices	19	72	5.6	275	1	21	7.0E-03	47	2.3E-02	3	4.5	4	0.3	1.1	21.2	4.1	45	0.6	3.6	14.6	2.1	12.0	2.8	0.2
17	301_F1_O_002	Offices	Shared Cellular Offices	9	33	5.6	275	1	9	3.2E-03	36	1.8E-02	3	4.5	2	0.3	1.1	22.3	4.1	40	0.6	3.9	15.6	2.2	5.7	2.9	0.2
18	301_F1_O_003	Offices	Shared Cellular Offices	14	51	5.6	230	2	28	9.5E-03	28	1.4E-02	3	4.5	3	0.6	2.1	25.7	4.1	33	0.8	5.2	19.0	1.4	11.2	3.7	0.3
19	301_F1_O_004	Offices	Shared Cellular Offices	11	39	5.6	185	1	13	4.4E-03	26	1.3E-02	3	4.5	2	0.3	1.2	28.5	4.1	31	0.3	4.1	21.8	0.9	3.2	1.4	0.2
20	301_F1_O_005	Offices	Shared Cellular Offices	14	53	5.6	185	1	13	4.4E-03	43	2.2E-02	3	4.5	3	0.2	0.9	28.6	4.1	31	0.2	5.2	21.9	0.9	3.2	1.0	0.3
21	301_F1_O_006	Offices	Shared Cellular Offices	12	45	5.6	185	1	11	3.8E-03	41	2.1E-02	3	4.5	3	0.3	0.9	30.0	4.1	33	0.2	3.8	23.3	0.9	2.7	1.0	0.2
22	301_F1_O_007	Offices	Shared Cellular Offices	10	38	5.6	185	1	9	3.2E-03	39	2.0E-02	3	4.5	2	0.2	0.9	30.8	4.1	32	0.2	3.8	24.1	0.9	2.3	1.1	0.2
23	301_F1_O_008	Offices	Shared Cellular Offices	14	53	5.6	185	1	13	4.4E-03	43	2.2E-02	3	4.5	3	0.2	0.9	28.3	4.1	33	0.2	4.9	21.6	0.9	3.3	1.1	0.3
24	301_F1_O_009	Offices	Shared Cellular Offices	11	39	5.6	5	1	13	4.4E-03	36	1.8E-02	3	4.5	2	0.3	1.2	17.0	4.1	56	0.2	3.3	10.3	0.6	2.3	1.0	0.2
25	301_F1_O_010	Offices	Shared Cellular Offices	11	39	5.6	5	1	13	4.4E-03	36	1.8E-02	3	4.5	2	0.3	1.3	16.9	4.1	55	0.2	3.7	10.2	0.7	2.6	1.1	0.2
26	301_F1_O_011	Offices	Shared Cellular Offices	14	51	5.6	5	1	11	3.8E-03	45	2.3E-02	3	4.5	3	0.2	0.8	17.3	4.1	56	0.2	3.6	10.6	0.8	2.6	0.9	0.2
27	BH_F0_O_M01	Meeting Room	Meeting Rooms	49	148	1.5	195	2	48	1.9E-02	43	1.7E-02	3	2.5	20	0.3	1.0	22.2	3.1	54	0.4	4.9	15.5	1.3	17.7	0.9	0.3
28	BH_F0_O_001	Offices	Shared Cellular Offices	28	84	1.5	285	2	45	1.8E-02	21	1.8E-02	3	4.5	6	0.5	1.6	18.8	3.1	43	0.9	6.6	12.2	1.7	21.5	3.5	0.4
29	BH_F0_O_002	Offices	Shared Cellular Offices	21	62	1.5	150	1	12	4.8E-03	43	1.0E-02	3	4.5	5	0.2	0.6	22.8	3.1	41	0.2	4.1	16.1	1.0	3.5	0.8	0.2
30	BH_F0_O_003	Offices	Shared Cellular Offices	23	69	1.5	150	1	13	5.3E-03	44	1.8E-02	3	4.5	5	0.2	0.6	23.0	3.1	42	0.2	4.1	16.3	1.0	3.8	0.8	0.2
31	BH_F0_O_004	Offices	Shared Cellular Offices	27	82	1.5	105	2	38	1.5E-02	26	1.0E-02	3	4.5	6	0.5	1.4	20.6	3.1	46	0.5	5.5	13.9	1.0	10.5	1.7	0.3
32	BH_F0_O_005	Offices	Shared Cellular Offices	13	40	1.5	60	1	13	5.0E-03	32	1.3E-02	3	4.5	3	0.3	0.9	17.7	3.1	51	0.2	4.2	11.1	0.6	2.0	0.7	0.2
33	BH_F0_O_006	Offices	Shared Cellular Offices	41	123	1.5	285	2	46	1.8E-02	32	1.3E-02	3	4.5	9	0.4	1.1	17.4	3.1	50	0.4	5.1	10.7	1.1	14.0	1.5	0.3
34	BH_F1_O_001	Offices	Shared Cellular Offices	14	43	4.5	285	2	23	9.1E-03	33	1.3E-02	3	4.5	3	0.5	1.6	17.8	3.9	49	0.8	3.8	11.1	1.5	9.6	3.0	0.2
35	BH_F1_O_002	Offices	Shared Cellular Offices	14	43	4.5	330	1	12	4.8E-03	34	1.3E-02	3	4.5	3	0.3	0.8	17.8	3.9	51	0.4	2.5	11.2	1.5	5.0	1.6	0.1
2695	RC2_F2_O_007	Offices	Open Offices	63	171	8.0	230	2	26	5.8E-03	75	1.7E-02	13	6.0	11	0.2	0.4	15.4	4.5	50	1.8	7.3	8.7	11.5	84.3	8.0	0.4
2696	RC2_F2_O_008	Offices	Shared Cellular Offices	13	35	8.0	95	1	6	1.3E-03	37	8.2E-03	13	4.5	3	0.2	0.5	16.2	4.5	48	0.4	6.7	9.5	2.5	4.2	1.5	0.4
2697	RC2_F2_O_009	Offices	Shared Cellular Offices	13	35	8.0	95	1	6	1.3E-03	37	8.2E-03	13	4.5	3	0.2	0.5	16.3	4.5	49	0.4	6.6	9.6	2.5	4.2	1.5	0.4
2698	RC2_F2_O_010	Offices	Shared Cellular Offices	14	39	8.0	95	1	7	1.5E-03	38	8.4E-03	13	4.5	3	0.2	0.5	16.2	4.5	50	0.4	6.5	9.5	2.5	4.7	1.5	0.4
2699	RC2_F2_O_011	Offices	Shared Cellular Offices	14	39	8.0	95	1	7	1.5E-03	38	8.4E-03	13	4.5	3	0.2	0.5	16.0	4.5	50	0.4	6.6	9.3	2.5	4.7	1.5	0.4
2700	RC2_F2_O_012	Offices	Shared Cellular Offices	14	39	8.0	95	1	7	1.5E-03	38	8.4E-03	13	4.5	3	0.2	0.5	15.8	4.5	50	0.4	6.8	9.1	2.5	4.8	1.5	0.4
2701	RC2_F2_O_013	Offices	Shared Cellular Offices	16	43	8.0	50	2	11	2.4E-03	35	7.8E-03	13	4.5	4	0.3	0.7	15.2	4.5	51	0.9	7.4	8.6	3.6	10.8	3.1	0.4
2702	RC2_F2_O_014	Offices	Shared Cellular Offices	35	94	8.0	50	2	32	7.0E-03	32	7.0E-03	13	4.5	8	0.3	0.9	14.9	4.5	52	1.4	7.6	8.2	4.0	35.6	4.6	0.4
2703	RC2_F2_O_015	Offices	Shared Cellular Offices	29	77	8.0	5	1	13	2.8E-03	45	1.0E-02	13	4.5	6	0.2	0.5	15.3	4.5	52	0.9	6.9	8.6	5.2	18.5	2.9	0.4
2704	RC2_F2_O_016	Offices	Shared Cellular Offices	23	61	8.0	5	1	10	2.2E-03	43	9.4E-03	13	4.5	5	0.2	0.5	15.4	4.5	51	0.8	6.7	8.8	4.9	13.7	2.7	0.4
2705	RC2_F2_O_017	Offices	Shared Cellular Offices	14	37	8.0	5	1	6	1.3E-03	39	8.5E-03	13	4.5	3	0.2	0.5	15.5	4.5	51	0.8	6.9	8.8	4.9	8.3	2.8	0.4
2706	RC2_F2_O_018	Offices	Shared Cellular Offices	15	41	8.0	5	1	7	1.5E-03	39	8.6E-03	13	4.5	3	0.2	0.5	15.5	4.5	51	0.8	6.8	8.9	5.0	9.4	2.8	0.4
2707	RC2_F2_O_019	Offices	Shared Cellular Offices	15	41	8.0	5	1	7	1.5E-03	39	8.6E-03	13	4.5	3	0.2	0.5	15.5	4.5	51	0.8	6.8	8.9	5.0	9.5	2.8	0.4
2708	RC2_F2_O_020	Offices	Single Cellular Office	10	27	8.0	5	1	6	1.3E-03	30	6.7E-03	13	9.0	1	0.2	0.6	15.4	4.5	50	1.1	7.2	8.7	5.0	8.5	7.6	0.4
2709	RC2_F2_O_021	Offices	Shared Cellular Offices	13	35	8.0	185	1	6	1.3E-03	37	8.2E-03	13	4.5	3	0.2	0.5	16.7	4.5	48	1.1	7.1	10.0	6.1	10.3	3.6	0.4
2710	RC2_F2_O_022	Offices	Shared Cellular Offices	14	39	8.0	185	1	7	1.5E-03	38	8.4E-03	13	4.5	3	0.2	0.5	16.8	4.5	48	1.1	7.1	10.1	6.0	11.4	3.6	0.4
2711	RC2_F2_O_023	Offices	Shared Cellular Offices	14	39	8.0	185	1	7	1.5E-03	38	8.4E-03	13	4.5	3	0.2	0.5	16.6	4.5	48	1.1	7.2	9.9	6.1	11.5	3.6	0.4
2712	RC2_F2_O_024	Offices	Shared Cellular Offices	13	35	8.0	185	1	6	1.3E-03	37	8.2E-03	13	4.5	3	0.2	0.5	16.2	4.5	48	1.1	7.6	9.5	6.2	10.5	3.7	0.4
2713	RC2_F2_O_025	Offices	Single Cellular Office	11	30	8.0	275	1	6	1.3E-03	33	7.3E-03	13	9.0	1	0.2	0.5	14.8	4.5	52	2.5	7.6	8.1	12.4	21.1	16.9	0.4
2714	RC2_F2_O_026	Offices	Single Cellular Office	11	30	8.0	275	1	6	1.3E-03	33	7.3E-03	13	9.0	1	0.2	0.5	14.8	4.5	52	2.6	7.6	8.1	13.1	22.2	17.8	0.4
2715	RC2_F2_O_027	Offices	Shared Cellular Offices	14	39	8.0	185	1	13	2.8E-03	29	6.4E-03	13	4.5	3	0.3	0.9	18.3	4.5	50	0.6	5.8	11.6	1.7	6.1	1.9	0.3
2716	RC2_F3_EF_DL01	Educational Facilities	Computer Rooms	43	118	11.3	50	2	27	6.0E-03	46	1.0E-02	13	4.0	11	0.2	0.6	15.2	4.8	52	0.8	7.8	8.5	3.6	27.1	2.5	0.5
2717	RC2_F3_EF_SR01	Educational Facilities	Computer Rooms	100	272	11.3	5	1	30	6.7E-03	108	2.4E-02	13	4.0	25	0.1	0.3	15.4	4.8	52	0.6	7.3	8.7	5.3	45.3	1.8	0.4
2718	RC2_F3_O_001	Offices	Single Cellular Office	10	26	11.3	230	2	17	3.7E-03	17	3.7E-03	13	9.0	1	0.6	1.7	14.6	4.8	50	5.0	8.7	7.9	7.8	36.8	34.0	0.5
2719	RC2_F3_O_002	Offices	Shared Cellular Offices	15	41	11.3	275	1	12	2.5E-03	32	7.0E-03	13	4.5	3	0.3	0.8	15.0	4.8	50	2.9	7.7	8.3	10.2	32.7	9.8	0.5
2720	RC2_F3_O_003	Offices	Shared Cellular Offices	15	41	11.3	275	1	11	2.4E-03	31	6.9E-03	13	4.5	3	0.3	0.7	15.4	4.8	49	2.7	7.5	8.7	10.3	30.9	9.3	0.4
2721	RC2_F3_O_004	Offices	Shared Cellular Offices	15	41	11.3	275	1	12	2.5E-03	32	7.0E-03	13	4.5	3	0.3	0.8	15.4	4.8	49	2.9	7.6	8.7	10.3	33.1	9.9	0.4
2722	RC2_F3_O_005	Offices	Single Cellular Office	10	26	11.3	275	1	8	1.8E-03	26	5.7E-03	13	9.0	1	0.3	0.8	15.2	4.8	48	3.3	7.8	8.5	10.6	23.9	22.0	0.5
2723	RC2_F3_O_006	Offices	Shared Cellular Offices	22	60	11.3	95	1	14	3.1E-03	41	9.1E-03	13	4.5	5	0.2	0.6	16.0	4.8	50	0.4	6.6	9.3	1.5	6.0	1.2	0.4
2724	RC2_F3_O_007	Offices	Single Cellular Office	10	26	11.3	95	1	6	1.3E-03	33	7.3E-03	13	9.0	1	0.2	0.6	15.7	4.8	49	0.5	7.6	9.0	2.0	3.3	3.1	0.4
2725	RC2_F3_O_008	Offices	Shared Cellular Offices	15	40	11.3	275	1	12	2.7E-03	30	6.6E-03	13	4.5	3	0.3	0.8	15.5	4.8	49	3.3	7.7	8.8	10.8	36.5	11.2	0.5
2726	RC2_F3_O_009	Offices	Shared Cellular Offices	13	35	11.3	275	1	11	2.4E-03	28	6.3E-03	13	4.5	3	0.3	0.8	15.5	4.8	49	3.5	8.1	8.8	11.3	34.2	11.8	0.5
2727	RC2_F3_O_010	Offices	Open Offices	80	218	11.3	320	2	57	1.3E-02	61	1.3E-02	13	6.0	13	0.3	0.7	14.8	4.8	52	1.8	8.2	8.1	6.9	109.5	8.2	0.5
2728	RC2_F3_O_011	Offices	Shared Cellular Offices	17	46	11.3	5	1	12	2.5E-03	33	7.3E-03	13	4.5	4	0.3	0.7	15.0	4.8	53	1.2	7.3	8.3	4.7	15.0	4.0	0.4
2729	RC2_F3_O_012	Offices	Shared Cellular Offices	25	67	11.3	50	2	25	5.5E-03	32	7.0E-03	13	4.5	6	0.4	1.0	14.8	4.8	53	1.6	8.0	8.1	4.3	29.6	5.4	0.5

The time-series outputs provide necessary information on the concentrations, temporal variation, and dynamics of $PM_{2.5}$ in indoor environments. Figure 4.8 shows the hourly indoor $PM_{2.5}$ concentration of four randomly selected zones in the Regent Court building using the baseline Q_{50} of 10 $m^3/h/m^2$. These zones show a south-facing room on the first floor (Z30), a south-east room on the second floor (Z122), a north-facing room on the first floor (Z96), and a north-west facing room on the third floor (Z87). Figure 4.9 shows the temporal and spatial variation of infiltrated $PM_{2.5}$. This could be due to the mediating effects of environmental and building characteristics on the ingress of $PM_{2.5}$ from outdoor sources. This highlights the necessity of achieving a model resolution at the room level for such complex buildings over time.

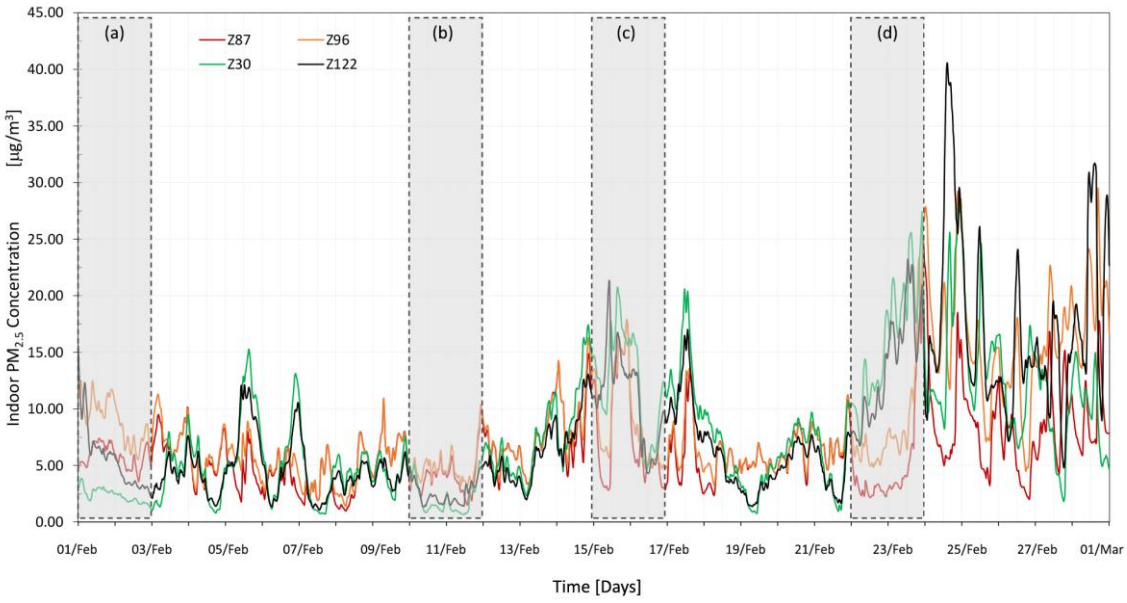


Figure 4.8: Hourly indoor $PM_{2.5}$ concentration in four different zones in the Regent Court Building, showing the spatial and temporal variability in concentrations within the same building. The Q_{50} of this building is 10 $m^3/h/m^2$.

Zooming into (a), (b), (c), and (d) in Figure 4.8, it can be seen in Figures 4.9(a) to 4.8(d) that the concentrations of infiltrated $PM_{2.5}$ can vary significantly throughout the day in different locations/zones within the same building. The spatial variability shown here is essential as it causes exposure disparities among building users in different zones/rooms within the same building. As such, studying the causes of the spatial variability on indoor $PM_{2.5}$ becomes essential. This can only be achieved through constructing models of a high spatial resolution at the room level.

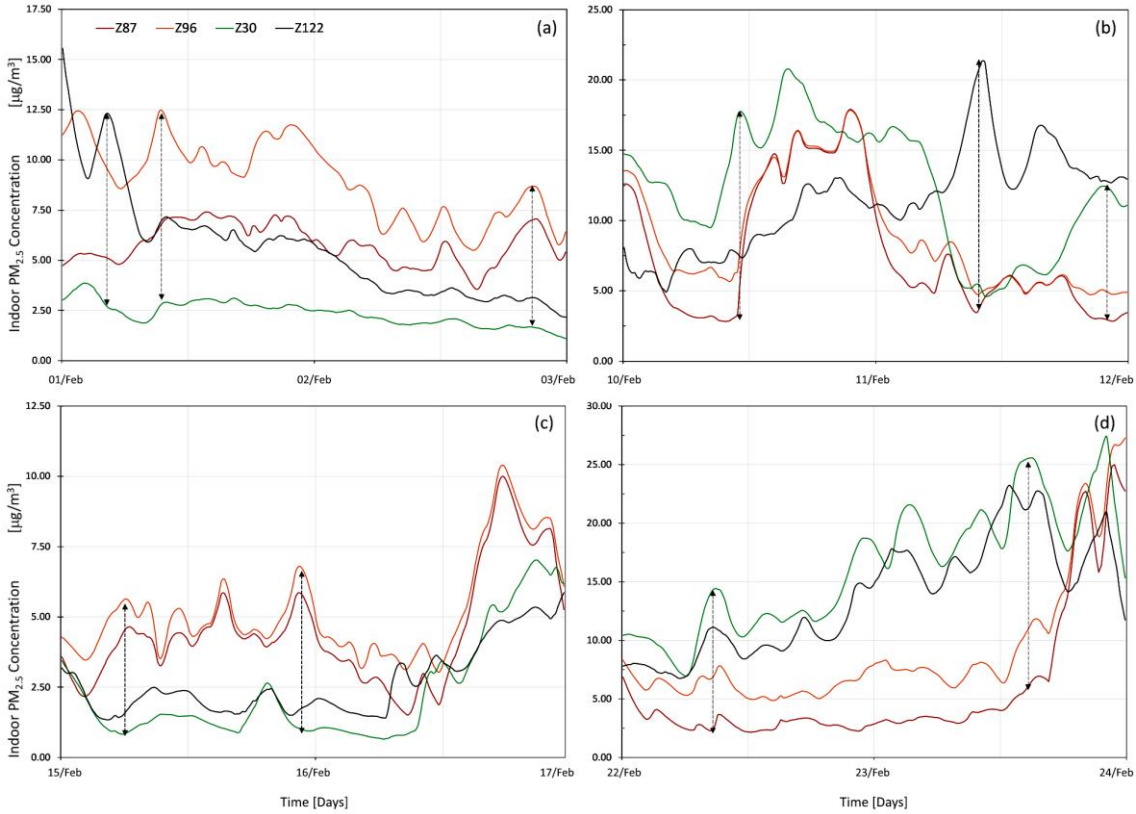


Figure 4.9 (a-d): Hourly indoor $PM_{2.5}$ concentration in four different zones in the Regent Court Building as data samples showing the spatial (zone location) and temporal (black arrows ↔) variability in concentrations within the same building.

Unsurprisingly, the time series data of multiple buildings reveals a temporal and spatial variability in indoor $PM_{2.5}$ concentrations in different zones in different buildings. Figure 4.10 shows the time series of hourly indoor $PM_{2.5}$ concentrations in four randomly selected zones in the four buildings (BH, AT, ADC, RC) using the baseline Q_{50} values in Section 3.3.2. The four Office zones represent a south-facing room in the BH building, a north-facing room in the ADC, a south-east-facing room in the Arts Tower, and a south-facing room in the Regent Court. Figures 4.11 (a)-(d) show the trends of infiltrated $PM_{2.5}$ concentrations in the selected timeframes. Interestingly, it can be noticed that the trends of infiltrated $PM_{2.5}$ vary throughout the day towards the end of the month when they exhibit a similar trend between the 22nd and 24th of February (Figure 4.10(d)). These plots of time series outputs can be seen as evidence showing that $PM_{2.5}$ in individual zones of the same space type (Office in this case) are sometimes similar in trend (overall peaks and troughs) but differ in location. The implications of such variability are discussed in detail in Chapter 8.

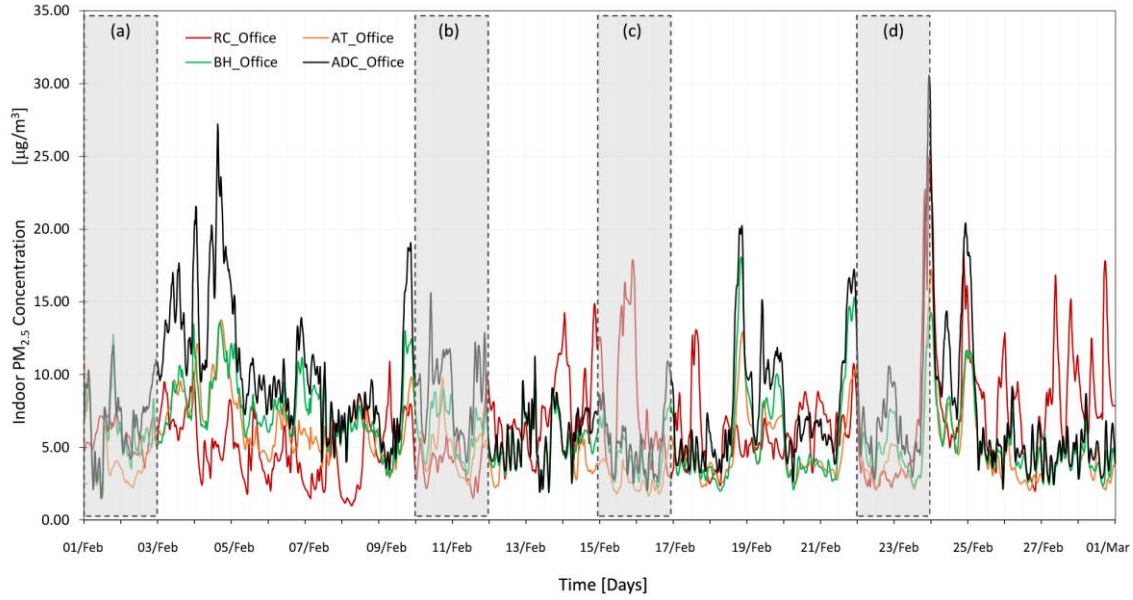


Figure 4.10: Hourly indoor $PM_{2.5}$ concentration levels in four randomly selected office zones in Regent Court (RC), Arts Tower (AT), Academic Development Centre (ADC), and Barber House (BH); zoom-in plots of (a) to (d) are shown in Figure 4.10 (a-d)

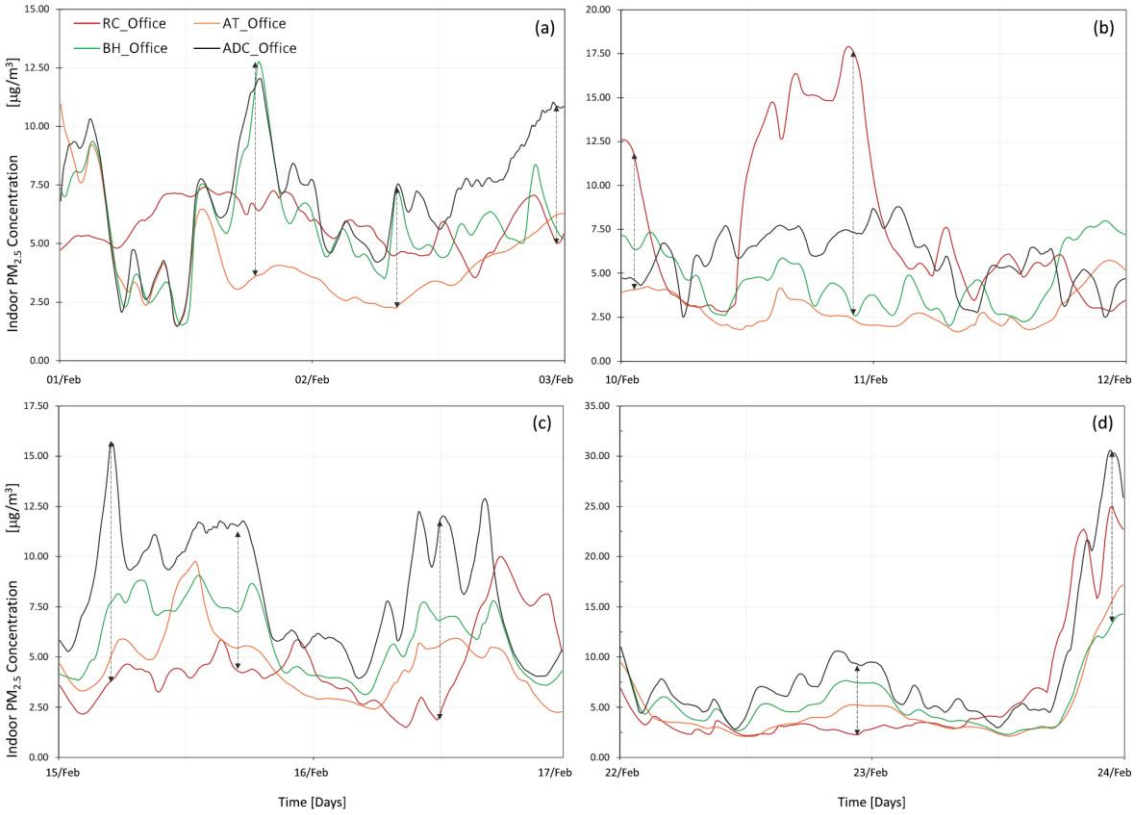


Figure 4.11 (a-d): Hourly indoor $PM_{2.5}$ concentration levels in the Office zones of RC, AT, BH, and ADC over the four periods in February 2019 ((black arrows ↔ highlighting the temporal variation)

4.3.2 Baseline Concentrations of Indoor $PM_{2.5}$

In order to examine the relationships between infiltrated $PM_{2.5}$ concentrations and different environmental and zone characteristics variables, the co-simulation outputs were resampled to generate average concentrations of infiltrated $PM_{2.5}$ over the heating season (Nov – April) and the year (annual). This allows for a better pairing of indoor $PM_{2.5}$ and several input variables, as seen in Table 4.4. Although previous studies have reported that building characteristics (e.g., building type, age, and floor level) can influence indoor air pollution, the building envelope airtightness Q_{50} can also modify the distribution of population exposure to infiltrated $PM_{2.5}$ from outdoor sources across an urban area. Using the baseline values of Q_{50} identified previously in Section 3.3.2 (ADC = 13 $m^3/h/m^2$, BH = 13 $m^3/h/m^2$, AT = 10 $m^3/h/m^2$, and the RC = 10 $m^3/h/m^2$), the heating season concentrations and annual concentrations of infiltrated $PM_{2.5}$ are summarised in Table 4.5 below.

Table 4.5: Descriptive Statistics for the Baseline Concentrations of Infiltrated $PM_{2.5}$ Over the Heating Season and the (*Annual Average Concentrations*).

Building	Q_{50} ($m^3/h/m^2$)	Number of Zones	Min-Max (µg/m^3)	Mean (µg/m^3)	Median (µg/m^3)	*Variance	**Standard Deviation
BH	13	20	5.16-9.24	7.04	6.95	1.39	1.14
			(8.11-14.06)	*(10.83)*	*(10.80)*	*(2.79)*	*(1.61)*
ADC	13	26	3.33-8.15	6.86	6.79	1.74	1.13
			(5.28-12.61)	*(9.97)*	*(9.61)*	*(1.59)*	*(1.38)*
AT	10	185	3.63-9.39	6.02	5.96	1.06	1.02
			(5.59-13.82)	*(9.02)*	*(8.85)*	*(2.04)*	*(1.42)*
RC	10	224	4.53-8.53	6.56	6.59	0.55	0.74
			(7.71-13.15)	*(10.38)*	*(10.43)*	*(1.01)*	*(1.00)*

*Variance: a statistical measure that quantifies the degree of dispersion or spread in a dataset by calculating the average of the squared differences between each data point and the dataset's mean. **Standard Deviation: is a measure of the amount of variation or dispersion in a dataset, representing the square root of the variance.

The results demonstrate the annual concentrations of $PM_{2.5}$ in the different buildings, reflecting the long-term exposure to *infiltrated* $PM_{2.5}$. Building BH exhibited an annual $PM_{2.5}$ concentration range of 8.11-14.06 µg/m³, with a mean concentration of 10.83 µg/m³. Building ADC showed a slightly lower annual $PM_{2.5}$ concentration range of 5.28-12.61 µg/m³, with a mean concentration of 9.97 µg/m³, although they share the same Q_{50} = 13 $m^3/h/m^2$. Building AT had an annual $PM_{2.5}$ concentration range of 5.59-13.82 µg/m³, with a mean concentration of 9.02 µg/m³. Building RC demonstrated an annual $PM_{2.5}$ concentration range of 7.71-13.15 µg/m³, with a mean concentration of 10.38 µg/m³.

Interestingly, the observed differences in $PM_{2.5}$ concentrations between the AT and RC buildings can be attributed to several factors, including variations in building design and the potential pollutant dispersion within the buildings. As a high-rise building, it is common for the AT to experience a gradient of $PM_{2.5}$ concentrations across different zones due to differences in floor levels, outdoor pollutant intrusion, and variations in ventilation effectiveness at different heights. The range of annual $PM_{2.5}$ concentrations, from 5.59 μg/m³ to 13.82 μg/m³, indicates a considerable variation in the exposure levels experienced by occupants within the building. In contrast, the RC, also sharing the same airtightness value (Q_{50} = 10 $m^3/h/m^2$), demonstrated a narrower range of annual $PM_{2.5}$ concentrations, from 7.71 μg/m³ to 13.15 μg/m³. The relatively minor range suggests a more consistent distribution of $PM_{2.5}$ concentrations across the zones within the building.

The standard deviation values indicate the spread of $PM_{2.5}$ concentrations around the mean, reflecting the variability within each building. Buildings BH, ADC, AT, and RC had standard deviations of 1.61 μg/m³, 1.38 μg/m³, 1.42 μg/m³, and 1.00 μg/m³, respectively, suggesting different levels of variability in $PM_{2.5}$ concentrations. Additionally, the variance values provide further insight into the dispersion of $PM_{2.5}$ concentrations within each building. The calculated variances for Buildings BH, ADC, AT, and RC were 2.79 μg/m³, 1.59 μg/m³, 2.04 μg/m³, and 1.01 μg/m³, respectively, indicating the degree of $PM_{2.5}$ concentration variability across different zones within the buildings. These results align with previous findings on the variability of $PM_{2.5}$ concentrations within a building due to building design and environmental factors (Elliot et al., 2000).

Finally, the results highlight the significant proportion of zones within each building that exceeded the WHO recommended long-term exposure limit of 10 μg/m³ for $PM_{2.5}$ concentrations. The BH had 77% of its zones exceeding the limit, accounting for 15 zones. Similarly, the ADC Building had 61% of its zones above the recommended limit, totalling 15 zones. The AT showed 60% of its zones exceeding the WHO limit, with 111 zones surpassing the recommended threshold. Finally, the RC Building had 70% of its zones above the limit, encompassing 157 zones. Although these findings are subject to uncertainty due to the selection of Q_{50} values, they underscore the substantial IAQ concerns in the studied buildings, indicating a considerable number of zones with $PM_{2.5}$ concentrations exceeding the recommended WHO long-term exposure limit.

4.3.3 Impact of Q_{50} on the Concentrations of Indoor $PM_{2.5}$

This section examined the impact of improving the airtightness of the building envelope represented by Q_{50} values on the concentrations of infiltrated $PM_{2.5}$ over the heating season and for a year. Initial observations from the co-simulation revealed such variations when the Q_{50} was varied between 3-13 $m^3/h/m^2$, see Table 4.6. It can be noted that the average heating season concentrations of infiltrated $PM_{2.5}$ were 3.94 ± 0.98 $\mu g/m^3$ when Q_{50} = 3 $m^3/h/m^2$. When the Q_{50} was selected to represent a leaky building, i.e., Q_{50} = 13 $m^3/h/m^2$, the average heating season concentrations were about 57% higher (6.9 ± 1.14) $\mu g/m^3$. The data also revealed a clear relationship between the Q_{50} value and the infiltration air change rates ACH_{INF} during the heating season. As the Q_{50} value increased, the ACH_{INF} also increased, indicating a higher air exchange rate between indoor and outdoor environments. Table 4.7 shows that the mean ACH_{INF} progressively increased from 0.38 h^{-1} for Q_{50} = 3 $m^3/h/m^2$ to 1.37 h^{-1} for Q_{50} = 13 $m^3/h/m^2$. This represents a substantial increase of approximately 260.5% in the ACH_{INF} over the heating season due to a leaky building envelope, see Figure 4.12.

Table 4.6: Descriptive Statistics for the Concentrations of Infiltrated $PM_{2.5}$ by Building Q_{50} Over the Heating Season and the (*Annual Average Concentrations*).

Q_{50} ($m^3/h/m^2$)	**Number of Zones**	**Min-Max ($\mu g/m^3$)**	**Mean ($\mu g/m^3$)**	**Median ($\mu g/m^3$)**	***Variance**	****Standard Deviation**
3	455	0.61-7.19	3.94	3.90	0.97	0.98
		(2.16-11.29)	*(6.86)*	*(6.85)*	*(1.85)*	*(1.36)*
5	455	0.98-7.96	4.86	4.94	1.16	1.08
		(2.97-12.23)	*(8.09)*	*(8.18)*	*(2.12)*	*(1.46)*
7	455	1.34-8.43	5.54	5.65	1.36	1.17
		(3.65-12.90)	*(8.99)*	*(9.12)*	*(2.44)*	*(1.56)*
9	455	1.70-8.64	5.92	6.11	1.31	1.14
		(4.27-13.24)	*(9.51)*	*(9.77)*	*(2.29)*	*(1.51)*
11	455	2.03-8.98	6.57	6.76	1.25	1.12
		(4.81-13.70)	*(10.41)*	*(10.61)*	*(2.02)*	*(1.42)*
13	455	2.33-9.24	6.90	7.12	1.30	1.14
		(5.28-14.06)	*(10.87)*	*(11.11)*	*(2.09)*	*(1.45)*

*Variance: a statistical measure that quantifies the degree of dispersion or spread in a dataset by calculating the average of the squared differences between each data point and the dataset's mean. **Standard Deviation: is a measure of the amount of variation or dispersion in a dataset, representing the square root of the variance.

Table 4.7: Descriptive Statistics for the ACH_{INF} by Building Q_{50} Over the Heating Season (Nov – April)

Q_{50} ($m^3/h/m^2$)	Number of Zones	Min-Max (h^{-1})	Mean (h^{-1})	Median (h^{-1})	*Variance	**Standard Deviation
3	455	0.06-1.57	0.38	0.32	0.05	0.23
5	455	0.11-2.22	0.60	0.50	0.14	0.37
7	455	0.16-2.97	0.84	0.72	0.26	0.51
9	455	0.20-3.63	0.97	0.81	0.32	0.57
11	455	0.24-4.32	1.17	0.97	0.52	0.72
13	455	0.28-5.02	1.37	1.14	0.71	0.84

*Variance: a statistical measure that quantifies the degree of dispersion or spread in a dataset by calculating the average of the squared differences between each data point and the dataset's mean. **Standard Deviation: is a measure of the amount of variation or dispersion in a dataset, representing the square root of the variance.

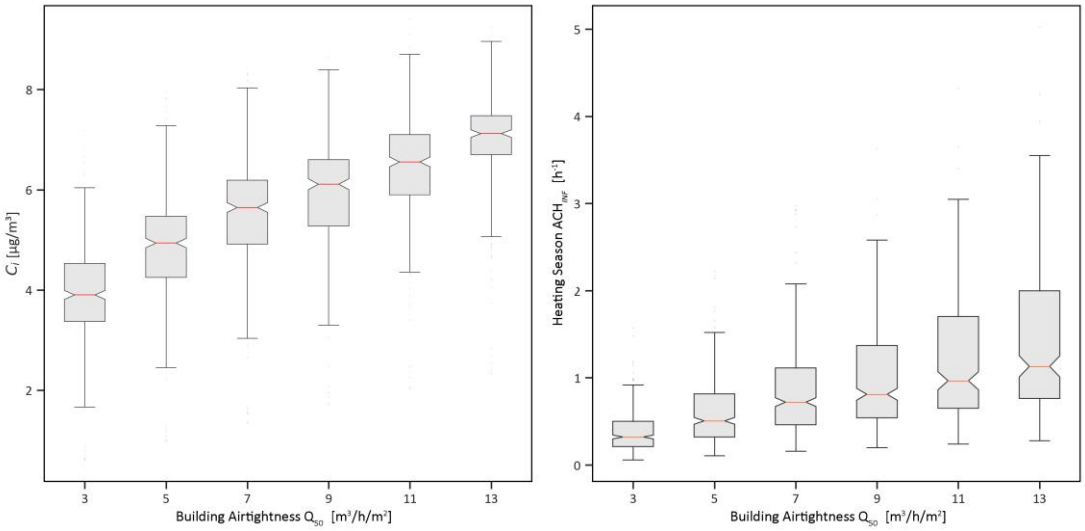


Figure 4.12: Box plots of the Heating Season Concentrations of Infiltrated $PM_{2.5}$ C_i and the ACH_{INF} stratified by the Building Envelope Airtightness (Q_{50})

The mean concentrations of infiltrated $PM_{2.5}$ varied among the different Q_{50} values. For a Q_{50} value of 3 $m^3/h/m^2$, the mean concentration was 6.86 ± 1.36 µg/m^3 µg/m3. As the Q_{50} value increased, the mean concentrations of $PM_{2.5}$ also increased. The highest mean concentration of 10.87 ± 1.45 µg/m^3 µg/m3 was observed for a Q_{50} value of 13 $m^3/h/m^2$. These findings suggest that improving the airtightness of the building envelope, as represented by tighter building envelopes (low Q_{50} values), tends to result in lower average concentrations of infiltrated $PM_{2.5}$.The standard deviation values, which indicate the dispersion of data points around the mean, ranged from 1.36 µg/m3 to 1.45 µg/m3 across the Q_{50} values of 3 to 13 $m^3/h/m^2$,

respectively. This indicates a moderate level of variability in the annual concentrations of $PM_{2.5}$ within each Q_{50} group. However, the differences in standard deviation among the Q_{50} groups were relatively small.

