L-Università ta' Malta
Faculty of Science

Department
of Chemistry

Modelling Indoor Household $PM_{2.5}$ using Positive Matrix Factorization and Machine Learning Algorithms

Renato Camilleri

November 2024

Supervisor: Prof. Noel Aquilina
Co-supervisor: Prof. Alfred J Vella

A dissertation presented to the Faculty of Science in fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Malta

L-Università ta' Malta

FACULTY/INSTITUTE/CENTRE/SCHOOL Science

DECLARATION OF AUTHENTICITY FOR DOCTORAL STUDENTS

(a) Authenticity of Thesis/Dissertation

I hereby declare that I am the legitimate author of this Thesis/Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedure

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code A 18.03.2018 and SCI-2023-00065.

X As a Ph.D. student, as per Regulation 66 of the Doctor of Philosophy Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Sacred Theology student, as per Regulation 17 (3) of the Doctor of Sacred Theology Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Music student, as per Regulation 26 (2) of the Doctor of Music Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

□ As a Professional Doctorate student, as per Regulation 55 of the Professional Doctorate Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

Dedication

To my family,

my wife Genevieve and my daughter Judith,

whose unwavering support over the years has made this work possible.

Acknowledgements

I express my deepest gratitude to my supervisor, Prof. Noel Aquilina, and my co-supervisor, Prof. Alfred J. Vella, for their expert guidance, constructive feedback, and encouragement throughout this project.

Special appreciation goes to all participants involved in the sampling campaigns.

I thank the Ministry for Education, Sport, Youth, Research and Innovation, which funded this study through the Tertiary Education Scholarship Scheme 2018.

Furthermore, I wish to acknowledge Prof. James J. Schauer, Robel Kebede, and Erin Mani from the Department of Civil and Environmental Engineering, College of Engineering, University of Wisconsin, USA, for the OC/EC analysis. I also thank Jody Potter at the Department of Natural Resources and the Environment, University of New Hampshire, Durham, for the ion chromatography analysis and Prof. Roy M Harrison and Christopher Stark from the Division of Environmental Health and Risk Management, School of Geography, Earth and Environmental Sciences, University of Birmingham, United Kingdom for the ICP-MS analysis.

I extend my thanks to Dr. Monique Borg Inguanez and Dr. David Suda from the Department of Statistics and Operations Research at the University of Malta for their insightful discussions on data processing methodologies. I am also grateful to Dr. Sara Camilleri for her logistical support during the first sampling campaign.

Finally, I am indebted to my family, mother, friends, and colleagues for their steadfast support and assistance throughout this study.

Table of Contents

List of Tables

List of Figures

Abstract

People tend to spend most of their time indoors, yet the concentration and composition of indoor fine particulate matter ($PM_{2.5}$) remain poorly understood in the Maltese Islands, with existing receptor modelling studies focusing solely on ambient air. The first objective of this study was to carry out long-term indoor air sampling of $PM_{2.5}$ followed by chemical characterisation, in order to identify and quantify the main natural and anthropogenic sources of indoor $PM_{2.5}$ at an urban background site in Malta using Positive Matrix Factorisation (PMF). The second objective was to explore the use of Machine Learning (ML) algorithms to model and predict indoor $PM_{2.5}$ concentrations in several households in Malta and Gozo.

PMF was used to identify and quantify the major sources of indoor $PM_{2.5}$. Quartz and PTFE filters were collected and analysed gravimetrically and chemically using ICP-MS, IC, and an OC-EC aerosol analyser to determine concentrations of $PM_{2.5}$, 18 elements, 5 ions, organic carbon (OC), and elemental carbon (EC). Eight contributing factors were identified, seven outdoor sources and one indoor source, contributing 68% and 26% to indoor $PM_{2.5}$, respectively. Cooking and e-cigarette use were the main contributors to the indoor factor. Uniquely for Malta, a fireworks factor was isolated indoors, responsible for most of the measured Sb and Ba, raising concerns due to the toxicity of these elements. An RF-SHAP model was integrated with the indoor PMF model to investigate the influence of key drivers on indoor $PM_{2.5}$ concentrations. An outdoor PMF analysis was also conducted, and a corresponding RF-SHAP model (CV RMSE: 2.79 μg m^{-3}; R^2: 0.80) was used to refine the outdoor source contributions. Transboundary contributions (Saharan dust and ammonium sulfate) were higher outdoors (58%) than indoors (33%) due to reduced infiltration when windows are closed. Local anthropogenic sources (Industrial, Fireworks, Traffic, Shipping) contributed more to outdoor $PM_{2.5}$ (33%) than indoor (25%), with increased indoor infiltration during warmer months coinciding with peak fireworks activity.

For the second objective, continuous PM measurements were taken using aerosol spectrometers at seven non-smoking residences. RF and XGBoost models were developed to predict indoor $PM_{2.5}$ at six-hourly intervals. At sites with low indoor PM generation, predictions were mainly influenced by outdoor PM_1 levels. At a site with high indoor emissions, indoor relative humidity was a key predictor, especially during cooking. The RF model performed best overall (RMSE: 30.65 μg m^{-3}; IOA: 0.66).

Research Aims, Objectives and Justifications

This PhD study aims to provide a better understanding of indoor $PM_{2.5}$ and its sources in Malta through the use of source apportionment via Positive Matrix Factorisation and Machine Learning algorithms.

This will be achieved through two objectives:

a. The first objective is to conduct a long-term indoor and outdoor air sampling, to chemically characterize $PM_{2.5}$, monitor meteorological parameters and identify its main natural and anthropogenic sources and seasonal variation using source apportionment receptor modelling via Positive Matrix Factorization (PMF). Special interest is being given to the effect of the seasonal burning of fireworks across the Maltese Islands and whether this has an impact on the indoor air quality of residences that are not located in the vicinity of fireworks displays.

b. The second objective is to use modern Machine Learning algorithms, including Random Forest (RF) and Extreme Gradient Boosting (XGBoost), to model and predict indoor $PM_{2.5}$ concentrations in several residences in Malta and Gozo. This requires short term continuous monitoring of indoor and outdoor PM concentrations at several residences in Malta and Gozo.

Limited research has been conducted so far on indoor airborne fine particulate matter in Malta (Aquilina and Camilleri 2022), hence, there is little understanding of the levels and composition of indoor fine particulate matter (PM). This study will provide a better insight of the sources of and exposure to fine PM in Malta and given that most people spend most of their time indoors, it can be used to provide more accurate exposure calculations to fine PM.

Research Outputs Arising from this Thesis

The indoor PMF analysis in Chapter 3 has been published in Atmospheric Environment:
Camilleri, R., Vella, A. J., Harrison, R. M., & Aquilina, N. J. (2022). Source apportionment of indoor $PM_{2.5}$ at a residential urban background site in Malta. Atmospheric Environment, 278, 119093. https://doi.org/10.1016/j.atmosenv.2022.119093

The validation of the microwave-assisted acid digestion method in Chapter 2 has been published in Heliyon:
Camilleri, R., Stark, C., Vella, A. J., Harrison, R. M., & Aquilina, N. J. (2023). Validation of an optimised microwave-assisted acid digestion method for trace and ultra-trace elements in indoor $PM_{2.5}$ by ICP-MS analysis. Heliyon, 9(1), e12844.
https://doi.org/10.1016/j.heliyon.2023.e12844

Chapter 4, Application of Machine Learning Algorithms in Predicting Indoor Residential $PM_{2.5}$ Concentrations has been published in Atmospheric Pollution Research:
Camilleri, R., Harrison, R. M., & Aquilina, N. J. (2025). Application of machine learning algorithms in predicting indoor residential $PM_{2.5}$ concentrations, Atmospheric Pollution Research, 16(10), 102609.
https://doi.org/10.1016/j.apr.2025.102609.

Abbreviations

Autocorrelation Function (ACF)
Black smoke (BS)
Certificate Reference Material (CRM)
Chemical Mass Balance (CMB)
Convolutional Neural Network (CNN)
Cross Validation (CV)
Deep Learning (DL)
Energy Dispersive X-ray Fluorescence Spectrometry (ED-XRF)
EPA, Source apportionment (SA) Machine Learning (ML)
Extreme Gradient Boosting (XGBoost)
Fuel oil (FO)
Generalised Additive Model (GAM)
Generlised Linear Model (GLM)
Index of Agreement (IOA)
Ion Chromatography (IC)
k-Means Clustering (kMC)
Least Absolute Shrinkage and Selection Operator (Lasso)
Long and Short-Term Memory (LSTM)
Long-range transport (LRT)
Multiple Linear Regression (MLR)
Organic Carbon (OC)
Particulate Matter (PM)
Positive Matrix Factorisation (PMF)
Random Forest (RF)
Root Mean Square Error (RMSE)
Secondary Inorganic Aerosols (SIAs)
Secondary Sulfates (SS)
Spectral Clustering (SC)
Standard Reference Material (SRM)
Support Vector Machines (SVM)
Thermal optical analysis (TOA)
Total optical transmittance (TOT)
Total Organic Carbon (TOC)
Urban (U)
Urban Background (UB)
Water insoluble organic carbon (WIOC)
Water soluble organic carbon (WSOC)

Chapter 1

Literature Review

1.1 Particulate Matter

Airborne particulate matter (PM) is a term used to define solid and liquid particles suspended in the air. It is a complex mixture of substances such as dust, smoke, aerosols, and liquid droplets, which may vary in size and chemical composition. The distribution of PM in the environment is influenced by several factors like particle size and meteorological conditions such as wind speed, direction, temperature and humidity, with particles with an aerodynamic diameter of less than 10 microns (PM_{10}) travelling long distances due to their small size. Particulate matter is made up of a variety of chemical constituents, including organic carbon (OC) compounds and elemental carbon (EC), trace metals, ions (predominantly sulfate, nitrate and chloride), major and trace elements and crustal material (such as soil particles or sand). PM can be either directly emitted from anthropogenic or natural sources (primary PM) or else formed through a series of physico-chemical processes (including oxidation, coagulation and deposition) in the atmosphere by volatile organic compounds (VOCs), PAHs, ammonia, oxides of sulfur (SO_x) and oxides of nitrogen (NO_x) (secondary PM).

Particulate matter, unlike other air pollutants such as NO_2, SO_2 and O_3, represents a complex mixture of air pollutants whose health impacts vary depending on particle size and composition. Given its complex nature, this literature review will primarily focus on the fine fraction of particulate matter, namely $PM_{2.5}$, its formation and its outdoor and indoor sources.

1.1.1 PM size distribution

Airborne PM varies in size and composition and is typically defined by the aerodynamic diameter. Particles with an aerodynamic diameter greater than 2.5 μm are considered coarse particles (typically $PM_{2.5–10}$), while those with an aerodynamic diameter equal to or less than 2.5 μm ($PM_{2.5}$) are considered fine particles. Particles smaller than 0.1 μm in aerodynamic diameter are designated ultrafine particles (UFP), with nanoparticles having aerodynamic diameters equal to or smaller than 50 nm (R.E Hester et al. 2016).

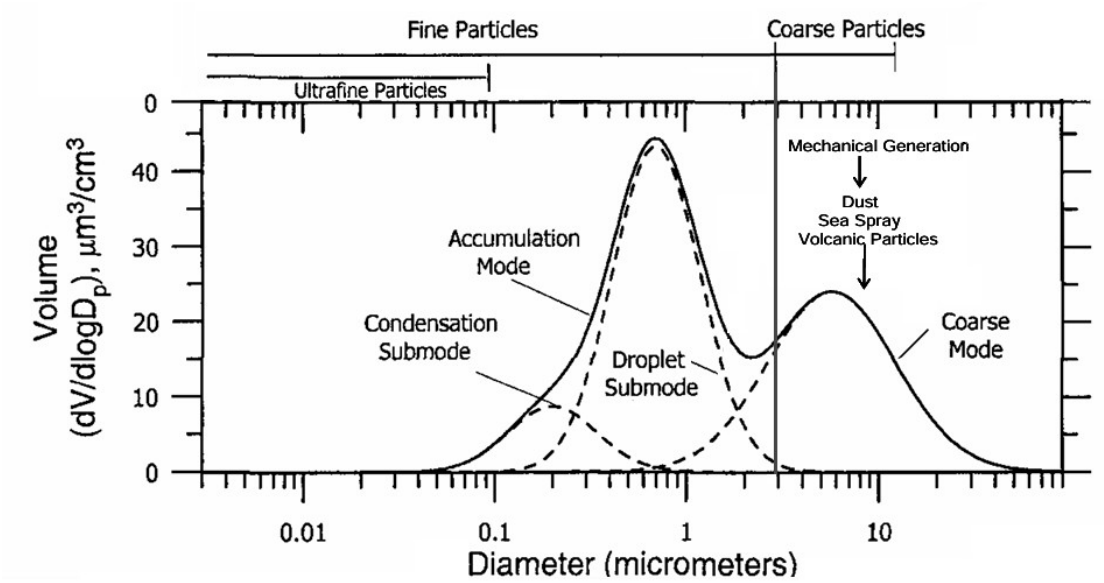


Figure 1.1 Volume distribution of atmospheric particles showing different formation modes (adapted from Seinfeld and Pandis 2016)

This classification arises as different-size fractions typically have distinct sources and formations, as summarised in Table 1.1. The mass distribution is dominated mainly by two modes (Figure 1.1): the accumulation mode (fine particles < 0.2 μm) and the coarse mode (coarse particles > 2 μm). The fine particles result from primary emissions, the condensation of secondary sulfates, nitrates, and organic compounds from the gas phase, and the coagulation of smaller particles. Generally, two generation modes are present for fine particles: the droplet sub-mode, which is created during the cloud processing of some of the accumulation-mode particles and the condensation sub-mode, which is the result of primary emissions and the growth of smaller particles by coagulation and condensation (Seinfeld and Pandis 2016). Coarse particles are mechanically generated primarily by wind and include dust, sea salt, volcanic particles, sand, and pollens. The coarse particles are mostly of primary origin.

Different size fractions also have distinctive health effects since the lung penetration and subsequent assimilation into the body highly depend on the aerodynamic diameter (Brown et al. 2013). Finer particles can penetrate deeper into the respiratory tract as they escape the natural anatomical and physiological mechanisms designed to stop particles from reaching the lungs (Kim et al. 2015). Ultrafine particles with diameters less than 100 nm can become assimilated into the body as they are transferred and deposited into organs within the body. Hence, epidemiological studies are showing that very fine particles (PM_1) may have an even more significant health impact than $PM_{2.5}$ (N. Li et al. 2003; Zwozdziak et al. 2016), which

results in increased mortality, more visits to the emergency room and more lost time from school and work.

Table 1.1 Generation of particulate matter depending on its sizes (adapted from Araujo 2011; Seinfeld and Pandis 2016)

Particle size	Generation mode	Half-life in the atmosphere
Coarse particles $PM_{2.5–10}$	Mechanical action (grinding, and abrasion), evaporation of sprays and suspension of dusts	Minutes to hours
Fine particles $PM_{2.5}$	Primary emission, vapour-to-particle condensation and coagulation (accumulation mode)	Days to weeks

1.1.2 Indoor $PM_{2.5}$

Previous research has demonstrated that outdoor particles can influence indoor air quality through mechanisms such as building ventilation systems and openings in the building structure. Instances of outdoor particulate matter infiltration indoors may involve anthropogenic sources like exhaust and non-exhaust emissions from traffic, shipping and secondary inorganic aerosols (Zhao et al. 2006; Jorquera et al. 2018), together with natural sources like sea salt, volcanic emissions and Saharan dust (Tofful and Perrino 2015; Matthaios et al. 2021).

However, the indoor PM concentrations are not solely due to outdoor infiltration. Indoor sources of PM can arise naturally or from human activities, primarily through indoor combustion (stoves and fireplaces) and cooking. Other significant indoor sources include frying, smoky cooking events, and the use of incense (Chao and Cheng 2003; Wallace et al. 2003; Barraza et al. 2014; Macneill et al. 2014). Smoking can also be a significant contributor to indoor $PM_{2.5}$ in residential homes. In a study in 101 homes in the USA, smoking elevated indoor $PM_{2.5}$ concentrations by 37 μg m^{-3} (Wallace et al. 2003) and contributed 14.2 μg m^{-3} increase in $PM_{2.5}$ exposure to Chinese urban residents (Hu and Zhao 2022). Movement within indoor spaces by occupants can lead to the re-suspension and deposition of particles (Molnár et al. 2014), especially in homes with wall-to-wall carpet rather than smooth flooring (Fromme 2012). The complexity of the process contributing to PM concentrations in the indoor environment is shown in Figure 1.2.

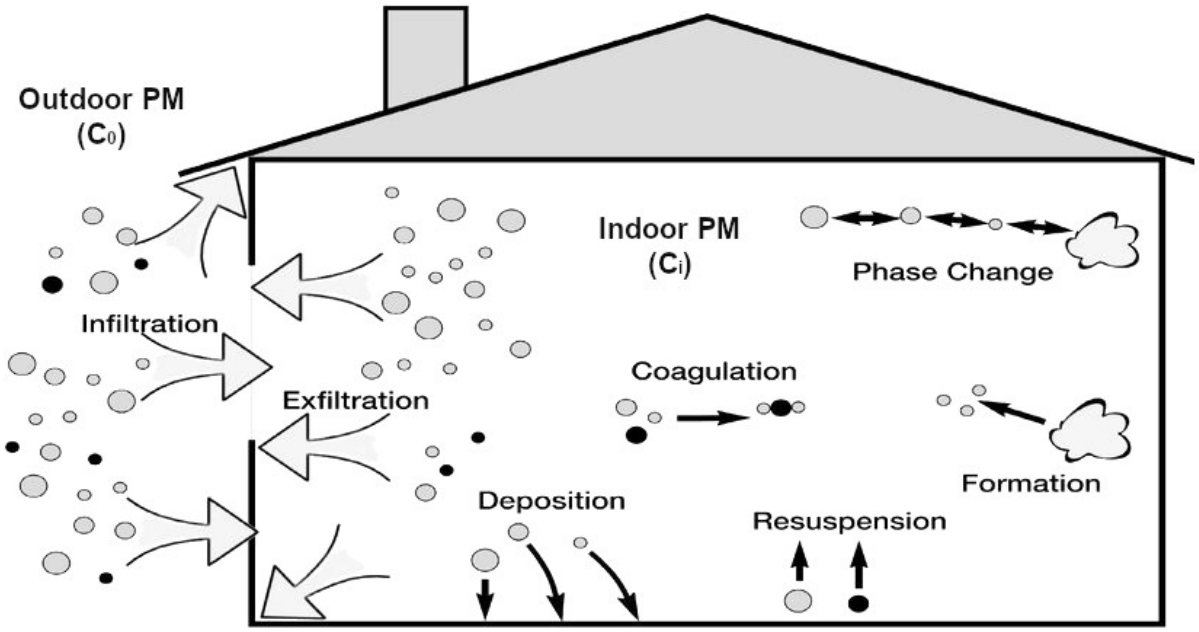


Figure 1.2 Processes that impact the indoor concentration of particulate matter (adapted from Fromme 2012)

In order to properly apportion sources to source profiles from receptor source apportionment models, a proper understanding of the principal and potential sources present is required, together with the main chemical components contributed by each source. This literature review will, therefore, focus on the main local outdoor and indoor particulate matter sources and their chemical constituents. A review of the main receptor models used for source apportionment and a discussion on which ones are best suited to this study is also included. This process will assist in interpreting the source profiles generated by the receptor model and assigning these profiles to valid sources.

1.2 Particulate Matter Sources

1.2.1 Outdoor sources of particulate matter

Primary particulate matter can be formed from both anthropogenic and natural sources. The contribution of different sources varies depending on location, season and time of day, and different sources will also produce primary particulate matter of different size fractions. Table 1.2 summarises the sources of primary particles based on their size fraction.

Table 1.2 Sources of origin and main components of primary coarse, fine and ultrafine particulate matter (adapted from R.E Hester et al. 2016)

PM fraction	Sources	Main components
Coarse particles ($PM_{10-2.5}$)	Tyre and brake wear	Cu, Sb, Zn
	Traffic resuspension	Road dust (Inorganic and organic components
	Construction activities	Ca, Si, Al, Ti and Fe
	Agricultural activities	Soil and OC
	Natural fires and volcanoes	Volcano ash and burned OC
	Biological sources	Plant debris, pollen and fungal spores
	Sea spray	Na, Cl, Mg and SO_4^{2-}
	Saharan dust (Mediterranean region)	Si, Ti, Al, Ca, Mg, Na, K and Fe (Amato et al. 2016)
Fine ($PM_{2.5}$) and ultrafine ($PM_{0.1}$) particles	Diesel fuelled combustion engines	EC
	Biomass combustion	OC, PAHs, K
	Combustion power plants and manufacturing	Ni, As, V, Cd, SO_4^{2-}
	Maritime traffic	EC, OC, Ni, V and SO_4^{2-}
	Metal processing activities	Cd, Pb, Cr, Zn
	Fireworks	EC, OC, Cl^-, K, Mg, Al, Cu, Sr, Ba and Pb (Wang et al. 2007; Tsai et al. 2012)

1.2.1.1 Natural sources

The primary natural sources include windblown dust, sea salt, wildfires, biological particulate matter and volcanoes. On a global scale, natural sources are the main contributors to airborne aerosols, with a contribution of around 98% to all primary and secondary particle formation, of which 84% by mass is derived from sea spray and 13% from mineral dust (M. Viana et al. 2014). The main contributor to the Maltese islands is windblown dust, which includes local

sea spray and, more importantly, transboundary Saharan dust (Scerri et al. 2016) and, to a certain extent, volcanoes. Volcanic ash from major eruptions of Mt Etna has been recorded to affect Malta, albeit being quite rare (Azzopardi et al. 2013).

Sea salt is a major contributor to local dust under 10 μm and, as expected, significantly affects the aerosol contribution of the Maltese islands. The contribution of sea salts depends on many factors, such as water surface temperature, wind, wave height and salinity (Spada et al. 2013). In Malta, the contribution of marine aerosols to PM_{10} was quantified at 19 %, with the highest being in winter and the lowest in summer, possibly due to the higher surface wind speeds and waves experienced during the winter months (Scerri et al. 2016).

The contribution of fresh marine aerosols becomes less significant with particle size. Marine aerosols in Lecce, Italy, were registered in particle diameters between 0.75 μm and 13 μm, with the highest contribution being at a particle size of 4.2 μm (D. Contini et al. 2014). This was also observed in a coastal station in Malta, where fresh sea salt only contributed 4.6% to the total $PM_{2.5}$ mass. Again, the highest contribution was recorded in winter, with the lowest being in summer (Scerri et al. 2018).

Another significant source of aerosol particles in Malta is transboundary windblown Saharan dust. Saharan sources are one of the most active sources of dust in the world (R.E. Hester et al. 2016). Saharan dust is frequently transported across the Mediterranean and southern Europe, sometimes reaching the United Kingdom. These events significantly affect Malta, located only 350 km north of Africa. The contribution of Saharan dust via long-range transport was estimated at 20% of the PM_{10} mass in Malta, with the highest contribution being in Spring (Scerri et al. 2016). The contribution to the total aerosol mass decreases for the finer $PM_{2.5}$ fraction, with Saharan dust contributing 15% by mass to the $PM_{2.5}$ in Msida, a central Mediterranean coastal town (Scerri et al. 2018). This decrease is not as significant as for marine aerosols, and Saharan dust can be expected to be a major contributor to local fine dust. However, little is known about how Saharan dust events affect indoor air quality in Malta, as local studies have focused on the impact of Saharan dust on ambient air quality. This shortfall will be addressed in this study.

1.2.1.2 Power generation and residential combustion

Anthropogenic combustion of fossil fuels and biomass generates significant primary airborne particulate matter. Residential combustion in Europe is the most significant source of particulate matter in European countries (EU-28 region), emitting 36% of PM_{10} and 51% of $PM_{2.5}$ in 2017 (European Environmental Agency 2018). Whilst this may be true for the EU-28 region, residential biomass burning most probably does not contribute as significantly to airborne particulate matter in Malta. Unfortunately, there is a paucity of data concerning this. The State of Environment report issued by the Environment and Resources Authority does not include any data on residential combustion. Moreover, residential combustion was not one of the sources in a source apportionment study carried out in the coastal town of Msida (Scerri et al. 2018). Nonetheless, residential fireplaces are increasing in popularity, and to date, the type of fuel that can be burnt in residential fireplaces remains unregulated.

Power generation is another, albeit much smaller, emitter of airborne particles in Europe, with only 4% of $PM_{2.5}$ for the EU-28 region. There has also been a significant reduction (−69%) in the $PM_{2.5}$ emissions from power generation over the period from 2000−2017 (European Environmental Agency 2018). This trend should also apply to Malta as electricity generation over this period has shifted from solely relying on heavy fuel oil (HFO) to natural gas. Moreover, as of 2014, Malta has obtained a considerable portion of its electrical energy via an interconnector to Sicily. No local emissions are generated from electricity bought from mainland Europe. The power plant is in Delimara, situated on the southeastern tip of the Maltese Islands. Emissions from the plant are typically blown offshore by the prevailing north-westerly winds. In fact, a study carried out in Birżebbuġa found that emissions from the power plant do not have a significant effect on the PM levels of Marsaxlokk and Birżebbuġa. The highest particulate matter concentrations were recorded during high wind speeds from a south-easterly direction, indicating a transboundary source for these high concentrations (Barnes 2013).

1.2.1.3 Maritime emissions

Maritime emissions contribute to fine particulate matter in Malta (Scerri et al. 2018). The Malta-Sicily channel is one of the busiest sailing corridors of the Mediterranean, with over 30,000 tankers and large ships sailing through this channel annually (Ellul et al. 2013). Hence, a significant fraction of maritime emissions are of transboundary origin. Malta is a maritime

hub with a substantial number of large vessels (12,833 for 2018) moving through Malta's ports and 330 Cruise Liners visiting the Grand Harbour in 2018 (TM 2018). As of May 2025, the Mediterranean Sea will be designated as an Emission Control Area for sulfur oxides, requiring ships to burn fuels with less than 0.1% sulfur content, which is expected to decrease $PM_{2.5}$ generated from maritime emissions via secondary particle formation (Directorate-General for Mobility and Transport 2022 Dec 16).

Local emissions from these ships, especially during hotelling, can be significant since most ships still use heavy fuel oil during these operations, with PM emissions from this fuel ranging from 0.56 to 2.12 g kWh^{-1}. Moreover, densely populated areas surround the Grand Harbour, and emissions from these operations are directly emitted into residential areas within these towns and villages. In fact, 5% of $PM_{2.5}$ in Msida was attributed to shipping activities and contributed 20% of the black carbon measured in the area (Scerri et al. 2018) even though the sampling site at Msida is located about 2 km NW of the Grand Harbour and therefore upwind of the prevailing winds in the Maltese islands. To decrease maritime emissions in the Grand Harbour, in 2024, Shore-to-Ship power systems were installed in the Grand Harbour Cruise liner terminal (Infrastructure Malta 2024 Mar 15), yet ships will only be obliged to use the technology by 2030 when the EU FIT for 55 regulations comes into force.

1.2.1.4 Vehicular emissions

Vehicular emissions are a significant source of primary particulate matter, including exhaust and non-exhaust emissions (R.E Hester et al. 2016). The latter include particulate emission due to brake, tyre and road surface wear, and particle re-suspension (Thorpe and Harrison 2008). Besides primary PM, exhaust emissions also contribute significantly to secondary PM formation in densely populated areas. The combustion of petrol and diesel in internal combustion engines is typically one of the major mechanisms for PM formation (R.E Hester et al. 2016). A source profile for traffic sources is typically characterised by very high EC / BC components and loadings from Ca, Cu, Fe, Mn and Sb (Daniela Cesari, Donateo, et al. 2016; Charron et al. 2019). These elements result from non-exhaust emissions with Cu, Fe, Mn, and Sb originating from brake pad wear, with Ca coming from the resuspension of roadside dust (Charron et al. 2019).

Road transport accounted for 10% of PM_{10} and 11% of $PM_{2.5}$ emissions in 2017 for the EU-28 region (European Environmental Agency 2018). Unlike power generation and residential combustion, vehicular emissions are significant in Malta. Traffic contributed 27.3% of $PM_{2.5}$ in the coastal town of Msida (Scerri et al. 2018). This is slightly lower when compared to urban sites in Spain, which registered 39 % – 44% and up to 53% in Barcelona (Viana et al. 2007).

Indoor source apportionment studies in urban sites typically identify traffic as a major source of indoor $PM_{2.5,}$ with traffic (motor vehicles and suspended street dust) found to contribute 39% of indoor $PM_{2.5}$ in Santiago, Chile (Barraza et al. 2014) and 21% for PM_{10} in Lisbon (Almeida-Silva et al. 2016a). Given the relatively high contribution of traffic to $PM_{2.5}$ in Malta, traffic could be one of the major indoor $PM_{2.5}$ sources in Malta and requires investigation and quantification.

1.2.1.5 Airport emissions

The Malta International Airport (MIA) reported a significant increase in aircraft movements over the past years, with annual aircraft movement for 2019 reaching 51,910 (MIA 2019) compared to just under 30,000 a decade ago. Although many studies deal with urban air pollution, very few deal with particulate matter pollution from airports. Aircraft movement at airports can generate particulate emissions from jet fuel combustion (exhaust emissions) and non-exhaust emissions, including brake and tyre wear and dust resuspension during take-off and landing. Exhaust emissions are responsible for particulate matter with higher OC. OC measured in the particulate matter at airports can exceed concentrations at urban sites (Amato et al. 2010). Cu and Sb are produced from brake wear, while Ba, Zn and Mo result from pneumatic tyre wear during landings (Amato et al. 2010; Masiol and Harrison 2014; Almeida-Silva et al. 2016a). Amato showed that the concentrations of Cu and Sb, which are used as tracers for dust emitted from brake pads, followed the same trend as the daily number of landing on the runway at El Prat airport in Barcelona (Amato et al. 2010). The studies did not show that the bulk PM measured at the airport was due to aircraft movement, but higher concentrations for OC (2 $\mu g\ m^{-3}$) and other trace elements such as Sb, Cu, Ba, Zn and Mo were recorded at the airport when compared to simultaneous measurements at the urban background site.

1.2.1.6 Fireworks

Across the Maltese islands, over 30 pyrotechnic manufacturing sites produce fireworks (Malta Tourism Authority 2023) for the local religious festivals (Festa), 85 of which are celebrated during the summer months between the end of June and September (Pace and Vella 2019). The main oxidisers used in fireworks are nitrates of potassium, strontium and barium; potassium chlorate, potassium perchlorate and barium chlorate. Aluminium, magnalium (a 50/50 alloy of magnesium and aluminium) and magnesium are the main metallic fuels locally used in fireworks. Titanium is also used to a lesser extent to create brilliant white sparks. Several metallic salts are also used for colour generation, and these include salts of sodium (yellow), strontium (red), barium (green) and copper (blue). Other toxic metallic salts such as dilead(II) lead(IV) oxide and antimony(III) sulfide are also used. The latter is used in considerable quantities as a tinder in perchlorate flash compositions.

Although fireworks are not commonly listed as a major PM source, they should be mentioned due to Malta's unique nature. Approximately 140 tons of fireworks-related material were estimated to have been burnt in the 2012−2014 period in Malta (Camilleri and Vella 2016). This amount is equivalent to 633 kg per square kilometre per year. Unfortunately, no statistics on the actual yearly consumption of fireworks in Malta are kept as local fireworks factories produce most of the fireworks burnt in Malta, and these are not obliged to submit the actual amounts being produced. Therefore, the amounts must be estimated from the amount of chemicals imported to be used to manufacture fireworks. In comparison, 93,700 tons of fireworks were burnt in the US during the same period (APA 2024). This amounts to only 9.9 kg per square kilometre per year and includes consumer and display fireworks. Therefore, the amount of local burnt fireworks per land area is conservatively at least 60 times that of the US. The figure quoted for Malta is only for the chemicals, while that for the US is for the whole fireworks device. Hence, the value quoted above for Malta underestimates the mass of burnt fireworks.

Moreover, these figures reveal a significant variation in cultural practices. In the US, the volume of consumer fireworks used surpasses that of display fireworks by factors ranging from 7 to 10. This pattern is comparable to that in mainland Europe but contrasts sharply with the situation in Malta, where nearly all firework activities are carried out as display pyrotechnics by licensed enthusiasts (Camilleri and Vella 2016).

Studies have shown that over 38 tons of PM_{10} are emitted from fireworks each year, making fireworks a significant source of PM in Malta (Camilleri and Vella 2016). In fact, fireworks were picked up as a source in a source apportionment study carried out in Msida, with fireworks contributing to 2.3% of the annual $PM_{2.5}$ at the sampling site (Scerri et al. 2018). The contribution by fireworks increases to 6.0% during summer when most fireworks are burnt in Malta.

Several research projects have been carried out recently in Malta that have assessed the impact of the burning of fireworks on ambient air (Camilleri and Vella 2010; Vella et al. 2015; Camilleri and Vella 2016; Pace and Vella 2016). These studies have focused on the metallic content emitted by fireworks (Camilleri and Vella 2010) and, more recently, the amount of perchlorate generated from fireworks burning (Vella et al. 2015). The amount of perchlorate in indoor dustfall has also been studied, but there is no information on indoor fine particulate matter, PM_{10} and $PM_{2.5}$, which make up the respirable and, hence, more dangerous fraction of airborne dust in Malta.

The increase in metals in PM_{10} associated with fireworks burning was reported to decrease significantly from the summer to autumn months in Malta (Camilleri and Vella 2010). In fact, the concentration of the metals in PM_{10} that were analysed, namely, aluminium, copper, barium, strontium and antimony, decreased between the July−August and September–October periods. This decrease is accompanied by a significant decrease in village feasts, indicating that firework displays significantly contribute to Malta's outdoor dust load, especially during the summer months.

Another study conducted during the 2006 fireworks festival (Camilleri 2008), which took place in May before the *Festa* season started, measured the daily PM_{10} concentrations at a location downwind from the fireworks display. A 50% increase (60 $\mu g\ m^{-3}$) in the daily PM_{10} concentration was recorded on the day of the fireworks display. The study also showed a significant increase in all the metals in PM_{10} related to the fireworks. Among these, a 65 times increase was recorded for Ba, and a 106 times increase was recorded for Sb. It is important to note that a 50% increase in the PM_{10} concentration is a daily average. It can be reasonably assumed that the PM_{10} concentration during the actual display, which lasted less than two hours, would have been much higher, as confirmed in other studies on particulate exposure during fireworks displays, such as the 2007 Montreal International Fireworks Competition

(Joly et al. 2010). This shows that any site downwind from the fireworks display can potentially register high PM_{10} concentrations with particles laden with potentially toxic components. The daily increase in PM_{10} was also shown to increase by 2-6 times during the Diwali festival in India (Barman et al. 2008). Fine particulate concentrations measured by (Roberta Vecchi et al. 2008) over a four-hour sampling period reached a maximum of 70 μg m^{-3} during the 2006 FIFA World Cup, and the hourly total particulate matter after a National day celebration reached 180 μg m^{-3} (Moreno et al. 2007). It was also reported that particulate exposure of spectators and residents to outdoor fireworks displays in the fireworks plume less than 2 km from the launch site may be exposed to levels as high as 10,000 μg m^{-3} of fine particulate, albeit for very short periods (Joly et al. 2010). Levels can be sustained for above 1000 μg $PM_{2.5}$ m^{-3} during the 45-minute displays.

Besides particulate metallic content that is typically associated with fireworks, namely K, Mg, Al, Cu, Sr, Ba, Sb and Pb (Wang et al. 2007; Roberta Vecchi et al. 2008; Camilleri and Vella 2010), ions, OC and EC are also formed during fireworks displays (Tsai et al. 2012; Jiang et al. 2015). Ion analysis can be useful in tracing the impact of fireworks on $PM_{2.5}$ dust. A study by Tian et al. 2014 reviewed the total, direct and indirect impacts of fireworks in a megacity in China during the Chinese New Year. This was calculated using both Positive Matrix Factorisation (PMF) and Chemical Mass Balance (CMB) based on peak analysis. It was also noted that K^+ and Mg^{2+} might be effective fireworks tracers. The direct fireworks' contribution to the $PM_{2.5}$ portion was calculated to be higher than that for PM_{10}. In fact, higher non-sea salt (nss) Cl^- was found in $PM_{2.5}$ (72%) than in PM_{10} (57%). This was attributed to the fact that sea salt contributes a higher amount to the coarse fraction of particulate matter (Keuken et al. 2013).

Therefore, K^+ ion concentrations and Cl^-/Na^+ together with OC/EC ratios are suggested as indicators of fireworks burning due to their significant variation in concentration and for the latter in their mass ratios. The Cl^-/Na^+ is found to increase during fireworks displays due to the contribution of chloride ions from the chlorine content in fireworks compositions. K^+ ion concentrations increase significantly during fireworks displays due to the significant use of potassium salts (KNO_3, $KClO_3$ and $KClO_4$) in fireworks. Although K^+ ions are also an indicator of biomass burning (J. Li et al. 2003; Salameh et al. 2015), their use as an indicator of fireworks, especially in areas where biomass burning is not prevalent, should be considered.

Indoor and outdoor $PM_{2.5}$ associated chlorate and perchlorate were investigated in Jinan, China, where the average outdoor and indoor perchlorate in $PM_{2.5}$ were 4.18 ng m^{-3} and 3.54 ng $m^{-3,}$ respectively. The peak concentration of perchlorate in $PM_{2.5}$ was recorded on the Spring Festival Eve, where the outdoor and indoor concentrations rose to 173.76 ng m^{-3} and 125.14 ng $m^{-3,}$ respectively (L. Yao et al. 2015). This suggests that significant indoor $PM_{2.5}$ dust containing perchlorate took place through infiltration even though the windows were closed during sampling. Hence, remaining indoors is not an effective way to limit exposure to aerosol pollutants from fireworks-generated dust. As expected, in the absence of an indoor source, the indoor-to-outdoor (I/O) ratio is less than one (0.85) for the average concentrations of perchlorate in $PM_{2.5}$. The I/O ratios for perchlorate ranged from 0.12 to 16.47, meaning that some samples have an I/O ratio greater than 1. Since the indoor sampling site had no internal source of perchlorate, it suggests that indoor diffusion conditions are poorer than exfiltration; hence, perchlorate-laden $PM_{2.5}$ dust would accumulate indoors and remain at high concentrations (L. Yao et al. 2015). This is of great significance to the Maltese Islands as people spend at least 80−90% of their time in indoor environments (Klepeis et al. 2001; L. Yao et al. 2015; Bo et al. 2017), and indoor pollutants may have significant effects even at low concentrations.

Notwithstanding this, little is known about the indoor concentrations of PM_{10} and $PM_{2.5}$ and their metallic content during fireworks displays, where the residence time of these particles can increase due to the lower air exchange in the indoor environment. Due to the frequency of the fireworks displays and the small geographical area of the Maltese Islands, the burning of fireworks could potentially increase the fine particulate matter laden with toxic particles in indoor air in Malta.

1.2.1.7 Secondary inorganic aerosols (SIAs)

Unlike the primary sources discussed previously, these particles are of secondary origin, resulting from gas-to-particle conversion in the atmosphere. These include secondary nitrates and sulfates, largely formed from the gaseous precursors NH_3, SO_2 and nitrogen oxides (NO_x).

The formation of secondary sulfates in PM starts with the oxidation of SO_2, emitted from sources such as fossil fuel combustion, industrial processes and natural sources, homogeneously by hydroxyl radicals and heterogeneously on droplets and particle surfaces

(Seinfeld and Pandis 2016), forming SO_3 and sulfuric(VI) acid. Sulfuric(VI) acid reacts with NH_3 gas present in the atmosphere to form ammonium sulfate through gas-to-particle conversion. The increased photooxidation rates of SO_2 during the summer months explain the increase in particulate sulfate during the summer months (Song et al. 2001; Kim and Hopke 2008). NH_3 is primarily emitted from agricultural activities (92%), including emissions from livestock and the use of high nitrogen content synthetic fertilisers, but it is also released in urban areas from catalysed gasoline engines (European Environmental Agency 2018).

Secondary sulfates, predominantly as ammonium sulfate, are a significant source of ambient $PM_{2.5,}$ particularly in the Mediterranean. A study in Italy found that 97% of sulfates in $PM_{2.5}$ at both the urban and urban background sites were non-sea salt sulfate (D. Cesari et al. 2016). In a meta-analysis of ambient PM source apportionment (SA) models in Europe, sulfate was reported in 87 records and made up 20 ± 7% of the PM mass concentrations, while nitrate was recorded in 72 studies and constituted 16 ± 6%. In a global SA review, on average, mixed SIA (nitrate and sulfate) contributed to 29% of the ambient $PM_{2.5}$ concentrations (Hopke et al. 2020). In a coastal town in Malta, 23.4% of ambient $PM_{2.5}$ was apportioned to ammonium sulfate, mainly attributed to the long-range transport of air parcels from continental Europe, particularly Italy (Scerri et al. 2018).

1.2.2 Indoor sources of fine particulate matter

Indoor particulate matter has two main sources. The first is of intrinsic indoor origin, and the second is from outdoor infiltration. Outdoor infiltration depends on particle penetration and ventilation conditions. Outdoor penetration depends on particle size and decreases with an increase in particle size (Abt et al. 2000). Hence, the penetration of $PM_{2.5}$ is expected to be higher than that of PM_{10}.

Infiltration can also occur through either natural or mechanical ventilation. The latter can be either filtered or unfiltered. Filtered ventilation is typically found in commercial environments such as offices, hospitals, retail, and large residential complexes. Since indoor environments differ significantly, source apportionment studies typically focus on a single type of indoor environment. Mass balance receptor models such as PMF assume that the particulate matter at all receptor sites originates from the same sources. Since this study is focused on the residential environment, the main indoor sources of $PM_{2.5}$ in households will be discussed. Figure 1.3

summarises the main sources of indoor dust for residential homes, and Table 1.3 summarises the main indoor sources identified in PMF SA studies of $PM_{2.5}$ and the common tracers used to assign source profiles to these sources.

Outdoor $PM_{2.5}$ was found to be a significant contributor to indoor particle concentrations, especially when the air exchange rates are high (Abt et al. 2000; Qing et al. 2005; Barraza et al. 2014). I/O ratios less than 1 indicate no indoor sources, and air pollutants originate mainly from outdoor sources. Zhu et al. 2005 concluded that ultrafine particles (70-100 nm) showed average I/O ratios of 0.6 under infiltration conditions (0.97 h^{-1} air exchange rate, AER). Open windows with an AER of 5.37 h^{-1} AER resulted in I/O ratios very close to 1.0 across all particle sizes. Mechanical ventilation (2.80 h^{-1} AER) with fibreglass filters decreased the I/O ratio. Similar results were reported by Lomboy et al. 2015, for $PM_{2.5}$ in a hospital with mechanical ventilation giving lower I/O ratios than natural ventilation.

Table 1.3 Sources of indoor origin and traces used in PMF SA of $PM_{2.5}$ (adapted from Saraga et al. 2023)

PM fraction	Sources	Main Components / Tracers
Fine ($PM_{2.5}$)	Smoking / ETS	OC, EC, K, Cd, benzo(k)fluoranthene, benzo(ghi) perylene
	Dust resuspension	Ca, Al, Fe, Mg, Mn, Cr, Cu, Ba, Ti
	Cooking	As, Zn, Ca, Mn, Fe, Cd, Cr, Cu
	Combustion	EC, OC, K, Zn, As, Pb PAHs

Indoor-generated particles can either be emitted indoors (primary) or formed through gas-phase precursors via condensation reactions (secondary). However, it is exceedingly difficult to generalise indoor particulate matter sources even within the same city, as this predominantly depends on human activities in the indoor environment. Barraza et al. 2014, found that the socioeconomic status of residences within the same city was significantly associated with differences in indoor $PM_{2.5}$ concentrations. Indoor sources were found to contribute half of the measured indoor $PM_{2.5}$. Improving indoor air quality via filtration of outdoor air, reducing household AER, avoiding indoor smoking or candle burning and controlling fumes generated from cooking through proper hood extraction can significantly impact indoor air quality.

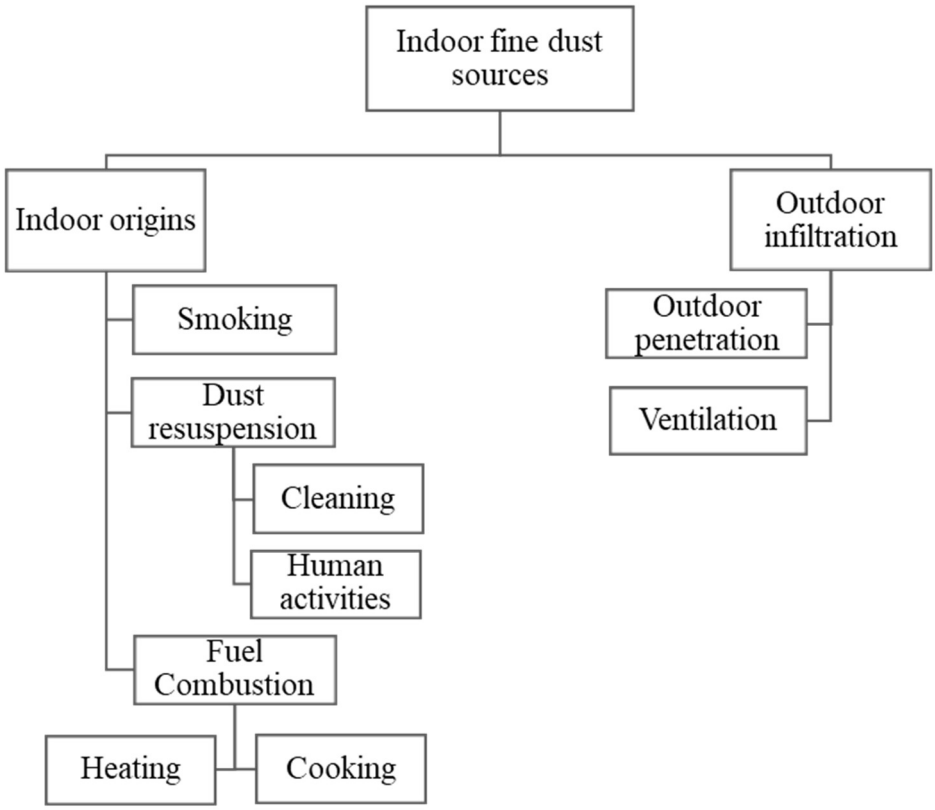


Figure 1.3 Summary of the main sources of indoor dust for residential homes

1.2.2.1 Tobacco smoke and e-cigarettes

Tobacco smoke is a significant source of indoor $PM_{2.5}$. Studies have shown that indoor $PM_{2.5}$ concentrations significantly increase with indoor smoke. The median indoor $PM_{2.5}$ levels in a study in Beijing residences were higher than outdoor levels for smoking residences for both the heating and non-heating seasons (Yang et al. 2018). Moreover, Semple et al. 2012, found that the indoor $PM_{2.5}$ concentrations in Irish or Scottish homes were higher than 35 $\mu g/m^3$ for 60.8 % of the sampling time. In this study, smoking was the main contributor compared to the other source types, which included coal, wood and peat heating and gas cooking. Other studies have shown indoor $PM_{2.5}$ concentrations increase by more than 30 $\mu g/m^3$ in consistently smoking households (Neas 2000; Wallace et al. 2003). Indoor source profiles include BC, Cd and K and, to a lesser extent, Ca, Fe, Pb and Zn as species derived from smoking (Tunno et al. 2016).

In recent years, the use of e-cigarettes has gained significant popularity. These are used as an alternative to conventional cigarettes. Studies have shown that e-cigarettes show a significant

decrease in particulate matter compared to conventional cigarettes, with a 10-fold decrease in total exposure to particulate elements being registered (Saffari et al. 2014). Albeit much smaller contributions than conventional cigarettes, e-cigarettes still registered concerning emission factors for K (7765 ng h^{-1}), Zn (1142 ng h^{-1}) and Pb (96 ng h^{-1}). Ti and Cr were emitted from e-cigarette emissions, yet conventional cigarettes did not produce these metals. Even organic species such as ethanal (0.26% that of cigarettes) and methanal (3.4%) showed significantly lower contributions to indoor concentrations for e-cigarettes when compared to conventional cigarettes, and BC emissions were non-detectable (Ruprecht et al. 2017).

1.2.2.2 Cooking

Besides smoking, cooking can strongly contribute to residential fine particulate matter (Wallace et al. 2003; Li et al. 2016; Huang et al. 2017). Different cooking modes, such as frying/grilling/baking/oven or toaster use, have different particulate emissions. Frying/sauteing was shown to be the cooking mode that produced the highest amount of fine particulate matter (Wallace et al. 2003). The fuel used during cooking significantly affects the amount of fine particulate matter given off indoors, with lower $PM_{2.5}$ concentrations in households using electric/LPG cooking fuel (Huang et al. 2017). A 40-70% reduction of indoor $PM_{2.5}$ was registered in suburban residences in China that used gas/electricity for cooking compared to those using coal as a fuel (Li et al. 2016; Tunno et al. 2016). Most modern households use electricity or gas for cooking, contributing little to indoor fine particulate matter. Cooking source profiles include Ca, Fe and Mn and, to a lesser extent, Al and Zn (Tunno et al. 2016).

1.2.2.3 Resuspension of Indoor Dust

This source of indoor $PM_{2.5}$ is primarily of outdoor origin when settled dust that has either infiltrated indoors or brought indoors on shoes and clothes is resuspended due to various indoor activities, such as walking, cleaning, or even air currents generated by ventilation systems. Resuspension of indoor dust was found to contain the elements Al, Ca, K, Cu and Fe (Zhao et al. 2006; Tunno et al. 2016). The presence of Cu was attributed to the use of electric motors such as fans or vacuum cleaners, as these activities would result in the resuspension of indoor dust.

1.2.2.4 Sources of emergent pollutants

Perfluorinated compounds (PFCs) and polybrominated diphenyl ethers (PBDEs) are emergent persistent organic pollutants (POPs) that have become ubiquitous in the indoor environment (Anake and Nnamani 2022) and are known for bioaccumulative effects in addition to their adverse health effects including neurotoxicity, thyroid toxicity and cancer (Chen et al. 2019). PBDEs are compounds that are covalently bonded to polymer-based products to increase ignition resistance, while PFCs are typically used as lubricants and detergent materials. Indoor dust, particularly $PM_{2.5}$, acts as a sink for these POPs due to the large surface area of fine PM. Evidently, no source apportionment studies have been carried out so far on these emergent pollutants in $PM_{2.5}$ in the indoor environment.

Four sources of PBDEs in indoor dust in Turkey have been identified using PMFv5. The PBDEs sources were the emission of commercial pentaBDE (43%), the commercial use of octa- (27%) and decaBDE (14%) and the combination of debromination and insulation (15%), thereby suggesting that the best way to reduce indoor PBDEs is to reduce the commercial use of pentaBDE (Basaran and Yılmaz Civan 2021). This study acknowledges that measuring these POPs in $PM_{2.5}$ in addition to indoor dust would have enabled a proper measurement of the air quality of indoor environments. Measuring POPs in PM is challenging in terms of sampling and analytical measurement (Anake and Nnamani 2022) and even more so for indoor $PM_{2.5}$. Analysing indoor dust for PBDEs resulted in half the analysed PBDEs below the LOD. This problem would be exacerbated for indoor $PM_{2.5}$.

1.3 Receptor Modelling

Source apportionment (SA) using receptor models (RMs) enables the identification of pollution sources and quantifying the amount each source contributes at the receptor site. The fundamental principle of receptor modelling is the mass conservation between the emission source and the receptor site. A chemical mass balance (CMB) analysis can be used to identify and apportion sources of atmospheric pollutants. RMs identify sources by solving equation 1.1.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \qquad i = 1,2,\dots,n \quad j = 1,2,\dots,m \qquad \textbf{Eq. 1.1}$$

x_{ij} = concentration (μg m^{-3}) of PM species *j* measured in sample *i*;

g_{ik} = contribution of source type *k* to sample *i*;

f_{kj} = concentration (μg m^{-3}) of PM species *j* in source type *k*;

e_{ij} = the residual term, i.e. the difference between the measured and the fitted value.

n = number of samples

m = number of PM species

p = number of independent sources

Several SA models are available, the most common ones being Principal Component Analysis (PCA), Unimix, Chemical Mass Balance (CMB) and Positive Matrix Factorisation (PMF). RMs use the measured pollutant concentration at the receptor site, although the type of data inputs in various models varies. There are three kinds of input: receptor ambient pollutant mass concentrations, source profiles and meteorological data (wind speed and direction). Extended receptor models have been developed to process additional information about the sampling period or conditions. The models would also include information such as seasons, weekdays and precipitation (Belis et al. 2013).

Receptor models based on equation 1.1 depend on two main assumptions, although not to the same extent. The first assumption is that source profiles do not change significantly over time, or if they do change, they do so in a reproducible manner. For example, CMB depends on this, while PMF, being a factor analysis, depends on the source profile changing uniformly from one sampling interval to the next. The second assumption is that receptor species do not react chemically or undergo phase changes during the transport from the emission source to the receptor site. When these assumptions are not fulfilled, then source contributions by the RMs may be under/overestimated (Belis et al. 2013).

With models such as PCA, PMF and UNIMIX, little knowledge is required about pollution sources before the model can be run. On the other hand, models such as CMB require significant knowledge about pollution sources. CMB requires source profiles for well-defined local sources, which might not be readily available (Belis et al. 2013). In Malta, there are no well-defined source profiles that can be used for CMB. The clear advantage of CMB is that

the dataset can be quite small since the source profiles are provided in the model. In theory, the model can be run for just one sample, although a minimum of fifteen samples is typically suggested to increase the statistical performance of the model (Belis et al. 2014a).

1.3.1 Positive Matrix Factorisation (PMF)

PMF requires little to no knowledge of the pollution sources. This model does not require any *a priori* knowledge of source profiles, and hence, in cases such as Malta, where no previous study has been carried out on local source profiles, PMF is, therefore, superior to CMB.

The solutions require multiple iterations implemented in a computer program, with the most used software being the EPA PMF.

$$Q(E) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{x_{ij}-\sum_{k=1}^{p} g_{ik}f_{kj}}{\sigma_{ij}}\right)^2 = Minimum \qquad \textbf{Eq. 1.2}$$

The residual term (e_{ij}) is derived from the difference between x_{ij} (the measured amount) and the sum of the product for $g_{ik}f_{kj}$ (the modelled amount). Therefore, the main task of the iteration is to minimise the Q-value, which is given by the sum of squares of the residual term (e_{ij}) weighted inversely with the error estimates (σ_{ij}) of the data points. The Q-value represents the goodness-of-fit of the model, and the model carries out iterations to obtain the lowest value of Q. A stable solution to the equation is reached when further iterations to decrease Q prove ineffective. Both PMF and CMB require an uncertainty estimation of each data point.

PMF is an advanced factor analysis technique based on the work of Paatero and Tapper (1994). It imposes non-negativity constraints in the factor computational process. This receptor model has been applied for SA of fine PM in ambient air (Kim and Hopke 2008; Belis et al. 2013; D. Cesari et al. 2016; Manousakas et al. 2017) and has also been used for indoor SA (Minguillón et al. 2012; Barraza et al. 2014; Molnár et al. 2014; Suryawanshi et al. 2016).

The model can be written as

$$X = G \cdot F + E \qquad \textbf{Eq. 1.3}$$

where the species concentration matrix X (that is, the known n by m matrix for m number of PM species in n samples) can be factorised into two matrices, G, the factor contribution matrix (given by the $n \times p$) and F, the factor profile matrix (given by the $p \times m$ matrix of source profiles), where p represents the number of factors extracted. E is the residual matrix, which is the difference between the measured X and the model $G \times F$ and hence is the unapportioned part.

The source profiles and source contributions provided by the model are not assigned to specific sources, and this assessment has to be carried out by the analyst. Source profiles are assigned to probable sources depending on the pollutants present in the source profile and the percentage contribution of each pollutant to that source profile.

PCA has frequently been used in air quality to carry out factor analysis. PCA can be used to estimate source contributions by carrying out a multilinear regression analysis of the principal components against the particulate matter mass. This method can be used as an exploratory tool or to obtain a preliminary picture of source contributions, as other models, such as PMF, offer a more robust analysis (Belis et al. 2014; Daniela Cesari, Amato, et al. 2016). PMF offers a significant advantage over factor selection by PCA, as the quantification of source contributions with PCA can be more inaccurate. In PCA, source contributions (in terms of mass) are subject to uncertainty due to the absence of the non-negativity constraint. PMF restricts all G (or g_{ik}) elements and F (f_{kj}) to positive values. On the other hand, in PCA, the output could show negative source contributions (in terms of mass), which has no physical meaning.