The impact of improving the Q_{50} airtightness value on the percentage of zones exceeding the WHO long-term indoor $PM_{2.5}$ limit of 10 $\mu g/m^3$ was evaluated in this study. Figure 4.13 below shows a cumulative distribution function of the annual concentrations of $PM_{2.5}$ due to varying the airtightness value Q_{50}. As the Q_{50} increased, there was a noticeable reduction in the percentage of zones exceeding the WHO limit. For the highest Q_{50} value considered, 13 $m^3/h/m^2$, the percentage of exceedance was 82%, with a total of 373 zones exceeding the limit. A decrease in the Q_{50} value to 11 $m^3/h/m^2$ resulted in a slightly lower but still significant percentage of exceedance of 77%, with 350 zones surpassing the threshold.

Further improvement in airtightness to a Q50 value of 9 $m^3/h/m^2$ led to a substantial decrease in the percentage of exceedance to 41%, with 186 zones exceeding the WHO limit. As the airtightness improved with a Q_{50} value of 7 $m^3/h/m^2$, only 22% of zones exceeded the limit, with 100 zones surpassing it. The trend of decreasing exceedance percentages continued as the Q_{50} value decreased further. For a Q_{50} value of 5 $m^3/h/m^2$, the percentage of exceedance was reduced to 4%, with only 25 zones exceeding the limit. The lowest percentage of exceedance was observed for the Q_{50} value of 3 $m^3/h/m^2$, where only 1% of zones exceeded the WHO limit, with 5 zones surpassing it.

These findings highlight the significant influence of improving airtightness, as represented by tighter building envelopes, in reducing the percentage of zones exceeding the WHO recommended limit for long-term indoor $PM_{2.5}$ exposure. By implementing airtightness interventions, the infiltrated levels of $PM_{2.5}$ can be effectively controlled, thereby enhancing IAQ and potentially minimising health risks associated with $PM_{2.5}$ exposure. It is important to note that while improving airtightness is crucial in reducing the percentage of exceedance, other factors, such as outdoor air quality and local pollutant sources, should also be considered when addressing indoor $PM_{2.5}$ concentrations.

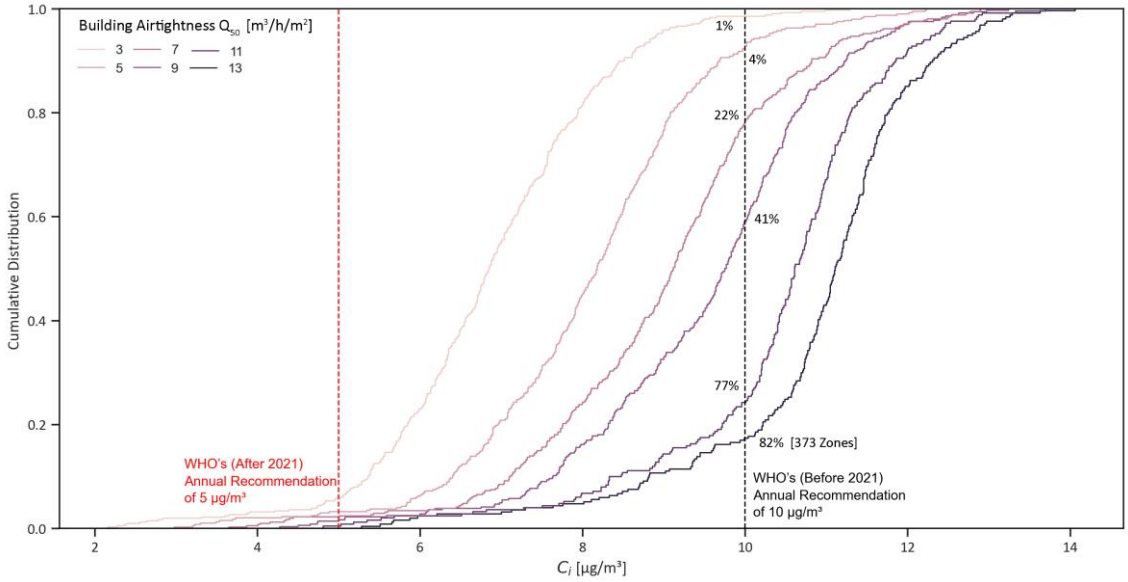


Figure 4.13: A Cumulative Distribution Function CDF showing the percentage of zones with an annual average concentration of infiltrated $PM_{2.5}$ above the WHO permissible levels of 10 µg/m^3 (before 2021)

4.3.4 Indoor/Outdoor (I/O) $PM_{2.5}$ Ratio

The I/O ratio is utilised to assess the disparity between indoor $PM_{2.5}$ concentrations and the corresponding outdoor concentrations and gauge indoor sources' strength within buildings. Indoor $PM_{2.5}$ concentrations are affected by the infiltration of outdoor $PM_{2.5}$ into buildings and indoor sources (Huang et al., 2007). Numerous factors, such as building design, location, and various indoor activities, can lead to significant variations in the I/O ratio. Table 4.6 summarises the heating season and annual $PM_{2.5}$ I/O ratio when stratified by building Q_{50}. It was found from the co-simulation results that the annual $PM_{2.5}$ I/O ratio ranged between 0.26-1.03 for all zones, with an average of 0.66 ± 0.06 when Q_{50} = 3 m^3/h/m^2. This indicates a high spatial variability of infiltrated $PM_{2.5}$ concentrations within the same building and across buildings sharing the same Q_{50}. This spatial disparity in infiltrated $PM_{2.5}$ highlights the importance of considering the model resolution of "individual zones" as an essential factor in estimating the population exposure in HEI buildings.

Table 4.8: Descriptive Statistics for the Indoor/Outdoor Ratio of Infiltrated $PM_{2.5}$ by Building Q_{50} Over the Heating Season and the (Annual Average I/O Ratio).

Q_{50} ($m^3/h/m^2$)	Number of Zones	Min-Max	Mean	Median	*Variance	**Standard Deviation
3	455	0.04-0.42	0.23	0.23	<0.01	0.06
		(0.26-1.03)	*(0.66)*	*(0.66)*	*(0.01)*	*(0.11)*
5	455	0.06-0.47	0.29	0.29	<0.01	0.06
		(0.35-1.11)	*(0.76)*	*(0.50)*	*(0.01)*	*(0.12)*
7	455	0.08-0.49	0.33	0.33	<0.01	0.07
		(0.42-1.16)	*(0.83)*	*(0.84)*	*(0.02)*	*(0.13)*
9	455	0.10-0.51	0.35	0.36	<0.01	0.07
		(0.48-1.19)	*(0.88)*	*(0.81)*	*(0.02)*	*(0.12)*
11	455	0.12-0.53	0.39	0.40	<0.01	0.07
		(0.53-1.22)	*(0.95)*	*(0.97)*	*(0.01)*	*(0.11)*
13	455	0.14-0.54	0.41	0.42	<0.01	0.07
		(0.57-1.25)	*(0.99)*	*(1.01)*	*(0.01)*	*(0.11)*

*Variance: a statistical measure that quantifies the degree of dispersion or spread in a dataset by calculating the average of the squared differences between each data point and the dataset's mean. **Standard Deviation: is a measure of the amount of variation or dispersion in a dataset, representing the square root of the variance.

The relationship between indoor and outdoor $PM_{2.5}$ concentrations was examined in different zones within the RC building, specifically Zone (1) and Zone (2), located on the first floor but in different locations. The analysis focused on the determination coefficient (R^2) values, which provide insights into the degree of correlation between indoor and outdoor $PM_{2.5}$ concentrations using the daily temporal scale when the Q_{50} was 3 and 7 $m^3/h/m^2$, see Figure 4.14. For Zone (1), with a Q_{50} value of 7 $m^3/h/m^2$, the R^2 value between outdoor and indoor $PM_{2.5}$ concentrations was 0.89. This suggests that approximately 89% of the variation in indoor $PM_{2.5}$ concentrations can be attributed to changes in outdoor $PM_{2.5}$ levels, indicating a more robust correlation than the Q_{50} value of 3 $m^3/h/m^2$ ($R^2 = 0.85$). Similarly, for Zone (2), with a Q_{50} value of 7 $m^3/h/m^2$, the R^2 value was 0.83, indicating a significant correlation between outdoor and indoor $PM_{2.5}$ concentrations when compared to the Q_{50} of 3 $m^3/h/m^2$ ($R^2 = 0.72$).

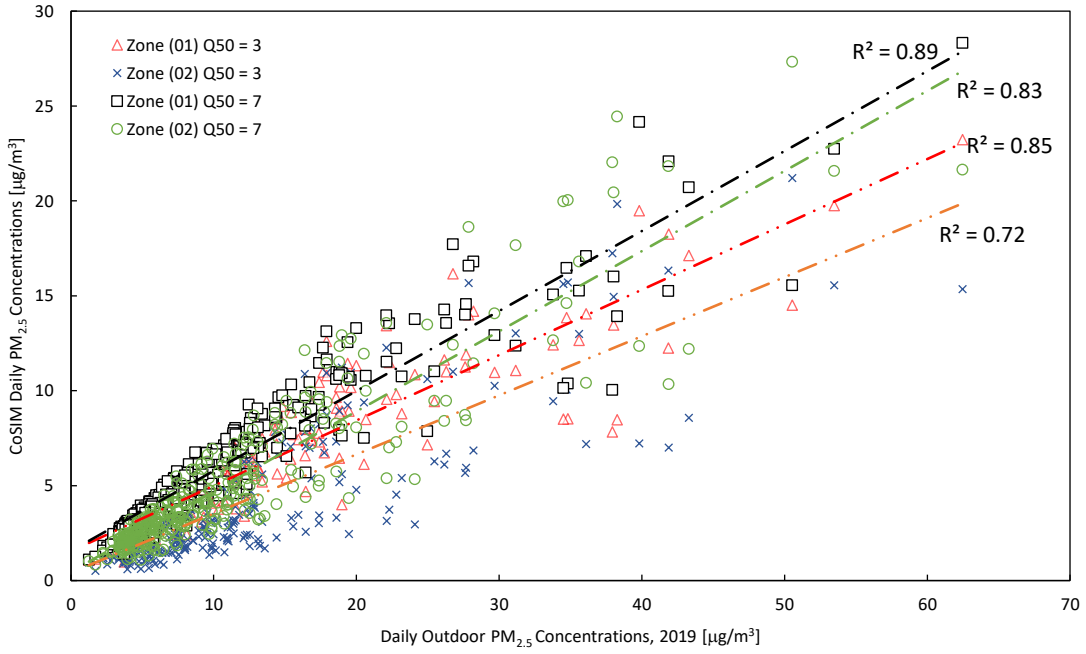


Figure 4.14: Scatter Plots of the Daily Outdoor $PM_{2.5}$ concentrations and the Daily Infiltrated $PM_{2.5}$ in 2 Different Zones in the Regent Court Building when Q_{50} = 3 and 7 $m^3/h/m^2$

The relationship between outdoor and indoor $PM_{2.5}$ concentrations was also analysed in different zones within the AT building. Specifically, Zone (1) on the 3rd floor and Zone (2) on the 12th floor were investigated to understand the influence of floor level on the infiltration of $PM_{2.5}$ indoors, see Figure 4.15. It can be noticed that when Zone (1) had a Q_{50} value of 7 $m^3/h/m^2$, and exhibited a strong correlation between outdoor and indoor $PM_{2.5}$ concentrations, with an R^2 of 0.95. Similarly, Zone (2), with the same Q_{50} value of 7 $m^3/h/m^2$, showed a moderately strong relationship with an R^2 value of 0.79. Following the improvement in airtightness, where the Q_{50} value was reduced to 3 $m^3/h/m^2$, changes in the relationship between outdoor and indoor $PM_{2.5}$ concentrations were observed. In Zone (1), the R^2 value decreased slightly to 0.90. This indicates that although the relationship between outdoor and indoor $PM_{2.5}$ concentrations remains strong, the improvement in airtightness led to a slight reduction in the strength of this relationship. Similarly, in Zone (2) on the 12th floor, the R^2 value decreased further to 0.66, indicating a weaker relationship between outdoor and indoor $PM_{2.5}$ concentrations after the airtightness improvement.

These findings suggest that while airtightness improvements can effectively reduce the infiltration of $PM_{2.5}$ particles into indoor environments, the impact may vary depending on other factors such as building height. The decrease in the R^2 values observed in Zone (1) and Zone (2) after tightening the envelope (i.e., Q_{50} = 3 $m^3/h/m^2$) suggests that height can also play a role in the infiltration of $PM_{2.5}$, with higher floors potentially experiencing a more significant reduction in the correlation between outdoor and indoor concentrations following airtightness improvements.

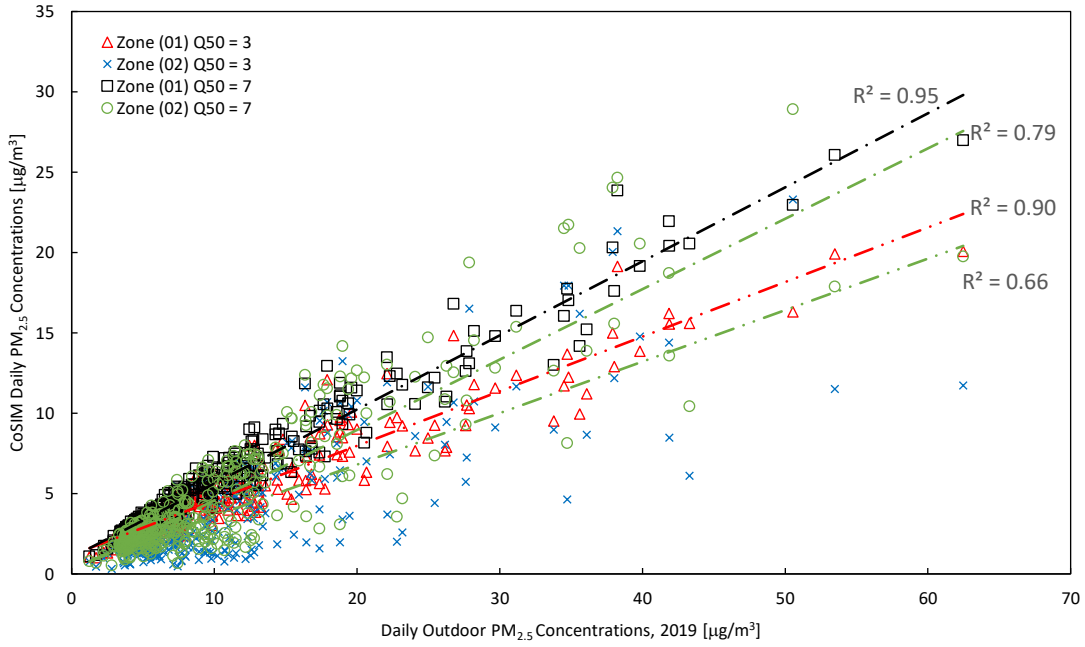


Figure 4.15: Scatter Plots of the Daily Outdoor $PM_{2.5}$ concentrations and the Daily Infiltrated $PM_{2.5}$ in 2 Different Zones in the Arts Tower Building when Q_{50} = 3 and 7 $m^3/h/m^2$

4.4 Validation of the Co-Simulation Results

As previously mentioned, CONTAM and EnergyPlus have been extensively used in research for over thirty years. In order to verify their applicability for research and analysis, several analytical and empirical validation efforts (Emmerich & Hirnikel, 2001b) have been conducted on different building types and locations (L. C. Ng et al., 2012), and pollutants (L. J. Underhill et al., 2018). Additionally, validation enhances the model's credibility by ensuring simulation predictions are more closely aligned with actual observations. Statistical evaluation of indoor air quality models can be found in the American Society for Testing and Materials (ASTM) D5157-19 standard

guide[8]. The topics presented in the ASTM guide are establishing evaluation objectives, selecting datasets for evaluation, statistical methods for analysing model performance, and considerations when using these methods. ASTM D5157 provides three statistical tools to assess the accuracy of IAQ predictions, and two additional statistical tools are provided to assess bias.

In order to determine whether predictions are in agreement with observations (measurements), the following measures are used:

- Predictions and measurements should have a correlation coefficient of 0.9 or greater.
- The regression line between predictions and measurements should have a slope between 0.75 and 1.25 and an intercept of less than 25% of the average concentration.
- The normalised mean square error (NMSE) is less than 0.25 and can be calculated as follows:

$$\text{NMSE} = \sum_{i=1}^{N} (C_{pi} - C_{oi})^2 / (\bar{C}_o \bar{C}_p) \tag{4.2}$$

Where *N* is the number of observations (measurements) in the datasets, C_p is the predicted concentration and C_o is the observed concentration.

As affected by the COVID-19 pandemic restriction, it was not possible to validate the IAQ models developed in this research by following the ASTM guide for validation. The field measurement phase of the research was initially planned to occur between January 2020 and January 2021; however, due to the UK Government's restrictions (i.e., nationwide lockdowns), field measurement data collection was not possible during this period. It was also challenging to gather occupancy-related data through surveys since access to every university building was severely restricted.

In the summer of 2022, another doctoral student at the School of Architecture (UoS) conducted environmental research in the Arts Tower. The research involved collecting indoor environmental data (temperature, humidity, and CO_2) inside an office on the 9th floor. The measurements were taken on a 5-min timestep between 11:00 AM on 11/05/2022 and 11:00 AM on 17/05/2022, resulting in a sample size of 1,720 data points. The sensor and data logger (HOBO UX100-003)

[8] https://www.astm.org/d5157-19.html

were placed in the middle of the room and were 1 meter above floor level. A survey was also used to collect information from the occupants of that office. This included their behaviour, opening and closing windows and doors, level of clothing, and perception of the room temperature (see Table 4.5). Alongside the Arts Tower data obtained from the EFM, the field measurements made available by the doctoral study were deemed suitable for a model validation exercise (indoor air temperature only) in this study.

It should be noted that the validation presented here was merely to demonstrate the author's understanding of the validation process and its importance. Without a statistical evaluation guide for simulated indoor air temperatures against field measurements, the ASTM standard was used in this exercise.

Table 4.9: Results of the validation (Indoor air temperatures, Floor 9, Arts Tower)

Properties of the room under investigation					
Date and Time of Measurements		11/05/2022 (11:00 AM) – 17/05/2022 (11:00 AM)			
Location		Room 9.02 on the 9th Floor of the Arts Tower Building			
Orientation	North	**Number of Occupants at the Time of Measurement**		1	
Room Area	46 m^2	**Heating Policy**	Off		
Q_{50}	10 $m^3/h/m^2$	**External Wall**	Double Glazed Curtain Wall U-Value 2.2 $W/m^2.K$		
		Windows	Closed		
		Results			
	Number of Samples	**Average Indoor Temperature**	**Standard Deviation**	**Correlation Coefficient**	**NMSE**
Co-Simulation Results	1,720	22.34 °C	1.029	0.76	0.48
Field Measurements	1,720	22.47 °C	0.986		

The overall correlation coefficient of 0.76 with an NMSE of 0.48 indicates some agreement between the simulated results and field measurements (Figure 4.17). The difference between the average values was only 0.07 °C, so these values were considered reasonable for indoor air temperature. However, they are lower than what the ASTM guide states.

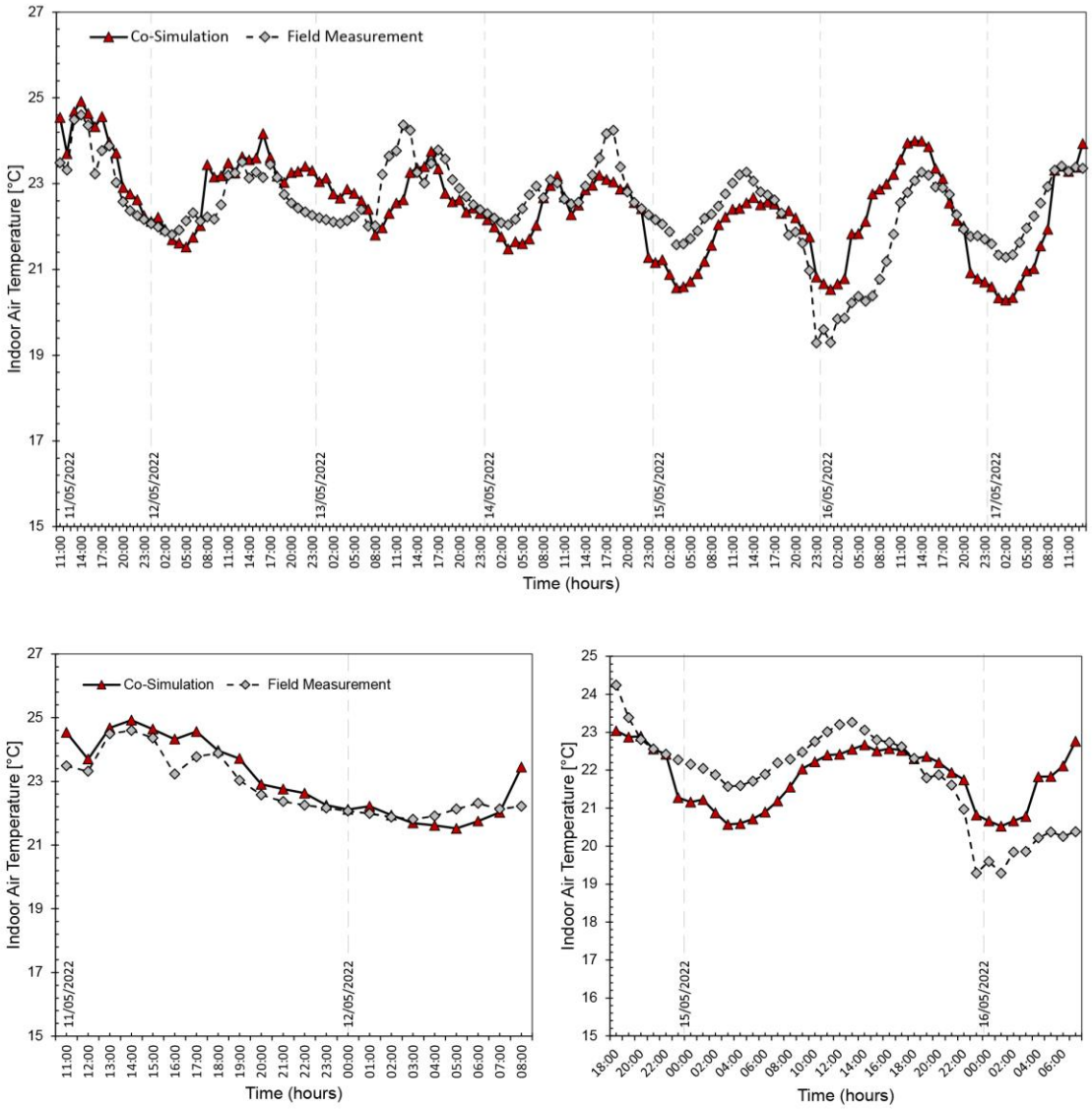


Figure 4.16: Simulated and observed indoor air temperature [°C]. Dates 11-17/05/2022, Location: North-facing office on the 9th Floor of the Arts Tower.

Chapter 5 Predictive Models: Metamodelling Roadmap

5.1 Introduction

Following the physics-based modelling and CONTAM-EnergyPlus co-simulation of the selected HEI buildings in Chapter 4, this Chapter presents a roadmap showing the key steps required for the development of machine learning (ML) based predictive metamodels. First, the chapter presents a study of identifying significant input variables through conducting sensitivity analysis in preparation for developing a metamodel. Next, a roadmap shows the key steps required for the development of the metamodels. The steps includes presenting the fundamentals and assumptions of the selected ML algorithms, selecting training and testing datasets, cross-validation techniques, hyperparameter optimisation (tuning), ML performance evaluation metrics, and interpretability of ML models. This chapter forms the procedural basis for generating and interpreting the results presented in Chapter 6.

5.2 Input and Output Data

5.2.1 Sensitivity Analysis

The sensitivity analyses are used to test the dependence of each output on the inputs. Using the methods reported in (Benjamin Jones et al., 2015) and (Das et al., 2014), we tested for *linear*, *monotonic*, and *non-monotonic* relationships between the inputs and the output. A linear relationship can be tested using: (i) Kendall's τ rank, (ii) Pearson's r product moment correlation coefficient, and (iii) linear regression. A monotonic relationship is tested with: (iv) Spearman's rank correlation coefficient and (v) the rank-transformed standardised variables, and the non-monotonic relationships can be tested using (vi) Kolmogorov-Smirnov and (vii) the Kruskal–Wallis quantile tests. The original tests were written in a Matlab code and is freely available

online[9] (Benjamin Jones, 2019). The code ranks the inputs by their significance, where the most important is ranked 1st. The code was replicated using Python, and follows the same ranking procedure.

Table 5.1: Summary of the sensitivity analyses applied in this chapter: Column (i) the category of the sensitivity analysis, (ii) the particular method within the category, (iii) the symbol, (iv) the type of correlation between input and output variables the analysis can detect, (v) the relevant outputs derived from applying the analysis, and (vi) the specific metric used for ranking the inputs (Based on (Das et al., 2014)

Category	Method	Symbol	Detection Ability	Relevant Output	Ranking Metric
Correlation	Pearson's Product Moment Correlation Coefficient	S_{Pear}	Linear	Coefficient between -1 (perfect negative linear correlation) and 1 (perfect positive linear correlation).	The magnitude of the correlation coefficient.
	Kendall's tau-b Correlation Coefficient	$S_{Kendall}$	Linear	Coefficient between 0 (no relationship) and 1 (perfect relationship).	The magnitude of the correlation coefficient.
	Spearman Correlation Coefficient	S_{Spear}	Monotonic	Coefficient between -1 (perfect negative monotonic correlation) and 1 (perfect positive monotonic correlation).	The magnitude of the correlation coefficient.
Regression	Linear Regression Coefficients	$S_{Regress}$	Linear	Coefficient between -1 (perfect negative linear correlation) and 1 (perfect positive linear correlation).	The magnitude of the regression coefficient.
	Rank-transformed standardised variables	$S_{RegRank}$	Monotonic	Coefficient between -1 (perfect negative monotonic correlation) and 1 (perfect positive monotonic correlation).	The magnitude of the regression coefficient.
Sample Comparison	Kolmogorov-Smirnov	S_{Kol}	Non-monotonic	Kolmogorov-Smirnov test statistic, ranging between 0 (no difference between samples) and 1 (maximum difference between samples).	Kolmogorov-Smirnov test statistic.
	Kruskal-Wallis	S_{KW}	Non-monotonic	Kruskal-Wallis test statistic, (value greater than 0), with higher values showing a greater difference between samples.	Kruskal-Wallis test statistic.

[9]Generic Global Sensitivity Analysis Code. http://dx.doi.org/10.13140/RG.2.2.21670.88644

All eighteen inputs and one output were retained for performing a sensitivity analysis initially, see Table 5.2. A fundamental assumption of the sensitivity analysis is that all the inputs tested are independent to avoid multicollinearity, therefore any correlated inputs should be eliminated. No transformation is applied to the input and output data, and all outliers were considered. The SA framework step checks the output and input averages over the heating season (Nov-Apr). The coefficients and p-values were calculated for each test, and the inputs were automatically ranked by the magnitudes and their significance based on the written Python code, and following (Benjamin Jones, 2019).

Table 5.2: Inputs retained, and output computed for the initial sensitivity analysis.

Inputs			**Outputs**
Zone Characteristics		**Indoor Environment**	
1. Floor Area (A, m^2)	9. Party Walls Gross Area (A_{pw}, m^2)	12. Indoor Temperature (Tin, °C)	Infiltrated $PM_{2.5}$ Concentrations (Ci, µg/xm^3)
2. Orientation	10. Building Airtightness @ 50Pa ($Q_{50,}$ $m^3/h/m^2$)	13. Infiltration ACH ($ACH_{INF,}$ h^{-1})	
3. Zone Height (H, m)	11. Air Permeability @ 4Pa ($Q_{4,}$ $m^3/h/m^2$)	14. Indoor Relative Humidity (RH, %)	
4. Number of Ex. Facades		15. Scaled Local Wind Speed (v, m/s)	
5. Area of Exposed Facades (A_{ef}, m^2)		16. Outdoor/Indoor Temperature Difference (ΔT, °C)	
6. Envelope: Volume Ratio (Aef:V)		17. Ventilation Rate Per Person (L/s/person)	
7. Envelope: Zone Area (Aef:A)		18. Total Ventilation Rate Per Zone (l/s)	
8. Total Permeable Area (L_{ef}, m^2)			

5.2.2 Multicollinearity (VIF)

The term multicollinearity refers to a condition in which the independent variables in a study exhibit a strong correlation. In situations where certain predictor variables are not independent, it would become difficult to determine or attribute the contributions of the various predictor variables to the response variable. Furthermore, when multicollinearity is left unaddressed, the variance of coefficient estimates can be increased, resulting in a broader range of confidence

intervals. As such, the interpretability of the model would become an issue. Consequently, obtaining statistically significant results from any subsequent analyses becomes more difficult.

Among the essential metrics for assessing multicollinearity is the variance inflation factor (VIF). A VIF directly measures the ratio between the variance of the entire model and the variance of a model containing only the variable in question, see Eq. (5.1). Simply put, it measures how much a feature contributes to the variance of the coefficients of the features included in the model. A VIF value of 1 indicates that the feature does not correlate with any other features. VIF values greater than 5 are considered high. Any feature with such VIF values (≥ 5) will likely contribute to multicollinearity (Table 5.3).

$$VIF = \frac{1}{1 - R^2} \tag{5.1}$$

Table 5.3: Variance Inflation Factor Analysis Threshold Values

VIF	Result
VIF ≤ 1	Not Correlated
1 < VIF > 5	Moderately Correlated
VIF ≥ 5	Highly Correlated

5.3 Algorithms Selection

The Generalised Additive Models (GAMs), Random Forest Regression (RFR), and Extreme Gradient Boosting (XGB) are three popular machine-learning algorithms used for predictive modelling in various applications. Based on the review of literature in Chapter 2, Section 2.7, these algorithms offer several benefits, such as the ability to handle complex and nonlinear data relationships, interactions, missing values, and outliers. Additionally, they provide insights into feature importance, making them useful for feature selection and variable importance analysis. However, the choice of the best algorithm for a particular problem depends on the nature of the data, analysis goals, and available resources for model development and implementation. Therefore, choosing GAMs, RFR, and XGB could provide a combination of flexibility, robustness, and predictive accuracy for the metamodel model development.

5.3.1 Generalised Additive Models (GAMs)

Generalised Additive Models (GAMs) have emerged as a leading class of models regarding prediction accuracy while also being simple enough for human users to understand and mentally simulate the underlying mechanism. GAMs are smooth semi-parametric models with a critical difference compared to Generalised Linear Models (GLMs) such as Linear Regression Models (LRMs), see Figure 5.1. In LRMs, the response variable (y) is defined by the sum of the linear combination of continuous variables (x). Each variable is given a weight,beta and added together to obtain a line that best fits the data, see Eq. (5.2).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \varepsilon_i \tag{5.2}$$

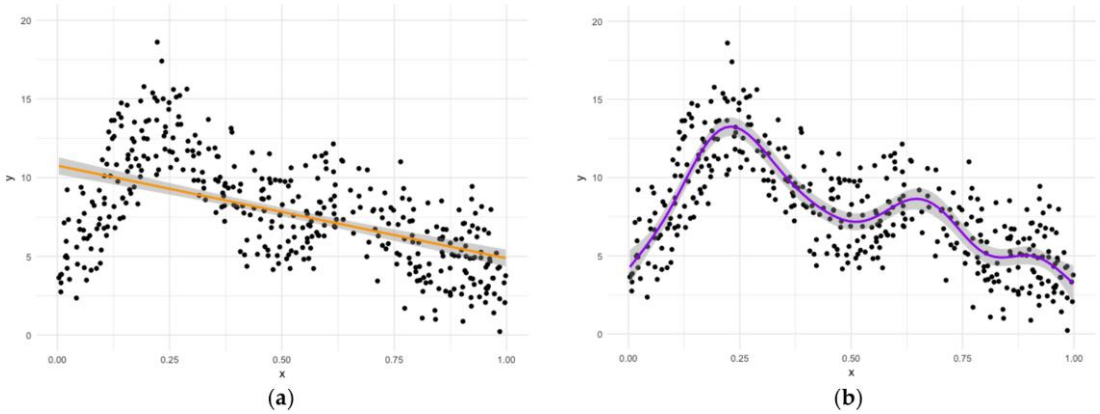


Figure 5.1: Scatter plot between two hypothetical variables x and y showing a nonlinear relationship in which (a) a linear model is fitted, (b) a GAM is fitted with splines (Y. Xu et al., 2021)

In GAMs, the assumption that y can be calculated using the linear combination of variables is dropped, allowing users to learn nonlinear features by replacing the term $\beta_i x_i$ with a flexible '*smooth function*', f(X_i) called a '*Spline*' (Wood, 2017), and the sum of multiple splines forms a GAM, see Eq.(5.3). Splines are real functions that are piecewise defined by polynomial functions (basis functions). The places where the polynomial pieces connect are called knots. The '*smooth function*', f(X_i) is composed of the sum of basis functions b and their corresponding regression coefficients β, see Eq. (4) and Figure 5.2.

$$y = \beta_0 + \mathrm{f}(X_1) + \cdots + \mathrm{f}(X_i) + \varepsilon_i \tag{5.3}$$

$$\mathrm{f}(x) = \sum_{i=1}^{q} b_i(x)\beta_i \tag{5.4}$$

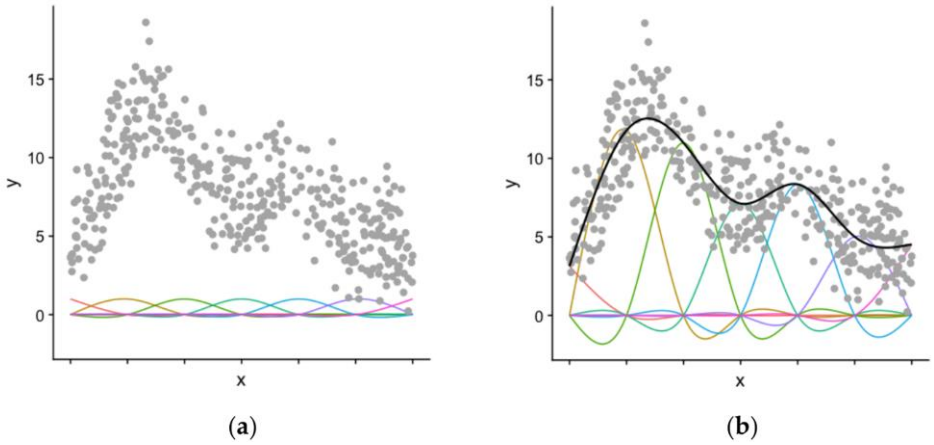


Figure 5.2: Incorporating the smoothing functions to GAM, (a) Basis functions with equal coefficients, (b) Basis functions multiplied by coefficients, each of which is a parameter in the model (Based on (Wood, 2017)).

By replacing the complicated parametric relationships in GLMs and LRMs with '*smooth functions*', it is possible to avoid cumbersome and unwieldy models. This flexibility and convenience, however, result in two new theoretical problems (Wood, 2017). The first is to select how a '*smooth function*' should be represented (*e.g.*, *cubic polynomial spline*) to give a curve that best fits the data, see Figure 5.3. The second is to determine the '*degree of smoothness*' by estimating the smoothing parameter (λ). The latter is defined as the '*wiggliness*' penalty used to penalise the basis coefficients for controlling the degree of smoothness and ensuring that knots are well spread.

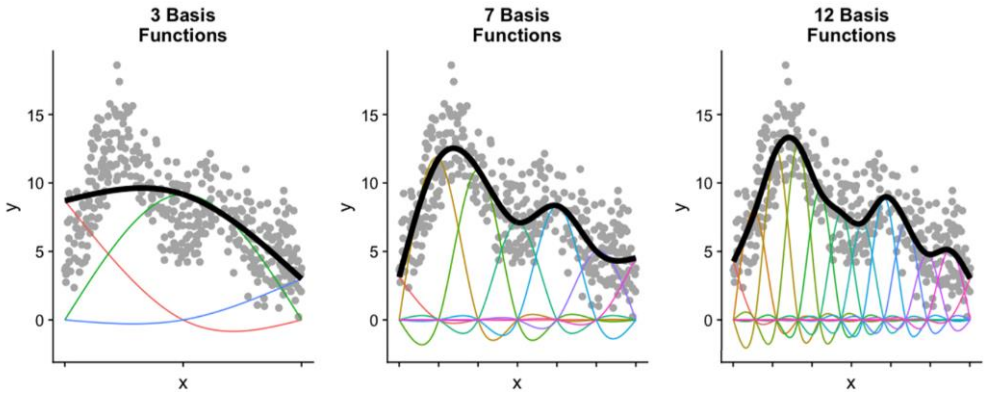


Figure 5.3: The effect of the number of basis functions on the shape of the line of best fit.

When λ is too high, the spline fits many of the data points poorly and does no better with the missing point. When λ is too low, the spline fits the noise as well as the signal, and the extra variability that this induces causes it to predict the missing datum rather poorly. For the intermediate λ, the spline fits the underlying signal quite well but smoothes through the noise. As a result, the missing data is reasonably well predicted. As such, by changing the value of λ, various models of different smoothness can be obtained. Figure 5.4 illustrates this but begs the question, which value of λ is 'best' or 'optimal'? This can be resolved through the Generalised Cross Validation (GCV) method to estimate model hyperparameters (Bottegal & Pillonetto, 2018); this is discussed in detail in Section 5.4.2

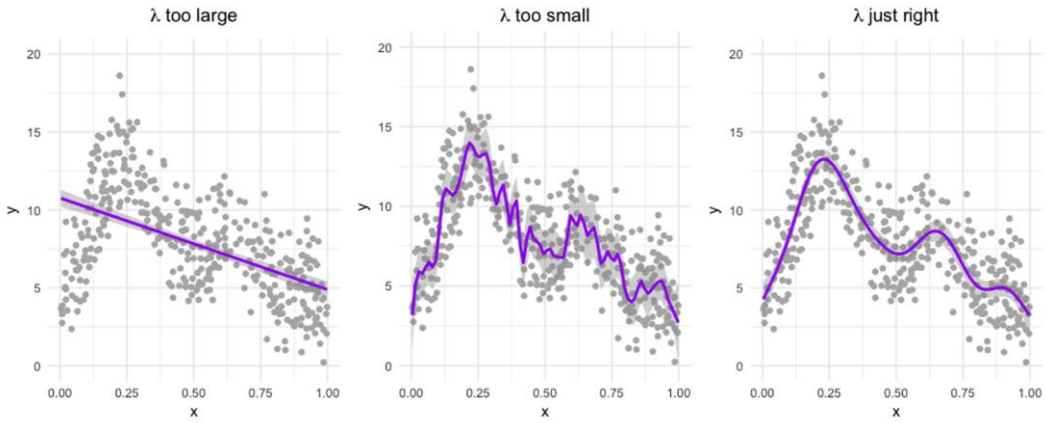


Figure 5.4: Penalised regression spline fits the response variable y vs the explanatory variable x using three values for the smoothing parameter, λ.

The 'pyGAM' 0.8.0 package (Servén & Brummitt, 2018) for building GAMs in 'Python 3.10.5 (Rossum & Drake, 2022b)' was applied here using the `gam.fit()` function to fit a GAM to the simulated monthly $PM_{2.5}$ concentration levels ($\mu g/m^3$) for the building's zones dataset.

5.3.2 Random Forest Regression (RFR)

A Random Forest Regression (RFR) is a supervised ML technique that is used to solve regression problems where nonlinear relationships between input features and the response variable (y) exist. This technique uses ensemble learning to solve complex problems by combining a group of decision trees (Brieman, 2001). RFR models use multiple Decision Trees (DTs) and a technique called *Bootstrap* and *Aggregation*, commonly known as '*Bagging*'. The idea is to combine multiple decision trees in determining the final prediction value rather than relying on individual DTs (Giussani, 2021), see Figure 5.5. In regression problems, the predicted response variable (y)

is the average prediction value across the DTs. As a result of the averaging, RFs are better than single DTs in prediction accuracy, and overfitting can be reduced.