Although PMF offers the advantage of SA without needing *a priori* knowledge of source profiles, several conditions must be met for the model to work effectively. PMF requires a large dataset. The user guide for the PMFv5 manual does not specify the exact number, but it suggests using over 100 samples with 10 to 20 species (Norris et al. 2014). Several studies have reported that 50 chemically analysed samples are the absolute minimum for PMF (Johnson et al. 2011; Belis et al. 2014). Moreover, increasing the number of species helps reduce collinearity between different source profiles. The need for a large dataset and the chemical analysis of multiple species significantly increases sampling and analytical costs. Typically, a minimum of fifteen elements, OC and EC, and NH_4^+, SO_4^{2-}, NO_3^- and Cl^- ions are required to

obtain a good mass balance for the model. PMF also requires significant chemical knowledge, as source profiles have to be assigned to sources.

The latest version of the PMF software, PMFv5, uses the Multilinear Engine (ME-2) to constrain PMF profiles and contributions. For example, ratios of *F* elements/ions such as Cl^-/Na^+ ratios can be used to indicate a particular source, such as sea salt. Custom expressions can be set up to build equations for the *F* and *G* elements based on chemical knowledge of the aerosols at the sampling site. These constraints impose external information on the PMF solution, forcing the concentration of species in the *F* matrix or the contributions in the *G* matrix to a known value, thereby reducing rotational ambiguity (Salameh et al. 2018; Scerri et al. 2019) and enabling the PMF solution to be solved with a smaller data set.

These recent developments have also been used to validate the use of PMFv5 for smaller datasets. A recent case study from a receptor in southern Italy (Scerri et al. 2019) tested the PMF source apportionment solution on a small dataset. The study focused on the apportionment of a combined data set from 29 PM_{10} and 33 $PM_{2.5}$ samples from a rural site in southern Italy. Constraints were added to the factor solution based on knowledge of the chemical composition of the aerosols in the affected region. Known species concentrations were forced onto the *F* matrix, thus solving rotational ambiguity problems. A 5-factor solution was resolved for the combined dataset. These factors were compared with Constrained Weighted Non-negative Matrix Factorisation (CW-NMF) and Chemical Mass Balance (CMB) models. Both models gave similar factor contributions to the PMF analysis, with the CW-NMF giving the closest results.

1.3.2 Indoor PMF source apportionment studies

While SA of ambient $PM_{2.5}$ is well-established (Belis et al. 2013; Hopke et al. 2020), indoor SA is still a developing field of study and presents significant challenges when compared to ambient SA studies (Saraga et al. 2024). Ambient $PM_{2.5}$ concentrations are regulated in several countries (U.S.C. 1990; Council directive 2008/50/EC 2008), but indoor $PM_{2.5}$ is not. Given that people spend most of their time indoors, residential, indoor SA studies are becoming important as they provide a better quantification of the exposure to $PM_{2.5}$ than ambient SA studies. Compared to ambient air studies, only a limited number of studies have carried out SA via PMF on indoor $PM_{2.5}$, and very few have been carried out in the residential environment.

Comparing data from different indoor SA studies is not always easy since the sources' absolute and relative contributions may differ as these depend on the indoor environment, building types and ventilation, SA method, urban environment, and seasonal and cultural activities. Nonetheless, certain important qualitative comparisons can be made, especially on the main sources affecting indoor environments from similar indoor and outdoor environments. A summary of indoor residential SA studies is given in Table 1.4 (a) for PMF and (b) for CMB. Table 1.5 includes a few SA studies on residential indoor $PM_{2.5}$ via PCA and factor analysis.

In a first integrated literature review on source apportionment of indoor air (Saraga et al. 2023), 127 studies met the review criteria, with 53% being carried out on indoor $PM_{2.5}$ and 48% of the total studies being carried out on residential buildings, followed by schools and universities (29%) and office buildings (11%). Only 31% of the total studies used PMF receptor models. Most studies used other SA techniques such as PCA, Multiple Linear Regression (MLR), CMB or enrichment factors. The scarcity in indoor residential PMF studies can be mainly attributed to the following factors:

1. PMF requires an extensive dataset. The logistics required to obtain a vast sample size and extensive chemical speciation required for PMF, combined with the financial costs of such sampling campaigns necessary for PMF SA, make such campaigns operationally complex. The logistical and financial problems are exacerbated when sampling is conducted indoors, as such campaigns typically require indoor and outdoor environments. This translates into multiple samplers being installed at the sampling site, making the sampling operation more complex, costly, and demanding for staff.

2. Different chemical species are required for a comprehensive PMF solution with organic, elemental and ion analysis typically being required. The analysis normally requires more than one filter medium, as quartz filters are required to analyse the organic fraction, and PTFE filters are required to analyse the inorganic fraction (elements and ions). Hence, multiple indoor $PM_{2.5}$ samplers are required. $PM_{2.5}$ samplers are bulky and noisy, making it difficult to find residents willing to participate in long-term sampling campaigns (Saraga et al. 2024).

3. Another hurdle to indoor PMF SA is intrinsic to the PMF solution. One of the main assumptions of PMF is that source profiles are assumed to be the same across all

samples. The indoor environment has a higher variability of indoor sources, especially those related to human activity, such as cooking and cleaning. Different cooking and cleaning methods can generate different profiles, while different lifestyles can affect emissions from cleaning and resuspension of indoor dust. House ventilation can affect the infiltration rate and the effect of outdoor sources on indoor air quality. Moreover, VOC emissions can vary between households, and the indoor chemistry can affect the oxidation potential, generating new SOA (Saraga et al. 2023). Nonetheless, several indoor PMF studies result from combined indoor datasets from several households. To mitigate variability of indoor sources, most of the indoor PMF studies limit the SA to a specific indoor environment such as residential homes (Barraza et al. 2014; Molnár et al. 2014; Matthaios et al. 2021), schools or office buildings (Saraga et al. 2010; Carrion-Matta et al. 2019), or else assess specific outdoor effects such as dust events and their effect on indoor $PM_{2.5}$ (Wang et al. 2024).

From the data summarised in Table 1.4a, most of the studies analysed indoor and outdoor residential environments, and some were part of personal exposure (PE) campaigns. All studies carried out a gravimetric determination of $PM_{2.5,}$ with several studies carrying out sampling on different filter media to analyse the organic and inorganic fractions. Trace metal analysis was carried out predominantly with either ICP-MS or XRF. To address the low indoor $PM_{2.5,}$ most indoor air samples are usually time-averaged, with 24-h, 48-h and even 96-h time-averages, which precludes the identification of peak events. Moreover, data is typically collected from a single room, most frequently the living room, which is typically not representative of a whole residence or building.

Only a couple of studies measured OC and EC analytically. Many studies measured BC using a non-destructive smoke stain reflectometer. In some studies, organic matter (OM) was calculated from OC using the empirical relationship, $OM = 1.6 \times OC$ (Shang et al. 2019), while sulfur analysed using XRF was assumed to be ammonium sulfate (Wang et al. 2024). In fact, none of the reviewed indoor residential SA via PMF studies measured ionic content via ion chromatography. Some studies had to combine samples or extend the sampling period beyond 24 hours to collect enough $PM_{2.5}$ for chemical analysis.

Indoor SA studies show that the outdoor contribution to indoor residential $PM_{2.5}$ varies considerably, from 20% to 85% of the indoor $PM_{2.5}$ concentration. The main outdoor sources

identified include traffic, sea salt, secondary sulfates, long-range transport, shipping and industry. The outdoor sources were analysed either through separate SA of the indoor and outdoor datasets or by combining indoor and outdoor datasets into one (Minguillón et al. 2012).

Most studies identified Smoking or Environmental Tobacco Smoke (ETS) as a $PM_{2.5}$ source, with contributions ranging from 2 – 10% of the indoor $PM_{2.5}$ concentration. In the absence of PAH (such as benzo(*k*)fluoranthene, benzo(*ghi*)perylene) analysis, ETS could not be resolved as an individual source profile and was typically combined with other indoor sources (Barraza et al. 2014; Matthaios et al. 2021), with high levels of potassium used as a marker for ETS. A PMF source profile in a SA study in Massachusetts, USA, was labelled Fireworks & ETS due to the elevated levels of K, Ba and Sr in the profile and the higher contributions of this source profile during the first week of July (Matthaios et al. 2021). In Santiago, Chile, cooking and ETS were combined as a single source profile (Barraza et al. 2014). In this study, the occupants' activity record assisted in identifying the presence of ETS in the source profile. Box plots were constructed for each indoor source contribution against a categorical variable that was 1 or 0, depending on the presence of smokers inside the household. One source profile showed a shift to higher concentrations for households with smokers.

Cleaning or cleaning combined with cooking (10 to 53%) are also recorded as indoor sources. Such sources are identified by high OC and high levels of Zn and Ca in the source profiles. Cleaning is typically identified by higher levels of Cl in the source profile (Barraza et al. 2014; Saraga et al. 2023). Cooking and combustion sources require further analysis of PAHs and VOCs to identify specific cooking and combustion sources. CMB, with the use of *a priori* source profiles is commonly used to identify specific indoor sources.

Indoor dust or indoor resuspension is also identified as a main contributor to indoor $PM_{2.5}$ (7 – 51%). This indoor source results from outdoor infiltration or dust brought indoors adhered to shoes and clothing, which is then resuspended via indoor activities. The source profile is typically identified from the abundance of crustal elements (Si, Ca, Fe, Al and Ti), although indoor dust resuspension represents higher percentages of PM_{10} concentrations when compared to $PM_{2.5}$ (Qian et al. 2014). Given the outdoor origin of the resuspended dust, it is not easy to identify indoor resuspension from outdoor dust based solely on the source profile. Separate identification is typically possible when both outdoor and indoor SA are carried out, and the indoor elemental source profile lacks correlation with the outdoor sources (Molnár et al. 2014).

Table 1.4a Summary of PMF studies for indoor residential $PM_{2.5}$

Study					Results			References
Receptor	**Period of Study**	**Type of Study**	**Chemical species analysed**	**Analysis**	**I-O concentrations ($\mu g\ m^{-3}$)**	**Indoor PMF source contribution to $PM_{2.5}$ as % ($\mu g\ m^{-3}$)**		
North Carolina (USA)	June 2000 to May 2001	I-O residential PE (non-smoking)	PM OC-EC Elements	Gravimetric TOA EDXRF		Soil Traffic Secondary sulfate Secondary nitrate ETS Personal care & activity Cooking Cu factor	3.7 9.4 22.5 4.7 10 19.1 53.6 1.2	(Zhao et al. 2006)
Gothenburg* (Sweden)	2002 April 2nd – June 7th Sep 26th – Nov 6th 2003 March 27th –June 12 th October 7th – 30th	I-O residential PE	$PM_{2.5}$ BS Elements PM_{rest} (59%)	Gravimetric Smoke Stain Reflectometer EDXRF Calculated	Outdoor 7.8 Indoor 9.7	LRT Marine Traffic Soil resuspension Indoor resuspension Indoor Cu	8 (0.8) 5 (0.5) 22 (2.2) 4 (0.4) 51 (5) 5 (0.5)	(Molnár et al. 2014)
Barcelona† (Spain)	Nov 2008 to Nov 2009	I-O residential PE (pregnant women)	$PM_{2.5}$ BS Major & trace elements PAHs	Gravimetric Smoke Stain Reflectometer ICP-AES ICP-MS GC-MS	Outdoor 20 Indoor 24	Sulfate FO + sea salt Mineral (Crustal) Smoking Traffic Industry Unassigned	10 2 5 2 4 1 75	(Minguillón et al. 2012)

Study					Results			References
Receptor	**Period of Study**	**Type of Study**	**Chemical species analysed**	**Analysis**	**I-O concentrations (µg m^{-3})**	**Indoor PMF source contribution to $PM_{2.5}$ as % (µg m^{-3})**		
Massachusetts (USA)	2006 - 2010	Indoor residential	$PM_{2.5}$	Gravimetric	Indoor 7.64	Indoor mixed dust Salt aerosol Fireworks and ETS Outdoor pollution	17 14 7 62	(Matthaios et al. 2021)
			BC	Smoke Stain Reflectometer				
			Elements	EDXRF				
Santiago (Chile)	2012 Nov 2nd – June 7th	I-O residential	$PM_{2.5}$	Gravimetric	Outdoor 18	Sulfates Cooking and ETS Traffic	16 28 26	(Barraza et al. 2014)
			OC-EC	TOT (NIOSH870)	Indoor 21	Cleaning and cooking Indoor dust Street dust	11 7 13	
			Elements	XRF				
Cyprus	2019 -2021 Feb – June Sep – Dec	I-O residential	$PM_{2.5}$	Gravimetric	Outdoor 15.1	SS Ca rich particles Biomass burning	48 31.9 10.4	(Wang et al. 2024)
		$PM_{2.5}$ and PM_{10}	BC BrC	SootScan Optical Transmisso-meter	Indoor 13.8	Traffic Mixed factor	5 4.7	
			Elements	XRF				

Notes: *Calculated as not given directly in the study

† The high percentage of unassigned mass is mainly because this study did not analyse ions.

Table 1.4b Summary of CMB studies for indoor residential $PM_{2.5}$

Receptor	Period of Study	Type of study	Chemical species analysed	Analysis	I-O concentrations (µg m^{-3})	Indoor CMB source contribution to $PM_{2.5}$ as %		References
Hong Kong††	Oct 1999 – March 2020	Indoor residential $PM_{2.5}$ and PM_{10}	Elements	PIXE		Smoking Cooking Incense burning Human activities Outdoor	6.7 68.1 5.8 13.2 31.9	(Chao and Cheng 2003)
Los Angeles††† **(USA)**	Summer 2005 – Winter 2006	I-O Retirement homes	$PM_{2.5}$ 92 organic compounds Elements WSOC OC-EC	Gravimetric GC-MS ICP-MS Sievers TOC analyzer Semi-continuous OC-EC analyzers	Outdoor 22−27 Indoor 16−23	Vehicular Shipping Other WSCOC Sulfate Wood smoke Resuspended dust Other WIOC	39 1 6 18 3 11 20	(Hasheminassab et al. 2014)

Notes: †† Results are from a 96h averaged sampling period.

††† Five filters were combined for the GC-MS.

Table 1.5 Source contributions from factor analysis of indoor residential $PM_{2.5}$

Study					Results			References
Receptor	**Period of Study**	**Type of Study**	**Chemical species analysed**	**Analysis**	**I-O concentrations (µg m^{-3})**	**Percentage variance**		
Braddock and Clairton (USA)	2011 July 25th − Sep 13th 2012 Jan 30th − March 5th	Indoor PE residential (asthmatic cohort)	$PM_{2.5}$ BC Elements NO_2	Gravimetric Smoke stain Reflectometer ICP-MS Passive sampling UV-VIS	Indoor Summer 25.8 Winter 18.9	Summer Smoking, Cooking and Steel Making Motor Vehicle, smoking Coal, Braking Cooking Coal Winter Cooking Smoking Cooking Coal, Brake tyre	40 20 12 7 5 48 18 12 6	(Tunno et al. 2016)
Beijing (China)	2013 August – October Nov–March (2014)	I-O residential	$PM_{2.5}$ Elements	Gravimetric ICP-MS	Outdoor 72.5 Indoor 57.6	Combustion Traffic Cooking Dust and soil Smoking	22 21 14 11 9	(Yang et al. 2018)

1.4 Extraction and Analysis of Trace Metals in $PM_{2.5}$

Quantifying trace metals in indoor fine fractions, $PM_{2.5}$ and $PM_{1,}$ poses analytical challenges as small amounts of PM are collected from 24-hour samples when operating low-volume samplers. Increasing the sampled volume via high-volume samplers is usually not possible as these samplers are not practical for many indoor environments. Indoor $PM_{2.5}$ samples typically contain less than 500 µg of particulate mass and are usually associated with low elemental concentrations for most analytes due to lower indoor emissions. These drawbacks could be compensated with longer sampling times, resulting in a lower time resolution, which averages out peak events even further. Traditional analytical methods using Energy Dispersive X-ray Fluorescence Spectrometry (ED-XRF) have been used effectively for the elemental composition of PM (Mazzei et al. 2008; Scerri et al. 2018; Seibert et al. 2020). This technique offers the advantage of little sample manipulation and, hence, high sample throughput; however, such methods suffer from relatively high detection limits, making them unsuitable for detecting trace or ultra-trace quantities.

ICP-MS coupled with microwave-assisted acid digestions can provide a reliable elemental analysis at trace levels required for environmental samples of $PM_{2.5}$ (Kulkarni et al. 2007; Celo et al. 2010; Manousakas et al. 2014). ICP-MS offers the advantage of detection limits in the trace (µg L^{-1}) and ultra-trace (ng L^{-1}) levels, a wide dynamic range and multi-elemental analysis together with high sample throughput. Most ICP-MS systems are equipped for liquid sample introduction, thus requiring digestion of the PM. Several digestion techniques are available, including hot acid digestion and microwave-assisted acid digestion. Microwave digestion offers the advantages of a closed system with rapid (20 - 60 minute) digestion cycles at high pressure (16 to 20 atm) and hence is preferred over longer hot acid digestion at atmospheric pressure. The acid combination used depends on the sample matrix and elements of interest. Mixtures of concentrated mineral and oxidising acids have been used, including HNO_3, HNO_3/H_2O_2, HNO_3/HCl, HNO_3/HF and $HNO_3/HF/H_3BO_3$. The $HNO_3/HF/H_3BO_3$ mixture has been shown to provide near-total digestion of most elements, even in the presence of aluminosilicate matrices and refractory compounds (Karthikeyan et al. 2006; Celo et al. 2010; Manousakas et al. 2014).

The standard method for measuring Pb, Cd, As and Ni in the PM_{10} fraction of suspended PM is UNI EN 14902:2005, and uses a HNO_3 (70% w/v) and H_2O_2 (30% w/v) mixture in a 4:1 ratio. Including hydrogen peroxide increases the oxidative potential of the mixture and deoxidises NO_x back into nitrate. This reaction has the added benefit of reducing the pressure generated by the digestion of organic matter present in the PM sample, which could be significant in PM_{10} samples, thus avoiding overpressure in the digestion vessel. The downside is that H_2O_2 introduces potential contamination from Al, Mg, Zn, Fe, and Na present in commercially available 30–50% H_2O_2, which would increase the detection limit for these elements. Hence, using hydrogen peroxide is not desirable in the digestion of indoor $PM_{2.5}$ samples.

Certain acids in the digestion matrix could compromise the ICP-MS analysis due to the formation of species with m/z ratios that could mask elements of interest. Hydrochloric acid is avoided in ICP-MS analysis due to polyatomic chloride interferences (Karthikeyan et al. 2006). The use of hydrofluoric acid is therefore required for near-total digestion. HF achieves the decomposition of silicate matrices by converting them into H_2SiF_6. Even when used in excess, the loss of volatile fluorides, BF_3, SiF_4 and SeF_4 is typically unavoidable. High digestion acid concentrations introduce matrix-related interferences and increase the corrosion of ICP-MS parts, most importantly the skimmer cones. Therefore, samples analysed by ICP-MS typically require dilution to keep the concentration of digestion acids as low as possible. High quantities of HF, whilst providing near-total digestion and effective means to quantify elements such as K, Ti, Al, Sb and Cr, require specialised HF-resistant sample introduction systems. Minimising the amount of HF used in the digestion procedure is desirable as this avoids using specialised equipment.

Elements such as K, Al, Ti and Sb are important for identifying sources such as Saharan dust events and fireworks in the Central Mediterranean (Camilleri 2008; Scerri et al. 2018). High recovery efficiencies for these elements are therefore required. On the other hand, several SA techniques, such as PMF, require the analysis of hundreds of samples, and fast, reliable digestion techniques with minimal acid use are essential.

Studies (Lee et al. 2003; Celo et al. 2010) have shown that nitric acid alone was unsuitable for the effective analysis of some metals of interest, including K, Ti, Al and Sb. Hence, the investigation into a microwave-assisted acid digestion procedure that would give the best

elemental extraction from $PM_{2.5}$ deposited on PTFE filters whilst maintaining low detection limits is desirable. The optimal mixture of nitric, hydrofluoric and boric acids that provides near-total digestion at the lowest detection limits requires further investigation.

1.5 Modelling $PM_{2.5}$ Using Machine Learning Algorithms

1.5.1 Introduction

Machine Learning (ML) uses advanced computer algorithms that enable computers to learn from and make decisions based on data, and these data-driven models have been rapidly developing and applied in air pollution research (Y. Li et al. 2023). These data-driven algorithms constitute statistical techniques that rely on historical data to predict future values. ML modelling techniques have been recently adopted to capture the complex relationships with the data due to their ease of use and low computational resources (Yuchi et al. 2019; Du et al. 2021; Gu et al. 2021; Kumar et al. 2022).

There are two primary categories in machine learning: supervised and unsupervised learning. In supervised learning, the objective is to fit a model that relates the response to the predictor as accurately as possible to make future predictions about the response or else to understand the relationship between the response and the predictors (James 2013). Many classical machine learning algorithms, such as linear and logistic regression, fall under this category. Modern supervised algorithms include the Generalised Additive Model (GAM), Tree-Based models (Random Forest and Boosting), and Deep Learning algorithms such as Support Vector Machines (SVM).

On the other hand, unsupervised learning deals with algorithms without a response variable that can supervise the analysis. That is, there is no specific response variable to predict, and hence, there is no predefined outcome. Clustering (such as K-Means Clustering and Hierarchical Clustering) and PCA are examples of unsupervised learning. Modern neural networks such as Autoencoders are also unsupervised machine learning algorithms. A study evaluated two clustering algorithms, k-means clustering (kMC) and spectral clustering (SC), as potential receptor models for SA in comparison to PMF (Kumar et al. 2022). The application of kMC produced unsatisfactory results compared to PMF, but SC showed a statistically

significant correlation with the PMF results. The authors noted that these data-driven algorithms require large datasets for training the model, and accuracy improves with an increase in the size of the available data. Hence, they are not very useful for smaller datasets.

An increased frequency of ML modelling applications, particularly for $PM_{2.5}$ after 2017, has been recorded (Y. Li et al. 2023). Most research involving modern ML models to predict or understand $PM_{2.5}$ have mainly focused on supervised machine learning algorithms, specifically Tree-Based models (Brokamp et al. 2017; Chen et al. 2018; Y. Zhang et al. 2020; Wang et al. 2022; Zhang et al. 2022). Although Linear Regression (LR) has been widely used to calculate and predict the variability of air pollutants and has been used to predict the concentrations of various pollutants at several locations (Wanget al. 2013; Huang et al. 2018; Gu et al. 2021; Lesar and Filipčić 2021; Gulati et al. 2023), these models are limited to linear relationships between the predictor variables and the response variable, and yet, many atmospheric interactions are complex and nonlinear (James 2013; Wang et al. 2022). GAM is an extension of linear regression but introduces flexibility in the model to allow for the modelling of more complex relationships. On the other hand, in Tree-Based models, the model is trained on the training data via an ensemble of decision trees that can handle nonlinear relationships between the predictor variables and the response variable. One significant advantage of algorithms such as Random Forest (RF) or Extreme Gradient Boosting (XGBoost) is their capacity to comprehend intricate and nonlinear correlations between independent variables and the response variable, even with limited training data (Brokamp et al. 2017).

Moreover, Tree-Based models are less of a "black-box" method when compared to Artificial Neural Networks (ANNs), as the learning process can be more easily explained, investigated, and interpreted. They have been shown to offer superior performance compared to conventional statistical and air quality models, and they achieve this by minimizing variance/bias and errors in datasets with many predictor variables (Y. Zhang et al. 2020). Deep learning has been used for $PM_{2.5}$ prediction, including deep neural network, long and short-term memory (LSTM), and convolutional neural network (CNN) algorithms, although missing values, which can be quite high in air sampling campaigns (Joharestani et al. 2019), are a challenging problem with some deep learning methods such as LSTM and CNN.

1.5.2 Modelling Indoor $PM_{2.5}$

Research has demonstrated that residential $PM_{2.5}$ concentrations serve as an effective substitute for personal exposure, especially for vulnerable populations who predominantly remain indoors (Klepeis et al. 2001), and can assist in accurately assessing $PM_{2.5}$ exposure, thereby minimizing exposure measurement errors in epidemiological studies. Although direct measurement of indoor $PM_{2.5}$ concentrations yields the most precise data, such methods are labour-intensive and expensive, rendering them impractical for large-scale studies. Moreover, modelling indoor $PM_{2.5}$ concentrations using chemical models is complicated when applied on a large spatial scale. Chemical models use chemical laws to model the concentration of $PM_{2.5}$ from the chemical processes that generate particles. The multiple and nonlinear interactions are complex and challenging, requiring a detailed emissions inventory, which usually varies at different sampling locations (Pan 2018; Dai et al. 2023). Consequently, data-driven models constitute statistical techniques that rely on historical data to predict future values, and these approaches can enhance the limited availability of direct indoor $PM_{2.5}$ measurements.

In recent years, besides the modelling of outdoor $PM_{2.5}$ (Feng et al. 2015; Chen et al. 2018; Pan 2018; Ma et al. 2020), several studies have investigated the use of modern ML algorithms to predict indoor residential $PM_{2.5}$ concentration (Ryan et al. 2015; Yuchi et al. 2019; Xu et al. 2020; Men et al. 2023; Yu et al. 2024). While most studies focused on estimating hourly indoor $PM_{2.5,}$ the study by Yuchi et al. 2019 predicted weekly indoor $PM_{2.5}$ concentrations, with most of the studies conducted in Chinese cities. All these studies investigate residential indoor $PM_{2.5}$ except for Yu et al. 2024 in Australia, which considered both commercial and residential environments.

Most of these studies measured indoor $PM_{2.5}$ concentrations using low-cost sensors to obtain a large dataset. Typically, 20 – 40 residences were investigated, but two studies sampled more than 500 households, which would have been near impossible with particle spectrometers that are typically used at ambient monitoring stations as this equipment is bulky and expensive. Hence, the use of low-cost sensors was deemed necessary for these studies. Typically, low-cost sensors approximate $PM_{2.5}$ concentrations as these sensors tend to show lower accuracy at high PM concentrations, low temperatures (<10 ^{o}C), and high humidity (>65%). In the study by Xu et al. 2020, concurrent gravimetric $PM_{2.5}$ sampling was carried out, and a five-day weighted mass concentration was used to post-correct the hourly light scattering $PM_{2.5}$

concentrations. As for outdoor PM concentrations, some studies used official data from the closest ambient monitoring stations (Yu et al. 2024) or used the same low-cost sensors (Xu et al. 2020; Men et al. 2023).

All studies requested the participants to answer a standardized structured questionnaire, including questions about building age and size, floor level, ventilation type, number of inhabitants, combustion methods used for cooking and/or indoor heating, and the occurrence of indoor smoking. This information and meteorological parameters (outdoor and indoor temperature and relative humidity and air pressure) were used as potential predictor variables in the models to determine indoor $PM_{2.5}$ concentrations. RF was the most common ML algorithm used, and the $PM_{2.5}$ predictions obtained from RF were typically compared to those obtained when the same dataset was analysed using MLR.

Although MLR offers a simple methodology and ease of application, RF regression provides the advantage of not requiring any assumptions about data distribution. MLR, however, necessitates a linear relationship between the predictor and response variables. Furthermore, the response variable in MLR must follow a normal distribution and must also satisfy the homogeneity of variance (Yuchi et al. 2019; Shi et al. 2023). These stringent requirements for MLR are frequently unmet, as numerous chemical and meteorological processes contributing to the formation of indoor $PM_{2.5}$ exhibit non-linear relationships. Most studies reported that regression RF models performed much better and gave predicted indoor $PM_{2.5}$ concentration with higher R^2 values and lower root mean square error, RMSE compared to MLR. RF models also provided better predictions in the Personal Air study (RIOPA), where both LR and RF methods were used to predict elements in 48-h personal $PM_{2.5}$ samples using indoor/outdoor concentration and house characteristics (Ryan et al. 2015).

Even though RF regression models outperformed MLR models, they still struggled to accurately predict extreme concentrations (Wang et al. 2022). Nonetheless, RF regression models and other modern machine-learning algorithms warrant further investigation. These advanced modelling techniques hold potential for the quantitative estimation of indoor $PM_{2.5}$ exposure in large-scale studies.

Chapter 2

Methodology: Materials, Instrumentation & Analytical Methods

2.1 Introduction

This chapter will detail the methodology used for the two sampling campaigns carried out in this study. The first part covers the sampling campaign for the daily sampling of $PM_{2.5}$ from 23/06/2018 to 24/06/2019 at an urban background site in Malta. This will include the gravimetric analysis of $PM_{2.5,}$ the microwave digestion and ICP-MS elemental analysis of indoor and outdoor samples, and the validation of the microwave digestion procedure. The extraction and analysis of ionic species in indoor samples via ion chromatography and the OC-EC analysis of indoor samples via thermal-optical transmittance are also included. The data derived from this campaign was used for the SA of $PM_{2.5}$ via PMF. The second part covers the sampling campaign for the continuous indoor and outdoor PM measurements from 01/01/2021 to 15/05/2024 at several residences in Malta and Gozo. The data was used to predict indoor $PM_{2.5}$ using machine learning algorithms.

2.2 Daily $PM_{2.5}$ Sampling at an Urban Background Site in Malta

2.2.1 Equipment used for daily $PM_{2.5}$ sampling and analysis

Single filter low volume air samplers – (Leckel LVS-3): These samplers can maintain a constant flow rate of 2.3 m^3 h^{-1}. The flow rate is controlled by means of a temperature-compensated orifice plate according to Bernoulli's law, and conversion into the operating volume flow (m^3 h^{-1}) is performed according to the gas law. These samplers were calibrated according to standard EN12341.

Sequential low volume air samplers – (Leckel SEQ 47/50): The samplers operate in the same manner of the LVS-3 model but with automatic changing of filters. The filter cassette holds a maximum of 16 filter cassettes. Once the sampling period finishes, the sampled filter is immediately transferred to a refrigerated stack inside the sampler and kept at a temperature of < 10 ºC.

$PM_{2.5}$ inlet – (Leckel). This inlet conforms to CEN standard EN12341 for $PM_{2.5}$ monitoring.

Volumetric flow laboratory calibrator – (BGI TetraCal). Volumetric flow rate calibrator used to calibrate the LVS.

Weather station – (Davis Vantage Vue Indoor and Outdoor weather station).

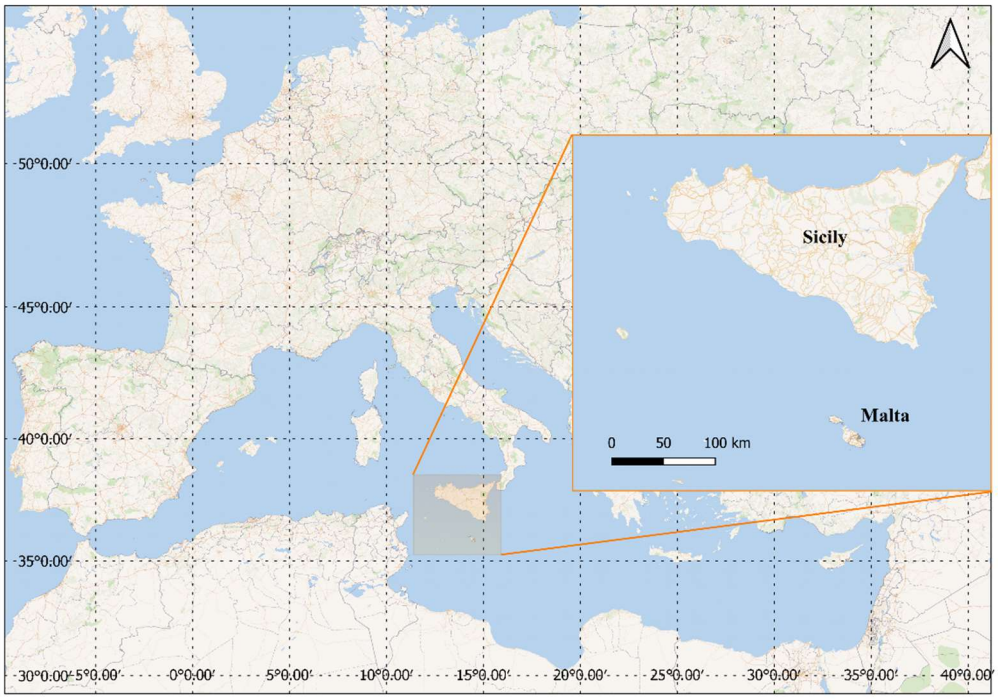
Microbalance – (Mettler Toledo XP2U with a readability of 0.1μg).

Standard weights – (Mettler Toledo 100 mg F1 wire weight).
RH and humidity monitor – (Extech, SD800).
Air handling system – (Airedale, EasiCool).
Microwave digester – (Milestone, EthosUP with SK-15 rotor and 7ml PFA microvessels).
ICP-MS – (Perkin Elmer)
Lab OC-EC Aerosol Analyser – (Sunset)
Petri slides – (Millipore)

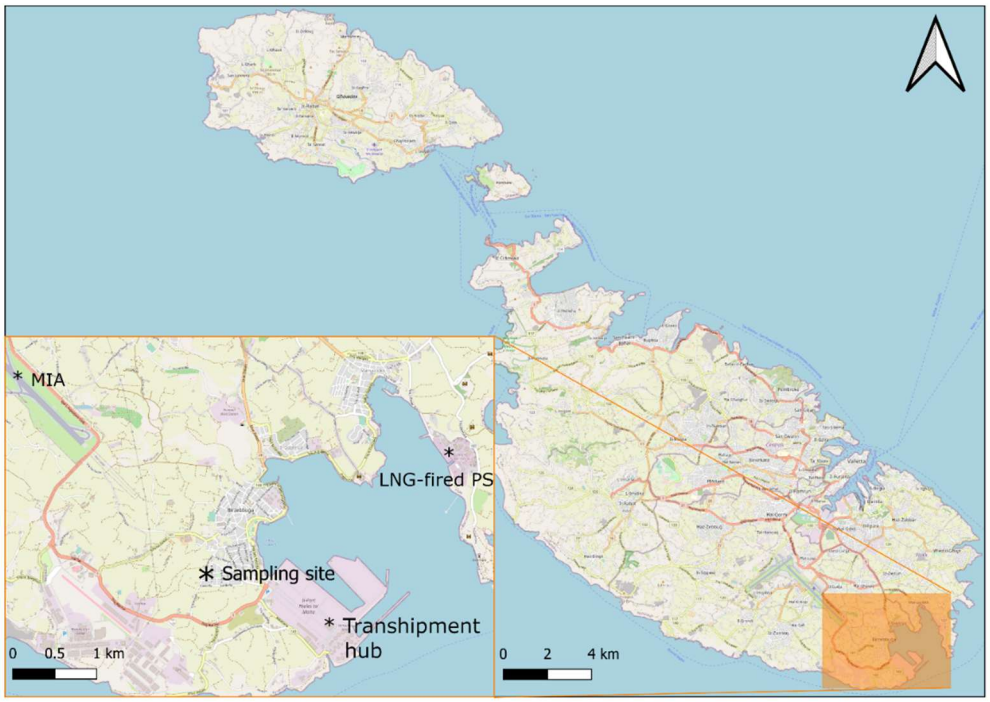
2.2.2 Site selection

The Maltese Islands are an archipelago located in the central Mediterranean Sea. The archipelago is geographically small (316 km^2) and densely populated (*ca.* 500 000), with the two largest inhabited islands being Malta and Gozo.

The sampling was carried out in a three-storey building occupied by three residents in Birżebbuġa, a village of approximately 10,000 residents, located in the southeastern part of Malta (Figure 1) at an urban background site. The building is located on the outskirts of the village, in a residential area comprised of two to three-storey buildings. No retail or industrial activity is carried out in the vicinity. The roads in the area are two-lane residential streets with parking allowed on either side of the street. The area around Birżebbuġa was deemed to be an urban background site since its pollution level is not significantly determined by any single source. Such a description of the built conglomeration was used in an assessment report of air quality in Malta (Stacey and Bush 2002). Urban background sites represent the integrated contribution from many sources, including all upwind sources. The Malta International Airport (at Gudja) lies at approximately 2 km northwest of the site. The liquified natural gas (LNG)-fired power station (at Delimara) and a transhipment hub (at Birżebbuġa port) are in an easterly and south-easterly direction at 1.5 and 2 km, respectively.



(a)



(b)

Figure 2.1 (a) Map indicating the geographic location of Malta and (b) showing the Birżebbuġa sampling site in relation to the Malta International Airport (MIA), the Transhipment hub and the LNG-fired power station.

The sampling site is situated downwind of most anthropogenic and natural pollution sources in Malta, considering the prevalent north-westerly winds and should, therefore, be an ideal representative of the pollution exposure for the general urban population.

2.2.3 $PM_{2.5}$ sampling procedure

Indoor and outdoor $PM_{2.5}$ sampling were carried out concurrently. Indoor $PM_{2.5}$ was collected using two single-filter low-volume air samplers (Leckel LVS-3) in the residence's living room. The living room is connected to the kitchen and dining room via a small window of 0.3 m^2. One sampler was loaded with PTFE filters (Whatman, $PM_{2.5}$ air monitoring membrane), and the other was loaded with quartz fibre filters (Whatman, QM-H, 47 mm).

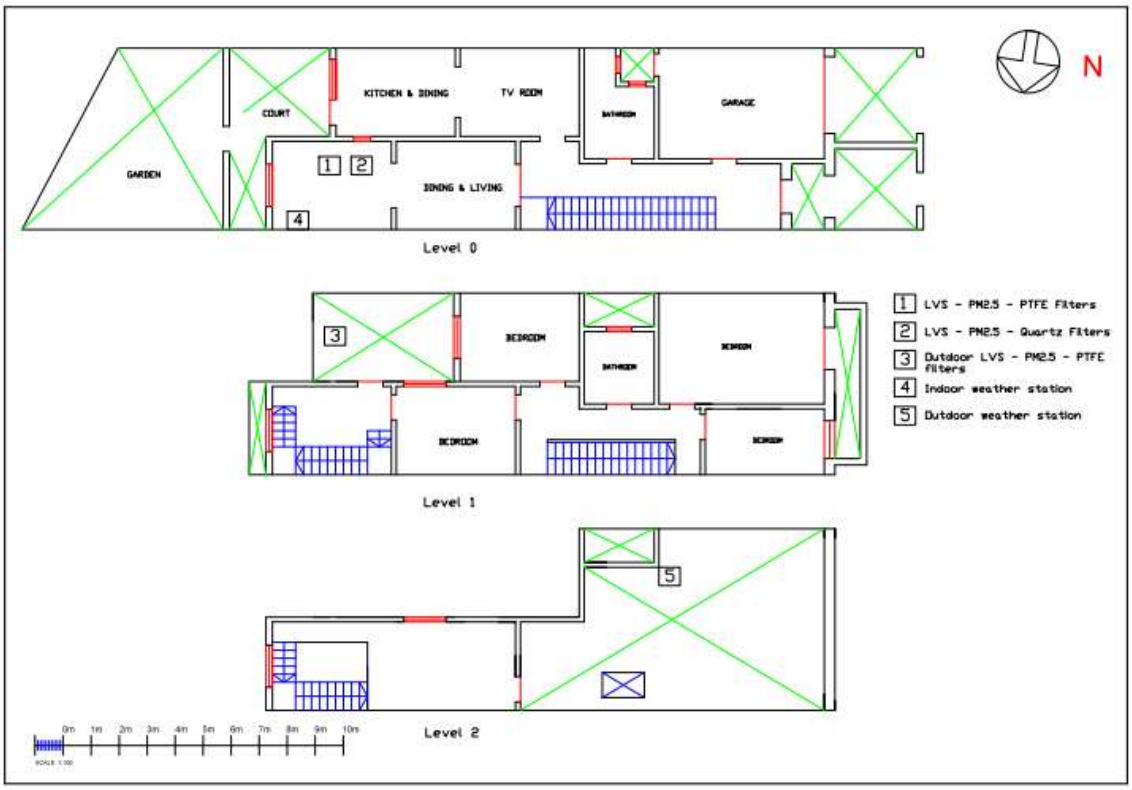


Figure 2.2 Schematic diagram of the Birżebbuġa sampling site detailing the position of the samplers and weather stations.

Outdoor $PM_{2.5}$ was collected using a sequential low-volume air sampler (Leckel SEQ 47/50) placed on the terrace on the first floor of the residence and located at least 1.5 m from the adjacent walls, as shown in Figure 2.2. Both indoor and outdoor samples were stored at a temperature of < 10 ^{o}C prior to final weighing.

The samplers were located within a 3 m vertical distance from each other within the residence. The vertical variation of $PM_{2.5}$ concentrations was studied over a 65 m elevation (Zauli Sajani et al. 2018). The study recorded a slight gradient with a 4% decrease in $PM_{2.5}$ concentrations in winter at the highest sampling point. This decrease went up to 11% in summer, and the difference was statistically significant. $PM_{2.5}$ values within the first 15 m of elevations were less than 4% different and statistically insignificant. Based on this information, any differences in the data collected by the two samplers at the Birżebbuġa sampling site should not be related to the difference in elevation between them.

The sampling campaign was conducted between June 2018 and June 2019, and 159 days were sampled. The 24-hour sampling started at midnight three times a week. Sampling was conducted on Mondays, Thursdays and Saturdays (except for the Santa Maria festival in mid-August 2018, where sampling was done daily from 13th to 16th August).

2.2.4 Meteorological data

Indoor and outdoor weather data, including non-sampling days, were collected during the whole period. The data included wind speed, wind direction, and rainfall. Less than 10% of the sampling days experienced more than 5 mm of rainfall; hence, there were not enough data points to assess the impact of rainfall on indoor $PM_{2.5}$. During sampling days, residents kept a daily log of indoor activities such as cooking, heating/cooling, cleaning, smoking, and the opening or closing of outdoor windows. A summary of these activities is given in Table A.1.

2.2.5 Gravimetric analysis

PTFE filters were conditioned and weighed following the weighing parameters detailed in EN 12341:2014 (CEN 2014). The weighing room temperature was set at 19 ºC to 21 ºC, and the relative humidity was maintained at 45% to 50%. The weighing procedure is detailed in Table 2.1.

Table 2.1 Weighing procedure for PTFE filters

Weighing		Result	Conditioning period	Acceptance criteria
Prior to sampling	First weighing	$m_{u,1}$	≥ 48 h	
	Second weighing	$m_{u,2}$	≥ 12 h	$\lvert m_{u,1} - m_{u,2} \rvert \leq 40$ µg
			All samples fulfilled this condition. The unsampled mass filter mass, m_s, was taken as the average of $m_{u,1}$ and $m_{u,2}$.	
After sampling	First weighing of sampled filter	$m_{s,1}$	≥ 48 h	
	Second weighing of sampled filter	$m_{s,2}$	≥ 24 h	$\lvert m_{s,1} - m_{s,2} \rvert \leq 60$ µg
			All samples fulfilled this condition. The sampled mass filter mass, m_c, was taken as the average of $m_{c,1}$ and $m_{c,2}$.	

Weighing room blanks and field blanks were kept and analysed as per EN 12341:2014. All field blanks registered a mass difference of ≤ 60 µg before and after sampling.

The $PM_{2.5}$ concentrations were calculated according to the formula:

$$C_{PM_{2.5}} = \frac{m_s - m_u}{\varphi_a . t} \qquad \textbf{Eq. 2.1}$$

c = $PM_{2.5}$ concentration in µg m^{-3}, where the volume is taken at ambient conditions near the inlet during sampling.

m_s = sampled filter mass in µg

m_u = unsampled filter mass in µg

φ_a = the flow rate at ambient conditions in $m^3\ h^{-1}$

t = sampling time in h

2.2.6 Filter analyses

Microwave-assisted acid digestion was carried out on the PTFE filters, and the digest was analysed by ICP-MS to calculate the concentrations of Al, As, Ba, Ca, Cu, Cr, Fe, K, Mg, Mn, Na, Ni, Pb, Sb, Sr, Ti, V and Zn. ICP-MS analysis was conducted at the School of Geography, Earth and Environmental Studies, University of Birmingham, UK.

Quartz fibre filters were split in half, and one half was sent to the Wisconsin State Lab of Hygiene, Madison, USA, for OC-EC analysis using thermal-optical transmittance while the remaining half of the sampled QF filters were used for ion chromatography (IC) analysis of $PM_{2.5,}$ which was carried out at the Water Quality Analysis Laboratory, University of New Hampshire, Durham, USA.

2.2.7 Microwave digestion and ICP-MS analysis

2.2.7.1 Reagents and materials

Deionised water with a resistivity higher than 18.2 MΩ·cm (Elga Purelab system) was used for this work. Concentrated hydrofluoric acid (48% w/v, Honeywell, TraceSELECT), concentrated nitric acid 68% w/v (Honeywell, TraceSELECT) and boric acid (99.9999%, Merck, Suprapur) were used directly without any purification. NIST standard reference material (SRM) 1648a Urban PM (SRM 1648A 2015) was used in this study, and it was handled, stored and used as per instruction in the certificate of analysis. The ICP-MS calibration standard solutions were prepared via dilution of 1 mg L^{-1} standards for ICP-MS or 1000 mg L^{-1} standards for ICP (Sigma Aldrich, TraceCERT).

2.2.7.2 Validation of the Digestion Procedure

Microwave-assisted acid digestion of the PTFE filters was carried out using a closed-vessel microwave digestor (Ethos UP^{TM}, Milestone). The oven is equipped with a 2.45 GHz microwave supply of 0 – 1.8 kW output and a 15-position rotor for 100 mL PTFE vessels. Acid digestion was carried out using 7 mL PTFE ultra-trace inserts. Each 100 mL PTFE vessel was equipped with a PTFE ultra-trace insert holder capable of carrying three PTFE ultra-trace inserts, increasing the sample throughput threefold. Digestion can be carried out with 2 – 3 mL of acid in the ultra-trace insert compared to a minimum of 10 mL for the 100 mL vessels, which allows for a fast, reliable digestion method with minimal losses and low detection limits.

The internal pressure and temperature were monitored in a control vessel equipped with pressure and temperature sensors. The external temperature of each vessel was monitored via an IR sensor to check for any runaway reactions in any of the vessels.

This validation aimed to identify the acid digestion procedure that would give the best elemental extraction from $PM_{2.5}$ PTFE filters whilst providing the lowest limit of detection (LOD). This study used the Ethos UP^{TM} multiple ultra-trace insert system to compare three digestion procedures. This system allows rapid digestion of multiple samples within the standard microwave digestion rotor. The efficiency of tested procedures was evaluated by digesting SRM 1648a. The selected method was then applied to 318 $PM_{2.5}$ samples. Multiple digestions using the SRM 1648a and the selected method were carried out to confirm the selected method's robustness.

Methods similar to those proposed by (Lee et al. 2003; Kulkarni et al. 2007; Celo et al. 2010) were optimised for our laboratory equipment and conditions to identify the optimal acid digestion procedure. The use of highly corrosive hydrofluoric acid was minimised to maintain the final concentration of HF below 0.1% v/v: this enables the use of standard quartz ICP torches for the analyses of the digests without significant damage to the equipment. The digestion methods used are described in Table 2.2. To establish the most suitable method for $PM_{2.5}$ samples, SRM 1648a dust was digested with different acid mixtures, and the percentage recoveries for each element were calculated.

Approximately 5 mg of SRM 1648a were weighed on a microbalance (Mettler Toledo, Model XP2U Ultra Micro Balance) using a weighing paper. The powder was then transferred to a 7 mL PTFE ultra-trace insert. Although this sample size is much smaller than 100 mg as recommended by NIST, it has been shown to provide good analytical reproducibility (Celo et al. 2010). The ultra-trace inserts were then transferred to a fume cupboard, and 2.5 mL of 69% v/v HNO_3 were added using a pipette. Three different digestion methods were tested. The volume of acid added to each ultra-trace insert and the digestion cycles are detailed in Table 2.2. Before digestion, for methods 2 and 3, the corresponding volume of 50% v/v hydrofluoric acid was added to each ultra-trace insert using a pipette. The ultra-trace inserts were capped, and a sample holder loaded with three ultra-trace inserts was lowered in the 100 mL PTFE vessel, which was pre-filled with 15 mL of deionised water. Immersion in this “water blanket” ensures a stable and gradual increase in temperature inside the ultra-trace insert. The small

volumes of acid used in the ultra-trace insert are prone to develop hotspots. The water blanket around the ultra-trace insert avoids this. The 100 mL PTFE vessels were covered, placed in the protection shields and tightened using a torque wrench. Each vessel was ramped to 200 ºC in twenty minutes and kept at 200 ºC and 16 bar for another twenty minutes.

Table 2.2 Proposed Microwave-Assisted Digestion Methods.

Method	Digestion Solution	Step	Power (W)	Ramp time (min)	Temp (ºC)	Max Pressure (bar)	Hold Time (min)
1	2.5 mL HNO_3	1	1200	20	200	16	20
2[*]	2.5 mL HNO_3 + 3 µL HF	1	1200	15	200	16	20
		2	Cooling to < 50 ºC (45 min) followed by the opening of the vessels				
	24 µL 5% m/v H_3BO_3	3	1200	20	200	16	20
3[†]	2.5 mL HNO_3 + 30 µL HF	1	1200	15	200	16	20
		2	Cooling to < 50 ºC (45 min) followed by the opening of the vessels				
	240 µL 5% m/v H_3BO_3	3	1200	15	200	16	20

Notes: [*] This method is similar to the procedure used by (Celo et al. 2010).
[†] In this method, higher quantities of HF and H_3BO_3 are used to test if higher extraction efficiencies are obtained with the 10-fold increase in HF quantity adopted.

All vessels were then cooled for a minimum of 45 minutes to a temperature lower than 50 ºC and opened. For methods 2 and 3, the corresponding volume of 5% w/v H_3BO_3 solution was added to each ultra-trace insert using a pipette. These vessels were re-sealed and ramped again to 200 ºC in twenty minutes and kept at 200 ºC and 16 bar for another twenty minutes. H_3BO_3 is required to remove excess HF by forming the complex tetrafluoroboric acid, HBF_4, thereby aiding with the resolubilisation of insoluble fluorides of Mg, Al and Ca. A sample of the microwave digestion profile is shown in Figure A.1.

For each digestion cycle, a reference vessel was used to monitor the temperature and pressure of the digestion process. A reagent blank for each method was added to the reference vessel, which was filled with the same volume of reagents used in the sample vessels.

The contents of each ultra-trace insert were transferred into a 50 mL volumetric flask. The ultra-trace insert was then washed five times with deionised water. Each wash was transferred to the volumetric flask, and the volume made up to the mark. The contents of the flask were homogenised, and the solution was stored in propylene bottles at 4 ºC prior to ICP-MS analysis. Each SRM 1648a digestion solution was diluted 5-fold before ICP-MS analysis to bring the elemental concentrations approximately in range to what is expected for the indoor $PM_{2.5}$ samples.

Cleaning Procedure: After each digestion, the microvessels and the 100 ml Teflon vessels were rinsed three times with DI water and transferred to a 10% HNO_3 acid bath for 48 hrs. All vessels were then rinsed three times with DI water and allowed to air dry.

2.2.7.3 ICP-MS analysis

The ICP-MS analysis was carried out at the School of Geography, Earth and Environmental Sciences, University of Birmingham, UK, using a Perkin Elmer Nexion 300x ICP-MS equipped with a glass sea spray nebuliser and glass cyclonic spray chamber. The elements were analysed using the helium kinetic energy discrimination (KED) mode to reduce polyatomic interferences derived from the plasma via a collision cell.

Table 2.3 ICP-MS setup

Parameter	Values
Sample flow rate into the nebuliser	0.3 ml min^{-1}
Nebuliser gas flow rate	0.94 ml min^{-1}
Plasma auxiliary gas flow rate	1.2 ml min^{-1}
Plasma gas flow rate	18 ml min^{-1}
ICP RF power	16000 W
Helium Collision cell flow rate	4 ml min^{-1}

Internal standardisation

Internal standardisation was carried out using the internal standard stock solution of Li-6 and Rh-103, which was added automatically by the system to account for instrument drift and physical interferences. The same amount of internal standard solution added to the samples was added to the reagent blank.

Calibration

Reagent blank: The same amount of internal standard solution that was added to the samples was added to the reagent blank. This solution was used to determine if the elements to be analysed or if any other interferences were present as contaminants in the laboratory environment, reagents, or apparatus.

Calibration blanks: The method blank was used as the blank for the calibration standards. Calibration was carried out for the following elements Na-23, Mg-24, Al-27, K-39, Ca-43, V-51, Cr-52, Mn-55, Fe-57, Ni-60, Cu-63, Zn-66, As-75, Sr-88, Cd-111, Sb-121, Ba-138 and Pb-208. Table 2.4 lists the concentrations used for the calibration curves.

Table 2.4 Concentrations of the calibration standard (in ppb) used for the elements analysed by ICP-MS

Cal STD	Na-23	Mg-24	Al-27	Si-28	K-39	Ca-43	Ti-48
1	6.25	12.5	125	125	12.5	125	6.25
2	12.5	25	250	250	25	250	12.5
3	25	50	500	500	50	500	25
4	50	100	1000	1000	100	1000	50
5	100	200	2000	2000	200	2000	100

Cal STD	V-51	Cr-52	Mn-55	Fe-57	Ni-60	Cu-63	Zn-66
1	0.625	0.625	1.25	125	0.625	1.25	6.25
2	1.25	1.25	2.5	250	1.25	2.5	12.5
3	2.5	2.5	5.0	500	2.5	5.0	25
4	5.0	5.0	10.0	1000	5.0	10	50
5	10	10	20.0	2000	10	20	100

Cal STD	As-75	Sr-88	Cd-111	Sb-121	Ba-138	Pb-208
1	0.625	0.625	0.625	0.625	0.625	12.5
2	1.25	1.25	1.25	1.25	1.25	25
3	2.5	2.5	2.5	2.5	2.5	50
4	5.0	5.0	5.0	5.0	5.0	100
5	10	10	10	10	10	200

Method detection limit (MDL)

The limits of detection for each analyte for the three digestion procedures were calculated from reagent blanks that were digested using identical amounts of reagents used in each digestion

procedure. The LOD was calculated from the reagent blanks of each digestion procedure set at three times the standard deviation of a reagent blank and was calculated from the calibration plot using equation 2.2:

$$MDL = \frac{I_o.3\sigma}{m_c} \qquad \textbf{Eq. 2.2}$$

where I_o is the net intensity of the reagent blank (cps), m_c is the gradient of the calibration curve, and σ is the standard deviation based on 11 blank measurements with the confidence factor set at 3. The method detection limits are given in Table A.2.

2.2.7.4 Quality control and instrument performance

Maintaining low and stable blank concentrations and detection limits is essential in analysing trace and ultra-trace metals. The concentration of the elements in the blanks is affected by the quality of reagents used, laboratory procedures used for digestion and dilution and most importantly, cleaning procedures. To obtain consistent low MDLs and blank concentrations, the ultra-trace inserts and the 100 mL PTFE vessels were cleaned according to the described procedure.

For the analyses of the $PM_{2.5}$ samples, a calibration standard was run with every 20 samples to ensure that the instrument remained within the calibration parameters. If any of the elements' concentrations varied by more than ±15% from the actual concentration of the calibration standard, the standard was re-tested. The instrument was re-calibrated if the concentration was again outside the ±15% limit. Reagent blanks, method blanks and field blanks were analysed for each digestion run as part of the quality control procedure. The blank concentrations were monitored to ensure the concentrations were below the detection limits. The results for the blanks are given in Table A.3.

2.2.7.5 Method Validation Results

The standard digestion method for analysis of Pb, Cd, As and Ni in PM_{10} (EN 14902:2005) uses a digestion mixture of 80% HNO_3 (68% v/v) and 20% H_2O_2 (30% v/v). However, if a total dissolution of several trace elements, including Na, Al, K, Ti, Cr and Sb, is required, digestion in the presence of HF is required (Celo et al. 2010). This was confirmed in this study as the

recovery efficiencies for Na, Al, K, Ti, Cr, and Sb increased significantly with a digestion mixture of HNO_3/HF/H_3BO_3 as shown in Table 2.5. The correct contribution of these elements from crustal, fireworks and vehicular sources cannot be determined accurately without near-total digestion.

The presence of small amounts of hydrofluoric acid can have a significant impact on the extraction of these elements. The mean percentage recoveries, given in Table 2.3, show that adding HF to the digestion procedure in method 2 increased the recovery efficiency for Al, Sb, Cr, Ni, K and Na by approximately 30% over method 1. With a tenfold increase in the quantity of HF used in method 3, all elements recorded a recovery efficiency greater than 70%. This confirms that the use of HF is required for the complete solubilisation of samples containing inorganic silicates, while H_3BO_3 is required after the HNO_3/HF digestion to complex unreacted HF and re-solubilise insoluble fluorides such as CaF_2 and MgF_2. Table 2.3 also shows that the complexation property of HF acid explains the increased recoveries of Sb compared to HNO_3 digestions (Celo et al. 2010).