Additionally, compared to DT, the RFR algorithm searches for the best feature from a random subset of features, hence its name 'Random Forest'. This adds extra randomness to trees growing in a random forest. As a result of '*Feature Randomness*', random forest decision trees are uncorrelated, see Figure 5.6. The correlation between the trees is the key. This helps the trees to protect each other from their individual errors, providing they don't consistently error in the same direction. Even though some trees may be wrong, many other trees will be correct, so as a group, the trees can move along in the correct direction.

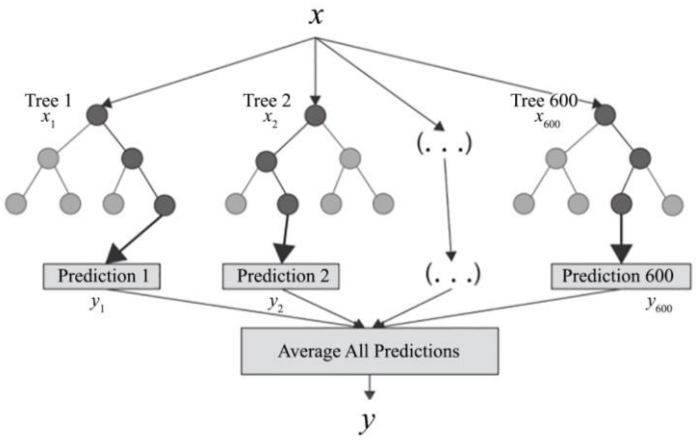


Figure 5.5: Architecture of the Random Forest Regressor Model showing the constructed decision trees and the classes' average as the predicted value of all trees.

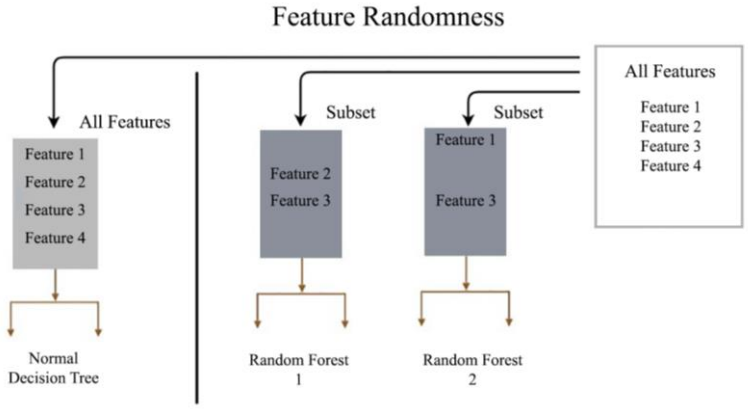


Figure 5.6: An individual tree's node splitting is determined by a random subset of features (Feature Randomness)

A potential downside of the RFR is extrapolation. While RFs have greater genericity than LRMs and can be applied to complex nonlinear problems, they can lead to nonsensical predictions if extrapolation domains are used. When using RFR, the predicted values are never outside the training set values for the response variable. Since the RFR always predicts the average of the values seen previously, the average of a given sample can never go beyond the highest and lowest values of the sample. This is due to the RFR's inability to detect trends that would enable it to extrapolate values outside the training set. In such a scenario, the regressor assumes that the prediction will be close to the maximum value of the training set, see Figure 5.7.

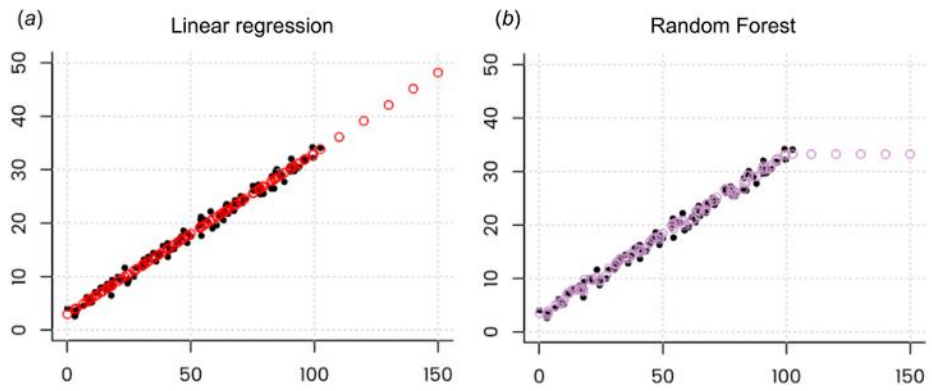


Figure 5.7: Illustration of the extrapolation problem of Random Forest (Hengl et al., 2018).

Despite this behaviour, the RFR will be applied to the simulated monthly $PM_{2.5}$ concentration levels ($\mu g/m^3$) and I/O $PM_{2.5}$ ratios for the building's zones dataset and then compared to the fitted GAM model. The scikit-learn Python ML library version 1.1.1 (Pedregosa et al., 2011) provides an implementation of RFR. Additionally, similar to all ML models, hyperparameter tuning is essential to control the behaviour of the fitted model. The RFRs hyperparameter tuning will be discussed in detail in Section 5.4.2.

5.3.3 Extreme Gradient-Boosted Decision Trees (XGB)

XGB is an ensemble tree-based model that follows the gradient boosting framework principle (T. Chen & Guestrin, 2016). It is used for supervised ML problems, where a dataset with multiple features x is used to predict a response variable y. XGB is an *iterative* decision trees algorithm with multiple decision trees. Although similar to RF in its architecture, XGB depends on a *boosting* technique, not *bagging*. It means that the algorithm tries to improve the error from previous trees. Rather than training all of the models independently from one another, *boosting* trains models in a sequential form, with each new model being trained to correct the errors made

by the previous ones, see Figure 5.8. Models are added sequentially until no further improvements can be made. The predicted value of the XGB model is the sum of all predicted values across the decision trees, see Eq. (5.5).

$$y_i = \sum_{k=1}^{n} f_k(x_i) \qquad f_k \in F \tag{5.5}$$

where F means the space of regression trees, f_k corresponds to a tree, so $f_k(x_i)$ is the result of tree k, and y_i is the predicted value of the ith instance x_i.

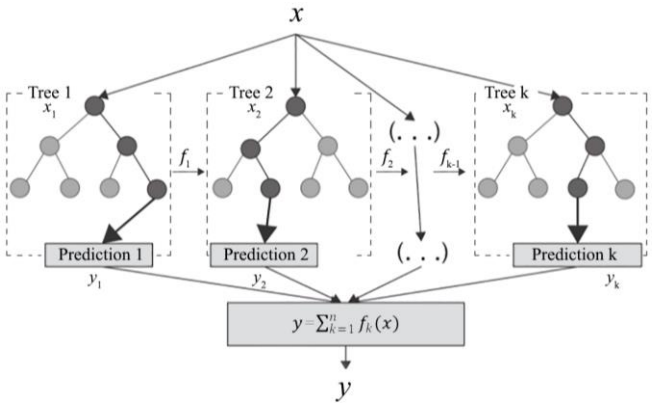


Figure 5.8: Architecture of the XG-Boost Model showing the constructed decision trees by imposing regularisation and providing parallel tree boosting.

Although XGB can be computationally efficient, i.e., fast to execute, and highly effective, perhaps even more so than any other open-source implementation, hyperparameter tuning can be challenging as it usually leads to extensive grid search experiments, which is discussed in Section 5.4.2**.** Table 5.5 summarises the three ML algorithms (models) as discussed above.

Table 5.4: Comparison of the ML models selected for this study

Algorithm Name	Family	Type of Response Variable y	Linearity of the Model	Strengths	Weaknesses
***Linear Regression Model (LRM)**	Regression Models	Continuous	Linear	1) Determines the best predictor of the variable of interest 2) The simplicity of the model to be used for predictions 3) Detects Outliers	1) It does not deal with nonlinear problems if data is not transformed linearly. 2) Requires more observations than variables 3) Multi-collinearity 4) Outliers can seriously bias the regression coefficients.
Generalised Additive Models (GAM)	Regression Models	Continuous and Categorical	Linear and Nonlinear	1) Ability to model highly complex non-monotonic and nonlinear relationships. 2) Able to deal with categorical predictors 3) High Interpretability and confidence intervals 4) Feature selection with p-values. 5) Fast cross-validation via GCV. 6) Controlled extrapolation.	1) Computational Complexity when working with hyperparameters. 2) Prone to overfitting if the sample size is small.
Random Forest Regressor (RFR)	Decision Trees	Continuous and Categorical	Linear and Nonlinear	1) High prediction accuracy by reducing the variance in predictions. 2) Reduces overfitting in comparison to single decision trees. 3) Deals with missing data automatically 4) Data normalisation is not required 5) Works well with continuous and categorical data	1) Less Interpretable compared to GAMs and LRM and fails to determine the significance of each variable. 2) Multiple hyperparameters to tune. 3) Requires high computation power to build numerous trees. 4) Not efficient in extrapolation
Extreme Gradient Boost (XGB)	Decision Trees	Continuous and Categorical	Linear and Nonlinear	1) Uses the power of parallel processing 2) High prediction accuracy 3) Supports regularisation to prevent overfitting 4) Deals with missing data automatically 5) Allows for cross-validation 6) Effective Tree pruning to prevent negative loss in the split.	1) Less Interpretable compared to GAMs and LRM and fails to determine the significance of each variable. 2) Challenging hyperparameter tuning.

*Linear Regression was not fitted to the project's datasets; however, it was included in this comparative table for comparison only.

5.4 The Learning Roadmap

5.4.1 Cross-Validation

Cross-validation (CV) is a statistical approach to estimating ML model performance. The purpose of this method is to evaluate the likelihood that the results of a predictive ML model will generalise to a new and unseen sample of data. The dataset must be resampled into training and testing/validation datasets to perform a CV. A comparison can be made between the training and testing/validation datasets to measure the difference in results. This ensures that overfitting or selection bias is flagged with the training datasets. There is a range of different CV techniques utilised in ML; Monte Carlo CV (MCCV), Leave-One-Out CV (LOOCV), *k*-fold CV (L. Xu et al., 2018), etc.; *k*-fold CV is a non-exhaustive method that is widely used in applied ML and is therefore selected for CV in this work.

In *k*-fold CV, the original dataset is separated into a test/hold-out set for the ML model's final evaluation. This typically constitutes about 30% of the original dataset as test data, and the remaining 70% is divided into *k* folds (training subsets). During CV, each iteration uses one of the *k* folds as the validation set while the remaining folds act as the training set. The process is repeated until every fold has been used as a validation set. Figure 5.9 shows what this process looks like for a 5-fold CV.

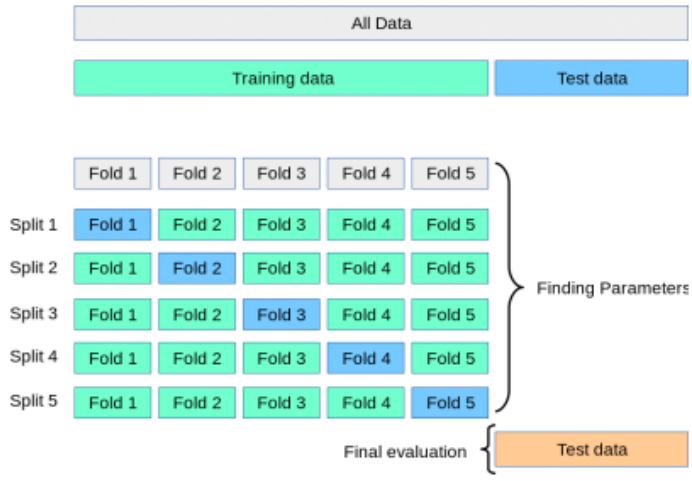


Figure 5.9: *k*-Fold Cross Validation for a *k*=5

It is better to determine how well the ML model might perform on data it has never seen before by training and testing it *k* times on different subsets of the same training data. For example, a *k*-fold CV can show how well the ML model performs compared to one-time trials by scoring the model after every iteration and calculating the average of all scores.

5.4.2 Hyperparameter Optimisation

Hyperparameters are defined as the settings of an ML algorithm that can be adjusted to optimise its performance. As opposed to model parameters, which are learned during model training and cannot be changed arbitrarily, hyperparameters can be altered by the user before the model is trained. To determine the ideal settings for hyperparameters, it is best to try various combinations and evaluate each model's performance. It is, however, essential to keep in mind that evaluating each model only on the training set can lead to overfitting, which is one of the most fundamental problems in ML. Table 5.5 lists the hyperparameters used in GAM, RFR, and XGB.

Table 5.5: List of hyperparameters selected to tune for each ML algorithm (GAM, RFR, and XGB)

Generalised Additive Model (GAM)	Random Forest Regressor (RFR)	Extreme Gradient Boost Regressor (XGB)
Smoothing parameter (λ)	The No. of Decision Trees in the forest (n_estimators)	The No. of Decision Trees in the forest (n_estimators)
	The maximum depth of the individual trees (max_depth)	The maximum depth of the individual trees (max_depth)
	The minimum samples to split on at an internal node (min_samples_split)	The Learning Rate (learning_rate)
	Minimum number of leaf nodes (min_samples_leaf)	Fraction of Columns to be randomly sampled per tree (colsample_bytree)
	Number of random features *(max_features)*	Fraction of Observations to be sampled per tree (subsample)

When applied to ML problems, hyperparameters that perform well on one may perform poorly on the others. Therefore, it is recommended to select systematic methods to tune hyperparameters for each ML model. This can be done using a Random Search CV (RSCV) in Scikit-Learn for the RFR and XGB Models and the Generalised CV (GCV) in pyGAM for the GAM modelling. Both methods are based on defining a grid of hyperparameter ranges and randomly sampling from this grid during fitting, then performing a *k*-fold CV with each combination of values. Even though

the R^2 value for the models can sometimes be affected (reduced), it is expected that performing both RSCV and GCV can improve the generalisation of results and reduce overfitting.

5.4.3 Performance Evaluation Metrics for Regressions

It is crucial to have accurate predictive models since they determine the quality of predictions that serve as scientific evidence for policy and decision-making. Many regression models rely on distance metrics to determine the convergence to the best result. Several metrics can be used to determine the overall accuracy of a predictive model. Prediction accuracy is defined as a model's ability to minimise the overall error between actual values and predicted values see Eq. (5.6). In simple words, the error can be defined as any deviation from the actual value.

$$error = y\ (actual) - \hat{y}\ (predicted) \tag{5.6}$$

Using this error, it is possible to derive many different metrics that can provide more insights into a model's prediction accuracy. The most commonly known performance evaluation metrics include the Mean Square Error (MSE), the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Coefficient of Determination. (R^2). The R^2 value represents how much the model can explain variation in the dependent variable. The R^2 is calculated by dividing the sum of squared prediction error by the total sum of squares that replace the calculated prediction with the mean. R^2 is a value between 0 and 1, with a higher value indicating a better fit between prediction and actual value. R^2 is a good measure of how well a model fits the dependent variables; however, it does not take overfitting into account.

As opposed to R^2, MSE is an absolute measure of the goodness of the fit. In MSE, the sum of the squares of prediction error is divided by the number of data points. The prediction error equals the difference between the actual output and the predicted output. It gives an absolute measure of how far the predictions are off. A single result does not give many insights, but it gives an actual number to compare with other model results and helps select the best regression model. The RMSE is the square root of MSE. Because MSE values can sometimes be too large to compare easily, RMSE is used more commonly than MSE. Additionally, the square root makes it easy to interpret since it is brought back to the same level of prediction error as its square root.

Finally, MAE is the average of all absolute differences between actual and predicted values, i.e., the average absolute values of errors where all individual errors have equal weight. MAE can range from 0 to ∞ and produces an error in units of the variable of interest, making it easy to interpret. It is essential to point out that both MAE and RMSE are negatively oriented scores, meaning lower values are better.

5.4.4 Interpretability of ML Models

For many applications, understanding why a model makes a particular prediction can be just as important as the prediction's accuracy. Often, the most accurate models for large datasets are complex models which even experts have trouble interpreting why the model outputs as they are, such as ensemble models or deep learning. By having an interpretable model, one can understand what the model is learning, what other information it possesses, how it makes decisions, and how it justifies those decisions in the context of the real-world problem under study.

The interpretability of an ML model refers to its ability to associate a cause with an effect. However, many ML models face this challenge due to the lack of a consistent approach or metric for measuring the importance of predictor variables and quantifying their contribution (Gu et al., 2021). For example, in a GAM model, the predictor variables selected in the final prediction model are generally regarded as necessary, and their coefficients of the partial determination indicate the percentage of explained variability. While in the RFR and XGB models, the importance of a predictor variable is measured by the permutation importance method. This method measures whether the model's score increases or decreases when the predictor variable is randomly shuffled. In other words, those models assess the importance of predictor variables differently, and each will only reflect the model's predictive power. As a result, there is no quantification of the fractional contribution of each predictor variable to the predicted outcome, and the model comparisons are inconsistent, resulting in an inability to interpret the results.

SHapley Additive exPlanations (SHAP) is used as a unifying framework for interpreting and comparing ML models to quantify the marginal contributions of each predictor variable in a model (Christoph, 2020). Based on a game theoretic approach, it computes the contribution of each predictor variable to the prediction in terms of Shapley values from coalitional game theory, which can explain the prediction from any ML model (Lundberg & Lee, 2017). Shapley value can determine the contribution of each predictor variable to a prediction by estimating the average

marginal contribution across all possible coalitions. The benefits of using SHAP values can be summed up as follows:

(1) Global Interpretation: In addition to showing the importance of a particular feature, SHAP values also indicate whether the feature is impacting predictions positively or negatively.
(2) Local interpretability can be achieved by calculating SHAP values for every prediction and determining how the features contribute to the prediction. In contrast, other techniques only display aggregated results for the entire dataset.
(3) It is possible to explain a wide range of models with SHAP values, including linear models (e.g., linear regression and GAM), tree-based models (e.g., RFR and XGB) and neural networks. In contrast, other techniques can only explain a limited number of models.

Several versions of SHAP are available to accommodate different models' architectures regarding computation time. In this study, the Shapley values were computed using scikit-learn 0.23.1 and the SHAP library in Python.

5.4.5 ML Models Evaluation

After developing the ML models based on a reduced set of input parameters, it is imperative to determine the adequacy of the prediction in comparison to the simulated results. Performance metrics described earlier in Section 5.4.3 will be used to assess the performance of GAM, RFR, and XGB models against simulated monthly indoor $PM_{2.5}$ concentration levels and aggregated indoor $PM_{2.5}$ concentration levels over the heating season. To ensure the generalisability of the predictive models, further evaluation will be conducted using new unseen data that was not used for training, validating, or testing the models. This data results from modelling the ICoSS building in E+ and CONTAM, which has not been used in any of the training and testing developments.

Chapter 6 Sensitivity Analysis and Model Predictions

6.1 Introduction

This chapter demonstrates the implementation of the ML roadmap described in Chapter 5. The chapter begins by presenting the results of the Sensitivity Analyses (SA) framework applied to the heating season concentrations of infiltrated $PM_{2.5}$. Then, an in-depth analysis of all the variables in the dataset will be conducted to identify the essential variables associated with the heating season concentrations of infiltrated $PM_{2.5}$. Several statistical analyses will also quantify the correlation between the dependent and independent variables. As a result of the strength of the relationship, each variable will be assigned a rank. Finally, the outcome of the SA will be used as input to each ML algorithm (i.e., GAM, RFR, and XGB).

Successful implementations of most ML models require the specification of optimal training methods, the use of a sufficient amount of training and testing data, and the utilisation of a broad range of computational resources. For this purpose, each ML algorithm is fitted using a set of default hyperparameter values. Then, a 3-fold cross-validation (CV) technique is used to determine each algorithm's optimal set of hyperparameters. There is a good chance that a well-fitted model will not match the available data perfectly, and predictions are bound to contain some errors (prediction errors). Nevertheless, it should be able to predict the outcome (e.g., heating season concentrations of infiltrated $PM_{2.5}$) fairly accurately. It should also consider the dataset's overall shape to ensure that the interpolated predictions are reasonably accurate. Thus, a 10-fold CV technique was used to evaluate each model by calculating the Root Mean Square Error (RMSE) across each fold separately. Then, the average RMSE across the 10-folds was used as a benchmark value to compare the performance of the tuned models. As mentioned in Section 5.4.1, the GAM method provides an internal Generalised Cross Validation (GCV), while the RFR and the XGB require a standalone k-fold CV. Lastly, this chapter concludes by assessing the

performance of each model on an unseen dataset. Then, based on the model performance metrics mentioned in Section 5.4.3, the prediction error and model accuracy were evaluated and compared to the simulated datasets.

6.2 Results of the Sensitivity Analyses Framework

The results of these analyses enable the identification of the most influential input variables to develop a predictive metamodel so that they can be targeted for attention when designing new buildings or for future field measurement campaigns in the case of epistemic uncertainties in influential model inputs. Scatter plots in Figures 6.1 and 6.2 are utilised to visually demonstrate the correlation between the modelled heating season concentrations of infiltrated $PM_{2.5}$ (C_i) and each of the eighteen input variables (see Table 5.2). These plots are based on the analysis of the data for 2,729 zones. A positive non-linear relationship can be seen between the heating season concentration of infiltrated $PM_{2.5}$ C_i and the infiltration ACH_{INF}, the zone air permeability rate Q_4, the building airtightness Q_{50}, and the ventilation rates per person. There is a strong negative non-linear relationship between C_i and the indoor/outdoor temperature difference ΔT which seems to be more scattered for zones with higher Q_{50}. Furthermore, a moderate negative non-linear relationship can be seen between C_i and the scaled wind speed v. There is no clear relationship between C_i and the zone area (A_z), zone volume (V_z), the number of exposed facades (N_{ef}), relative humidity RH, and zone orientation φ. Finally, the zone effective leakage area (L_{ef}) and the zone exposed area-to-zone volume (A_{ef}:V_z) show an exponential positive relationship with C_i.

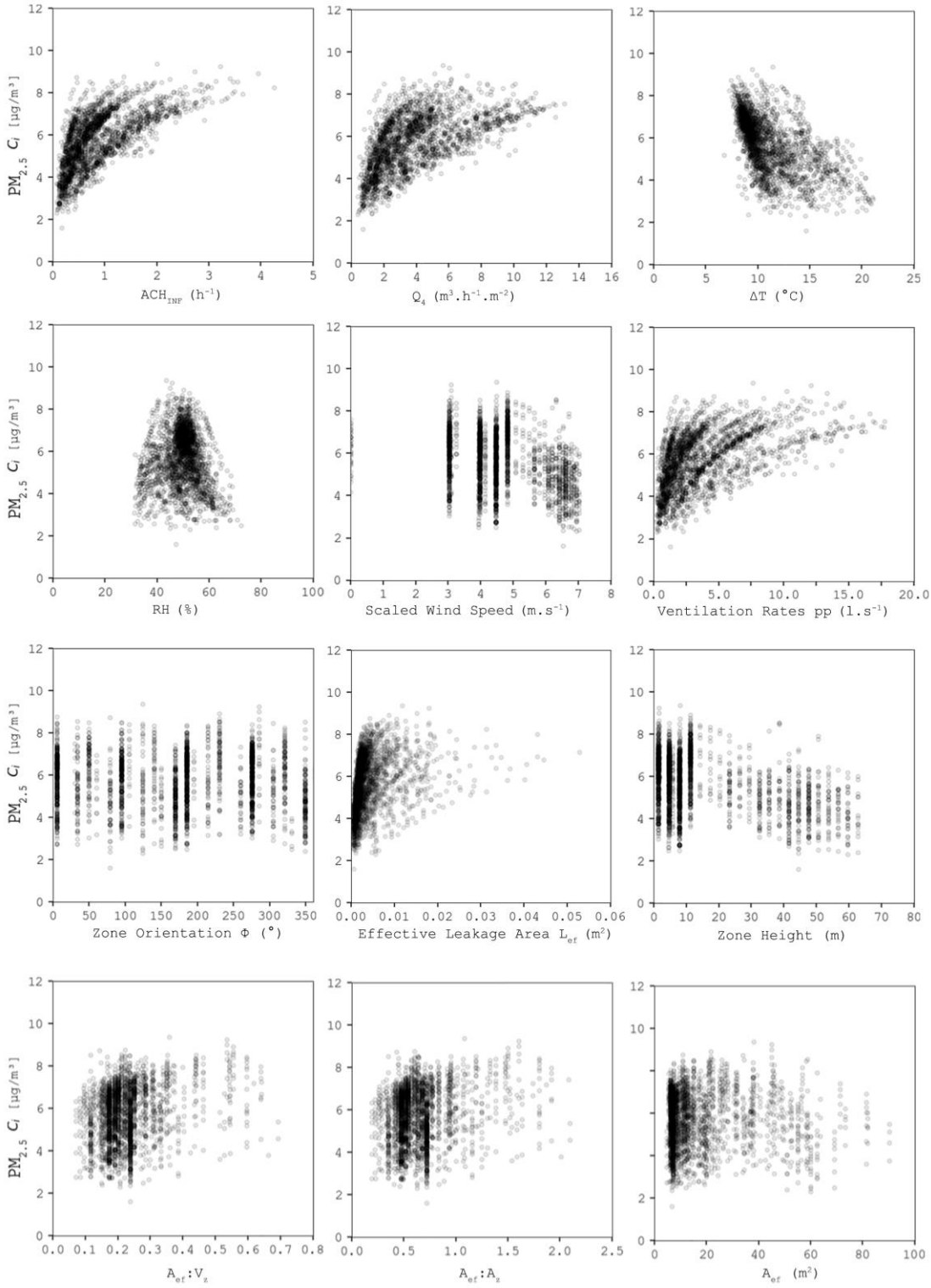


Figure 6.1: Scatter Plots of Each Input versus the *Heating Season* Concentrations of Infiltrated $PM_{2.5}$

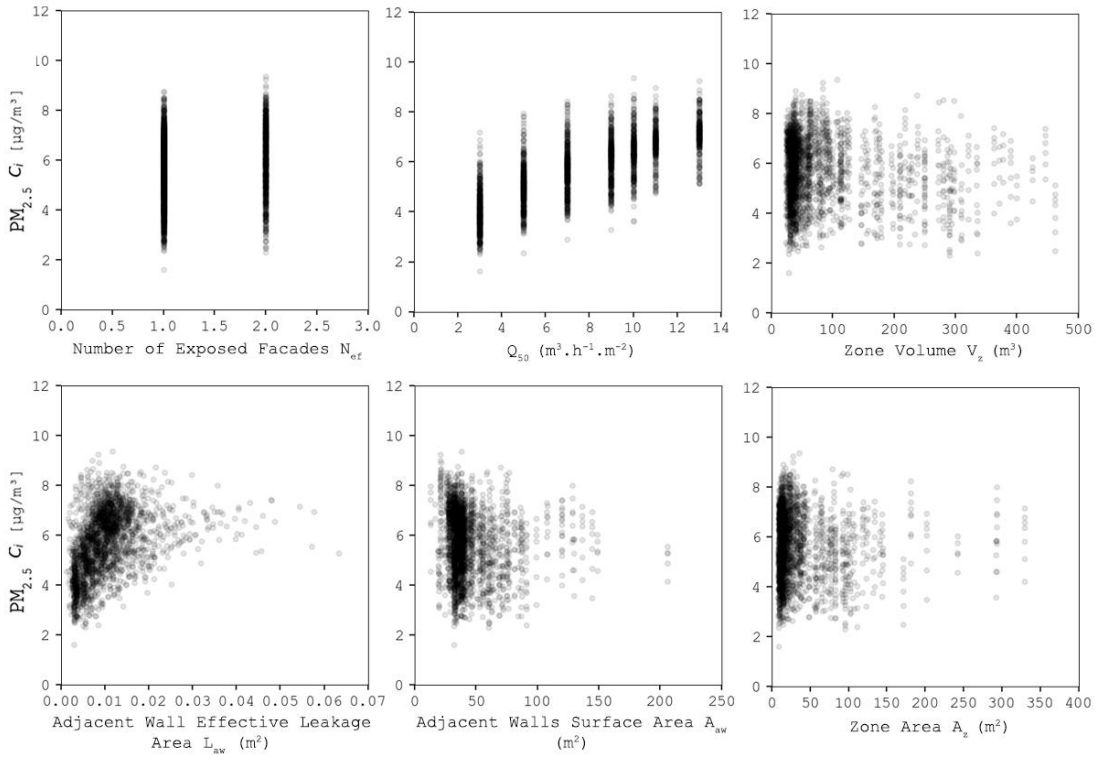


Figure 6.2: Continued Scatter Plots of Each Input versus the *Heating Season* Concentrations of Infiltrated $PM_{2.5}$

Tables 6.1, 6.2, and 6.3 present the results of various sensitivity analyses for the concentrations of infiltrated $PM_{2.5}$ during the heating season. These tables display the ranking of input variables according to their level of importance, enabling the selection of the most relevant input variables in a prioritised order. The tables manifest that the Infiltration ACH_{INF}, ΔT, v, Q_4, zone height H and the ventilation rate per person V are among the top six variables of significant importance to C_i in all tests, albeit with some variability in their respective rankings. Correlation tests (S_{Pears} and S_{Spear}) and regression tests ($S_{Regress}$ and $S_{Regress\text{-}Rank}$) rank Q_{50} as a variable with no importance. However, group comparison tests (S_{Kol}, S_{KW2}, and S_{KW5}) and the correlation test $S_{Kendall}$ rank Q_{50} as the variable with the most importance (rank = 1). This discrepancy between tests is attributed to zones within the same building sharing the same Q_{50}, thus considered a categorical variable. However, it is evident in Figure 6.3 that selecting a low Q_{50} value (e.g. Q_{50} = 3 $m^3/h/m^2$) reflecting a tight building envelope contributes to a low C_i, and vice versa (when Q_{50} > 5 $m^3/h/m^2$). As such, Q_{50} is selected as the variable with the highest impact on the concentrations of infiltrated $PM_{2.5}$ during the heating season.

As one delves further into the rankings, there appears to be considerably less uniformity in the rankings. This observation can be elucidated by scrutinising the *p*-values associated with the computed correlation coefficients (such as S_{Pears}, S_{Spear}, S_{Kol}, and S_{KW}). The *p*-values are less than 0.05 for the six most critical variables and comparatively higher for the rest. Consequently, the remaining input variables do not exhibit significant influence at the widely used 5% significance level, and it would be illogical to compare them. Hence, the coefficients and *p*-values may be utilised to construct a reduced set of the most crucial input variables.

Table 6.1: Test Statistics for Correlation applied to *Heating Season* Concentrations of *Infiltrated* $PM_{2.5}$ (1 is the highest rank), with the value of relevant output and sig. *p*-value.

Input	$S_{Kendall}$	*p*-value	Rank	S_{Pears}	*p*-value	Rank	S_{Spear}	*p*-value	Rank
Zone Area (A_z)	-0.010	0.426	17	-0.060	0.005	13	-0.016	0.419	17
Zone Volume (V_z)	-0.028	0.041	16	-0.063	0.002	12	-0.041	0.042	14
Zone Height (H_z)	-0.201	0.000	6	-0.390	0.000	5	-0.284	0.000	5
Zone Orientation (φ)	-0.010	0.500	18	-0.035	0.087	16	-0.014	0.490	18
Number of Exposed Façade (N_{ef})	0.114	0.000	12	0.145	0.000	9	0.140	0.000	11
Area of Exposed Façade (A_{ef})	0.174	0.208	9	0.009	0.645	18	0.026	0.194	16
Effective Leakage Area (L_{ef})	-0.042	0.002	15	-0.029	0.168	17	-0.063	0.002	12
Area of Adjacent Walls (A_{aw})	-0.118	0.000	11	-0.105	0.000	11	-0.172	0.000	9
Effective Leakage Area (L_{aw})	-0.192	0.000	8	-0.140	0.000	10	-0.282	0.000	7
Area of Exposed Façade to Volume Ratio (A_{ef}:V_z)	0.136	0.000	10	0.239	0.000	7	0.199	0.000	8
Area of Exposed Façade to Area Ratio (A_{ef}:A_z)	0.105	0.000	13	0.190	0.000	8	0.152	0.000	10
Zone Indoor/Outdoor Temperature Difference (ΔT)	-0.456	0.000	3	-0.569	0.000	3	-0.634	0.000	3
Zone Outdoor Wind Speed (v_z)	-0.201	0.000	7	-0.355	0.000	6	-0.284	0.000	6
Relative Humidity (RH)	-0.044	0.001	14	-0.050	0.014	15	-0.061	0.003	13
Infiltration ACH_{INF}	0.499	0.000	2	0.644	0.000	1	0.683	0.000	1
Building Airtightness Q_{50}	0.588	0.000	1	-0.056	0.006	14	-0.041	0.043	15
Zone Air Permeability Rate Q_4	0.450	0.000	4	0.577	0.000	2	0.634	0.000	2
Ventilation Rate / Person	0.397	0.000	5	0.516	0.000	4	0.563	0.000	4

Table 6.2: Test Statistics for Regression applied to *Heating Season* Concentrations of *Infiltrated* $PM_{2.5}$ (1 is the highest rank), with the value of relevant output and sig. *p*-value.

Input	$S_{Regress}$	*p*-value	Rank	$S_{Regress\text{-}Rank}$	*p*-value	Rank
Zone Area (A_z)	0.004	0.005	13	0.000	0.419	17
Zone Volume (V_z)	0.004	0.002	12	0.002	0.042	14
Zone Height (H_z)	0.152	0.000	5	0.081	0.000	5
Zone Orientation (φ)	0.001	0.087	16	0.000	0.49	18
Number of Exposed Façade (N_{ef})	0.021	0.000	9	0.020	0.000	11
Area of Exposed Façade (A_{ef})	0.000	0.645	18	0.001	0.194	16
Effective Leakage Area (L_{ef})	0.001	0.168	17	0.004	0.002	12
Area of Adjacent Walls (A_{aw})	0.011	0.000	11	0.030	0.000	9
Effective Leakage Area (L_{aw})	0.020	0.000	10	0.080	0.000	7
Area of Exposed Façade to Volume Ratio (A_{ef}:V_z)	0.057	0.000	7	0.040	0.000	8
Area of Exposed Façade to Area Ratio (A_{ef}:A_z)	0.036	0.000	8	0.023	0.000	10
Zone Indoor/Outdoor Temperature Difference (ΔT)	0.324	0.000	3	0.402	0.000	3
Zone Outdoor Wind Speed (v_z)	0.126	0.000	6	0.081	0.000	6
Relative Humidity (RH)	0.003	0.014	15	0.004	0.003	13
Infiltration ACH_{INF}	0.415	0.000	1	0.466	0.000	1
Building Airtightness Q_{50}	0.003	0.006	14	0.002	0.043	15
Zone Air Permeability Rate Q_4	0.333	0.000	2	0.402	0.000	2
Ventilation Rate / Person	0.266	0.000	4	0.317	0.000	4

Table 6.3: Test Statistics for Group Comparison applied to *Heating Season* Concentrations of *Infiltrated* $PM_{2.5}$ (1 is the highest rank), with the value of relevant output and sig. *p*-value.

Input	S_{Kol}	*p*-value	Rank	S_{KW2}	*p*-value	Rank	S_{KW5}	*p*-value	Rank
Zone Area (A_z)	0.087	0.000	16	0.634	0.426	16	11.12	0.025	17
Zone Volume (V_z)	0.092	0.000	14	3.77	0.052	14	16.23	0.008	16
Zone Height (H_z)	0.315	0.000	6	209.53	0.000	7	229.37	0.000	6
Zone Orientation (φ)	0.088	0.000	15	2.42	0.119	15	5.43	0.000	18
Number of Exposed Façade (N_{ef})	0.085	0.000	17	22.93	0.000	11	76.12	0.000	12
Area of Exposed Façade (A_{ef})	0.050	0.096	18	0.008	0.926	18	33.70	0.000	14
Effective Leakage Area (L_{ef})	0.110	0.000	13	17.62	0.000	12	48.68	0.000	13
Area of Adjacent Walls (A_{aw})	0.138	0.000	11	37.42	0.000	10	95.51	0.000	11
Effective Leakage Area (L_{aw})	0.242	0.000	8	138.87	0.000	8	191.31	0.000	8
Area of Exposed Façade to Volume Ratio (A_{ef}:V_z)	0.169	0.000	10	38.70	0.000	9	149.40	0.000	9
Area of Exposed Façade to Area Ratio (A_{ef}:A_z)	0.125	0.000	12	15.48	0.000	13	110.94	0.000	10
Zone Indoor/Outdoor Temperature Difference (ΔT)	0.518	0.000	2	818.07	0.000	2	949.38	0.000	3
Zone Outdoor Wind Speed (v_z)	0.315	0.000	7	209.55	0.000	6	229.36	0.000	7
Relative Humidity (RH)	0.202	0.000	9	0.519	0.471	17	18.77	0.000	15
Infiltration ACH_{INF}	0.495	0.000	3	742.45	0.000	3	1064.9	0.000	2
Building Airtightness Q_{50}	0.584	0.000	1	1005.3	0.000	1	1320	0.000	1

Zone Air Permeability Rate Q_4	0.478	0.000	4	681.60	0.000	4	943	0.000	4
Ventilation Rate / Person	0.375	0.000	5	491.62	0.000	5	713.85	0.000	5

The assessment of multicollinearity is imperative in the development of metamodels, as the results of the multicollinearity test provide valuable information about the degree of intercorrelation among the input variables. Based on the sensitivity analyses above, Q_{50}, Q_4, and *V* are ranked among the first five most important variables and were tested for multicollinearity using the Variable Inflation Factor (VIF) method (see Chapter 5). Table 6.4 shows the results of the correlation and regression tests. Generally, a pressure differential of 4 Pa is conventionally considered representative of natural ventilation (including infiltration), and it is desirable to ascertain the leakage Q_4 under such conditions. Nevertheless, evaluating the leakage at such low pressures is perceived to be prone to substantial errors resulting from the wind and buoyancy-induced pressures generated during the test. In the UK, Q_{50} is commonly used in building regulations and standards to represent buildings' airtightness (Gillott et al., 2016). Figure 6.3 (Right) illustrates a high correlation between Q_4 and $Q_{50,}$ and therefore Q_{50} is selected while eliminating Q_4 and *V* from the selection.

Table 6.4: Testing for Multicollinearity using Correlation and Regression tests between Q_{50}, Q_4, and *V*

Input	S_{Pears}	*p*-value	S_{Spear}	*p*-value	$S_{Regress}$	*p*-value	$S_{Regress\text{-}Rank}$	*p*-value
Q_{50} and Q_4	0.566	0.000	0.598	0.000	0.320	0.000	0.358	0.000
Q_{50} and *V*	0.698	0.000	0.735	0.000	0.487	0.000	0.540	0.000
Q4 and *V*	0.380	0.000	0.412	0.000	0.144	0.000	0.170	0.000

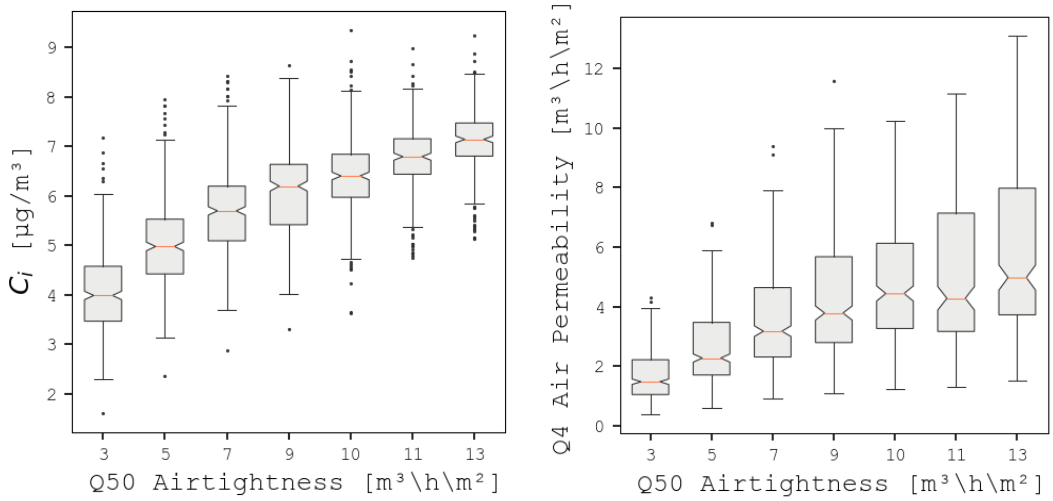


Figure 6.3: Box Plots of Left: Heating Season Concentrations of Infiltrated $PM_{2.5}$ and Right: Zone Air Permeability Q_4 plotted for each Building Airtighness Value Q_{50}

Based on the elimination of Q_4 and *V*, variables not included in the initial set of important variables were further tested. This includes looking into space geometry variables such as A_{ef}:V_z. It is clear from the sensitivity analyses that Aef:V_z stands at the centre of the ranking, with its rank varying between 7 – 10 for different tests. This variable is considered an essential variable as it normalises the area of exposed façade for each zone A_{ef} to the internal volume of the zone V_z and is likely to be essential for C_i following (J. Taylor et al., 2014b). Table 6.5 shows the results of the multicollinearity test between various space geometry (A_{ef}:V_z and A_{ef}:A_z) and building envelope variables (A_{ef}, N_{ef}, and L_{ef}). Based on the multicollinearity test and the sensitivity analysis results in Tables 6.1, 6.2, and 6.3, A_{ef}:V_z was included, while other variables were eliminated from the final set.