Compared to similar studies, slightly lower percentage recoveries were reported for Sb, Ni, Cr and Ti for both methods 2 and 3 compared to similar digestion methods as reported by (Kulkarni et al. 2007) and (Celo et al. 2010), as shown in Table 2.3. Therefore, method 2 or method 3 is needed for near-total digestion even though digestions with HF increased the detection limit for most elements compared to HNO_3 digestions.

Si cannot be determined in HF digestions since gaseous SiF_4 formed in the reaction is lost during digestion. Adding HF also increases the reagent blank concentrations for Si, as acidic attack on the borosilicate volumetric glassware becomes significant. The reagent blank values for Si for method 3 were more than 200 µg/L. Therefore, further Si analysis was not performed.

The LODs given in Table 2.6, for digestion methods 2 and 3 were slightly higher than the values reported for method 1, which included only HNO_3 acid in the digestion procedure. Adding HF and H_3BO_3 in methods 2 and 3 increased the digestion cycle duration and detection limits as additional reagents were used. Nonetheless, the LODs for both methods 2 and 3 are sufficiently low for elemental analysis of $PM_{2.5}$ from outdoor and indoor environments. Table 2.6 shows that the LODs for methods 2 and 3 are very similar, and for some elements, they are

lower than the LODs reported for a digestion procedure, which is very similar to method 2 by Celo et al. 2010.

Based on the elemental recoveries and LODs, method 3 was chosen as the preferred digestion method for near-total digestion of $PM_{2.5}$. It offers significantly better percentage recoveries for most elements than digestion with HNO_3, especially for Sb, Cr, Ni, K and Na. It also offers better percentage recoveries than method 2 for Ti and Sb. Even though method 3 shows a marginal increase of the LOD for some elements (Mg, Al, Ti, Fe, Ni, Zn, Sr and Ba) when compared to method 2, the LODs of method 3 are still reasonably low, and in fact lower for most elements when compared to the extraction method employed by (Kulkarni et al. 2007) and (Celo et al. 2010). Given that the quantity of aluminosilicates in $PM_{2.5}$ varies significantly and is especially important during Saharan dust events that occur in the central Mediterranean region, method 3 provides for better near-total digestion when compared with method 2.

Table 2.7 compares the LODs for digestion method 3, given as ng m^{-3}, to $1/10^{th}$ the mean elemental concentration for indoor samples collected in Barcelona (Minguillón et al. 2012). The Barcelona dataset was used as no local indoor elemental dataset is available. Barcelona was chosen due to its regional and climatic similarities to Malta, and it should provide a suitable comparison. Even though the comparison is being carried out with $1/10^{th}$ of the indoor mean elemental concentration, all elemental concentrations, except for Sb, were higher than the LOD of method 3, which suggests that most of the elemental concentrations collected at Birżebbuġa are expected to exceed the detection limit for digestion method 3.

Table 2.5 Percentage elemental recoveries of SRM 1648a by ICP-MS after microwave-assisted acid digestion and their respective standard deviations shown in brackets for each digestion method carried out in this study and similar other studies

Element	Method 1		Method 2		Method 3		Celo et al. 2010	Celo et al. 2010	Kulkarni et al. 2007
	2.5 mL HNO_3		2.5 mL HNO_3, 3µl HF, 24µl H_3BO_3		2.5 mL HNO_3, 30µl HF, 240µl H_3BO_3		Conc HNO_3	1.5 mL HNO_3, 3µl HF, 24µl H_3BO_3	3 mL HNO_3, 300µl HF, 2.4mL H_3BO_3
Na	56.2	(1.7)	116	(12.3)	110	(5.4)			88 (7)
Mg	101	(2.8)	90.5	(0.7)	89.0	(2.1)			101 (4)
Al	60.8	(0.1)	88.8	(3.7)	91.4	(1.6)	26 (8)	86 (10)	104 (8)
Si	13.9	(24.5)	27.1	(66.6)	2.50	(37.8)			
K	39.4	(0.5)	72.5	(0.6)	74.3	(0.8)			95 (5)
Ca	108	(0.5)	106	(4.7)	106	(0.5)			
Ti	35.7	(3.1)	53.2	(8.7)	78.4	(1.1)	17 (15)	73 (10)	107 (4)
V	87.3	(0.4)	93.9	(0.6)	96.2	(0.6)	70 (2)		101 (6)
Cr	44.1	(11.4)	70.7	(8.9)	73.0	(3.9)	23 (1)	56 (14)	81 (8)
Mn	92.8	(0.5)	98.6	(1)	101	(1.3)	84 (2)	99 (7)	96 (5)
Fe	85.1	(1.2)	93.3	(0.4)	92.4	(0.7)	72 (1)	92 (10)	98 (4)
Ni	66.5	(7.7)	75.9	(10.9)	77.2	(18.8)	76 (2)	93 (10)	99 (4)
Cu	89.8	(1)	88.1	(2.5)	91.0	(2.7)	83 (1)	91 (5)	91 (7)
Zn	90.7	(4.5)	85.5	(8.4)	77.0	(2.4)	94 (2)	94 (4)	101 (9)
As	86.0	(3.4)	83.8	(0.9)	82.8	(0.3)	91 (1)	102 (4)	109 (7)
Sr	122	(5.9)	101	(4.7)	99.0	(4.1)			
Cd	82.7	(0.5)	84.3	(0.1)	83.2	(0.7)	91 (1)	99 (5)	112 (4)
Sb	42.6	(8)	70.7	(11)	79.0	(0.7)	16 (11)	101 (9)	86 (3)
Pb	90.0	(0.9)	87.6	(1.7)	83.8	(0.1)	94 (1)	98 (4)	108 (5)

Table 2.6 Limits of Detection (LODs) calculated from the reagent blank for each method

	LOD for Method			LOD (HNO_3HF, H_3BO_3) (Celo et al. 2010)*	LOD (HNO_3HF, H_3BO_3) (Kulkarni et al. 2007)
	1	2	3		
Element	**(µg L^{-1})**			**(µg L^{-1})**	**(µg L^{-1})**
Na	0.004	0.118	0.096		0.921
Mg	0.070	0.408	2.539		0.398
Al	1.144	1.440	7.420	7.6	4.025
Si	2.556	28.86	51.93		
K	1.924	2.606	1.689		
Ca	17.77	40.06	23.47		
Ti	0.622	0.559	1.322	1.1	4.056
V	0.008	0.027	0.014	0.11	0.051
Cr	0.033	0.032	0.016	1.1	0.363
Mn	0.015	0.028	0.027	0.22	0.021
Fe	0.857	2.121	2.644	11	3.246
Ni	0.047	0.022	0.132	0.86	0.124
Cu	0.025	0.020	0.021	2.2	0.062
Zn	0.313	0.356	0.495	9.7	1.026
As	0.028	0.113	0.074	0.11	0.019
Sr	0.027	0.018	0.113	0.22	
Cd	0.011	0.022	0.010	0.0	0.035
Sb	0.014	2.201	0.858	0.43	0.026
Ba	0.010	0.012	0.018	0.22	0.436
Pb	0.027	0.236	0.196	0.65	0.086

Notes: ***** The LODs in this study were reported for filter blanks and hence would be expected to be marginally higher than LODs from reagent blanks.

Table 2.7 LODs for digestion method 3 in ng m^{-3} and $1/10^{th}$ the mean concentration for indoor samples in Barcelona.

Element	LODs for digestion method 3 (ng m^{-3})	$1/10^{th}$ the mean concentration (ng m^{-3}) for indoor samples in Barcelona (Minguillón et al. 2012)
Na	0.09	30
Mg	2.31	4.0
Al	6.75	11
K	1.54	20
Ca	21.3	40
Ti	1.20	3.2
V	0.01	1.2
Cr	0.01	0.29
Mn	0.02	1.4
Fe	2.40	20
Ni	0.12	0.74
Cu	0.02	1.6
Zn	0.45	5.5
As	0.07	0.11
Sr	0.10	0.17
Cd	0.01	0.04
Sb	0.78	0.10
Ba	0.02	0.41
Pb	0.18	0.83

2.2.7.6 Filter Digestion and Analysis

Method 3 was chosen for the digestion of samples and field blanks, which included 350 filters. The polypropylene ring of each PTFE filter was removed by cutting around the perimeter of the ring using ceramic scissors. After each use, the scissors were wiped with a clean, lint-free paper. The filter paper was then transferred into a numbered 7 ml PFA microvessel using plastic forceps and digested as per method 3.

Each digestion cycle comprised a reagent blank, method blank, field blanks and samples. The field blanks were filters that had undergone all the sampling procedures except for the air sampling by the LVS. These were digested using the same amount of reagents used for the samples. One field blank was carried out for every ten sampled filters.

The contents of each microvessel were transferred into a 50 ml volumetric flask. The microvessels were then washed five times with deionised water. Each wash was transferred to the volumetric flask, and the volume was then made up to the 50 ml mark. The contents of

the flask were homogenised, and the solution was then stored in propylene bottles at 4 °C before ICP-MS analysis.

For every three digestion cycles, three SRM 1648A digests were carried out for quality control. Table 2.8 gives the mean percentage recoveries of the twelve-quality control SRM digests carried out with method 3 during the sampling campaign. Except for Cr, all elements had acceptable recovery efficiencies of between 8 % − 120%. This data also shows marginally better recoveries for Ni, Ti and Sb for method 3 compared to data from Table 2.6. The results in Table 2.8 were obtained from a larger sample, which could explain the slightly higher recoveries.

Table 2.8 Percentage Recovery efficiencies for quality control SRM digests

N = 12	**SRM digests carried out during Filter analysis**				
Element	**Mean**	**(SD)**	**Element**	**Mean**	**(SD)**
Na	101	(12.4)	Fe	101	(5.6)
Mg	97.0	(11.2)	*Ni	92.2	(20.9)
Al	102	(4.2)	Cu	95.0	(9.9)
K	92.6	(11.2)	Zn	97.9	(8.1)
Ca	98.5	(9.4)	As	107	(7.3)
Ti	85.3	(5.6)	Sr	91.1	(8.3)
V	103	(7.2)	Cd	100	(5.1)
***Cr**	63.2	(4.6)	Sb	98.9	(5.3)
Mn	99.4	(6.9)	Pb	99.8	(4.0)

Notes: *Mean value obtained from 10 samples due to outlier values in 2 samples

2.2.8 OC-EC analysis

2.2.8.1 Pre-firing of blanks

Quartz fibre filters (Whatman, QM-H, 47 mm) were used for OC-EC analysis of $PM_{2.5}$ dust. These filters have low organic carbon (OC) concentrations (1.22 ± 1.11 μg cm^{-2}) even without pre-firing (Karanasiou et al. 2015). The quartz fibre filters were pre-fired to ensure minimal blank OC concentrations.

Aluminium containers used for the pre-firing process were themselves pre-fired at 500 °C for 3 hours (Karanasiou et al. 2015). The QM-H filters were placed in a pre-fired aluminium container and heated at 500 °C for 4 hours. The oven was cooled to room

temperature, and the filters were individually stored in pre-cooled Petri dishes (Millipore). Filters were handled with stainless steel tweezers. Each petri dish was numbered and sealed in PE zip-lock bags. The filters were stored at –18 ºC prior to being exposed.

2.2.8.2 Handling of QM-H filters

Prior to sampling, the filters were removed from the freezer and loaded into the sampling head at least 2 hours before the sampling start time. After sampling, the filters were placed back into the respective petri dishes, sealed inside the PE zip-lock bag, placed inside the freezer, and stored at –18 ºC.

A field blank was taken every ten samples. A pre-fired QM-H filter was loaded in the sampling head for 2 hours without any airflow passing through the filter. These filters were otherwise treated and analysed in the same way as the loaded filters. Transport of loaded filters and field blanks was carried out in insulated containers. The temperature was always kept below 0 ºC with ice packs.

2.2.8.3 OC-EC analysis

Prior to analysis, the filters were removed from the freezer and split into two using an aluminium template and stainless steel rotary cutter. The stainless steel cutter was cleaned with a new filter between individual cuts. The half filter was placed into new pre-cooled Petri dishes, sealed inside the PE zip-lock bag and shipped to the Wisconsin State Lab of Hygiene (WSLH), Madison, USA, for analysis. The samples were kept cold during shipment using ice packs.

A 1 cm^3 punch from all sampled filters and field blanks was taken and analysed using an OC-EC aerosol analyser. For every 10^{th} filter, a duplicate punch was analysed for quality control purposes, and the percentage change in total carbon (TC) between the two punches was recorded.

The NIOSH 870 OC-EC protocol used is a modification of NIOSH 5040, and is detailed in Table 2.9. Since elemental carbon is a strong absorber of light, the OC and EC split was determined by a laser beam that tracks light absorbance on the filter during the analysis. This makes it possible to separate elemental carbon that was originally present on the filter from

elemental carbon formed as a result of organic carbon charring (pyrolysis) during the analysis. During the first ramp, the temperature is increased to 870 ºC and purged with helium (OC4). The oven is then cooled to 550 ºC (EC1). At this stage, the carrier gas is switched to a helium/oxygen mix, followed by a second temperature ramp. The split-time differs from sample to sample, depending on the laser reading.

Table 2.9 NIOSH 870 OC-EC protocol

Step	Temperature (°C)	Duration (s)
OC1	310	80
OC2	475	60
OC3	615	60
OC4	870	90
EC1	550	45
EC2	625	45
EC3	700	45
EC4	775	45
EC5	850	120
EC6	870	120

The OC and EC concentrations in $PM_{2.5}$ were calculated using equations 2.3 and 2.4. The method detection limits are given in Table A.2.

$$C_{OC} = \frac{f_{OC}.A_f}{\varphi_a.t} \text{ and } C_{EC} = \frac{f_{EC}.A_f}{\varphi_a.t} \qquad \textbf{Eq. 2.3 and 2.4}$$

C_{OC} = OC concentration in $PM_{2.5}$ concentration in µg m^{-3} where the volume is taken at ambient conditions near the inlet during sampling.

C_{EC} = EC concentration in $PM_{2.5}$ concentration in µg m^{-3} where the volume is taken at ambient conditions near the inlet during sampling.

f_{OC} = OC filter concentration in µg cm^{-2}.

f_{EC} = EC filter concentration in µg cm^{-2}.

A_f = Sampled area for 47 mm filter which is equal to 13.85 cm^2.

φ_a = the flow rate at ambient conditions in m^3 h^{-1}

t = sampling time in h

2.2.9 Ion Analysis

2.2.9.1 Filter extraction

The remaining half of the sampled QF filters were used for ion chromatography (IC) analysis of $PM_{2.5}$. The filters were brought to room temperature and transferred to labelled polypropene centrifuge tubes. Digestion was carried out in 25 mL of deionised water with a resistivity of at least 18.2 MΩ.cm. The tubes were sonicated for 60 minutes at a temperature between 20 °C and 25 °C (Chow and Watson 1999). This was followed by 15 minutes of shaking, and the tubes were then left to rest overnight (12-24 hrs) at –4 °C. The contents were then filtered through a 0.45 µm PTFE syringe filter.

Cations (Na^+, K^+, Mg^{2+} and Ca^{2+}) were analysed via IC according to EPA 350.1, and anions (Cl^-, NO_3^-, SO_4^{2-}) were analysed via IC according to ASTMD6919-03. The ammonium ion nitrogen content was determined by colourimetry according to EPA 350.1. The method detection limits are given in Table A.2.

2.2.9.2 Recovery tests

Accuracy was evaluated by recovery of filters spiked with a known concentration from a standard solution prepared from a certified reference material (CRM). For cations, NH_4^+, Na^+, K^+, Mg^{2+} and Ca^{2+}, the multi-cation CRM standard (91286 *Trace*CERT®, Merck) was used, while the multi-anion CRM standard (69734 *Trace*CERT®, Merck) was used for anions Cl^-, NO_3^-, SO_4^{2-}. The anion standard was not diluted, while the cation standard was diluted with a 1:10 dilution. The 25 ml volumetric flask was made up with DI water to the mark. The spike concentration and recoveries are given in Table A4.

Three blank quartz filters were split in half, and each half was transferred to a 50 ml centrifuge tube. Three half-filters were spiked with 0.25 ml of the anion spike solution, and the other three were spiked with 0.25 ml of the cation spike solution. The filters were then allowed to air dry for 4 hours at room temperature. The spiked filters were then extracted as per the normal procedure.

2.2.10 Analytical Uncertainties

The uncertainty for ICP-MS analysis was assigned as the sum of the uncertainty of the recovery efficiency of the SRM analysis, μ_{Rm}, the uncertainty due to instrument drift, μ_{drift}, and the precision uncertainty, μ_P, as these three procedures are the main source of uncertainty in the analysis (Barwick et al. 1999; Hortellani et al. 2011). The μ_{Rm} was calculated according to equation 2.5.

$$\mu_{Rm} = Rm\sqrt{\left(\frac{\mu_{C_{RM}}}{C_{RM}}\right)^2 + \frac{s_{obs}^2}{nC_{obs}^2}} \qquad \textbf{Eq. 2.5}$$

In this equation, Rm is the mean observed analyte concentration (C_{obs}) divided by the actual SRM analyte concentration (C_{RM}), $\mu_{C_{RM}}$ is the uncertainty of the reference material SRM1648A, C_{RM} is the concentration of the SRM sample used for recovery (in µg L^{-1}), s_{obs} is the standard deviation of replicate analysis of the RM digest solution for n number of replicates and C_{obs} gives the mean analyte concentration obtained from replicate RM analysis (in µg L^{-1})

The acceptable instrument drift was set at ±15% of the internal standard (IS) concentration. If the IS concentration was greater than ±15%, re-calibration followed by re-analysis of the sample was carried out (Barwick et al. 1999; Hortellani et al. 2011). The μ_{drift} was calculated according to equation 2.6.

$$\mu_{drift} = \frac{15\%}{\sqrt{3}} = 8.66\% \qquad \textbf{Eq. 2.6}$$

Since there is no evidence of a lower probability measurement towards the limits of the ±15% range, this can be treated as a normal distribution and divided by the $\sqrt{3}$ to calculate the uncertainty related to instrument drift (Barwick et al. 1999). Precision uncertainty, μ_P, was calculated from the relative standard deviation of the ICP-MS measurements.

The analytical uncertainty, μ_{ij}, is given by the equation 2.7.

$$\mu_{ij} = C_{ij}\sqrt{\left(\frac{\mu_{RM}}{C_{obs}}\right)_{ij}^2 + \left(\frac{\mu_{drift}}{100}\right)_{ij}^2 + \left(\frac{\mu_p}{100}\right)_{ij}^2} \qquad \textbf{Eq. 2.7}$$

In this equation, μ_{ij} gives the analytical uncertainty, in μg L^{-1}, in sample *i* and element *j* for C_{ij}, measured concentration, in μg L^{-1}, in sample *i* and element *j*.

The analytical uncertainty, in μg L^{-1}, for ion chromatography was assigned as the sum of the uncertainty of the recovery efficiency of the SRM analysis and μ_{Rm}, the uncertainty due to instrument drift, μ_{drift}, as these two procedures are the main source of uncertainty in the analysis. This is given by the equation 2.8.

$$\mu_{ij} = C_{ij}\sqrt{\left(\frac{\mu_{RM}}{C_{obs}}\right)^2_{ij} + \left(\frac{\mu_{drift}}{100}\right)^2_{ij}} \qquad \textbf{Eq. 2.8}$$

For the OC/EC analysis, the analytical uncertainties for each sample were provided by the laboratory in μg cm^{-2}. All analytical uncertainties, $\mu_{ij,}$ were converted to ng m^{-3} and used to calculate the error estimates for PMF as described in equation 3.2.

2.2.11 Air exchange rate analysis

Air exchange rates at the Birżebbuġa sampling site were calculated by measuring the concentrations of a tracer gas with time. Carbon dioxide was used as the tracer gas. The model is based on the decay of CO_2 gas inside the sampling room based on (Aglan 2003).

The differential equation, which describes the flow rate of a tracer gas such as CO_2 based on mass balance factors, is given by equation 2.9.

$$\frac{\partial(VC_i(t))}{\partial t} = qC_o - qC_i(t) + S \qquad \textbf{Eq. 2.9}$$

V = the inside volume (m^3)

$C_i(t)$ = the indoor tracer gas concentration at a given time t (μg m^{-3})

C_0 = the outdoor background tracer gas concentration (μg m^{-3})

q = the flow rate (m^3 h^{-1})

S = the indoor tracer gas emission rate (μg h^{-1})

Rearranging equation 2.9 and integrating it over the initial and final CO_2 tracer gas concentrations and time gives equation 2.10, which describes the indoor CO_2 tracer gas concentration as a function of time, *t*.

$$C_i(t) = \left(C_0 + \frac{S}{q}\right) - \left(C_0 + \frac{S}{q} - C_{i0}\right) e^{-\frac{q}{V}(t-t_0)} \qquad \textbf{Eq. 2.10}$$

During the decay of the tracer gas, the emission rate of the tracer gas is zero, and equation 2.10 can be rearranged so that the indoor tracer gas concentration can be modelled using equation 2.11. In this equation, q/V is equal to the AER in h^{-1}.

$$C_i(t) = C_0 - \left(C_0 - C_{i0}\right) e^{-\frac{q}{V}(t-t_0)} \qquad \textbf{Eq. 2.11}$$

An R script (Vella Ryan 2019) was used to determine AER by fitting Equation 2.11 to the time series data for the decay of indoor CO_2 concentration. The minimum and maximum CO_2 concentrations were used for C_0 and ($C_0 - Ci_0$), respectively. The plots are shown in Figure A.2.

The CO_2 concentration inside the room was increased by releasing CO_2 gas inside the sampling room until a concentration of at least 5000 ppm was achieved. The concentrations were checked at least at three different points inside the room to ensure a homogeneous concentration inside the room and to ensure that equilibrium with the indoor environment was established. Once the concentration of CO_2 exceeded 5000 ppm, its concentration was measured every minute by a CO_2 data logger (Extech SD800). As the gas emission was stopped, the indoor CO_2 concentrations decreased. The readings were taken for 10 hours or until the CO_2 concentration fell to ambient levels of around 500 ppm. The decay of the CO_2 tracer gas inside the sampling room was then used to calculate the air exchange rate according to equation 2.11.

The AER for the sampling room was calculated when the outdoor windows were completely closed, slightly open (about 5 cm) and half-open. A duplicate reading was taken for each scenario.

2.3 Continuous Indoor and Outdoor PM measurements

2.3.1 Equipment

Indoor and outdoor PM concentrations were sampled using two calibrated aerosol spectrometers (Fidas Palas Fine Dust Monitoring System 200S). These allow a constant and stable measurement of particle concentration values and mass fractions for fine dust particles in the size range 180 nm – 18 μm. The calculated hourly PM_1, $PM_{2.5}$ and PM_{10} fractions were used for this analysis. The Fidas 200S is certified according to EN 16450 and EN 15267. The indoor and outdoor spectrometers were calibrated at each site using monodisperse particles (Palas MonoDust 1500). The sampling head is located 1.2m above ground for the indoor sampler and 1.5m above ground for the outdoor sampler. Indoor and outdoor weather parameters were measured using a Davis Vantage Vue indoor and outdoor weather station.

2.3.2 Site Selection

Sampling was carried out in six households in Malta. The location of the residences is shown in Figure 2.3, and the description of each sampling site is given in Table 2.10. Indoor samplers were placed in the living room of each site, but the outdoor samplers were placed either on balconies, roofs or in backyards. The position of the outdoor sampler depended on the best available and safe location permitted at each site. An information sheet describing the site characteristics was filled out for each sampling site, and the detailed site characteristics are given in Table 2.11. Continuous indoor and outdoor PM sampling was carried out for a minimum of 28 days at each site.

In 2021 in Malta, 35% of main residential dwellings are flats/apartments/penthouses, followed by 34% for terraced houses and ground floor maisonettes, with semi/fully-detached houses making up less than 5% of main residential dwellings (National Statistics Office 2021). In 2021, the most common occupancy for residential dwellings in Malta was two-person occupancy (28%), followed by one (27%), three (18%) and four-person (14%) occupancy. Although this study aimed to sample at least 12 residences, with a representative sample of the dwelling characteristics, the spatial distribution of residences in Malta and seasonal characteristics, only six sites were sampled. Unfortunately, several faults developed in both

Table 2.10 Description of the sampling sites.

Site number*	Location	Sampling period (mm/yy)	Sampling Days	Location of Outdoor Sampler	Number of residents	Type	Type
2†	Naxxar	01/21–02/21	53	Roof	2	Terraced 3-storey building	UB
3	Qrendi	08/23–09/23	28	First Floor Terrace	3	Terraced 3-storey building	UB
4	Fgura	11/23–01/24	36	Roof	4	Terraced 3-storey building	U
5	Naxxar	01/24–02/24	37	Back yard	4	Detached 2-storey building	UB
6	Gozo	02/24–03/24	37	Back yard	4	Terraced 3-storey building	UB
7	Sliema	04/24–05/24	41	Third Floor Terrace	4	3rd floor flat in a 4-storey building	U

Notes: *Site numbering was started at two to maintain consistency in this study, given that site one is the Birżebbuġa sampling site, where the $PM_{2.5}$ data collected was used for the PMF analysis.
†Data from site two was obtained from a previous study (Aquilina and Camilleri 2022)
Urban background (UB), Urban (U)

samplers that necessitated repairs and the instruments had to be sent for repairs and re-calibration several times abroad. This severely limited the sampling period, and in total, six sites were sampled, with data for site two obtained from Aquilina and Camilleri 2022. Hence, this study does not properly represent flats/apartments/penthouses and singly-occupied dwellings. As shown in Figure 2.3, most of the sampling occurred in the Northern, Northern Harbour and Southern Harbour districts, as these districts constitute 65% of the population in Malta and Gozo (National Statistics Office 2023).

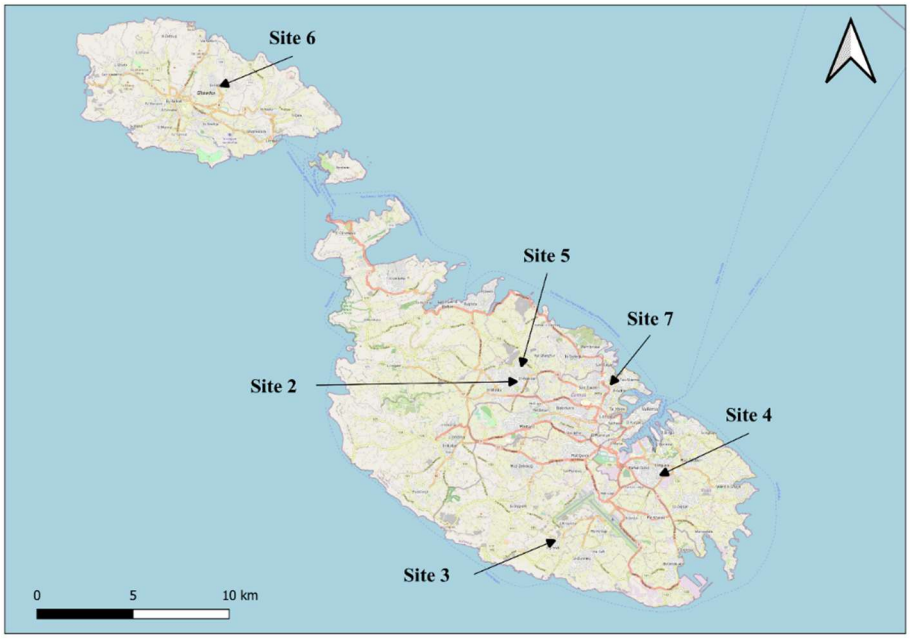


Figure 2.3 Map indicating the geographic location of the six sampling sites in Malta and Gozo.

Table 2.11 Site characteristics of the sampling sites

Site characteristics	2	3	4	5	6	7
Garage connected directly to residence	Yes	Yes	Yes	Yes	Yes	No
Cigarette smoking	No	No	No	No	No	No
e-cigarette smoking	No	No	No	No	No	No
Pets	No	Yes	Yes	Yes	Yes	No
Garden	No	Yes	Yes	Yes	Yes	No
Natural Ventilation	Yes	Yes	Yes	Yes	Yes	Yes
Open windows	No	Yes	Yes	Yes	Yes	Yes
Mechanical/filtered ventilation	No	No	No	No	No	No

Notes: The "Open windows" was marked yes when the residence left indoor windows slightly open whilst in the residence.

2.3.3 Meteorological Data and Indoor Activities.

Hourly indoor and outdoor meteorological data was collected during the sampling period. The data included wind speed and direction, atmospheric pressure, temperature, humidity and rainfall. During sampling days, residents were asked to keep a log of indoor activities that are known to generate an abnormal amount of dust, such as construction works inside or adjacent to the residence, open flames such as BBQs, wood fireplaces/ovens and candles and fireworks displays in the same town/village. A log of the time spent sweeping and vacuuming inside the residence was also requested. Table A.5 gives a sample of the log of these activities.

Chapter 3

Source Apportionment of Indoor $PM_{2.5}$ at a Residential Urban Background Site in Malta

3.1 Introduction

The Maltese Islands are an archipelago located in the central Mediterranean Sea. The archipelago is geographically small and densely populated (*ca.* 500 000), with the two largest inhabited islands being Malta (316 km^2) and Gozo (67 km^2). The daily levels of outdoor particulate matter (PM) measured at the stations representing Malta's most trafficked areas and urban background areas exceed the Air Quality Directive limit values of $PM_{2.5}$ and PM_{10} several instances during the year. This is a recalcitrant problem in Malta (Scerri et al. 2018; Fenech and Aquilina 2020). Receptor modelling on $PM_{2.5}$ in a coastal town in Malta by Scerri et al. (2018) identified traffic, shipping and fireworks as the main anthropogenic sources contributing over 35% of the $PM_{2.5}$ at that site. Anthropogenic activities were not limited to local emissions, with a significant contribution by ammonium sulfate (23.6%) attributed to regional/long-range transport (Viana et al. 2008). However, no research has been conducted so far on the chemical speciation and source apportionment of airborne indoor $PM_{2.5}$ in Malta.

The correlation of ambient air pollution to health diseases and increased morbidity and mortality is widely established (Brunekreef and Holgate 2002; IARC 2015; World Health Organization 2016) and traditionally, studies have been focused on outdoor air quality. Since people spend most of their time indoors (Klepeis et al. 2001; L. Yao et al. 2015; Bo et al. 2017), indoor air quality has been given significant attention in the past decade. Many studies have been conducted on indoor air quality, but these studies are generally related to the natural, industrial or urban effects on indoor air quality (Macneill et al. 2014; Krasnov et al. 2015; Almeida-Silva et al. 2016b).

This work aims to use PMF to characterise and quantify the sources affecting indoor $PM_{2.5}$ levels at an urban background site, where the pollution levels are not significantly determined by any single source. Urban background sites represent the integrated contribution from many sources, including all upwind sources. Due to the historic Catholic religious influence in the Maltese Islands, all towns and villages celebrate their local patron saint during a week called the *festa* or feast. Several social and religious activities are generally organised during the feast, during which a substantial amount of fireworks are let off. Special interest is being given to the effect of the seasonal burning of fireworks across the Maltese Islands and whether this has an impact on the indoor air quality of residences that are not located in the vicinity of fireworks displays. Over 30 local pyrotechnic manufacturing sites produce fireworks for the religious

feasts, 85 of which are celebrated during the summer months between the end of June and September. In addition, a three-day international fireworks festival is held in April and serves as a main tourist attraction during a period of relative pyrotechnic inactivity. $PM_{2.5}$ and PM_{10} originating from fireworks have been recorded in ambient air across the Maltese islands (Camilleri 2008; Vella et al. 2015; Camilleri and Vella 2016; Scerri et al. 2018), but there is a dearth of data both locally and abroad on the indoor levels of fine particulate matter originating from fireworks.

3.2 Meteorological data

Figure 3.1 summarises the wind direction logged at the site during the sampling period. It shows wind directions and speeds typical of the Maltese islands, with over 55% of the wind blowing from a westerly to a north-westerly direction. Since the site is located in the south-east of Malta, it is effectively downwind from pollution sources over the Maltese islands and major fireworks displays when the wind blows from W–NW. The Delimara power station and the Birżebbuġa Free Port are around 2 km in an easterly to south-easterly direction from the sampling site.

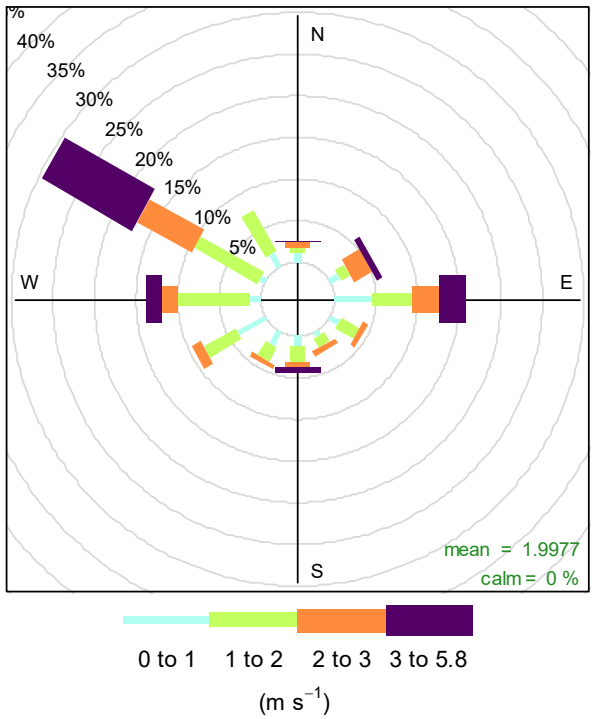


Figure 3.1 Windrose for the Birżebbuġa sampling site, June 2018 – June 2019

The mean monthly indoor and outdoor $PM_{2.5}$ concentrations by wind direction are given in Figure 3.2. While a decrease is recorded for both indoor and outdoor concentrations during the winter period, the indoor concentrations were lower than the outdoor ones during the winter period, most probably due to lower infiltration of fine dust. This can be attributed to the lower air exchanges from November till May, as outdoor windows and doors were kept closed during this period. It also suggests that there are limited indoor sources inside the residence as the indoor $PM_{2.5}$ concentration were consistently lower than 5 µg m^{-3} during this period.

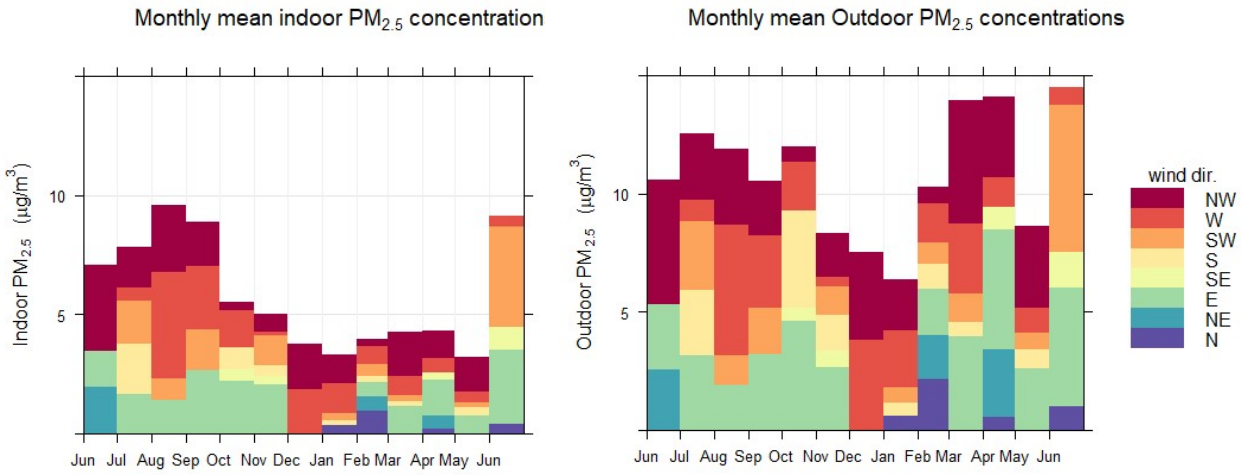


Figure 3.2 Monthly mean indoor and outdoor $PM_{2.5}$ concentrations

For both indoor and outdoor $PM_{2.5}$, the highest $PM_{2.5}$ concentrations were recorded when the wind was either from the south-westerly to north-westerly or easterly directions.

3.3 Descriptive statistics

Descriptive statistics for the daily indoor and outdoor concentrations of $PM_{2.5}$, OC, EC, elements and ions for concentrations are given in Table 3.1. Indoor $PM_{2.5}$ concentration ranged from 1.8 to 18.6 µg m^{-3}, with outdoor $PM_{2.5}$ ranging from 3.2 to 47.6 µg m^{-3}. On average, outdoor concentrations were 5.1 µg m^{-3} higher than indoor concentrations. All elements, except for As, recorded a higher outdoor daily mean concentration. The concentrations of Ca and Zn were below the method detection limit (MDL) for more than half the samples.

Table 3.1 Descriptive statistics of concentrations of chemical species. *Concentrations for $PM_{2.5}$, OC and EC (in μg m^{-3}) and elemental concentrations (in ng m^{-3}). † The number of samples for EC and OC was 158.

Pollutant (N=159)†	Indoor concentrations*							Outdoor concentrations*						
	Mean	SD	Geometric mean	Q1	Q2	Q3	% of samples below the MDL	Mean	SD	Geometric mean	Q1	Q2	Q3	% of samples below the MDL
$PM_{2.5}$	5.7	3.1	5.0	3.4	4.5	7.6	0	10.8	5.6	9.8	7.5	9.6	12.4	
OC	2.5	0.7	2.5	2.0	2.4	3.0	0							
EC	0.4	0.2	0.3	0.20	0.30	0.46	0							
Na	91	98	60	31	53	101	0	380	510	220	100	220	440	0
Mg	19	21	12	7.1	12	24	0	71	93	46	24	46	80	0
Al	35	63	17	8.3	16	34	12	91	209	31	14	26	74	14
K	67	110	30	7.2	34	76	33	130	190	72	41	73	117	8
Ca	60	77	32	15	15	84	61	170	323	74	15	84	179	43
Ti	2.7	4.0	1.5	0.75	1.5	2.7	4	6.2	15	2.2	0.9	1.9	4.8	3
V	2.6	3.4	1.5	0.73	1.7	3.1	0	5.9	5.9	3.7	1.8	4.0	8.3	0
Cr	0.28	0.8	0.16	0.10	0.19	0.29	18	0.36	0.42	0.24	0.15	0.27	0.43	18
Mn	0.82	0.9	0.53	0.29	0.51	1.1	3	1.5	2.5	0.86	0.45	0.84	1.4	1
Fe	62	54	45	26	42	78	0	71	156	33	16	29	59	0
Ni	1.2	1.5	0.51	0.085	0.81	1.7	25	2.3	2.4	1.1	0.5	1.7	3.4	25
Cu	1.3	2.4	0.55	0.28	0.66	1.2	10	1.4	1.8	0.97	0.58	0.96	1.50	1
Zn	2.6	3.9	1.2	0.45	0.45	3.8	58	6.7	9.6	3.40	1.73	4.70	8.10	32
As	0.26	0.2	0.17	0.11	0.16	0.28	6	0.18	0.17	0.14	0.12	0.15	0.21	25
Sr	0.59	1.5	0.19	0.040	0.18	0.54	31	1.7	2.9	0.71	0.34	0.64	1.46	8
Cd	0.030	0.034	0.020	0.0090	0.0090	0.039	48	0.080	0.14	0.046	0.026	0.048	0.088	25
Sb	0.7	1.6	0.23	0.097	0.23	0.	3	1.1	1.8	0.51	0.23	0.52	0.96	1
Ba	1.6	5.5	0.24	0.078	0.23	0.77	11	3.1	7.6	0.86	0.31	0.78	1.8	1
Pb	1.9	2.9	1.1	0.58	1.0	2.1	0	5.2	15	2.7	1.6	2.8	4.7	0
Cl^-	80	87	54	30	52	86	16	11	5.6	9.8	7.5	9.6	12	
NO_3^-	210	310	98	68	120	220	0							
SO_4^{2-}	1500	1400	1000	510	930	2000	0							
NH_4^+	530	440	390	230	340	680	0							

3.4 Air exchange rates and $PM_{2.5}$ I/O ratios

Table 3.2 shows the air exchange rates in the sampling room at the sampling site. The plots showing the actual and modelled CO_2 decay rate are given in Figure A.2.

Table 3.2 Air exchange rates in the sampling room with windows (a) closed (b) slightly open, and (c) half-open.

	Sampling condition	AER run 1 (h^{-1})	AER run 2 (h^{-1})
a	Sampling room window closed	0.20	0.19
b	Sampling room window slightly open	0.52	0.52
c	Sampling room window half-open	1.62	1.63

The small indoor window connecting the sampling room to the kitchen was left completely open during all the AER tests. Therefore, when the exterior windows were closed, a part of the AER could result from the exchange of air with the indoor environment of the house.

The results clearly show little air exchange when the outdoor windows are closed. This represents the environment inside the sampling room from November till May. The outdoor windows were kept completely closed during this period; hence, the indoor environment at the sampling site showed a limited air exchange with the outdoor environment.

Windows were kept either partially or completely open during the hotter months. As the air exchange rates increase when windows are opened, indoor air becomes more subject to outdoor air pollution. Effectively, the air inside the sampling room is exchanged more than 35 times daily when the windows are half-open.

Indoor/Outdoor (I/O) ratios calculated for $PM_{2.5}$ concentrations at the sampling sites support the evidence of the AER data. A significant decrease in the I/O ratios is registered during the period when the outdoor windows were closed, as shown in Figure 3.3.

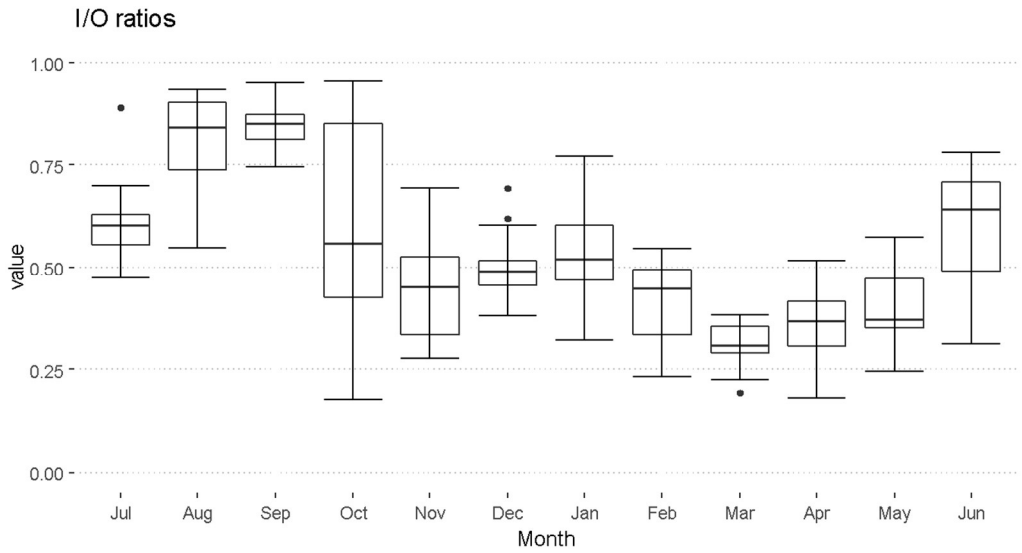


Figure 3.3 $PM_{2.5}$ I/O ratios from July 2018 to June 2019.

Table 3.3 Mean ± standard deviation (SD) for $PM_{2.5}$ I/O ratio

Period	Mean I/O ratio	SD
June 18 – October 18, June 19	0.70	0.10
November 18 – May 19	0.46	0.12

When windows were slightly or completely open, the indoor $PM_{2.5}$ concentrations were at an average of 0.70 compared to the outdoor ones (Table 3.3). This decreased to 0.46 during the period when the windows were closed. This suggests that there are minimal indoor sources for $PM_{2.5}$ in the house, and hence, the main pollution sources that affect the indoor environment seem to be of outdoor origin, especially when the windows are open during the hot summer months.

3.5 Exploratory Statistical Analysis

3.5.1 Analysis of the outlier values for daily $PM_{2.5}$ concentrations

A statistical analysis was carried out to identify which daily $PM_{2.5}$ concentrations were statistical outliers. Seven days were identified as outliers for outdoor $PM_{2.5}$ concentrations and three for indoor concentrations. HYSPLIT back trajectories were run for all these dates to investigate these outliers (Figure B.1). The trajectories were run at an arrival height of 500, 1000 and 1500 m above ground level (AGL). The 1500 m trajectory was considered to analyse Saharan dust transport as the air mass at this height is in the upper part of the boundary layer,

and surface artefacts are therefore avoided. Moreover, this height can represent the mean transport wind at a synoptic scale within the upper boundary layer (Salvador et al. 2007; Scerri et al. 2019).

Table 3.4 Data for days with outlier indoor and outdoor $PM_{2.5}$ concentrations. The outlier $PM_{2.5}$ concentrations are highlighted in grey.

Date	Indoor $PM_{2.5}$ ($\mu g\ m^{-3}$)	Outdoor $PM_{2.5}$ ($\mu g\ m^{-3}$)	Wind Direction	Saharan dust event	Windows
13/08/18	14.9	16.9	248	No	Open
29/10/18	8.4	47.6	170	Yes	Closed
25/03/19	6.4	20.9	287	No	Closed
22/04/19	5.9	28.6	101	Yes	Closed
25/04/19	6.8	37.2	67	Yes	Closed
13/06/19	12.7	20.5	79	Yes	Open
15/06/19	18.2	23.9	235	Yes	Open
17/06/19	18.6	25.9	95	Yes	Open

The trajectories show that six of these outliers were influenced by air masses that passed over Saharan dust in the previous four days. The outlier recorded on the 13th of August 2018 for indoor concentrations does not originate from a Saharan dust event. Given the westerly wind direction, the increase in $PM_{2.5}$ is most likely due to fireworks burnt during this time of year, when six villages in Malta celebrate the Feast of the Assumption by burning several fireworks displays during the whole week. On this day, indoor concentrations for K, V, As, Cd, Sb and Ba were registered as outliers. K, Sb and Ba are all associated with fireworks emissions. Heavy rainfall was recorded on the 14th of August, which decreased the indoor $PM_{2.5}$ concentrations slightly for the 14th and 15th of August. These days registered slightly lower indoor $PM_{2.5}$ levels but significantly higher concentrations in metals associated with fireworks emissions.

The Saharan dust event in June 2019 translated into high indoor and outdoor concentrations. The outdoor windows at the residence were open during this period, thus allowing high infiltration. Other Saharan dust events did not translate into significantly higher indoor concentrations, and this can be attributed to the lower infiltration of fine dust when the exterior windows are closed. This suggests that even though the outdoor fine PM concentrations increase significantly and typically exceed the daily limits during these events, indoor fine particulate matter concentrations do not increase substantially if the infiltration of fine particles

is kept low. Thus, high pollution events such as Saharan dust events will have minimal effect on indoor air pollution if the windows are closed.

The Saharan source profile identified through the PMF analysis of $PM_{2.5}$ dust collected from Mediterranean cities identifies Si, Ti, Al, Ca, Mg, Na, K and Fe (Amato et al. 2016; Scerri et al. 2018). The Saharan dust events were accompanied by higher elemental concentrations for all these elements. Other elements, such as Ba and Sr, also showed a significant increase, as shown in Figure 3.4. During these dust events, the concentrations of these elements increase by at least ten times over the yearly median concentration. Such an increase was not registered for Sb and Cu.

Although Ba and Sr also increased during the Saharan dust event, the sharp increase in Cu and Sb was observed only when fireworks were let off, and this could be used to distinguish these fireworks pollution events from the natural origins of Saharan dust events. Further investigation is required, but such data could be used to constrain the PMF model to aid the separation of fireworks and Saharan dust/crustal source profiles.

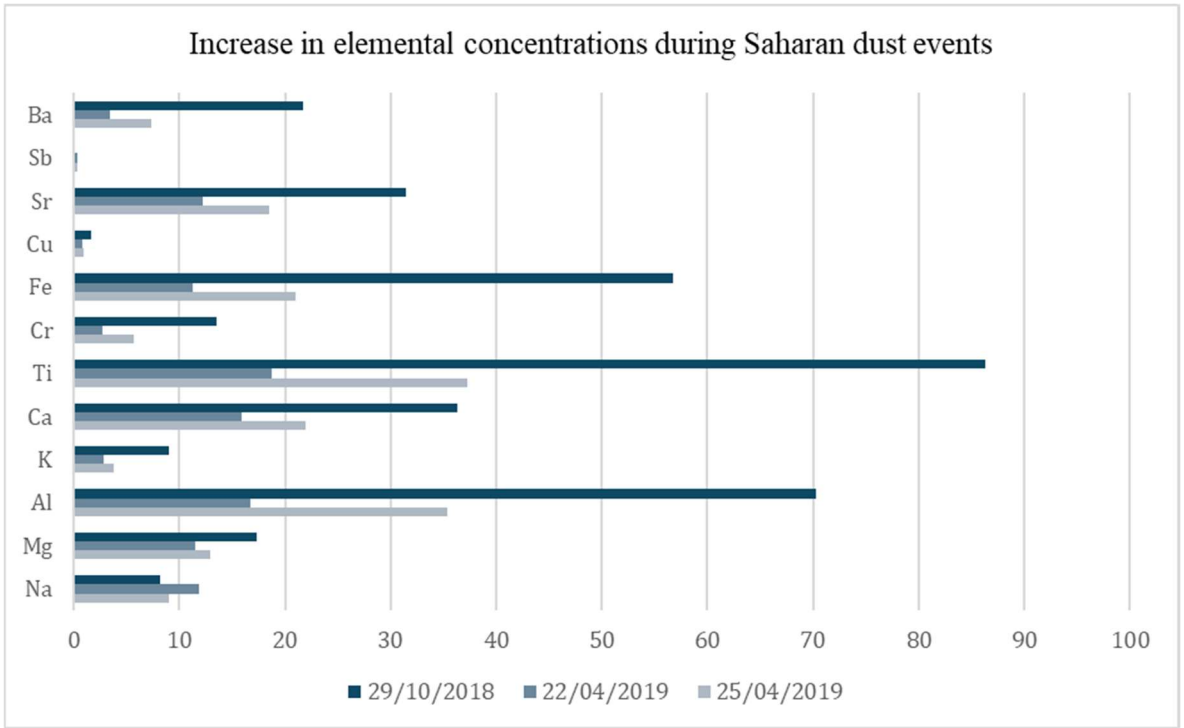
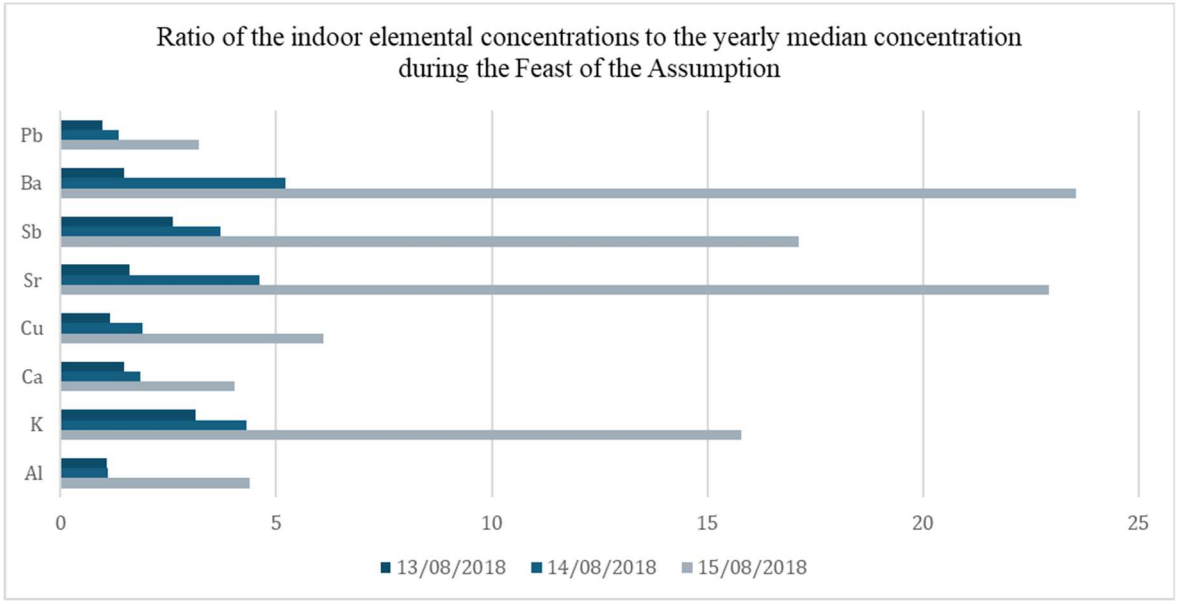
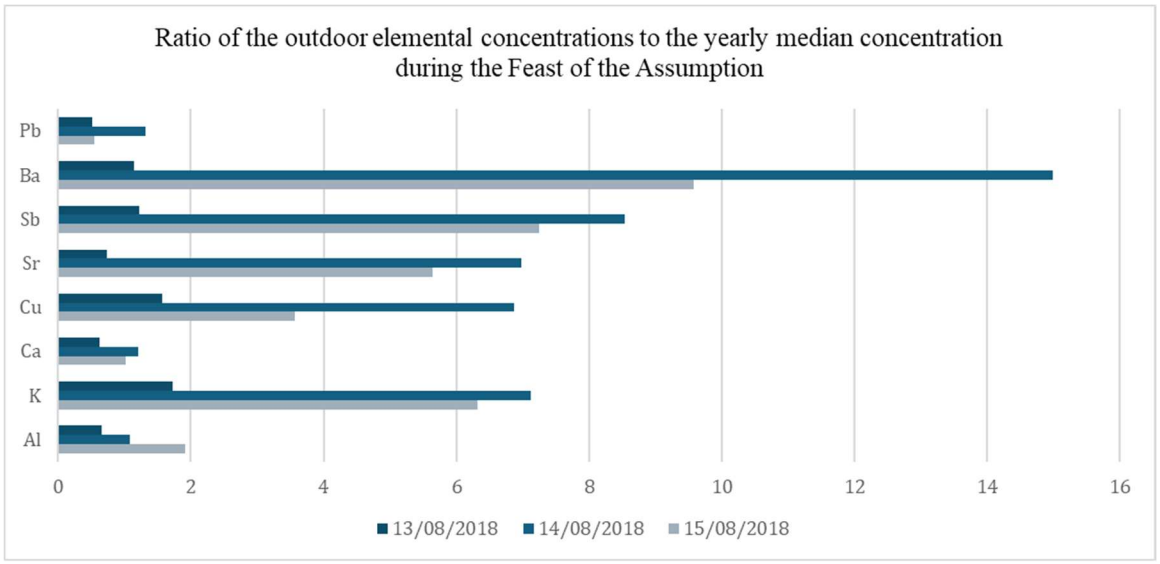


Figure 3.4 The outdoor elemental percentage increase over the yearly median concentration during Saharan dust events.

The indoor elemental concentration for toxic elements such as Sb and Ba increases by an order of magnitude over the yearly median concentration, as shown in Figure 3.5a, during mid-August. It is important to emphasize that there were no fireworks displays during this period at the sampling site since the Feast of the Assumption is not celebrated in Birżebbuġa. The closest fireworks displays were at least 3.5 km away from the sampling site.



(a)



(b)

Figure 3.5 Ratio of the (a) indoor and (b) outdoor elemental concentrations to the yearly median concentration during the mid-August Feast of the Assumption.

Table 3.5 I/O ratios for mid-August 2018.

	I/O ratio								
	$PM_{2.5}$	**Al**	**K**	**Ca**	**Cu**	**Sr**	**Sb**	**Ba**	**Pb**
13/08/2018	0.88	0.59	0.95	0.71	0.72	0.79	1.25	0.80	0.65
14/08/2018	0.74	0.37	0.32	0.46	0.28	0.24	0.25	0.22	0.35
15/08/2018	1.01	0.82	1.32	1.21	1.70	1.47	1.39	1.53	2.00
16/08/2018	0.91	0.40	0.51	1.19	0.49	0.43	0.46	0.34	0.48

In the absence of indoor sources, the I/O ratio of a species should be less than 1. Except for Al, all the I/O ratios for the elements for August 15th listed in Table 3.5 are greater than 1, which means that the indoor concentrations are higher than the outdoor ones. The fact that indoor concentrations of these elements are typically low suggests that there are no significant indoor sources. Given the periodic higher outdoor concentrations, an explanation for the high I/O ratios could be that these elements are being concentrated in the indoor environment. The prolonged burning of fireworks in various locations during mid-August generates pollutants only at specific periods during the day, mostly during late evening when most fireworks displays are held. When the plumes reach the household, they increase the outdoor concentration of these pollutants and, subsequently, indoor concentrations via infiltration. The infiltration can be high if windows are open. Outdoor pollutant dispersion is generally higher, and the plume is quickly carried away, especially at wind speeds greater than 2 m s^{-1} recorded on the 15th of August.