Table 6.5: Testing for Multicollinearity using Correlation and Regression tests between space geometry and building envelope variables.

Input	S_{Pears}	*p*-value	S_{Spear}	*p*-value	$S_{Regress}$	*p*-value	$S_{Regress\text{-}Rank}$	*p*-value
A_{ef}:V_z and A_{ef}:A_z	0.973	0.000	0.968	0.000	0.947	0.000	0.937	0.000
A_{ef} and L_{ef}	0.866	0.000	0.788	0.000	0.749	0.000	0.620	0.000
A_{ef} and N_{ef}	0.640	0.000	0.667	0.000	0.408	0.000	0.445	0.000
A_{ef}:V_z and N_{ef}	0.411	0.000	0.290	0.000	0.169	0.000	0.085	0.000
A_{ef}:A_z and A_{ef}	0.105	0.000	0.184	0.000	0.011	0.000	0.033	0.000
A_{ef}:V_z and A_{ef}	0.051	0.013	0.114	0.000	0.003	0.013	0.012	0.000
A_{ef}:V_z and L_{ef}	0.041	0.047	0.047	0.020	0.002	0.047	0.002	0.020

Following the sensitivity analysis, the final list of inputs retained for the metamodel development includes five variables: ACH_{INF}, ΔT, v, Q_{50}, and A_{ef}:V_z. The detailed results of the VIF analysis for multicollinearity are presented in Tables B.1 and B.2 - Appendix B.

6.3 Development of a Metamodel as a $PM_{2.5}$ Predictor

ML techniques have shown tremendous potential in predicting the concentrations of indoor $PM_{2.5}$ with high accuracy (see Chapter 2). The primary objective of this section is to present the metamodel development process utilising three different ML algorithms, namely Generalized Additive Models (GAM), eXtreme Gradient Boosting (XGB), and Random Forest Regression (RFR), to predict the heating season concentrations of infiltrated $PM_{2.5}$ using six input variables: ACH_{INF}, ΔT, H, v, and A_{ef}:V_z. The dataset comprises the CONTAM-EnergyPlus co-simulations of the four selected university buildings (N= 2729 zones), which were randomly divided into a training set (70%, n=1910) and a testing set (30%, n=819). The GAM metamodel employed

Generalized Cross-Validation (GCV) as an internal cross-validation method, while k-fold cross-validation was employed for the XGB and RFR metamodels. The performance of each metamodel was evaluated based on root mean squared error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE) metrics. The metamodels were further evaluated using a hold-out dataset to identify the model with the best performance. The hold-out dataset was obtained from the CONTAM-EnergyPlus co-simulation of the ICoSS Building (n = 42 zones) – the fifth selected building for this research.

Furthermore, the predicted concentrations of infiltrated $PM_{2.5}$ generated by the developed metamodel will be employed to estimate the annual population exposure of the selected HE buildings to infiltrated $PM_{2.5}$, with comparisons made against the World Health Organisation's annual permissible exposure levels of 5 $\mu g/m^3$. Finally, the resultant percentages of an exceedance will be computed to identify spaces where the HEI populace is at a greater risk of exposure. The ensuing chapter shall deliberate upon these findings.

6.3.1 Training the Algorithms

In order to obtain reliable predictions of the heating season concentrations of infiltrated $PM_{2.5}$, 6 models were fitted [(3 $Models_{pre\text{-}HPT}$ + 3 $Models_{post\text{-}HPT}$) [10] * Heating Season Dataset = 6 Models]. Each model was fitted using the five input variables identified by the sensitivity analyses and one response variable, see Section 6.2. The Pearson Correlation Coefficient (R-value) was calculated for the training results. Ideally, the outputs acquired from each model should match the targets, i.e., the desired model outputs. Hence, a slope of 45° implies perfect fitting. In the case of validation datasets, when a model achieves an error value close to the average error of validation datasets, the training ceases immediately. In detail, an algorithm's performance on the dataset was verified using a multi-level CV technique before and after hyperparameter tuning (pre- and post-HPT). Initially, each algorithm is set to its default hyperparameter values, and a 3-fold randomised search CV is used to determine the most optimal hyperparameters. Afterwards, the dataset was divided into 10-folds of validation subsets so that the RMSE of each subset could be calculated and averaged over the 10-folds. Following that, the averaged RMSE score across the 10-folds was compared to the RMSE score of the training dataset. Finally, this method was applied to each algorithm pre-HPT and post-HPT to track the performance of all algorithms.

[10] pre-HPT – pre hyperparameter tuning; post-HPT – post hyperparameter tuning.

As shown in Figure 6.4, both the $RFR_{post\text{-}HPT}$ and $XGB_{post\text{-}HPT}$ models have achieved very high R-values, with R = 0.999, and thus, very high correlation with CoSIM datasets. This is followed by the $GAM_{post\text{-}HPT}$, with R = 0.910. The regression line fit between the predicted and CoSIM values for the training dataset is indicated in Figure 6.4. It can be seen that the highest regression value (R^2) for the training dataset using $XGB_{post\text{-}HPT}$ is above $R^2 = 0.999$. This is followed by the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, with $R^2 = 0.998$ and 0.830, respectively. As a result, it is observed that the $XGB_{post\text{-}HPT}$ model predicts the heating season concentrations of infiltrated $PM_{2.5}$ very closely to the CoSIM values.

Moreover, most of the input data points of the $XGB_{post\text{-}HPT}$ fall closer to the regression fit line compared to the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, as depicted in Figures 6.4. It can be seen from the scatter plots that most regression points are located along the diagonal line, where some regression points deviate from the fitting line. Comparing the model performance metrics across $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ shows that the $XGB_{post\text{-}HPT}$ gives less prediction error than $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$. As seen in Table 6.6, the value of the RMSE and MAE is 0.017 and 0.010, respectively; thus, it gives the least prediction error. This is followed by the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, with RMSE = 0.025 and 0.543, and MAE = 0.020 and 0.410, respectively. These results demonstrate that the $XGB_{post\text{-}HPT}$ has the highest prediction accuracy of 99.81% compared to $RFR_{post\text{-}HPT}$ (99.70%) and $GAM_{post\text{-}HPT}$ (92.45%) on the training datasets.

Table 6.6: Model Evaluation Metrics for Fitted GAM, RFR, and XGB before and after HPT (*Heating Season CoSIM* Dataset vs *Training* Dataset)

	GAM		**RFR**		**XGB**	
Performance Metric	pre-HPT	post-HPT	pre-HPT	post-HPT	pre-HPT	post-HPT
R	0.855	0.910	0.775	0.999	0.840	0.999
R^2	0.730	0.830	0.600	0.999	0.700	0.999
RMSE	0.680	0.543	0.830	0.025	0.725	0.017
MAE	0.510	0.410	0.650	0.020	0.570	0.010
Model Accuracy	90.30%	92.45%	87.70%	99.70%	85.20%	99.81%

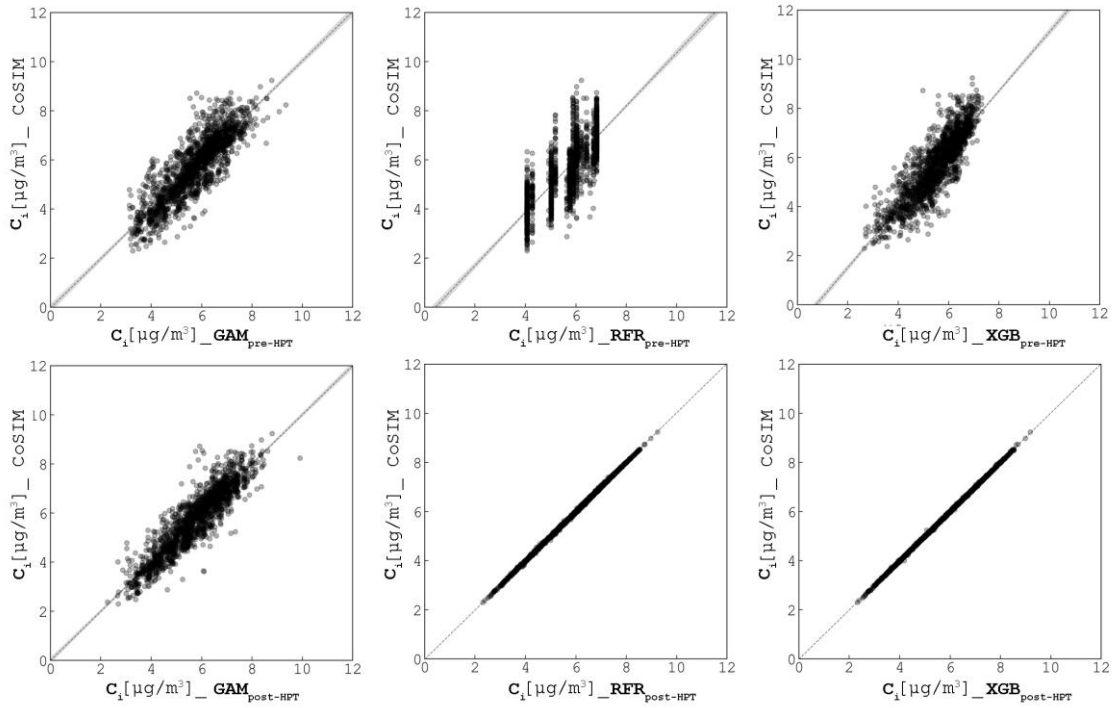


Figure 6.4: Regression plots of *Heating Season CoSIM* datasets vs *training* dataset for Fitted GAM, RFR, and XGB; top: pre-HPT and bottom: post-HPT

6.3.2 Testing the Models

The models are tested for their prediction ability using the CoSIM testing dataset (30%, n=819 zones). Both the $RFR_{post\text{-}HPT}$ and $XGB_{post\text{-}HPT}$ models have achieved very high R values, with R = 0.935 and 0.960, respectively. Thus, very high correlation with CoSIM datasets. This is followed by the $GAM_{post\text{-}HPT}$, with R = 0.900. The regression line fit between the predicted and CoSIM values for the testing dataset is indicated in Figure 6.5. It can be seen that the highest regression value (R^2) for the testing dataset using $XGB_{post\text{-}HPT}$ is above R^2 = 0.920. This is followed by the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, with R^2 = 0.880 and 0.815, respectively. As a result, it is observed that the $XGB_{post\text{-}HPT}$ model predicts the heating season concentrations of infiltrated $PM_{2.5}$ very closely to the CoSIM values.

Moreover, most of the input data points of the $XGB_{post\text{-}HPT}$ fall closer to the regression fit line compared to the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, as depicted in Figures 6.6. It can be seen from the scatter plots that most regression points are located along the diagonal line, where some regression points deviate from the fitting line. Comparing the model performance metrics across $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ shows that the $XGB_{post\text{-}HPT}$ gives less prediction error than $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$. As seen in Table 6.7, the value of the RMSE and MAE is 0.370 and

0.260, respectively, and thus, gives the least prediction error. This is followed by the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$, with RMSE = 0.450 and 0.560, and MAE = 0.315 and 0.440, respectively.

These results demonstrate that the $XGB_{post\text{-}HPT}$ has the highest prediction accuracy of 95.00% compared to $RFR_{post\text{-}HPT}$ (93.90%) and $GAM_{post\text{-}HPT}$ (91.70%) on the testing datasets. Furthermore, from Figure 6.5, it is demonstrated that a generally acceptable agreement between the predicted data and the CoSIM data has been achieved using the testing datasets.

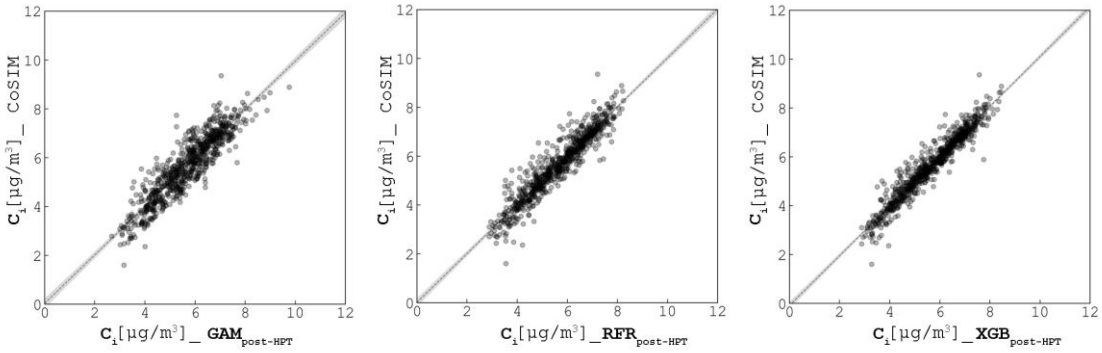


Figure 6.5: Regression plots of *Heating Season* CoSIM Datasets vs *Testing* Dataset for Fitted $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$

Table 6.7: Model Evaluation Metrics for Fitted $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ (*Heating Season* CoSIM Dataset vs *Testing* Dataset)

Performance Metric	**$GAM_{post\text{-}HPT}$**	**$RFR_{post\text{-}HPT}$**	**$XGB_{post\text{-}HPT}$**
R	0.900	0.935	0.960
R^2	0.815	0.880	0.920
RMSE	0.560	0.450	0.370
MAE	0.440	0.315	0.260
Model Accuracy	91.70%	93.90%	95.00%

6.4 Model Explanations using SHAP

As previously explained in Section 5.4.4, SHAP was used as a unifying framework for quantifying the marginal contributions of each predictor variable in each model. Doing so makes it possible to determine why a model makes specific predictions, i.e. interpretable predictive metamodels.

Figure 6.6 compares the SHAP values of the most influential input variables on the heating season concentrations of ifiltrated $PM_{2.5}$ in $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ with each dot representing one test data point (zone). The input variables were sorted in each plot by the sum of the absolute average SHAP values over all test points. These values illustrate the distribution of each input variable's impact on the prediction. The summary plot of the SHAP values for each variable consist of points transitioning from a blue colour representing a low value of an input variable (e.g. Q_{50} = 3 $m^3/h/m^2$) , to a red colour, representing a higher value of an input variable (e.g. Q_{50} = > 5 $m^3/h/m^2$). The x-axis represents the SHAP value impact on model output i.e, the concentrations of infiltrated $PM_{2.5}$ with positive SHAP values representing a positive contribution on the prediction and negative values representing a negative contribution of the prediction.

Table 6.8 below summarises Figure 6.6. It is obvious that all three models capture the impact of Q_{50} on the concentrations of infiltrated $PM_{2.5}$ as it was ranked first in all three models. However, the percentage of contribution that Q_{50} has on infiltrated $PM_{2.5}$ varies across all models. For example the $GAM_{post\text{-}HPT}$ model shows that Q_{50} explains almost 40% of the variation in infiltrated $PM_{2.5}$ concentrations, meanwhile it explains 32.5% and 33.6% for $XGB_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$, respectively. Additionally, it is notable that the ranking of the other input variables in $RFR_{post\text{-}HPT}$ and $XGB_{post\text{-}HPT}$ is similar, meanwhile it differs in $GAM_{post\text{-}HPT}$. This behaviour is expected, as both RFR and XGB are tree based algorithms where they use decision trees as their base learners and capture variable interactions. However, GAM requires the interactions to be predetermined in the construction of the metamodel.

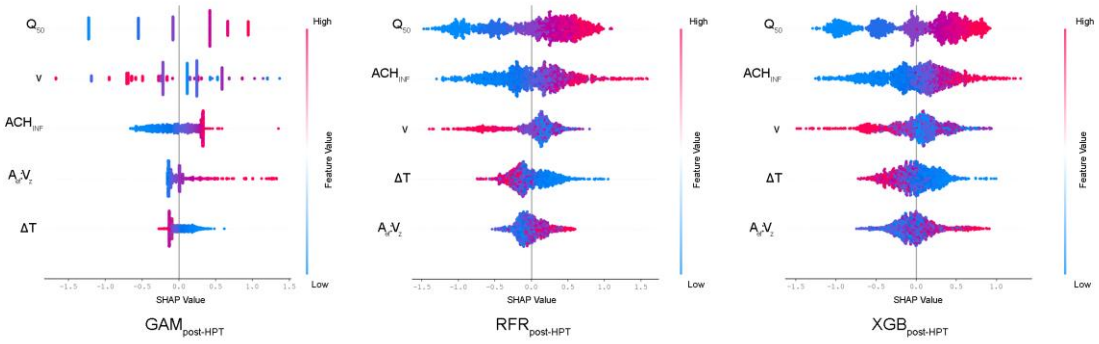


Figure 6.6: Importance and Threshold of Features for the heating season concentrations of infiltrated $PM_{2.5}$

Table 6.8: The *absolute average* SHAP-Value and calculated variations explained by each input variable shows on the *heating season* concentrations of infiltrated $PM_{2.5}$, with the variable's rank between parentheses.

Model	**A_{ef}:V_z**		**ACH_{INF}**		**ΔT**		**Q_{50}**		**v**	
	SHAP	%	SHAP	%	SHAP	%	SHAP	%	SHAP	%
$GAM_{post\text{-}HPT}$	+0.17 (4)	11.3	+0.24 (3)	15.9	+0.11 (5)	7.3	+0.60 (1)	39.7	+0.39 (2)	25.8
$RFR_{post\text{-}HPT}$	+0.15 (5)	9.9	+0.39 (2)	25.7	+0.21 (4)	13.8	+0.51 (1)	33.6	+0.26 (3)	17.1
$XGB_{post\text{-}HPT}$	+0.19 (5)	12.6	+0.34 (2)	22.5	+0.21 (4)	13.9	+0.49 (1)	32.5	+0.28 (3)	18.5

Figures 6.7 and 6.8 show a heat map plot of the individual impact a variable has on the heating season concentrations of infiltrated $PM_{2.5}$. The colors in these plots represents the SHAP value with red color representing positive SHAP values and blue representing negative values. The plots shows the variables contribution in pushing the concentrations of infiltrated $PM_{2.5}$ from the base value (mean concentration of the dataset Ci = 5.73 µg/m^3) to the actual predicted value. In other words, variables that contribute positively to the prediction are represented by the colour red, while those that have a negative impact are depicted by the colour blue. The x-axis represents the zones in the dataset (zone ID). It is clear from the plots that both RFR and XGB show the number of variables marked in red is nearly equal to the number of variables marked in blue, indicating a nearly balanced distribution of variables that play both positive and negative impact on infiltrated $PM_{2.5}$.

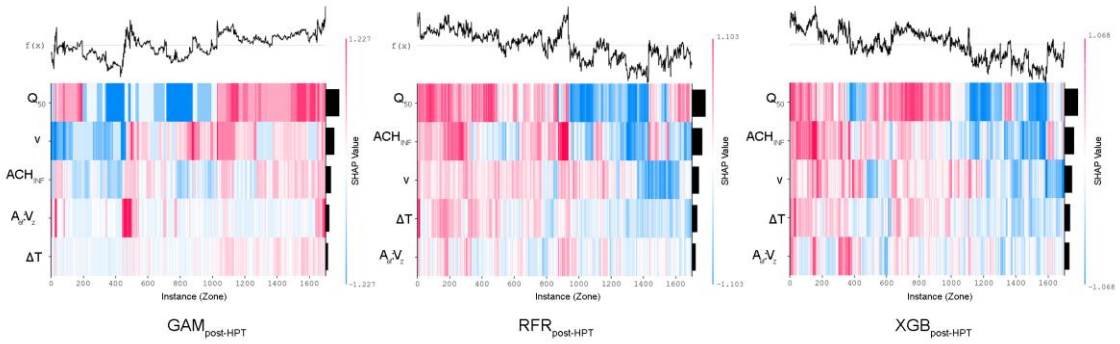


Figure 6.7: Overall Impact of input variables on the heating season concentrations of infiltrated $PM_{2.5}$ (f(x) =mean $\boldsymbol{Ci}$ of the sample = 5.73 µg/m^3)

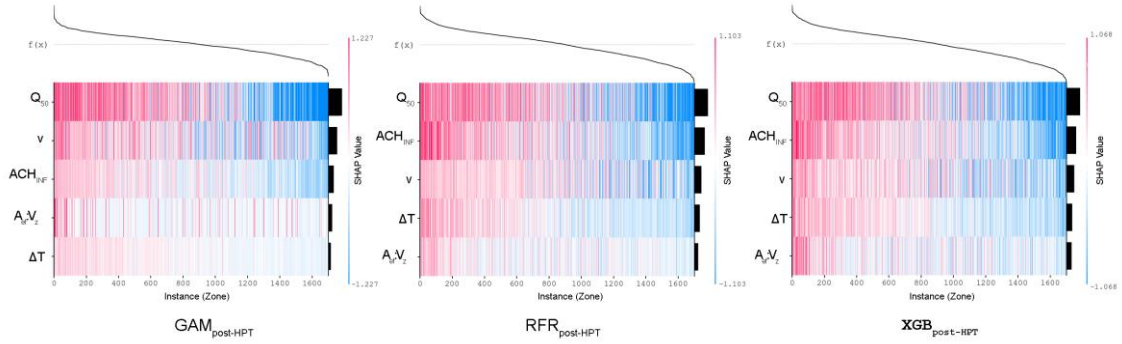


Figure 6.8: The ordered Overall Impact of input variables on the heating season concentrations of infiltrated $PM_{2.5}$ (f(x) =mean $\boldsymbol{Ci}$ of the sample = 5.73 μg/m^3)

Figure 6.9 further demonstrates how those variables impact the heating season concentrations of infiltrated $PM_{2.5}$ in randomly selected zones for the $XGB_{post\text{-}HPT}$ metamodel. For example, in Sample 1, at the bottom, indicated with "base value" = 5.73 μg/m^3, is the mean concentration of the dataset, and [f(X)] = 3.83 μg/m^3 is the predicted concentration by the $XGB_{post\text{-}HPT}$ for that specific zone. The plot demonstrates the changes in the dataset's average concentration resulting from the inclusion or exclusion of specific variables. The colours and size of the variables indicate the impact (+/-SHAP value) a variable has on the concentrations with the values of variables given for each choice that contributes in pushing the value to or away from the average concentration. When looking at more samples, it is worth noting that even though the f(x) might be same for different zones, the contributions of variables vary.

In general it is important to highlight that SHAP plays a crucial role in facilitating local interpretation of prediction effects by offering a comprehensive and dependable methodology for discerning the contribution of individual variables within a metamodel. By addressing the "why" behind a metamodel's specific prediction for a given instance, i.e., a zone / building , SHAP aids in unravelling the underlying rationale. Moreover, SHAP values provide insights into both the direction and magnitude of feature effects on predictions, highlighting the influential variables that drive predictions towards higher or lower values. This analytical capability fosters a deeper understanding of the decision-making process employed by the metamodel. Consequently, SHAP serves as a vital tool for promoting transparency, trust, and interpretability in ML models.

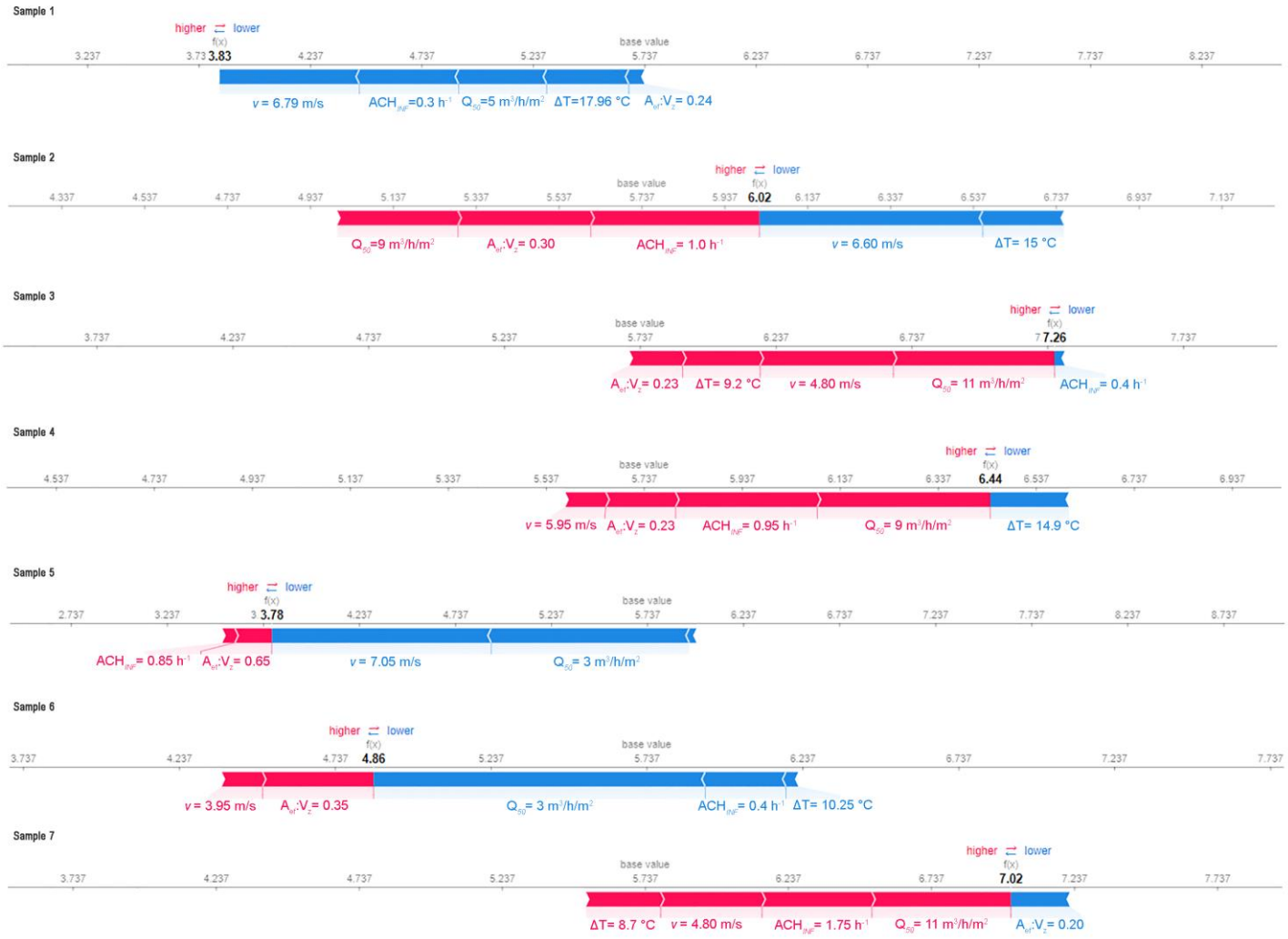


Figure 6.9: Examples of individual effects of variables on the heating season concentrations of infiltrated $PM_{2.5}$

6.5 Metamodels Evaluation of Unseen Data

It is essential to have a generalisable model to predict data not included in the training and testing phases of the model development. Thus, this additional step is not intended to improve the accuracy of the model's predictions by using a hold-out dataset. Nevertheless, it can increase the confidence that models developed can accurately predict unseen data. For example, suppose a model performs poorly on unseen data; in that case, this could indicate that the model suffers from high variance due to overfitting and that the models were developed more complexly than necessary.

For this reason, this section presents the results of evaluating $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ on the hold-out dataset, i.e., the ICoSS building CoSIM dataset. The aim is to see how each model performs in predicting the concentrations of infiltrated $PM_{2.5}$ in ICoSS over the heating season. The models were evaluated based on the RMSE, MAE, and R^2 values between the predicted results and the CoSIM results of the ICoSS building.

As shown in Figure 6.10, the $XGB_{post\text{-}HPT}$ model has achieved very high R values, with R = 0.950, and thus, almost accurate predictions. This is followed by the $GAM_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$, with R = 0.935 and 0.790, respectively. The regression line fit between the predicted and CoSIM values for the hold-out dataset is indicated in Figure 6.10. It can be seen that the highest regression value (R^2) for the evaluation datasets using $XGB_{post\text{-}HPT}$ is above R^2 = 0. 905. The $GAM_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$ follow this with R^2 = 0.860 and 0.620, respectively. As a result, it is observed that the $XGB_{post\text{-}HPT}$ model predicts the infiltrated $PM_{2.5}$ concentration very closely to the CoSIM values.

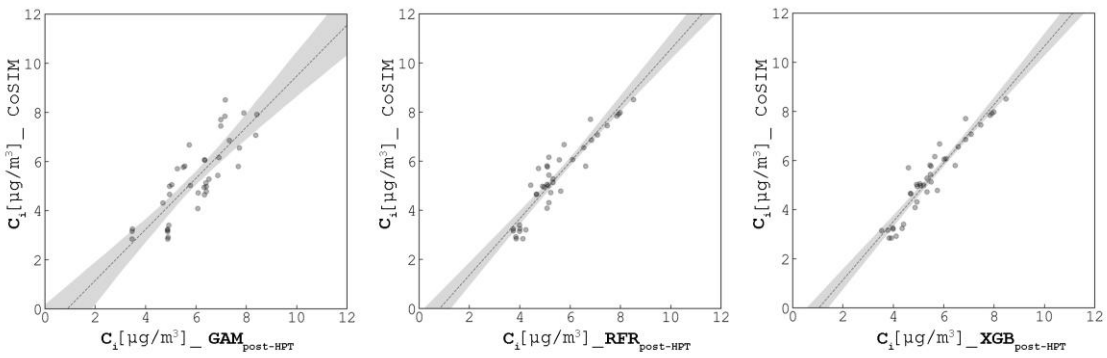


Figure 6.10: Regression Plots of *Heating Season* 'ICoSS' *CoSIM* Datasets vs *Prediction* Dataset for $GAM_{post\text{-}HPT}$, $RFsR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$

Moreover, most of the input data points of the $XGB_{post\text{-}HPT}$ fall closer to the regression fit line compared to the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$ in all months, as depicted in Figure 6.10. It can be seen from the scatter plots that most regression points are located along the diagonal line, where some regression points deviate from the fitting line. Comparing the model performance metrics across $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ shows that the $XGB_{post\text{-}HPT}$ gives less prediction error than $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$. As seen in Table 6.9, the value of the RMSE and MAE is 0.580 and 0.410, respectively; thus, it gives the least prediction error. This is followed by the $GAM_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$, with RMSE = 0.618 and 1.150, and MAE = 0.460 and 0.955, respectively.

Table 6.9: Model Evaluation Metrics for Fitted $GAM_{post\text{-}HPT}$, $RFR_{post\text{-}HPT}$, and $XGB_{post\text{-}HPT}$ on the ICoSS *Heating Season* Dataset

Performance Metric	**$GAM_{post\text{-}HPT}$**	**$RFR_{post\text{-}HPT}$**	**$XGB_{post\text{-}HPT}$**
R	0.790	0.935	0.950
R^2	0.620	0.860	0.905
RMSE	1.150	0.618	0.580
MAE	0.955	0.460	0.410
Model Accuracy	78.59%	88.85%	90.60%

It is evident from these results that the $XGB_{post\text{-}HPT}$ is the most accurate on the evaluation datasets, with a prediction accuracy of 92.60%, compared to the $RFR_{post\text{-}HPT}$ (88.85%) and the $GAM_{post\text{-}HPT}$ (78.59%). Furthermore, when comparing the models' performance on the testing datasets with their performance on the hold-out dataset, $XGB_{post\text{-}HPT}$ exhibits nearly identical performance in terms of R^2 and prediction accuracy. In contrast, the $RFR_{post\text{-}HPT}$ and $GAM_{post\text{-}HPT}$ show a 5% and 13.1% reduction in the prediction accuracy, respectively. A possible reason is that the XGBoost algorithm incorporates regularisation terms into its objective function to regulate the complexity of the model and facilitate column sampling. These measures are implemented to prevent the model from overfitting the training data and to optimise computation time. Additionally, XGBoost follows a specific tree construction approach, building all possible subtrees from the top down, and then performs reverse pruning from the bottom up. This sequential pruning strategy ensures that the model avoids getting trapped in local optimal solutions.

6.6 Summary

In this chapter, a novel methodology for predicting indoor $PM_{2.5}$ concentration levels within different rooms within an HEI building stock was developed by testing different machine learning algorithms. Namely, Generalised Additive Models (GAM), Random Forest Regression (RFR), and Extreme Gradient Descent Boosted Trees (XGB). The heating season concentrations of infiltrated $PM_{2.5}$ concentration used to train, test, and evaluate the models were based on the outputs of the co-simulation framework described in Chapter 4. The data was obtained for November, December, January, February, March, and April, covering the entire heating season as specified in the University of Sheffield Heating Policy (Section 3.3.5).

Prior to the development of the ML models, a sensitivity analysis framework applied to the CoSIM outputs revealed the sensitivity of infiltrated $PM_{2.5}$ concentrations to different independent variables (Q_{50}, ACH_{INF}, ΔT, v, and A_{ef}:V_z). These variables were used to develop a predictive metamodel so that they can be targeted for attention when designing new buildings. For predicting the infiltrated $PM_{2.5}$ concentrations, each algorithm was fitted using five input variables and one response variable. Each model's performance on the dataset was verified using a multi-level CV technique before and after hyperparameter tuning on the training and testing datasets.

According to the results, $XGB_{post\text{-}HPT}$ achieved an R^2 value higher than 0.92 for the heating season concentrations of infiltrated $PM_{2.5}$ on training and testing datasets, with a model prediction accuracy greater than 95%. The $XGB_{post\text{-}HPT}$ then appears to be a powerful tool for making predictions by achieving high levels of accuracy. Furthermore, the results indicate that the developed $XGB_{post\text{-}HPT}$ model is slightly more predictive than the $GAM_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$ models.

Following this, SHAP was used as a common framework for quantifying the marginal contributions of each input variable to heating season concentrations of infiltrated $PM_{2.5}$. By estimating the contributions of the variables locally and globally, it was possible to explain/interpret the variations in $PM_{2.5}$ concentrations when certain variables were combined. This is especially valuable when explaining complex models like XGB, where traditional feature importance measures may not be sufficient.

The models were then evaluated using a hold-out dataset to test their generalisability to unseen data. Evaluation results from the fifth building (ICoSS) study show that $XGB_{post\text{-}HPT}$ had higher prediction accuracy than $GAM_{post\text{-}HPT}$ and $RFR_{post\text{-}HPT}$, with an R^2 above 0.90 for predicting the infiltrated $PM_{2.5}$ concentrations. In conclusion, the study highlights the substantial importance of developing reliable but easy-to-use and interpretable metamodels that can effectively predict the levels of infiltrated $PM_{2.5}$ and test the impacts of increasing the airtightness of building envelope on the concentrations of infiltrated $PM_{2.5}$ over the heating season. Finally, the results of this chapter can pave the way to move ahead with estimating the population exposure to infiltrated $PM_{2.5}$ in different microenvironments. They can inform decision-making for future master planning (Chapter 7)

Chapter 7 Microenvironmental Modelling of Population Exposures

7.1 Introduction

Chapter 6 described the development of the $XGB_{post\text{-}HPT}$ model for best predicting the heating season concentrations of infiltrated $PM_{2.5}$ in different UoS building stock zones. Based on the model's predictions for various zones, this chapter uses a microenvironment modelling approach to estimate the average Personal Exposure (E_i) to infiltrated $PM_{2.5}$ and the average Population-Weighted Exposure (PWE) to infiltrated $PM_{2.5}$ for different microenvironments. A microenvironment can be defined as a three-dimensional space in which pollutant levels are uniform or exhibit constant statistical properties over time (Watson et al., 1988). To conduct the epidemiological analyses, it is necessary to consider the determinants of *time-activity* patterns. Personal exposures vary due to variations in personal time-activity fractions (Lane et al., 2015). Several studies have shown that exposure varies as individuals move through various microenvironments, including the home, office, car, bus, and outdoors. (Gulliver & Briggs, 2004; Nasir & Colbeck, 2009). Due to the health risks associated with some demographic groups, it is justified and necessary to consider a wide range of time-activity fractions. The study by Elliot et al. (Elliot et al., 2000) found that even individuals working within the same building will be exposed to varying levels of air pollution based on the patterns of their daily activities. In this regard, we propose four categories of occupancy profiles that might reflect Similar Time-Activity groups (STGs) in HEI buildings (Klepeis, Nelson, Ott, Robinson, Tsang, Switzer, Behar, et al., 2001). Here, the definition of STGs in HEIs in the UK is based on the data from the UK Higher Education Statistics Agency (HESA) (HESA, 2021). The HESA divides HEI buildings users into students, academic staff, and non-academic staff.

7.2 Similar Time-Activity Groups (STGs)

This thesis project examined offices, studios, lecture halls, labs, workshops, and study rooms in specific university buildings as examples of locations that can be defined, under appropriate conditions, as microenvironments. Taking the University of Sheffield as an example, those spaces can be classified into four main types of microenvironments: *offices*, *educational facilities*, *shared facilities*, and *circulation areas*. As shown in Table 7.1, the time-weighting fractions for individuals moving through various microenvironments in a university building define the total amount of time spent in each microenvironment (Klepeis, Nelson, Ott, Robinson, Tsang, Switzer, Behar, et al., 2001). The time taken for individuals to move between rooms is approximately ten minutes, which is regarded as negligible in this study. It is imperative to note, however, that the time-weighting factors are only for demonstration and may not reflect the actual behaviour of the population. In this study, individuals were assumed to spend the entire working hours in a university building, thus removing any external environments as a cause of variability in personal exposure (e.g., homes or transport). Furthermore, internal sources were not taken into account in the exposure study. Therefore, the results reflect the exposure to indoor $PM_{2.5}$ infiltrated from external sources.

Table 7.1: Assumed Typical Time Fractions Spent in Each Microenvironment for Different HEI Building Users

Microenvironment	Academic Staff	Administration Staff	Undergraduate Students	Post Graduate Students
Offices	0.4705 [197 min]	0.941 [395 min]	0.00	0.600 [360 min]
Educational Facilities	0.4705 [197 min]	0.00	0.784 [470 min]	0.134 [80 min]
Circulation	0.016 [10 min]	0.016 [10 min]	0.016 [10 min]	0.016 [10 min]
Shared Facilities	0.035 [15 min]	0.035 [15 min]	0.200 [120 min]	0.250 [150 min]

Personal exposure to $PM_{2.5}$ was estimated from the predicted heating season concentrations of the $XGB_{post\text{-}HPT}$ model added to the CoSIM non-heating season concentrations for building users in four categories: (a) an 'academic' with an average concentration of $PM_{2.5}$ in the offices, educational facilities and shared facilities using time weighting factors of 0.4705, 0.4705, and 0.035 respectively; (b) the exposure experienced by an 'administrative staff' who occupies the office and shared facilities using weighting factors of 0.941 and 0.035, respectively; (c) the exposure of an 'undergraduate student' who never enters the office microenvironment and mainly spends time in the educational facilities and shared facilities with weighting 0.784 and 0.20, respectively, and (d) the exposure of a 'post-graduate research student' (PGR) who occupies the

offices, educational facilities and shared facilities with weighting 0.60, 0.134 and 0.25, respectively. For each category of builder users, the population-weighted indoor exposure to $PM_{2.5}$ across the selected buildings investigated was calculated from the predictions using the weightings that reflect the maximum design occupancy for each zone in the buildings and assuming that all zones were fully occupied.