On the other hand, indoor concentrations may decrease at a slower rate as this depends on the building's air exchange rates. This is significant as it exposes residents to higher concentrations of these toxic elements for extended periods. Yao et al. 2015 has shown a similar increase in I/O ratios for perchlorate in Jinan, China, ranging from 0.12 to 16.47. Here, too, no indoor sources for perchlorate were detected, and I/O ratios higher than 1 could only be explained by the extensive fireworks displays close to the sampling site coupled with the fact that indoor pollutant diffusion was poorer than outdoor ones, which resulted in the accumulation of this pollutant indoors.

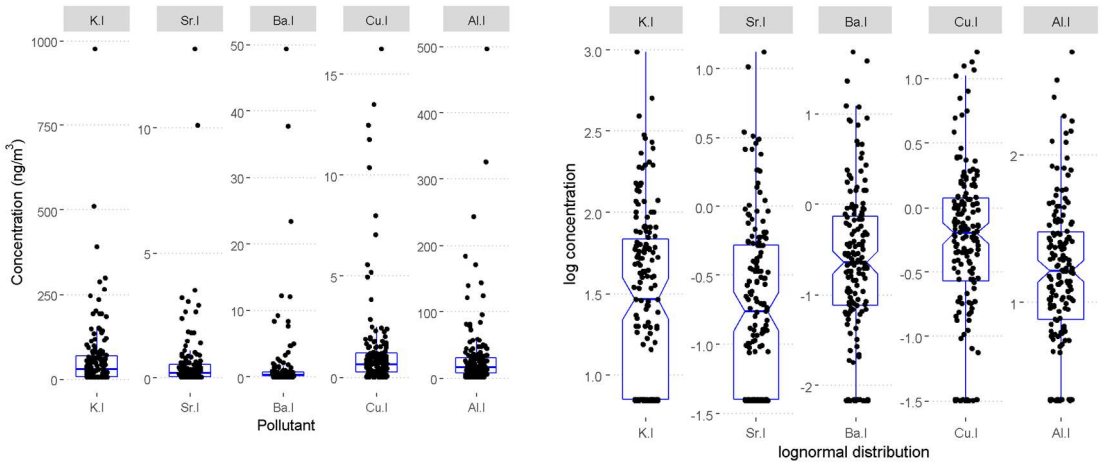
3.5.2 Tests for Normal Distribution of Data

The data was tested with the Shapiro-Wilk test for normal distribution. The null hypothesis is that the data is normally distributed. The p-value for all the data is significantly lower than the 0.05 acceptance criteria at 95% confidence level; hence, the data is not normally distributed.

Table 3.6 Shapiro-Wilk test for normally distributed data for selected elements

Analyte	p-value (Shapiro-Wilk test)	Analyte	p-value (Shapiro-Wilk test)
$PM_{2.5}$	6.6e-10	**Sr**	< 2.2e-16
Na	1.365e-15	**Ni**	3.382e-15
Cu	< 2.2e-16	**V**	2.29e-16
Al	< 2.2e-16	**Sb**	< 2.2e-16
K	< 2.2e-16	**OC**	2.751e-07
Ba	< 2.2e-16	**EC**	8.132e-09

In fact, Figure 3.6a clearly shows the presence of outliers in the data, confirming that the data is skewed. A lognormal distribution is more representative of the data, and Figure 3.6b clearly shows a reduction in the number of outliers in the data.



(a) **(b)**

Figure 3.6 Box plots for (a) normal and (b) lognormal distributed indoor data

3.5.3 Wilcoxon rank-sum tests for indoor and outdoor samples

The Wilcoxon rank-sum tests can be used as a non-parametric test for equal distribution and should be used when the hypothesis of normality is not justified (Reinmann et al. 2008). The

null hypothesis tested is that the underlying distributions of a chemical species in indoor and outdoor environments are equal.

Table 3.7 Outcome of the Wilcoxon rank-sum tests for two samples.
* shows paired medians present significant differences ($p < 0.05$)

Species	p-value	Higher median	Species	p-value	Higher median
$PM_{2.5}$	0.0000*	Outdoor	**Cu**	0.0000*	Outdoor
Na	0.0000*	Outdoor	**Zn**	0.0000*	Outdoor
Mg	0.0000*	Outdoor	**As**	0.0440*	Minimal indoor
Al	0.0000*	Outdoor	**Sr**	0.0000*	Outdoor
K	0.0000*	Outdoor	**Cd**	0.0000*	Outdoor
Ca	0.0000*	Outdoor	**Sb**	0.0000*	Outdoor
Mg	0.0000*	Outdoor	**Ba**	0.0000*	Outdoor
Fe	0.0002*	Indoor	**Pb**	0.0000*	Outdoor
Ni	0.0000*	Outdoor			

The Wilcoxon rank-sum tests were run for the outdoor and indoor species. The results are shown in Table 3.7. The alternative hypothesis ($p < 0.05$) that paired medians represent significant differences was accepted for all species. All the chemical species, except Fe and As, had significant ($p < 0.05$) outdoor medians higher than indoor ones, indicating that these species have minimal indoor sources. On the other hand, Fe had a significant ($p < 0.05$) higher indoor median, which suggests that indoor sources could contribute to this element.

3.6 The PMF solution

3.6.1 Error estimates for PMF

Values equal to or below the MDL were expressed as half the detection limit (Polissar et al., 1998). The uncertainties for these samples were calculated as a fixed fraction of the MDL given by equation 3.1 (Norris et al., 2014; Polissar et al., 1998). In doing so, a high uncertainty loading is given to values below the MDL in the PMF solution.

$$\sigma_{ij} = \frac{5}{6}(\overline{MDL_j}) \qquad \textbf{Eq. 3.1}$$

In equation 3.1, σ_{ij}, gives the PMF uncertainty, in ng m^{-3}, in sample *i* and element *j* and $\overline{MDL_j}$ is the mean MDL, in ng m^{-3}, for element *j*. Although PMF allows an equation-based

uncertainty calculation, this approach was not used for this analysis as it would not capture errors associated with specific samples and analytes. Therefore, for concentrations greater than the MDL, the PMF uncertainties for each analyte concertation were calculated according to equation 3.2 (Polissar et al., 1998; Reff et al., 2007) where σ_{ij} gives the PMF uncertainties, in ng m^{-3}, in sample *i* and element *j* and μ_{ij} is the analytical uncertainty, in ng m^{-3}, in sample *i* and element *j*.

$$\sigma_{ij} = \mu_{ij} + \frac{\overline{MDL_J}}{3} \qquad \textbf{Eq. 3.2}$$

3.6.2 Interpreting the PMF solution

The analytes were screened for their signal-to-noise ratios (S/N) (Paatero and Hopke 2003; Zhao et al. 2006). In PMF, analytes can be considered bad, weak, and good. In this solution, no analyte was classified as bad. Analytes with a S/N lower than four were down-weighted to weak. These include the analytes Na^+, K, Ca, Cr, Ni, Cu, Zn, As, Sr, Cd, Sb, Cl^- and NO_3^-. EC was also down-weighted to weak, even though it had a S/N ratio higher than 4, as a robust solution could not be obtained without down-weighting EC.

As PMF is essentially a descriptive model, determining the optimal number of factors fitted in the model remains subjective (R Vecchi et al. 2008), and choosing the number of factors is a compromise. The PMF solution was run for different factors ranging from 5 to 10. All base runs for the different number of factors solutions converged. The final number of factors was chosen depending on the results of the Q(Robust) and Q(True) values for each run, the intra-run residuals, the bootstrap (BS) analysis and the Bootstrap Displacement (BD-DSP) analysis (Norris et al. 2014; Brown et al. 2015).

In the 5- and 7-factor solutions, the intra-run residuals showed several base runs with significantly different fits. A very high Q/Qexp for most species also characterised the run with the lowest Q(Robust), suggesting a poor fit. Moreover, many factors were not appropriately mapped in the BS analysis, and some were left unmapped. The 8-factor solution produced near-identical intra-run residuals for the base runs and very low (<1) Q/Q_{exp} for most species. Furthermore, all factors were mapped over 95% in the BS analysis for the base run. The 9-factor solution also produced very similar intra-run residuals for the base runs but had high Q/Q_{exp} for most elements for the base run. In the BS analysis of the base run, all factors were

mapped over 95%. The 10-factor solution gave a different intra-run residual for the base runs, indicating a poorly fitted solution. The BS resulted in one factor with less than 50% mapping, indicating that a phantom factor was created, which translates into the fragmentation of a real single source over more than one factor.

The 8- and 9-factor solutions were retained as they provided the best solution for the data and were investigated further. The same constraints were applied to both solutions to decrease rotational ambiguity. Five data points sampled in August during fireworks events were pulled up to maximum for the fireworks factor. When this was carried out, the modelled $PM_{2.5}$ concentrations were closer to the sampled concentrations for these days. Given that no fireworks displays occur in winter, three data points that showed a slight contribution by the "Fireworks" factor were set to zero.

Moreover, species known to be present in fireworks, namely, chloride, nitrate, barium and antimony, were pulled up to maximum for this factor. Lead was also pulled up to maximum for the "Industrial" factor. These constraints resulted in an increase in dQ(robust) of 5.0% for the 8-factor solution and 7.1% for the 9-factor solution. This was deemed acceptable as a more interpretable solution was provided, with the Q/Q_{exp} decreasing below 1 for most species. Moreover, all factors were mapped 99% in the constrained BS analysis, and no factor swaps were registered in the displacement analysis, which indicates a robust solution with minimal rotational ambiguity.

The BS-DISP analysis on the constrained run was used to compare the robustness of the 8- and 9-factor solutions. The species selected to be displaced in BS-DISP were OC, Mg, V, Fe, Ba and SO_4^{2-} as these are the key species required for factor identification. Norris et al. (2014) suggest a small number of variables in BS-DISP to speed up the computation analysis. Three-factor swaps were registered for the 8-factor solution, while 16-factor swaps were registered for the 9-factor solution. This suggested that the 8-factor solution was more robust with a lower rotational ambiguity. The main difference between the 8- and 9-factor solution is the splitting of the indoor factor into two separate factors, one explaining the indoor OC and the other factor explaining the indoor EC.

The BS-DISP analysis showed a more robust 8-factor solution with the lowest rotational ambiguity, which was accepted as the solution that provided the largest number of physically

interpretable solutions. The solution achieved a good reconstruction of the $PM_{2.5}$ mass with the modelled mass explaining 93% of the observed mass with a 0.98 coefficient of determination (R^2) for the plot of the modelled against the observed $PM_{2.5}$ given in Figure 3.7 and RMSE of 1.1 µg m^{-3}.

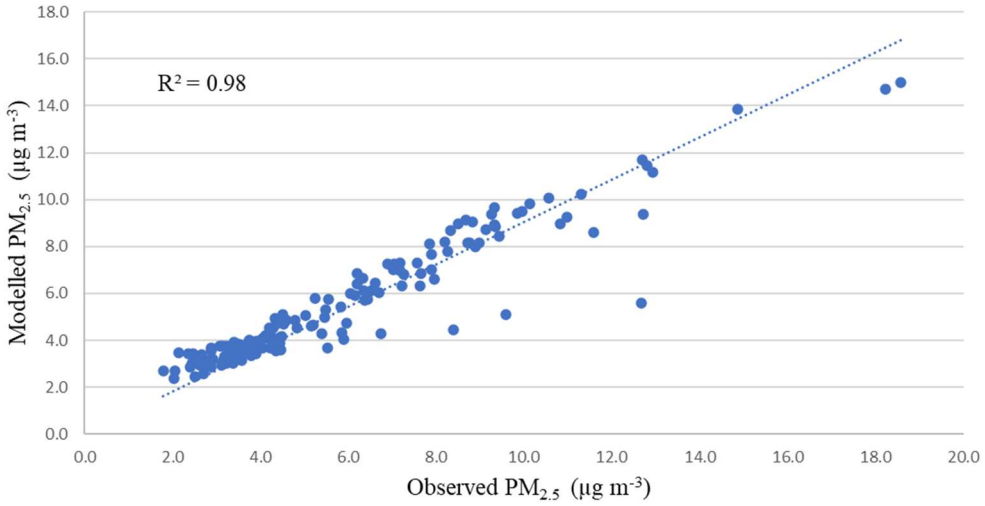


Figure 3.7 Plot of PMF-modelled $PM_{2.5}$ concentrations against observed indoor $PM_{2.5}$ concentrations.

3.6.3 PMF results for Indoor $PM_{2.5}$

The normalised source contributions from the PMF solution were converted into meaningful contributions in ng m^{-3} by multiplying the normalised source contributions against the respective source profile masses. Although there are non-negativity constraints, PMFv5 allows minimal negative source contribution; hence, slightly negative source contributions were pulled up to zero before being converted into mass since negative mass values are meaningless in terms of source contributions. The modelled PMF was compared to the observed gravimetric $PM_{2.5}$ mass, and the un-modelled mass was calculated. The mass source contributions were used to calculate polar plots for each source using the OpenAir package in R-studio (Carslaw and Ropkins 2012). Seasonal contributions were also calculated, and seasonal variations were analysed using the Kruskal-Wallis rank-sum test and the Mann-Whitney U-test in R-Studio.

The profiles for the 8-factor solution are shown in Figure 3.8. The uncertainty levels are set to the 5^{th} and 95^{th} percentile of the BS runs for the constrained solution. The yearly and seasonal contributions by each factor to indoor $PM_{2.5}$ are shown in Figures 3.10a and 3.10b, respectively.

The polar plots for the source contributions were plotted using the OpenAir package in R-Studio (Carslaw and Ropkins 2012) and are shown in Figure 3.10.

Factor 1 contributed 1505 ± 774 ng m^{-3} of the indoor $PM_{2.5}$ (26%) at the receptor site. The polar plot in Figure 3.10 shows that the contributions from this factor are not linked to any particular wind direction factor, and the highest concentrations are at the receptor site. This factor was therefore labelled "indoor". The lowest contribution was registered in summer, and the highest during autumn and winter. The seasonal difference in average contributions between summer and the other seasons is statistically significant ($p<0.01$) but not statistically significant between autumn, winter and spring ($p>0.01$). This could possibly be explained by the lower exfiltration from autumn to spring due to the lower air exchange rate (AER) at the sampling site, given that homes are generally kept more closed during this period. The mean AER measured at the sampling site ranged from 0.195 h^{-1} with closed windows to 1.625 h^{-1} with partially opened windows. The activity summary in Table A.1 shows that 40% of the sampling days were carried out with the residence's windows either partially or entirely open. Windows were kept closed from mid-October 2018 to mid-May 2019, meaning windows were only open during hot summer months. Given the higher AERs during the summer months, outdoor sources are expected to contribute more significantly to indoor $PM_{2.5}$ during this period due to the higher infiltration.

This factor is responsible for half of the OC at the receptor site, with an OC/EC ratio of 11.9. The presence of lower amounts of EC in this factor is mainly due to EC being a primary pollutant generated from incomplete combustion, and unlike OC, it is most likely caused by outdoor sources, especially since there was no indoor solid fuel combustion or cigarette smoking. A poor correlation was registered between OC and EC sampled indoors, with the Pearson correlation for log-transformed data at $r = 0.007$. This is also reflected in the PMF results, as indoor EC is mainly obtained from traffic (45.6%) while most of the OC (50.6%) is of indoor origin, which confirms different sources for OC and EC and highlights that most of the EC is of outdoor origin.

OC/EC ratios have been used to characterise emissions of carbonaceous aerosols and the formation of secondary organic carbon (SOC) (Pio et al. 2011). Indoor OC/EC ratios for Chinese cities ranged from 3.9 to 4.5 (Zhang et al. 2020), which suggested the formation of SOC. This factor's OC/EC ratios are much higher than that. Pio et al. 2011 reported an OC/EC

minimum ratio of 0.6 to 0.8 for ambient $PM_{2.5}$ in urban background environments with higher ratios associated with OC of secondary origins. Therefore, this factor's high OC/EC ratios indicate the presence of SOC resulting probably from gas-to-particle condensation from indoor combustion related to cooking, e-cigarette smoking, gas heating, and secondary organic aerosols formed from oxidation of cleaning agents. The activity summary in Table A.1 shows that a flueless butane gas heater was used during 40% of the sampling days. The gas heater was only used during the colder months when outdoor windows were kept closed, resulting in higher indoor OC due to low exfiltration. Cooking using an electric oven/toaster was carried out on most sampling days, whilst frying was carried out during 26% of sampling days. Different cooking modes, such as frying/grilling/baking/oven or toaster use, have different particulate emissions, with frying/sauteing producing the highest amount of fine particulate matter (Wallace et al. 2003). E-cigarettes were smoked indoors during 94% of the sampling days. The presence of high levels of Cr and the presence of Ti in this factor indicates that the source of these elements could be from heating elements, including those of e-cigarette smoke. Ti and Cr were found to be emitted from e-cigarette heated filaments, yet conventional cigarettes did not produce these metals (Ruprecht et al. 2017). Given that only e-cigarettes were smoked daily, this explains the presence of high levels of Cr in this factor.

While e-cigarettes also contribute to indoor organic compounds such as ethanal (0.26 % that of cigarettes) and methanal (3.4%), albeit significantly lower when compared to conventional cigarettes, they do not contribute any BC emissions (Ruprecht et al. 2017). This is also consonant with the low EC detected in the indoor factor and the high OC/EC ratios.

Factor 1 also accounts for 26.0% of Na and 26.9% of chloride at the site. These can be attributed to indoor dust resuspension from human activities and cleaning activities such as sweeping and vacuum cleaning, which were carried out during 16% and 8% of the sampling days, respectively. This factor is also the highest contributor of nitrate (31.1%). This may be attributed to using an unvented gas heater on 40% of the sampling days. The generation of NO_x from unvented gas combustion is well known (World Health Organization 2010; Arata et al. 2018). Indoor nitrate production occurs due to a lack of NO_2 photolysis. When sufficient O_3 is present from outdoor infiltration, NO_x is oxidised to nitrate. This conversion occurs after combustion stops and ceases to supply NO (Arata et al. 2018).

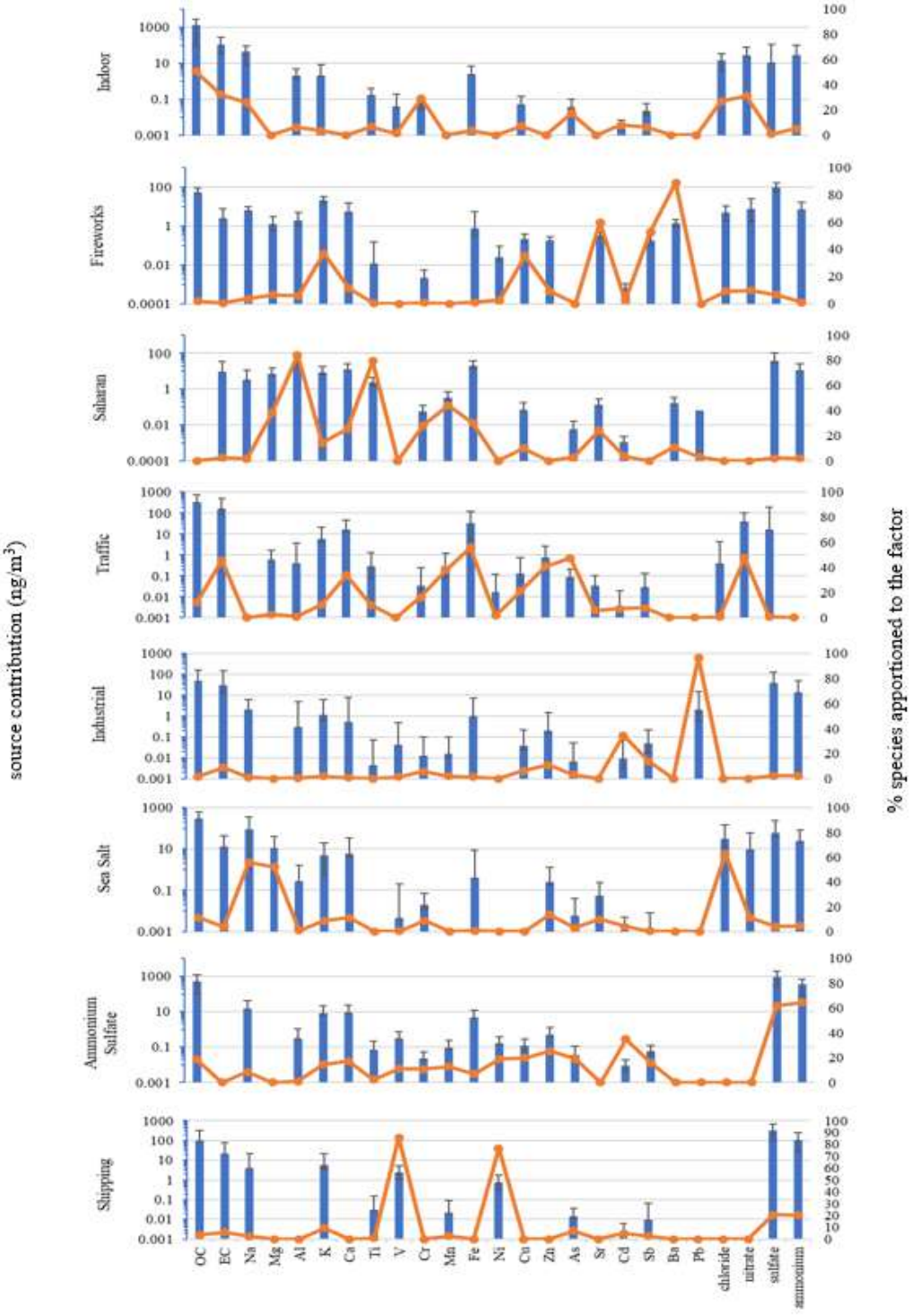


Figure 3.8 Factor profiles (bars) and percentage species apportioned to each factor (line)

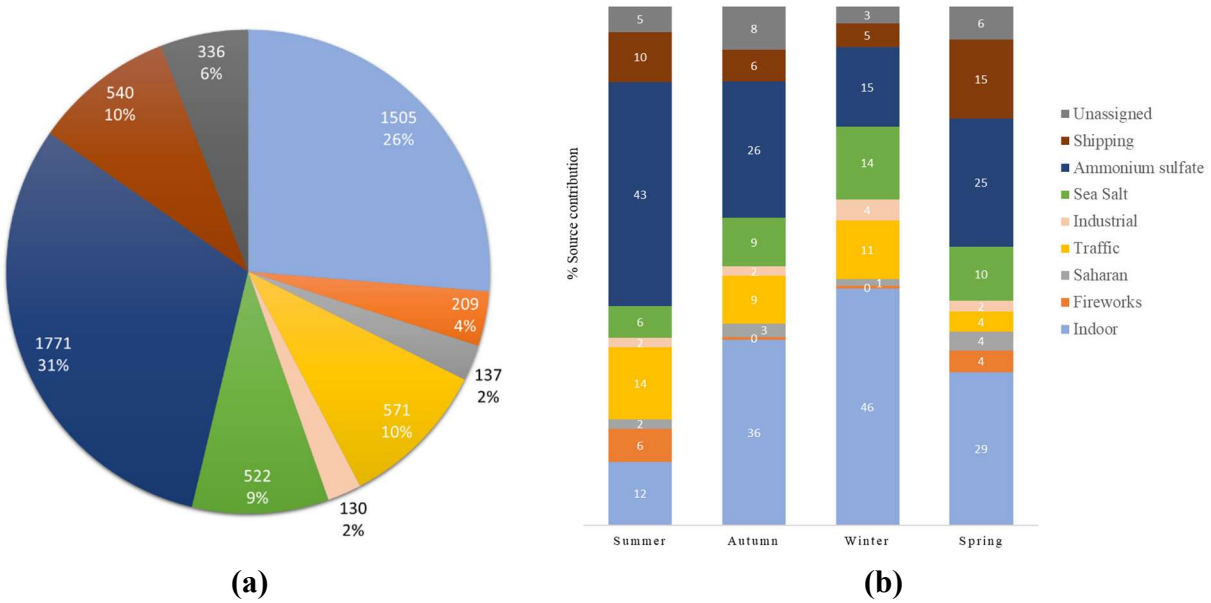


Figure 3.9 **(a)** Annual mean source contributions to indoor $PM_{2.5}$. The values are given as percentages and in ng/m^3. **(b)** Seasonal percentage source contributions to indoor $PM_{2.5}$

Factor 2 is the highest source contributor for Sr (60%) and Ba (35%). It is characterised by Cu, Sr, Ba, K, Cl^- and NO_3^-. The contribution from this factor is seasonal and is a source contributor only between June and September, which matches the local patron saints' feast season. Minimal contributions are noticed from October to May. The difference between the average contributions between the periods June to September and October to May is statistically significant ($p<0.01$). The polar plot shows the highest emissions are mainly from a NW to SW direction, as expected due to the receptor site being downwind from fireworks displays in NW to SW winds. This factor is characterised by peak emissions during 14 - 26 July (feasts in Tarxien, Valletta and Zurrieq), the five feasts celebrated on 13-15 August, 01 and 03 September (feasts in Zurrieq) and 13 and 15 June 2019 (feasts in Mqabba, Żejtun and Marsa). Reference is made to Figure B.3 for the indication of the areas letting off fireworks in the abovementioned periods with respect to the sampling site. This data is similar to that shown in a study by (Camilleri and Vella 2010). The increase in metals in PM_{10} associated with fireworks burning was reported to decrease significantly from the summer to autumn months in Malta (Camilleri and Vella 2010). Similarly, this study recorded a significant decrease in fireworks-related $PM_{2.5}$ at the receptor site from the summer to autumn months.

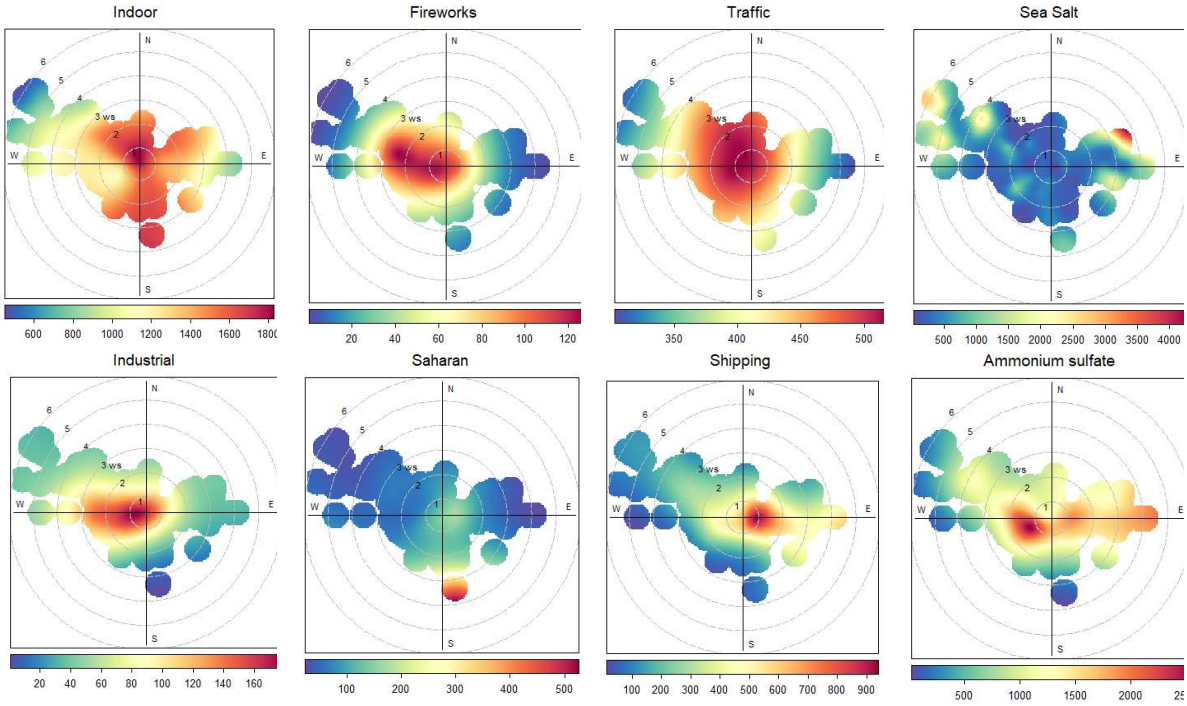


Figure 3.10 Polar plot for the source contributions in ng m^{-3}.

A sudden increase in K concentrations can be used to identify fireworks events(Zhang et al. 2018) due to the prevalence of large amounts of K salts in pyrotechnic compositions. This factor explains 37% of the potassium at the site. During the feast of 13-15 August, Ba concentrations exceed 25 ng m^{-3} for several days, representing a 30-fold increase over the annual median concentration for this metal. In contrast, the roadside median concentration of Ba in Marylebone Road in London, which is mainly derived from brake wear, was recorded at 15.6 ng m^{-3} (Gietl et al. 2010). This observation is significant since no feast was celebrated in Birżebbuġa in mid-August, and it is a clear indication that the metal-laden PM from nearby villages' fireworks displays dispersed substantially to areas close by.

This source contributes 209 ± 769 ng/m^3 to the indoor $PM_{2.5}$ (4%) at the site. A 6% contribution is registered for the summer. This is a substantial contribution, given that this particulate load is not spread evenly throughout the year but occurs at specific peaks in relation to fireworks activities.

Factor 3 is characterised by the elements Al, Mg, Ca, Ti, Fe, Mn and Sr, which are considered markers of Saharan dust (Rodríguez et al. 2002; Viana et al. 2008) and is responsible for 84.5%

of the Al, 39.5% of the Mg, 79.5% of the Ti and 44.3% of the Mn at the site. This factor registered peaks at least five times the normalised contribution on 06 – 08 September, 29 October and 13 – 17 June. HYSPLIT back trajectories were run for all these dates at several heights above ground level (Figure SI.2). The 1500 m trajectory was considered for the analysis of Saharan dust transport as the air mass at this height is in the upper part of the boundary layer, and surface artefacts are thus avoided. Moreover, this height can be considered representative of the mean transport wind at a synoptic scale within the upper boundary layer (Salvador et al. 2007). The high factor contributions were all associated with air masses passing over the North African continent before reaching the receptor site. This is also confirmed by the polar plot for this factor, which shows the highest contributions associated with S to SE winds.

Except for the September 2018 peaks, the other peaks coincided with Saharan episodes notified by Malta to the European Commission in its report for deduction of the contribution of natural sources from PM_{10} levels (ERA 2019; ERA 2020). Saharan dust episodes only resulted in very high indoor concentration when the residence windows were open. Although the contribution of Saharan dust to PM_{10} was calculated as being in excess of 100 $\mu g\ m^{-3}$ on the 29th of October 2018 (ERA 2019), the indoor concentration for $PM_{2.5}$ remained below 10 $\mu g\ m^{-3}$. In contrast, when the windows were open, a PM_{10} Saharan dust contribution of 25-31 $\mu g\ m^{-3}$ in June 2019 (ERA 2020) resulted in an indoor $PM_{2.5}$ concentration of 18 $\mu g\ m^{-3}$. Given that the residence windows were mostly closed throughout the year, this factor only contributes 137 $\pm$ 268 ng/m^3 to the indoor $PM_{2.5}$ (2%). This contrasts with the contributions of Saharan dust to outdoor $PM_{2.5}$ at Msida, which were recorded at 15.1% (Scerri et al. 2018).

Factor 4 is the highest source contributor for EC (45.6%), NO_3^- (47.9%), Zn (40.9%), Fe (55.9%) and As (47.5%). It is also a main contributor to Ca, Cu, and Mn. These elements and compounds are known to be different markers of traffic particulate emissions. Fe, Cu, Zn Cr and Sb are considered tracers for brake-wear (Schauer et al. 2006; Amato et al. 2011), Zn for tyre wear (Amato et al. 2011), OC, EC and NO_3^- are considered tailpipe emissions (Schauer et al. 2006) with As being emitted from tailpipe emissions of diesel engines (Talebi and Abedi 2005). The presence of Ca indicates the contribution of re-suspended dust, given that the geology of outcropping rocks in Malta is dominated by limestone (Cassar and Vella 2003). This factor, therefore, represents both exhaust and non-exhaust vehicular emissions.

OC/EC ratios at urban background sites in Portugal, Spain and the UK were recorded at 2.3 – 4.0 (Pio et al. 2011), with lower ratios (<2) being registered for urban roadside sites. The traffic factor at the receptor site has an OC/EC ratio of 2.0, which is similar to OC/EC ratios for the urban background sites. The Cu/Sb mass concentration ratio for this factor is 5.2. This ratio indicates that the most likely source is brake-wear particles, with this ratio being like that reported for the fine particulate fraction in other studies. (Pant and Harrison 2013) found that the Cu/Sb mass ratio in brake-wear particles depends on the PM size fraction. Several Cu/Sb mass ratios for $PM_{2.5}$ fractions were reported, including 10 (Barcelona), 2.17 ± 0.83 (Taiwan) and 2.5 (Palermo). This reaffirms that this factor also explains the fine PM generated from non-exhaust emissions at this urban background location.

The polar plot in Figure 3.10 shows that the highest concentrations of this factor are at the receptor site, indicating that this factor represents traffic emissions in the vicinity of the site. This factor contributes 571 ± 600 ng/m^3 (10%) to the indoor $PM_{2.5}$. This is substantially lower than that reported by (Scerri et al. 2018) in Msida, where the traffic factor constituted 27.3% of the roadside outdoor $PM_{2.5}$. This is expected, given that the measurements at Birżebbuġa referred to indoor values and the site's characteristics are very different. The highest percentage contribution was recorded in summer, with the traffic factor responsible for 14% of the indoor $PM_{2.5}$. The average contribution by traffic in summer is significantly different ($p<0.01$) from that measured in the other seasons, which can be explained by higher infiltration rates and also higher traffic flows in summer close to the receptor site, given that Birżebbuġa is an important seaside holiday destination during the summer months.

Factor 5 is the highest source contributor for Pb (93.9%) and Cd (34.7%) and contributes 9.1% of EC at the site. The presence of high levels of Pb and Cd in this factor indicates industrial emissions. The polar plot shows a single source located in a W-SW direction from the receptor site. The Hal Far industrial zone is located SW of the receptor site at 1.5 km. Steel manufacturing, hot dip galvanising and hot asphalt plants located in this area could be the sources of these heavy metals and EC. Hot dip galvanising is known for its emissions of Zn, Pb and Cd (National Pollutant Inventory 1999).

Factor 5 is the smallest contributor at the site with 130 ± 187 ng/m^3 (2%) and shows the highest contributions during winter. Nonetheless, this factor indicates that industrial activity has only a slight impact on the indoor $PM_{2.5}$ levels at the receptor site.

Factor 6 is characterised by high Na, Mg and Cl levels. It is the most important source of Na (55%) and chloride (62%) at the site and contributes 522 (9%) ± 578 ng/m^3 of the indoor $PM_{2.5}$ at the receptor site. The Cl/Na ratio of 0.38 is considerably lower than the reported mass ratio for sea spray at 1.8 (Moller 1990; Seinfeld and Pandis 2016). However, the Na/Mg ratio stands at 8.8, nearly identical to the ratio of these elements in seawater at 8.36 (Seinfeld and Pandis 2016). This suggests that these elements are of marine origin, so this factor was labelled as "Sea Salt".

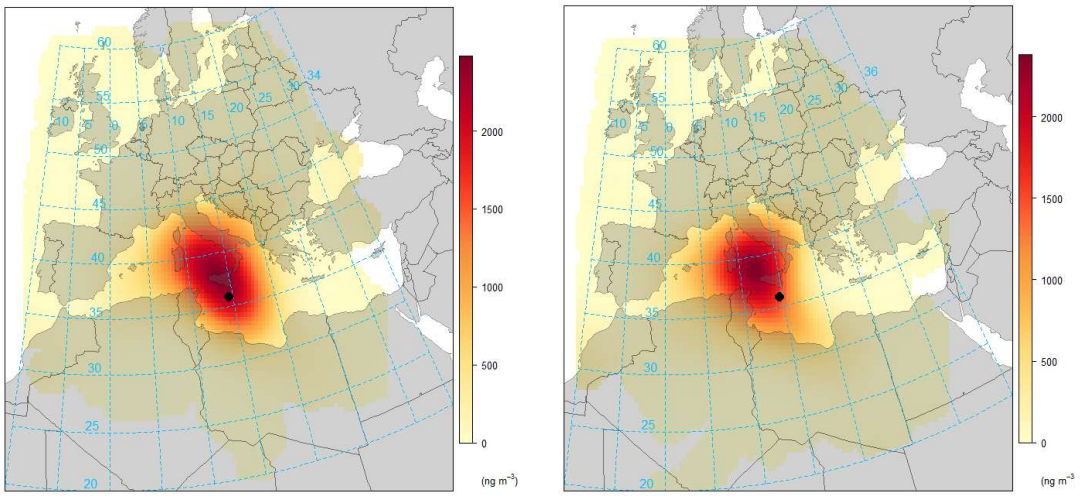
The presence of significant nitrate and sulfate in this factor is coupled with low chloride levels. The deficit of chloride ions can be explained by the volatilisation of chloride when NaCl reacts with acidic species such as HNO_3 and H_2SO_4 (Rodríguez et al. 2002; Galindo et al. 2008). In these chemical reactions, the chloride ions are replaced by nitrate and sulfate ions, forming aged sea salt. Hence, this factor represents aged sea salt. Given that the receptor site is affected by emissions from the nearby harbour and other offshore sources, the fresh sea spray could interact with pollutants such as NO_x and SO_2, producing particles enriched in nitrate and sulfate. This factor is, in fact, responsible for 11% of the nitrate at the site. The polar plot shows that this factor is not particularly dependent on wind direction. High source contributions are recorded with increased wind speed, irrespective of wind direction. This again indicates sea salt as the source, given that sea salt aerosol mass concentration is known to increase with an increase in wind speed (Ovadnevaite et al. 2012). The plots also show that higher source contributions for this factor are associated with NW and E wind directions. Thus, sea salt particles in these air masses react with products of oxidation and hydration of NO_x and SO_2, which are generated both locally from traffic and harbour areas and/or transported to form aged sea salt.

Factor 7 is the most important factor at the site, contributing 1771 ± 1877 ng/m^3 (31%) of the $PM_{2.5}$ and is characterised by high amounts of NH_4^+ and SO_4^{2-} ions. The NH_4^+/ SO_4^{2-} mass ratio for this factor is 0.36, which is nearly identical to the mass ratio of $(NH_4)_2SO_4$ at 0.376, suggesting that ammonium sulfate is the main component in this source, and hence the factor was labelled "ammonium sulfate". Given that there are no known major indoor sources of fine SO_4^{2-} (Rivas et al. 2015), this factor can be considered to be of outdoor origin. The mass contribution of 3.79 μg m^{-3} for the ammonium sulfate factor for indoor air during the summer

at the receptor site is nearly identical to that recorded by Scerri et al. for outdoor air in Msida (3.6 μg m^{-3}), which confirms the outdoor origins of this factor.

As expected, this factor shows a seasonal variation, with the highest contribution registered in summer (43%) and the lowest in winter (15%). The increased photooxidation rates of SO_2 during the summer months explain the increase in particulate sulfate during the summer months (Song et al. 2001; Kim and Hopke 2008). In a study on aerosol sulfates in the western Mediterranean (Galindo et al. 2008), the correlation coefficients between monthly average concentration and the percentages of secondary sulfates in $PM_{2.5}$ to radiation and temperature were 0.92 and 0.88, respectively, with the highest sulfate concentration recorded in summer months.

The difference between the average percentage contribution during different seasons was statistically significant ($p<0.01$) except for autumn and spring. The percentage values are being used, rather than the actual mass contributions, to compensate for any seasonal changes due to changes in infiltration. Given that the oxidation of SO_2 to sulfuric acid is slow (Querol et al. 1998), secondary sulfates are used as indicators of aged air masses (Manousakas et al. 2017; Scerri et al. 2018) and therefore include contributions from long-range transport (Viana et al. 2008). Secondary aerosols originating from local or long-range transport are difficult to separate in factor profiles as they tend to share the same chemical markers (Viana et al. 2008). Figure 3.11 shows that air masses associated with the highest contributions for factor 7 pass across Mt. Etna in Sicily, an active volcano with an estimated SO_2 emission of 1.8 kt/day (Carn et al. 2017). Given that the highest contributions are associated with air masses originating from Italy, it further corroborates this factor's transboundary nature. This factor is also responsible for 34.7% of the Cd and 19.1% of the As at the site, which suggests a transboundary contribution of these toxic elements.



(a) (b)

Figure 3.11 Concentration field for the ammonium sulfate source contribution (in ng m^{-3}) at an arrival height of (a) 500 m and (b) 1000 m.

Factor 8 is the highest source contributor for V (85.8%) and Ni (76.5%) and contributes 540 ± 767 ng/m^3 (10%) of indoor $PM_{2.5}$. Ni and V are found in higher concentrations in heavy fuel oils commonly used by the shipping industry. V/Ni ratios for heavy oil combustion calculated by means of PMF were 3.2 ± 0.8 (Mar Viana et al. 2014) in all PM fractions. For Mediterranean harbours, V/Ni ratios ranging from 2.5 to 5 are markers of shipping emissions (Viana et al. 2009), with non-vessel-derived (land) sources having a V/Ni ratio of 12 (Mar Viana et al. 2014). Given that the V/Ni ratio for this factor is 3.2, it was labelled as "shipping". A similar V/Ni ratio of 2.5 was also recorded for the contribution by shipping to another coastal town in Malta (Scerri et al. 2018).

The polar plot for this factor indicates that the highest source contributions are associated with winds from the easterly direction. This further confirms shipping activities from the Marsaxlokk harbour as the primary source of this factor. The transhipment hub located in this harbour experiences high shipping activity. Since the power station in Delimara is LNG-powered, shipping can be assumed to be the main source of this factor. Moreover, the bunkering area at Hurd's Bank and the waiting area for Marsaxlokk and Valletta harbours are also located around 15 km offshore in an ENE direction from the sampling site.

The actual and percentage source contributions for this factor are higher in summer and spring, suggesting that the higher contributions are due to the increased AERs during the warmer months and increased emissions arriving at the site. This is similar to that reported by (Manousakas et al. 2017) in Greece and (Scerri et al. 2018) in Malta and is possibly due to increased shipping activity in the warmer months.

This factor is responsible for 20.7 % of the sulfate at the receptor site. The high levels of sulfate in this factor can be explained by direct emissions of SO_3 by ships, given that ships burn fuel oil with higher sulfur content and oxidation of SO_2 at high flue temperatures is catalysed by V_2O_5 which deposits on the heat exchangers and other stack surfaces. The emitted SO_3 then shows as sulfates at the receptor site following condensation (Kim and Hopke 2008; Pandolfi et al. 2011).

Many SA studies include an unassigned fraction. In this study, this amounted to 336 ± 1056 ng m^{-3} (6%). The seasonal variation in the unassigned mass was not statistically significant ($p>0.01$). Unassigned mass can be due to structural and absorbed water in the sampled dust remaining during sample conditioning (Yubero et al. 2011) and due to analytical uncertainties.

3.6.4 Comparison with other studies

A limited number of studies have carried out PMF receptor modelling on residential indoor $PM_{2.5}$ (Larson et al. 2004; Minguillón et al. 2012; Barraza et al. 2014; Molnár et al. 2014; Bari et al. 2015; Suryawanshi et al. 2016; Lai et al. 2019). Table 3.8 briefly summarises the SA and the percentage contribution in several indoor PMF studies compared to this study. A PMF study on ambient urban background in another coastal Mediterranean city (D. Cesari et al. 2016) and the only other $PM_{2.5}$ PMF SA carried out in Malta (Scerri et al. 2018) are also included. The sources' absolute (and/or relative) contribution may differ between these studies as these depend on the climate classification, meteorology, topography, urban fabric, building types and ventilation conditions. Although quantitative comparisons can be difficult, certain inferences can be noted.

The residential indoor contribution across these studies ranged from 16.5% to 66.5%, with our study registering the second-lowest indoor contribution at 26%. Studies by Zhao et al. 2006 in the USA and Barraza et al. 2014 in Chile identified cooking as the primary contributor to indoor

$PM_{2.5}$. This study also concludes that cooking is a major source of the indoor factor, and similar to these studies, such profiles included high levels of OC. The other indoor contributions were Environmental Tobacco Smoke and cleaning activities characterised by K and Cl ions. Since no smoking occurred at our receptor site, tobacco smoking was not recorded as a contributor, and this could be one of the reasons for the lower indoor contribution when compared to the indoor sites in the USA and Chile.

All indoor studies recorded a contribution from traffic, although this varies significantly between studies (3.5% to 38.8%). The highest contribution was registered in Santiago (38.8%), given that sampling at this site was conducted in an urban environment. Interestingly the percentage traffic contribution at our receptor site was only slightly lower than that observed in Lecce for an ambient background site (16.5%) but considerably lower than the 27.6% at the ambient traffic site in Msida.

Secondary Inorganic Aerosols (SIAs) are the most important contributor to indoor residential sites (15.5% to 54%), with the Birzebbuġa site registering one of the highest percentage contributions (31%) for this source. This value was nearly identical to ambient $PM_{2.5}$ in Msida (29.7%). This suggests that the air masses in Malta are considerably affected by pollution from long-range/transboundary pollution, given the very similar contributions in other coastal Mediterranean cities, with Lecce (UB) 24.3%, Thessaloniki (UB) 33.0% (Daniela Cesari, Donateo, et al. 2016; Diapouli et al. 2017).

Although in Europe, SO_2 emissions have decreased during the past 25 years, the relatively lenient limits on the sulfur content of marine fuel in the Mediterranean compared with fuels for land transport means that we now expect that under business-as-usual scenarios, emissions from maritime activities will equal those from land-based sources in the EU (Schembari et al. 2012). SO_x emissions from ships in the Mediterranean account for over 50% of all the European sea regions (Jalkanen et al. 2016). Therefore, with the Malta-Sicily channel being one of the busiest sailing corridors of the Mediterranean, with over 30,000 tankers and large ships sailing through this channel annually (Ellul et al. 2013), maritime emissions constitute a significant contributor both directly, with an estimated 10% contribution at the receptor site, and indirectly via particulate sulfate through secondary formation. Ammonia from land-based emissions in mainland Europe reacts with the products of oxidation and hydration of SO_2 emitted from ships,

resulting in increased secondary sulfate concentrations along the shipping routes and coastal areas in the central and western Mediterranean (Aksoyoglu et al. 2016).

Fireworks were detected as a separate source profile in Malta, with a 2.9% contribution for ambient air in Msida and 4.0% for indoor air in Birzebbuġa. The only other PMF SA study which calculated the direct and indirect impacts of fireworks was carried out in a megacity in China during the Chinese New Year (Tian et al. 2014). However, this resulted from a short three-week sampling campaign carried out specifically during the Chinese lantern festival, with fireworks contributing 24.1% of the $PM_{2.5}$ during this period. An indoor PMF source profile in a SA study in Massachusetts, USA, was labelled Fireworks & ETS due to the elevated levels of K, Ba, and Sr in the profile and the higher contributions of this source profile were recorded during the first week of January and July. Some contributions were attributed to plumes from outdoor New Year's and Independence Day fireworks infiltrating indoors. (Matthaios et al. 2021). However, the identification of fireworks as a unique source in long sampling campaigns remains purely a Maltese phenomenon.

Approximately 140 tons of fireworks-related material were estimated to have been burnt in the 2012−2014 period in Malta (Camilleri and Vella 2016). This amount is equivalent to 633 kg per square kilometre per year. In comparison, 93,700 tons of fireworks were burnt in the US during the same period (APA 2024), which amounts to only 9.9 kg per square kilometre per year. The figure in the US is for the entire firework device (including packaging and inert structural material).

These findings confirm that fireworks affect the indoor air quality in Malta, and when the prevalence of pyrotechnic activity is high, the potential for harm to human health cannot be ignored. Even though fireworks contribute 4% of the total $PM_{2.5}$ at the receptor site, they are responsible for most of the Ba (88.9%) and Sb (52.9%). They are also responsible for 35.5% of the Cu measured at the site. Ba causes bronchoconstrictor effects and interferes with the heart and renal functions (Hicks et al. 1986; Kravchenko et al. 2014), and it is suspected to be responsible for the significant increase in asthma during the Diwali festival in India (Murty 2000). Moreover, Ba derived from fireworks is expected to be more bioavailable since it is present as water-soluble compounds (Steinhauser et al. 2008) in contrast with the far less soluble barite from brake pads.

The presence of Sb in indoor $PM_{2.5}$ is also problematic, given that antimony trioxide is classified as a possible carcinogen to humans (Group 2B) (IARC 1989; von Uexküll et al. 2005). Pyrotechnic-derived PM is also associated with the presence of unreacted perchlorate (Vella et al. 2015), given that potassium perchlorate is the primary oxidiser used in pyrotechnic compositions in Malta (Camilleri and Vella 2016). Perchlorate is known to interfere with thyroid function as it inhibits the uptake of iodine into the thyroid (Clewell et al. 2004), although the effect of human chronic exposure to low doses of environmental perchlorate is still unclear and the subject of study (Gold et al. 2013; Pace and Vella 2016).

The fact that pyrotechnic metals are present in the indoor environment is of greater concern as this alone increases the potential for human exposure to these toxic species (L. Yao et al. 2015; Bo et al. 2017). Moreover, one can expect that the highest episodic indoor concentrations are registered during the night, when people sleep, since most fireworks displays in Malta occur between 7 pm and 11 pm. Fireworks-derived particulate matter could be the easiest controllable source of particulates on the islands and would be achieved by properly controlling the quantity and type of burnt fireworks. However, the enthusiasm for fireworks manufacture and display exhibited by the pyrotechnic community and their numerous local supporters is so great that it will be difficult to restrict or regulate the activity. Unfortunately, in recent years, fireworks production in Malta is estimated to have increased by over 170% (Vella et al. 2015).

The WHO Annual Air Quality Guideline (AQG) level of 5 $\mu g\ m^{-3}$ (World Health Organization 2021) for $PM_{2.5}$ applies to both outdoor and indoor environments. The mean indoor $PM_{2.5}$ concentrations at the receptor site are slightly higher (5.7 $\mu g\ m^{-3}$) than the WHO AQG level, while the 24-hour limit of 15 $\mu g\ m^{-3}$ was exceeded twice, with shipping and fireworks being responsible for most of the contributions during these two days. Given the high mean indoor concentrations (8.8 $\mu g\ m^{-3}$) during the summer months, it would appear that reduction of emissions from local fireworks and from shipping in the Mediterranean would lead to an improved indoor air quality in Malta's homes in coastal areas.

Table 3.8 Factors defined by PMF for indoor residential studies. Urban Background sites are listed as (UB)

Receptor	Type	$PM_{2.5}$ ($\mu g\ m^{-3}$)	Source contribution to $PM_{2.5}$ % ($\mu g\ m^{-3}$)								Article
			Indoor	Crustal	Traffic	Shipping*	SIA	Sea Salt	Industrial	Others	
Lecce (Italy)	Ambient (UB)	14.4	-	17.6 (2.5)	16.5 (2.4)	-	29.7 (4.3)	3.6 (0.5)		32.5 (4.7)	(D. Cesari et al. 2016)
Msida (Malta)	Ambient Traffic	15.1	-	15.1 (2.3)	27.6 (4.1)	5.0 (0.8)	23.6 (3.6)	4.6 (0.7)		24.4 (3.7)	(Scerri et al. 2018)
Birzebbuġa (Malta)	Indoor residential	5.7	26 (1.5)	2 (0.1)	10 (0.6)	10 (0.5)	31 (1.8)	9 (0.5)	2 (0.1)	10 (0.5)	**This Study**
Barcelona (Spain)	Indoor residential			5	4		10	2	1	78	(Minguillón et al. 2012)
Gothenburg †† (Sweden)	Indoor (UB)	8.5	16.5 (1.4)		3.5 (0.3)		54 (4.6)	16.5 (1.4)		9.5 (0.8)	(Molnár et al. 2014)
North Carolina (USA)	Indoor residential		66.5 †	3.0	7.6		21.9			1.0	(Zhao et al. 2006)
Santiago (Chile)	Indoor residential		45.7 (10) †		38.8 (8.5)		15.5 (3.4)				(Barraza et al. 2014)

Notes: * Include Heavy Fuel Oil (HFO) combustion

† Includes Second-hand Smoke and personal care. Cooking is the major contributor

†† Calculated as not given directly in the study

Others includes other anthropogenic factors and unassigned fractions. Traffic includes road dust. Crustal includes Saharan

3.7 Conclusion

This study carried out receptor modelling via PMF on indoor $PM_{2.5}$ at a residential location in an urban background site in Malta. Indoor environments have not been extensively characterised via PMF, and few such studies have, to date, been conducted in the Mediterranean region. This study identified eight factors, including seven outdoor sources, namely fireworks (4%), Saharan dust (2%), road traffic (10%), industrial (2%), sea salt (9%), ammonium sulfate (31%) and shipping (10%). An indoor factor was also identified, which contributed to 26% of indoor $PM_{2.5}$. Cooking and e-cigarette smoking were identified as the main contributors to the indoor factor.

Indoor $PM_{2.5}$ at this urban background receptor site is mainly affected by outdoor sources, which generate 68% of indoor $PM_{2.5}$. The contribution by outdoor sources of 7.3 $\mu g\ m^{-3}$ (83%) is much higher in summer in both absolute and relative terms. The decrease to 1.8 $\mu g\ m^{-3}$ (51%) in winter can partially be attributed to lower infiltration as the residence's windows were kept closed during the colder months. The anthropogenic contribution to indoor $PM_{2.5}$ during the summer months is 75%, with SIAs (mainly as ammonium sulfate), traffic, shipping and fireworks being the most important contributors.

Reducing the anthropogenic contribution would assist indoor $PM_{2.5}$ concentrations in falling below the ambient WHO AQG for $PM_{2.5}$. Establishing an emissions-controlled area could reduce pollutant emissions from ships in the Mediterranean region, and in 2025, the Mediterranean Sea will be designated as an Emission Control Area for sulfur oxides, requiring ships to burn fuels with less than 0.1% sulfur content, which is expected to decrease $PM_{2.5}$ generated from maritime emissions via secondary particle formation (Directorate-General for Mobility and Transport 2022 Dec 16). However, transboundary ammonia emissions from mainland Europe must also decrease to reduce particulate ammonium sulfate formation in Maltese residences. Locally, the burning of social (non-essential) fireworks could be readily controlled to decrease exposure to toxic elements in the indoor residential environment, but apparently, the opposite trend appears more likely. This work is the first step in untangling and quantifying the principal sources of indoor fine particulate matter as defined by $PM_{2.5}$. It also presents a better picture of the exposure status to $PM_{2.5}$ of home dwellers in Malta.

Chapter 4

Application of Machine Learning Algorithms in Predicting Indoor Residential $PM_{2.5}$ Concentrations

4.1 Introduction

Residential $PM_{2.5}$ concentrations can be used as a useful substitute for personal exposure, especially for elderly vulnerable populations who tend to spend most of their time inside their homes (Klepeis et al. 2001; Xu et al. 2020), increasing the accuracy of $PM_{2.5}$ exposure and thereby minimizing exposure measurement errors in epidemiological studies. Direct accurate measurement of indoor particulate matter concentrations is labour-intensive and expensive, rendering continuous monitoring impractical. Therefore, it is essential to develop models that can predict indoor household $PM_{2.5}$ concentrations and their spatial and temporal variation without overfitting.

Given the multiple and nonlinear interactions leading to the formation of fine particulate matter (Pan 2018; Dai et al. 2023), data-driven models have proven useful modelling techniques for indoor $PM_{2.5}$ as they rely on historical data to predict future values. Data-driven modelling techniques have been recently adopted to capture the complex relationships with the data due to their ease of use and low computational resources (Yuchi et al. 2019; Du et al. 2021; Gu et al. 2021; Kumar et al. 2022). These approaches can enhance the limited availability of direct indoor $PM_{2.5}$ measurements.

This study aims to use modern ML algorithms, RF and XGBoost models to predict indoor residential $PM_{2.5}$ concentrations at six-hourly averages. The performance of these tree-based models is compared to traditional GLM regression with Lasso regularisation using R^2 (coefficient of determination) and the RMSE (Root Mean Square Error) as performance indicators to determine the model performance. The most important predictor variables to accurately model indoor $PM_{2.5}$ concentration in residences in Malta are also investigated.

4.2 Data handling

The hourly $PM_{2.5}$ concentrations and meteorological data were used to calculate the daily and six-hourly averages using the OpenAir package in R-Studio (Carslaw and Ropkins 2012). The six-hourly averages represented the following time periods: 00 to 06 (night), 06 to 12 (morning), 12 to 18 (afternoon) and 18 to 00 (late evening). The data was split per sampling site, and the "na.interp" function from the forecast package (version 1.5.2) in RStudio (Kuhn 2008) was used to impute the data for the missing values of PM in the time-series data of each

sampling site. This was done to impute the missing values resulting from the periods when the samplers stopped working. The number of instances of imputed data per site is given in Table C.1. Indoor and outdoor PM concentrations identified as significant outliers, due to exceptional events such as construction activities within the residence recorded in activity sheets by occupants, were excluded from the dataset. These values were imputed as per the procedure described above. This was done as the objective of this study is to test the ability of the ML models to predict indoor $PM_{2.5}$ under normal living conditions. In a large enough dataset and with such extraordinary events noted, such conditions might be used as predictor variables. However, due to the limited dataset available, only a few such occurrences were registered in this study.

Six-hourly averages were used to model indoor $PM_{2.5}$ rather than hourly data, as the objective is to model the indoor concentrations using primarily outdoor variables. Using six-hourly averages minimises extreme indoor concentrations, as shown in Tables 4.1 and 4.2, sacrificing time-resolved detail yet still providing four distinct periods that can be associated with different human activities during the day.

Due to a possible time-lagged relationship between indoor and outdoor variables (Yu et al. 2024), and given that the observations were collected over a time period, lagged variables were added as new columns were added to the dataset that accounted for a lag of one, that represents data from the previous six hours sampling period. These lagged variables for wind speed outdoor PM_1 (pm1.L), outdoor $PM_{2.5}$ (pm2.5.L), outdoor PM_{10} (pm10.L) (ws.L), wind direction (wd.L), outdoor temperature (Temp.O.L), indoor temperature (Temp.I.L), outdoor humidity (Humidity.O.L) and indoor humidity (Humidity.I.L) were added. These lagged variables assist the models in capturing temporal dynamics. Time-based features, namely sampling time and day, were also added as predictor variables to aid in capturing temporal dynamics.

The dataset was split into 80% training and 20% testing datasets, ensuring an 80:20 ratio was retained for all the data points from each sampling site. The testing data was set aside to verify the method's accuracy. Ignoring the variability introduced by splitting the training and test datasets can lead to biased model performance; hence, the training and testing datasets were analysed to ensure similar data distribution.

4.3 Prediction Models

4.3.1 Tree-based models

Random Forest (RF) is an ensemble learning method that consists of a collection of decision trees where each tree is constructed using a randomly selected subset of the training data and variables. Combined decisions make the predictions from a sequence of base models. Combining predictions from multiple regression trees reduces the overfitting and improves accuracy. Hence, the final model *g*, for a given input vector *x*, can be represented as the average of the output of the individual decision trees, where *N* is the number of trees and f_i is the output of the i-th tree and is described in equation 4.1.

$$g_{(x)} = \left(\frac{1}{N}\right) * \sum f_{i(x)} \qquad \textbf{Eq. 4.1}$$

RF is an improvement over bagging as the correlation between individual trees is reduced since, during the construction of the multiple decision trees, at each split, a random subset of predictors (mtry) is selected from the total set of predictors (p). A baseline value of mtry is typically approximated to $\sqrt{p}$, although a better method is to optimise a value for mtry using hyperparameter tuning. Each split in the decision tree is then restricted to use only one of these (mtry) predictors. The advantage is that a few predictors are not allowed to dominate the decision trees, thereby making the average of the resulting trees more reliable (James 2013).

The XGBoost algorithm is based on gradient boosting, which consists of a machine learning algorithm that builds a robust predictive model by combining the predictions of multiple weak decision trees. XGBoost is also an ensemble learning method, but each subsequent model corrects the errors made by the previous ones. XGBoost includes the Lasso (Least Absolute Shrinkage and Selection Operator) regularisation term, which limits model complexity and helps prevent overfitting.

XGBoost differs from RF regression in that it is based on gradient boosting whereby the prediction from weak decision trees is built sequentially, with each tree trying to correct the errors made by the previous ones. In RF regression, the prediction from the ensemble of decision trees is combined through averaging. Each tree is trained independently, and the RF model introduces randomness during the construction of individual trees.

A repeated 10-fold cross-validation (CV) was carried out on the training dataset with three repeats. Repeated CV obtain a more reliable estimate of the model's performance since it reduces the effect of variability when randomly splitting the data into K-folds. The model is trained on the K-1 fold and tested on the remaining folds. This process is then repeated ten times. This 10-fold CV is then repeated three times. In each repetition, the folds are randomly shuffled and re-divided. Hyperparameter tuning using grid search was carried out while performing CV. Grid search runs an exhaustive search over specified hyperparameters to optimise the algorithm using the Root Mean Square Error (RMSE) as the evaluation metric, selecting the best combination of values that gave the best RMSE.

For RF regression, hyperparameter tuning for mtry was carried out using grid search. XGBoost was tuned for the maximum number boosting iteration (nrounds), the learning rate at which the model learns patterns in data (eta) with lower values resulting in slower computation, maximum tree depth (max_depth) and regularisation parameter that controls overfitting (gamma). The data was then analysed using the caret package (version 6.0-92) in RStudio (Kuhn 2008).

The model's suitability was evaluated using the R^2 and the RMSE for the concentrations predicted by the model on the testing dataset. The formula for R^2 is given in equation (4.3).

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \textbf{Eq. 4.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \textbf{Eq. 4.3}$$

In equation 4.2 and 4.3, y_i is each observed concentration (μg m^{-3}) of $PM_{2.5}$ in the testing dataset, and $\hat{y}_i$ is the predicted concentration (μg m^{-3}) of $PM_{2.5}$ for the testing dataset and *n* is the number of observations in the testing dataset. The formula for RMSE is shown in equation (4.4).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \textbf{Eq. 4.4}$$

$$IOA = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} = 0 \leq IOA \geq 1 \qquad \textbf{Eq. 4.5}$$

The Index of Agreement (IOA), described for equation 4.5, will also be given as it provides a normalized value of the prediction error and is derived from the mean square error (MSE) ratio to potential error. Although the IOA provides a holistic evaluation of model performance, it is still susceptible to the influence of extreme outlier values, as it relies on squared differences between observed and predicted data and potentially skewing the assessment of model performance (AgriMetSoft 2024).

4.3.2 Generalised Linear Model (GLM)

The same data handling procedures as RF were applied to the same training dataset and the predictor variables were tested using a Generalised Linear Model (GLM) on the training dataset in R-studio.

The GLM is a multiple linear regression model that analyses the relationship between a dependent variable and one or several independent variables. This model can be represented by equation 4.6 where Y represents the dependent variable, the indoor $PM_{2.5}$ concentration, X_i are the independent predictor variables, β_i are the regression coefficients representing the relationship between the dependent variable and each independent predictor variable, and $\boldsymbol{\varepsilon}$ is the residual term.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i + \varepsilon \qquad \textbf{Eq. 4.6}$$

The GLM was run using R-Studio's glmnet package (version 4.1-8) (Friedman et al. 2010). The glmnet package runs an elastic net regularisation method for fitting GLMs and performs variable selection. A tuning grid search was run to optimise the regularisation parameter (lambda) that controls the Lasso regularisation penalties. The Lasso regularisation adds a penalty corresponding to the absolute value of the coefficients to the loss function, enabling the model to shrink some coefficients to zero. In doing so, both variable selection and regularisation are carried out at the same time.

Given that GLM was run with Lasso, the data was standardised (scaled) and for each variable, each value is subtracted by the mean of that variable and then divided by the SD of the variable.

$$z = \frac{value - mean}{SD} \qquad \textbf{Eq. 4.7}$$

The z-scores described in equation 4.7 were calculated for each variable using the training dataset. The testing dataset was scaled using the z-scores of the training dataset. Lasso regularisation imposes penalties on the magnitude of the coefficients. Hence, a variable with a large variance, having a smaller coefficient is penalised less. Standardisation before regression minimises this effect and ensures a more equitable penalisation across all variables. Even though most studies log-transform the data for MLR (Yuchi et al. 2019; Xu et al. 2020), in this study, the data was standardised for GLM to get a fairer Lasso regularisation across the predictor variables.

As with the tree-based algorithms, the model suitability was evaluated using the R^2 and RMSE for the values predicted by the model on the testing dataset.

4.3.3 Predictor Variables

All models were tested using the following predictor variables:

1. Outdoor PM: PM_1 (pm1O), $PM_{2.5}$ (pm2.5O), PM_{10} (pm10O).
2. Indoor and outdoor meteorological parameters: Wind speed (ws), wind direction (wd), outdoor temperature (Temp.O), indoor temperature (Temp.I), outdoor humidity (Humidity.O) and indoor humidity (Humidity.I).
3. Lagged variables (denoted by a ".L"): These are the variables for the previous 6-hr interval. New columns were added to the dataset that accounted for the lag-1 data from the previous six hours.
4. Sampling information and time-based variables: Site (as a categorical variable), sampling time and sampling day.

A detailed list of the 25 predictor variables is found in Table C.2. Due to the limited number of sampling sites, each sampling site was used as a separate categorical variable. On larger datasets, similar sampling sites could be grouped into a single categorical variable, for example,

grouping terraced houses as a categorical variable. Due to the scarcity of sweeping and vacuuming in the activity logs of several sites, these variables were not included in the models.

4.4 Results

4.4.1 Descriptive statistics

Descriptive statistics for the hourly and six-hourly indoor and outdoor $PM_{2.5}$ concentrations are given in Tables 4.1 and 4.2, respectively. Using the six-hourly averages reduces the maximum values recorded for indoor $PM_{2.5,}$ with the greatest decreases recorded for site 7. All sites had lower mean and median indoor $PM_{2.5}$ concentrations except for site 7.

Table 4.1 Descriptive statistics for the hourly indoor (I) and outdoor (O) $PM_{2.5}$ concentrations ($\mu g\ m^{-3}$).

Site	$PM_{2.5}$	n	Mean	SD	Geometric mean	Min	25th	50th	75th	Max
2	I	1272	8.7	9.7	7.0	1.3	4.8	6.6	9.2	134
	O	1272	14	13	10.6	1.7	6.9	11	15	109
3	I	648	15	13	11.9	2.6	7.5	12	19	178
	O	648	17	12	13.9	3.5	8.8	13	22	118
4	I	853	10	13	7.7	1.9	5.0	7.7	11	212
	O	698	10	5.9	8.9	2.5	5.9	8.4	14	34.9
5	I	888	8.9	6.1	7.8	2.5	5.4	7.7	11	87.9
	O	871	11	8.1	9.6	1.9	6.4	9.5	13	89.0
6	I	894	11	13	8.1	1.8	5.3	8.0	11	139
	O	722	12	18	9.0	2.0	6.3	8.3	11	153
7	I	984	55	120	21.5	2.8	9.4	16	43	1285
	O	803	11	5.7	9.4	1.5	6.8	9.2	14	77.3

Table 4.2 Descriptive statistics for the six-hourly indoor (I) and outdoor (O) $PM_{2.5}$ concentrations (μg m^{-3}).

Site	$PM_{2.5}$	n	Mean	SD	Geometric mean	Min	25th	50th	75th	Max
2	I	212	8.7	8.1	7.2	1.6	5.1	6.8	9.4	83.6
	O	212	14	12	11	2.5	7.3	11	15.0	82.4
3	I	108	5	10	12	3.1	7.8	12	19.9	69.8
	O	108	17	9.6	14	3.7	9.5	13	23.0	46.6
4	I	144	11	12	8.4	2.4	5.6	7.8	11.3	95.9
	O	120	10	5.7	9.1	3.1	6.2	8.5	13.5	29.3
5	I	148	8.9	4.9	8.0	3.3	5.7	7.6	10.5	35.2
	O	147	11	7.1	9.8	3.2	6.7	9.5	13.1	43.5
6	I	149	11	12	8.4	2.2	5.8	8.5	11.1	85.6
	O	121	12	18	9.2	2.4	6.6	8.3	11.0	126
7	I	164	55	100	25	3.9	11	18	51.4	627
	O	154	11	5.3	9.6	1.5	6.9	9.4	14.6	29.6

For sites 4 to 7, the outdoor observations were less than the indoor ones due to malfunctions in the outdoor instrument. These observations were imputed as previously described. I/O ratios can be calculated from Table 4.1, showing that sites 2 to 6 have values lower than 1 (0.64 to 0.99), while site 7 has an I/O ratio of 5.2.

The six-hourly data was tested with the Shapiro-Wilk test for normal distribution. The null hypothesis being that the data is normally distributed. The p-value for all the data is significantly lower than the 0.05 acceptance criteria at 95% confidence level; hence, the data is not normally distributed. Figure 4.1 shows the distribution of $PM_{2.5}$ on a $\log_{10}$ scale. While outdoor concentrations show similar distributions at all sites, indoor concentrations at site seven differ from other sites, with the box plot for site seven showing significantly right-skewed data for this site.

4.4.1.1 Kruskal-Wallis and Wilcoxon rank-sum tests

The Kruskal-Wallis test can be used as a non-parametric test for equal distribution between multiple datasets and should be used when the hypothesis of normality is not justified (Reinmann et al. 2008). In this test, the null hypothesis is that the medians of each group are the same, meaning that all groups come from the same distribution. The alternative hypothesis

($p < 005$) was accepted, showing that at least one of the sites has a different median, meaning at least one has a significantly different distribution than the others.

Table 4.3 Kruskal-Wallis test tests for indoor and outdoor $PM_{2.5}$ for sites 2 to 7.

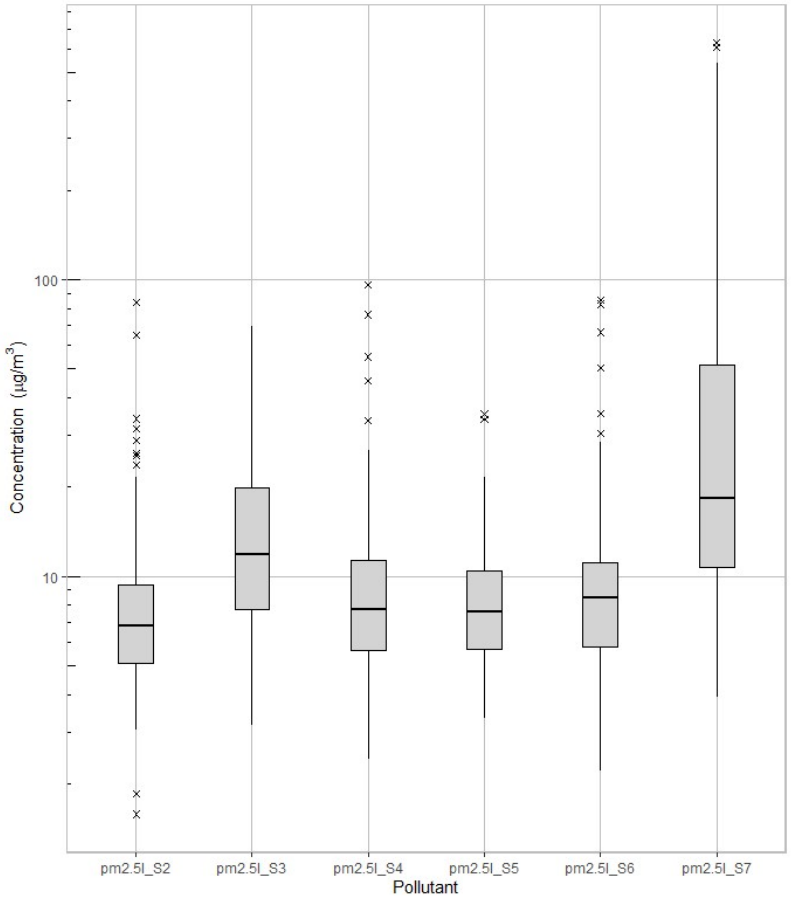
Site	Pollutant	chi-squared	dF	p-value (Kruskal-Wallis test)
2 to 7	Indoor $PM_{2.5}$	201	5	<0.0001
2 to 7	Outdoor $PM_{2.5}$	51.5	5	<0.0001

The Wilcoxon rank-sum tests were carried out for indoor and outdoor $PM_{2.5}$ for the different combinations of sampling site pairs. The Wilcoxon rank-sum test is also a non-parametric test for equal distribution and should be used when the hypothesis of normality is not justified (Reinmann et al. 2008). The null hypothesis tested is that the underlying distributions of the $PM_{2.5}$ concentrations between the two sites are equal and that paired medians do not represent significant differences.

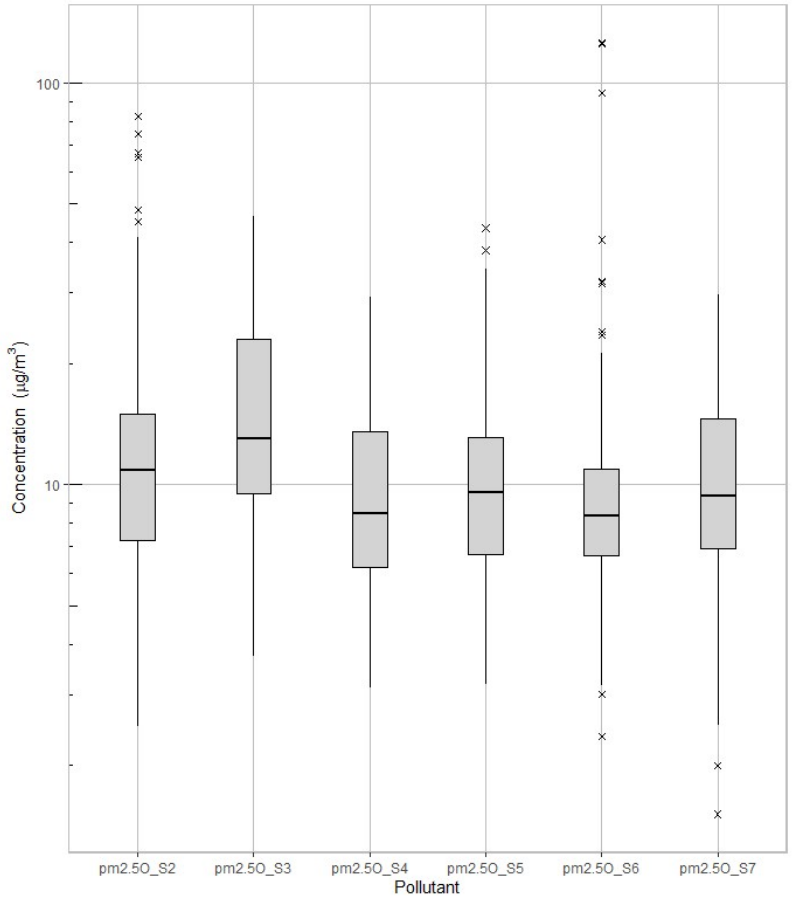
For indoor $PM_{2.5}$, the null hypothesis ($p > 0.05$) was only accepted for site 4 with sites 5 and 6 and between sites 5 and 6, indicating that sites 4, 5 and 6 show similar distributions of indoor $PM_{2.5}$ concentrations.

The null hypothesis for the distribution of outdoor $PM_{2.5}$ was accepted for more sites, indicating similar distribution of outdoor $PM_{2.5}$ at several sites. Site 2 showed similar outdoor $PM_{2.5}$ distributions with sites 5 and 7, site 4 with sites 5, 6 and 7, site 5 with 6 and 7 and site 6 with site 7. Site 3 did not show any similar distributions of outdoor $PM_{2.5}$ concentrations, with site 3 being the only site sampled in summer.

Given the higher similarities found in the distribution of outdoor $PM_{2.5}$ concentrations between sites, indoor characteristics such as cooking, combustion, and cleaning must contribute to the variability of indoor concentrations across the sampling sites.



(a)



(b)

Figure 4.1 Box plots with $\log_{10}$ scale for six-hourly (a) indoor and (b) outdoor $PM_{2.5}$ concentrations. Outliers are shown with an x.

4.4.2 Grid search results

Due to the significant difference in indoor $PM_{2.5}$ concentrations at site 7, the models were first tested with sites 2 to 6 and then run again for all sites. Hence, results for two sets of models will be given, one set for sites 2 to 6 and the other for all sites. Table 4.4 gives the optimal hyperparameter values obtained using grid search on the CV of the three models using RMSE as the evaluation metric.

Table 4.4 Optimal hyperparameter values for the three models

Sites	GLM	RF	XGBoost
2 to 6	Lambda = 0.01	mtry = 17	Nrounds =100 max_depth = 2 eta = 0.1 gamma = 0
All	Lambda = 0.01	mtry = 11	Nrounds =150 max_depth = 2 eta = 0.01 gamma = 0

4.4.3 Results of the Prediction Models

For sites 2 to 6, the GLM model gave the best CV RMSE, marginally better than XGBoost and RF, as shown in Table 4.5, but had a higher RMSE for the training dataset than the testing one. Typically, the testing RMSE is higher as the model is tested on new data, and this warrants investigation, given that training and testing distributions were checked for similar distributions. Hence, the model is not expected to perform worse on the training dataset. Moreover, the Lasso regularisation for the model was also relatively small, which did not significantly penalise the model's complexity. High Lasso regularisation could penalise model complexity, leading to poorly fit training data and a better fit for unseen data.

Given the nature of the data being analysed, where data was being collected over different periods, and since GLMs assume that the data is not a time series, the residuals resulting from fitting the model on the training set were checked for serial correlation (autocorrelation). The residuals from the GLMs were calculated, and an autocorrelation function (ACF) plot was computed on the residuals. A significant spike at lag-1 was registered for both GLM models, indicating first-order autocorrelation, where the current

residual correlates with the previous residual. The Box-Ljung test was also computed at lag-1 on the residuals, giving a p-value <0.01. Hence, the null hypothesis that the data is independently distributed and no autocorrelation exists at lag-1 was rejected. The ACF plots and the Box-Ljung test indicate significant autocorrelation in the residuals, violating the assumption of independence for linear models (i.e. residuals are not correlated with each other).

Even though lagged variables were introduced as predictors to capture temporal dependencies, significant serial correlation was present, violating one of the assumptions of linear models. Hence, the model could be missing key temporal dynamics in the data and should not be used.

Table 4.5 RMSE (in μg m^{-3}) and R^2 for CV, training and testing datasets for sites 2 to 6

	CV		**Training data**			**Testing data**		
	RMSE	R^2	RMSE	IOA	R^2	RMSE	IOA	R^2
GLM	5.05	0.58	4.95	0.86	0.61	3.94	0.92	0.75
RF	5.29	0.55	2.37	0.97	0.94	3.87	0.92	0.77
XGBoost	5.09	0.58	3.90	0.92	0.76	3.91	0.93	0.76

Table 4.6 RMSE (in μg m^{-3}) and R^2 for CV, training and testing datasets for all sites

	CV		**Training data**			**Testing data**		
	RMSE	R^2	RMSE	IOA	R^2	RMSE	IOA	R^2
GLM	29.00	0.28	34.53	0.45	0.18	33.03	0.52	0.27
RF	24.82	0.49	15.35	0.94	0.89	30.65	0.66	0.35
XGBoost	27.92	0.35	29.20	0.65	0.57	34.89	0.38	0.22

Although tree-based models are not inherently designed for time series, they can be adapted through feature engineering, such as lagged variables and time-based features. For sites 2-6, both XGBoost and RF performed well on the testing dataset with near identical RMSE and R^2 and IOAs >0.9. CV RMSE for XGBoost (5.09 μg m^{-3}) was marginally better than RF (5.29 μg m^{-3}). XGBoost performance indicators for the training and testing dataset are identical, indicating that the model does not overfit the data. RF performed significantly better on the training data, with 2.37 μg m^{-3}, 0.94 R^2 and a very high IOA of 0.97. Although the lower RMSE for the training dataset when compared to the CV RMSE for RF could possibly indicate overfitting of the model on the

training dataset, RF provided the best RMSE and R^2 on the unseen test dataset, indicating that the model has a good bias-variance trade-off.

An RF model with the same predictor variables was tested with a time-series cross-validation, having 100 observations in each CV training set. Grid search gave an optimal mtry of 4 with a CV R^2 and RMSE of 0.374 and 5.36 μg m^{-3}, respectively. The lower CV performance indicators compared to the previous RF model can be attributed to the lower sample size in each CV training set. The training set R^2 and RMSE were 0.74 and 4.21 μg m^{-3}, showing that the time-series cross-validation did not improve the performance indicators of the RF model on the training dataset.

The variable importance plots for models run for sites 2-6 in Figure 4.2 show outdoor PM_1 and $PM_{2.5}$ as the topmost important predictor variables for both RF and XGBoost. Lagged variables show higher variable importance in RF, suggesting the model's ability to explain inherent temporal dynamics, which could explain the model's better performance on the training dataset.

The Sites variable shows low importance in both tree-based models, suggesting that specific site characteristics are not particularly significant in determining indoor $PM_{2.5}$ for sites 2 to 6. Both Tree-based models show that outdoor PM_1 is the most important predictor of indoor $PM_{2.5,}$ suggesting that indoor $PM_{2.5}$ is significantly affected by outdoor fine particles.

When the models were tested on data from all sites (Table 4.6), RF showed the best performance indicators for the CV, training and testing datasets, with the CV R^2 being only slightly lower than the value for the RF model for sites 2 to 6 as shown in Table 4.5. The RMSE for these models is expected to be much higher than sites 2 to 6 due to the significantly higher indoor $PM_{2.5}$ concentrations recorded for site 7. The GLM model showed very poor performance indicators. In this GLM, the very low R^2 (0.18) for the testing dataset can be explained by the several observed indoor $PM_{2.5}$ concentrations in excess of 400 μg m^{-3} which had high prediction errors. The testing datasets did not have such extreme values, so the model had a better R^2 for the testing dataset. The XGBoost model also displayed poor performance indicators for the CV and the testing data, with these values being slightly better than the GLM.

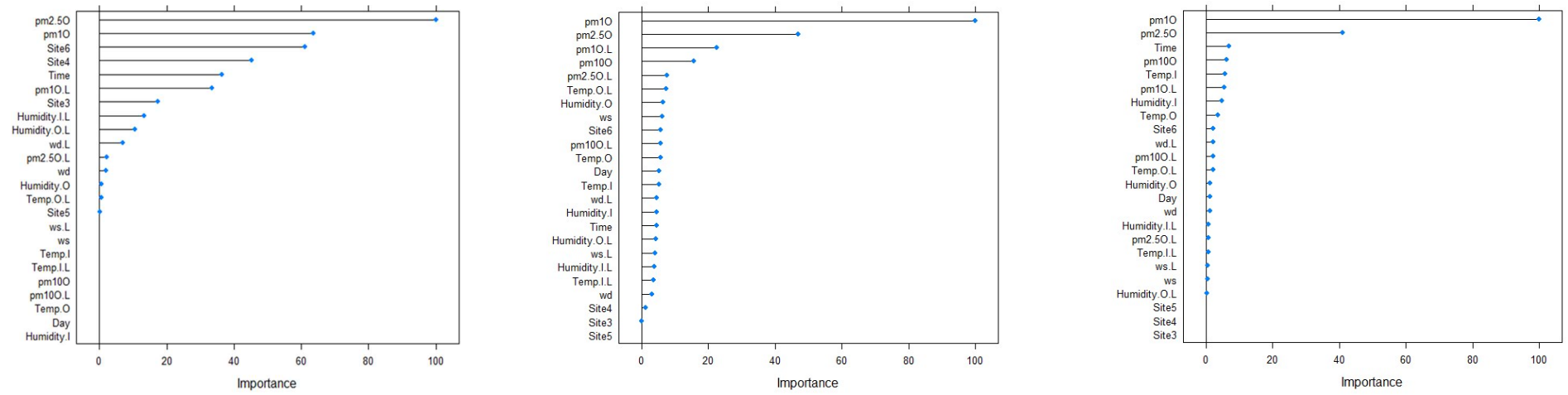
These results indicate that the RF model is more resistant to outliers, showing the highest IOA (0.66) from the models when tested on all sites. The high R^2 and IOA for the training dataset indicate that the RF model can robustly model the data, including outlier values. The low R^2 and high RMSE for testing data for the RF model can be mainly attributed to a single high observed indoor $PM_{2.5}$ concentration with a very high prediction error. When this value is excluded, the R^2 increases to 0.54. This is significant given the limited dataset and the fact that only outdoor PM and indoor and outdoor meteorological variables were used.

Figure 4.6 shows that the RF model has very low prediction errors for sites 2 to 6, with the prediction errors increasing significantly for site 7, thereby explaining the high RMSE values recorded for these models. Table 4.7 gives the descriptive statistics for the predicted indoor $PM_{2.5}$ by the three models and the observed indoor $PM_{2.5}$. While XGBoost shows better predictions at Q1 and Q2, RF gives better predictions at higher concentrations. In fact, GLM and XGBoost fail to properly predict very high indoor $PM_{2.5}$ concentrations.

Figure 4.4 shows the variable importance plots of the models for all sites, with the three models showing notably different important predictor variables. The RF models, which had the best performance indicators, used indoor and outdoor humidity as the most important predictors. In this model, the outdoor PM concentrations and their lagged values are quite low in importance. The RF model used many more variables than GLM and XGBoost. This explains why the RF model had better performance indicators than XGBoost, which only used indoor humidity, site 7 and lagged outdoor temperature as variables with significant importance.

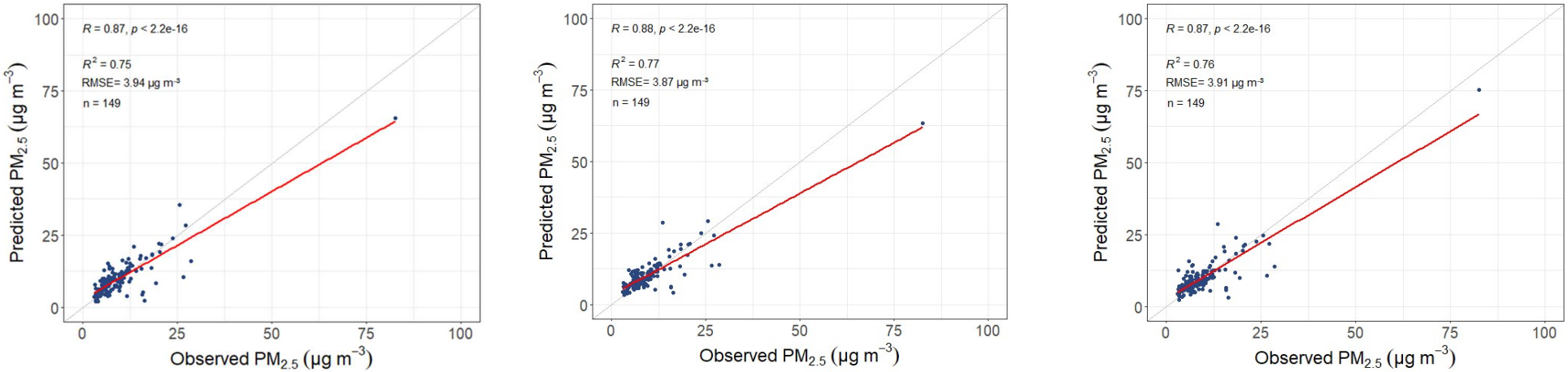
Table 4.7 Indoor $PM_{2.5}$ ($\mu g\ m^{-3}$) model statistics for all sites.

Metric	Observed Indoor $PM_{2.5}$	Predicted $PM_{2.5}$ GLM	Predicted $PM_{2.5}$ RF	Predicted $PM_{2.5}$ XGBoost
Mean	16	16	16	12
SD	38	15	26	13
Q1	6.0	5.5	7.2	6.5
median	8.8	12	10	8.2
Q3	14	21	16	11
maximum	600	66	360	180



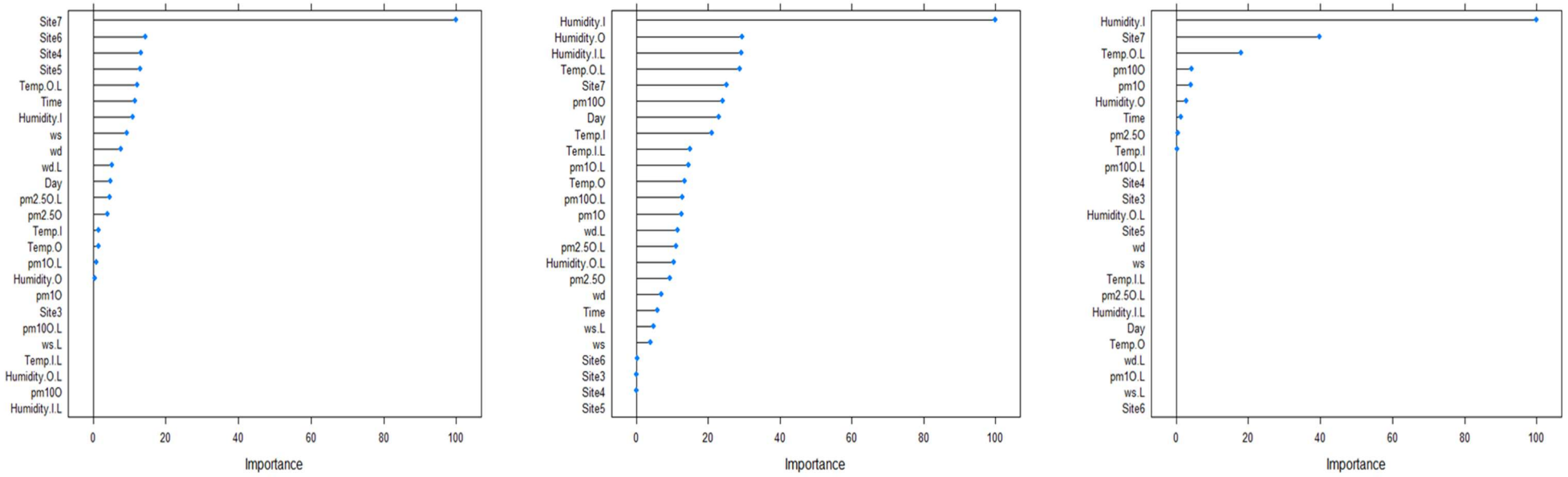
(a) (b) (c)

Figure 4.2 Variable importance plots for (a) GLM (b) RF and (c) XGBoost models for sites 2 to 6



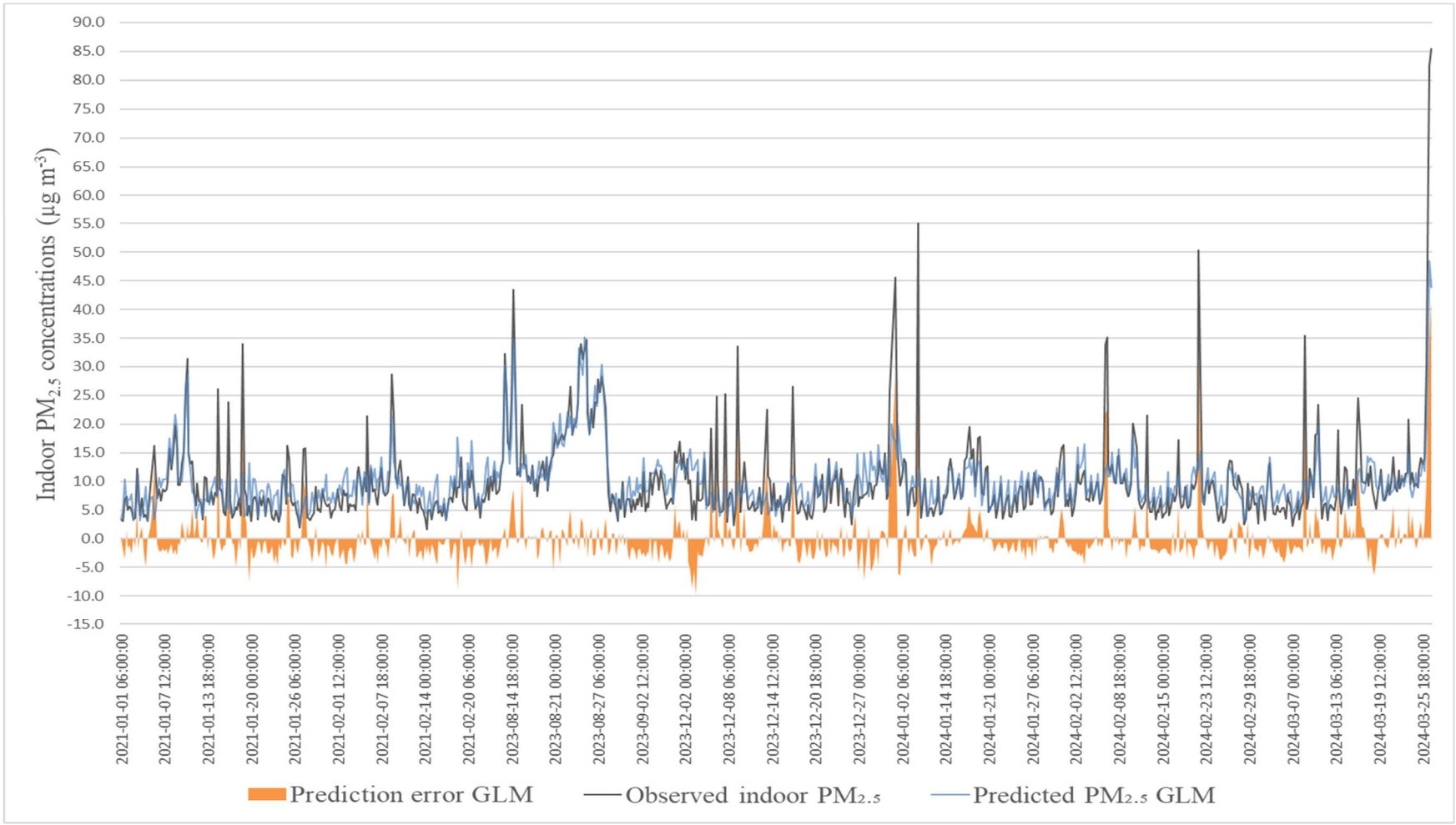
(a) (b) (c)

Figure 4.3 Scatter plots of predicted indoor $PM_{2.5}$ versus observed value for (a) GLM (b) RF and (c) XGBoost on the test dataset. The red line shows a linear regression between predicted and observed $PM_{2.5}$ values for the test dataset. The grey line represents y=x.

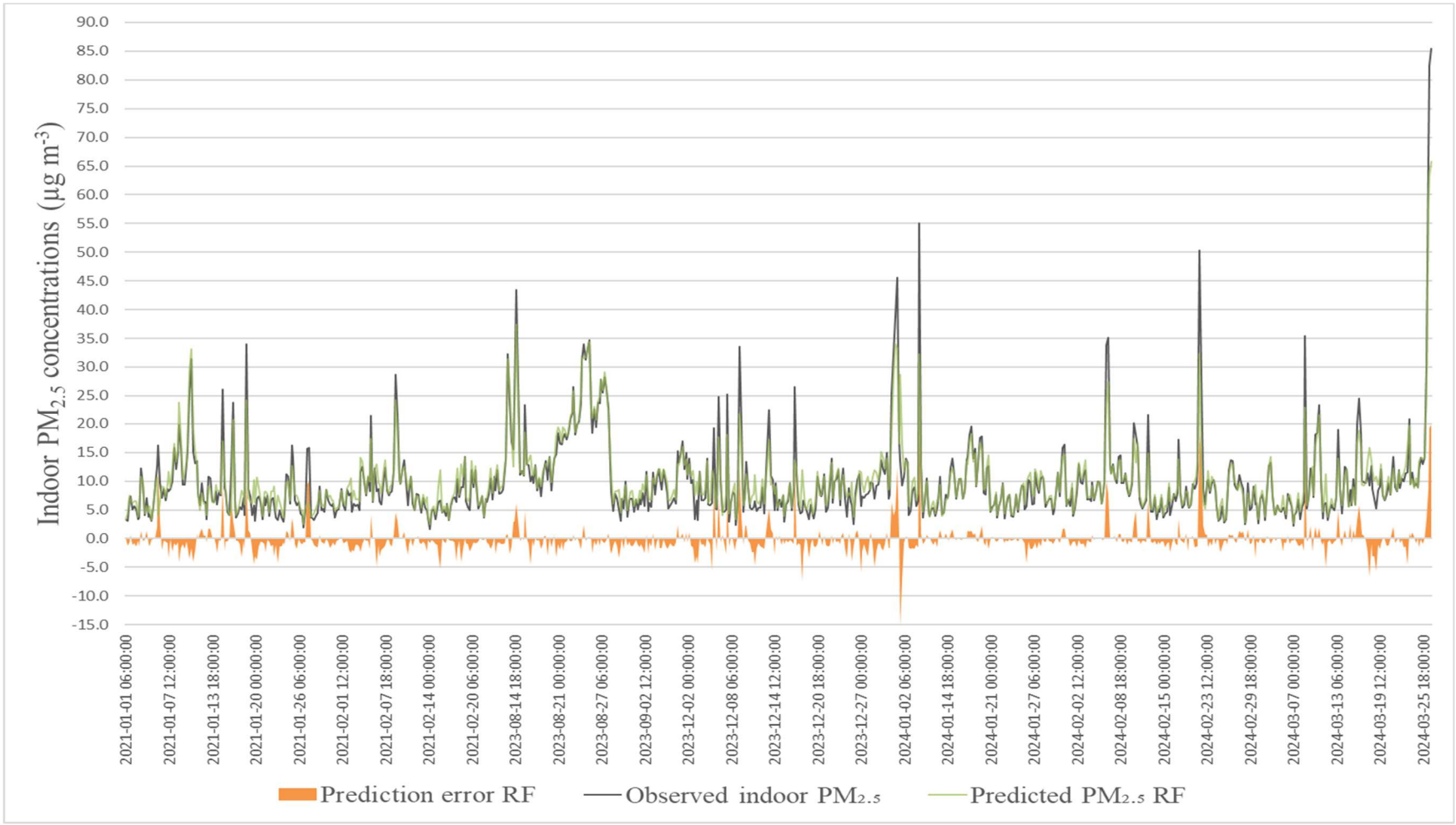


(a) (b) (c)

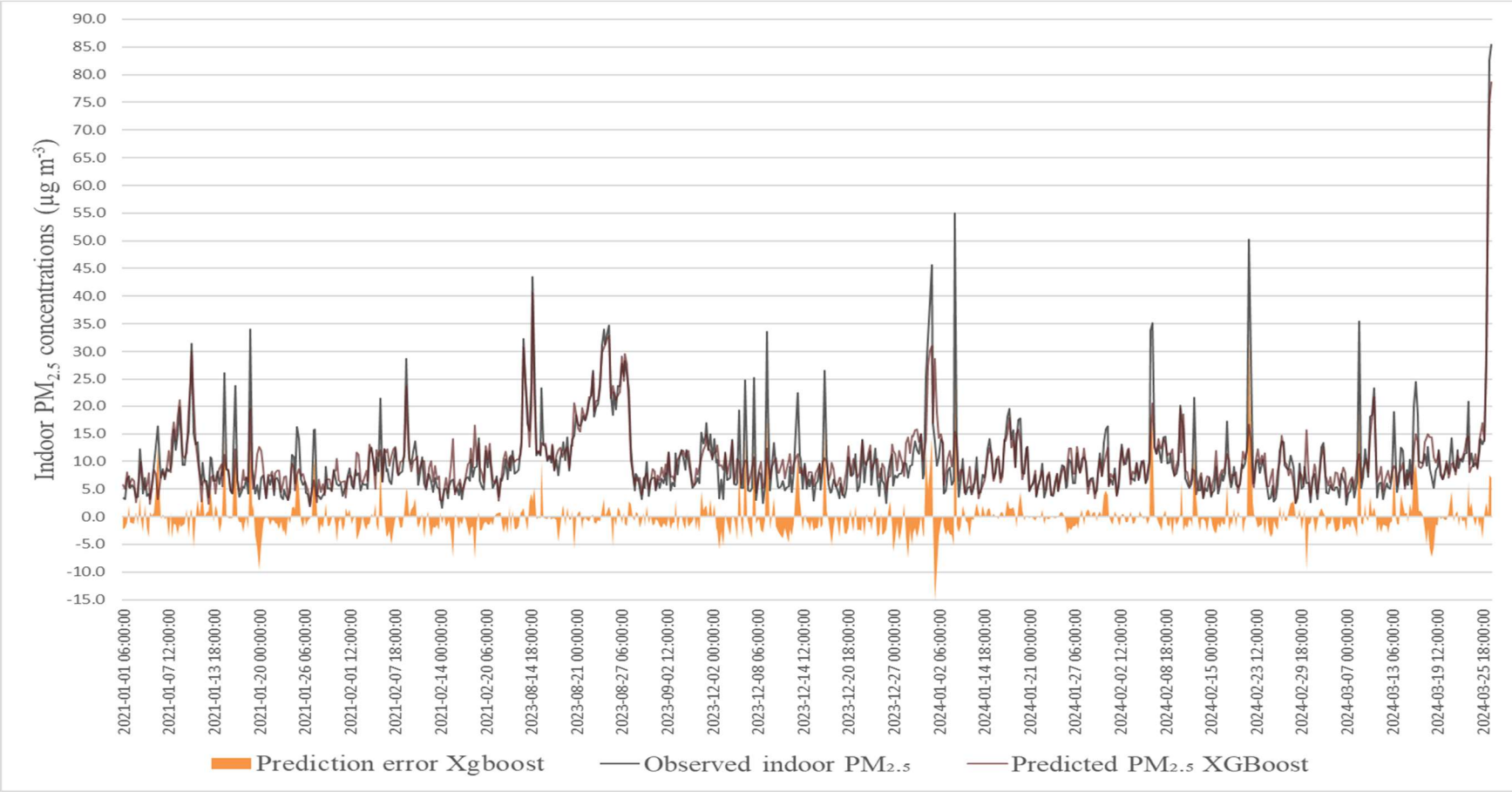
Figure 4.4 Variable importance plots for (a) GLM (b) RF and (c) XGBoost models for all sites


Indoor $PM_{2.5}$ concentrations ($\mu g\ m^{-3}$)
90.0
85.0
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
2021-01-01 06:00:00
2021-01-07 12:00:00
2021-01-13 18:00:00
2021-01-20 00:00:00
2021-01-26 06:00:00
2021-02-01 12:00:00
2021-02-07 18:00:00
2021-02-14 00:00:00
2021-02-20 06:00:00
2023-08-14 18:00:00
2023-08-21 00:00:00
2023-08-27 06:00:00
2023-09-02 12:00:00
2023-12-02 00:00:00
2023-12-08 06:00:00
2023-12-14 12:00:00
2023-12-20 18:00:00
2023-12-27 00:00:00
2024-01-02 06:00:00
2024-01-14 18:00:00
2024-01-21 00:00:00
2024-01-27 06:00:00
2024-02-02 12:00:00
2024-02-08 18:00:00
2024-02-15 00:00:00
2024-02-23 12:00:00
2024-02-29 18:00:00
2024-03-07 00:00:00
2024-03-13 06:00:00
2024-03-19 12:00:00
2024-03-25 18:00:00
Prediction error GLM
Observed indoor $PM_{2.5}$
Predicted $PM_{2.5}$ GLM

(a)


Indoor PM2.5 concentrations (µg m-3)
90.0
85.0
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
2021-01-01 06:00:00
2021-01-07 12:00:00
2021-01-13 18:00:00
2021-01-20 00:00:00
2021-01-26 06:00:00
2021-02-01 12:00:00
2021-02-07 18:00:00
2021-02-14 00:00:00
2021-02-20 06:00:00
2023-08-14 18:00:00
2023-08-21 00:00:00
2023-08-27 06:00:00
2023-09-02 12:00:00
2023-12-02 00:00:00
2023-12-08 06:00:00
2023-12-14 12:00:00
2023-12-20 18:00:00
2023-12-27 00:00:00
2024-01-02 06:00:00
2024-01-14 18:00:00
2024-01-21 00:00:00
2024-01-27 06:00:00
2024-02-02 12:00:00
2024-02-08 18:00:00
2024-02-15 00:00:00
2024-02-23 12:00:00
2024-02-29 18:00:00
2024-03-07 00:00:00
2024-03-13 06:00:00
2024-03-19 12:00:00
2024-03-25 18:00:00
Prediction error RF
Observed indoor PM2.5
Predicted PM2.5 RF

(b)



(c)

Figure 4.5 Time series plots showing model prediction $PM_{2.5}$ concentration, observation $PM_{2.5}$ concentration and prediction error for (a) GLM (b) RF (c) XGBoost for sites 2 to 6.

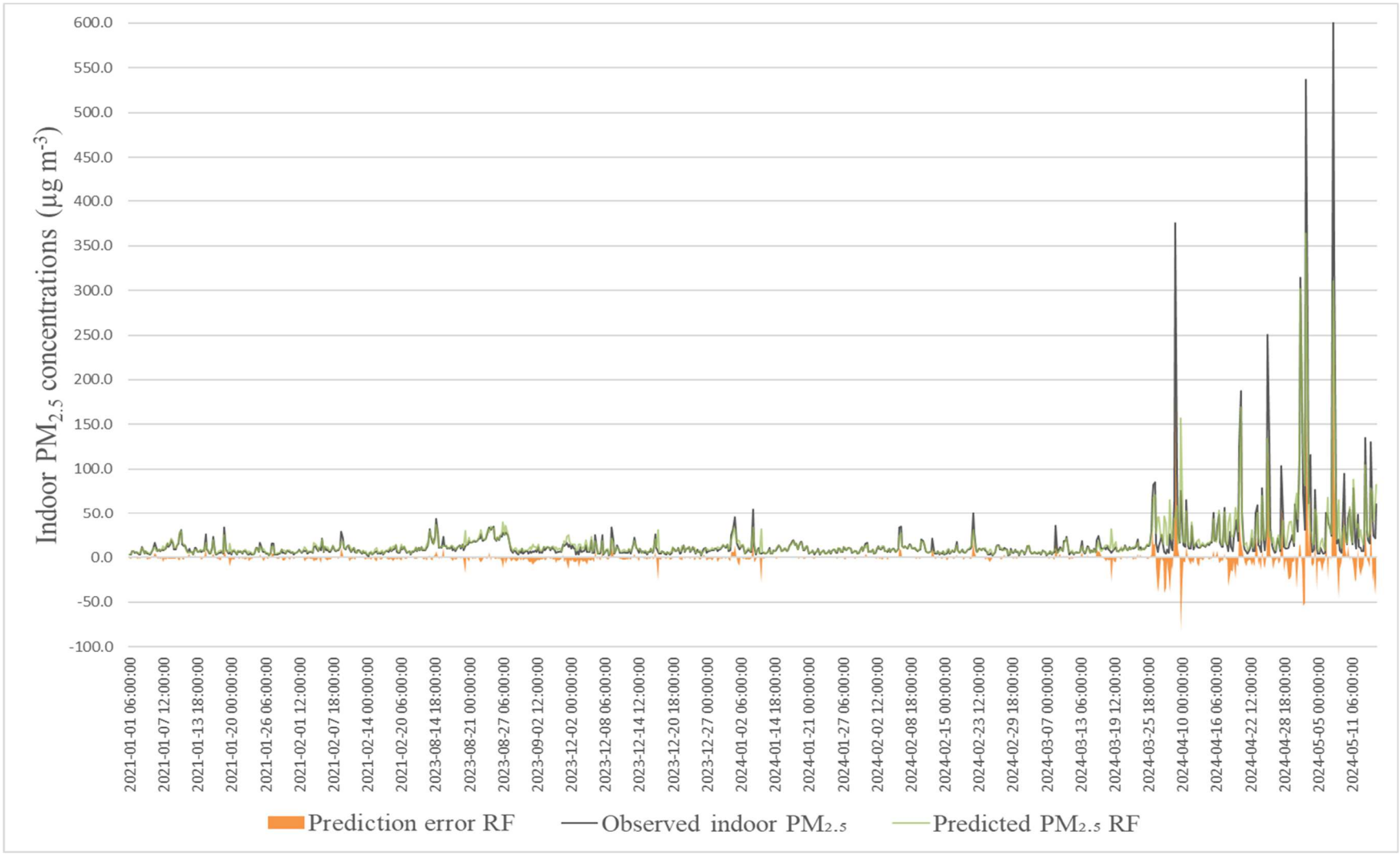


Figure 4.6 Time series plots showing model prediction $PM_{2.5}$ concentration, observation $PM_{2.5}$ concentration and prediction error for RF for sites all sites.

4.5 Discussion

Even though the accuracy indicators, RMSE and R^2, for the tree-based models are only marginally better than those for the GLM for sites 2 to 6, the presence of autocorrelation in the GLM model prevents the use of this model in predicting indoor $PM_{2.5}$ concentrations. Figure 4.5b clearly shows that the RF model for sites 2 to 6 shows the lowest prediction error when the model is used to predict the indoor $PM_{2.5}$ for these sites. High prediction errors are recorded only when indoor concentrations exceed 30 $\mu g\ m^{-3}$. The most important predictor was outdoor PM_1, with lagged outdoor PM concentrations and outdoor temperature also showing high variable importance in the RF model. The RF models suggest a greater impact of finer outdoor particles on indoor $PM_{2.5,}$ which can be explained by the higher infiltration of PM_1 compared to $PM_{2.5}$ due to its smaller aerodynamic diameters (Fromme 2012). Very fine particles are typically generated from combustion emissions and the condensation of secondary sulfates, nitrates, and organic compounds from the gaseous phase (Seinfeld and Pandis 2016). Therefore, outdoor sources, such as exhaust emissions and particles formed through secondary formation, may impact indoor PM more than other sources at these sites and keeping in mind that PM_1 has greater lung penetration efficiency (Brown et al. 2013), a greater health impact would probably be envisaged.

When considering all sites, the RF model is significantly better than GLM and XGBoost. Site 7 had a significant indoor generation, and this is shown in Table 4.1, with the indoor hourly median concentration 1.7 times the outdoor one. Unlike the other sampling sites, site 7 was a third-floor apartment with no external hood extraction on the cooking hobs. The peaks in indoor $PM_{2.5}$ were recorded during the evening, which coincides with the time when cooking was carried out at this residence and pan frying and searing were frequently used. These cooking techniques are known to generate the highest amount of fine particles (Wallace et al. 2003) and can increase indoor fine PM by over ten times the average concentration (Zhao et al. 2014). The absence of external hood extraction exacerbated this problem as hood extraction is known to decrease indoor $PM_{2.5}$ generated from cooking (Kang et al. 2019). The residents only noted it on the activity log when visible smoke was seen inside the house while cooking. Three data points corresponding to these events in the evening period were removed from the dataset.

Nonetheless, in site 7, 29% of the sampling days resulted in values exceeding 95.9 μg m^3, the maximum concentration recorded at the other sampling sites. In total, 16 six-hourly averages from 164 showed values higher than 95.9 μg m^{-3}, eight of which were during the evening period and four during the night. The hourly data also shows that peaks exceeded 1000 μg m^{-3} several times, and these $PM_{2.5}$ levels required several hours (between 4 and 12 hours) to return below the mean indoor concentration.

An I/O ratio of 5.19 was recorded for site 7, confirming that most of the $PM_{2.5}$ is of indoor origin. The data indicates that the high mean indoor $PM_{2.5}$ concentrations at site 7 result from short-lived episodes that substantially increase the indoor fine PM concentrations. This is mostly due to indoor particle formation related to cooking, with the lack of external hood extraction delaying the extraction of these particles to the outside and, hence, the rapid decay of these fine particles' concentration. Besides the high levels of fine PM, which are known to be harmful to human health (World Health Organization 2021), frying increases the amount of combustion-related emissions in the indoor environment, particularly well-known carcinogenic pollutants such as particle-phase polycyclic aromatic hydrocarbons (Z. Yao et al. 2015), which would exacerbate health problems, both in the short and long term.

The RF model for all sites used indoor RH as the most important variable, and the model's high R^2 on the training dataset suggests that indoor RH is being used as a good predictor variable for the very high indoor concentrations recorded at site 7. Indoor RH has been documented to increase during and after various cooking processes, especially when exhaust hoods are turned off (Zhao et al. 2014). Further sampling campaigns should include a detailed log on the time and type of cooking to investigate the use of indoor RH as a predictor variable for indoor generated fine PM from cooking.

Unfortunately, site 7 was the only apartment in a multi-storey building sampled in this study, and hence, the effect of other apartments in the multi-storey building cannot be assessed. Floor level has been reported to be an influencing factor in indoor air quality (Xu et al. 2020), and further campaigns should include a representative sample of sites from apartments at different floor levels.