7.3 Personal Exposure to Indoor $PM_{2.5}$

Two methods have been widely employed to calculate personal exposures to indoor pollutants as individuals move through a series of microenvironments within a building. First, the "time-weighted average" exposure can be determined by multiplying the average concentration of $PM_{2.5}$ by the percentage of time spent in each microenvironment (Kousa et al., 2001). An alternative method is to use a "time-activity profile" in which the exposure at each time interval is equal to the concentration in that microenvironment at the time interval in question and then calculate the cumulative exposure by dividing it by the total time (Dimitroulopoulou et al., 2001). According to the second approach, a more detailed time-activity profile would require information about the period and duration spent by the individual(s) in each microenvironment. Since the model developed can only predict the average $PM_{2.5}$ concentrations indoors, the time-weighted personal exposure method for indoor pollution levels will be used here. The general form of the equation used to calculate the personal exposure from a specific microenvironment is defined as (Watson et al., 1988):

$$\mathrm{E}_i = C_j t_{ij} \tag{7.1}$$

where E_i is the time-weighted personal exposure for person i over the specified time fraction; C_j is the average pollutant concentration in microenvironment j; t_{ij} is the aggregate time that person i spends in microenvironment j. Eq (7.1), can be modified to calculate the time-weighted integrated personal exposure EI_i by combining the average $PM_{2.5}$ concentrations at different microenvironments with the time fractions defined. The general form of the equation used to calculate the integrated exposure from various microenvironments is defined as (Watson et al., 1988):

$$\mathrm{EI}_i = \sum_{j}^{J} C_j t_{ij} \tag{7.2}$$

where J is the total number of microenvironments that person i moves through during the specified time period.

Using Eq. (7.1), the personal exposure for STGs can be estimated using the annual indoor $PM_{2.5}$ concentrations. Nonetheless, it is essential to note that if Eq (7.1) is used along with the annual indoor $PM_{2.5}$ concentrations, a number of assumptions are made: that person *i* is present in microenvironment *j* during the time (t_{ij}) and the concentration C*j* in that microenvironment *j* remains constant and homogeneous; that the short-term variations in indoor concentrations that could vary substantially due to variations in air change rates and outdoor $PM_{2.5}$ concentrations are ignored; that only a few microenvironments are necessary to characterise personal exposure adequately, but in reality, it is uncertain how many microenvironments are required to make accurate estimations.

In this study, the annual personal exposure to indoor $PM_{2.5}$ was calculated using Eq. (7.1) and considered three scenarios based on the airtightness of the buildings: Baseline Q_{50}, Q_{50} = 7 $m^3/h/m^2$, and Q_{50} = 3 $m^3/h/m^2$. Table 7.2 shows the annual personal exposure to indoor $PM_{2.5}$ using the baseline Q_{50}. The results show that the annual personal exposure of an individual working in the office microenvironment (E_{Off}) can range from 5.28 – 14.06 µg/m^3 with an average of 10 (± 1.45) µg/m^3. As seen in Table 7.2, administrative staff had the highest annual personal exposure to $PM_{2.5}$, with E_{Admin} = 9.53 (± 1.28) µg/m^3 due to the duration spent in the office microenvironment across all the selected buildings. When comparing academic staff's annual personal exposure (E_{Aca}) and PGRs' annual personal exposure (E_{PGR}), their annual personal exposure is 4.33 (± 1.66) µg/m^3 and 6.55 (± 1.24) µg/m^3, respectively. This result suggests that administrative staff are under a higher risk of exposure to indoor $PM_{2.5}$ from outdoor sources.

Table 7.2: Personal time-weighted exposure to annual indoor $PM_{2.5}$ in different microenvironments using the baseline building airtightness values Q_{50*}

Microenvironment Type	Annual $PM_{2.5}$ Concentration (C_j, µg/m^3)	Sub-Category	Annual $PM_{2.5}$ Concentration (C_j, µg/m^3)	Time Fraction (t_{ij})	C_j x t_{ji} (µg/m^3)
Offices Microenvironment	10.00 (± 1.45)	Administration Offices	10.12 (± 1.28)	0.941	9.53
		Academic Offices	9.23 (± 1.66)	0.470	4.33
		PGRs Offices	10.92 (± 1.24)	0.600	6.55
Educational Facilities Microenvironment **	9.36 (± 1.34)	Workshops	9.20 (± 0.99)	0.784 (0.47)	7.21 (4.32)
		Lecture Halls	10.00 (± 1.03)	0.784 (0.47)	7.84 (4.70)
		Labs	9.21 (± 0.56)	0.784 (0.47)	7.22 (4.34)
		Studios	9.35 (± 8.83)	0.784 (0.47)	7.33 (4.39)
Shared Facilities Microenvironment **	9.38 (± 1.58)	Study Rooms	9.38 (± 1.58)	0.784 (0.30)	7.35 (2.81)

*Q_{50} = 13 $m^3/h/m^2$ for the BH and ADC buildings, and 10 $m^3/h/m^2$ for the AT and RC buildings (Section 3.3.2).

**The time fraction 0.784 is used for an undergraduate student spending most of the time in the educational facilities microenvironment. t_{ij} = 0.47 and 0.30 in parentheses are used for Academics and PGRS, respectively.

On the other hand, the results show that an individual's annual exposure to the educational facilities microenvironment (E_{Edu}) can range from 6.54 – 12.45 µg/m^3 with an average of 9.36 (± 1.34) µg/m^3. In detail, the educational facilities microenvironment includes studios, lecture halls, workshops, and labs. It is clear that based on the sub-categorisation, lecture halls had the highest annual personal exposure to $PM_{2.5}$ with $E_{lecture}$ =7.84 (± 1.03) µg/m^3. This is followed by studios 7.33 (± 0.83) µg/m^3, and labs and workshops sharing a similar value of 7.2 (± 0.75) µg/m^3. These values are estimated assuming an undergraduate student spends most of the time daily in this microenvironment. Finally, the results show that the annual personal exposure of an individual in the shared facilities microenvironment (E_{Shared}) during can range from 5.66 – 12.29 µg/m^3 with an average of 9.38 (± 1.58) µg/m^3.

Table 7.3 demonstrates the impact of improving Q_{50} on the annual personal exposure to indoor $PM_{2.5}$ in different microenvironments. It can be noticed that improving the Q_{50} to represent buildings with moderate airtightness (Q_{50}= 7 m^3/h/m^2 scenario), there was a decrease in the annual personal exposure to $PM_{2.5}$ in most microenvironments compared to the Baseline Q_{50} scenario. The results show that the annual personal exposure of an individual working in the office microenvironment exhibited an 8% reduction with (E_{Off}) ranging from 3.65 – 12.90 µg/m^3 with an average of 9.15 (± 1.55) µg/m^3. Additionally, the administrative staff ME experienced a 4% reduction with E_{Admim}= 9.15 (± 1.31) µg/m^3.

Table 7.3: Personal time-weighted exposure to annual indoor $PM_{2.5}$ in different microenvironments using a Q_{50} = 7 m^3/h/m^2

Microenvironment Type	**Annual $PM_{2.5}$ Concentration (C_j, µg/m^3)**	**Sub-Category**	**Annual $PM_{2.5}$ Concentration (C_j, µg/m^3)**	**Time Fraction (t_{ij})**	**C_j x t_{ji} (µg/m^3)**
Offices Microenvironment	9.17 (± 1.55)	Administration Offices	9.73 (± 1.31)	0.941	9.15
		Academic Offices	7.58 (± 1.34)	0.470	3.56
		PGRs Offices	9.63 (± 1.01)	0.600	5.78
Educational Facilities Microenvironment *	8.15 (± 1.03)	Workshops	8.31 (± 1.17)	0.784 (0.47)	6.51 (3.90)
		Lecture Halls	8.66 (± 1.08)	0.784 (0.47)	6.79 (4.07)
		Labs	8.00 (± 1.39)	0.784 (0.47)	6.28 (3.76)
		Studios	8.08 (± 0.94)	0.784 (0.47)	6.35 (3.80)
Shared Facilities Microenvironment *	7.86 (± 1.71)	Study Rooms	7.86 (± 1.71)	0.784 (0.30)	6.16 (2.36)

*The time fraction 0.784 is used for an undergraduate student spending most of the time in the educational facilities microenvironment. t_{ij} = 0.47 and 0.30 in parentheses are used for Academics and PGRS, respectively.

Both academic staff's annual personal exposure (E_{Aca}) and PGRs' annual personal exposure (E_{PGR}) exhibited an approximately 15% reduction in their annual personal exposure to 3.56 (± 1.34) μg/m^3 and 5.78 (± 1.01) μg/m^3, respectively. This result suggests that administrative staff remain under a higher risk of exposure to indoor $PM_{2.5}$ from outdoor sources. Other MEs exhibited a similar trend in reduction in annual personal exposure, with the annual personal exposure in educational facilities E_{Edu} and shared facilities E_{Shared} both reducing by 12% and 15% respectively.

Further improvements on the building envelope airtightness value Q_{50} = 3 m^3/h/m^2 highlighted a more significant decrease in the annual personal exposure to indoor $PM_{2.5}$ compared to the previous scenarios. In Table 7.4, the offices microenvironment exhibited a 30% reduction in personal exposure with E_{Off} ranging between 2.16 – 11.29 μg/m^3 and an average of 7.03 (± 1.34) μg/m^3 when compared to the baseline Q_{50} scenario. Additionally, improving the Q_{50} of the buildings reduced the annual personal exposure in educational facilities E_{Edu} and shared facilities E_{Shared} by 34% and 39% respectively. These results highlight the impact that improving the Q_{50} of buildings has on the annual personal exposure to indoor $PM_{2.5}$.

Table 7.4: Personal time-weighted exposure to annual indoor $PM_{2.5}$ in different microenvironments using a Q_{50} = 3 m^3/h/m^2

Microenvironment Type	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Sub-Category	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Time Fraction (t_{ij})	C_j x t_{ji} (μg/m^3)
Offices Microenvironment	7.03 (± 1.34)	Administration Offices	7.42 (± 1.19)	0.941	6.98
		Academic Offices	6.11 (± 1.20)	0.470	2.87
		PGRs Offices	7.67 (± 1.07)	0.600	4.60
Educational Facilities Microenvironment *	6.21 (± 0.96)	Workshops	6.63 (± 1.62)	0.784 (0.47)	5.20 (3.10)
		Lecture Halls	6.59 (± 0.97)	0.784 (0.47)	5.15 (3.05)
		Labs	5.91 (± 1.34)	0.784 (0.47)	4.63 (2.78)
		Studios	6.22 (± 0.91)	0.784 (0.47)	4.88 (2.90)
Shared Facilities Microenvironment *	5.80(± 1.50)	Study Rooms	5.80(± 1.50)	0.784 (0.30)	4.55 (1.75)

*The time fraction 0.784 is used for an undergraduate student spending most of the time in the educational facilities microenvironment. t_{ij} = 0.47 and 0.30 in parentheses are used for Academics and PGRS, respectively.

Using the annual indoor $PM_{2.5}$ concentration levels for the three Q_{50} scenarios (baseline Q_{50}, Q_{50}=7, and Q_{50}= 3) , the personal time-weighted, integrated exposure (EI_i) i.e., the total exposure from all microenvironments for the four categories of building users can be calculated using Eq. (7.2). This allows for the estimation of relative contribution from specific microenvironments to an individual's time-weighted integrated exposure. Tables 7.5, 7.6 and 7.7 show the calculated EI_i using the time fractions described for STGs in Section 7.2. The results show that for all three

scenarios, a PGR student is at higher risk from exposure to infiltrated $PM_{2.5}$ with EI_{Stu} = 10.15, 8.84, and 6.88 µg/m^3, respectively. This is followed by EI_{Adm}, EI_{Stu}, and EI_{Aca} with values of EI_{Adm} = 9.53, 9.15 and 7 µg/m^3, EI_{Stu} = 9.22, 7.97, and 6.03 µg/m^3, and EI_{Aca} = 9.07, 7.68, and 6 µg/m^3 for baseline Q_{50}, Q_{50} = 7 m^3/h/m^2, and Q_{50} = 3 m^3/h/m^2, respectively.

Table 7.5: Examples of the relative contributions from specific microenvironments to an STGs annual time-weighted, integrated exposure to indoor $PM_{2.5}$ from external sources using the **baseline airtightness values Q_{50}**. s

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, µg/m^3)	Time Fraction (t_{ij})	$C_j * t_{ji}$ (µg/m^3)	Microenvironment Contribution (%)
Administrative Staff	Offices	10.12	0.941	9.53	100 %
			$EI_{Adm} = \sum C_j . t_{ij}$ = 9.53 µg/m^3		
Academic Staff	Offices	9.23	0.470	4.34	47.8 %
	Educational Facilities	9.36	0.470	4.40	48.6 %
	Shared Facilities	9.38	0.035	0.33	3.6 %
			$EI_{Aca} = \sum C_j . t_{ij}$ = 9.07 µg/m^3		
PGR Student	Office	10.92	0.600	6.55	65.0 %
	Educational Facilities	9.36	0.134	1.25	12.0 %
	Shared Facilities	9.38	0.250	2.35	23.0 %
			$E_{IPGR} = \sum C_j . t_{ij}$ = 10.15 µg/m^3		
Undergraduate	Educational Facilities	9.36	0.784	7.34	79.6 %
	Shared Facilities	9.38	0.200	1.88	20.4 %
			$E_{IStu} = \sum C_j . t_{ij}$ = 9.22 µg/m^3		

Table 7.6: Examples of the relative contributions from specific microenvironments to an STGs annual time-weighted, integrated exposure to indoor $PM_{2.5}$ from external sources (**Q_{50} = 7 m^3/h/m^2**)

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, µg/m^3)	Time Fraction (t_{ij})	$C_j * t_{ji}$ (µg/m^3)	Microenvironment Contribution (%)
Administrative Staff	Offices	9.73	0.941	9.15	100 %
			$EI_{Adm} = \sum C_j . t_{ij}$ = 9.15 µg/m^3		
Academic Staff	Offices	7.58	0.4705	3.56	46.3 %
	Educational Facilities	8.15	0.4705	3.84	50.1 %
	Shared Facilities	7.86	0.0350	0.28	3.6 %
			$EI_{Aca} = \sum C_j . t_{ij}$ = 7.68 µg/m^3		
PGR Student	Office	9.63	0.600	5.77	65.4 %
	Educational Facilities	8.15	0.134	1.10	12.36 %
	Shared Facilities	7.86	0.250	1.97	22.24 %
			$E_{IPGR} = \sum C_j . t_{ij}$ = 8.84 µg/m^3		
Undergraduate	Educational Facilities	8.15	0.784	6.39	80.26 %
	Shared Facilities	7.86	0.200	1.58	19.74 %
			$E_{IStu} = \sum C_j . t_{ij}$ = 7.97 µg/m^3		

Table 7.7: Examples of the relative contributions from specific microenvironments to an STGs annual time-weighted, integrated exposure to indoor $PM_{2.5}$ from external sources (**Q_{50} = 3 $m^3/h/m^2$**)

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, µg/m^3)	Time Fraction (t_{ij})	C_j * t_{ji} (µg/m^3)	Microenvironment Contribution (%)
Administrative Staff	Offices	7.42	0.941	7.00	100 %
			$EI_{Adm} = \sum C_j . t_{ij}$ = 7.00 µg/m^3		
Academic Staff	Offices	6.11	0.4705	2.87	47.9%
	Educational Facilities	6.21	0.4705	2.93	48.7%
	Shared Facilities	5.80	0.0350	0.20	3.4%
			$EI_{Aca} = \sum C_j . t_{ij}$ = 6.00 µg/m^3		
PGR Student	Office	7.67	0.600	4.60	66.8%
	Educational Facilities	6.21	0.134	0.83	12.1%
	Shared Facilities	5.80	0.250	1.45	21.1%
			$E_{IPGR} = \sum C_j . t_{ij}$ = 6.88 µg/m^3		
Undergraduate	Educational Facilities	6.21	0.784	4.87	80.8%
	Shared Facilities	5.80	0.200	1.16	19.2%
			$E_{IStu} = \sum C_j . t_{ij}$ = 6.03 µg/m^3		

Figure 7.1 shows the average annual time fractions and the contributions of each microenvironment to EI_i to indoor $PM_{2.5}$ from outdoor sources for the four categories of STGs. The results show that the Educational Facilities ME_{EF}, has the highest contribution to the EI_{Stu} with values of 79.60 %. This reflects the percentage of time (78.4%) an undergraduate student could spend in the ME_{Edu}. With the assumption that an undergraduate student spends 20% of time in the Shared Facilities ME, the results show that the ME_{SF} contributes to about 20.40% of EI_{Stu} to indoor $PM_{2.5}$. On the other hand, with the assumption that administration staff spend on average 94.1% of their time in the offices ME (ME_{Off}), the results show that the ME_{Off} fully contributes (100%) to the EI_{Adm}. When comparing EI_{Adm} and EI_{Aca}, it is noticed that academic staff are subject to exposures from two main microenvironments: the ME_{Off} and ME_{Edu}. The time spent in both MEs is assumed to be equal (47.05%), and therefore the results show that the ME_{Edu} contributes to 48.6% of EI_{Aca} and ME_{Off} contributes to 47.8% of EI_{Aca}. These results indicate that academic staff are at higher risk of exposure to indoor $PM_{2.5}$ from the time spent in educational facilities than their own offices. Moreover, the results show that the ME_{Off} contributes to about 65% to the EI_{PGR}. This is followed by ME_{Shared} and ME_{Edu}, with contributions of 23% and 12%.

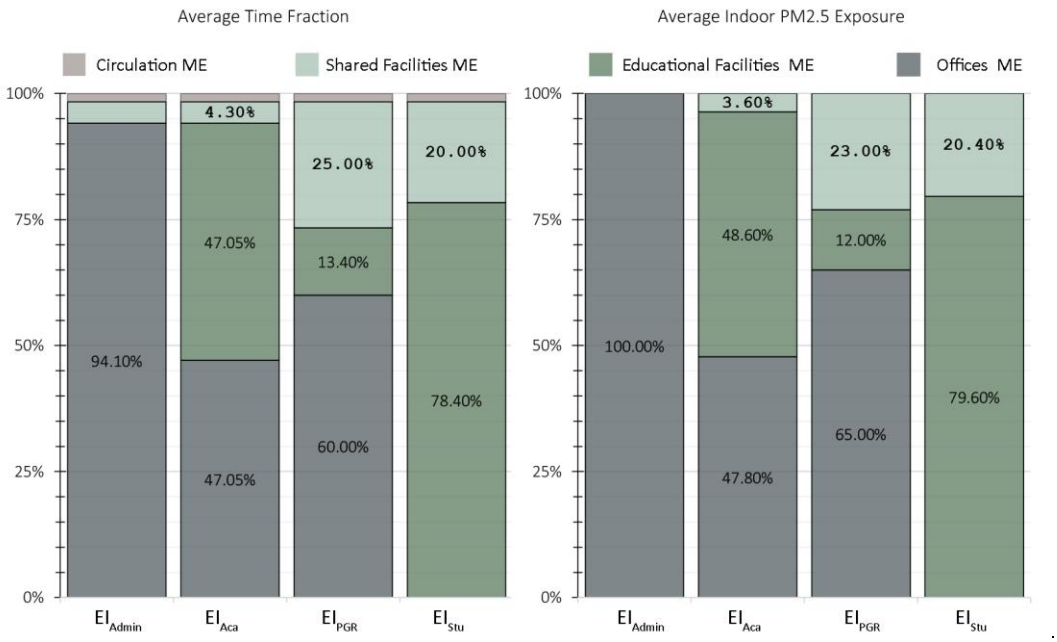


Figure 7.1: Time-Activity Fractions and Contributions of each microenvironment to annual indoor $PM_{2.5}$ from outdoor sources for different building users.

7.4 Population-weighted Exposure to indoor $PM_{2.5}$

7.4.1 Population-weighted Exposure to indoor $PM_{2.5}$ in Microenvironments

Following the calculation of the personal exposure as a function of annual average indoor concentration of $PM_{2.5}$ from outdoor sources in indoor microenvironments and the average fraction of time spent in those microenvironments (Eqs. (7.1) and (7.2)) , the Population-Weighted Exposure (PWE) metric can be calculated using Eq. (7.3) (Abdul Shakor et al., 2020; Aunan et al., 2018):

$$\mathrm{PWE}_i = \frac{1}{P_i}\sum_i C_i P_i \tag{7.3}$$

where P_i is the population, C_i is the annual average indoor $PM_{2.5}$ concentration and i refers to a microenvironment (offices, educational facilities and shared facilities). The Population-weighted annual average concentrations can provide better estimates of population exposures because they give proportionately greater weight to the indoor $PM_{2.5}$.-The heating season and annual average

indoor $PM_{2.5}$ levels concentrations were obtained from the $XGB_{post\text{-}HPT}$ model added to the CoSIM non-heating season concentrations. The population for each microenvironment represents the maximum number of occupants in each zone and summed up based on the categorisation of that zone to which microenvironment it belongs. This data was obtained from the Estates and Facilities Management (EFM) and presented in Table 7.8.

Table 7.8: Demography and Area Characteristics of the Microenvironments and Subcategories in investigated Buildings

`	**Sub_Category**	**Maximum Occupancy**	**Area (m^2)**	**Occupancy Density (m^2/person)**
Offices	Administration Offices	859	4184.4	4.5
	Academic Offices	315	2862.5	9
	PGR Offices	245	1490.2	6
Educational Facilities	Lecture Halls	794	1219.6	1.5
	Labs	478	1194.5	2.5
	Studios	861	2156.6	2.5
	Workshops	320	802.2	2.5
Shared Facilities	Study & Computer Rooms	782	1,954.7	2.5

It is imperative to note that different microenvironments within higher education buildings may have varying levels of indoor air pollution. By quantifying the population exposure, it is possible to identify specific areas or microenvironments that have higher concentrations of $PM_{2.5}$. This information can help prioritise interventions and targeted mitigation strategies to reduce exposure and improve the air quality in those areas. Moreover, the findings from quantifying population exposure can inform the design and operation of HEI buildings. It highlights the importance of considering airtightness and ventilation strategies to minimize $PM_{2.5}$ concentrations in areas where students, faculty, and staff spend significant amounts of time. This knowledge can guide building professionals in implementing effective measures to optimise IAQ and create healthier learning and working environments.

The analysis considered three scenarios based on the airtightness of the buildings: Baseline Q_{50}, Q_{50} = 7 $m^3/h/m^2$, and Q_{50} = 3 $m^3/h/m^2$. The microenvironments investigated in this study included PWE_Adm (Administrative areas), PWE_Aca (Academic areas), PWE_PGR (Postgraduate Research areas), PWE_Work (Workshops), PWE_Lect (Lecture rooms), PWE_Labs (Laboratories), PWE_Studio (Studios), and PWE_Study (Study areas). Table 7.9 presents the results of the population exposure to indoor $PM_{2.5}$ from outdoor sources for the three Q_{50} scenarios. For the Baseline Q_{50} scenario, the annual population exposure to $PM_{2.5}$ varied across

the microenvironments. The PWE_PGR exhibited the highest exposure with a value of 10.92 µg/m^3, followed by PWE_Adm with 10.12 µg/m^3, PWE_Lect with 10 µg/m^3, PWE_Study with 9.38 µg/m^3, PWE_Studio with 9.35 µg/m^3, PWE_Aca with 9.23 µg/m^3, PWE_Labs with 9.21 µg/m^3, and PWE_Work with 9.2 µg/m^3.

Table 7.9: Average Heating Season and Annual Indoor $PM_{2.5}$ Concentration (unit µg/m^3) in different ME for the baseline Airtightness Q_{50}, Q_{50} = 7 m^3/h/m^2, and Q_{50} = 3 m^3/h/m^2

Microenvironment		**Baseline Q_{50}**		**Q_{50} = 7 m^3/h/m^2**		**Q_{50} = 3 m^3/h/m^2**	
Main	**Sub**	Heating Season	Annual	Heating Season	Annual	Heating Season	Annual
Offices	Administration Offices	6.44 (±0.88)	10.12 (± 1.28)	6.08 (±0.97)	9.73 (± 1.31)	4.35 (±0.88)	7.42 (± 1.19)
	Academic Offices	5.80 (±1.37)	9.23 (± 1.66)	4.50 (±1.05)	7.58 (± 1.34)	3.38 (±1.37)	6.11 (± 1.20)
	PGR Offices	7.02 (±0.65)	10.92 (± 1.24)	6.02 (±0.51)	9.63 (± 1.01)	4.52 (±0.75)	7.67 (± 1.07)
Educational Facilities	Workshops	5.92 (±0.63)	9.20 (± 0.99)	5.09 (±0.87)	8.31 (± 1.17)	3.78 (±1.24)	6.63 (± 1.62)
	Lecture Halls	6.32 (±0.66)	10.00 (± 1.03)	5.31 (±0.77)	8.66 (± 1.08)	3.72 (±0.71)	6.59 (± 0.97)
	Labs	5.92 (±0.18)	9.21 (± 0.56)	5.05 (±0.74)	8.00 (± 1.39)	3.67 (±0.66)	5.91 (± 1.34)
	Studios	6.30 (±1.16)	9.35 (± 8.83)	4.91 (±0.67)	8.08 (± 0.94)	3.49 (±0.68)	6.22 (± 0.91)
Shared Facilities	Study and Computer Rooms	5.75 (±1.41)	9.38 (± 1.58)	4.60 (±1.42)	7.86 (± 1.71)	3.10 (±1.09)	5.80(± 1.50)

Figure 7.2 demonstrates the impact of improving Q_{50} on the population exposure to indoor $PM_{2.5}$ in different microenvironments. It can be noticed that improving the Q_{50} to represent buildings with moderate airtightness (Q_{50}= 7 m^3/h/m^2 scenario), there was a decrease in the annual population exposure to $PM_{2.5}$ in most microenvironments compared to the Baseline Q_{50} scenario. The PWE_Adm showed a 4% decrease to 9.73 µg/m^3 from 10.12 µg/m^3. This is attributed to most administrative spaces located in buildings with a baseline Q_{50} of 10 m^3/h/m^2. The , PWE_Aca decreased to 7.58 µg/m^3 with a reduction in population exposure of 1.75 µg/m^3 (19%). The population exposure in postgraduate research areas exhibited a 12% reduction with PWE_PGR = 9.63 µg/m^3. Further reductions were also exhibited in PWE_Studio, PWE_Study, PWE_Work, and PWE_Lect, with population exposures of 8.08, 7.86, 8.31, and 8.66 µg/m^3, respectively.

Further improvements on the building envelope airtightness value Q_{50} = 3 $m^3/h/m^2$ highlighted a more significant decrease in the annual population exposure to $PM_{2.5}$ compared to the previous scenarios. The microenvironments with the highest population exposure values in this scenario were PWE_PGR with 7.67 µg/m^3, PWE_Adm with 7.42 µg/m^3, and PWE_Lect with 6.59 µg/m^3. The microenvironments with the lowest population exposure values were PWE_Labs with 5.91 µg/m^3, PWE_Study with 5.8 µg/m^3, and PWE_Aca with 6.11 µg/m^3. These results demonstrate that as the airtightness of the buildings improved (from Baseline Q_{50} to Q_{50} = 3 $m^3/h/m^2$), the annual population exposure to $PM_{2.5}$ decreased across various microenvironments. This indicates that tighter building envelopes can contribute to reducing indoor air pollution and potentially improve indoor air quality.

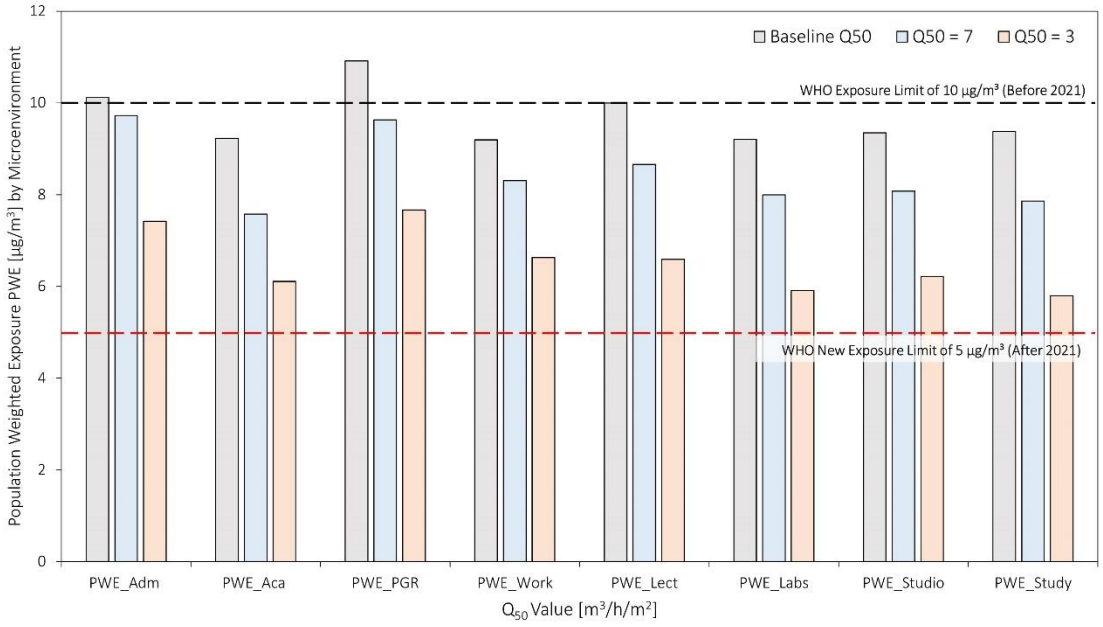


Figure 7.2: Population weighted exposure to indoor $PM_{2.5}$ from outdoor sources in different microenvironments for three scenarios of building airtightness Q_{50} values

Comparing the obtained population exposure values to the WHO annual exposure guidelines of 10 µg/m^3 (before 2021) and 5 µg/m^3 (after 2021) for $PM_{2.5}$ provides valuable insights into the potential health implications of IAQ within the studied microenvironments. In the Baseline Q_{50} scenario, the microenvironments of PWE_Adm, PWE_PGR, and PWE_Lect exceeded the WHO guideline of 10 µg/m^3, indicating a higher risk of adverse health effects for individuals occupying these spaces. However, it is important to note that the other microenvironments, including PWE_Aca, PWE_Work, PWE_Labs, PWE_Studio, and PWE_Study, exhibited population exposure levels below the 10 µg/m^3 threshold. In the Q_{50} = 7 $m^3/h/m^2$ scenario, the population

exposure values for all microenvironments fell below the WHO guideline of 10 µg/m^3, suggesting a lower overall health risk. Similarly, in the Q_{50} = 3 $m^3/h/m^2$ scenario, the population exposure values remained below the 10 µg/m^3 threshold for all microenvironments, indicating a further reduction in potential health risks.

Moreover, when comparing the obtained exposure values to the more stringent WHO guideline of 5 µg/m^3, it is evident that all microenvironments in the Baseline Q_{50} and Q_{50} = 7 $m^3/h/m^2$ scenarios surpassed this threshold. However, in the Q_{50} = 3 $m^3/h/m^2$ scenario, some microenvironments, such as PWE_Labs, PWE_Studio, and PWE_Study, demonstrated population exposure levels close to the WHO guideline of 5 µg/m3, suggesting a potentially lower health risk in these areas. These comparisons emphasize the importance of considering the WHO guidelines for $PM_{2.5}$ exposure in indoor environments and highlight the need for measures to further mitigate indoor air pollution. Implementing interventions to enhance building envelope airtightness, adopting appropriate ventilation strategies, and employing effective air filtration systems could help ensure compliance with the WHO guidelines and promote healthier indoor environments within higher education buildings.

It is important to consider the specific characteristics and conditions of each microenvironment when interpreting these findings. It is worth noting that the microenvironments classified as administrative and PGR spaces (PWE_Adm and PWE_PGR) rely on natural ventilation. These spaces exhibited relatively high population exposure values across all scenarios. These results indicate that if such improvements on the building envelope airtightness cannot be achieved, the use of mixed-mode ventilation systems could be necessary to mitigate indoor $PM_{2.5}$ concentrations in these areas. Additionally, it is important to note that the results for lecture halls (PWE_Lect) might overestimate the actual population exposure to $PM_{2.5}$. Lecture halls typically have dedicated mechanical ventilation systems that aim to dilute pollutants and maintain acceptable IAQ. However, this specific aspect was not part of the current study and warrants further investigation.

Nonetheless, the findings suggest that there is room for improvement in the IAQ of lecture halls within higher education buildings. Strategies to enhance the airtightness of these spaces or introduce more efficient ventilation systems, such as mixed-mode or mechanical ventilation, could potentially reduce population exposure to $PM_{2.5}$. These interventions may help ensure compliance with established guidelines, such as the World Health Organization (WHO)

recommendations for $PM_{2.5}$ exposure, which can contribute to the overall well-being and health of students, faculty, and staff. Further research and investigations are necessary to comprehensively explore the broader implications, potential interventions, and specific factors influencing indoor air quality and population exposure to $PM_{2.5}$ in each microenvironment within the studied setting.

7.4.2 Total Population-weighted Exposure to indoor $PM_{2.5}$

The annual population exposure to indoor $PM_{2.5}$ was determined using Eq (7.3), which accounts for the population size of the selected buildings. In the Baseline Airtightness Q_{50} scenario, the annual population exposure to $PM_{2.5}$ was determined to be 9.6 $\mu g/m^3$. This indicates that, on average, the individuals within the selected HEI buildings were exposed to an annual $PM_{2.5}$ concentration of 9.6 µg/m3. Under the Moderate Airtightness Q_{50} scenario, the annual population exposure to $PM_{2.5}$ decreased to 8.5 $\mu g/m^3$. This reduction in population exposure suggests that implementing moderate airtightness measures, with a Q_{50} value of 7 $m^3/h/m^2$, led to a decrease in indoor $PM_{2.5}$ concentrations and subsequent exposure levels for the population. Furthermore, in the Tighter Building Envelope Q_{50} scenario, the annual population exposure to $PM_{2.5}$ further decreased to 6.5 $\mu g/m^3$. This finding indicates that adopting a tighter building envelope with a Q_{50} value of 3 $m^3/h/m^2$ resulted in a significant reduction in indoor $PM_{2.5}$ concentrations and subsequent population exposure.

Comparing the obtained population exposure values to the WHO guidelines of 10 $\mu g/m^3$ provides further context and insights into the potential health implications. In the Baseline Airtightness Q_{50} scenario, the population exposure was almost equal to the WHO guideline of 10 $\mu g/m^3$, indicating a potential health risk. However, with the implementation of moderate airtightness measures (Q_{50}=7 $m^3/h/m^2$), the population exposure decreased, remaining below the 10 $\mu g/m^3$ threshold with an approximately 11.5% reduction. In the tighter building envelope Q_{50} scenario, the population exposure further decreased, falling below the 10 $\mu g/m^3$ WHO exposure limit. The reduction in population exposure compared to the Baseline Airtightness Q_{50} scenario was approximately 32.3%.

These results highlight the importance of implementing measures to improve the airtightness of HEI buildings. Transitioning from baseline airtightness to moderate or tighter building envelopes can lead to substantial reductions in population exposure to indoor $PM_{2.5}$. The findings emphasise

the potential health benefits associated with airtightness improvements and support the need for adherence to the WHO guidelines. In conclusion, the findings demonstrate the effectiveness of enhancing the airtightness of HEI buildings in reducing population exposure to indoor $PM_{2.5}$. Implementing moderate or tighter building envelopes can significantly contribute to achieving compliance with WHO guidelines and promoting healthier indoor environments within the HEI building stock.

Chapter 8 Discussion

In Chapter 3, a methodological framework was proposed to estimate the heating season infiltrated $PM_{2.5}$ concentration in a HEI building stock from a reduced set of input parameters. The outputs from the modelling framework were then used to provide estimates on the annual population-weighted exposure to indoor $PM_{2.5}$ for similar time-activity groups (STGs) and the whole population within a HEI building stock. The modelling methods and techniques were applied to five buildings selected from the UoS stock showing how the indoor $PM_{2.5}$ concentrations and exposures could be estimated. This chapter summarises and discusses the main outcomes and contributions of the research: [1] the data sources identified for the HEI Building Stock IAQ modelling (Chapter 3), [2] the IAQ-Thermal co-simulation approach and the model inputs (Chapter 4), [3] the sensitivity analysis framework and metamodels developed (Chapters 5 and 6), and [4] assessment of the microenvironmental modelling approach to estimate exposures to $PM_{2.5}$ (Chapter 7). Finally, the chapter address the research limitations and suggests recommendations for future work.

8.1 The data sources for institutional building stock IAQ modelling

Previous studies in building stock IAQ and energy modelling have shown that the obligatory data requirements are of two types: *data demand* and *data robustness* (Abdalla & Peng, 2021). *Data demand* specifies the scope, amount and type of input data required to achieve satisfactory prediction accuracy and consistency (G. Sousa et al., 2017b). For example, modellers have used different data sources to calculate or simulate the energy consumption attributable to the constituents of housing stock (Bennadji et al., 2022) or to estimate the population exposure to indoor air pollutants at a housing stock level (Das et al., 2014; Symonds et al., 2016). As these studies were targeted at the residential building stock, they relied on national population and housing censuses (e.g., English Housing Survey) that can provide essential statistical information on household details including the demographic and socioeconomic characteristics (*e.g.*, income,

education, employment status, age, gender, *etc*.), and building characteristics (*e.g.*, built-up area, number of rooms, number of storeys, fuel sources, heating/cooling systems, *etc*.). In cases where building data is unavailable from census surveys, records of *building permits* may contain information about building floor areas and building ages.

In comparison to residential building stocks, collating a representative database of buildings for a HEI building stock can be challenging due to lack of existent systematic survey data. There is a high level of heterogeneity often observed in HEI buildings in terms of their geometry, sizes, functions, constructions, and building uses. As such, this research was designed with *selected* buildings rather than *representative* buildings (i.e., archetypes) where the *statistical significance* of archetypes can be calculated. How to develop statistically representative archetypes of a HEI building stock is beyond the scope of this research. Here, the strategy was to work with an initial selection of buildings while collecting as much data as possible.

The Higher Education Statistics Agency (HESA) collects various self-reported statistics from the HEIs in the UK. Although HESA's primary aim is to collect information on university finance, students and staff, energy consumption data has also been collected since 2001/02 as part of the Estates Management System dataset. However, apart from the general statistics of HEIs in the UK, including age, gender, occupation, and level of study, HESA does not provide detailed information about HEI building stocks. In this research, the University's Estates and Facilities Management (EFM) was accessed to obtain data and information about building characteristic (e.g., geometries (areas and volumes), year of construction, ventilation method, materials and construction details (U-values), and heating policy (heating thermostat set point).

Although the data provided by the EFM was deemed sufficient for this research, there were various challenges that need to be addressed. First, the EFM does not contain an up-to-date record on all buildings as they go under refurbishments/interventions as part of the UoS Energy Policy. This could affect the quality and accuracy of the developed metamodels. Second, there is lack of consistency in the data collected among buildings, and assumptions in model inputs were unavoidable (e.g., baseline building envelope airtightness Q_{50}). This suggests that the EFM should develop new databases for compiling data/information of building properties consistently to enable accurate assessment and prediction of an HEI's indoor air quality performance.

8.2 Necessity of a Multi-zone Indoor air-thermal Coupling Approach

Indoor air pollution can be estimated with building simulations if sufficient data are unavailable, and several techniques are available for first-order approximations. Single-zone mass balance models represent all indoor spaces within a building as a single volume of air, and they are the simplest models of building IAQ (Jung et al., 2011). These first-order techniques have been used to examine the infiltration of outdoor $PM_{2.5}$ (Fazli et al., 2021; Logue et al., 2015; Rosofsky et al., 2019), the efficacy of efforts to dilute airborne contaminants of indoor origin (L. Ng et al., 2021), and to predict impacts of large-scale planning efforts (Abdalla & Peng, 2021). Single-zone models, however, are less suitable for evaluating and predicting IAQ-related health impacts since they do not capture significant heterogeneity caused by building-specific factors that could drive significant variations in indoor concentrations.

In contrast, multi-zone models better represent the compartmentalised nature of indoor spaces (Abdalla & Peng, 2021) and offer the ability to evaluate exposure and health impacts of specific IAQ interventions (Lindsay J. Underhill et al., 2020). As shown in Chapter 4, the multi-zone IAQ-Thermal coupled simulation approach can now be implemented through the freely available CONTAM-EnergyPlus toolset. While the coupled CONTAM-EnergyPlus models of HEI buildings allow for simultaneous thermal, airflow and contaminant transport simulations, it also revealed several issues. One key issue identified here is the impact of building envelope airtightness Q_{50} value on zonal infiltration ACH_{INF} and the concentrations indoor $PM_{2.5}$.

As shown in Table 4.7, the ACH values ranged from 0.28 – 5.02 h^{-1} during the heating season, corresponding to buildings with a Q_{50} of 13 $m^3/h/m^2$. The differences in ACH then translated into the differences in indoor $PM_{2.5}$ concentration levels, ranging from 2.33 – 9.24 $\mu g/m^3$ with the highest concentration corresponding to an ACH of 5.02 h^{-1} (Figure 4.12). Therefore, when simulating IAQ in a building, the envelope leakage rate must be carefully selected, as it can have significant impact on the predicted airflows and infiltrated $PM_{2.5}$. In light of the limited data available on building envelope leakage in the HEI building stock, the selection of these values presents a significant challenge for both the analysis of air quality and that of airflow. Although constant infiltration airflow rates have been used in energy simulation, it is unlikely to account for the effects of weather on infiltration. Thus, treatment of infiltration that is more dynamic based should be applied.