Several studies have used machine learning algorithms to predict air quality, and many have shown that RF is superior to MLR, giving higher performance indicators. Such results are

expected when considering that several predictors, such as temperature and relative humidity, are not expected to have a linear relationship with indoor $PM_{2.5}$. This study demonstrates this by the variable importance plots shown in Figure 4.4. While both tree-based models had indoor RH as the most important variable, the GLM model shows the categorical variable site 7 as the most significant variable, with the GLM performing poorly compared to RF.

The Relationship of Indoor, Outdoor and Personal Air (RIOPA) study used RF models to predict elemental concentrations in personal $PM_{2.5}$ based on all outdoor and indoor elemental concentrations (Ryan et al. 2015). The study involved the continuous 48-h sampling of several participants in the USA. R^2 was used as a performance indicator, and the study reported better performance by RF regression models than by linear models, attributing this to the ability of the tree-based model to account for complex and non-linear interactions. Home characteristics were reported to not significantly improve the RF model's predictions of concentrations in personal $PM_{2.5,}$ although this has to be viewed in the context that the model was predicting 48-hour averaged elemental concentrations. These long averages dilute emissions from specific sources, reducing the importance of home characteristics as a predictor variable.

Indoor $PM_{2.5}$ were predicted in two megacities in China using MLR and RF regression (Xu et al. 2020), CV R^2 of 0.78 and RMSE of 24.3 μg m^{-3} were reported for the RF regression model with poorer performance from the MLR models. This study used outdoor $PM_{2.5}$, meteorological parameters and site information to predict indoor $PM_{2.5}$. Outdoor $PM_{2.5}$ and lagged outdoor $PM_{2.5}$ were the most important variables, although time of day, distance to the nearest road, floors and window opening were also recorded as important predictor variables.

In a study predicting indoor fine particulate matter concentrations in a highly polluted city in Mongolia, Yuchi et al. 2019 reported a RF R^2 of 0.938, which is comparable to the 0.94 in this study for the training dataset, and a 0.22 μg m^{-3} RMSE, but the CV R^2 fell to 0.489 for a 10-fold CV with the RMSE increasing to 0.51 μg m^{-3}. A total of 22 predictor variables were used, including outdoor $PM_{2.5}$ concentrations, meteorological parameters, house characteristics, residents' behaviours and spatial variables (elevation and land use). In this study, MLR with Lasso and RF gave similar performance indicators. It is important to note several differences in this study. This study relied on estimates from low-cost sensors and used two air-quality monitoring sites to provide the outdoor concentrations. More importantly, this study predicted indoor weekly $PM_{2.5}$ concentrations. With extended time-averaged concentrations, any peaks

resulting from short-lived indoor emissions are lost; hence, time-resolved information about influencing variables is not possible.

A study predicting hourly indoor fine PM across 24 distinct buildings in Australia (Yu et al. 2024) reported RMSE of as low as 0.1 μg m^{-3} and high R^2 >0.9 for both RF and XGBoost on unseen data. This study used low-cost sensors and a very large dataset and included indoor PM_{10} and lagged $PM_{2.5}$ as predictor variables, which were the most important variables for both models. These indoor PM variables explain the high accuracy indicators reported in the study by Yu and colleagues.

In another study in Rural North China (Men et al. 2023), an RF model was used to assess the impact of household energy on indoor $PM_{2.5}$. The study reported a very high average indoor $PM_{2.5}$ concentration of 120 μg m^{-3}. The study included 1602 rural households, and indoor and outdoor PM concentrations were measured using extremely cheap, low-cost sensors. Such a large sampling group would not have been possible without cheap sensors; nonetheless, using such sensors raises doubts about the accuracy of the PM concentrations. PM concentrations from low-cost sensors can report concentrations significantly lower than high-end equipment used in air quality stations, and the variations depend on meteorology, aerosol chemical and size characteristics (Mukherjee et al. 2017). In this study, the RF model (R^2 of 0.85, RMSE 64 μg m^{-3}) outperformed the GLM (R^2 of 0.21, RMSE 148 μg m^{-3}). Although the RF model significantly outperformed the GLM model, this study had several limitations. The CV and testing dataset accuracy indicators were not included, and the GLM model was trained on data that was not log-transformed or scaled. Lasso or ridge regularisations were not used to optimise the linear model.

Compared to the available literature, this study differed by sampling only a relatively small number of residences, albeit for a relatively long time at each site, with a combined dataset of 919 six-hourly averaged observations. All other studies had larger datasets; hence, better results are expected as data-driven models require large amounts of data to explain the relationships between variables. These large datasets were only possible due to the use of low-cost sensors. Unlike all other studies, this study used calibrated optical spectrometers used in air quality stations and certified according to EN 16450 and EN 15267. Hence, indoor and outdoor PM concentrations recorded in this study are known to be reliable. Apparently, this is the only study that has used outdoor PM_1 as a predictor variable and in which concurrent

outdoor PM_1 was the most important predictor variable for sites 2 to 6, indicating that indoor fine PM is predominantly of outdoor origin at these sites. This study reaffirmed the importance of lagged variables in predicting indoor $PM_{2.5,}$ with both RF models indicating that lagged variables are important predictor variables.

4.6 Conclusion

This study reported low RMSE and high R^2 for the testing data for sites 2-6 for both RF and XGBoost, with GLM also showing good performance indicators for testing data, but serial correlation at lag-1 was recorded. When the models were tested with the data from all sampling sites, RF gave the lowest prediction error consonant with results from similar studies (Xu et al. 2020; Men et al. 2023) and with an IOA of 0.66. This RF model managed to predict indoor $PM_{2.5}$ from sites that showed that their indoor $PM_{2.5}$ had different distributions that were statistically significant. This could be partly due to hyperparameter tuning, which was not always present in other studies.

In sites 2 to 6, the prediction of indoor $PM_{2.5}$ is mostly dependent on the outdoor PM_1 fraction. For site 7, indoor-generated $PM_{2.5}$ was the main contributor, and high concentrations were most likely due to cooking (frying and searing) in the absence of external hood extraction. For this site, the RF model used indoor RH as a good predictor variable for the very high indoor $PM_{2.5}$ concentrations recorded at this site. For site 7, the indoor $PM_{2.5}$ concentration was 3.6 times higher than the World Health Organisation short-term (24-hour) Air Quality Guideline level of 15 $\mu g\ m^{-3}$ (World Health Organization 2021), which was exceeded during 83% of the sampling days.

There are several limitations to this study. Using a six-hour rather than hourly average limits the ability of the models to analyse the variables' effects on the intraday variation of indoor $PM_{2.5}$. The time variable, which represented the four six-hour averages during the day, was not predominant in both RF models, showing limited intraday variability in indoor $PM_{2.5}$ concentration. However, this can be partly attributed to averaging the hourly PM concentrations, which might have reduced the apparent intraday variability.

Although the dataset included nearly a thousand individual observations, these observations were obtained from only six residences. The six residences were spatially distributed around Malta and Gozo but are not a representative sample of the local distribution of residential buildings in Malta, making it difficult to extrapolate these results to people living across different household characteristics. More data collected from flats/apartments/penthouses on different floors in multi-storey buildings, which have become prevalent in Malta (National Statistics Office 2023), would be needed for a comprehensive assessment. The repeated measurement from each residence resulted in serial correlations in the GLM model. In this scenario, Time Series models that account for autocorrelation should be investigated.

A shorter sampling period per site, not exceeding ten days, should be considered in further studies as it was challenging to find participants, especially when using bulky and noisy equipment. The shorter sampling period should also make it easier for participants to take detailed logs of indoor activities.

Chapter 5

Application of Interpretable Machine Learning: An RF-SHAP Analysis

5.1 Introduction

Data-driven modelling techniques have been recently adopted to capture the complex relationships with the data due to their ease of use and low computational resources (Yuchi et al. 2019; Gu et al. 2021; Tang et al. 2021; Kumar et al. 2022). The data-driven algorithms are not limited to linear relationships between the predictor variables and the response variable and may be better suited for atmospheric interactions given the complex and nonlinear relationships present. Tree-based ML algorithms, such as RF models, are trained on the data via an ensemble of decision trees that can handle nonlinear relationships between the predictor variable and target variables. One disadvantage of machine learning algorithms, such as RF, is that they offer less interpretability when compared to MLR. However, to reduce $PM_{2.5}$ levels effectively, it is necessary to evaluate and quantify the contributing sources, but there is no consistent approach or metric to measure the importance of different factors across various data-driven models.

For instance, in MLR models, the coefficients of the selected predictors are considered important and indicate the percentage of explained variability. In RF models, variable importance is measured by the variation in model performance when a predictor variable is randomly shuffled. While these approaches provide insights into variable importance within each model, they do not quantify the fractional contribution of each variable to the predicted concentration. This lack of consistency makes it difficult to compare models in terms of interpretability.

This is where SHapley Additive exPlanations (SHAP) values (Lundberg and Lee 2017) can be employed. SHAP values provide a way to interpret the predictions made by the RF model. Unlike the variable importance metric provided by the RF model, SHAP values measure the contribution of each input variable to the final prediction and can be used to identify the most important variables that influence the predictor variable. Visualizing SHAP values can give insights into the relationship between the predictor variables and response variable and, hence help understanding the direction and magnitude of their effects.

In this context, using RF and SHAP values on $PM_{2.5}$ datasets can provide a powerful tool for both accurate predictions and interpretation of the underlying factors that influence $PM_{2.5}$ levels (Gu et al. 2021; Hou et al. 2022; Zhang et al. 2022; T. Li et al. 2023) Such methods can be

applied to indoor $PM_{2.5}$ (Men et al. 2023) and can aid in modelling and understanding sources to indoor $PM_{2.5}$ which can be used to design effective air quality management strategies for the protection of public health against adverse effects of air pollution.

5.1.1 Applications of Interpretable ML

Two applications of interpretable machine learning will be tested:

1. The first application of interpretable machine learning being tested is the combination of the PMF analysis, which offers a mass balance model to determine the sources at the specific receptor site, and RF regression. Given that the PMF model solves a mass balance equation, other important variables cannot be included in the model, and hence, the qualitative and quantitative effects of meteorological variables on the sources at the site cannot be analysed directly using PMF (Zhang et al. 2022). To this end, a RF model, in conjunction with SHapley Additive exPlanations analysis (RF-SHAP), using sources identified via the indoor PMF SA and meteorological data as predictors for indoor $PM_{2.5}$ is used to determine the impacts of the PMF sources and meteorological factors on indoor $PM_{2.5}$ at the Birżebbuġa sampling site.

2. The second application of RF-SHAP is used to quantify the sources of outdoor $PM_{2.5}$ at the receptor site in Birżebbuġa. Quantifying outdoor $PM_{2.5}$ sources assists and validates the interpretation of indoor source apportionment results and is included in several SA studies on indoor PM (Barraza et al. 2014; Saraga et al. 2023; Wang et al. 2024). Therefore, this analysis aims to run a PMF analysis on outdoor $PM_{2.5}$ at the Birżebbuġa site to understand better and supplement the information obtained from the indoor PMF analysis.

   Due to a limited number of PM samplers available and limited funding, only the elemental content of the outdoor $PM_{2.5}$ samples, besides the gravimetric analysis, could be tested. On average, the outdoor elemental content analysed via ICP-MS accounted for only 8% of the gravimetric $PM_{2.5}$ mass. Although metals are good receptor species for the PMF solution due to their limited chemical or phase changes from emission to the receptor site (Belis et al. 2013), they only typically constitute a minor fraction of the $PM_{2.5}$ mass. The combined OM, EC and ionic (sulfate, nitrate, chloride and

ammonium ions) content in $PM_{2.5}$ exceeds the elemental content by an order of magnitude (R.E Hester et al. 2016; Hopke et al. 2020). Running a PMF solution with only the elemental content is limiting and leads to incomplete source characterisation where the source contribution estimates (SCEs) from the receptor modelling may be underestimated or overestimated. Major sources, such as SIAs, would be excluded, and other known local sources with high OC and EC content, such as exhaust emissions, will be underestimated, while other sources, such as sea salt, Saharan dust and fireworks, might be overestimated. Hence, the indoor EC, OC and ionic content will be used to estimate outdoor concentrations, and a PMF analysis combined with an RF-SHAP analysis will be used to quantify the sources of outdoor $PM_{2.5}$ at the receptor site.

5.1.2 Using SHAP Values

Based on game theory, SHAP values can be used to explain and interpret ML algorithms and provide a fair way to distribute model output among the predictor variables. SHAP values explain the contribution of different predictor variables by attributing the difference between the actual prediction and the baseline value to individual variables. The baseline value is given by the mean prediction of the model in the dataset. Typically, this is the mean of the training dataset, but since the RF-SHAP is being used to explain the influence of variables on $PM_{2.5}$ in this application, it is being run on the entire dataset. Hence, in this analysis, the baseline value (ϕ_0), which is shown as *E[f(x)]* in the plots, is the mean prediction of the whole dataset.

This baseline value (ϕ_0) is used as an initial point for interpreting individual predictions. The baseline value calculated by the kernelshap package (version 0.4.1) in RStudio is not necessarily an equal representation of all features in the model; however, ϕ_0 represents the expected value of the model's output, averaged over the dataset where it serves as the baseline prediction of the model when none of the variables is taken into account explicitly. Each individual SHAP value, ϕ_{ij}, for variable j and observation x_i, is interpreted as the variable value x_{ij} that contributed ϕ_{ij} towards the prediction, for instance, x_i, compared to the average prediction for the dataset.

A positive SHAP value for a variable pushes the prediction higher, whereas a negative SHAP value pushes the prediction lower. SHAP values are considered additive, as the sum of the

SHAP values for all variables equals the difference between the baseline value and the prediction (Molnar 2022).

5.2 Indoor $PM_{2.5}$ Analysis

5.2.1 Methodology

The daily indoor PMF source contributions, the meteorological parameters, and the reported activities from the activity log from the Birżebbuġa sampling campaign were combined into one dataset. The RF model was tested using the caret package (version 6.0-92) in RStudio (Kuhn 2008) using the following predictor variables:

1. Meteorological parameter: Wind speed (ws), wind direction (wd), outdoor temperature (Temp.O), indoor temperature (Temp.I), outdoor RH (Humidity.O), indoor RH (Humidity.I) and atmospheric pressure (Pressure).
2. Indoor PMF source contributions: Indoor, Saharan, Sea Salt, Ammonium Sulfate, Traffic, Fireworks, Industrial and Shipping.
3. Reported activities from the daily activity log. These were included as categorical variables in the model. House windows closed (windows 1), house windows open (windows 2), no frying (frying 1), frying (frying 2)

Given that the objective of this analysis is to interpret the effects of the predictor variables on indoor $PM_{2.5}$, the dataset was not split into training and testing sets, but the RF model coupled with SHAP interpretation (RF-SAHP) was trained on the entire dataset. Repeated 10-fold cross-validation was carried out to assess the model's performance using RMSE as the evaluating metric, and grid search was used for hyperparameter tuning of mtry.

The SHAP values were calculated using the kernelshap package (version 0.4.1) in RStudio and then using the shapviz package (version 0.9.3) to create the visualizations. The shapviz package uses the dataset that includes all the corresponding variables for visualization only and can also be used for factor variables such as "windows" or "frying". Hence, in the visualisation of the SHAP values using the shapviz package, the dataset containing the variables is only used as an explanation dataset and is not used to calculate the SHAP values.

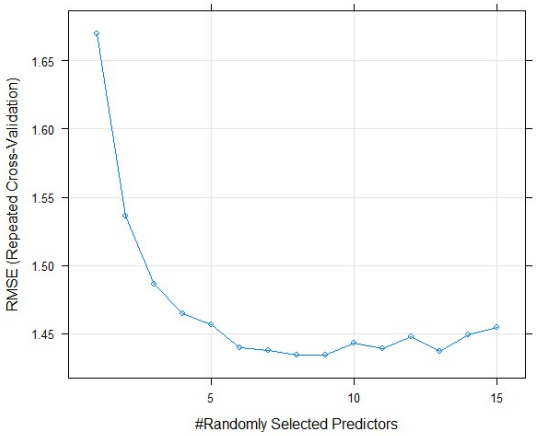
5.2.2 Results

5.2.2.1 The RF model

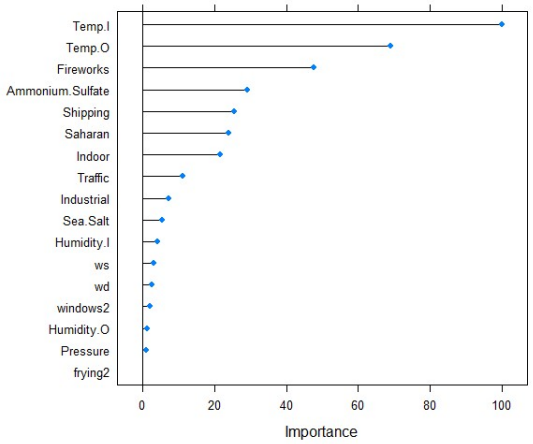
Figure 5.1(a) shows that an mtry of 9 gives a model with the lowest CV RMSE at 1.43 μg m^{-3}, having an R^2 of 0.80. These values show a good fit for the RF model, indicating that it interprets the data closely. Figure 5.1(b) gives the variable importance of the selected RF model, with indoor and outdoor temperatures being the two variables with the highest variable importance. This means these two variables were the most randomly chosen in building the RF model. Some variables with very low importance are excluded from the plot. The selected RF model was used to predict the indoor $PM_{2.5}$ on the whole dataset. Figure 5.1(c) shows an RMSE of 0.62 μg m^{-3}, an IOA of 0.99 an R^2 of 0.97.

Given that the model was trained using the entire dataset, compared to the CV values, a lower RMSE and higher R^2 are expected when the model is tested on the entire dataset since the model is being evaluated using the same data on which it was trained. During 10-fold cross-validation, the dataset is divided into ten parts, and for each of the ten iterations, the model is trained on nine parts and tested on the remaining parts. The performance metrics are then averaged over the ten iterations to give a more realistic estimate of the model's performance on unseen data. Hence, the CV RMSE is based on a model with some missing data and is a better metric for model performance. The testing R^2 is then expected to be higher than CV R^2 for the same reason.

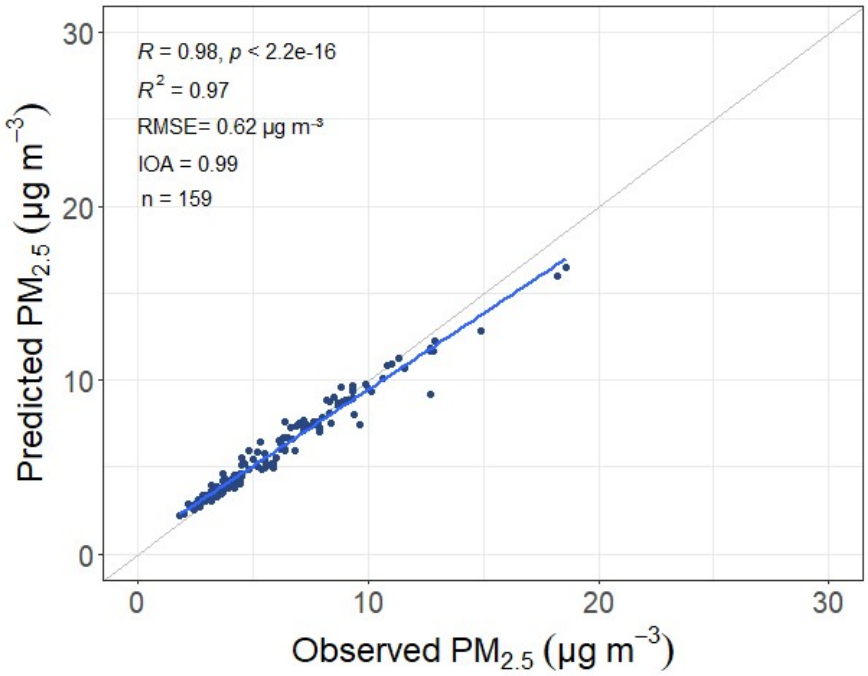
This disparity between the prediction metrics of the training dataset and cross-validation metrics underscores the importance of using cross-validation for model evaluation. CV provides a more accurate measure of how the model will perform on new, unseen data, helping to ensure that the model has not been overfitted to the training data. Hence, the cross-validation metrics provide a better performance estimate for the model.



(a)



(b)



(c)

Figure 5.1 **(a)** RMSE values for the repeated 10-fold cross-validation results and hyperparameter tuning **(b)** Variable importance plot for the RF model and **(c)** Plot of predicted against observed $PM_{2.5}$ for the RF model when it is used to predict the whole data set.

5.2.2.2 SHAP Results

The overall impact of each variable on the entire sampling period

Mean SHAP values can be used to interpret the model globally, and the SHAP importance plot in Figure 5.2 gives the average absolute SHAP value across all samples for each variable and hence shows the quantifies the impact of each variable in the model.

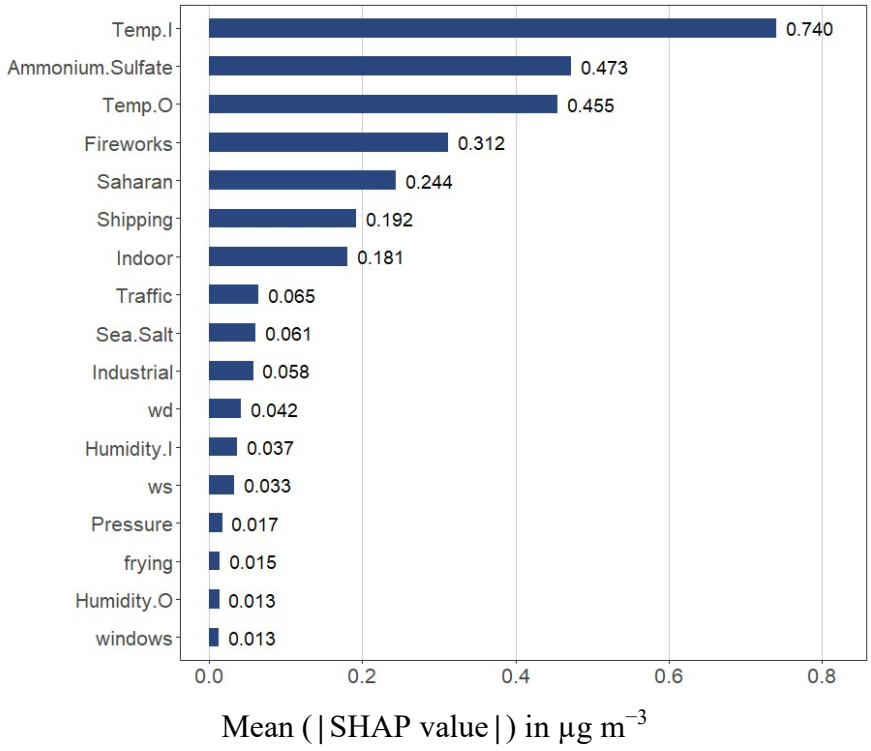


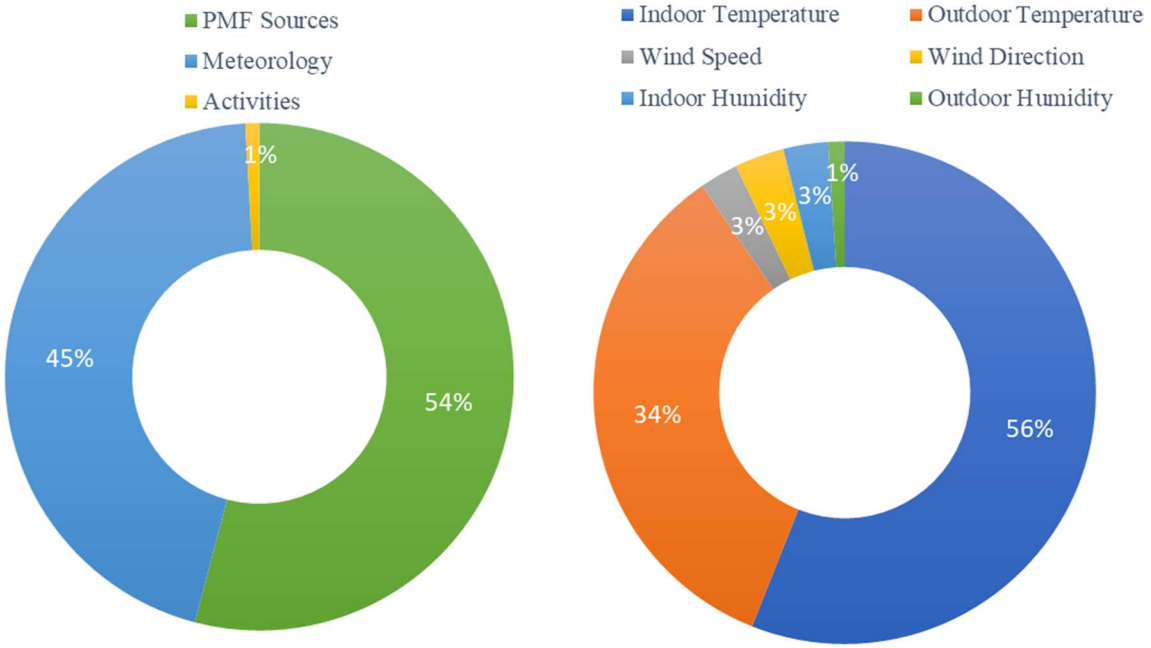
Figure 5.2 SHAP importance plot

The SHAP importance plot is not identical to the RF variable importance plot in Figure 5.1b. The latter is a plot of the model's variable importance, which measures the change in the RF model performance when a variable is randomly shuffled. Moreover, it does not quantify the fractional contribution of each variable to the predicted $PM_{2.5}$ concentration, which information is only given by the SHAP importance plot.

As shown in Figure 5.3a, the effects of meteorological factors evaluated by the RF-SHAP accounted for about 2.57 μg m^{-3} of variations in indoor $PM_{2.5}$ concentrations, with 45% of the total impact. For the impacts by meteorological factors, indoor temperature was the dominant variable (56% of meteorological impact), followed by outdoor temperature (34% of meteorological impact), with the other variables showing minimal impact.

The seasonal variation of indoor $PM_{2.5}$ concentrations can explain the significant impact of indoor and outdoor temperatures. As shown in Table 5.1, the seasonal change in indoor $PM_{2.5}$ concentration is more significant than the outdoor one. The mean indoor $PM_{2.5}$ concentration during winter is 42% of the mean summer concentration. This is partly due to the higher outdoor concentrations due to increased source contribution from SIA, such as ammonium sulfate and fireworks, but more significantly due to the higher infiltration during summer as windows were left completely open during the hotter summer months.

A study on outdoor $PM_{2.5}$ in Tianjin, China (Zhang et al. 2022), also recorded temperature as the main meteorological factor driving variations in outdoor $PM_{2.5}$ concentrations from an RF-SHAP analysis and attributed this to the significant effect of temperature on secondary particle formation. Although RH is known to influence the formation of aerosols through liquid pathway formation (Zhang et al. 2022), and has been shown as an important variable in indoor $PM_{2.5}$ prediction in Chapter 4, the low impact of RH on driving indoor $PM_{2.5}$ concentration variations could be explained by the daily averages taken in this study which could minimise effects from short-lived episodes such as cooking.



(a) **(b)**

Figure 5.3 The percentage contribution on indoor $PM_{2.5}$ as evaluated by the RF-SHAP (a) for PMF sources, meteorological factors, and household activities (b) for meteorological parameters only.

Table 5.1 Seasonal mean and median concentrations in μg m^{-3} for indoor and outdoor $PM_{2.5}$ at Birżebbuġa.

	Indoor $PM_{2.5}$		**Outdoor $PM_{2.5}$**	
	Mean (μg m^{-3})	Median (μg m^{-3})	Mean (μg m^{-3})	Median (μg m^{-3})
Summer	8.8	8.7	12	12
Autumn	5.1	4.4	9.4	8.3
Winter	3.7	3.7	9.1	8.2
Spring	5.2	3.9	13	11

Ammonium sulfate, which is of transboundary origin, is the leading outdoor source driving the variation of indoor $PM_{2.5}$ concentrations at the site, with fireworks, Saharan dust and shipping all showing higher impacts than indoor sources. This confirms that indoor $PM_{2.5}$ is significantly affected by outdoor sources at this sampling site. The fact that fireworks and Saharan dust are, on average, more important impact drivers than indoor sources can be attributed to the acute nature of these sources. Although typically short-lived, they contribute high $PM_{2.5}$ concentration during these periods, increasing the absolute mean contribution.

The SHAP summary beeswarm plot in Figure 5.4 provides another global interpretation of the model where the distribution of SHAP values for each variable is shown whilst highlighting the importance and the nature of the effect of each variable, whether positive or negative, on the prediction. Figure 5.4 confirms that the top six variables show a significant right skewness, which explains their higher mean SHAP contribution. Moreover, high feature values are associated with high SHAP contribution in the RF model for indoor and outdoor temperatures. This confirms the higher contributions to indoor $PM_{2.5}$, predominantly from ammonium sulfate, fireworks, Saharan dust and shipping during the hotter summer. While ammonium sulfate and fireworks emissions increase during the hotter summer months, the overall increased contribution is also associated with the higher infiltration of outdoor dust during the hotter summer months, given that the house is naturally ventilated.

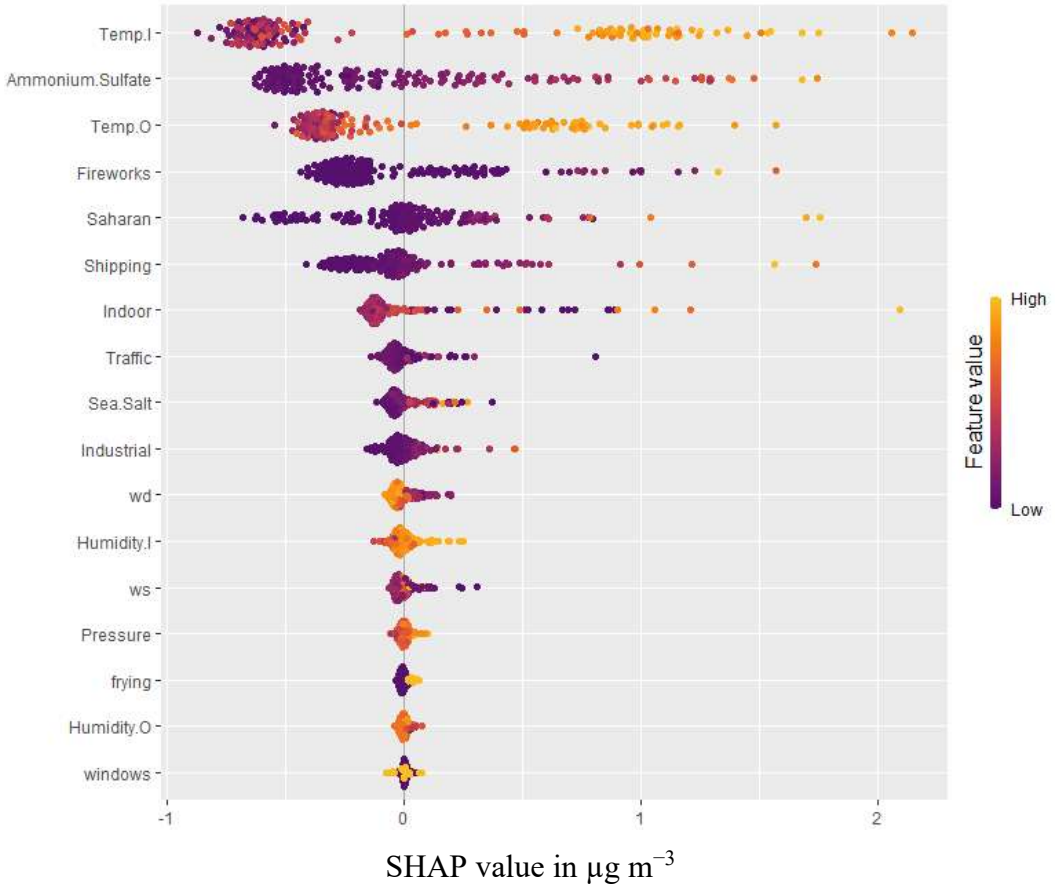


Figure 5.4 SHAP summary plot. The feature value represents the range (low to high) of each predictor variable.

Figure 5.5 gives the 2D dependency plot for the predictor variables ammonium sulfate, outdoor temperature and fireworks, respectively, with indoor temperature. The 2D dependency plots calculate the SHAP values for two interacting features, where the predicted outcome changes as a function of the values of the two specific predictor variables while considering their interactions. This displays the impact and contribution of each variable pair towards the RF model's predictions.

The 2D dependency plot confirms the increased contribution of ammonium sulfate as temperature increases. Contributions by fireworks take place during the hotter summer months when indoor temperatures exceed 25 ºC at the site, as most fireworks displays in Malta are held between May and September. The outdoor and indoor temperature 2D dependence plot gives a clear picture of the effect of natural ventilation in this house, with the associated contribution of these two variables increasing by over 2 µg m^{-3} with an increase in temperature.

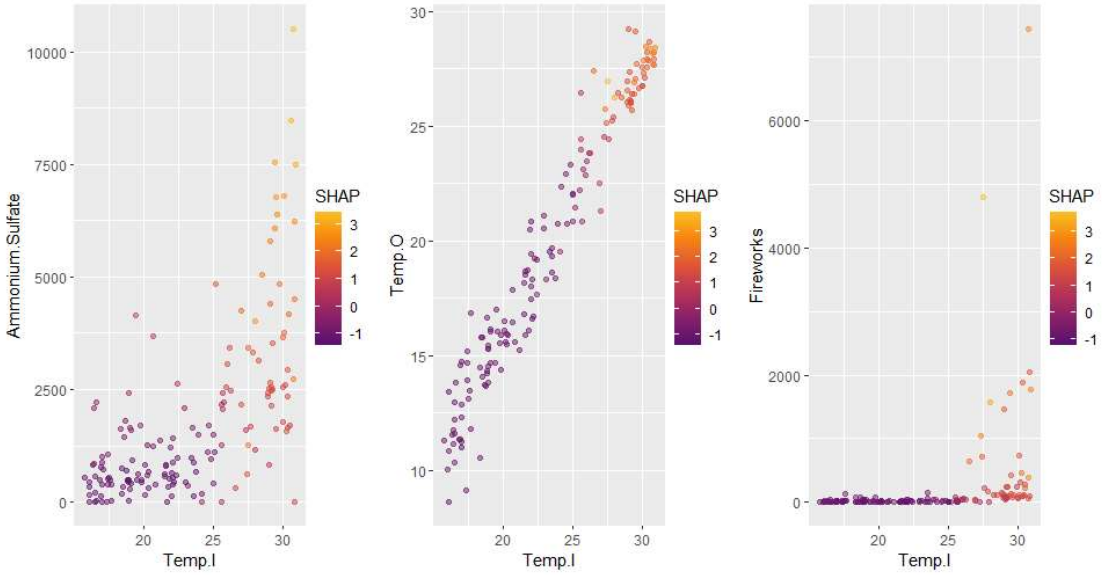


Figure 5.5 SHAP 2D Dependency plot for the predictor variables ammonium sulfate, outdoor temperature and fireworks, respectively, with indoor temperature. The SHAP values in these plots are the sum of the SHAP values from the two variables.

Given that Figure 5.6 shows no particular relationship between indoor temperature and indoor source contribution, the increased indoor $PM_{2.5}$ must be associated with increased outdoor $PM_{2.5}$ concentrations through increased infiltration due to higher AER when the windows are open. In fact, this plot shows that the highest combined SHAP contributions are at higher indoor temperatures, not at high indoor source contributions.

Although Figure 5.4 shows that frying is associated with mildly higher SHAP values, Figure 5.6 shows no significant impact on the indoor source contribution from frying, suggesting that the kitchen is well-ventilated. In fact, the house had an external hood extractor over the cooking hobs, which is known to mitigate the generation of fine particles from frying. Again, given that the data in this model analyses $PM_{2.5}$ daily averages, any short temporal increase in indoor $PM_{2.5}$ concentrations due to frying is diminished as the increase is averaged out over the 24-hour period.

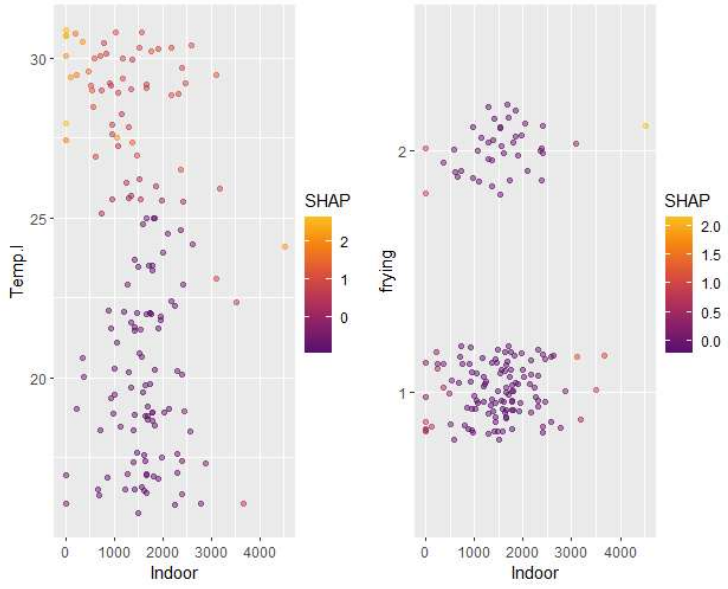
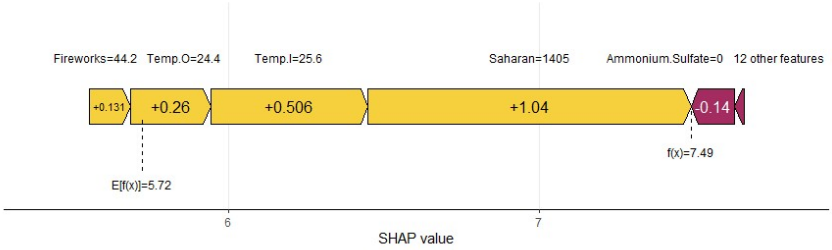


Figure 5.6 SHAP 2D Dependency plot for the predictor variables indoor temperature and frying, respectively, with the indoor source variable.

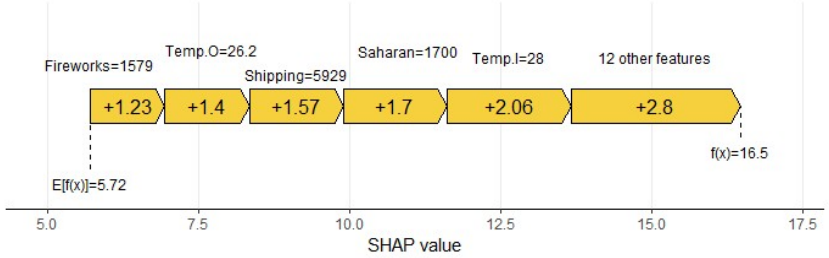
Local interpretation

SHAP can be used to carry out local interpretations of the RF model. Examining the SHAP values for individual predictions can increase the understanding of how variables influence a particular prediction. This removes the "black box" reputation that is associated with many modern ML algorithms, as it provides a quantitative way to analyse single predictions.

The baseline ensures that the SHAP values fairly distribute the difference between the observed prediction, representing the mean prediction among the contributing features. The prediction for any individual observation can be seen as starting from the baseline value. The SHAP values of the individual features are adjusted to the baseline to reach the actual prediction for the specific observation (Lundberg and Lee 2017).



(a)



(b)

Figure 5.7 SHAP force Plots for (a) 29/10/2018 and (b) 17/06/2019

Figure 5.7 shows the SHAP force plot for the local interpretation of two days with known Saharan dust events. The first one, on 29/10/18, shows a Saharan dust event with closed windows. In this instance, the predicted value is marginally higher than the baseline value, and hence, most variables have negative SHAP values because they are not contributing features. Saharan dust is shown with a positive SHAP value, highlighting the effect of this source on this day. The second one, on 17/06/19, shows a very different picture. Saharan dust is still shown to have a high positive SHAP value, but since the predicted value is significantly higher than the baseline, other features, including shipping, have positive SHAP values. The high indoor temperature and its associated contribution show that high infiltration rates were present at the site, which explains the higher indoor concentrations. Moreover, on 17/06/19, both shipping and Saharan dust were major contributing sources. These two examples show that force plots can be useful tools to interpret the impact derived from the RF model of individual sources on the $PM_{2.5}$ concentrations for specific observation.

5.3 PMF and RF-SHAP Analysis on Outdoor $PM_{2.5}$

5.3.1 Methodology

Given that only the elemental content was available for the outdoor $PM_{2.5}$ samples gathered at the Birżebbuġa sampling site, the following procedure will be followed to estimate the outdoor $PM_{2.5}$ OC, EC and ionic content from the indoor concentrations. Using the single zone mass balance model, which describes households as completely mixed flow reactors, the indoor PM concentrations depend on the outdoor PM concentration according to equation 5.1 (Abt et al. 2000; Barraza et al. 2014) :

$$C_i = \frac{PaC_o}{a+k} + \frac{Q_i}{V(a+k)} \equiv F_{INF}C_o + C_{ig} \qquad \textbf{Eq. 5.1}$$

where *Ci* and *Co* are the measured indoor and outdoor $PM_{2.5}$ (µg m^{-3}), respectively, *a* is the household air exchange rate (AER, h^{-1}), *P* is a dimensionless penetration coefficient and *k* the rate of removal of PM by surface deposition or reaction (h^{-1}). The term *Pa/(a +k)* is the infiltration factor (F_{inf}), and it quantifies the fraction of outdoor $PM_{2.5}$ that infiltrates indoors. Q_i is the rate of indoor generation and resuspension of PM (µg h^{-1}), and *V* is the household volume (m^3). Q_i */V(a+k)* is the indoor-generated concentration, C_{ig}.

From equation 5.1, the infiltration factor can be described by equation 5.2:

$$F_{INF} = \frac{C_i - C_{ig}}{C_o} \qquad \textbf{Eq. 5.2}$$

Given that C_{ig} for OC, EC and ionic content can be derived from the contribution of the indoor source profile from the indoor PMF analysis, the outdoor concentration can be calculated from equation 5.3:

$$C_o = \frac{C_i - C_{ig}}{F_{INF}} \qquad \textbf{Eq. 5.3}$$

To increase the effectiveness of the outdoor PMF solution, outdoor concentrations of chemical species OC, EC, chloride, nitrate, sulfate and ammonium ions were estimated from indoor ionic and OC/EC concentrations and the indoor PMF analysis. Equation 5.4 was used to estimate the outdoor concentrations of these chemical species.

$$y_{ij} = \frac{x_{ij} - (g_i f_j)}{r_i} \qquad \textbf{Eq. 5.4}$$

y_{ij} = estimated outdoor concentration (ng m^{-3}) for sample *i* and chemical species *j*

f_j = concentration (ng m^{-3}) of PM species j in the indoor source profile

g_i = indoor factor contribution to sample *i*

x_{ij}= indoor concentration (ng m^{-3}) of PM species *j* measured in sample *i*;

r_i = $PM_{2.5}$ indoor-outdoor ratio

The assumption for calculating concentrations for OC, EC, chloride, nitrate, sulfate and ammonium ions, was that the concentrations calculated by the indoor PMF solution from the indoor source profile are of indoor origin. Hence, these calculated indoor concentrations could be deducted from the total indoor concentration of these ions, resulting in the indoor concentrations that originated from outdoor sources.

The y_{ij} concentrations from equation 5.4 were combined with the analysed outdoor elemental concentrations to calculate the determined outdoor $PM_{2.5}$ concentrations. Given that most of this outdoor $PM_{2.5}$ concentration is estimated and not analysed, it will be referred to as determined outdoor $PM_{2.5}$ concentration. On average, this determined outdoor concentration comprises 11% of the analysed outdoor $PM_{2.5}$ elemental content and 89% of the estimated concentrations for OC, EC and ionic content.

The $PM_{2.5}$ I/O ratio was used in equation 5.4 instead of F_{INF} as described in equation 5.3, because when F_{INF} was used, the determined outdoor $PM_{2.5}$ concentrations exceeded the observed (gravimetric) outdoor $PM_{2.5}$ concentration for 51% of the observations. The yearly average was 123%, ranging from 100% (summer) to a maximum of 164% (autumn). This was not deemed acceptable due to the significantly over-estimated outdoor $PM_{2.5}$ concentrations.

Hence, the F_{INF} was substituted with the I/O $PM_{2.5}$ ratio, r_i, which provided a conservative estimate of y_{ij}. When outdoor $PM_{2.5}$ concentrations were estimated with concentrations derived from equation 5.4, the determined outdoor $PM_{2.5}$ concentration exceeded the observed outdoor $PM_{2.5}$ concentration for only 4% of the observations. Using this method, the

determined outdoor $PM_{2.5}$ concentrations were, on average, 70 % of the gravimetric $PM_{2.5}$ concentration.

Table 5.2 Annual and seasonal percentage determined outdoor concentration and the percentage of sampling days with open windows.

	Percentage determined outdoor concentration	Percentage of sampling days with open windows
Summer	87	98
Autumn	64	33
Winter	57	0
Spring	70	29
Annual	70	40

Table 5.2 shows the seasonal variation of the percentage determined outdoor $PM_{2.5}$ concentrations, with a minimum of 57% in winter and a maximum of 87% in summer. This correlated to the change in the percentage of sampling days with open windows, a metric for outdoor PM infiltration. The AER at the sampling site increased eight times when the windows were open (Table 3.2), making the indoor $PM_{2.5}$ more influenced by outdoor sources. Hence, the calculation in equation 5.4 becomes more effective at high AER. Conversely, during colder winter days, with closed windows, indoor $PM_{2.5}$ is less affected by outdoor sources, and hence, the estimation decreases in accuracy.

Using F_{INF} would have significantly overestimated the determined outdoor $PM_{2.5}$ concentrations. This would have overestimated the ionic and OC EC contributions for the outdoor PMF analysis, potentially producing biased source profiles. The more conservative calculation with the $PM_{2.5}$ I/O ratio gives a more careful yet substantive determination of outdoor concentrations that could be used for the PMF analysis. An RF-SHAP analysis was used to fit the outdoor PMF source contributions to the observed $PM_{2.5}$ concentrations. The RF model via SHAP interpretation was then used to determine the contribution of outdoor sources at the sample site.

5.3.2 Interpretation of the 6- to 8-factor outdoor PMF solution

The analytes were screened for their signal-to-noise ratios (S/N) (Paatero and Hopke 2003; Zhao et al. 2006). Analytes with a S/N lower than 4 were down-weighted to weak, which included the analytes Ca, Cr, As, Cd, Sb, Cl^- and NO_3^-.

PMF is essentially a descriptive model, and determining the optimal number of factors that are fitted in the model remains subjective (R Vecchi et al. 2008). Since the PMF analysis was run after the indoor one and the best indoor PMF mass balance analysis was resolved with an 8-factor solution, the outdoor data was tested with a 6- to 8-factor solution. All base runs for the different number of factors solutions converged. The final number of factors was chosen depending on the results of the Q(Robust) and Q(True) values for each run, the intra-run residuals, the Bootstrap (BS) analysis and the Bootstrap Displacement (BD-DSP) analysis (Norris et al. 2014; Brown et al. 2015).

For the 6-factor solution, the intra-run residuals showed several base runs with significantly different fits. Most factors were mapped more than 90% of the time in the BS analysis of the run with the lowest Q(Robust), although several factors were not mapped. A very high Q/Qexp for several species, including OC, K, Mn, Ni, Cu, Zn, Sr, Ba, sulfate ions and ammonium ions, characterised the run with the lowest Q(Robust), suggesting a poor fit.

The 8-factor solution produced near-identical intra-run residuals for all base runs except one. A high Q/Qexp for some species, including K, Mn, Cu, Zn, Sr and Pb, were calculated. The BS resulted in two factors with 60% and 14% mapping, respectively, indicating that a phantom factor was created, which translates in the fragmentation of a real single source over more than one factor.

The 7-factor solution produced near-identical intra-run residuals for all base runs, showing that all base runs had similar fits. A high Q/Qexp was calculated for some species, including OC, K, Ni, Cu, Zn, Sr, Pb and ammonium ions. The BS analysis of the base run resulted in all factors being mapped at 100% except factor 3 (Fireworks), which was mapped at 83%, which is to be expected as this source factor is only present during a restricted number of sampling days. The 7-factor solution was retained as it provided the best solution for the data, and it was investigated further. Constraints were applied to this PMF solution to decrease rotational ambiguity.

Five data points were sampled in August during fireworks events, and two data points in June 2019 were pulled up to the maximum for the fireworks factor (factor 3). When this was carried out, the modelled $PM_{2.5}$ concentrations were closer to the sampled concentrations for these

days. Given that no fireworks displays occur in winter, five data points that showed a slight contribution by the "Fireworks" factor were set to zero. Moreover, species known to be present in fireworks, namely, chloride, nitrate, barium and antimony, were pulled up to maximum for this factor. These constraints resulted in an increase in dQ(robust) of 3.1% for the 7-factor solution. This was deemed acceptable as a more interpretable solution was provided, with the Q/Qexp decreasing below 3 for most species. Moreover, all factors were mapped at 100% in the constrained BS analysis, and no factor swaps were registered in the displacement analysis. This indicated a robust solution with minimal rotational ambiguity.

The BS-DISP analysis on the constrained run was used to compare the robustness of the 7-factor solutions. The species selected to be displaced in BS-DISP were OC, V, Ba and SO_4^{2-} as these are the key species required for factor identification. Norris et al.(2014) suggest that a small number of variables in BS-DISP should be used to speed up the computation analysis. Six swap counts were registered for the lowest dQmax level, indicating that the solution is not perfectly constrained. The factor swaps were primarily between the Fireworks and Saharan factors, which is to be expected due to the transient and infrequent nature of these sources. Two-factor swaps were also registered for the Shipping factor.

Obtaining a perfectly constrained model in the BS-DISP model is challenging, given that fireworks are present for a limited number of sampling days. Moreover, this PMF solution is being carried out on the determined outdoor concentration fraction that accounted for 70% of the gravimetric $PM_{2.5}$ concentration. In comparison, the analysed indoor concentration accounted for 101% of the gravimetric $PM_{2.5}$ concentration, and hence, the outdoor PMF solution is expected to be less robust than the indoor one due to the lower fraction of the analysed $PM_{2.5}$ concentration used for PMF analysis. Nonetheless, the 7-factor solution was accepted as the most robust solution, providing the largest number of physically interpretable solutions. The solution achieved a reconstruction of the $PM_{2.5}$ mass with the modelled mass explaining 64% of the observed gravimetric mass with a 0.77 coefficient of determination (R^2) for the plot of the modelled against observed $PM_{2.5}$ and RMSE of 4.6 μg m^{-3} (Figure 5.8).

The RMSE for the modelled against observed $PM_{2.5}$ decreased to 3.0 μg m^{-3} for sampling days with open windows and increased to 5.6 μg m^{-3} for sampling days with closed windows. This indicates that the model performance is better for sampling days with higher AER, which is expected given the method used for the PMF solution.

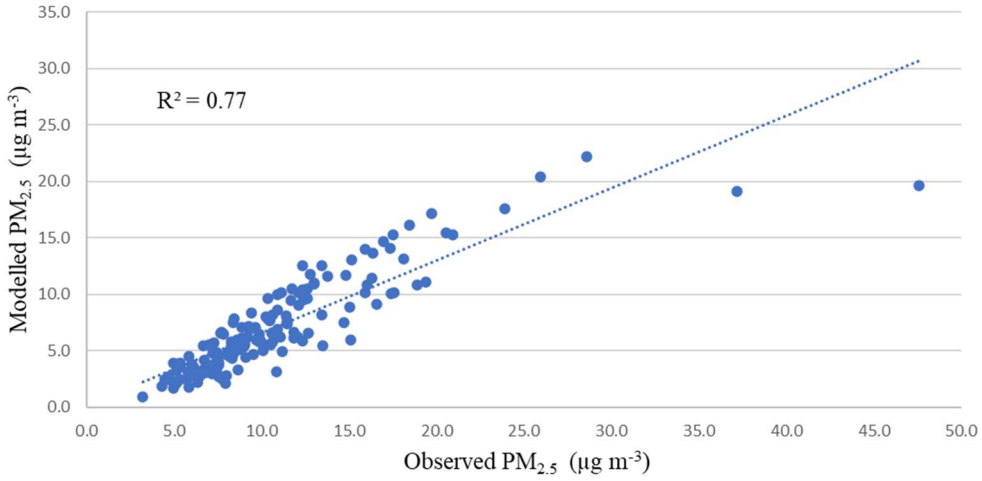
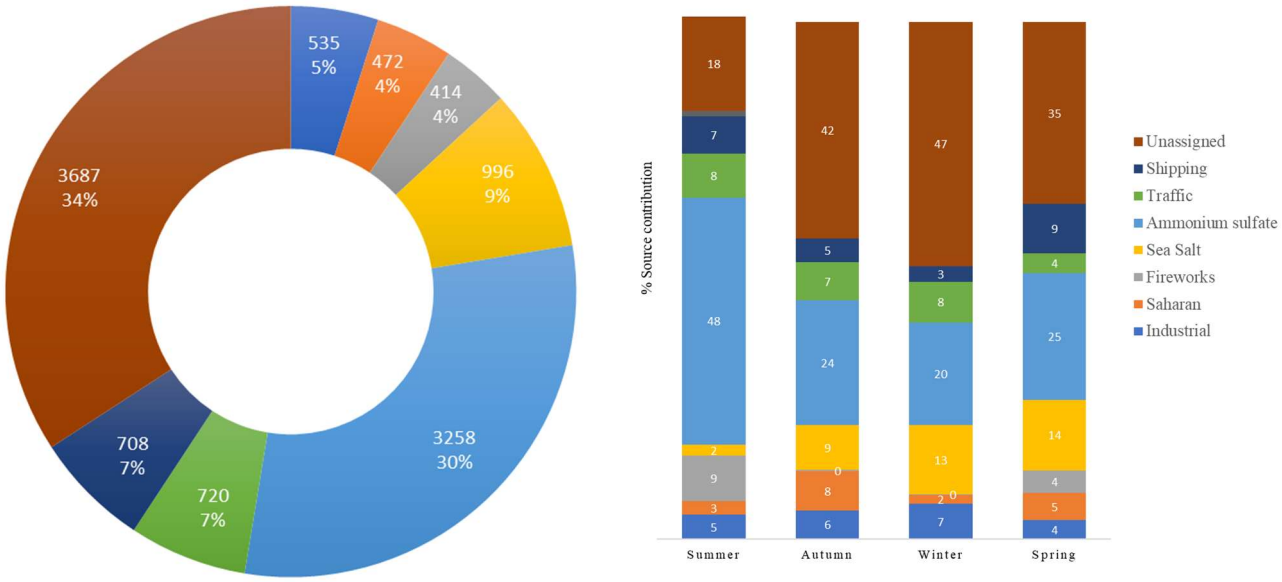


Figure 5.8 Plot of PMF-modelled $PM_{2.5}$ concentrations against observed outdoor $PM_{2.5}$ concentrations.

5.3.3 PMF results for the outdoor $PM_{2.5}$ solution

The profiles for the 7-factor solution are shown in Figure 5.10. The uncertainty levels are set to the 5th and 95th percentile of the BS runs for the constrained solution. The yearly and seasonal contributions by each factor to outdoor $PM_{2.5}$ as calculated by the PMF solution are shown in Figures 5.8a and 5.8b, respectively.