A multi-zone airflow and indoor air quality model such as CONTAM considers buildings as interconnected networks. The airflow rates are then calculated according to the relationship between flow and pressure, similar to the relationship between heat transfer and temperature differences in thermal models. In this way, multi-zone building airflow models can more accurately estimate the pressure relationships between different building zones, which are influenced by (a) building geometry, (b) exposure to the ambient environment, (c) Interzone leakages, and (d) exhaust fan airflows. On the other hand, a building energy/thermal model takes into account thermal loads in different building zones, system efficiency in meeting these loads, and types and sizes of equipment. Energy models generally define zones based on their similarity and differences in thermal loads, despite the importance of building geometry, exposure to the outside, and fan airflows in energy calculations. As a result, these thermal zones alone may not provide an adequate model of building airflows.

Hence, there is a clear case of support for coupled multi-zone IAQ-Thermal models (as in CONTAM-EnergyPlus coupling) for estimating indoor $PM_{2.5}$ concentration. The most appropriate coupling method will depend on the degree of coupling of the airflow-thermal problem. The more prominent indoor airflow-thermal interaction, such as in naturally ventilated buildings where large temperature gradients may exist and are essential drivers of airflow, the more sophisticated the coupling method will need to be. It was evident from this research that it is necessary to adopt a coupled simulation approach in the context of UK HEI building stocks.

The results of the sensitivity analysis (Section 6.2) show that the indoor – outdoor temperature differences (ΔT, $^{\circ}C$) during the heating season has a strong negative non-linear relationship with a $S_{Pears} = -0.46$ and $S_{Spear} = -0.57$ with the concentrations of infiltrated $PM_{2.5}$ for all tested Q_{50} values ($3 \leq Q_{50} \leq 13$ $m^3/h/m^2$). In UK university buildings, the indoor temperature is typically controlled by an institutional heating policy that centres on the outdoor air temperature. As such, to capture the dynamic interaction between indoor-outdoor temperature differences and outdoor/indoor $PM_{2.5}$, an IAQ-Energy coupled multi-zone simulation approach is necessary and achievable.

8.3 Implications for Non-Domestic Building Stock IAQ Modelling

This study suggests that higher education institutions (HEIs) in the UK and abroad could deliver building planning and design strategically towards improving IAQ. First, a university's facility management can coordinate efforts to collect building and environmental data for IAQ modelling. An advantage like this is challenging to realise in a building stock with many owners. Second, a university typically owns and manages a sizable number of diverse and complex buildings (e.g., a current portfolio/stock of 85 academic buildings at the University of Sheffield). Therefore, indoor $PM_{2.5}$ concentrations and exposures in a HEI building stock can exhibit substantial variations not seen in housing stock IAQ studies. Third, to address fully the influences of diverse (heterogeneous) factors and activities on IAQ, the scope of targeted performance indicators (e.g., indoor $PM_{2.5}$) and potential interventions (e.g., likely parameters of building retrofitting) should be reviewed regularly to inform HEI stock IAQ model development.

8.3.1 Spatial and Temporal Variations in indoor $PM_{2.5}$ Concentrations

Based on the results of the coupled simulation of the 4 selected buildings within a HEI building stock of the UoS, it was evident that there was spatial variability in infiltrated $PM_{2.5}$ concentrations across different rooms/spaces within each individual building. In Section 4.3.1, the analysis of the indoor $PM_{2.5}$ time-series data across randomly selected zones supported the use of a high spatial resolution modelling approach. Previous research found that individuals working within the same building will be exposed to varying levels of indoor pollution based on the patterns of their daily activities (Elliot et al., 2000). Ferguson et.al., found disparity in exposure to indoor pollutants based on socio-economic groups, highlighting the importance of investigating indoor air contaminant concentrations across different building types and settings, rather than isolated examples. However, these studies omit the impact of building characteristics on the spatial variation of indoor $PM_{2.5}$ concentrations. More recently, (Milando et al., 2022b) examined the influence of building characteristics, HVAC type, and model resolution on indoor exposure to $PM_{2.5}$ in different housing typologies in Boston. The results indicated that variations in model resolution (room or floor-level model resolution) modified the differences in indoor-sourced $PM_{2.5}$ exposure.

In contrast, the same study found that model resolution (e.g., single zone model) may adequately approximate indoor $PM_{2.5}$ exposures from outdoor-sourced $PM_{2.5}$. However, their findings might be accepted when modelling the IAQ of domestic building stocks. In fact, previous research on

housing stock IAQ models used low resolution models i.e., single zone models (Fazli & Stephens, 2018) , or simplified multizone models (Das et al., 2014). However, in a HEI building stock, it was evident from the findings that high resolution model was required for two main reasons: 1) the scale and geometry of HEI buildings cannot be compared to that of residential buildings. 2) The representation of several building physics phenomena in the residential housing stock and its impact on the estimated indoor air pollution levels had been considered adequate (e.g., not considering inter-zonal airflow). However, in HEI buildings, the complexity of an airflow network and the level of details required to adequately model the IAQ of HEI building stock is unknown.

In fact, there had been limited attempts to model the IAQ of a non-domestic building stock. One recent study that can be acknowledged here is the UK Classrooms Archetype Stock Model (Schwartz et al., 2021). that was conducted to model the IAQ and overheating in UK Classrooms using an archetype stock modelling approach. However, in this study, the developed classroom archetypes were represented by thermal zones rather than detailed airflow networks. As such, this disregards how buildings interact with influential factors on IAQ such as ambient weather conditions (outdoor/indoor temperature difference, wind speed and direction), which increases infiltration. Additionally, the spatial variability of indoor air pollution as a result of building characteristics was not captured. Lastly, a detailed evaluation of the model's sensitivity to various input variables was not carried out.

Here the novelty of this research was highlighted by the development of a high-resolution HEI building stock IAQ model to account for the spatial variability in infiltrated $PM_{2.5}$ concentrations during the heating season. It is apparent that the number and type of input parameters in the metamodel reflects the spatial variability in infiltrated $PM_{2.5}$ concentrations as a function of a zone's morphological and indoor environmental characteristics. Here, the identification of the key parameters becomes relevant for room/space under investigation within a HEI building stock. However, due to the selection of a model resolution of individual rooms/spaces in the metamodels development process there was variability in the morphological and indoor environmental parameters between the spaces. To quantify this variability and identify the key parameters influencing the indoor $PM_{2.5}$ concentration levels, a sensitivity analysis framework was applied.

8.3.2 Simulated Average Indoor $PM_{2.5}$ Concentrations and I/O Ratios

The results obtained from the three different scenarios of Q_{50} (Baseline Q_{50}, Q_{50} = 7 $m^3/h/m^2$, and Q_{50} = 3 $m^3/h/m^3$) provide valuable insights into the average $PM_{2.5}$ I/O ratios during the heating and non-heating seasons. As discussed in Section 4.3.4, the I/O ratio represents the ratio of indoor $PM_{2.5}$ concentrations to outdoor $PM_{2.5}$ concentrations, indicating the extent to which outdoor pollutants infiltrate indoor environments or determines indoor sources strengths. The findings highlight the significance of infiltration rates (ACH_{INF}) and outdoor $PM_{2.5}$ concentrations in determining the IAQ. Moreover, the average indoor $PM_{2.5}$ concentrations during the heating season compared to the non-heating season further reinforce the importance of these factors.

Based on the results presented in Table 4.8, during the heating season (November - April), when the primary source of airflow is infiltration, the $PM_{2.5}$ I/O ratio is lower for the Q_{50} = 3 $m^3/h/m^2$ scenario (0.23 ±0.06) compared to the Baseline Q_{50} scenario (0.37 ±0.06), see Figure 8.1. This indicates that reducing the infiltration rate significantly decreases the penetration of outdoor $PM_{2.5}$ particles, resulting in lower indoor concentrations. The average indoor $PM_{2.5}$ concentrations align with these findings, showing an apparent decrease from the Baseline Q_{50} scenario (6.30 ±1.07 µg/m³) to the Q_{50} = 3 $m^3/h/m^2$ scenario (3.94 ±0.98µg/m³); see Figure 8.2. Conversely, during the non-heating season (May - October), when natural ventilation (P = 1.0, Section 3.3.7) is the primary source of airflow, the $PM_{2.5}$ I/O ratios remain relatively stable across the scenarios. However, it is crucial to note that the average indoor $PM_{2.5}$ concentrations are lower during the non-heating season than the heating season for all scenarios. This implies that although natural ventilation contributes to an increased air change rate, the impact of outdoor $PM_{2.5}$ concentrations on IAQ is lower during this season. The average indoor $PM_{2.5}$ concentrations support this observation, with lower values during the non-heating season across all scenarios.

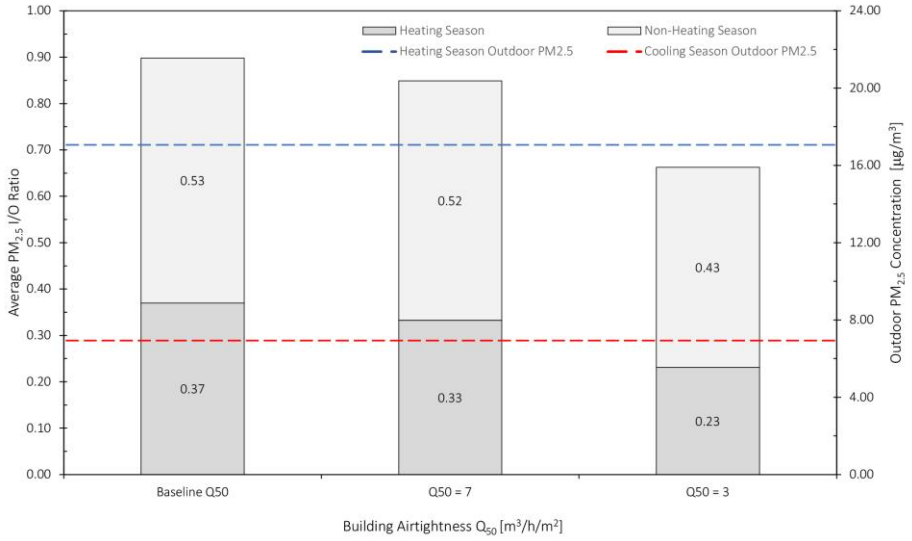


Figure 8.1: Average seasonal $PM_{2.5}$ I/O ratios under three scenarios of Q_{50}, and seasonal variation in outdoor $PM_{2.5}$ concentration [$\mu g/m^3$]

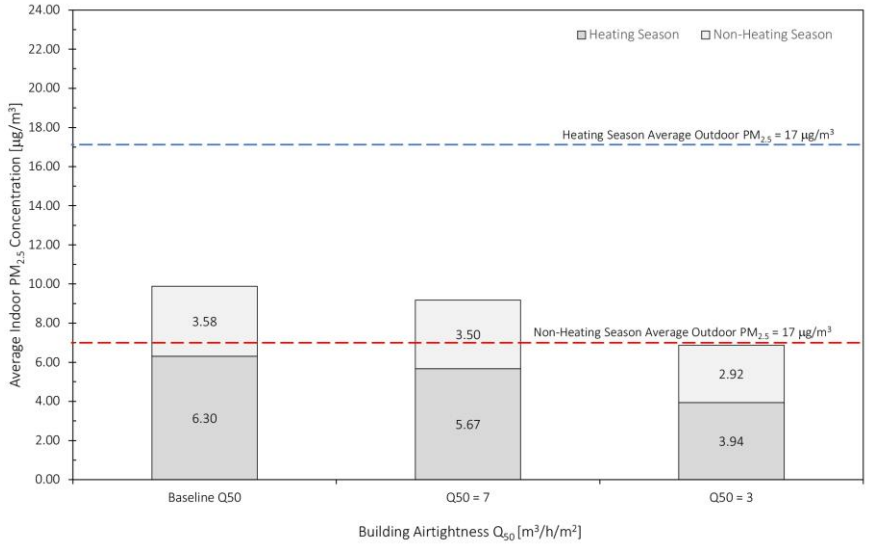


Figure 8.2: Average seasonal indoor $PM_{2.5}$ concentration [$\mu g/m^3$] under three scenarios of Q_{50}, and seasonal variation in outdoor $PM_{2.5}$ concentration [$\mu g/m^3$]

Considering the $PM_{2.5}$ I/O ratios and average indoor $PM_{2.5}$ concentrations, it becomes apparent that the heating season is the critical period to focus on for mitigating indoor $PM_{2.5}$ concentrations. The higher $PM_{2.5}$ I/O ratios during the heating season indicate a more significant infiltration of outdoor pollutants into indoor spaces, leading to elevated indoor $PM_{2.5}$ concentrations. Moreover, for all scenarios, the average indoor $PM_{2.5}$ concentrations during the heating season are consistently higher than those during the non-heating season. Therefore, efforts to improve IAQ

should prioritise measures that reduce infiltration rates and lower outdoor $PM_{2.5}$ concentrations, particularly during the heating season. This could include enhancing building envelope insulation, implementing air filtration systems, and minimising air leakage. Additionally, initiatives aimed at reducing outdoor $PM_{2.5}$ emissions and improving outdoor air quality are crucial for limiting the impact of outdoor pollution on indoor environments.

The results highlight the relationship between infiltration rates (ACH_{INF}), outdoor $PM_{2.5}$ concentrations, and IAQ during the heating and non-heating seasons. The findings underscore the need to focus on the heating season due to the higher concentrations of outdoor $PM_{2.5}$ outweighing those during the non-heating season. Lowering infiltration rates, reducing outdoor $PM_{2.5}$ concentrations, and promoting IAQ measures are essential to ensure healthier indoor environments, particularly during the heating season when occupants are more exposed to indoor pollutants.

Comparing $PM_{2.5}$ I/O to the findings of other UK and international studies provides valuable insights into the consistency and relevance of results. However, without relevant studies on $PM_{2.5}$ I/O ratios in HEI buildings, findings from studies on residential and office buildings are used. In relation to other UK studies, it has been reported that indoor sources can contribute significantly to indoor $PM_{2.5}$ concentrations in residential settings, with some studies even finding $PM_{2.5}$ I/O ratios close to or greater than one (Jones et al., 2000; Lai et al., 2006). Internationally, studies conducted in European countries have indicated $PM_{2.5}$ I/O factors ranging from 0.30 to 0.70 (Hanninen et al., 2011), while broader international studies have reported values between 0.30 and 0.82 (Chen and Zhao, 2011). These findings align reasonably well with the simulated $PM_{2.5}$ I/O ratios, which estimated values of 0.42 to 1.16 (0.83 ± 0.13) for Scenario 1 (Q_{50} =7 $m^3/h/m^2$) and 0.26 to 1.03 (0.66 ± 0.11) for Scenario 2 (Q_{50} = 3 $m^3/h/m^2$).

Examining the relationship between indoor and outdoor $PM_{2.5}$ concentrations in office environments, the statistical behaviour derived from the CoSIM results for different Q_{50} values are in agreement with existing evidence suggesting that outdoor air serves as the primary source of particles in office environments, while indoor sources may make a minor contribution in some instances (Matson, 2005; Morawska et al., 2017). The reported I/O ratios in office buildings have varied from 0.10 to 1.35, with a mean $\pm\sigma$ value of 0.62 ± 0.24 across all buildings (Zhu et al., 2015). Notably, Zhu et al. (2015) observed $PM_{2.5}$ I/O ratios of 0.44 and 0.62 for the summer and autumn seasons, respectively, supporting that outdoor $PM_{2.5}$ sources are the major contributors to

indoor $PM_{2.5}$ levels. Consistent with these findings, other relevant studies have reported average $PM_{2.5}$ I/O ratios ranging from 0.4 to 0.9 in an office room in Beijing (Zhao et al., 2015), a ratio of 0.86 in an office in Xi'an, China (Niu et al., 2015), and an I/O ratio of 0.62 ± 0.14 in a study encompassing four offices in Milan, Italy (Sangiorgi et al., 2013). Consequently, the present results demonstrate satisfactory consistency with previous office-based investigations.

8.3.3 Sensitivity Analysis

The sensitivity analysis framework applied in this work followed the work in (Das et al., 2014) however, used deterministic inputs rather than following a probabilistic sampling of input variables. However, in order to apply the sensitivity analysis framework on a HEI building stock, some modifications and assumption were required. First, in this research the focus on the spatial variation of infiltrated $PM_{2.5}$ concentrations caused by a single factor, while neglecting the interactions between the factors. Second, the results of the co-simulations were resampled to a seasonal and annual temporal resolution. The reason for this was to allow for a better pairing of the simulated results with several zonal morphological and indoor environmental characteristics (see Table 4.4) As such, it was assumed that the variation of infiltrated $PM_{2.5}$ concentrations across the heating season and non-heating season was sufficient to capture the temporal variation in this research.

The sensitivity analysis framework was used to determine the relationships between each of the inputs and the outputs. Scatter plot of inputs versus the output illustrate the relationships between the individual inputs and the output for visual inspection. They are shown to be highly complex, and so the type of correlation becomes more difficult to interpret; see Figure 6.2. Nevertheless, the results were yet useful for identifying the inputs that are more important and more related for the development of the metamodels (see Tables 6.1-6.3) . *p*–values can be interpreted for significance, and so testing the relationship between the datasets. Given the relatively small sample size (n=2729 zones), this interpretation can be meaningless, since the chance of finding significance increases with the sample size (Gigerenzer, 2004). Furthermore, the statistical significance of *p*–values is arbitrary, and so in the sensitivity analysis framework, the focus is on the nature and the magnitude of the effects (Fenton & Neil, 2018). In consequence, for reporting *p*–values, the exact level of significance was given rather than its interpretation.

Comparing the various sensitivity analysis methods for identifying key explanatory variables is particularly essential when input and output variables have non-linear correlations. It was evident

from the sensitivity analysis framework that the relationship between the infiltrated $PM_{2.5}$ concentrations and several influential parameters can be characterised as monotonic or non-linear. Here, direct comparison of the sensitivity analysis results to previous results in other building stock studies is challenging, given the context of stock under investigation (e.g., a HEI building stock), and the potential differences in building structures and behavioural patterns. This said, there was some similarities in findings across the literature. For instance, the infiltration ACH_{INF} has been associated with the outdoor sourced $PM_{2.5}$ levels in indoor environments (Das et al., 2014; Wichmann et al., 2010). Additionally, indoor/outdoor temperature difference ΔT and the scaled wind speed v can lead to variations in infiltrated $PM_{2.5}$ (Jonathon Taylor et al., 2015). Not only are they influential in the settlement of particles (lower concentrations in indoor air) (Lv et al., 2017; Zhang et al., 2019) but as thermal comfort indices, their variation will also alter users behaviour (Zhang et al., 2019). Another similarity was the area of exposed surface to internal volume area (J. Taylor et al., 2014b).

Alongside these findings, the process proved to be a useful technique, at least in regard to exploratory analysis, to help identify key features and potential relationships for the selection of candidate ML algorithms as there were disparities in the shape and magnitude of the relationships between the concentrations of infiltrated $PM_{2.5}$ and several explanatory parameters (e.g., linear/non-linear and positive/negative). In fact, it became evident for a HEI building stock the need to combine sensitivity analysis methods with the development of metamodels to capture the non-linear and non-monotonic relations between the variables and to be able to reproduce them.

8.4 Development of a Heating Season Metamodel for IAQ

This research proposed a metamodeling framework to rapidly estimate the spatial variations of infiltrated $PM_{2.5}$ concentrations in an HEI building stock from a set of key explanatory variables.. A metamodel can also reduce the number of inputs needed to determine the desired outputs if combined with sensitivity analysis. This study examined three metamodels: the GAM, the RF, and the XGBoost, which can reproduce nonlinear and non-monotonic relationships between inputs and output (indoor $PM_{2.5}$ concentrations). In this study, the metamodels were developed to account for the spatial variability in infiltrated $PM_{2.5}$ concentrations during the heating season by decomposing buildings selected from an HEI building stock into a structured cohort of individual spaces or rooms. The findings from the Co-simulation underscore the need to focus on the heating

season due to the higher concentrations of indoor and outdoor $PM_{2.5}$ outweighing those during the non-heating season.

Developing a metamodel to predict the concentrations of infiltrated $PM_{2.5}$ during the heating season is important in the context of IAQ assessment and building envelope design. The findings from this research, where a metamodel was constructed using the XGB algorithm and a reduced set of parameters, provide valuable insights into the implications of increasing the airtightness of a building envelope Q_{50} on the concentrations of infiltration $PM_{2.5}$. The metamodel was trained and tested using simulated data generated through the coupled CONTAM-EnergyPlus simulation approach, considering a parameter space of Q_{50} values ranging from 3 to 13 $m^3/h/m^2$. The results of the sensitivity analysis highlighted Q_{50}, infiltration air change rates ACH_{INF}, indoor/outdoor temperature difference ΔT, scaled wind speed v, and Aef:V as the most important parameters influencing the infiltrated $PM_{2.5}$ concentrations.

The prediction capability of the developed metamodel was assessed by testing the impacts of changing Q_{50} from 7 to 3 $m^3/h/m^2$ using a holdout dataset (Section 6.5). The results demonstrated a high level of accuracy, with an R^2 value of 0.91 and a model accuracy of 90.6% in predicting the concentrations of infiltrated $PM_{2.5}$. These findings have several important implications.

First, the metamodel offers a valuable tool for predicting and understanding the behaviour of infiltrated $PM_{2.5}$ concentrations in buildings during the heating season. By considering a reduced set of parameters identified through sensitivity analysis, the metamodel provides a simplified yet effective approach to estimate the impact of changing airtightness levels Q_{50} on IAQ. This capability is crucial for stakeholders such as university EFM, policy makers, and engineers, who can utilise the metamodel to assess the potential consequences of different building envelope airtightness values and make informed decisions regarding building design, maintenance, and energy efficiency measures.

Second, the identified influential parameters contribute to a deeper understanding of the factors affecting infiltrated $PM_{2.5}$ concentrations. This knowledge enhances the ability to control and manage IAQ effectively, particularly in a complicated and heterogeneous building stock. For example, policymakers can use this information to develop regulations and guidelines that address the most critical parameters, leading to improved IAQ standards and healthier indoor environments.Third, the high accuracy and predictive capability of the metamodel validate its

robustness and reliability in estimating infiltrated $PM_{2.5}$ concentrations. This reliability allows for the exploration of plausible scenarios beyond the observed range of Q_{50} values during model training. Users can rely on the metamodel to evaluate the effects of extreme or hypothetical situations, assisting in the development of innovative building designs and the assessment of potential regulatory changes.

Overall, the development of the metamodel and the interpretation of its findings underscore the importance of considering airtightness and related parameters when addressing IAQ issues. The metamodel provides a valuable tool for stakeholders to predict and investigate the impacts of changing building envelope airtightness Q_{50} on infiltrated $PM_{2.5}$ concentrations. This knowledge can guide decision-making processes, promote energy-efficient building designs, and contribute to the creation of healthier and more sustainable indoor environments.

8.5 Interpretation and Explanation of Metamodels

A key contribution of this thesis was the application of SHAP values to increase the transparency of the metamodels and statistically quantify the contribution of each input variable to the predicted indoor $PM_{2.5}$ concentration levels. Here the significance of this is highlighted by answering the research questions 1 and 3 listed in Section 1.4.

Previous studies showed that the SHAP values yielded more reliable results than other measures (e.g. feature importance and Gini importance measure) (Aldrich, 2020; Gu et al., 2021). In fact, the application of SHAP in the domain of IAQ is limited. As such this research is considered the first attempt to apply the SHAP values on the domain of IAQ. Moreover, previous attempts to model the IAQ of building stocks have relied on the coefficients of linear regression to highlight the importance of features to indoor $PM_{2.5}$ levels (Molina, Jones, et al., 2020). This could be assumed acceptable when the relationships between the inputs and outputs are linear. However, following the application of the sensitivity analysis framework, the priority of variables was represented by ranking them indicating most influential to least influential based on the correlation coefficients and significance p-value. It was impossible to quantify the contribution of each variable to the predicted infiltrated $PM_{2.5}$ concentrations using these coefficients for one particular reason. It was clear that the relationships were non-linear, i.e., the changes in the infiltarted $PM_{2.5}$ concetrations do not change in direct proportion to changes in any of the inputs.

Other studies have used ANNs, Support Vector Machine (SVM) and Radial Basis Functions (RBF) algorithms to model the IAQ in the residential building stock (Das et al., 2014; Symonds et al., 2016). Although these algorithms can capture the non-linearities in the relationships between the inputs and outputs. They are black box models, in which the contribution of each input variable to the predicted outcome cannot be quantified. Here, the importance of applying the SHAP values in this research was highlighted by quantifying the global and local importance of a feature to the infiltrated $PM_{2.5}$ concentrations. Additionally, the SHAP values were used to quantify the variation of infiltrated $PM_{2.5}$ concentrations as a result of varying an individual variable. In Table 6.8 the calculated percentage of variations in infiltrated $PM_{2.5}$ concentrations highlighted the key variables that can inform future plans for enhancing the IAQ in a HEI building stock. However, it is important to highlight that while SHAP shows the contribution or the importance of each feature on the prediction of the model, it does not evaluate the quality of the simulated results or the prediction itself.

8.6 Microenvironment Modelling for Exposure Assessment

Total cumulative exposures are determined by the concentration of a pollutant at a particular location i (also known as a microenvironment) and the amount of time spent there ($C_i\ t_i$). The basis for the time-activity profiles for each population group in an HEI building stock was assumed to define a set of time-activity profiles for sub-populations based on the HESA data. It was thus possible to explore the impact of contrasting time-activity patterns on personal exposure to $PM_{2.5}$ from outdoor sources. Despite the fact that time-activity patterns are much more variable within each group in practice, the time-activity profiles presented in this thesis are mainly intended to serve as a demonstration. Here, time-activity profiles were developed for four groups using the Similar Time Activity Groups (STGs) defined in Section 7.2**.**

In order to determine whether population exposure to indoor $PM_{2.5}$ in HEI building stock can be reduced, it is necessary to be able to determine the factors influencing indoor $PM_{2.5}$ concentrations, in the same manner as these factors have been assessed for housing stocks. However, there is a substantially smaller volume of data assessing population exposure to $PM_{2.5}$ in HEI building stocks. As such, this research uses the infiltrated $PM_{2.5}$ concentrations produced by the XGBoost metamodel added to the CoSIM non-heating season indoor $PM_{2.5}$ concentration to quantify the population exposure to indoor $PM_{2.5}$ as a result of improving the Q_{50} of the

buildings. The analysis considered three scenarios based on the airtightness of the buildings: Baseline Q_{50}, Q_{50} = 7 $m^3/h/m^2$, and Q_{50} = 3 $m^3/h/m^2$

The outcomes of the research demonstrate how changes in airtightness Q_{50} can affect the relative exposures of occupants within higher education institution (HEI) buildings, specifically in terms of annual indoor $PM_{2.5}$ concentrations. Through comparative analysis of distinct scenarios, it is conceivable to ascertain the extent to which the safeguarding intervention of modifying airtightness, as denoted by the manipulation of the Q_{50} value, can foster healthier indoor environments or potentially yield adverse impacts on occupants within university buildings.

The results indicate that implementing measures to improve the airtightness of HEI buildings can significantly impact population exposure to indoor $PM_{2.5}$. In the Baseline Airtightness Q_{50} scenario, the annual population exposure to indoor $PM_{2.5}$ was determined to be 9.6 µg/m3, which is close to the WHO guideline of 10 µg/m3. This finding suggests that occupants within the selected HEI buildings may be at potential health risk due to elevated $PM_{2.5}$ levels. However, with the implementation of moderate airtightness measures (Q_{50}=7 $m^3/h/m^2$), the annual population exposure to $PM_{2.5}$ decreased to 8.5 µg/m^3, representing an approximately 11.5% reduction in exposure compared to the Baseline Airtightness Q_{50} scenario (Section 7.4.2). This reduction in exposure indicates that implementing moderate airtightness measures would help occupants by lowering their exposure to indoor $PM_{2.5}$ concentrations, moving them further away from potential health risks.

Furthermore, in the Tighter Building Envelope Q_{50} scenario, the annual population exposure to $PM_{2.5}$ decreased even further to 6.5 µg/m3. This reduction represents a significant decrease in exposure compared to both the Baseline Airtightness Q_{50} and Moderate Airtightness Q_{50} scenarios, with a reduction of approximately 32.3% compared to the Baseline Airtightness Q_{50} scenario. These findings highlight the effectiveness of adopting a tighter building envelope, indicated by a lower Q_{50} value of 3 $m^3/h/m^2$, in reducing population exposure to indoor $PM_{2.5}$ concentrations. In light of the stricter target set by the WHO of 5 µg/m3 for $PM_{2.5}$ concentrations, it is noteworthy to mention that the findings of this study suggest that relying solely on tighter Q_{50} values may not be sufficient to meet this stringent guideline. While implementing a tighter building envelope with reduced Q_{50} values resulted in significant reductions in population exposure to indoor $PM_{2.5}$, as discussed previously, achieving compliance with the WHO target of 5 µg/m^3 may require additional interventions or strategies beyond airtightness improvements

alone. For example, integrating advanced air purification strategies, such as portable air purifiers, into indoor spaces can further reduce indoor $PM_{2.5}$ concentrations (Fermo et al, 2021). This observation underscores the complexity of addressing indoor $PM_{2.5}$ concentrations and highlights the importance of adopting a comprehensive approach that encompasses multiple factors and mitigation measures to ensure the attainment of stricter air quality standards.

Comparing the exposure findings in the this study (HEI Building Stock) under the baseline Q_{50} scenario to the results presented by the INDAIR probabilistic mode (Dimitroulopoulou et al., 2006) provides valuable insights into the indoor $PM_{2.5}$ concentrations and their sources in different environments. In the this study, the annual population exposure to indoor $PM_{2.5}$ was determined to be 9.6 µg/m_3 in the Baseline Airtightness Q_{50} scenario. This finding indicates a moderate level of exposure, influenced primarily by infiltration of outdoor $PM_{2.5}$ investigated in the study. In contrast, the study by Dimitroulopoulou et al. (2006) focused on UK residential settings and reported an estimated annual indoor average $PM_{2.5}$ concentration of 19.78 mg/m^3. This concentration was significantly higher compared to both the outdoor $PM_{2.5}$ levels (13.0 mg/m3) and the exposure levels observed in the HEI building stock. Another study by (Shrubsole et al., 2012) reported that indoor $PM_{2.5}$ levels derived from outdoor air (excluding indoor sources) were less than half of the corresponding outdoor levels in the London housing stock. However, the average household member experiences an estimated an annual average indoor $PM_{2.5}$ concentration of 28.4 mg/m^3, which was over twice the concentration observed in outdoor air (13.0 mg/m3). In the context of a 2050 refurbishment scenario in their study (Q_{50} reductions to 3 $m^3/h/m^2$ and the utilisation of properly installed and optimally functioning MVHR equipment) resulted in a reduction in household average annual exposure to total $PM_{2.5}$ (from indoor and outdoor sources) from 28.4 mg/m^3 to 9.6 mg/m^3.

The comparison between the the findings in this study and the aforementioned studies highlights the variation in indoor $PM_{2.5}$ concentrations across different environments and building types. The findings of this study indicate relatively lower exposure levels in a HEI building stock compared to residential building stock. However, it is important to consider that the effectiveness of the implemented interventions, such as Q_{50} reductions may vary depending on the specific building characteristics and occupant activities. Further investigations are warranted to explore the effectiveness of these interventions in various building types and to assess their applicability in achieving the desired exposure reductions, particularly when aiming to meet stricter guidelines such as the WHO standards of 5 µg/m3 for $PM_{2.5}$.

8.7 Limitations and Further Work

There are a number of limitations in this research which points to the areas for future work. First, conventional methods of building engineering calculation tend to be deterministic, predetermined values (defaults) are often used without tackling uncertainty (Panagopoulos et al., 2011). In general, uncertainties in building stock modelling are of three sources: (1) the heterogeneity within a building stock (e.g. an extensive range of building characteristics), (2) the first-order or *aleatoric* uncertainties where different simulation outputs are probable given the same building, and (3) the second-order or *epistemic* uncertainties where input parameters can take different values in light of new data or knowledge (Ferguson et al., 2020).

Increasingly, uncertainty quantification has been introduced to housing stock energy and IAQ modelling. Based on generating distributions of predictions followed by sensitivity analyses, Das, Shrubsole, Jones et al. developed a probabilistic framework for quantifying uncertain parameters in housing stock IAQ modelling (Das et al., 2014). More recently, Molina et al. applied a similar framework to the Chilean housing stock to quantify the uncertainties in indoor pollutant concentration levels, ventilation, and infiltration (Molina, Jones, et al., 2020). However, these frameworks were specific to the housing stock studies. In this thesis, a deterministic modelling framework was applied that does not consider probable fluctuations of some input parameters of any initial conditions and the solution was one and only (Renard et al., 2013). In contrast, stochastic models attempt to quantify some or all of the parameters by probabilistic distributions rather than single assumed definitive values.

Das et al. showed that uncertain input data could be contaminants related, such as ambient concentrations, generation rates, and deposition rates (Das et al., 2014). Booth et al. stated that any building stock model should provide information about the potential risks associated with proposed interventions by displaying a distribution of confidence levels due to the diverse sources of uncertainty (Booth et al., 2012). To do so, mathematical and statistical methods are available for evaluating uncertainties in model inputs and outputs (Molina, Kent, et al., 2020). Table 8.1 summarises the sources of uncertain input parameters in the building stock IAQ modelling literature. To improve the quality of predictions in HEI stock IAQ modelling, the issue of uncertainty needs to be addressed.

Table 8.1: Sources of uncertain input parameters in building stock IAQ modelling.

Sources	Descriptive Parameters	Key References
Environment and Climate	Spatial-Temporal Variations of Ambient Contaminant Concentrations, Wind Speed and Direction, Local Outdoor Temperature and Terrain Properties	(Dias & Tchepel, 2018; Benjamin Jones et al., 2015; Malkawi & Augenbroe, 2004)
Physical Characteristics	Building Geometry and Layout (e.g., Block Aspect Ratio), Space/Zone Volume, Material Properties, Orientation, Building/Room Height, and Number of Exposed Facades	(G. Sousa et al., 2017a) (Das et al., 2014)
Building Physics	Zone Pressure, Local Zone Temperature, Air Temperature Stratification, Wind Pressure Coefficients, Building Envelope Airtightness, Ventilation and Infiltration Rates, Flow Path Discharge Coefficient, Airflow Exponent *n*, and Flow Path Area	(Booth et al., 2012; Cóstola et al., 2008; Herring et al., 2016; Benjamin Jones et al., 2015; Molina, Jones, et al., 2020)
Building Components / Systems	HVAC Runtimes, HVAC Supply and Return Flow Rates, Air Exchange Rates (AER), Filter Efficiency and Removal Rate, Combustion Sources and Emission Rates	(Benjamin Jones et al., 2015)
Occupants and Activity	Building Population, Time-Activity-Location Factor, and Occupancy Schedules (HVAC Runtime, Window Opening Area and Time)	(Ben & Steemers, 2018)
Contaminant Properties	Contaminant Generation Rates (e.g., emission rate), Source Strength, Contaminant Sinks, Contaminant Penetration Factor, and Deposition Rates	(Dimitroulopoulou et al., 2006)

Secondly, there are input parameters in IAQ modelling that can vary according to stock variability and/or measurement uncertainty, such as wind pressure coefficients on the building envelope, discharge coefficient of flow paths, temperature stratification, occupant behaviour, and building envelope airtightness (Cóstola et al., 2008; Herring et al., 2016; Yan et al., 2015). The wind pressure coefficients are of particular importance here. Based on Swami and Chandra's correlation, housing stock IAQ models have been developed utilising single average wind pressure coefficients for low-rise rectangular buildings. However, using a single average pressure coefficient over an HEI building's entire facade may not be appropriate. The uncertainty of wind-induced pressures at the building envelope is often more significant in dense urban environments because of sheltering and turbulence caused by other structures. Due to the difficulty in obtaining and presenting results for multizone cases with complex ventilation elements, this thesis project acknowledged the limitations of this approach.

It is also necessary to consider how best to represent building leakage in addition to the average surface pressures. Based on the data collected by (L. C. Ng et al., 2013), this was demonstrated by a multizone model infiltration study. According to the study, it is necessary to divide the exterior wall leakage on individual floors into three parts, representing the leakage of each wall's lower, middle, and upper thirds, to understand the stack effect better. Accordingly, the difficulties associated with assessing outdoor-indoor exchanges and their reliance on the representation and location of leakage components are evident.

Numerical wind field calculation (by coupling CFD with CONTAM, for instance) may derive building-specific wind pressure coefficients for any site and any oncoming wind angle as an alternative to wind tunnel data. It is also possible that they can provide pressures directly to the CONTAM models, eliminating the need to generate pressure coefficients in an intermediate step (Wang et al., 2010). When an urban wind field model is already being used to predict pressure distributions on building facades, this approach becomes particularly attractive in generating building-specific wind pressure coefficients. However, it remains to be seen how to average wind pressure distributions and relate them to building leakage distributions or determine the specific pressures associated with ventilation components on the building envelope.

Thridly, the more training samples available, the better the metamodel approximates the original model. Unfortunately, due to time constraints and data availability, creating as many samples as desired was impossible. Therefore, it was crucial to determine how accurate the metamodel must be. Nevertheless, the model's accuracy will depend on its purpose: predictive models tend to require high levels of accuracy. To ensure the accuracy of the metamodels, validation data should not be used in the training process, and the metamodel should only be used within the range of training data values.

In future work, it is recommended that the sample size of buildings and zones included in the training and testing of the metamodels be increased to reflect the accuracy of the metamodel and the calculation time. It is anticipated that increasing the sample size will allow other parameters not addressed in this thesis to be included. For example, the five selected buildings in this research were naturally ventilated buildings with localised exhaust fans. Based on the data received by the UoS EFM, this represents the majority of buildings in the UoS building stock under investigation. However, this study did not investigate mechanically ventilated buildings or where localised mechanical ventilation systems could be used (e.g. lecture halls), where the influence of outdoor environmental characteristics is reduced due to a pressurised building envelope. It is important to acknowledge that the CoSimulation models employed in this study focus primarily on airflow resulting from infiltration, as well as extract systems in kitchens and bathrooms. Mechanical ventilation, which plays a crucial role in providing filtered air to dilute particles that ingress due to infiltration, is not explicitly considered in these models. This omission is particularly noteworthy in the case of certain spaces such as lecture halls that are likely to be mechanically ventilated.

The exclusion of mechanical ventilation from the models has the potential to impact the results significantly. It is plausible that considering these additional airflows could lead to lower exposure results, as the presence of mechanical ventilation systems can effectively mitigate indoor $PM_{2.5}$ concentrations by supplying filtered air and facilitating dilution of infiltrated particles. However, it is important to note that the investigation of mechanical ventilation and its influence on exposure outcomes was beyond the scope of this study, which primarily focused on the effects of infiltration.

Therefore, it is acknowledged that further investigation is warranted to explore the potential impacts of mechanical ventilation on exposure results. Future research endeavours should consider incorporating mechanical ventilation systems into the modelling framework to comprehensively assess the combined effects of infiltration and mechanical ventilation on indoor $PM_{2.5}$ concentrations and population exposure. By expanding the scope of analysis to encompass these additional factors, a more comprehensive understanding of the indoor air quality dynamics and potential exposure reductions can be achieved. Furthermore, scientific evidence suggest that existing naturally ventilated buildings in the UoS building stock might require interventions (adding HVAC systems) due to climate change. This will have an impact on the parameters to be included in the development of the metamodels.

Lastly, the microenvironments identified in this thesis were based on the space types described in the UoS energy policy. The approach used in defining the microenvironments here might not be suitable if they are defined as three-dimensional spaces in which pollutant levels are uniform or exhibit constant statistical properties over time. Because indoor $PM_{2.5}$ concentrations vary significantly within and across buildings in HEIs, a more comprehensive statistical approach is required to calculate microenvironments in HEIs. One possible approach is to classify the time series of indoor $PM_{2.5}$ in all zones and over the simulation period based on the similarities and patterns of time-series profiles. This allows for identifying the zones that exhibit similarities in the behaviour of indoor $PM_{2.5}$ and identify potential, influential parameters for better classification of microenvironments.

Chapter 9 Conclusions

This thesis conducts an exploratory investigation involving five existing buildings within the higher education building stock of the University of Sheffield (UoS): The Arts Tower (AT), Regent Court Building (RC), Academic Development Centre (ADC), Barber House (BH), and the Interdisciplinary Centre of Social Sciences (ICoSS). The primary goal is to develop a data-driven model to rapidly assess and quantify the potential effects of enhancing the airtightness of the building envelope (Q_{50}) on the annual average population exposure to indoor fine particulate matter ($PM_{2.5}$) from outdoor sources. To achieve this goal, the study employs building physics-based modelling using the integrated co-simulation framework of CONTAM and EnergyPlus as the primary source of information on indoor $PM_{2.5}$ concentrations. Furthermore, this approach enables the examination and evaluation of the effects of varying Q_{50} within a specific range on the concentrations of indoor $PM_{2.5}$, allowing for testing and analysis of its influence. This novel comprehension establishes a methodological framework that enables the assessment and formulation of regulations on higher education institution (HEI) building stocks. This framework can be integrated with prevailing energy-related policies to mitigate campus energy consumption and carbon dioxide (CO_2) emissions effectively. Consequently, it furnishes compelling empirical evidence supporting integrating IAQ considerations with energy-related initiatives.

In conclusion, this study has offered comprehensive insights into the complex dynamics of indoor $PM_{2.5}$ concentrations in HEI buildings, shedding light on the impact of building envelope airtightness (Q_{50}) and its implications for population exposure. By systematically investigating a range of factors and employing various analytical approaches, this research has contributed to a deeper understanding of IAQ and its relationship with building design and environmental variables.The study commenced by establishing the foundation for its investigations by identifying existing data sources pertinent to HEI buildings, with a particular emphasis on the Q_{50} parameter. This initial step provided the essential data framework for the subsequent analyses, ensuring a robust and data-driven investigation. Through this approach, the research laid the

groundwork for assessing the interplay between Q_{50}, IAQ, and the well-being of occupants within HEI buildings.

The results of this research unveiled a diverse landscape of indoor $PM_{2.5}$ concentrations across the examined buildings, highlighting the considerable influence of building design and environmental variables. The analysis not only underscored the heterogeneity of $PM_{2.5}$ levels within each building but also revealed the troubling finding that a significant proportion of zones exceeded the recommended long-term exposure limit set by the World Health Organization (WHO). This raised pertinent concerns about IAQ in HEI buildings and its potential health implications, emphasising the critical importance of addressing these issues.

Furthermore, a detailed examination of floor levels in one of the buildings, notably the AT, presented interesting findings. The analysis indicated that airtightness improvements can indeed effectively reduce $PM_{2.5}$ infiltration. However, it also revealed that the impact of these improvements may vary based on factors such as building height. This underscores the need for a nuanced approach to IAQ management, considering the specific characteristics of individual buildings. In pursuit of answering the research question, the study further delved into a sensitivity analysis framework. By identifying the critical variables influencing infiltrated $PM_{2.5}$ concentrations, the research paved the way for the development of a predictive metamodel. This metamodel, particularly the $XGB_{post\text{-}HPT}$, exhibited remarkable performance, achieving a high level of accuracy in predicting $PM_{2.5}$ concentrations. This predictive tool holds great promise for architects, engineers, and policymakers aiming to design buildings that prioritise IAQ.

Lastly, the research ventured into the realm of microenvironmental modelling to quantify population exposure to indoor $PM_{2.5}$. The findings demonstrated the tangible impacts of improving building envelope airtightness on population exposure. Notably, the study revealed that the adoption of a tighter building envelope (Q_{50}=3 $m^3/h/m^2$) led to a substantial reduction in annual population exposure, marking a promising stride towards healthier indoor air quality. However, the study also wisely cautioned that achieving full compliance with stringent WHO guidelines for $PM_{2.5}$ may necessitate additional interventions beyond airtightness improvements, underlining the need for a multifaceted approach to address IAQ concerns comprehensively.

In summary, this comprehensive research endeavour has contributed significantly to the field of building science, IAQ management, and public health. It has elucidated the multifaceted relationship between building design, airtightness, and indoor $PM_{2.5}$ concentrations, offering

practical insights and tools for mitigating the potential health risks associated with poor IAQ. As we move towards a future where sustainability and well-being are paramount in architectural design, the findings of this study will undoubtedly play a pivotal role in shaping healthier indoor environments for all.

References

Abdalla, T., & Peng, C. (2021). Evaluation of housing stock indoor air quality models: A review of data requirements and model performance. *Journal of Building Engineering*, *43*(May), 102846. https://doi.org/10.1016/j.jobe.2021.102846

Abdul Shakor, A. S. ad, Pahrol, M. A., & Mazeli, M. I. (2020). Effects of Population Weighting on PM10 Concentration Estimation. *Journal of Environmental and Public Health*, *2020*. https://doi.org/10.1155/2020/1561823

Aldrich, C. (2020). Process variable importance analysis by use of random forests in a shapley regression framework. *Minerals*, *10*(5), 1–17. https://doi.org/10.3390/min10050420

Altan, H., Ward, I., Mohelnikova, J., & Vajkay, F. (2009). An internal assessment of the thermal comfort and daylighting conditions of a naturally ventilated building with an active glazed facade in a temperate climate. *Energy and Buildings*, *41*(1), 36–50. https://doi.org/10.1016/j.enbuild.2008.07.009

Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M., & Pope, C. A. (2018). Ambient PM2.5 Reduces Global and Regional Life Expectancy. *Environmental Science and Technology Letters*, *5*(9), 546–551. https://doi.org/10.1021/acs.estlett.8b00360

Arup. (2012). *University of Sheffield Energy Strategy*. *May*.

ASHRAE. (2019). ANSI/ASHRAE Standard 62.1-2019 Ventilation for Acceptable Indoor Air Quality. In *American Society of Heating, Refrigerating, and Air-Conditioning Engineers* (Issue 62.1).

ASTM. (1999). Standard E779: Standard Test Method for Determining Air Leakage by Fan Pressurization. In *American Society of Testing and Materials*.

Aunan, K., Ma, Q., Lund, M. T., & Wang, S. (2018). Population-weighted exposure to PM2.5 pollution in China: An integrated approach. *Environment International*, *120*, 111–120. https://doi.org/10.1016/j.envint.2018.07.042

Ausati, S., & Amanollahi, J. (2016). Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM2.5. *Atmospheric Environment*, *142*, 465–474. https://doi.org/10.1016/j.atmosenv.2016.08.007

Axley, J. (2007). Multizone airflow modelling in buildings: History and theory. *HVAC and R Research*, *13*(6), 907–928. https://doi.org/10.1080/10789669.2007.10391462

Axley, J. . (1988). Progress Toward A General Analytical Method for Predicting Indoor Air Pollution in Buildings, Indoor Air Quality Modelling Phase III Report. In *NBSIR 88-3814*. National Bureau of Standards (U.S.).

Axley, J. W. (1987). Indoor Air Quality Modeling Phase II Report. In *NBSIR 87-3661*. National Bureau of Standards (U.S.).

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 907. https://doi.org/10.1186/s12889-017-4914-3

Ben, H., & Steemers, K. (2018). Household archetypes and behavioural patterns in UK domestic energy use. *Energy Efficiency*, *11*(3), 761–771. https://doi.org/10.1007/s12053-017-9609-1

Bennadji, A., Seddiki, M., Alabid, J., Laing, R., & Gray, D. (2022). Predicting Energy Savings of the UK Housing Stock under a Step-by-Step Energy Retrofit Scenario towards Net-Zero. *Energies*, *15*(9), 1–18. https://doi.org/10.3390/en15093082

Birchby, D., Stedman, J., Whiting, S., & Vedrenne, M. (2019). *Air Quality Damage Cost* (Issue 2). https://uk-air.defra.gov.uk/assets/documents/reports/cat09/1902271109_Damage_cost_update_2018_FINAL_Issue_2_publication.pdf

Blochwitz, T., Otter, M., Arnold, M., Bausch, C., Clauss, C., Elmqvist, H., Junghanns, A., Mauss, J., Monteiro, M., Neidhold, T., Neumerkel, D., Olsson, H., Peetz, J.-V., & Wolf, S. (2011). The Functional Mockup Interface for Tool independent Exchange of Simulation Models. *Proceedings from the 8th International Modelica Conference, Technical Univeristy, Dresden, Germany*, *63*, 105–114. https://doi.org/10.3384/ecp11063105

Bolton, P. (2022). *Higher Education Student Numbers* (Issue July).

Booth, A. T., Choudhary, R., & Spiegelhalter, D. J. (2012). Handling uncertainty in housing stock models. *Building and Environment*, *48*(1), 35–47. https://doi.org/10.1016/j.buildenv.2011.08.016

Bottegal, G., & Pillonetto, G. (2018). The Generalized Cross Validation Filter. *Automatica*, *90*, 130–137. https://doi.org/10.1016/j.automatica.2017.12.054

Bravo-Linares, C., Ovando-Fuentealba, L., Orellana-Donoso, S., Gatica, S., Klerman, F., Mudge, S. M., Gallardo, W., Pinaud, J. P., & Loyola-Sepulveda, R. (2016). Source identification, apportionment and toxicity of indoor and outdoor PM2.5 airborne particulates in a region characterised by wood burning. *Environmental Science: Processes and Impacts*, *18*(5), 575–589. https://doi.org/10.1039/c6em00148c

Brieman, L. (2001). Random Forests. *Journal of Machine Learning*, *45*, 5–32. https://doi.org/http://dx.doi.org/10.1023/A:1010933404324

Brookes, D., Stedman, J., Kent, A., Whiting, S., Rose, R., & Williams, C. (2021). *Technical report on UK supplementary assessment under The Air Quality Directive (2008/50/EC), The Air Quality Framework Directive (96/62/EC) and Fourth Daughter Directive (2004/107/EC) for 2019* (Issue 1).

Bruhns, H. R., Steadman, P., & Marjanovic, L. (2006). A preliminary model of non-domestic energy use for England and Wales. *COBRA 2006 - Proceedings of the Annual Research Conference of the Royal Institution of Chartered Surveyors*, 7–8.

BSI. (2001). EN13829:2001.Thermal Performance of Buildings e Determination of Air Permeability of Buildings e Fan Pressurization Method. In *British Standards Institution*.

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., Apte, J. S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim, S. S., Kan, H., Walker, K. D., Thurston, G. D., Hayes, R. B., … Spadaro, J. V. (2018). Global estimates of mortality associated with longterm exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(38), 9592–9597. https://doi.org/10.1073/pnas.1803222115

Carrié, F. R., & Leprince, V. (2016). Uncertainties in building pressurisation tests due to steady wind. *Energy and Buildings*, *116*, 656–665. https://doi.org/10.1016/j.enbuild.2016.01.029

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., & Forastiere, F. (2013). Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environmental Health Perspectives*, *121*(3), 324–331. https://doi.org/10.1289/ehp.1205862

Chakraborty, R., Heydon, J., Mayfield, M., & Mihaylova, L. (2020). Indoor air pollution from residential stoves: Examining the flooding of particulate matter into homes during real-world use. *Atmosphere*, *11*(12). https://doi.org/10.3390/atmos11121326

Challoner, A., Pilla, F., & Gill, L. (2015). Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings.

International Journal of Environmental Research and Public Health, *12*(12), 15233–15253. https://doi.org/10.3390/ijerph121214975

Chapman, P. M. (2007). Determining When Contamination is Pollution — Weight of Evidence Determinations for Sediments and Effluents. *Environment International*, *33*(4), 492–501. https://doi.org/10.1016/j.envint.2006.09.001

Chelani, A. B., & Devotta, S. (2006). Nonlinear Analysis and Prediction of Coarse Particulate Matter Concentration in Ambient Air. *Journal of the Air and Waste Management Association*, *56*(1), 78–84. https://doi.org/10.1080/10473289.2006.10464432

Chen, A., Jacobsen, K. H., Deshmukh, A. A., & Cantor, S. B. (2015). The evolution of the disability-adjusted life year (DALY). *Socio-Economic Planning Sciences*, *49*(March), 10–15. https://doi.org/10.1016/j.seps.2014.12.002

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Choi, M.-L., Jae Lim, M., Kwon, Y.-M., & Chung, D.-K. (2017). A Study on the Prediction Method of Emergency Room (ER) Pollution Level based on Deep Learning using Scattering Sensor. *Journal of Engineering and Applied Sciences*, *12*(10), 2560–2564. https://doi.org/10.36478/jeasci.2017.2560.2564

Christoph, M. (2020). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*.

CIBSE. (2005). *Natural ventilation in non-domestic buildings: CIBSE Application Manual AM10:1997*. 99.

CIBSE. (2018). Environmental design CIBSE Guide A. In *The Chartered Institution of Building Services Engineers, London.* (Issue June).

CIBSE. (2022). *TM23 Testing Buildings for Air Leakage*.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., … Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Coleman, J. R., & Meggers, F. (2018). Sensing of indoor air quality—characterization of spatial and temporal pollutant evolution through distributed sensing. *Frontiers in Built*

Environment, *4*(August), 1–12. https://doi.org/10.3389/fbuil.2018.00028

COMEAP. (2009). Long-Term Exposure to Air Pollution: Effect on Mortality. In *A report by the Committee on the Medical Effects of Air Pollutants*.

COMEAP. (2018). *Associations of long-term average concentrations of nitrogen dioxide with mortality*.

Connolly, E., Fuller, G., Baker, T., & Willis, P. (2013). Update on Implementation of the Daily Air Quality Index. *Department for Environment Food & Rural Affairs*, *April*. https://uk-air.defra.gov.uk/assets/documents/reports/cat14/1304251155_Update_on_Implementation_of_the_DAQI_April_2013_Final.pdf

Conte, S. D., & de Boor, C. (1972). *Elementary Numerical Analysis: An Algorithmic Approach*.

Cooper, E. W., Etheridge, D. W., & Smith, S. J. (2007). Determining the adventitious leakage of buildings at low pressure. Part 2: Pulse technique. *Building Services Engineering Research and Technology*, *28*(1), 81–96. https://doi.org/10.1177/0143624406072331

Cóstola, D., Blocken, B., & Hensen, J. L. M. (2008). Uncertainties due to the use of surface averaged wind pressure coefficients. *Proceedings of the 29th AIVC Conference, Kyoto, Japan*, *2008*, 14–16.

Cserbik, D., Chen, J. C., McConnell, R., Berhane, K., Sowell, E. R., Schwartz, J., Hackman, D. A., Kan, E., Fan, C. C., & Herting, M. M. (2020). Fine particulate matter exposure during childhood relates to hemispheric-specific differences in brain structure. *Environment International*, *143*(December 2019), 105933. https://doi.org/10.1016/j.envint.2020.105933

Dales, R., Miller, D., & McMullen, E. (1997). Indoor air quality and health: validity and determinants of reported home dampness and moulds. *International Journal of Epidemiology*, *26 1*, 120–125.

Das, P., Shrubsole, C., Jones, B., Hamilton, I., Chalabi, Z., Davies, M., Mavrogianni, A., & Taylor, J. (2014). Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling. *Building and Environment*, *78*, 171–182. https://doi.org/10.1016/j.buildenv.2014.04.017

Davis, J. A., & Nutter, D. W. (2010). Occupancy diversity factors for common university building types. *Energy and Buildings*, *42*(9), 1543–1551. https://doi.org/10.1016/j.enbuild.2010.03.025

DEFRA. (2012). *Fine Particulate Matter (PM2.5) in the United Kingdom*. http://uk-air.defra.gov.uk/assets/documents/reports/cat11/1212141150_AQEG_Fine_Particulate_Matter_in_the_UK.pdf

DEFRA. (2020). *Air Pollution in the UK 2019* (Issue September). https://uk-air.defra.gov.uk/library/annualreport/viewonline?year=2019_issue_1&jump=5#report_pdf

DEFRA (Deparment for Environment Food and Rural Affairs) & DfT (Department for Transport) UK. (2017). *UK Plan for Tackling Roadside Nitrogen Dioxide Concentrations: Detailed Plan* (Issue July). www.nationalarchives.gov.uk/doc/open-government-licence/version/3/%0Awww.gov.uk/government/publications

DEFRA (Deparment for Environment Food and Rural Affairs) UK. (2018). *Air Pollution in the UK 2017* (Issue September).

Department for Environment Food and Rural Affairs. (2007). *The Air Quality Strategy for England , Scotland , Wales and Northern Ireland: Volume 1*. *1*, 1–56. https://doi.org/CM 4548, SE 2000/3, NIA 7

Diapouli, E., Chaloulakou, A., & Koutrakis, P. (2013). Estimating the concentration of indoor particles of outdoor origin: A review. *Journal of the Air and Waste Management Association*, *63*(10), 1113–1129. https://doi.org/10.1080/10962247.2013.791649

Dias, D., & Tchepel, O. (2018). Spatial and temporal dynamics in air pollution exposure assessment. *International Journal of Environmental Research and Public Health*, *15*(3). https://doi.org/10.3390/ijerph15030558

Dimitroulopoulou, C., Ashmore, M. R., Byrne, M. A., & Kinnersley, R. P. (2001). Modelling of indoor exposure to nitrogen dioxide in the UK. *Atmospheric Environment*, *35*(2), 269–279. https://doi.org/10.1016/S1352-2310(00)00176-X

Dimitroulopoulou, C., Ashmore, M. R., Hill, M. T. R., Byrne, M. A., & Kinnersley, R. (2006). INDAIR: A probabilistic model of indoor air pollution in UK homes. *Atmospheric Environment*, *40*(33), 6362–6379. https://doi.org/10.1016/j.atmosenv.2006.05.047

Dols, W., & Polidoro, B. (2020). *CONTAM User Guide and Program Documentation Version 3.4*. Technical Note (NIST TN), National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/https://doi.org/10.6028/NIST.TN.1887r1

Dols, W. Stuart, Emmerich, S. J., & Polidoro, B. J. (2016). Coupling the multizone airflow and contaminant transport software CONTAM with EnergyPlus using co-simulation. *Building Simulation*, *9*(4), 469–479. https://doi.org/10.1007/s12273-016-0279-2

Dols, W. Stuart, Milando, C. W., Ng, L., Emmerich, S. J., & Teo, J. (2021). On the benefits of whole-building IAQ, ventilation, infiltration, and energy analysis using co-simulation between CONTAM and EnergyPlus. *Journal of Physics: Conference Series*, *2069*(1). https://doi.org/10.1088/1742-6596/2069/1/012183

Dols, W. Stuart, & Polidoro, B. J. (2020). *CONTAM User Guide and Program Documentation - Version 3.4 (NIST Technical Note 1887)*.

Dols, W.S., & Polidoro, B. J. (2015). *CONTAM User Guide and Program Documentation Version 3.2. Tech. rept.* NIST.

ECA. (2003). European Collaborative Action on '"Urban air, Indoor Environment and human Exposure"' Ventilation, Good Indoor Air Quality and Rational Use of Energy. Report No 23. In *Pollution Atmospherique*.

Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, *49*, 591–603. https://doi.org/10.1016/j.enbuild.2012.03.010

Elbayoumi, M., Ramli, N. A., Faizah, N., & Yusof, F. (2015). Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM 2 . 5 e 10 and PM 2 . 5 concentrations in naturally ventilated schools. *Atmospheric Pollution Research*, *6*(6), 1013–1023. https://doi.org/10.1016/j.apr.2015.09.001

Elliot, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. (2000). *Spatial epidemiology: methods and applications.* Oxford University Press.

Emmerich, S. J., & Hirnikel, D. (2001a). Validation of multizone IAQ modeling of residential-scale buildings: A review. *ASHRAE Transactions*, *107 PART 2*(January 2001), 619–628.

Emmerich, S. J., & Hirnikel, D. (2001b). Validation of multizone IAQ modeling of residential-scale buildings: A review. *ASHRAE Transactions*, *107 PART 2*, 619–628.

Emmerich, S. J., Ng, L. C., & Dols, W. S. (2019). Simulation analysis of potential energy savings from air sealing retrofits of U.S. Commercial buildings. *ASTM Special Technical Publication*, *STP 1615*, 61–70. https://doi.org/10.1520/STP161520180021

Etheridge, D. (2011). *Natural ventilation of buildings: theory, measurement and design*. John Wiley & Sons.

Evans, S., Liddiard, R., & Steadman, P. (2017). 3DStock: A new kind of three-dimensional model of the building stock of England and Wales, for use in energy analysis. *Environment and Planning B: Urban Analytics and City Science*, *44*(2), 227–255. https://doi.org/10.1177/0265813516652898

Everett, C. P. (2013). Sheffield arts tower: Rejuvenation of a II* listed structure. *Proceedings of the Institution of Civil Engineers: Structures and Buildings*, *166*(1), 38–48. https://doi.org/10.1680/stbu.10.00072

Fazli, T., Dong, X., Fu, J. S., & Stephens, B. (2021). Predicting U.S. Residential Building Energy Use and Indoor Pollutant Exposures in the Mid-21st Century. *Environmental Science & Technology*, *55*(5), 3219–3228. https://doi.org/10.1021/acs.est.0c06308

Fazli, T., & Stephens, B. (2018). Development of a nationally representative set of combined building energy and indoor air quality models for U.S. residences. *Building and Environment*, *136*, 198–212. https://doi.org/10.1016/J.BUILDENV.2018.03.047

Fenton, N., & Neil, M. (2018). *Risk Assessment and Decision Analysis with Bayesian Networks*. Chapman and Hall/CRC. https://doi.org/10.1201/b21982

Ferguson, L., Taylor, J., Davies, M., Shrubsole, C., Symonds, P., & Dimitroulopoulou, S. (2020). Exposure to indoor air pollution across socio-economic groups in high-income countries: A scoping review of the literature and a modelling methodology. *Environment International*, *143*(November 2019), 105748. https://doi.org/10.1016/j.envint.2020.105748

Fernando, G., Plata, V., & Agustiniana, U. (2020). *Use Of Non-Industrial Environmental Sensors And Machine Use Of Non-Industrial Environmental Sensors And Machine Learning Techniques In Telemetry For Indoor Air Pollution*. *August*.

Feustel, H. E. (1999). COMIS-an international multizone air-flow and contaminant transport model. *Energy and Buildings*, *30*(1), 3–18. https://doi.org/10.1016/S0378-7788(98)00043-7

Gaidajis, G., & Angelakoglou, K. (2009). Indoor air quality in university classrooms and relative environment in terms of mass concentrations of particulate matter. *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering*, *44*(12), 1227–1232. https://doi.org/10.1080/10934520903139936

García-Tobar, J. (2019). Weather-dependent modelling of the indoor radon concentration in two dwellings using CONTAM. *Indoor and Built Environment*, *28*(10), 1341–1349. https://doi.org/10.1177/1420326X19841119

Ghiassi, N., & Mahdavi, A. (2017). Reductive bottom-up urban energy computing supported by multivariate cluster analysis. *Energy and Buildings*, *144*, 372–386. https://doi.org/10.1016/j.enbuild.2017.03.004

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gillott, M. C., Loveday, D. L., White, J., Wood, C. J., Chmutina, K., & Vadodaria, K. (2016). Improving the airtightness in an existing UK dwelling: The challenges, the measures and their effectiveness. *Building and Environment*, *95*, 227–239.

https://doi.org/10.1016/j.buildenv.2015.08.017

Giussani, A. (2021). Applied Machine Learning with Python. In *Applied Machine Learning* (2nd Editio). Bocconi University Press. https://books.google.co.uk/books?id=6dtozgEACAAJ

González-Martín, J., Kraakman, N. J. R., Pérez, C., Lebrero, R., & Muñoz, R. (2021). A state–of–the-art review on indoor air pollution and strategies for indoor air pollution control. *Chemosphere*, *262*. https://doi.org/10.1016/j.chemosphere.2020.128376

Gramsch, E., Cáceres, D., Oyola, P., Reyes, F., Vásquez, Y., Rubio, M. A., & Sánchez, G. (2014). Influence of surface and subsidence thermal inversion on PM2.5 and black carbon concentration. *Atmospheric Environment*, *98*, 290–298. https://doi.org/https://doi.org/10.1016/j.atmosenv.2014.08.066

Grot, R. (1985). *Indoor Air Quality Modeling Phase Report Framework for Development of General Models*.

Gu, J., Yang, B., Brauer, M., & Zhang, K. M. (2021). Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models. *Atmospheric Environment*, *246*, 118125. https://doi.org/10.1016/j.atmosenv.2020.118125

Gulliver, J., & Briggs, D. J. (2004). Personal exposure to particulate air pollution in transport microenvironments. *Atmospheric Environment*, *38*(1), 1–8. https://doi.org/10.1016/j.atmosenv.2003.09.036

Haghighat, F. (1989). Air infiltration and indoor air quality models — a review. *International Journal of Ambient Energy*, *10*(3), 115–122. https://doi.org/10.1080/01430750.1989.9675130

Haghighat, Fariborz. (1996). A comprehensive validation of two airflow models - COMIS and CONTAM. *Indoor Air*, *6*(4), 278–288. https://doi.org/10.1111/j.1600-0668.1996.00007.x

Harrison, R., & Yin, J. (2004). *Characterisation of particulate matter in the United Kingdom: Literature Review*. *March*, 1–46.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*(8), e5518. https://doi.org/10.7717/peerj.5518

Hensen, J., & Lamberts, R. (2011). *Building Performance Simulation for Design and Operation*. London ; New York : Spon Press, 2011.

Hering, S. V., Lunden, M. M., Thatcher, T. L., Kirchstetter, T. W., & Brown, N. J. (2007). Using Regional Data and Building Leakage to Assess Indoor Concentrations of Particles of Outdoor Origin. *Aerosol Science and Technology*, *41*(7), 639–654.

https://doi.org/10.1080/02786820701368026

Herring, S. J., Batchelor, S., Bieringer, P. E., Lingard, B., Lorenzetti, D. M., Parker, S. T., Rodriguez, L., Sohn, M. D., Steinhoff, D., & Wolski, M. (2016). Providing pressure inputs to multizone building models. *Building and Environment*, *101*, 32–44. https://doi.org/10.1016/j.buildenv.2016.02.012

HESA. (2021). *HESA - Higher Education Statistics Agency*. 0–1. https://www.hesa.ac.uk/stats

Hinnells, M., & Shea, A. D. (2008). Transforming UK non-residential buildings : achieving a 60 % cut in CO2 emissions by 2050 Transforming UK non-residential buildings : achieving a 60 % cut in CO2 emissions by 2050. *Paper Presented at the IEECB*.

Holgate, S. T. (2017). 'Every breath we take: the lifelong impact of air pollution' – a call for action. *Clinical Medicine*, *17*(1), 8 LP – 12. https://doi.org/10.7861/clinmedicine.17-1-8

Huang, H., Cao, J. J., Lee, S. C., Zou, C. W., Chen, X. G., & Fan, S. J. (2007). Spatial Variation and Relationship of Indoor/Outdoor PM2.5 at Residential Homes in Guangzhou City, China. *Aerosol and Air Quality Research*, *7*(4), 518–530. https://doi.org/10.4209/aaqr.2007.03.0018

IEHIAS. (n.d.). *Indoor air pollution models | Integrated Environmental Health Impact Assessment System*. Retrieved October 5, 2020, from http://www.integrated-assessment.eu/eu/guidebook/indoor_air_pollution_models.html

Isiugo, K., Jandarov, R., Cox, J., Chillrud, S., Grinshpun, S. A., Hyttinen, M., Yermakov, M., Wang, J., Ross, J., & Reponen, T. (2019). Predicting indoor concentrations of black carbon in residential environments. *Atmospheric Environment*, *201*, 223–230. https://doi.org/10.1016/j.atmosenv.2018.12.053

Jafta, N., Barregard, L., Jeena, P. M., & Naidoo, R. N. (2017). Indoor air quality of low and middle income urban households in Durban , South Africa. *Environmental Research*, *156*(January), 47–56. https://doi.org/10.1016/j.envres.2017.03.008

Jamieson, S. (2008). A Partial Review of Mass Balance Models. *Department of Epidemiology and Biostatistics Imperial College London*, *Feustel 1999*, 1–9.

Jones, B, Phillips, G., O'Leary, C., Molina, C., & Hall, I. (2018). Diagnostic barriers to using PM2. 5 concentrations as metrics of indoor air quality. *Smart Ventilation for Buildings, July*. https://www.researchgate.net/profile/Benjamin_Jones10/publication/326573104_Diagnostic_barriers_to_using_PM25_concentrations_as_metrics_of_indoor_air_quality/links/5b5708f145851507a7c4c2ee/Diagnostic-barriers-to-using-PM25-concentrations-as-metrics-of-ind

Jones, Benjamin. (2019). *Generic Global Sensitivity Analysis Code*. https://doi.org/10.13140/RG.2.2.21670.88644

Jones, Benjamin, Das, P., Chalabi, Z., Davies, M., Hamilton, I., Lowe, R., Mavrogianni, A., Robinson, D., & Taylor, J. (2015). Assessing uncertainty in housing stock infiltration rates andassociated heat loss: English and UK case studies. *Building and Environment*, *92*, 644–656. https://doi.org/10.1016/j.buildenv.2015.05.033

Jones, Benjamin, Das, P., Chalabi, Z., Davies, M., Hamilton, I., Lowe, R., Milner, J., Ridley, I., Shrubsole, C., & Wilkinson, P. (2013). The effect of party wall permeability on estimations of infiltration from air leakage. *International Journal of Ventilation*, *12*(1), 17–29. https://doi.org/10.1080/14733315.2013.11683999

Jung, K. H., Bernabé, K., Moors, K., Yan, B., Chillrud, S. N., Whyatt, R., Camann, D., Kinney, P. L., Perera, F. P., & Miller, R. L. (2011). Effects of Floor Level and Building Type on Residential Levels of Outdoor and Indoor Polycyclic Aromatic Hydrocarbons, Black Carbon, and Particulate Matter in New York City. *Atmosphere*, *2*(2), 96–109. https://doi.org/10.3390/atmos2020096

Kim, M., Kim, Y., Sung, S., & Yoo, C. (2009). Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks. *ICCAS-SICE 2009 - ICROS-SICE International Joint Conference 2009, Proceedings*, 1688–1692.

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., Behar, J. V., Hern, S. C., & Engelmann, W. H. (2001). The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*, *11*(3), 231–252. https://doi.org/10.1038/sj.jea.7500165

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., & Behar, J. V. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology*, *11*(3), 231–252. https://doi.org/10.1038/sj.jea.7500165

Kousa, A., Monn, C., Rotko, T., Alm, S., Oglesby, L., & Jantunen, M. J. (2001). Personal exposures to NO2 in the EXPOLIS-study: Relation to residential indoor, outdoor and workplace concentrations in Basel, Helsinki and Prague. *Atmospheric Environment*, *35*(20), 3405–3412. https://doi.org/10.1016/S1352-2310(01)00131-5

Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J., Murith, C., Palacios, M., & Baechler, S. (2015). Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation : An application to Switzerland. *Science of the Total*

Environment, The, *505*, 137–148. https://doi.org/10.1016/j.scitotenv.2014.09.064

Lane, K. J., Levy, J. I., Scammell, M. K., Patton, A. P., Durant, J. L., Mwamburi, M., Zamore, W., & Brugge, D. (2015). Effect of time-activity adjustment on exposure assessment for traffic-related ultrafine particles. *Journal of Exposure Science & Environmental Epidemiology*, *25*(5), 506–516. https://doi.org/10.1038/jes.2015.11

Lee, M. C., Mui, K. W., Wong, L. T., Chan, W. Y., Lee, E. W. M., & Cheung, C. T. (2012). Student learning performance and indoor environmental quality (IEQ) in air-conditioned university teaching rooms. *Building and Environment*, *49*, 238–244. https://doi.org/https://doi.org/10.1016/j.buildenv.2011.10.001

Li, J., Liu, L., Le, T. D., & Liu, J. (2020). Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*, *2*(1), 13–15. https://doi.org/10.1038/s42256-019-0140-2

Li, X., & Wen, J. (2014). Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews*, *37*, 517–537. https://doi.org/10.1016/j.rser.2014.05.056

Lim, H., & Zhai, Z. J. (2017). Review on stochastic modeling methods for building stock energy prediction. *Building Simulation*, *10*(5), 607–624. https://doi.org/10.1007/s12273-017-0383-y

Liu, Z., Li, H., & Cao, G. (2017). Quick Estimation Model for the Concentration of Indoor Airborne Culturable Bacteria: An Application of Machine Learning. *International Journal of Environmental Research and Public Health*, *14*(8), 857. https://doi.org/10.3390/ijerph14080857

Logue, J. M., Sherman, M. H., Lunden, M. M., Klepeis, N. E., Williams, R., Croghan, C., & Singer, B. C. (2015). Development and assessment of a physics-based simulation model to investigate residential PM2.5 infiltration across the US housing stock. *Building and Environment*, *94*(P1), 21–32. https://doi.org/10.1016/j.buildenv.2015.06.032

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems*.

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://doi.org/10.1016/j.proeng.2017.10.002

Malkawi, A., & Augenbroe, G. (2004). Advanced building simulation. In A. Malkawi & G. Augenbroe (Eds.), *Taylor & Francis* (1st editio). Spon Press.

Mara, F. (2010). *Retrofit: HLM Architects reclads Sheffield Arts Tower*. https://www.architectsjournal.co.uk/specification/retrofit-hlm-architects-reclads-sheffield-arts-tower

Martínez-Comesaña, M., Eguía-Oller, P., Martínez-Torres, J., Febrero-Garrido, L., & Granada-Álvarez, E. (2022). Optimisation of thermal comfort and indoor air quality estimations applied to in-use buildings combining NSGA-III and XGBoost. *Sustainable Cities and Society*, *80*(October 2021), 103723. https://doi.org/10.1016/j.scs.2022.103723

McDowell, T. P., Emmerich, S., Thornton, J. W., & Walton, G. (2003). Integration of airflow and energy simulation using CONTAM and TRNSYS. *ASHRAE Transactions*, *109 PART 2*(March), 757–770. https://doi.org/10.1107/S1600536812032953

Megri, A. C., & Haghighat, F. (2007). Zonal modeling for simulating indoor environment of buildings: Review, recent developments, and applications. *HVAC and R Research*, *13*(6), 887–905. https://doi.org/10.1080/10789669.2007.10391461

Milando, C. W., Carnes, F., Vermeer, K., Levy, J. I., & Fabian, M. P. (2022a). Sensitivity of modeled residential fine particulate matter exposure to select building and source characteristics: A case study using public data in Boston, MA. *Science of the Total Environment*, *840*(June), 156625. https://doi.org/10.1016/j.scitotenv.2022.156625

Milando, C. W., Carnes, F., Vermeer, K., Levy, J. I., & Fabian, M. P. (2022b). Sensitivity of modeled residential fine particulate matter exposure to select building and source characteristics: A case study using public data in Boston, MA. *Science of The Total Environment*, *840*(February), 156625. https://doi.org/10.1016/j.scitotenv.2022.156625

Miller, S. L., & Nazaroff, W. W. (2001). Environmental tobacco smoke particles in multizone indoor environments. *Atmospheric Environment*, *35*(12), 2053–2067. https://doi.org/10.1016/S1352-2310(00)00506-9

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning* (Issue August). The MIT Press.

Molina, C. (2019). *A data analysis of the Chilean housing stock and the estimation of uncertainty in indoor air quality in Chilean houses* (Issue December). The University of Nottingham.

Molina, C., Jones, B., Hall, I. P., & Sherman, M. H. (2020). CHAARM : A model to predict uncertainties in indoor pollutant concentrations , ventilation and infiltration rates , and associated energy demand in Chilean houses. *Energy & Buildings*, 110539. https://doi.org/10.1016/j.enbuild.2020.110539

Molina, C., Kent, M., Hall, I., & Jones, B. (2020). A data analysis of the Chilean housing stock

and the development of modelling archetypes. *Energy and Buildings*, *206*, 109568. https://doi.org/10.1016/j.enbuild.2019.109568

Morrison, R. (2018). Energy system modeling: Public transparency, scientific reproducibility, and open development. *Energy Strategy Reviews*, *20*, 49–63. https://doi.org/10.1016/j.esr.2017.12.010

Mortimer, N. D., Ashley, A., & Rix, J. H. R. (2000). Detailed energy surveys of nondomestic buildings. *Environment and Planning B: Planning and Design*, *27*(1), 25–32. https://doi.org/10.1068/b2572

Nasir, Z. A., & Colbeck, I. (2009). Particulate air pollution in transport micro-environments. *Journal of Environmental Monitoring*, *11*(6), 1140. https://doi.org/10.1039/b821824b

Nazaroff, W. W. (2004a). Indoor particle dynamics. *Indoor Air, Supplement*, *14*(SUPPL. 7), 175–183. https://doi.org/10.1111/j.1600-0668.2004.00286.x

Nazaroff, W. W. (2004b). Indoor particle dynamics. *Indoor Air, Supplement*, *14*(SUPPL. 7), 175–183. https://doi.org/10.1111/j.1600-0668.2004.00286.x

Nemmar, A., Holme, J. A., Rosas, I., Schwarze, P. E., & Alfaro-Moreno, E. (2013). Recent Advances in Particulate Matter and Nanoparticle Toxicology: A Review of the In Vivo and In Vitro Studies. *BioMed Research International*, *2013*, 1–22. https://doi.org/10.1155/2013/279371

Ng, L. C., Musser, A., Persily, A. K., & Emmerich, S. J. (2013). Multizone airflow models for calculating infiltration rates in commercial reference buildings. *Energy and Buildings*, *58*, 11–18. https://doi.org/10.1016/j.enbuild.2012.11.035

Ng, L. C., Persily, A. K., & Emmerich, S. J. (2012). *NIST Technical Note 1734 Airflow and Indoor Air Quality Models of DOE Reference Commercial Buildings*. *October*.

Ng, L., Poppendieck, D., Polidoro, B., Dols, W. S., Emmerich, S., & Persily, A. (2021). Single-Zone Simulations Using FaTIMA for Reducing Aerosol Exposure in Educational Spaces. *National Institute of Standards and Technology Technical Note 2150*, 79. https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2150.pdf%0Ahttps://doi.org/10.6028/NIST.TN.2150

Niu, M., Wang, Y., Sun, S., & Li, Y. (2016). A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. *Atmospheric Environment*, *134*, 168–180. https://doi.org/10.1016/j.atmosenv.2016.03.056

Norbäck, D., & Nordström, K. (2008). Sick building syndrome in relation to air exchange rate, CO(2), room temperature and relative air humidity in university computer classrooms: an

experimental study. *International Archives of Occupational and Environmental Health*, *82*(1), 21–30. https://doi.org/10.1007/s00420-008-0301-9

Norbäck, D., Nordström, K., & Zhao, Z. (2013). Carbon dioxide (CO2) demand-controlled ventilation in university computer classrooms and possible effects on headache, fatigue and perceived indoor environment: an intervention study. *International Archives of Occupational and Environmental Health*, *86*(2), 199–209. https://doi.org/10.1007/s00420-012-0756-6

O'Leary, C., de Kluizenaar, Y., Jacobs, P., Borsboom, W., Hall, I., & Jones, B. (2019). Investigating measurements of fine particle (PM 2.5) emissions from the cooking of meals and mitigating exposure using a cooker hood. *Indoor Air*, *29*(3), 423–438. https://doi.org/10.1111/ina.12542

O'Leary, C., Jones, B., Dimitroulopoulou, S., & Hall, I. P. (2019). Setting the standard: The acceptability of kitchen ventilation for the English housing stock. *Building and Environment*, *166*(September), 106417. https://doi.org/10.1016/j.buildenv.2019.106417

OECD. (2016). The Economic Consequences of Outdoor Air Pollution. In *OECD Publishing* (Vol. 15, Issue 4). OECD. https://doi.org/10.1787/9789264257474-en

Oezkaynak, H., Xue, J., Weker, R., Butler, D., & Koutrakis, P. (1996). *Particle Team (PTEAM) Study: Analysis of the Data. Final Report, Volume 3*. https://www.osti.gov/biblio/402191

Office of Energy Efficient and Renewable Energy. (n.d.). *EnergyPlus | Department of Energy*. ENERGY.GOV. Retrieved February 26, 2020, from https://www.energy.gov/eere/buildings/downloads/energyplus-0

Oikonomou, E., Davies, M., Mavrogianni, A., Biddulph, P., Wilkinson, P., & Kolokotroni, M. (2012). Modelling the relative importance of the urban heat island and the thermal quality of dwellings for overheating in London. *Building and Environment*, *57*(July 2006), 223–238. https://doi.org/10.1016/j.buildenv.2012.04.002

Oladokun, M. G., & Odesola, I. A. (2015). Household energy consumption and carbon emissions for sustainable cities – A critical review of modelling approaches. *International Journal of Sustainable Built Environment*, *4*(2), 231–247. https://doi.org/10.1016/j.ijsbe.2015.07.005

Ott, W. R., Steinemann, A. C., & Wallac, L. a. (2006). *Exposure Analysis* (W. R. Ott, A. C. Steinemann, & L. A. Wallace (eds.); 1st Editio). Taylor & Francis. https://doi.org/10.1201/9781420012637

Oyedepo, S. O., Anifowose, E. G., Obembe, E. O., & Khanmohamadi, S. (2021). Energy-saving strategies on university campus buildings: Covenant University as case study. In D. Borge-

Diez & E. B. T.-E. S. F. and F. Rosales-Asensio (Eds.), *Energy Services and Management* (pp. 131–154). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-820592-1.00006-3

Pai, S. J., Carter, T. S., Heald, C. L., & Kroll, J. H. (2022). Updated World Health Organization Air Quality Guidelines Highlight the Importance of Non-anthropogenic PM 2.5 . *Environmental Science & Technology Letters*, *9*(6), 501–506. https://doi.org/10.1021/acs.estlett.2c00203

Panagopoulos, I. K., Karayannis, A. N., Kassomenos, P., & Aravossis, K. (2011). A CFD simulation study of VOC and formaldehyde indoor air pollution dispersion in an apartment as part of an indoor pollution management plan. *Aerosol and Air Quality Research*, *11*(6), 758–762. https://doi.org/10.4209/aaqr.2010.11.0092

Park, J. S., Jee, N.-Y., & Jeong, J.-W. (2014). Effects of types of ventilation system on indoor particle concentrations in residential buildings. *Indoor Air*, *24*(6), 629–638. https://doi.org/10.1111/ina.12117

Patil, S. L., Tantau, H. J., & Salokhe, V. M. (2008). *Modelling of tropical greenhouse temperature by auto regressive and neural network models*. *99*, 423–431. https://doi.org/10.1016/j.biosystemseng.2007.11.009

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Muller, A., Nothman, J., Louppe, G., Prenttenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*, *12*, 2825–2830. https://doi.org/https://doi.org/10.48550/arXiv.1201.0490

Persily, A. K., & Ivy, E. M. (2001). *Input Data for Multizone Airflow and IAQ Analysis.* 42.