(a) **(b)**

Figure 5.9 **(a)** Annual mean PMF source contributions to outdoor $PM_{2.5}$. The values are given as percentages and in ng m^{-3}. **(b)** Seasonal percentage PMF source contributions to outdoor $PM_{2.5}$

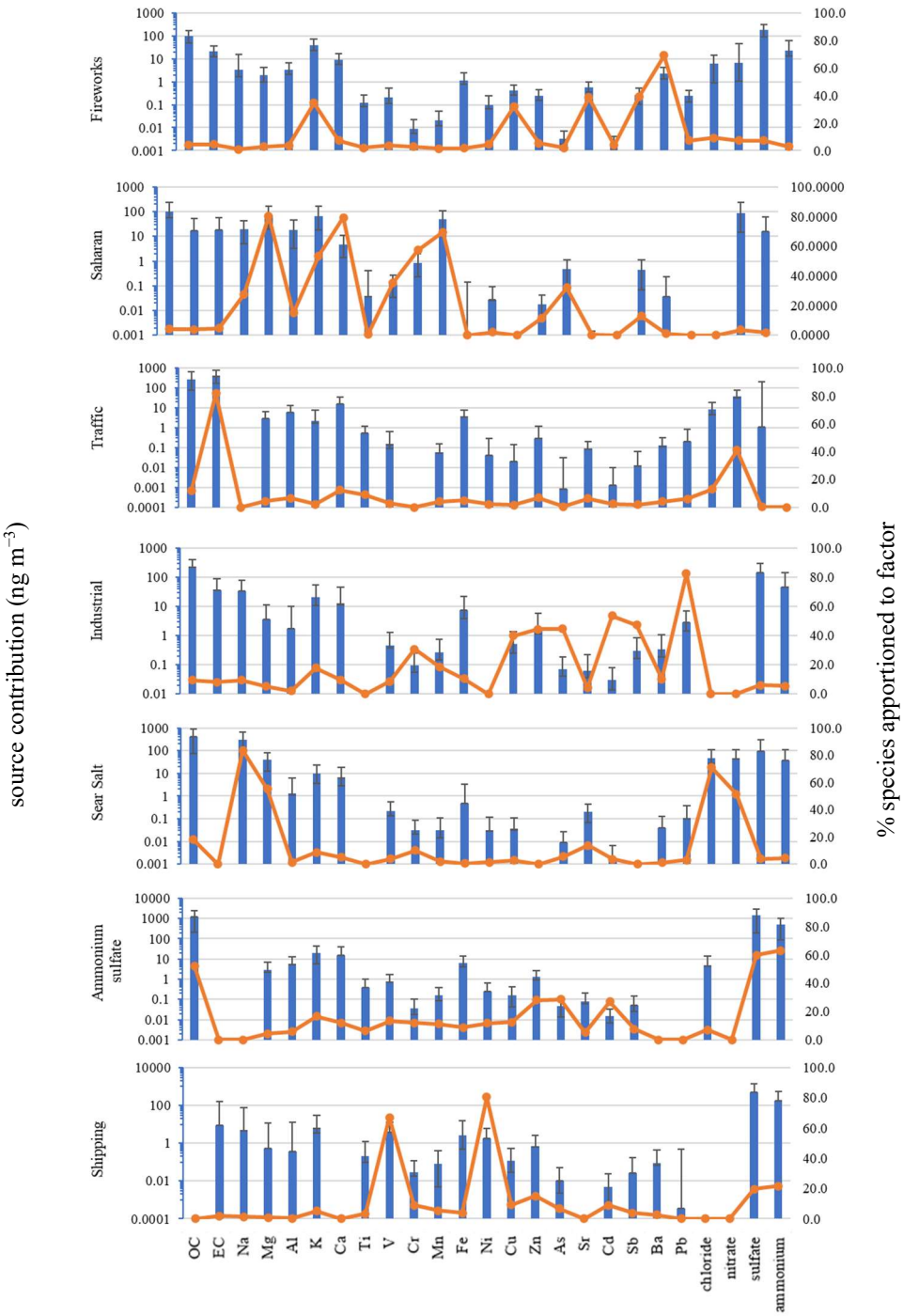


Figure 5.10 Outdoor factor profiles (bars) and percentage species apportioned to each factor (line)

5.3.3.1 RF-PMF analysis

An RF model coupled with SHAP interpretation (RF-SAHP) was carried out for the PMF outdoor solution to test if this approach can be used to account for the unassigned fraction of $PM_{2.5}$ and calculate an RF-SHAP percentage contribution of the outdoor PMF source factors. Two RF models were run with repeated 10-fold cross-validation over the entire dataset. Hyperparameter tuning for mtry using grid search was carried out. Model 1 included only the outdoor PMF source contributions as predictor variables, while Model 2 had the PMF source contribution and meteorological parameters as predictor variables. The results are shown in Figure 5.11.

Model 1 was run with an optimised mtry of 8 with the following predictor variables:

1. PMF source contributions: Saharan, Sea Salt, Ammonium sulfate, Traffic, Fireworks, Industrial and Shipping.

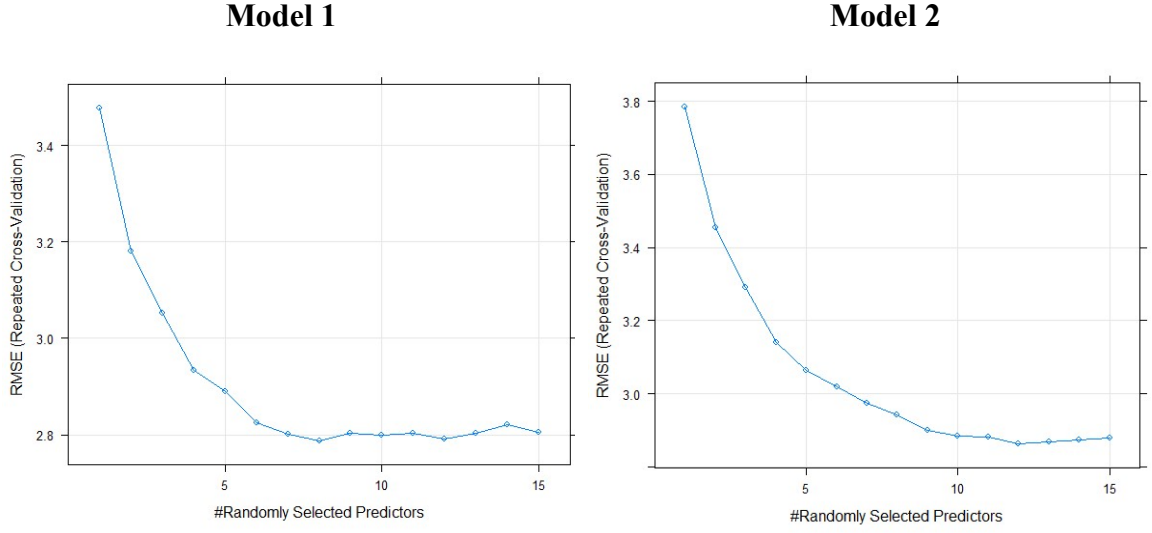
This model gave a CV RMSE of 2.79 µg m^{-3} with an R^2 of 0.80.

Model 2 was run with an optimised mtry of 12 with the following predictor variables:

1. Meteorological parameter: Wind speed (ws), wind direction (wd), outdoor temperature (Temp.O), indoor temperature (Temp.I), outdoor RH (Humidity.O), indoor RH (Humidity.I) and atmospheric pressure (Pressure).
2. PMF source contributions: Saharan, Sea Salt, Ammonium sulfate, Traffic, Fireworks, Industrial and Shipping.

This model gave a CV RMSE at 2.86 µg m^{-3}, having an R^2 of 0.80.

(a)



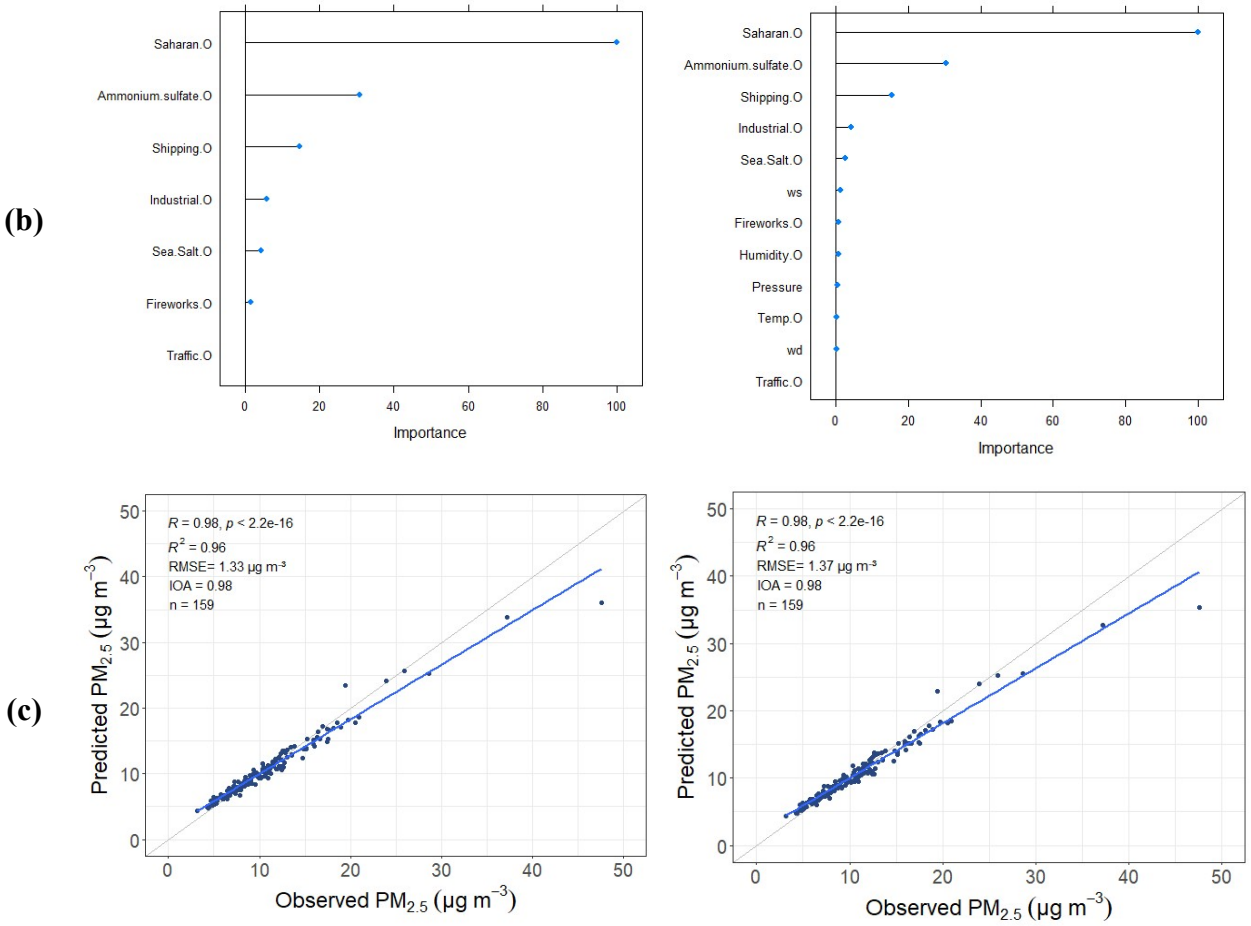


Figure 5.11 RF results for models 1 and 2 **(a)** CV RMSE for the various mtry values used in the hyperparameter tuning, **(b)** Variable importance plots and **(c)** Predicted vs observed $PM_{2.5}$ plots showing the RF model used to predict the whole dataset, including R^2 and RMSE values.

Model 1 was selected because it had a marginally better CV RMSE and predicted the observed outdoor $PM_{2.5}$ with an RMSE of 1.33 µg m^{-3}. Both models had the same IOA and R^2. This suggests that including meteorological parameters did not increase the model's performance. In fact, contrary to the indoor RF model, the variable importance plot for Model 2 in Figure 5.11b shows that the meteorological variables show negligible importance. The SHAP values were calculated on the RF Model 1 using the kernelshap package (version 0.4.1) in RStudio and then using the shapviz package (version 0.9.3) to create the visualizations.

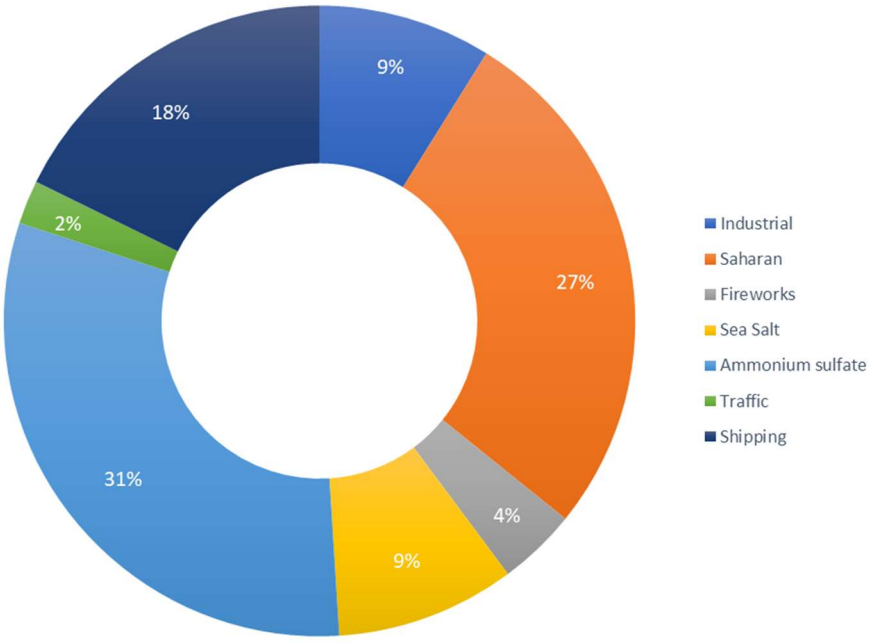
As with the indoor RF model, the SHAP values were calculated over the entire dataset since the RF-SHAP is being used to explain the contributions of the PMF factors for the whole

dataset. The mean SHAP values for the predictor variables were calculated using the RF model and are given in Figure 5.12b. These values were then used to compute the percentage contribution of the PMF source profiles. The percentage SHAP contributions were calculated using equation 5.5 and are shown in Figure 5.12a. Seasonal percentage SHAP source contributions to outdoor $PM_{2.5}$ are given in Figure D.3.

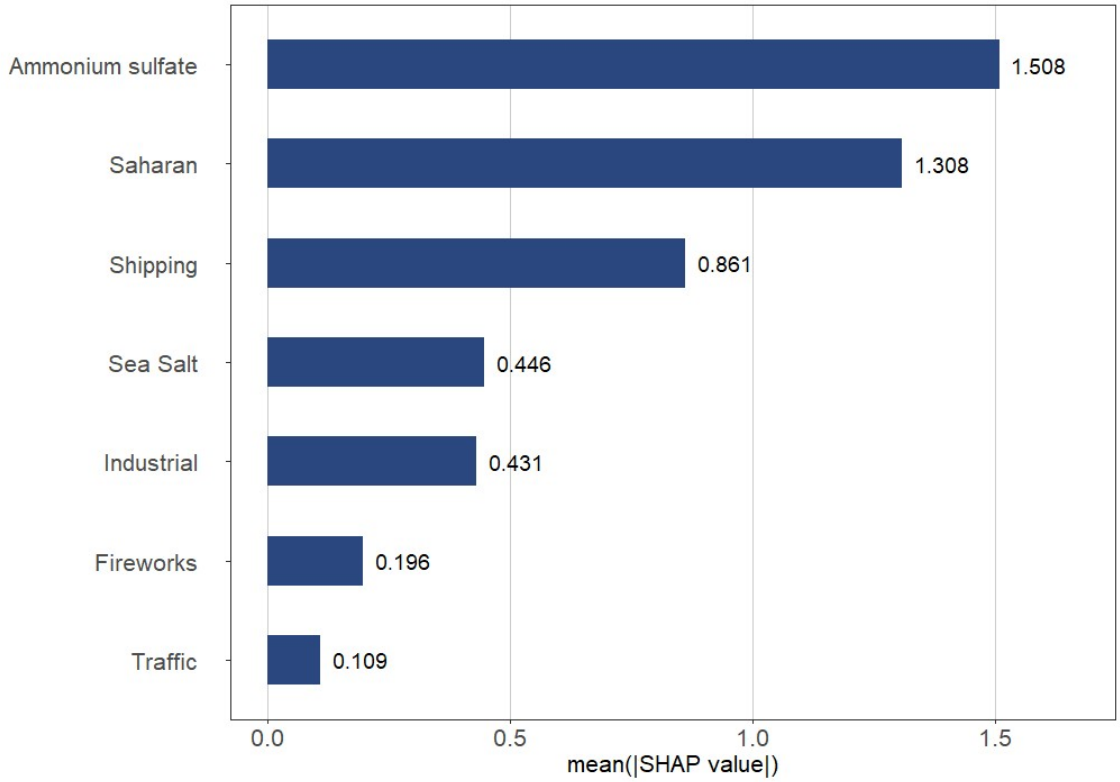
$$SHAP\ \%\ contribution = \left(\frac{\phi_j}{\sum \phi_j}\right) * 100 \qquad \textbf{Eq. 5.5}$$

Where ϕ_j is the annual mean SHAP contribution of PMF source factor k (ng m^{-3}) and $\sum \phi_j$ is the sum of the annual mean SHAP contributions for all factors.

The most pronounced difference between Figure 5.12a and Figure 5.9a is the increased contribution from the shipping (7 to 18%) and Saharan (4 to 27%) factors, which made up most of the difference for the unassigned fraction in the outdoor PMF solution. These two factors were investigated to confirm how the individual SHAP predictions contributed to an increased mean SHAP value and higher percentage contribution.



(a)



(b)

Figure 5.12 **(a)** Percentage SHAP source variable for outdoor $PM_{2.5}$. The values are given as percentages and in ng m^{-3}. **(b)** Mean SHAP values for the source variables of outdoor $PM_{2.5}$ in µg m^{-3}

Saharan factor

A considerably increased contribution by the Saharan variable in the RF model was registered on days with high PMF normalised contributions for the Saharan factor. This translated into high SHAP values for the Saharan variable in the SHAP calculation. Moreover, on most of these days, the modelled PMF $PM_{2.5}$ was significantly lower than the observed (gravimetric) $PM_{2.5}$ concentration. In fact, there were 20 days where the PMF normalised contribution for the Saharan factor was greater than 2 (i.e. days with a predominant contribution by the Saharan factor), and for these days, the unassigned $PM_{2.5}$ concentration was 6.6 ± 3.1 µg m^{-3} which is high considering that the mean outdoor $PM_{2.5}$ was 10.9 µg m^{-3}. Hence, the increased attribution to the Saharan variable by the RF model towards the $PM_{2.5}$ prediction was significant and explains why the RF-SHAP contribution is significantly higher than the one from the PMF solution.

The SHAP force plots in Figure 5.13 for three sampling days with confirmed Saharan dust events illustrate the substantially increased contribution to the Saharan factor by the RF-SHAP analysis.

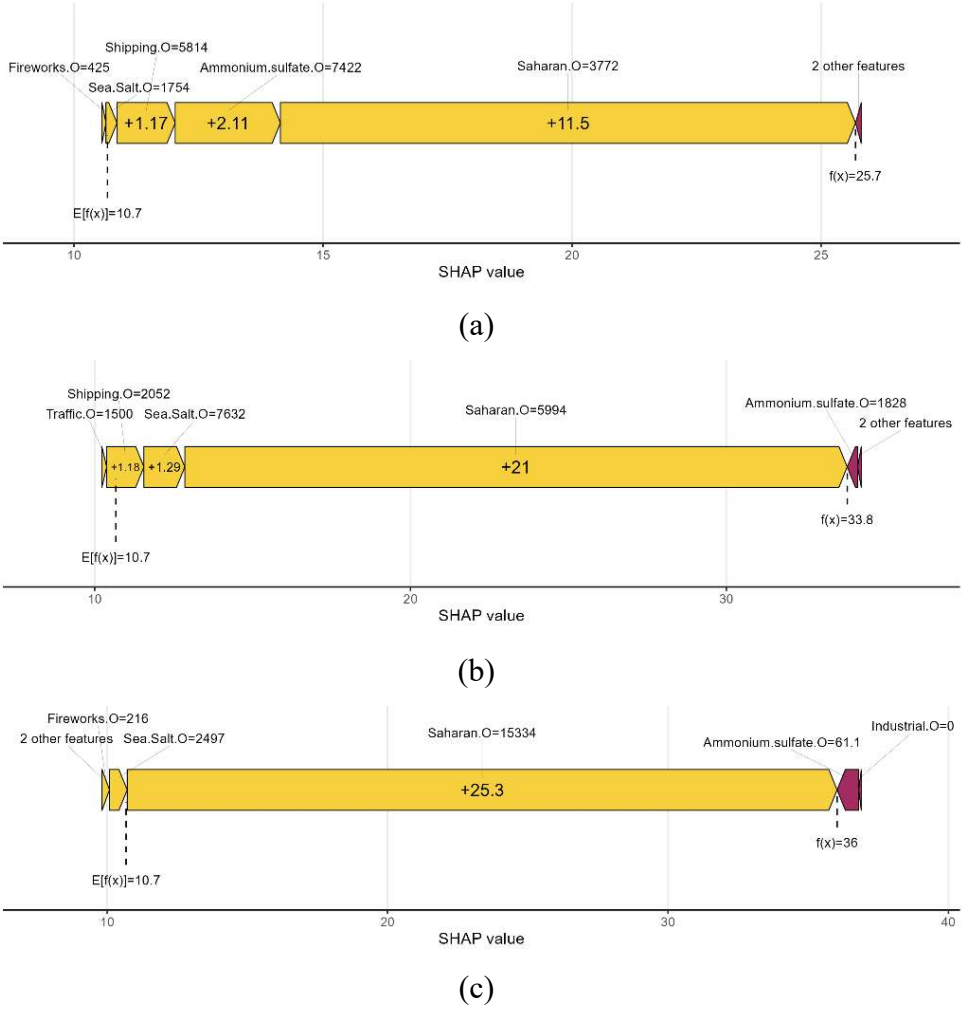


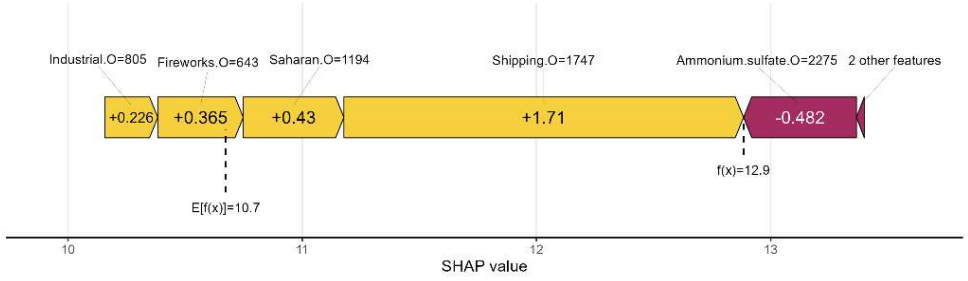
Figure 5.13 SHAP force plots for the Saharan variable for some of the days having a Saharan dust event confirmed via Hysplit analysis (a) 29/10/18 (b) 25/04/19, and (c) 17/6/19. SHAP values in μg m^{-3}, PMF contributions in ng m^{-3}.

Shipping factor

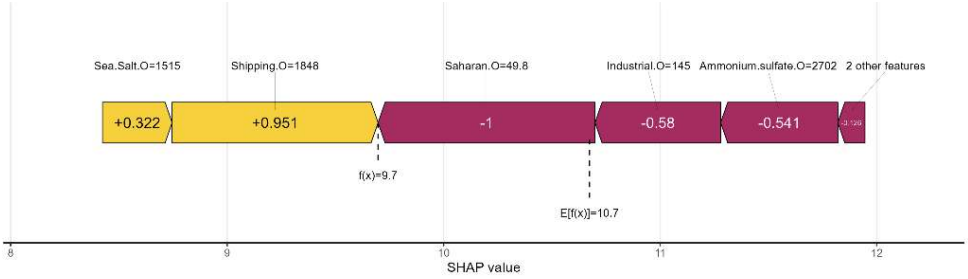
A similar increase was calculated for the shipping factor, albeit to a lower magnitude. There were 25 days when the PMF normalised contribution for the shipping factor was greater than 2; for these days, the unassigned $PM_{2.5}$ concentration was 4.4 ± 2.1 μg m^{-3}. Again, this explains the increased RF-SHAP contributions for the shipping factor compared to the one derived from the PMF solution. The SHAP force plots in Figure 5.14 for two sampling days with a high

contribution by shipping factor illustrate the substantially increased contribution to the shipping factor by the RF-SHAP analysis.

For Figure 5.14b, the RF prediction is lower than the baseline value and highlights variables with negative and positive SHAP values to bring the baseline value to the RF predicted value. Even though the SHAP contribution assigned to the shipping variable (0.951 µg m^{-3}) is lower than the PMF shipping contribution (1848 ng m^{-3}), it is important to underline that the SHAP contribution is over and above the baseline value and hence a positive SHAP value is still significant.



(a)



(b)

Figure 5.14 SHAP force plots for some days having a high contribution by shipping factor (a) 08/09/2018 and (b) 06/06/19. SHAP values in µg m^{-3}, PMF contributions in ng m^{-3}.

5.3.3.2 Discussion

Tables 5.3 and 5.4 show the percentage contributions from the indoor PMF solution and the outdoor RF-SHAP analysis. Local anthropogenic contribution to $PM_{2.5}$ is significant in both indoor (25%) and outdoor (33%). The outdoor $PM_{2.5}$ is dominated by transboundary dust, which is both anthropogenic (SIAs, mainly as ammonium sulfate) and natural (Saharan dust).

Table 5.3 Indoor and outdoor $PM_{2.5}$ source contributions

	Percentage Source Contribution	
Source Factor	Indoor (PMF)	Outdoor (RF-SHAP)
Indoor	26	N/A
Ammonium sulfate	31	31
Fireworks	4	4
Industrial	2	9
Saharan	2	27
Sea salt	9	9
Shipping	10	18
Traffic	10	2
Unassigned	6	N/A

Table 5.4 Indoor and outdoor $PM_{2.5}$ source contributions grouped by type of emission

	Percentage Source Contribution		Sources
	Indoor (PMF)	Outdoor (RF-SHAP)	
Indoor	26	N/A	
Natural	12	36	Sea salt and Saharan dust
Transboundary	33	58	Saharan dust and ammonium sulfate
Local Anthropogenic	25	33	Industrial, fireworks, Traffic and shipping

The outdoor source profiles for ammonium sulfate, shipping, Saharan dust, sea salt, and Fireworks are very similar to indoor ones and will not be discussed in detail again in this section. Fireworks is the highest source contributor for outdoor Sr (39%) and Ba (69%). It is characterised by the elements Cu, Sr, Ba, K, Cl^-, NO_3^- and SO_4^{2-}. Both PMF and RF-SHAP show a 4% contribution by this factor to outdoor $PM_{2.5}$.

The outdoor Saharan source profile is also characterised by the elements Al, Mg, Ca, Ti, Fe, Mn and Sr, which are considered markers of Saharan dust (Rodríguez et al. 2002; Viana et al. 2008) and is responsible for 80.4% of the outdoor Al, 79.5% of the Ti and 69.5% of the Fe at the site. The increase in normalised contributions matched the indoor Saharan factor, albeit to different magnitudes. For days with a high outdoor Saharan factor normalised contribution (>2), the outdoor and indoor normalised contributions were very similar in magnitude for days when the residence windows were open. For days with closed windows, the outdoor normalised contributions were, on average, seven times higher than the indoor ones.

Hence, although this factor only contributes 2% to the indoor $PM_{2.5}$, the outdoor contribution increases to 4% of outdoor $PM_{2.5}$ from the PMF analysis and 27% from the RF-SHAP analysis. The latter is higher than the contributions of Saharan dust to outdoor $PM_{2.5}$ at Msida, which were recorded at 15.1% (Scerri et al. 2018) and 17.6 % crustal contribution at an UB site in Italy (D. Cesari et al. 2016). The high percentage contribution from this source can be explained by the very high outdoor $PM_{2.5}$ concentrations recorded for days with high Saharan contributions. Moreover, the elemental contribution calculated by the PMF Saharan source does not account for the complete composition of the dust's mineral composition. In fact, in reconstituted mass (RM) solutions, minerals are typically calculated as 2.2Al + 2.49Si + 1.63Ca + 2.42Fe + 1.94Ti (R.E Hester et al. 2016). Thus, in the RF model the source contribution is being fit to the actual $PM_{2.5}$ mass and should give a better approximation of the actual contribution.

High Na and Cl levels characterise the outdoor sea salt source profile. It is the site's most important source for sodium (84%) and chloride (70.8%). The Cl/Na ratio of 0.15 is considerably lower than the reported mass ratio for sea spray at 1.8 (Moller 1990; Seinfeld and Pandis 2016). However, the Na/Mg ratio stands at 8.0, which is very close to the ratio of these elements in seawater at 8.36 (Seinfeld and Pandis 2016), suggesting that these elements are of marine origin. Like the indoor source contribution, this outdoor factor represents aged sea salt, given the presence of significant nitrate (contributing to 52% of the nitrate at the site) and sulfate, together with decreased chloride levels. Both PMF and RF-SHAP give a 9% contribution by this factor to outdoor $PM_{2.5,}$ which is notably lower than the 17.3% recorded for aged and fresh sea salt at Msida (Scerri et al. 2018) but higher than the mean 4 ± 3% reported for European countries (Belis et al. 2013).

Ammonium sulfate is the most important outdoor factor at the site, contributing 30% of the $PM_{2.5}$ and is characterised by elevated amounts of NH_4^+ and SO_4^{2-} ions. The NH_4^+/ SO_4^{2-} mass ratio for this factor is 0.35, which is nearly identical to the mass ratio of $(NH_4)_2SO_4$ at 0.375. This source factor is very similar to the indoor one, which is to be expected given that its main components, ammonium and sulfate ions, were estimated from indoor concentration. Similar to the indoor solution, this factor is responsible for 27.0% of the Cd and 28.7% of the As at the site, suggesting a transboundary contribution of these toxic elements. The PMF contribution (30%) and the RF-SHAP contribution (31%) are also nearly identical. A reconstructed mass

using the RF-SHAP value yields a contribution of 3.4 µg $m^{-3,}$ which is nearly twice the value for the indoor source contribution but similar to that recorded by Scerri et al. for outdoor air in Msida (3.6 µg m^{-3}), confirming that this factor is predominantly of outdoor and transboundary origin.

The shipping source profile is the highest source contributor for V (67.0%) and Ni (80.6%). The V/Ni ratio for this profile is 2.1, which is slightly lower than the indoor source profile (3.2) and similar to the 2.5 ratio recorded for the contribution by shipping in Msida (Scerri et al. 2018). The PMF contribution (7%) is much lower than the RF-SHAP contribution (18%). The 7% PMF outdoor source contribution is lower than the indoor value and is most probably an underestimation due to the high amount of unassigned mass. However, both values are much higher than the 5% contribution recorded by Scerri et al. in Msida. A reconstructed mass based on the RF-SHAP value yields a contribution of 2.0 µg m^{-3}, confirming that shipping emissions substantially impact both the outdoor and indoor $PM_{2.5}$ concentrations.

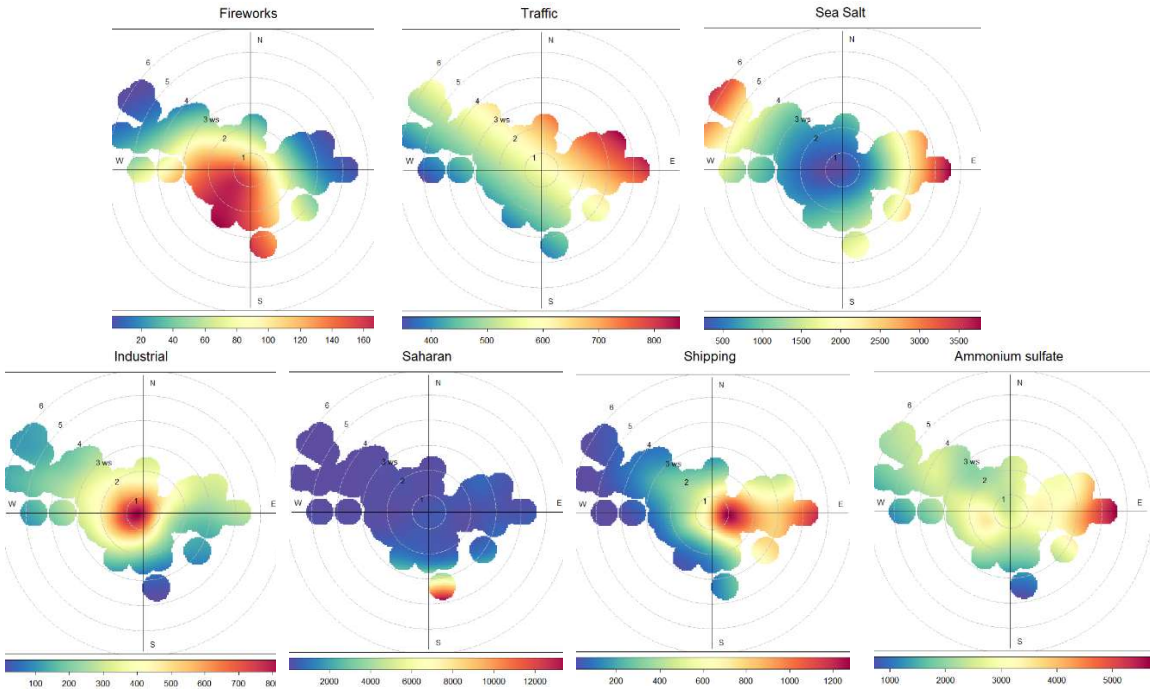


Figure 5.15 Outdoor polar plots for the source contributions in ng m^{-3}.

The polar plots for outdoor Saharan, shipping and ammonium sulfate are nearly indistinguishable from the indoor polar plots shown in Figure 3.10. This means that the distribution of these source contributions by wind speed and direction is similar to those determined by the indoor PMF solution. The indoor and outdoor polar plots for sea salt and

fireworks show minor differences. The increased contribution by sea salt at higher wind speeds is more defined in the outdoor polar plot. For fireworks, the highest outdoor contributions are from a south-westerly direction, while the highest indoor contributions are from a westerly direction.

The industrial factor is the highest source contributor for Pb and Cd in both indoor and outdoor PMF solutions, which have been recorded to be derived from industrial sites (Daniele Contini et al. 2014). The outdoor industrial factor accounts for 83% of Pb and 54% of Cd. The outdoor industrial factor is also responsible for 40% of Cu, 45% of Zn and As at the site, while the indoor one shows a lower contribution from these metals. The polar plots for this factor are also slightly different, with the indoor one having a more pronounced westerly contribution.

Interestingly, the polar plots for this source are mainly defined by Pb. In fact, the outdoor and indoor polar plots are essentially identical to the polar plots of outdoor and indoor Pb, respectively, as shown in Figure D.1. This can be attributed to the fact that Pb is the main chemical species contributing to the chemical profile for indoor and outdoor industrial factors. As discussed previously, the sampling site is close to an industrial area, with the Hal Far industrial zone located SW of the receptor site at 1.5 km. Steel fabrication, hot dip galvanising and hot asphalt plants in this area could contribute to the site's Zn, Pb and Cd metals and EC. The outdoor industrial factor contributes 0.54 ng m^{-3} (5%) to the $PM_{2.5}$ concentration at the sampling site. The RF-SHAP contribution for this factor increases to 9%, equivalent to 0.98 μg m^{-3}. The outdoor factor highlights that the impact of industrial activity has a more pronounced effect on outdoor $PM_{2.5}$ at this receptor background site, and it has only a minimal impact on indoor $PM_{2.5}$ levels.

The outdoor traffic factor also has a chemical profile similar to the indoor one. Both indoor and outdoor factors are the main contributors to EC and nitrate. The outdoor factor is responsible for 82% of EC, which is higher than the indoor contribution and 41% of the nitrate. The outdoor traffic factor has lower concentrations of Fe, Cu, Zn, Cr and Sb than the indoor factor. The polar plots for the indoor and outdoor factors vary significantly (Figures 5.14 and 3.11, respectively), with the main contribution for the outdoor factor coming from a NW to E direction and at higher wind speeds. However, on closer inspection, the polar plot for the outdoor factor is very similar to the indoor polar plot for the summer season shown in Figure D.2. Given that the main chemical species of this factor were estimated from indoor

concentrations and considering the better estimations during the summer seasons, the outdoor traffic factor shows a bias toward the summer's contribution and distribution of the indoor factor.

This factor was also downgraded from 7% to 2% in its percentage contribution to the outdoor $PM_{2.5}$ concentration in the RF-SHAP analysis. This fact highlights a disadvantage in using RF to augment the data obtained by PMF. A downgrading of the contribution by a source factor does not have a meaningful explanation and shows that other factors were slightly overestimated. This could be partly due to the low number of data points available, as machine learning algorithms tend to perform better with much bigger datasets. Given the very low importance of the traffic factor in the RF model, a low SHAP contribution is expected and can also be attributed to an intrinsic characteristic of how SHAP values are calculated. The SHAP value of a variable value is not the difference of the predicted value after removing the variable from the model. Given the set of variable values, the SHAP value is the contribution of a variable to the difference between the actual prediction and the mean prediction (Molnar 2022).

5.4 Conclusion

In this study, an RF-SHAP model was integrated with the PMF model and was utilized to investigate the influences of main driving factors on indoor $PM_{2.5}$ concentrations at an urban background site in Malta, using a comprehensive one-year dataset. The intricate nonlinear interactions between $PM_{2.5}$ and its driving factors, including emission sources and meteorological conditions, pose challenges in accurately assessing the impact of each individual factor on the variations in indoor $PM_{2.5}$ concentrations. The RF model correctly modelled the variations in indoor $PM_{2.5}$ (CV RMSE at 1.43 $\mu g\ m^{-3}$, an R^2 of 0.80), showing that during the sampling period, 54% of the driving factors were due to emission sources and 45% attributed to meteorological variables. Indoor and outdoor temperatures were demonstrated to be the most important meteorological variables. The 2D-plots showed that high indoor and outdoor temperatures are the driving factors behind high indoor $PM_{2.5}$ concentrations.

This study demonstrates that the RF regression coupled with SHAP can deliver accurate analyses for quantifying the important drivers of air pollution drivers across a range of

scenarios. The high importance of indoor and outdoor temperatures as predictor variables in this RF model and their high SHAP contributions suggest that these may be used as predictor variables to model indoor $PM_{2.5}$ and can be important predictor variables to model indoor PM concentrations in naturally ventilated houses where outdoor sources significantly influence indoor $PM_{2.5}$ concentrations.

This study also used an RF-SHAP model with good performance indicators (CV RMSE at 2.79 $\mu g\ m^{-3}$, an R^2 of 0.80) to calculate the outdoor contribution and augment the outdoor PMF analysis, with a dataset with only 159 observations. The transboundary contribution was found to be higher for outdoor $PM_{2.5}$ (58%) than indoor $PM_{2.5}$ (33%) and was attributed to the limited effect of Saharan dust events on indoor $PM_{2.5}$ when the windows are closed. The local anthropogenic contribution at this background site is significant in indoor (25%) and outdoor (33%) $PM_{2.5}$, mainly due to the high contributions of shipping and fireworks during the warmer period, which also coincides with greater infiltration indoors due to higher AERs. Outdoor contributions from natural sources are much higher than indoor ones, given that sea salt has higher contributions during the colder period, while some Saharan episodes also take place during the colder months when AERs are lower.

Shipping has a considerable effect, with an 18% contribution to outdoor $PM_{2.5}$. Establishing an emissions-controlled area is anticipated to mitigate pollutant emissions from maritime vessels within the Mediterranean region, which could significantly impact fine PM concentrations at this urban background site. By 2025, the Mediterranean Sea is set to be designated as an Emission Control Area (ECA) for sulfur oxides, which is projected to reduce $PM_{2.5}$ levels resulting from maritime emissions through the formation of secondary particles (Directorate-General for Mobility and Transport 2022 Dec 16).

The WHO annual Air Quality Guideline (AQG) level of 5 $\mu g\ m^{-3}$ (World Health Organization 2021) for $PM_{2.5}$ applies to both outdoor and indoor environments. The mean outdoor $PM_{2.5}$ concentrations at the receptor site are more than twice (10.8 $\mu g\ m^{-3}$) the WHO AQG level. The 24-hour limit of 15 $\mu g\ m^{-3}$ was exceeded 26 times, five of which fireworks were responsible for the high $PM_{2.5}$ levels. Again, local emissions from fireworks and emissions from shipping in the Mediterranean ought to be decreased in order to improve indoor air quality in coastal areas in Malta. On a local scale, the social practice of non-essential firework burning could

theoretically be readily managed to reduce exposure to toxic elements in both indoor and outdoor environments.

Chapter 6

Summary, Research Impacts and Future Work

6.1 Summary

This study addressed two primary aims to improve the understanding of indoor $PM_{2.5}$ in Malta. The first objective involved conducting long-term indoor air sampling to chemically characterize $PM_{2.5}$ and identify its main natural and anthropogenic sources at an urban background site using Positive Matrix Factorization (PMF). Through the analysis of quartz and PTFE filters, which were assessed gravimetrically and chemically for $PM_{2.5}$, 18 elements, 5 ions, organic carbon (OC), and elemental carbon (EC), the study identified eight factors influencing indoor $PM_{2.5}$. To augment the SA analysis, a PMF with RF-SHAP analysis was carried out to quantify the source contribution of outdoor $PM_{2.5}$ at the site. Another RF-SHAP analysis was integrated with the indoor PMF model and was utilized to investigate the influences of main driving factors on indoor $PM_{2.5}$ concentrations.

The second objective was to employ modern Machine Learning algorithms, specifically Random Forest (RF) and Extreme Gradient Boosting (XGBoost), to model and predict indoor $PM_{2.5}$ concentrations in several households in Malta and Gozo. Continuous PM sampling at seven residences provided data for modelling six-hourly averages of indoor PM2.5 concentrations. The models' performance was evaluated using RMSE and R^2 and identified key predictor variables influencing indoor $PM_{2.5}$ levels.

The summary highlights the main findings of this study for these two objectives:

PMF analysis of $PM_{2.5}$ at Birzebbuga (a residential urban background site)

- This study identified eight factors affecting indoor $PM_{2.5}$, including seven outdoor sources: fireworks (4%), Saharan dust (2%), traffic (10%), industrial emissions (2%), sea salt (9%), ammonium sulfate (31%), and shipping (10%). An additional indoor factor contributing to 26% of indoor $PM_{2.5}$ was also identified, with cooking and e-cigarette smoking being the primary contributors.
- During the summer months, anthropogenic contributions to indoor $PM_{2.5}$ reach 75%, with significant influences from secondary inorganic aerosols (SIAs), particularly ammonium sulfate, as well as traffic, shipping, and fireworks.

- The RF model accurately captured variations in indoor $PM_{2.5}$, with a CV RMSE of 1.43 μg m^{-3} and an R^2 of 0.80, revealing that 54% of the driving factors were linked to emission sources, while 45% were related to meteorological variables.
- Elevated indoor and outdoor temperatures were identified as key drivers of high indoor $PM_{2.5}$ concentrations. These were linked to higher outdoor source contributions during the warmer months and higher infiltration during the months due to increased AERs.
- The RF-SHAP model showed strong performance indicators (CV RMSE of 2.79 μg m^{-3} and an R^2 of 0.80) when used to estimate outdoor contributions and enhance the outdoor PMF analysis.
- Transboundary contribution is higher for outdoor $PM_{2.5}$ (58%) than indoor $PM_{2.5}$ (33%) due to the limited influence of Saharan dust events on indoor air quality when windows remain closed.
- The local anthropogenic contribution at this background site is significant in indoor (25%) and outdoor (33%) $PM_{2.5}$. The high indoor local anthropogenic contribution can be explained by the higher infiltration of $PM_{2.5}$ dust derived from these sources during the warmer period, which also coincides with greater emissions from fireworks.

Application of ML algorithms in predicting indoor residential $PM_{2.5}$ concentrations

- In sites with low indoor PM contributions (sites 2 to 6), the prediction of indoor $PM_{2.5}$ is mostly dependent on the outdoor PM_1 fraction.
- Low RMSE and high R^2 for the CV for both RF and XGBoost models and very high (>0.90) IOA on the testing data were reported for these sites.
- GLM also showed good performance indicators for testing data, but serial correlation at lag-1 was recorded.
- When the models were tested with the data from all sampling sites, the RF regression model gave the lowest prediction error, although the IOA decreased to 0.66.
- For site 7, indoor-generated $PM_{2.5}$ was the main contributor, and high concentrations were most likely due to cooking (frying and searing) in the absence of external hood extraction.
- At this site, the RF model used indoor RH as a good predictor variable for the very high indoor $PM_{2.5}$ concentrations recorded at this site.

6.2 Research Impacts

Generally, individuals spend most of their time indoors, and there is a limited understanding of the concentration and composition of indoor fine particulate matter in Malta. Although receptor modelling has been conducted in Malta, these studies have primarily concentrated on characterizing ambient air. Additionally, most investigations into the chemical speciation of fine dust associated with fireworks activities and related SA have been based on short-term sampling campaigns. This was the first study conducted in Malta using SA via PMF for outdoor and indoor $PM_{2.5}$ at an urban background site. Moreover, this study included the use of ML algorithms, RF and XGBoost, to predict indoor $PM_{2.5}$ in several residences. Hence, it presents a better picture of the exposure status to $PM_{2.5}$ among home dwellers in Malta.

As outlined in the conclusions, this study showed the significant impact of outdoor PM_1 on indoor $PM_{2.5}$ levels at sites with limited indoor fine PM sources. At sites with significant indoor generation from cooking, indoor $PM_{2.5}$ was 3.6 times the short-term (24-hour) AQG of the WHO. It would be ideal if proper extraction systems for domestic kitchens were to be regulated in order to minimise high exposure of home dwellers to indoor fine PM.

Considering that the anthropogenic contribution to indoor $PM_{2.5}$ during summer is 75%, being dominated by SIAs (mainly ammonium sulfate), traffic, shipping and fireworks, prioritising emission reductions from these sources is crucial for achieving indoor $PM_{2.5}$ concentrations below the WHO's annual AQG. The designation of the Mediterranean Sea as an Emission Control Area for sulfur oxides in 2025 will contribute to this end. Additionally, decreasing transboundary ammonia emissions from mainland Europe is necessary to mitigate the formation of particulate ammonium sulfate in Maltese residences. On a local level, the regulation of non-essential fireworks activity could significantly lessen the exposure to toxic elements within indoor residential settings; however, there seems to be a trend in the opposite direction.

6.3 Recommendations for future work

Given the limited number of residences sampled in this study, more data should be collected from flats/apartments/penthouses on different floors in multi-storey buildings, which are the types of habitation that have become prevalent in Malta. A representative sample of residence type and occupancy is needed to generate a model that can be used to extrapolate results to people living across different household characteristics in Malta.

The repeated measurement from each residence resulted in serial correlations in the GLM model. In this scenario, Time Series models that account for autocorrelation should be investigated. A shorter sampling period per site, not exceeding ten days, should be considered in further studies as it was challenging to find participants, especially when using bulky and noisy equipment. Moreover, the use of smaller and quieter PM spectrometers should be considered.

Disentangling the indoor contributions remains challenging. Measuring a vast array of chemical species over a long sampling campaign is expensive in terms of money and labour. This study included a novel use of RF-SHAP, where source contribution can be calculated by fitting an RF model to a PMF solution with a significant amount of unassigned mass. Such an approach can be applied to source contribution analysis, where the analytic fraction does not constitute a significant proportion of the PM mass. This approach can be applied to indoor PMF analysis where specific markers for indoor pollutants (such as POPs, cooking and combustion sources) are used. Such analyses result in PMF solutions with relatively high unassigned mass. RF-SHAP analysis on the PMF solution can be used to achieve the actual percentage source contributions by fitting the model on the observed indoor PM concentrations.

References

Abt E, Suh HH, Catalano P, Koutrakis P. 2000. Relative Contribution of Outdoor and Indoor Particle Sources to Indoor Concentrations. Environ Sci Technol 34:3579–3587.

Aglan HA. 2003. Predictive model for CO2 generation and decay in building envelopes. J Appl Phys 93:1287–1290.

AgriMetSoft. 2024. Index of Agreement. [accessed 2024 Jul 1]. AgriMetSoft.

Aksoyoglu S, Baltensperger U, Prévôt ASH. 2016. Contribution of ship emissions to the concentration and deposition of air pollutants in Europe. Atmos Chem Phys 16:1895–1906.

Almeida-Silva M, Faria T, Saraga D, Maggos T, Wolterbeek HT, Almeida SM. 2016a. Source apportionment of indoor PM10 in Elderly Care Centre. Environmental Science and Pollution Research 23:7814–7827.

Almeida-Silva M, Faria T, Saraga D, Maggos T, Wolterbeek HT, Almeida SM. 2016b. Source apportionment of indoor PM10 in Elderly Care Centre. Environmental Science and Pollution Research 23:7814–7827.

Amato F, Alastuey A, Karanasiou A, Lucarelli F, Nava S, Calzolai G, Severi M, Becagli S, Gianelle VL, Colombi C, et al. 2016. AIRUSE-LIFE+: A harmonized PM speciation and source apportionment in five southern European cities. Atmos Chem Phys 16:3289–3309.

Amato F, Moreno T, Pandolfi M, Querol X, Alastuey A, Delgado A, Pedrero M, Cots N. 2010. Concentrations, sources and geochemistry of airborne particulate matter at a major European airport. Journal of Environmental Monitoring 12:854–862.

Amato F, Pandolfi M, Moreno T, Furger M, Pey J, Alastuey A, Bukowiecki N, Prevot ASH, Baltensperger U, Querol X. 2011. Sources and variability of inhalable road dust particles in three European cities. Atmos Environ 45:6777–6787.

Anake WU, Nnamani EA. 2022. A Brief Review on Emerging Indoor Chemical Pollutants. In: IOP Conference Series: Earth and Environmental Science. Vol. 1054. Institute of Physics.

APA. 2024. U.S. Fireworks Consumption Figures 2000−2023. [accessed 2024 Oct 26]. https://www.americanpyro.com/industry-facts-figures.

Aquilina NJ, Camilleri SF. 2022. Impact of daily household activities on indoor PM2.5 and Black Carbon concentrations in Malta. Build Environ 207:108422.

Arata C, Zarzana KJ, Misztal PK, Liu Y, Brown SS, Nazaroff WW, Goldstein AH. 2018. Measurement of NO_3 and N_2O_5 in a Residential Kitchen. Environ Sci Technol Lett 5:595–599.

Araujo JA. 2011. Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis. Air Qual Atmos Health 4:79–93.

Azzopardi F, Ellul R, Presti M, Scollo S, Coltelli M. 2013. The effect of Etna volcanic ash clouds on the Maltese Islands. Journal of Volcanology and Geothermal Research 260:13–26.

Bari MA, Kindzierski WB, Wheeler AJ, Héroux MÈ, Wallace LA. 2015. Source apportionment of indoor and outdoor volatile organic compounds at homes in Edmonton, Canada. Build Environ 90:114–124.

Barman SC, Singh R, Negi MPS, Bhargava SK. 2008. Ambient air quality of Lucknow City (India) during use of fireworks on Diwali Festival. Environ Monit Assess 137:495–504.

Barnes J. 2013. Assessment of Air Quality in the Vicinity of Delimara Power Station, Malta.

Barraza F, Jorquera H, Valdivia G, Montoya LD. 2014. Indoor PM2.5 in Santiago, Chile, spring 2012: Source apportionment and outdoor contributions. Atmos Environ 94:692–700.

Barwick VJ, Ellison SLR, Fairman B. 1999. Estimation of uncertainties in ICP-MS analysis: A practical methodology. Anal Chim Acta 394:281–291.

Basaran B, Yılmaz Civan M. 2021. Investigating of primary components and source apportionment of persistent organic pollutants of indoor dust. International Journal of Environmental Science and Technology 18:2145–2160. [accessed 2024 Jul 17]. https://link-springer-com.ejournals.um.edu.mt/article/10.1007/s13762-020-02973-w.

Belis C a, Larsen BR, Amato F, Haddad I El, Favez O, Harrison RM, Hopke PK, Nava S, Paatero P, Prévôt A, et al. 2014. Air Pollution Source Apportionment with Receptor Models.

Belis CA, Karagulian F, Larsen BR, Hopke PK. 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos Environ 69:94–108.

Bo M, Salizzoni P, Clerico M, Buccolieri R. 2017. Assessment of Indoor-Outdoor Particulate Matter Air Pollution: A Review. Atmosphere (Basel) 8:136.

Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmos Environ 151:1–11.

Brown JS, Gordon T, Price O, Asgharian B. 2013. Thoracic and respirable particle definitions for human health risk assessment. Part Fibre Toxicol 10:12.

Brown SG, Eberly S, Paatero P, Norris GA. 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Science of the Total Environment 518–519:626–635.

Brunekreef B, Holgate ST. 2002. Air pollution and health. The Lancet 360:1233–42.

Camilleri R. 2008. A first study on airborne fine dust emissions from fireworks in Malta [Unpublished Dissertation for the degree of Master of Science]. Unpublished Dissertation for the degree of Master of Science at the University of Malta.

Camilleri R, Vella AJ. 2010. Effect of fireworks on ambient air quality in Malta. Atmos Environ 44:4521–4527.

Camilleri R, Vella AJ. 2016. Emission Factors for Aerial Pyrotechnics and Use in Assessing Environmental Impact of Firework Displays: Case Study from Malta. Propellants, Explosives, Pyrotechnics 41:273–280.

Carn SA, Fioletov VE, McLinden CA, Li C, Krotkov NA. 2017. A decade of global volcanic SO2 emissions measured from space. Sci Rep 7:44095.

Carrion-Matta A, Kang CM, Gaffin JM, Hauptman M, Phipatanakul W, Koutrakis P, Gold DR. 2019. Classroom indoor PM2.5 sources and exposures in inner-city schools. Environ Int 131:104968. [accessed 2021 Oct 31].

Carslaw DC, Ropkins K. 2012. OpenAir and R package for air quality data analysis. Environmental Modelling and Software 27–28:52–61.

Cassar J, Vella AJ. 2003. Methodology to identify badly weathering limestone using geochemistry : case study on the lower globigerina limestone of the Maltese Islands. Quarterly Journal of Engineering Geology and Hydrogeology 36:85–96.

Celo V, Dabek-Zlotorzynska E, Mathiew D, Okonskaia I. 2010. Validation of a Simple Microwave-Assisted Acid Digestion Method Using Microvessels for Analysis of Trace Elements in Atmospheric PM2.5 in Monitoring and Fingerprinting Studies. The Open Chemical and Biomedical Methods Journal 3:143–152.

CEN. 2014. EN 12341: Ambient air: Standard gravimetric measurement method for the determination of the PM10 or PM2.5 mass concentration of suspended particulate matter.

Cesari Daniela, Amato F, Pandolfi M, Alastuey A, Querol X, Contini D. 2016. An inter-comparison of PM10 source apportionment using PCA and PMF receptor models in three European sites. Environmental Science and Pollution Research 23:15133–15148.

Cesari Daniela, Donateo A, Conte M, Contini D. 2016. Inter-comparison of source apportionment of PM10 using PMF and CMB in three sites nearby an industrial area in central Italy. Atmos Res 182:282–293.

Cesari D., Giangreco A, Donateo A, Conte M, Merico E, Giangreco F, Contini D. 2016. An inter-comparison of PM2.5 at urban and urban background sites: Chemical characterization and source apportionment. Atmos Res 174–175:106–119.

Chao CY, Cheng EC. 2003. Source Apportionment of Indoor PM2.5 and PM10 in Homes. Indoor and Built Environment 11:27–37.

Charron A, Polo-Rehn L, Besombes JL, Golly B, Buisson C, Chanut H, Marchand N, Guillaud G, Jaffrezo JL. 2019. Identification and quantification of particulate tracers of exhaust and non-exhaust vehicle emissions. Atmos Chem Phys 19:5187–5207.

Chen G, Li S, Knibbs LD, Hamm NAS, Cao W, Li T, Guo J, Ren H, Abramson MJ, Guo Y. 2018. A machine learning method to estimate PM2.5 concentrations across China with remote

sensing, meteorological and land use information. Science of the Total Environment 636:52–60.

Chen J, Lu J, Ning J, Yan Y, Li S, Zhou L. 2019. Pollution characteristics, sources, and risk assessment of heavy metals and perfluorinated compounds in PM2.5 in the major industrial city of northern Xinjiang, China. Air Qual Atmos Health. [accessed 2024 Jul 17]. https://doi.org/10.1007/s11869-019-00706-8.

Chow JC, Watson JG. 1999. Ion Chromatography in elemental analysis of airborne particles. Elemetal Analysis of Airborne Particles 1:97–137.

Clewell RA, Merrill EA, Narayanan L, Gearhart JM, Robinson PJ. 2004. Evidence for Competitive Inhibition of Iodide Uptake by Perchlorate and Translocation of Perchlorate into the Thyroid. Int J Toxicol 23:17–23.

Contini Daniele, Cesari D, Donateo A, Chirizzi D, Belosi F. 2014. Characterization of PM10 and PM2.5 and Their Metals Content in Different Typologies of Sites in South-Eastern Italy. 10:435–453.

Contini D., Cesari D, Genga A, Siciliano M, Ielpo P, Guascito MR, Conte M. 2014. Source apportionment of size-segregated atmospheric particles based on the major water-soluble components in Lecce (Italy). Science of the Total Environment 472:248–261.