Persily, A., Musser, A., & Leber, D. (2006). *A collection of homes to represent the U.S. housing stock*. https://doi.org/10.6028/NIST.IR.7330

Pout, C. H. (2000). N-DEEM: The national nondomestic buildings energy and emissions model. *Environment and Planning B: Planning and Design*, *27*(5), 721–732. https://doi.org/10.1068/bst12

Public Health England. (2018). *Estimation of costs to the NHS and social care due to the health impacts of air pollution: summary report About Public Health England*. www.facebook.com/PublicHealthEngland

Renard, P., Alcolea, A., & Ginsbourger, D. (2013). Stochastic versus Deterministic Approaches. In *Environmental Modelling* (Issue 2013, pp. 133–149). John Wiley & Sons, Ltd.

https://doi.org/10.1002/9781118351475.ch8

Riley, W. J., McKone, T. E., Lai, A. C. K., & Nazaroff, W. W. (2002). Indoor particulate matter of outdoor origin: Importance of size-dependent removal mechanisms. *Environmental Science and Technology*, *36*(2), 200–207. https://doi.org/10.1021/es010723y

Robinson, S. (2008). Conceptual modelling for simulation Part I: Definition and requirements. *Journal of the Operational Research Society*, *59*(3), 278–290. https://doi.org/10.1057/palgrave.jors.2602368

Rosofsky, A., Levy, J. I., Breen, M. S., Zanobetti, A., & Fabian, M. P. (2019). The impact of air exchange rate on ambient air pollution exposure and inequalities across all residential parcels in Massachusetts. *Journal of Exposure Science & Environmental Epidemiology*, *29*(4), 520–530. https://doi.org/10.1038/s41370-018-0068-3

Ross, N. (2019). *Introduction to Generalized Additive Models*. 1–27. https://s3.amazonaws.com/assets.datacamp.com/production/course_6413/slides/chapter1.pdf%0Apapers3://publication/uuid/18079837-6A75-4E2B-9E60-1210A35AE4CE

Rossum, G. Van, & Drake, F. L. (2022a). *Python 3.10.5 Documentation* (3.10.5; pp. 1–1144).

Rossum, G. Van, & Drake, F. L. (2022b). *Python 3.10.5 Documentation* (3.10.5; pp. 1–1144). https://www.python.org/downloads/release/python-3105/

S Douglas, J. (2014). Energy Performance Analysis of University Buildings: Case Studies at Sheffield University, UK. *Journal of Architectural Engineering Technology*, *03*(03). https://doi.org/10.4172/2168-9717.1000129

Sarbu, I., & Pacurar, C. (2015). Experimental and numerical research to assess indoor environment quality and schoolwork performance in university classrooms. *Building and Environment*, *93*, 141–154. https://doi.org/https://doi.org/10.1016/j.buildenv.2015.06.022

Sarra, A., Fontanella, L., Valentini, P., & Palermi, S. (2016). Quantile regression and Bayesian cluster detection to identify radon prone areas. *Journal of Environmental Radioactivity*, *164*, 354–364. https://doi.org/10.1016/j.jenvrad.2016.06.014

Schneider, T., Alstrup Jensen, K., Clausen, P. A., Afshari, A., Gunnarsen, L., Wåhlin, P., Glasius, M., Palmgren, F., Nielsen, O. J., & Fogh, C. L. (2004). Prediction of indoor concentration of 0.5-4 μm particles of outdoor origin in an uninhabited apartment. *Atmospheric Environment*, *38*(37), 6349–6359. https://doi.org/10.1016/j.atmosenv.2004.08.002

Schwartz, Y., Korolija, I., Symonds, P., Godoy-Shimizu, D., Dong, J., Hong, S. M., Mavrogianni, A., Grassie, D., & Mumovic, D. (2021). Indoor Air Quality and Overheating in UK Classrooms – an Archetype Stock Modelling Approach. *Journal of Physics: Conference*

Series, *2069*(1), 012175. https://doi.org/10.1088/1742-6596/2069/1/012175

Schweizer, C., Edwards, R. D., Bayer-Oglesby, L., Gauderman, W. J., Ilacqua, V., Juhani Jantunen, M., Lai, H. K., Nieuwenhuijsen, M., & Künzli, N. (2007). Indoor time-microenvironment-activity patterns in seven regions of Europe. *Journal of Exposure Science and Environmental Epidemiology*, *17*(2), 170–181. https://doi.org/10.1038/sj.jes.7500490

Servén, D., & Brummitt, C. (2018). *pyGAM: Generalized Additive Models in Python* (0.8.0). Zenodo. https://doi.org/10.5281/zenodo.1208723

Seyedzadeh, S., Pour Rahimian, F., Oliver, S., Rodriguez, S., & Glesk, I. (2020). Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making. *Applied Energy*, *279*(May), 115908. https://doi.org/10.1016/j.apenergy.2020.115908

Sherman, M. H., & Dickerhoff, D. (1998). Air-tightness of US dwellings. *Transactions- American Society of Heating Refrigerating and Air Conditioning Engineers*, *104*, 1359– 1367.

Shimada, M., Okuyama, K., Okazaki, S., Asai, T., Matsukura, M., & Ishizu, Y. (1996). Numerical Simulation and Experiment on the Transport of Fine Particles in a Ventilated Room. *Aerosol Science and Technology*, *25*(3), 242–255. https://doi.org/10.1080/02786829608965394

Shrubsole, C., Ridley, I., Biddulph, P., Milner, J., Vardoulakis, S., Ucci, M., Wilkinson, P., Chalabi, Z., & Davies, M. (2012). Indoor PM2.5exposure in London's domestic stock: Modelling current and future exposures following energy efficient refurbishment. *Atmospheric Environment*, *62*(December), 336–343. https://doi.org/10.1016/j.atmosenv.2012.08.047

Skön, J., Johansson, M., Raatikainen, M., Leiviskä, K., & Kolehmainen, M. (2012). *Modelling Indoor Air Carbon Dioxide (CO2) Concentration using Neural Network*. *31*(1), 37–41.

Smith, J. D., Mitsakou, C., Kitwiroon, N., Barratt, B. M., Walton, H. A., Taylor, J. G., Anderson, H. R., Kelly, F. J., & Beevers, S. D. (2016). London Hybrid Exposure Model: Improving Human Exposure Estimates to NO 2 and PM 2.5 in an Urban Setting. *Environmental Science & Technology*, *50*(21), 11760–11768. https://doi.org/10.1021/acs.est.6b01817

Sokol, J., Cerezo Davila, C., & Reinhart, C. F. (2017). Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, *134*, 11–24. https://doi.org/10.1016/j.enbuild.2016.10.050

Soleimani-mohseni, B. T. *Æ*. M. (2007). *Artificial neural network models for indoor temperature prediction : investigations in two buildings*. 81–89. https://doi.org/10.1007/s00521-006-0047-9

Sousa, G., Jones, B. M., Mirzaei, P. A., & Robinson, D. (2017a). A review and critique of UK housing stock energy models, modelling approaches and data sources. *Energy and Buildings*, *151*, 66–80. https://doi.org/10.1016/j.enbuild.2017.06.043

Sousa, G., Jones, B. M., Mirzaei, P. A., & Robinson, D. (2017b). A review and critique of UK housing stock energy models, modelling approaches and data sources. *Energy and Buildings*, *151*, 66–80. https://doi.org/10.1016/j.enbuild.2017.06.043

Sousa, S., Martins, F., Alvim-Ferraz, M., & Periera, M. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, *22*(1), 97–103. https://doi.org/10.1016/j.envsoft.2005.12.002

Steadman, P., Bruhns, H. R., & Rickaby, P. A. (2000). An introduction to the national Non-Domestic Building Stock database. *Environment and Planning B: Planning and Design*, *27*(1), 3–10. https://doi.org/10.1068/bst2

Swami, M., & Chandra, S. (1987). Procedures for Calculating Natural Ventilation Airflow Rates in Buildings. *ASHRAE Final Report FSEC-CR-163-86*, 130. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Procedures+for+Calculating+Natural+Ventilation+Airflow+Rates+in+Buildings#0%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Procedures+for+calculating+natural+ventilation+airfl

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*(8), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Symonds, P., Taylor, J., Chalabi, Z., Mavrogianni, A., Davies, M., Hamilton, I., Vardoulakis, S., Heaviside, C., & Macintyre, H. (2016). Development of an England-wide indoor overheating and air pollution model using artificial neural networks. *Journal of Building Performance Simulation*, *9*(6), 606–619. https://doi.org/10.1080/19401493.2016.1166265

Tabares-velasco, P. C. (2013). Time Step Considerations When Simulating Dynamic Behavior of High-Performance Homes. *Thermal Performance of Exterior Envelopes of Whole Buildings XII, ASHRAE*, 10.

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., & Finn, D. (2015). Data driven approaches for prediction of building energy consumption at urban level. *Energy Procedia*, *78*, 3378–3383. https://doi.org/10.1016/j.egypro.2015.11.754

Taylor, J., Shrubsole, C., Davies, M., Biddulph, P., Das, P., Hamilton, I., Vardoulakis, S.,

Mavrogianni, A., Jones, B., & Oikonomou, E. (2014a). The modifying effect of the building envelope on population exposure to PM2.5 from outdoor sources. *Indoor Air*, *24*(6), 639–651. https://doi.org/10.1111/ina.12116

Taylor, J., Shrubsole, C., Davies, M., Biddulph, P., Das, P., Hamilton, I., Vardoulakis, S., Mavrogianni, A., Jones, B., & Oikonomou, E. (2014b). The modifying effect of the building envelope on population exposure to PM2.5 from outdoor sources. *Indoor Air*, *24*(6), 639–651. https://doi.org/10.1111/ina.12116

Taylor, Jonathon, Mavrogianni, A., Davies, M., Das, P., Shrubsole, C., Biddulph, P., & Oikonomou, E. (2015). Understanding and mitigating overheating and indoor PM2.5 risks using coupled temperature and indoor air quality models. *Building Services Engineering Research and Technology*, *36*(2), 275–289. https://doi.org/10.1177/0143624414566474

Taylor, Jonathon, Shrubsole, C., Biddulph, P., Jones, B., Das, P., & Davies, M. (2014). Simulation of pollution transport in buildings: The importance of taking into account dynamic thermal effects. *Building Services Engineering Research and Technology*, *35*(6), 682–690. https://doi.org/10.1177/0143624414528722

Taylor, Jonathon, Shrubsole, C., Symonds, P., Mackenzie, I., & Davies, M. (2019). Application of an indoor air pollution metamodel to a spatially-distributed housing stock. *Science of the Total Environment*, *667*(2), 390–399. https://doi.org/10.1016/j.scitotenv.2019.02.341

Taylor, S., Fan, D., & Rylatt, M. (2014). Enabling urban-scale energymodelling: A new spatial approach. *Building Research and Information*, *42*(1), 4–16. https://doi.org/10.1080/09613218.2013.813169

Thatcher, T. L., Lai, A. C. K., Moreno-Jackson, R., Sextro, R. G., & Nazaroff, W. W. (2002). Effects of room furnishings and air speed on particle deposition rates indoors. *Atmospheric Environment*, *36*(11), 1811–1819. https://doi.org/10.1016/S1352-2310(02)00157-7

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy & Buildings*, *49*, 560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Turk, B. H., Prill, R. J., Grimsrud, D. T., Moed, B. A., & Sextro, R. G. (1990). Characterizing the Occurrence, Sources, and Variability of Radon in Pacific Northwest Homes. *Journal of the Air and Waste Management Association*, *40*(4), 498–506. https://doi.org/10.1080/10473289.1990.10466705

Underhill, L. J., Fabian, M. P., Vermeer, K., Sandel, M., Adamkiewicz, G., Leibler, J. H., & Levy, J. I. (2018). Modeling the resiliency of energy-efficient retrofits in low-income multifamily

housing. *Indoor Air*, *28*(3), 459–468. https://doi.org/10.1111/ina.12446

Underhill, Lindsay J., Dols, W. S., Lee, S. K., Fabian, M. P., & Levy, J. I. (2020). Quantifying the impact of housing interventions on indoor air quality and energy consumption using coupled simulation models. *Journal of Exposure Science and Environmental Epidemiology*, *30*(3), 436–447. https://doi.org/10.1038/s41370-019-0197-3

Van Tran, V., Park, D., & Lee, Y. C. (2020). Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality. *International Journal of Environmental Research and Public Health*, *17*(8). https://doi.org/10.3390/ijerph17082927

Vardoulakis, S. (2009). Human exposure: Indoor and outdoor. *Issues in Environmental Science and Technology*, *28*, 85–107.

Venkatesan, P. (2016). WHO report: air pollution is a major threat to health. *The Lancet Respiratory Medicine*, *4*(5), 351. https://doi.org/https://doi.org/10.1016/S2213-2600(16)30014-5

W.S. Dols, B. J. P. (2007). *NIST Technical Note 1887 CONTAM User Guide and Program Documentation*. https://doi.org/10.1071/PP98146

Walton, G.N. (1989). AIRNET - A Computer Program for Building Airflow Network Modelling. *Nistir 89-4072*, *April*, 77.

Walton, George N. (1989). AIRNET - A Computer Program for Building Airflow Network Modelling. In *National Institute of Standards and Technology* (Issue April).

Wang, L., & Chen, Q. (2007). Analysis on the well-mixing assumptions used in multizone airflow network models. *IBPSA 2007 - International Building Performance Simulation Association 2007*, *2001*, 1485–1490.

Wang, L., Dols, W. S., & Chen, Q. (2010). Using CFD capabilities of CONTAM 3.0 for simulating airflow and contaminant transport in and around buildings. *HVAC and R Research*, *16*(6), 749–763. https://doi.org/10.1080/10789669.2010.10390932

Wate, P., Iglesias, M., Coors, V., & Robinson, D. (2020). Framework for emulation and uncertainty quantification of a stochastic building performance simulator. *Applied Energy*, *258*(February 2019), 113759. https://doi.org/10.1016/j.apenergy.2019.113759

Wate, Parag, Coors, V., Iglesias, M., & Robinson, D. (2019). Uncertainty assessment of building performance simulation. In *Urban Energy Systems for Low-Carbon Cities* (pp. 257–287). Elsevier. https://doi.org/10.1016/B978-0-12-811553-4.00007-X

Watson, A. Y., Bates, R. R., & Kennedy, D. (1988). Assessment of Human Exposure to Air Pollution: Methods, Measurements, and Models. In *Air Pollution, the Automobile, and*

Public Health. http://www.ncbi.nlm.nih.gov/books/NBK218147/

Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., & Mandin, C. (2019). Machine Learning and Statistical Models for Predicting Indoor Air Quality. *Indoor Air*, *29*(5), 704–726. https://doi.org/10.1111/ina.12580

WHO. (2012). *WHO releases country estimates on air pollution exposure and health impact*. https://www.who.int/en/news-room/detail/27-09-2016-who-releases-country-estimates-on-air-pollution-exposure-and-health-impact

WHO. (2013). Health Effects of Particulate Matter. In *Health Effects of Ambient Air Pollution*.

Wichmann, J., Lind, T., Nilsson, M. A. M., & Bellander, T. (2010). PM2.5, soot and NO2 indoor-outdoor relationships at homes, pre-schools and schools in Stockholm, Sweden. *Atmospheric Environment*, *44*(36), 4536–4544. https://doi.org/10.1016/j.atmosenv.2010.08.023

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd Editio). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

World Health Organization (WHO). (2021a). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. In *World Health Organization*.

World Health Organization (WHO). (2021b). *WHO Global Air Quality Guidelines*. https://apps.who.int/iris/handle/10665/345329

Xu, L., Hu, O., Guo, Y., Zhang, M., Lu, D., Cai, C. B., Xie, S., Goodarzi, M., Fu, H. Y., & She, Y. Bin. (2018). Representative splitting cross validation. *Chemometrics and Intelligent Laboratory Systems*, *183*(April), 29–35. https://doi.org/10.1016/j.chemolab.2018.10.008

Xu, Y., Shrestha, V., Piasecki, C., Wolfe, B., Hamilton, L., Millwood, R. J., Mazarei, M., & Stewart, C. N. (2021). Sustainability Trait Modeling of Field-grown Switchgrass (Panicum virgatum) Using UAV-based Imagery. *Plants*, *10*(12), 1–22. https://doi.org/10.3390/plants10122726

Yan, D., O'Brien, W., Hong, T., Feng, X., Burak Gunay, H., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings*, *107*, 264–278. https://doi.org/10.1016/j.enbuild.2015.08.032

Yang, L., Ye, M., & he, B. J. (2014). CFD simulation research on residential indoor air quality. *Science of the Total Environment*, *472*(January 2019), 1137–1144. https://doi.org/10.1016/j.scitotenv.2013.11.118

Yu, Y., Megri, A. C., & Jiang, S. (2019). A review of the development of airflow models used in building load calculation and energy simulation. *Building Simulation*, *12*(3), 347–363. https://doi.org/10.1007/s12273-018-0494-0

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S. B., Janes, C. R., Lanphear, B. P., Mccandless, L. C., Takaro, T. K., Venners, S. A., Webster, G. M., & Allen, R. W. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fi ne particulate matter concentrations in a highly polluted city *. *Environmental Pollution*, *245*, 746–753. https://doi.org/10.1016/j.envpol.2018.11.034

Zhang, J., Zhou, Z., Wang, C., Xue, K., Liu, Y., Fang, M., Zuo, J., & Sheng, Y. (2019). Research on the Influence of Indoor Relative Humidity on PM2.5 Concentration in Residential Buildings. *IOP Conference Series: Materials Science and Engineering*, *585*(1). https://doi.org/10.1088/1757-899X/585/1/012086

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049

Zheng, X., Cooper, E., Gillott, M., & Wood, C. (2020). A practical review of alternatives to the steady pressurisation method for determining building airtightness. *Renewable and Sustainable Energy Reviews*, *132*(January), 110049. https://doi.org/10.1016/j.rser.2020.110049

Zheng, X., & Wood, C. J. (2020). On the power law and quadratic forms for representing the leakage-pressure relationship – Case studies of sheltered chambers. *Energy and Buildings*, *226*, 110380. https://doi.org/10.1016/j.enbuild.2020.110380

Appendices

Appendix A. Selected Buildings Characteristics and Layouts

Table A.1: Area Schedule of Space Types within the Sampled Buildings (N is Number of Samples)

Space Type	Barber House		Academic Development Centre		Arts Tower		Regent Court		ICoSS	
	N	Area [m^2]	N	Area [m^2]	N	Area [m^2]	N	Area [m^2]	N	Area [m^2]
Offices	22	451.32	19	291.75	127	5604.84	203	4119	6	244.25
Educational Facilities	0	0	7	424	54	4260.28	23	1897	14	902.64
Shared Facilities	6	45.84	11	165.25	78	2334.56	31	459.69	11	123.44
Circulation	12	144.6	25	367.75	199	2668.88	91	2217.57	18	576.09
Services	5	232.72	9	102.5	137	1533.8	25	363.83	6	101.38
Total	**45**	**874.48**	**71**	**1351.25**	**595**	**16402.36**	**373**	**9057.09**	**55**	**1947.8**

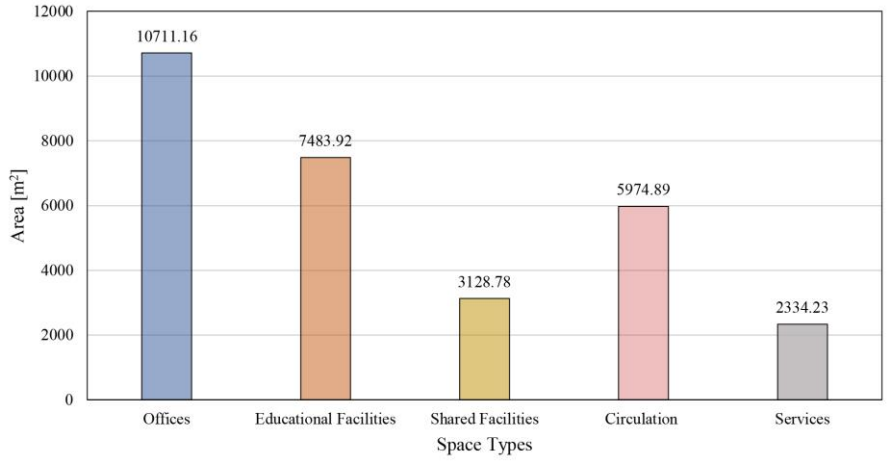


Figure A.1: Built-Up Area of Space Types within the Sampled Buildings

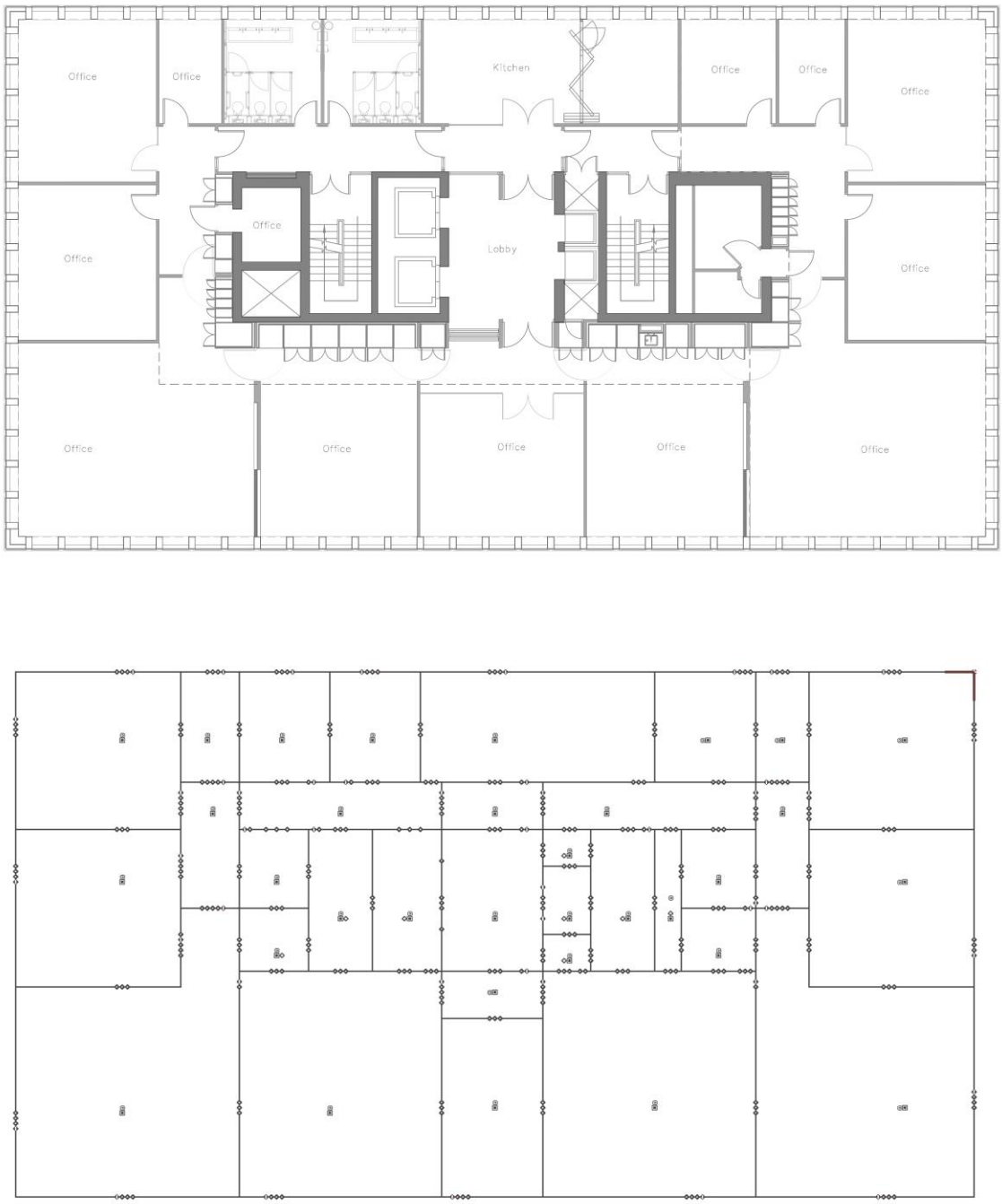


Figure A.2: The 9th Floor of the Arts Tower (AT) – Top: Original CAD Layout and Bottom: CONTAM Model

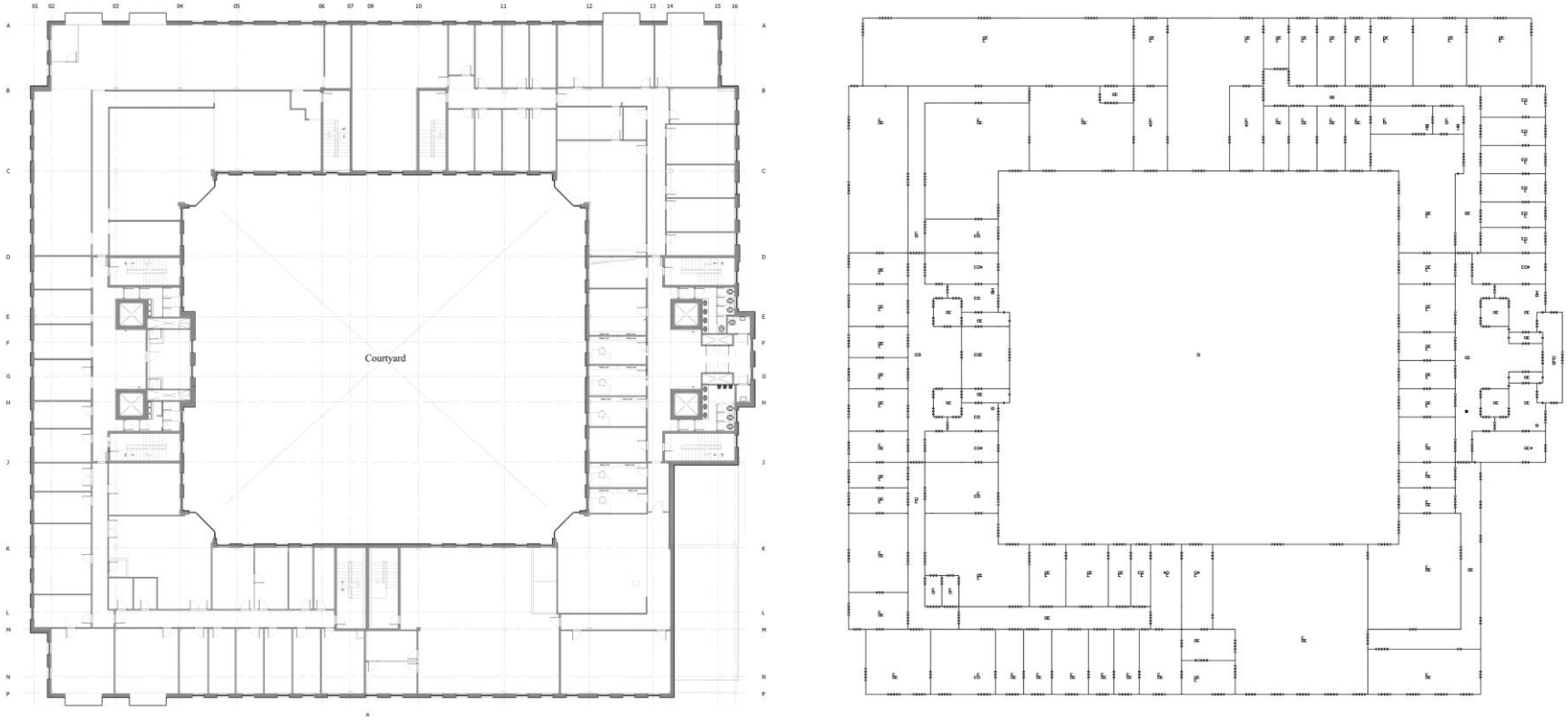


Figure A.3: The First Floor of the Regent Court (RC) Building – Left: Original CAD Layout and Right: CONTAM Model

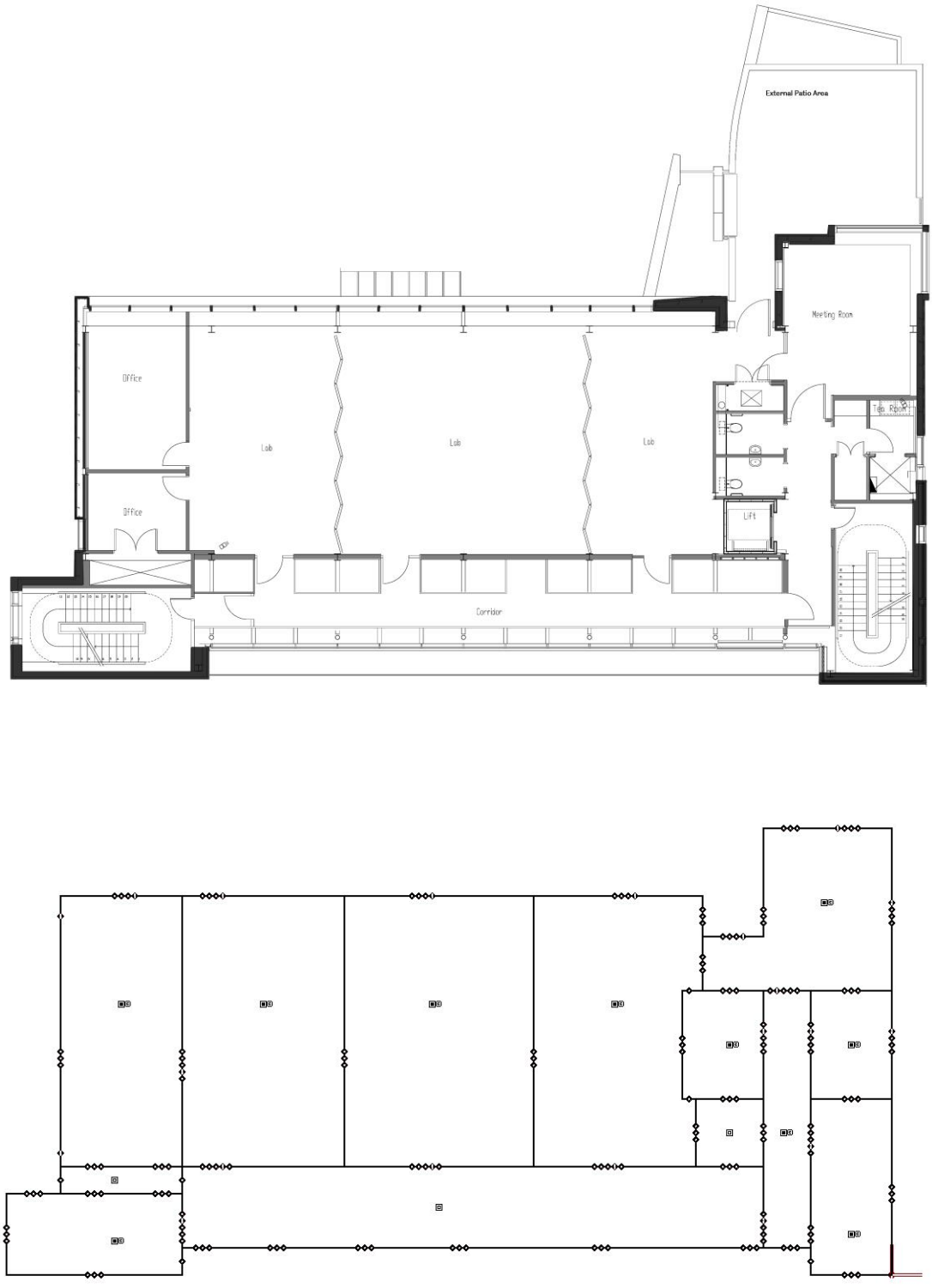


Figure A.4: Typical Floor of the ICoSS Building – Top: Original CAD Layout and Bottom: CONTAM Model

Appendix B. Results of the Sensitivity Analysis Framework

Table B.1: Correlation Matrix of all independent variables.

	A_z	V_z	H_z	j	N_{ef}	A_{ef}	L_{ef}	A_{aw}	L_{aw}	Q_{50}	A_{ef}:V_z	A_{ef}:A_z	DT	RH	υ	ACH_{INF}	Q_4
A_z	1.00																
V_z	0.98	1.00															
H_z	0.19	0.22	1.00														
j	0.10	0.11	0.19	1.00													
N_{ef}	0.46	0.45	0.09	0.01	1.00												
A_{ef}	0.90	0.91	0.24	0.08	0.64	1.00											
L_{ef}	0.80	0.80	0.13	0.06	0.57	0.87	1.00										
A_{aw}	0.85	0.88	0.16	0.07	0.26	0.67	0.60	1.00									
L_{aw}	0.59	0.59	-0.03	0.02	0.16	0.42	0.66	0.70	1.00								
Q_{50}	-0.05	-0.06	-0.22	-0.05	-0.04	-0.08	0.27	-0.03	0.61	1.00							
A_{ef}:V_z	-0.25	-0.24	-0.02	0.01	0.42	0.07	0.05	-0.48	-0.38	-0.05	1.00						
A_{ef}:A_z	-0.21	-0.17	0.05	0.01	0.43	0.13	0.10	-0.40	-0.34	-0.07	0.97	1.00					
DT	0.03	0.05	0.59	0.07	-0.03	0.07	-0.11	0.00	-0.31	-0.51	0.04	0.12	1.00				
RH	0.09	0.05	-0.50	-0.10	0.07	-0.01	-0.02	0.12	0.05	-0.03	-0.21	-0.28	-0.63	1.00			
υ	0.16	0.18	0.89	0.18	0.04	0.19	0.09	0.11	-0.05	-0.20	-0.03	-0.01	0.54	-0.47	1.00		
ACH_{INF}	-0.10	-0.11	-0.15	0.38	0.21	0.04	0.20	-0.24	0.10	0.48	0.44	0.39	-0.38	-0.06	-0.10	1.00	
Q_4	0.05	0.04	-0.15	0.43	0.01	0.02	0.21	0.02	0.37	0.57	-0.06	-0.10	-0.45	0.03	-0.10	0.83	1.00

Results of the Multicollinearity Analysis

Table B.2: Reduction of multicollinearity among independent variables using the Variance Inflation Factor Analysis (VIF). Each trial is associated with the variable eliminated (Target VIF <5)

Benchmark Values		**Trial 1: V_z & A_z**		**Trial 2: L_{aw} &A_{aw}**		**Trial 3: A_{ef}:A_z & L_{ef}**	
Variable	**VIF**	**Variable**	**VIF**	**Variable**	**VIF**	**Variable**	**VIF**
V_z	80.6	A_{ef}:V_z	39.1	A_{ef}:V_z	30.6	ACH_{INF}	13.4
A_{ef}:V_z	46.4	A_{ef}:A_z	29.9	A_{ef}:A_z	27.0	Q_4	11.5
A_z	41.1	L_{aw}	20.2	ACH_{INF}	13.4	H_z	5.8
A_{ef}:A_z	38.7	ACH_{INF}	17.3	Q_4	11.5	υ	5.1
A_{ef}	31.3	Q_4	14.2	A_{ef}	7.9	A_{ef}:V_z	4.9
A_{aw}	22.3	A_{aw}	14.2	L_{ef}	7.5	ΔT	4.0
L_{aw}	20.2	A_{ef}	13.9	H_z	6.4	RH	2.9
ACH_{INF}	17.5	L_{ef}	13.7	υ	5.9	N_{ef}	2.4
Q_4	14.4	Q_{50}	7.4	ΔT	4.1	Q_{50}	2.3
L_{ef}	13.7	H_z	6.5	Q_{50}	3.2	A_{ef}	2.0
Q_{50}	7.4	υ	5.9	RH	3.1	φ	1.6
H_z	6.5	ΔT	4.1	N_{ef}	2.5		
υ	6.0	RH	3.1	φ	1.6		
ΔT	4.2	N_{ef}	2.5				
RH	3.1	φ	1.6				
N_{ef}	3.0						
φ	1.6						

Trial 4: φ & N_{ef}		**Trial 5: Q_4**		**Trial 6: RH & A_{ef}**		**Trial 7: H_z**	
Variable	**VIF**	**Variable**	**VIF**	**Variable**	**VIF**	**Variable**	**VIF**
ACH_{INF}	13.3	H_z	5.6	H_z	5.4	ΔT	2.0
Q_4	11.0	υ	5.0	υ	5.0	ACH_{INF}	1.9
H_z	5.6	ΔT	3.9	ΔT	2.2	Q_{50}	1.6
υ	5.0	RH	2.7	ACH_{INF}	1.9	υ	1.5
A_{ef}:V_z	4.4	Q_{50}	2.1	Q_{50}	1.7	A_{ef}:V_z	1.4
ΔT	4.0	ACH_{INF}	2.0	A_{ef}:V_z	1.4		
RH	2.8	A_{ef}:V_z	1.5				
Q_{50}	2.1	A_{ef}	1.1				
A_{ef}	1.1						

Appendix C. Results of the Cross Validation

Generalised Cross Validation of GAMs using the RMSE, EDoF, and GCV Scores

Table C.1: Results of the Generalised Cross Validation of GAMs showing the RMSE, EDoF, and GCV Scores

	$GAM_{pre\text{-}HPT}$			$GAM_{post\text{-}HPT}$		
	RMSE	**EDoF**	**GCV**	**RMSE**	**EDoF**	**GCV**
Training Dataset	0.550	33.00	0.315	0.679	11.55	0.470

k-Folds Cross Validation of Random Forest Regressor (RFR) and Extreme Gradient Boosting (XGB)

Table C.2: RMSE Score for the 10-fold CV of RFR and XGB (pre-HPT and post-HPT)

Training Dataset	**RFR**		**XGB**	
K-Fold	**pre-HPT**	**post-HPT**	**pre-HPT**	**post-HPT**
1	0.92	0.39	0.85	0.31
2	0.74	0.39	0.85	0.32
3	0.88	0.44	0.97	0.39
4	0.82	0.45	0.92	0.31
5	0.86	0.45	1.03	0.40
6	0.87	0.47	0.98	0.44
7	0.81	0.43	0.87	0.35
8	0.85	0.43	0.91	0.34
9	0.86	0.50	1.09	0.40
10	0.84	0.50	0.92	0.40
Average	**0.85**	**0.45**	**0.94**	**0.37**

Appendix D. Datasets and Python Script for the GAM, RFR, and XGB Metamodels

All codes and datasets have been uploaded to Github and are licenced under Apache 2.0. Terms and conditions for use, reproduction, and distribution are clarified in the README file.

https://github.com/thaerKA1990/PhD_Repo.git