Council directive 2008/50/EC. 2008. Council directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Union. [accessed 2024 Jul 16]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008L0050.

Dai H, Liu Y, Wang J, Ren J, Gao Y, Dong Z, Zhao B. 2023. Large-scale spatiotemporal deep learning predicting urban residential indoor PM2.5 concentration. Environ Int 182:108343.

Diapouli E, Manousakas M, Vratolis S, Vasilatou V, Maggos T, Saraga D, Grigoratos T, Argyropoulos G, Voutsa D, Samara C, et al. 2017. Evolution of air pollution source contributions over one decade, derived by PM10 and PM2.5 source apportionment in two metropolitan urban areas in Greece. Atmos Environ 164:416–430.

Directorate-General for Mobility and Transport. 2022 Dec 16. New shipping fuel standards to reduce sulphur oxides in the Mediterranean by 80%. [accessed 2024 Aug 3]. https://environment.ec.europa.eu/news/new-shipping-fuel-standards-reduce-sulphur-air-pollutants-mediterranean-80-2022-12-16_en.

Du X, Yang J, Xiao Z, Tian Y, Chen K, Feng Y. 2021. Source apportionment of PM 2.5 during different haze episodes by PMF and random forest method based on hourly measured atmospheric pollutant. Environmental Science and Pollution Research 28:66978–66989.

Ellul R, Azzoparidi F, Smyth A, Saliba M, Azzopardi M. 2013. Shipping's Black Cloud. https://www.um.edu.mt/think/shippings-black-cloud/.

ERA. 2019. Justification report on the contribution of natural events to the PM10 limit values for 2018. Elsevier Ltd.

ERA. 2020. Justification report on the contribution of natural events to the PM10 limit values for 2019. Elsevier Ltd.

European Environmental Agency. 2018. European Union emission inventory report 1990-2017.

Fenech S, Aquilina NJ. 2020. Trends in ambient ozone, nitrogen dioxide, and particulate matter concentrations over the Maltese Islands and the corresponding health impacts. Science of the Total Environment 700:134527.

Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J. 2015. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos Environ 107:118–128. [accessed 2024 Jul 20].

Friedman JH, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33:1–22.

Fromme H. 2012. Particles in the Indoor Environment. In: Air Quality - Monitoring and Modeling. InTech.

Galindo N, Nicolás JF, Yubero E, Caballero S, Pastor C, Crespo J. 2008. Factors affecting levels of aerosol sulfate and nitrate on the Western Mediterranean coast. Atmos Res 88:305–313. [accessed 2021 Nov 5].

Gietl JK, Lawrence R, Thorpe AJ, Harrison RM. 2010. Identification of brake wear particles and derivation of a quantitative tracer for brake dust at a major road. Atmos Environ 44:141–146.

Gold EB, Blount BC, O'neill Rasor M, Lee JS, Alwis U, Srivastav A, Kim K. 2013. Thyroid hormones and thyroid disease in relation to perchlorate dose and residence near a superfund site. J Expo Sci Environ Epidemiol 23:399–408. https://doi.org/10.1038/jes.2012.90.

Gu J, Yang B, Brauer M, Zhang KM. 2021. Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models. Atmos Environ 246. https://doi.org/10.1016/j.atmosenv.2020.118125.

Gulati S, Bansal A, Pal A, Mittal N, Sharma A, Gared F. 2023. Estimating PM2.5 utilizing multiple linear regression and ANN techniques. Sci Rep 13, 22578. https://doi.org/10.1038/s41598-023-49717-7.

Hasheminassab S, Daher N, Shafer MM, Schauer JJ, Delfino RJ, Sioutas C, Del RJ, Sioutas C. 2014. Chemical characterization and source apportionment of indoor and outdoor fine particulate matter (PM2.5) in retirement communities of the Los Angeles Basin. Science of the Total Environment 490:528–537.

Hester R.E., Harrison RM, Querol X, editors. 2016. Airborne Particulate Matter: Sources, Atmospheric Process and Health. The Royal Society of Chemistry.

Hicks R, Caldas LQ, Dare PR, Hewitt PJ. 1986. Cardiotoxic and bronchoconstrictor effects of industrial metal fumes containing barium. Arch Toxicol Suppl 9:416–20.

Hopke PK, Dai Q, Li L, Feng Y. 2020. Global review of recent source apportionments for airborne particulate matter. Science of the Total Environment 740:140091. https://doi.org/10.1016/j.scitotenv.2020.140091.

Hortellani MA, Viera SantanaL, Sarkis JES, Azvedo J, Ulrich JC. 2011. ESTIMATE OF UNCERTAINTY OF MEASUREMENT IN THE DETERMINATION OF Zn, Cu, Hg AND Pb IN FISH MUSCLE. INAC 2011.

Hou L, Dai Q, Song C, Liu Bowen, Guo F, Dai T, Li L, Liu Baoshuang, Bi X, Zhang Y, et al. 2022. Revealing Drivers of Haze Pollution by Explainable Machine Learning. Environ Sci Technol Lett 9:112–119. [accessed 2024 Jul 20]. https://doi.org/10.1021/acs.estlett.1c00865.

Hu Y, Zhao B. 2022. Indoor sources strongly contribute to exposure of Chinese urban residents to PM2.5 and NO2. J Hazard Mater 426.

Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, Gu D, Liu Y. 2018. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environmental Pollution 242:675–683.

Huang Y, Du W, Chen Y, Shen G, Su S, Lin N, Shen H, Zhu D, Yuan C, Duan Y, et al. 2017. Household air pollution and personal inhalation exposure to particles (TSP/PM2.5/PM1.0/PM0.25) in rural Shanxi, North China. Environmental Pollution 231:635–643.

IARC. 1989. Antimony trioxide and antimony trisulfide. France.

IARC. 2015. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Outdoor Air Pollution.

Infrastructure Malta. 2024 Mar 15. Shore-to-Ship. https://www.infrastructuremalta.com/projects/shore-ship. [accessed 2024 Aug 3]. https://www.infrastructuremalta.com/projects/shore-ship.

Jalkanen JP, Johansson L, Kukkonen J. 2016. A comprehensive inventory of ship traffic exhaust emissions in the European sea areas in 2011. Atmos Chem Phys 16:71–84.

James G. 2013. An introduction to statistical learning: with applications in R. Springer.

Jiang Q, Sun YL, Wang Z, Yin Y. 2015. Aerosol composition and sources during the Chinese Spring Festival: Fireworks, secondary aerosol, and holiday effects. Atmos Chem Phys 15:6023–6034.

Joharestani MZ, Cao C, Ni X, Bashir B, Talebiesfandarani S. 2019. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere (Basel) 10.

Johnson TM, Guttikunda S, Wells GJ, Artaxo P, Bond TC, Russell AG, Watson JG, J. W. 2011. Tools for Improving Air Quality. Washington DC, USA.

Joly A, Smargiassi A, Kosatsky T, Fournier M, Dabek-Zlotorzynska E, Celo V, Mathieu D, Servranckx R, D'amours R, Malo A, et al. 2010. Characterisation of particulate exposure during fireworks displays. Atmos Environ 44:4325–4329.

Jorquera H, Barraza F, Heyer J, Valdivia G, Schiappacasse LN, Montoya LD. 2018. Indoor PM2.5 in an urban zone with heavy wood smoke pollution: The case of Temuco, Chile. Environmental Pollution 236:477–487.

Kang K, Kim H, Kim DD, Lee YG, Kim T. 2019. Characteristics of cooking-generated PM10 and PM2.5 in residential buildings with different cooking and ventilation types. Science of The Total Environment 668:56–66. [accessed 2024 Aug 1].

Karthikeyan S, Joshi UM, Balasubramanian R. 2006. Microwave assisted sample preparation for determining water-soluble fraction of trace elements in urban airborne particulate matter: Evaluation of bioavailability. Anal Chim Acta 576:23–30.

Keuken MP, Moerman M, Voogt M, Blom M, Weijers EP, Röckmann T, Dusek U. 2013. Source contributions to PM2.5 and PM10 at an urban background and a street location. Atmos Environ 71:26–35.

Kim E, Hopke PK. 2008. Source characterization of ambient fine particles at multiple sites in the Seattle area. Atmos Environ 42:6047–6056.

Kim KH, Kabir E, Kabir S. 2015. A review on the human health impact of airborne particulate matter. Environ Int 74:136–143.

Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzwe P, Behar J V, Hern SC, Engelmann WH. 2001. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. J Expo Sci Environ Epidemiol 11:231–252. https://doi.org/10.1038/sj.jea.7500165.

Krasnov H, Katra I, Novack V, Vodonos A, Friger MD. 2015. Increased indoor PM concentrations controlled by atmospheric dust events and urban factors. Build Environ 87:169–176.

Kravchenko J, Darrah TH, Miller RK, Lyerly HK, Vengosh A. 2014. A review of the health impacts of barium from natural and anthropogenic exposure. Environ Geochem Health 36:797–814.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. J Stat Softw 28:1–26.

Kulkarni P, Chellam S, Flanagan JB, Jayanty RKM. 2007. Microwave digestion-ICP-MS for elemental analysis in ambient airborne fine particulate matter: Rare earth elements and validation using a filter borne fine particle certified reference material. Anal Chim Acta 599:170–176.

Kumar V, Sahu M, Biswas P. 2022. Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms. Aerosol Air Qual Res 22:210240.

Lai AM, Carter E, Shan M, Ni K, Clark S, Ezzati M, Wiedinmyer C, Yang X, Baumgartner J, Schauer JJ. 2019. Chemical composition and source apportionment of ambient, household, and personal exposures to PM2.5in communities using biomass stoves in rural China. Science of the Total Environment 646:309–319.

Larson T, Gould T, Simpson C, Liu LJS, Claiborn C, Lewtas J. 2004. Source apportionment of indoor, outdoor, and personal PM2.5in seattle, washington, using positive matrix factorization. J Air Waste Manage Assoc 54:1175–1187.

Lee PKH, Brook JR, Dabek-Zlotorzynska E, Mabury SA. 2003. Identification of the Major Sources Contributing to PM2.5 Observed in Toronto. Environ Sci Technol 37:4831–4840.

Lesar TT, Filipčić A. 2021. The Hourly Simulation of PM2.5 Particle Concentrations Using the Multiple Linear Regression (MLR) Model for Sea Breeze in Split, Croatia. Water Air Soil Pollut 232.

Li J, Posfai M, Hobbs P V., Buseck PR. 2003. Individual aerosol particles from biomass burning in southern Africa : 2 , Compositions and aging of inorganic particles. 108:1–12.

Li N, Wang M, Nel A, Sioutas C, Cho A, Schmitz D, Misra C, Froines J, Sempf J, Oberley T. 2003. Ultrafine particulate pollutants induce oxidative stress and mitochondrial damage. Environ Health Perspect 111:455–460.

Li T, Cao S, Fan D, Zhang Yaqun, Wang B, Zhao X, Leaderer BP, Shen G, Zhang Yawei, Duan X. 2016. Household concentrations and personal exposure of PM2.5among urban residents using different cooking fuels. Science of the Total Environment 548–549:6–12.

Li T, Zhang Q, Peng Y, Guan X, Li L, Mu J, Wang X, Yin X, Wang Q. 2023. Contributions of various driving factors to air pollution events: Interpretability analysis from Machine learning perspective. Environ Int 173.

Li Y, Sha Z, Tang A, Goulding K, Liu X. 2023. The application of machine learning to air pollution research: A bibliometric analysis. Ecotoxicol Environ Saf 257.

Lundberg S, Lee S-I. 2017 May 22. A Unified Approach to Interpreting Model Predictions.

Ma J, Yu Z, Qu Y, Xu J, Cao Y. 2020. Application of the xgboost machine learning method in pm2.5 prediction: A case study of shanghai. Aerosol Air Qual Res 20:128–138.

Macneill M, Kearney J, Wallace L, Gibson M, Héroux ME, Kuchta J, Guernsey JR, Wheeler AJ. 2014. Quantifying the contribution of ambient and indoor-generated fine particles to indoor air in residential environments. Indoor Air 24:362–375.

Malta Tourism Authority. 2023. History of Maltese Fireworks. www.maltafireworksfestival.com. [accessed 2024 Aug 20]. www.maltafireworksfestival.com/history-of-maltese-fireworks.

Manousakas M., Papaefthymiou H, Diapouli E, Migliori A, Karydas AG, Bogdanovic-Radovic I, Eleftheriadis K. 2017a. Assessment of PM2.5 sources and their corresponding level of uncertainty in a coastal urban area using EPA PMF 5.0 enhanced diagnostics. Science of the Total Environment 574:155–164.

Manousakas Manousos-ioannis, Papaefthymiou H, Diapouli E, Migliori A, Karydas AG, Bogdanovic-Radovic I, Eleftheriadis K. 2017. Assessment of PM2.5 sources and their corresponding level of uncertainty in a coastal urban area using EPA PMF 5.0 enhanced diagnostics. Sci. Total Environ. 574:155–164.

Manousakas M., Papaefthymiou H, Diapouli E, Migliori A, Karydas AG, Bogdanovic-Radovic I, Eleftheriadis K. 2017b. Assessment of PM2.5 sources and their corresponding level of

uncertainty in a coastal urban area using EPA PMF 5.0 enhanced diagnostics. Science of the Total Environment 574:155–164.

Manousakas M, Papaefthymiou H, Eleftheriadis K, Katsanou K. 2014. Determination of water-soluble and insoluble elements in PM2.5 by ICP-MS. Science of the Total Environment 493:694–700.

Masiol M, Harrison RM. 2014. Aircraft engine exhaust emissions and other airport-related contributions to ambient air pollution: A review. Atmos Environ 95:409–455.

Matthaios VN, Liu M, Li L, Kang CM, Vieira CLZ, Gold DR, Koutrakis P. 2021. Sources of indoor PM2.5 gross α and β activities measured in 340 homes. Environ Res 197.

Mazzei F, D'Alessandro A, Lucarelli F, Nava S, Prati P, Valli G, Vecchi R. 2008. Characterization of particulate matter sources in an urban environment. Science of the Total Environment 401:81–89.

Men Y, Li Y, Luo Z, Jiang K, Yi F, Liu X, Xing R, Cheng H, Shen G, Tao S. 2023. Interpreting Highly Variable Indoor PM2.5 in Rural North China Using Machine Learning. Environ Sci Technol 57:18183–18192.

MIA. 2019. Annual Statistical Summary, Malta International Airport.

Minguillón MC, Schembari A, Triguero-Mas M, de Nazelle A, Dadvand P, Figueras F, Salvado JA, Grimalt JO, Nieuwenhuijsen M, Querol X. 2012. Source apportionment of indoor, outdoor and personal PM2.5 exposure of pregnant women in Barcelona, Spain. Atmos Environ 59:426–436.

Moller D. 1990. The Na/Cl ratio in rainwater and the seasalt chloride cycle. Tellus 42B:254–262.

Molnar C. 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. christophm.github.io/interpretable-ml-book/.

Molnár P, Johannesson S, Quass U. 2014. Source apportionment of PM2.5 using positive matrix factorization (PMF) and PMF with factor selection. Aerosol Air Qual Res 14:725–733.

Moreno T, Querol X, Alastuey A, Cruz Minguillón M, Pey J, Rodriguez S, Vicente Miró J, Felis C, Gibbons W. 2007. Recreational atmospheric pollution episodes: Inhalable metalliferous particles from firework displays. Atmos Environ 41:913–922.

Mukherjee A, Stanton LG, Graham AR, Roberts PT. 2017. Assessing the utility of low-cost particulate matter sensors over a 12-week period in the Cuyama valley of California. Sensors 17. https://doi.org/10.3390/s17081805.

Murty OP. 2000. Diwali toxicity. Journal of Forensic Medicine & Toxicology 17:23–26.

National Pollutant Inventory. 1999. Emission Estimation Technique Manual for Galvanizing.

National Statistics Office. 2021. Census of Population and Housing 2021 - Dwelling Characteristics. [accessed 2024 Aug 1]. https://nso.gov.mt/wp-content/uploads/Census-2021-Volume-2.pdf.

National Statistics Office. 2023. Regional Statistics Malta. [accessed 2024 Mar 1]. https://nso.gov.mt/themes_publications/regional-statistics-malta-2023-edition/.

Neas LM. 2000. Fine particulate matter and cardiovascular disease. Fuel Processing Technology 65:55–67.

Norris G, Duvall R, Brown S, Bai S. 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide.

Ovadnevaite J, Ceburnis D, Canagaratna M, Berresheim H, Bialek J, Martucci G, Worsnop DR, Dowd CO. 2012. On the effect of wind speed on submicron sea salt mass concentrations and source fluxes On the effect of wind speed on submicron sea salt mass concentrations and source fluxes. J Geophys Res 117.

Paatero P, Hopke PK. 2003. Discarding or downweighting high-noise variables in factor analytic models. In: Analytica Chimica Acta. Vol. 490. Elsevier. p. 277–289.

Pace C, Vella AJ. 2016. Escalation of Fireworks in Malta: Environmental Forensic Evidence from Perchlorate in Dust Fall. Xjenza 4:4–10.

Pace C, Vella AJ. 2019. Contamination of water resources of a small island state by fireworks-derived perchlorate: A case study from Malta. Environmental Pollution 250:475–481.

Pan B. 2018. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. In: IOP Conference Series: Earth and Environmental Science. Vol. 113. Institute of Physics Publishing.

Pandolfi M, Gonzalez-Castanedo Y, Alastuey A, de la Rosa JD, Mantilla E, de la Campa AS, Querol X, Pey J, Amato F, Moreno T. 2011. Source apportionment of PM10 and PM2.5 at multiple sites in the strait of Gibraltar by PMF: Impact of shipping emissions. Environmental Science and Pollution Research 18:260–269.

Pant P, Harrison RM. 2013. Estimation of the contribution of road traffi c emissions to particulate matter concentrations from fi eld measurements: A review. Atmos Environ 77:78–97.

Pio C, Cerqueira M, Harrison RM, Nunes T, Mirante F, Alves C, Oliveira C, Sanchez de la Campa A, Artíñano B, Matos M. 2011. OC/EC ratio observations in Europe: Re-thinking the approach for apportionment between primary and secondary organic carbon. Atmos Environ 45:6121–6132.

Polissar A V., Hopke PK, Paatero P, Malm WC, Sisler JF. 1998. Atmospheric aerosol over Alaska 2. Elemental composition and sources. Journal of Geophysical Research Atmospheres 103:19045–19057.

Qian J, Peccia J, Ferro AR. 2014. Walking-induced particle resuspension in indoor environments. Atmos Environ 89:464–481.

Qing YM, Turpin BJ, Korn L, Weisel CP, Morandi M, Colome S, Junfeng Z, Stock T, Spektor D, Winer A, et al. 2005. Influence of ambient (outdoor) sources on residential indoor and personal PM2.5 concentrations: Analyses of RIOPA data. J Expo Anal Environ Epidemiol 15:17–28.

Querol X, Alastuey A, Puicercus JA, Mantilla E, Ruiz CR, Lopez-Soler A, Plana F, Juan R. 1998. Seasonal evolution of suspended particles around a large coal-fired power station: Chemical characterization. Atmos Environ 34:719–731.

Reff A, Eberly SI, Bhave P V. 2007. Receptor Modeling of Ambient Particulate Matter Data Using Positive Matrix Factorization : Review of Existing Methods. 57:146–154.

Reinmann C, Filzmoser P, Garrett R, Dutter R. 2008. Statistical Data Analysis Explained - Applied Environmental Statistics with R. John Wiley & Sons.

Rivas I, Viana M, Moreno T, Bouso L, Pandolfi M, Alvarez-Pedrerol M, Forns J, Alastuey A, Sunyer J, Querol X. 2015. Outdoor infiltration and indoor contribution of UFP and BC, OC, secondary inorganic ions and metals in PM2.5in schools. Atmos Environ 106:129–138.

Rodríguez S, Querol X, Alastuey A, Plana F. 2002. Sources and processes affecting levels and composition of atmospheric aerosol in the western Mediterranean. Journal of Geophysical Research Atmospheres 107. https://doi.org/10.1029/2001JD001488.

Ruprecht AA, De Marco C, Saffari A, Pozzi P, Mazza R, Veronese C, Angellotti G, Munarini E, Ogliari AC, Westerdahl D, et al. 2017. Environmental pollution and emission factors of electronic cigarettes, heat-not-burn tobacco products, and conventional cigarettes. Aerosol Science and Technology 51:674–684.

Ryan PH, Brokamp C, Fan Z-H, Rao MB. 2015. Analysis of Personal and Home Characteristics Associated with the Elemental Composition of PM2.5 in Indoor, Outdoor, and Personal Air in the RIOPA Study. Res Rep Health Eff Inst:3–40.

Saffari A, Daher N, Ruprecht A, De Marco C, Pozzi P, Boffi R, Hamad SH, Shafer MM, Schauer JJ, Westerdahl D, et al. 2014. Particulate metals and organic compounds from electronic and tobacco-containing cigarettes: Comparison of emission rates and secondhand exposure. Environmental Sciences: Processes and Impacts 16:2259–2267.

Salameh D, Detournay A, Pey J, Pérez N, Liguori F, Saraga D, Bove MC, Brotto P, Cassola F, Massabò D, et al. 2015. PM2.5 chemical composition in five European Mediterranean cities: A 1-year study. Atmos Res 155. https://doi.org/10.1016/j.atmosres.2014.12.001.

Salameh D, Pey J, Bozzetti C, El Haddad I, Detournay A, Sylvestre A, Canonaco F, Armengaud A, Piga D, Robin D, et al. 2018. Sources of PM2.5 at an urban-industrial Mediterranean city, Marseille (France): Application of the ME-2 solver to inorganic and organic markers. Atmos Res 214:263–274.

Salvador P, Artíñano B, Querol X, Alastuey A. 2007. A combined analysis of backward trajectories and aerosol chemistry to characterise long-range transport episodes of particulate matter: The Madrid air basin, a case study. Sci Total Environ 390:495-506.

Saraga D, Duarte RMBO, Manousakas MI, Maggos T, Tobler A, Querol X. 2024. From outdoor to indoor air pollution source apportionment: Answers to ten challenging questions. TrAC - Trends in Analytical Chemistry 178.

Saraga D, Querol X, Duarte RMBO, Aquilina NJ, Canha N, Alvarez EG, Jovasevic-Stojanovic M, Bekö G, Byčenkienė S, Kovacevic R, et al. 2023. Source apportionment for indoor air pollution: Current challenges and future directions. Science of the Total Environment 900.

Saraga DE, Maggos TE, Sfetsos A, Tolis EI, Andronopoulos S, Bartzis JG, Vasilakos C. 2010. PAHs sources contribution to the air quality of an office environment: Experimental results and receptor model (PMF) application. Air Qual Atmos Health 3:225–234.

Scerri MM, Genga A, Iacobellis S, Delmaire G, Giove A, Siciliano M, Siciliano T, Weinbruch S. 2019. Investigating the plausibility of a PMF source apportionment solution derived using a small dataset : A case study from a receptor in a rural site in Apulia - South East Italy. Chemosphere 236:124376.

Scerri MM, Kandler K, Weinbruch S. 2016. Disentangling the contribution of Saharan dust and marine aerosol to PM10 levels in the Central Mediterranean. Atmos Environ 147:395–408.

Scerri MM, Kandler K, Weinbruch S, Yubero E, Galindo N, Prati P, Caponi L, Massabò D. 2018. Estimation of the contributions of the sources driving PM2.5levels in a Central Mediterranean coastal town. Chemosphere 211:465–481.

Schauer J, Lough GC, Shafer MM, Christensen WF, Arndt M, DeMinter JT, Park J-S. 2006. Characterization of Metals Emitted from Motor Vehicles.

Schembari C, Cavalli F, Cuccia E, Hjorth J, Calzolai G, Pérez N, Pey J, Prati P, Raes F. 2012. Impact of a European directive on ship emissions on air quality in Mediterranean harbours. Atmos Environ 61:661–669.

Seibert R, Nikolova I, Volná V, Krejčí B, Hladký D. 2020. Air pollution sources' contribution to PM2.5 concentration in the Northeastern Part of the Czech Republic. Atmosphere (Basel) 11.

Seinfeld JH, Pandis SN. 2016. Atmospheric Chemistry and Physics: from Air Pollution to Climate Change. 3rd ed. John Wiley & Sons, editor.

Semple S, Garden C, Coggins M, Galea KS, Whelan P, Cowie H, Sánchez-Jiménez A, Thorne PS, Hurley JF, Ayres JG. 2012. Contribution of solid fuel, gas combustion, or tobacco smoke to indoor air pollutant concentrations in Irish and Scottish homes. Indoor Air 22:212–223.

Shang J, Khuzestani RB, Tian J, Schauer JJ, Hua J, Zhang Yang, Cai T, Fang D, An J, Zhang Yuanxun. 2019. Chemical characterization and source apportionment of PM2.5 personal exposure of two cohorts living in urban and suburban Beijing. Environmental Pollution 246:225–236.

Shi Y, Du Z, Zhang J, Han F, Chen F, Wang D, Liu M, Zhang H, Dong C, Sui S. 2023. Construction and evaluation of hourly average indoor PM2.5 concentration prediction models based on multiple types of places. Front Public Health 11.

Song X, Polissar A V, Hopke PK. 2001. Sources of fine particle composition in the northeastern US. 35:5277–5286.

Spada M, Jorba O, Pérez García-Pando C, Janjic Z, Baldasano JM. 2013. Modeling and evaluation of the global sea-salt aerosol distribution: Sensitivity to emission schemes and resolution effects at coastal/orographic sites. Atmos Chem Phys 13:11735–11755.

SRM 1648A. 2015. Urban Particulate Matter; National Institute of Standards and Technology. U.S. Department of Commerce: Gaithersburg, MD, 18 September 2015.

Stacey B, Bush T. 2002. Preliminary Assessment of Ambient Air Quality in Malta.

Steinhauser G, Sterba JH, Foster M, Grass F, Bichler M. 2008. Heavy metals from pyrotechnics in New Years Eve snow. Atmos Environ 42:8616–8622.

Suryawanshi S, Chauhan AS, Verma R, Gupta T. 2016. Identification and quantification of indoor air pollutant sources within a residential academic campus. Science of the Total Environment 569–570:46–52.

Talebi SM, Abedi M. 2005. Determination of arsenic in air particulates and diesel exhaust particulates by spectrophotometry. J Environ Sci (China) 17:156–8.

Tang H, Chan WR, Sohn MD. 2021. Automating the interpretation of PM2.5 time-resolved measurements using a data-driven approach. Indoor Air 31:860–871.

Thorpe A, Harrison RM. 2008. Sources and properties of non-exhaust particulate matter from road traffic: A review. Science of the Total Environment 400:270–282.

Tian YZ, Wang J, Peng X, Shi GL, Feng YC. 2014. Estimation of the direct and indirect impacts of fireworks on the physicochemical characteristics of atmospheric PM10 and PM2.5. Atmos Chem Phys 14:9469–9479.

TM. 2018. Transport Malta Annual Report 2018.

Tofful L, Perrino C. 2015. Chemical composition of indoor and outdoor PM2.5 in three schools in the city of Rome. Atmosphere (Basel) 6:1422–1443.

Tsai HH, Chien LH, Yuan CS, Lin YC, Jen YH, Ie IR. 2012. Influences of fireworks on chemical characteristics of atmospheric fine and coarse particles during Taiwan's Lantern Festival. Atmos Environ 62:256–264.

Tunno BJ, Dalton R, Cambal L, Holguin F, Lioy P, Clougherty JE. 2016. Indoor source apportionment in urban communities near industrial sites. Atmos Environ 139:30–36.

von Uexküll O, Skerfving S, Doyle R, Braungart M. 2005. Antimony in brake pads-a carcinogenic component? J Clean Prod 13:19–31.

U.S.C. 1990. Clean Air Act. United States Congress.

Vecchi Roberta, Bernardoni V, Cricchio D, D'Alessandro A, Fermo P, Lucarelli F, Nava S, Piazzalunga A, Valli G. 2008. The impact of fireworks on airborne particles. Atmos Environ 42:1121–1132.

Vecchi R, Chiari M, Alessandro AD, Fermo P, Lucarelli F, Mazzei F. 2008. A mass closure and PMF source apportionment study on the sub-micron sized aerosol fraction at urban sites in Italy. 42:2240–2253.

Vella AJ, Chircop C, Micallef T, Pace C. 2015. Perchlorate in dust fall and indoor dust in Malta: An effect of fireworks. Science of the Total Environment 521–522:46–51.

Vella Ryan. 2019. A study on the dynamics of particulate matter infiltration in buildings. University of Malta.

Viana M, Amato F, Alastuey A, Moreno T, Gsrcia Dos Santos S, Herce MD, Fernandez-Patier R. 2009. Chemical Tracers of Particulate Emissions from Commercial Shipping. 43:7472–7477.

Viana Mar, Hammingh P, Colette A, Querol X, Degraeuwe B, Vlieger I de, van Aardenne J. 2014. Impact of maritime transport emissions on coastal air quality in Europe. Atmos Environ 90:96–105.

Viana M, Kuhlbusch TAJ, Querol X, Alastuey A, Harrison RM, Hopke PK, Winiwarter W, Vallius M, Szidat S, Prevot ASH, et al. 2008. Source apportionment of particulate matter in Europe : A review of methods and results. J Aerosol Sci 39:827–849.

Viana M., Pey J, Querol X, Alastuey A, de Leeuw F, Lükewille A. 2014. Natural sources of atmospheric aerosols influencing air quality across Europe. Science of the Total Environment 472:825–833.

Viana M, Querol X, Go T, Payo F, Maldonado JA, Ku N. 2007. Source apportionment of ambient PM 2 . 5 at five spanish centres of the european community respiratory health survey (ECRHS II). 41:1395–1406.

Wallace LA, Mitchell H, O'Connor GT, Neas L, Lippmann M, Kattan M, Koenig J, Stout JW, Vaughn BJ, Wallace D, et al. 2003. Particle concentrations in inner-city homes of children with asthma: The effect of smoking, cooking, and outdoor pollution. Environ Health Perspect 111:1265–1272.

Wang M, Beelen R, Basagana X, Becker T, Cesaroni G, De Hoogh K, Dedele A, Declercq C, Dimakopoulou K, Eeftens M, et al. 2013. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: The ESCAPE project. Environ Sci Technol 47:4357–4364.

Wang Y, Du Y, Fang J, Dong X, Wang Q, Ban J, Sun Q, Ma R, Zhang W, He MZ, et al. 2022. A Random Forest Model for Daily PM2.5 Personal Exposure Assessment for a Chinese Cohort. Environ Sci Technol Lett 9:466–472.

Wang Y, Koutrakis P, Michanikou A, Kouis P, Panayiotou AG, Kinni P, Tymvios F, Chrysanthou A, Neophytou M, Mouzourides P, et al. 2024. Indoor residential and outdoor sources of PM2.5 and PM10 in Nicosia, Cyprus. Air Qual Atmos Health 17:485–499.

Wang Y, Zhuang G, Xu C, An Z. 2007. The air pollution caused by the burning of fireworks during the lantern festival in Beijing. Atmos Environ 41:417–431.

World Health Organization. 2010. Selected Pollutants.

World Health Organization. 2016. Ambient air pollution: a global assessment of exposure and burden of disease.

World Health Organization. 2021. WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva.

Xu C, Xu D, Liu Z, Li Y, Li Ning, Chartier R, Chang J, Wang Q, Wu Y, Li Na. 2020. Estimating hourly average indoor PM2.5 using the random forest approach in two megacities, China. Build Environ 180.

Yang Y, Liu L, Xu C, Li N, Liu Z, Wang Q, Xu D. 2018. Source apportionment and influencing factor analysis of residential indoor PM2.5 in Beijing. Int J Environ Res Public Health 15:15–18.

Yao L, Yang L, Chen J, Toda K, Wang X, Zhang J, Yamasaki D, Nakamura Y, Sui X, Zheng L, et al. 2015. Levels, indoor-outdoor relationships and exposure risks of airborne particle-associated perchlorate and chlorate in two urban areas in Eastern Asia. Chemosphere 135:31–37.

Yao Z, Li J, Wu B, Hao X, Yin Y, Jiang X. 2015. Characteristics of PAHs from deep-frying and frying cooking fumes. Environmental Science and Pollution Research 22:16110–16120.

Yu W, Nakisa B, Loke SW, Stevanovic S, Guo Y, Rastgoo MN. 2024. Indoor PM2.5 forecasting and the association with outdoor air pollution: a modelling study based on sensor data in Australia. arXiv.org.

Yubero E, Carratalá A, Crespo J, Nicolás J, Santacatalina M, Nava S, Lucarelli F, Chiari M. 2011. PM10 source apportionment in the surroundings of the San Vicente del Raspeig cement plant complex in southeastern Spain. Environmental Science and Pollution Research 18:64–74.

Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, Naidan G, Ochir C, Legtseg B, Byambaa T, et al. 2019. Evaluation of random forest regression and multiple linear

regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environmental Pollution 245:746–753.

Zauli Sajani S, Marchesi S, Trentini A, Bacco D, Zigola C, Rovelli S, Ricciardelli I, Maccone C, Lauriola P, Cavallo DM, et al. 2018. Vertical variation of PM 2.5 mass and chemical composition, particle size distribution, NO 2 , and BTEX at a high rise building. Environmental Pollution 235:339–349.

Zhang J, Lance S, Freedman JM, Sun Y, Crandall BA, Wei X, James J. 2018. Fine particles from Independence Day fireworks events : chemical characterization and source apportionment. Atmos. Chem. Phys.:1–21.

Zhang Y, Vu T Van, Sun J, He J, Shen X, Lin W, Zhang X, Zhong J, Gao W, Wang Y, et al. 2020. Significant Changes in Chemistry of Fine Particles in Wintertime Beijing from 2007 to 2017: Impact of Clean Air Actions. Environ Sci Technol 54:1344–1352. [accessed 2024 Jul 20]. https://pubs-acs-org.ejournals.um.edu.mt/doi/full/10.1021/acs.est.9b04678.

Zhang Z, Gao Y, Yuan Q, Tan Y, Li H, Cui L, Huang Y, Cheng Y, Xiu G, Lai S, et al. 2020. Effects of indoor activities and outdoor penetration on PM2.5 and associated organic/elemental carbon at residential homes in four Chinese cities during winter. Science of the Total Environment 739.

Zhang Z, Xu B, Xu W, Wang F, Gao J, Li Y, Li M, Feng Y, Shi G. 2022. Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on PM2.5 pollution. Environ Res 212.

Zhao W, Hopke PK, Norris G, Williams R, Paatero P. 2006. Source apportionment and analysis on ambient and personal exposure samples with a combined receptor model and an adaptive blank estimation strategy. Atmos Environ 40:3788–3801.

Zhao Y, Li A, Gao R, Tao P, Shen J. 2014. Measurement of temperature, relative humidity and concentrations of CO, CO2 and TVOC during cooking typical Chinese dishes. Energy Build 69:544–561.

Zhu Y, Hinds WC, Krudysz M, Kuhn T, Froines J, Sioutas C. 2005. Penetration of freeway ultrafine particles into indoor environments. J Aerosol Sci 36:303–322.

Zwozdziak A, Sówka I, Zwozdziak J, Kwieci K. 2016. Influence of PM 1 and PM 2 . 5 on lung function parameters in healthy schoolchildren — a panel study. Environ Sci Pollut Res 23:23892–23901.

Appendix A: Chapter 2

Table A.1 Summary of indoor activities (from 159 sampling days).

Activity	Electric Oven	Use of Toaster	Frying	Open kitchen windows	Gas Heater	Dry Sweeping	Vacuum cleaner	e-cigarette smoking
Number of days	41	116	41	63	63	25	12	150
Percentage of sampling days	26	73	26	40	40	16	8	94

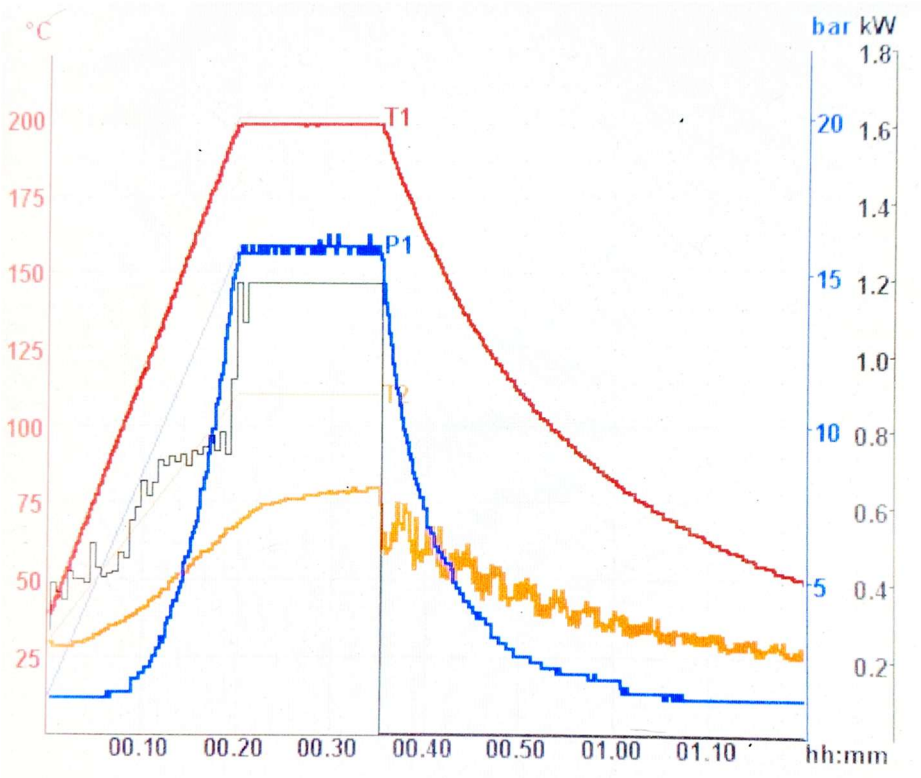


Figure A.1 Microwave digestion profile

Table A.2 Method detection limits (MDLs) for the $PM_{2.5}$ analyses in ng m^{-3} of sampled air.

Pollutant	MDL (ng m^{-3})	Pollutant	MDL (ng m^{-3})
OC	5.0	Cu	0.059
EC	5.0	Zn	0.82
Na^+	18.1	As	0.064
Mg	1.06	Sr	0.073
Al	3.99	Cd	0.016
K	13.00	Sb	0.021
Ca	26.91	Ba	0.012
Ti	0.24	Pb	0.017
V	0.015	Cl	22.7
Cr	0.066	NO^{3-}	16.1
Mn	0.048	SO_4^{2-}	54.4
Fe	0.78	NH_4^+	4.7
Ni	0.10		

Table A.3 Mean concentrations (µg/L) ± standard deviation (SD) for the Reagent blanks, Method blank, Indoor field blanks and Outdoor field blanks

Element	Reagent Blanks		Method Blanks		Indoor Field Blanks		Outdoor Field Blanks	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Na	-1.124	0.83	-1.678	1.49	-1.460	1.59	-1.005	2.19
Mg	0.948	0.44	0.154	1.42	-0.026	1.62	-0.204	1.72
Al	-3.587	4.44	-3.386	4.33	-2.646	3.75	-2.794	3.39
K	-0.849	7.88	0.873	6.93	-2.017	6.20	-2.252	7.07
Ca	-2.909	39.00	-23.536	56.95	-53.710	60.13	-49.878	60.97
Ti	-0.026	0.11	-0.040	0.12	-0.035	0.30	0.065	0.54
V	-0.007	0.00	-0.006	0.01	-0.002	0.02	-0.001	0.02
Cr	-0.196	0.02	0.011	0.04	0.002	0.08	0.037	0.10
Mn	-0.008	0.03	0.005	0.03	-0.023	0.04	-0.021	0.04
Fe	0.881	3.46	0.441	2.24	-0.776	0.74	-0.475	0.94
Ni	-0.111	0.34	0.067	0.38	-0.085	0.42	-0.340	0.37
Cu	-0.011	0.08	0.019	0.06	-0.023	0.16	0.056	0.08
Zn	0.018	1.58	0.893	3.52	-1.253	4.74	-0.284	3.54
As	-0.020	0.02	-0.012	0.02	-0.005	0.02	-0.007	0.02
Sr	0.078	0.08	-0.001	0.13	-0.051	0.14	-0.076	0.15
Cd	0.003	0.01	0.003	0.01	-0.002	0.00	-0.002	0.00
Sb	-0.001	0.01	-0.011	0.01	0.012	0.05	-0.001	0.03
Ba	-0.003	0.03	-0.039	0.06	-0.068	0.07	-0.064	0.06
Pb	-0.012	0.03	0.032	0.12	-0.044	0.02	-0.040	0.02

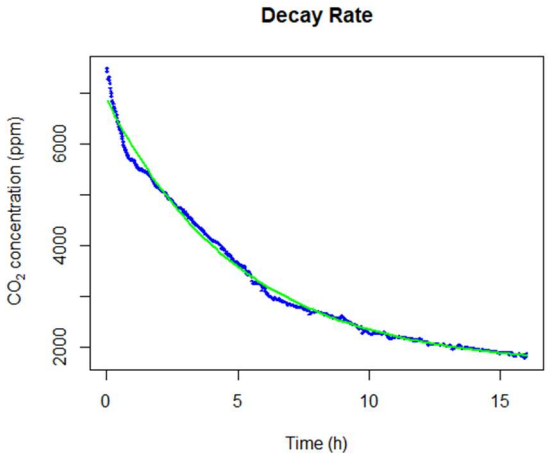
Table A.4 IC spike concentration and percentage recoveries

Ion	Standard Concentration	Spike concentration	Extracted concentration	Recovery efficiency	
	mg L^{-1}	mg L^{-1}	mg L^{-1}	%	SD
Cl^-	10	10	**0.10**	118	5.7
NO_3^-	20	20	**0.20**	104	2.3
SO_4^{2-}	20	20	**0.20**	125	1.7
Na^+	200	20	**0.20**	127	0.4
K^+	200	20	**0.20**	79	4.0
Mg^{2+}	200	20	**0.20**	95	10.7
Ca^{2+}	1000	100	**1.00**	75	18.7
NH_4^+	400	40	**0.40**	121	0.6

29/06/2019

Sampling room window closed

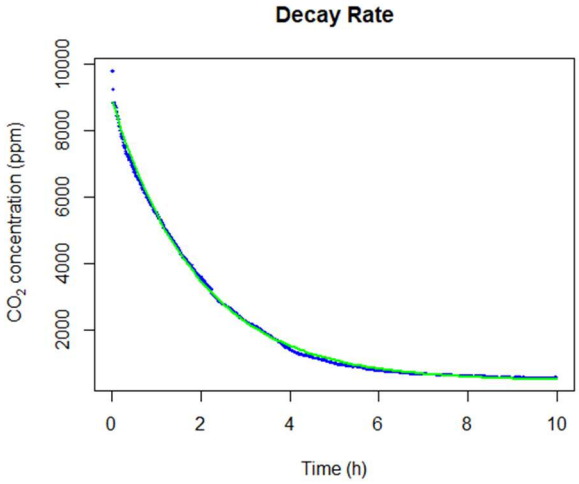
AER 0.20 h^{-1}



23/06/2019

Sampling room window closed

AER 0.52 h^{-1}



01/07/2019

Sampling room window closed

AER 1.62 h^{-1}

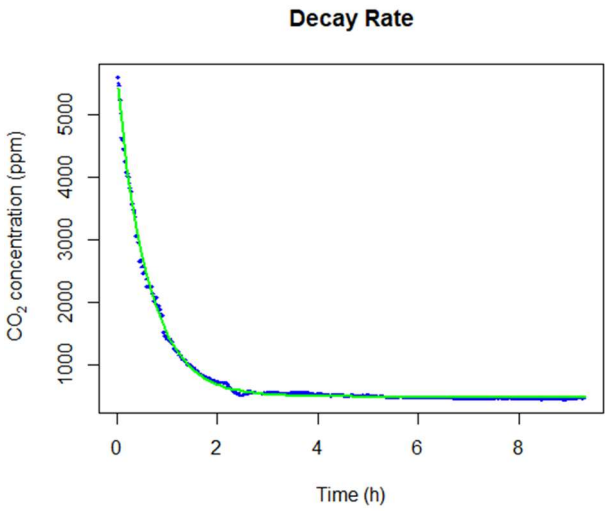


Figure A.2 Modelled and actual CO_2 concentrations and AER

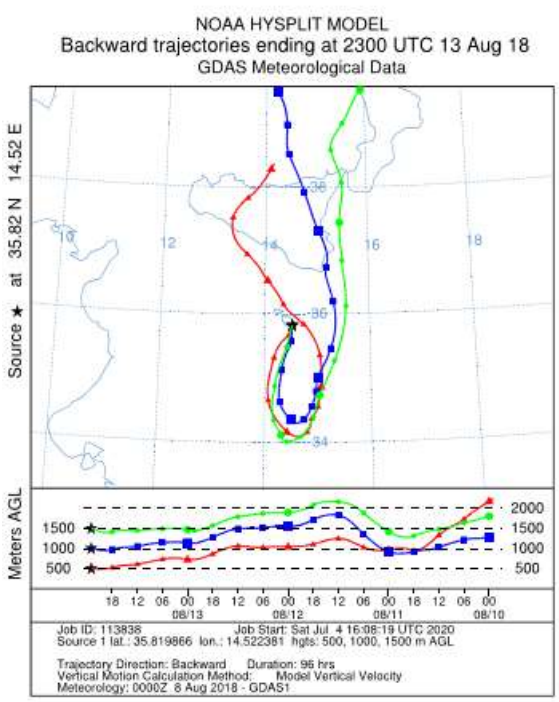
Table A.5 Sample of the logged of activities during the continuous indoor and outdoor PM measurements campaign

Date	Time	Duration	Location	Activity	Site
21/08/2023	18:45	1	Indoor	Construction work	3
31/08/2023	09:00	1	Indoor	Sweeping	3
03/09/2023	18:30	0.5	Outdoor	Yard sweeping	3
03/09/2023	19:00	1	Outdoor	Wood oven	3
08/09/2023	09:30	0.25	Indoor	Sweeping	3
13/10/2023	17:45	1	Indoor	Vacuum cleaner	3
15/10/2023	09:00	0.25	Indoor	Sweeping	3
25/11/2023	15:00	0.25	Indoor	Sweeping	3
26/11/2023	10:00	0.5	Indoor	Sweeping	3
26/11/2023	11:30	0.45	Indoor	Vacuum cleaner	3
07/12/2023			Outdoor	Fireworks	4
08/12/2023			Outdoor	Fireworks	4
13/12/2023	14:00	4	Indoor	Welding in Garage	4
29/12/2023	18:00	3	Indoor	Welding in Garage	4
31/12/2023	22:00	6	Indoor	Party inside the house	4

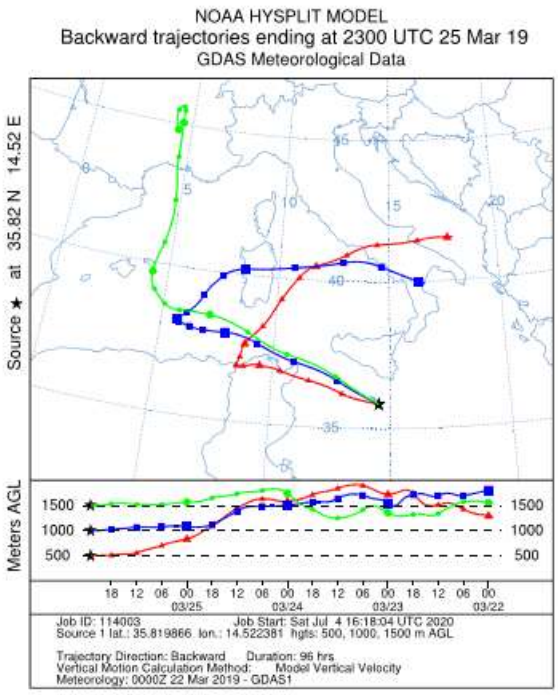
Appendix B: Chapter 3

Table B.1 Kruskal-Wallis rack sum tests and the Mann Whitney U-test results for the seasonal contributions of the some PMF factors.
For the fireworks factor, the difference between the average contributions between the periods June to September (fireworks burning) and October to May (no fireworks) was tested.

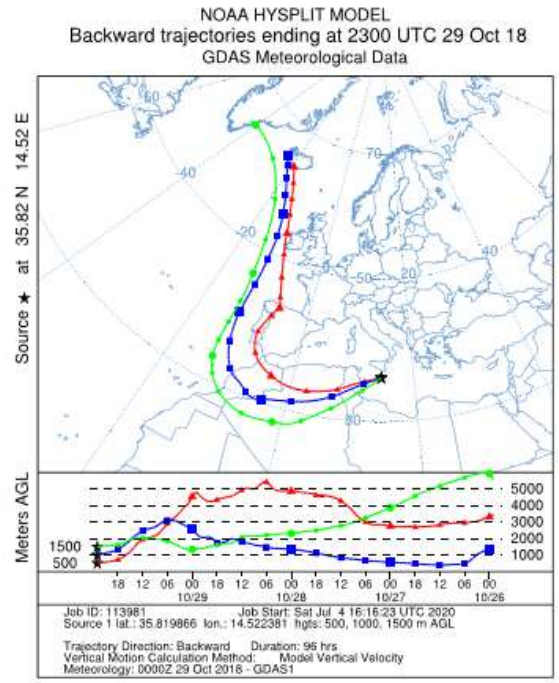
Indoor factor	**Kruskal-Wallis rank sum test (p-value)**	**0.000**		
	Mann Whitney U-test (p-value)	Autumn	Spring	Winter
	Spring	0.297	-	-
	Summer	0.000	0.002	-
	Winter	0.585	0.603	0.002
Traffic factor	Kruskal-Wallis rank sum test (p-value)	0.000		
	Mann Whitney U-test (p-value)	Autumn	Spring	Winter
	Spring	0.000	-	-
	Summer	0.000	0.000	-
	Winter	0.340	0.000	0.000
Ammonium sulfate factor (Percentage contributions)	Kruskal-Wallis rank sum test (p-value)	0.000		
	Mann Whitney U-test (p-value)	Autumn	Spring	Winter
	Spring	0.706	-	-
	Summer	0.001	0.001	-
	Winter	0.006	0.001	0.000
Fireworks factor (June to September and October to May periods)	Mann Whitney U-test (p-value)	0.000		
Unassigned fraction	Kruskal-Wallis rank sum test (p-value)	0.111		
	Mann Whitney U-test (p-value)	Autumn	Spring	Winter
	Spring	0.81	-	-
	Summer	0.21	0.16	-
	Winter	0.81	0.81	0.14



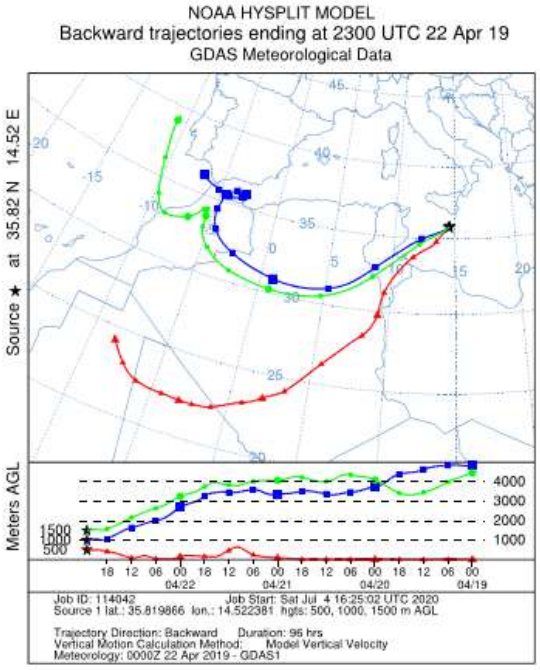
13/08/18



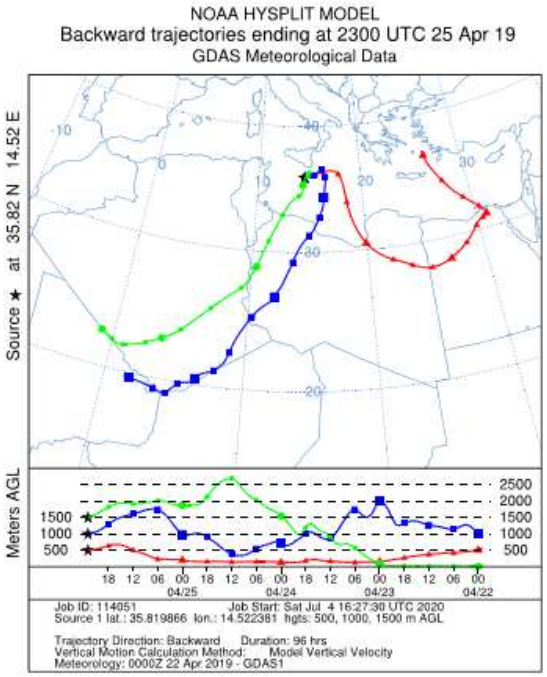
25/03/19



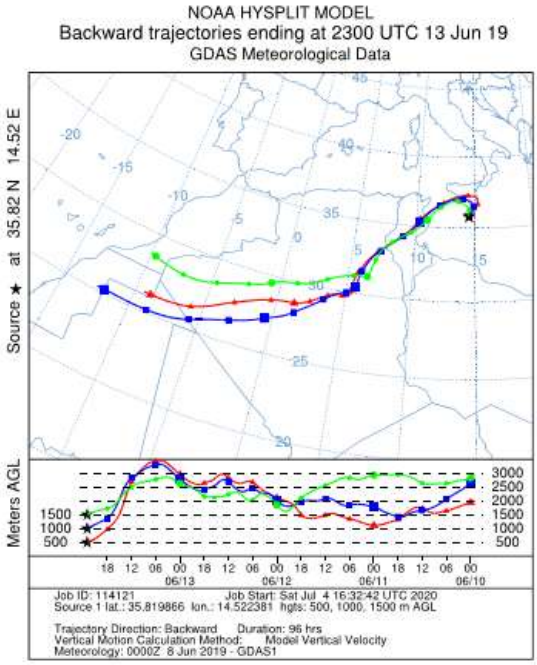
29/10/18



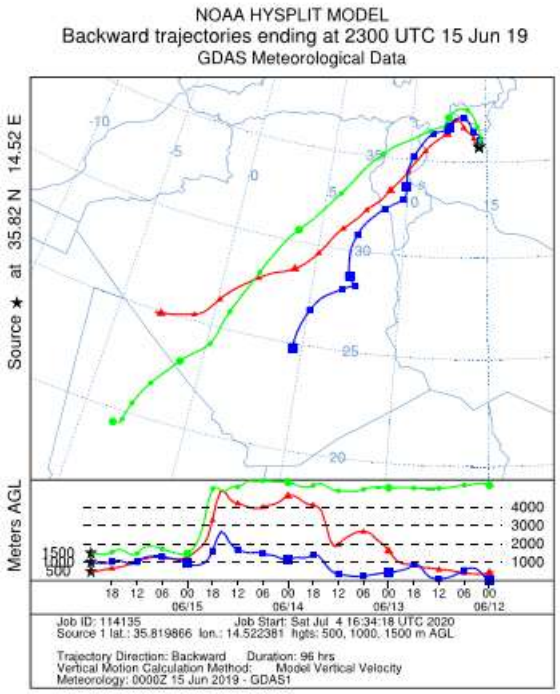
22/04/19



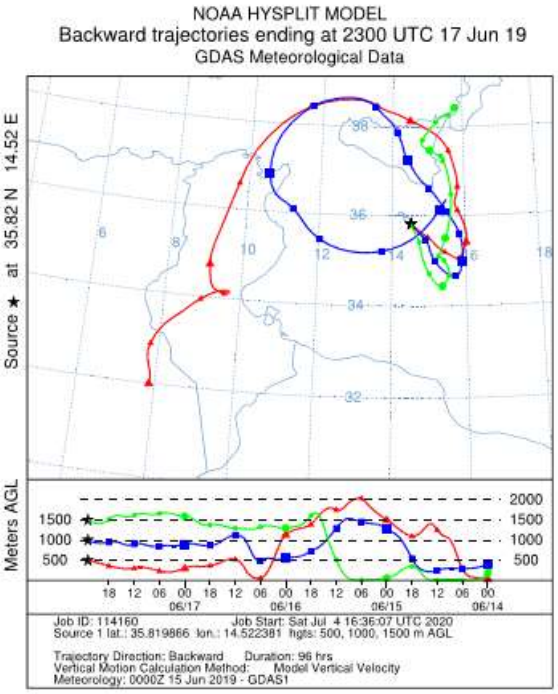
25/04/19



13/06/19



15/06/19



17/06/19

Figure B.1 NOAA HYSPLIT back trajectories

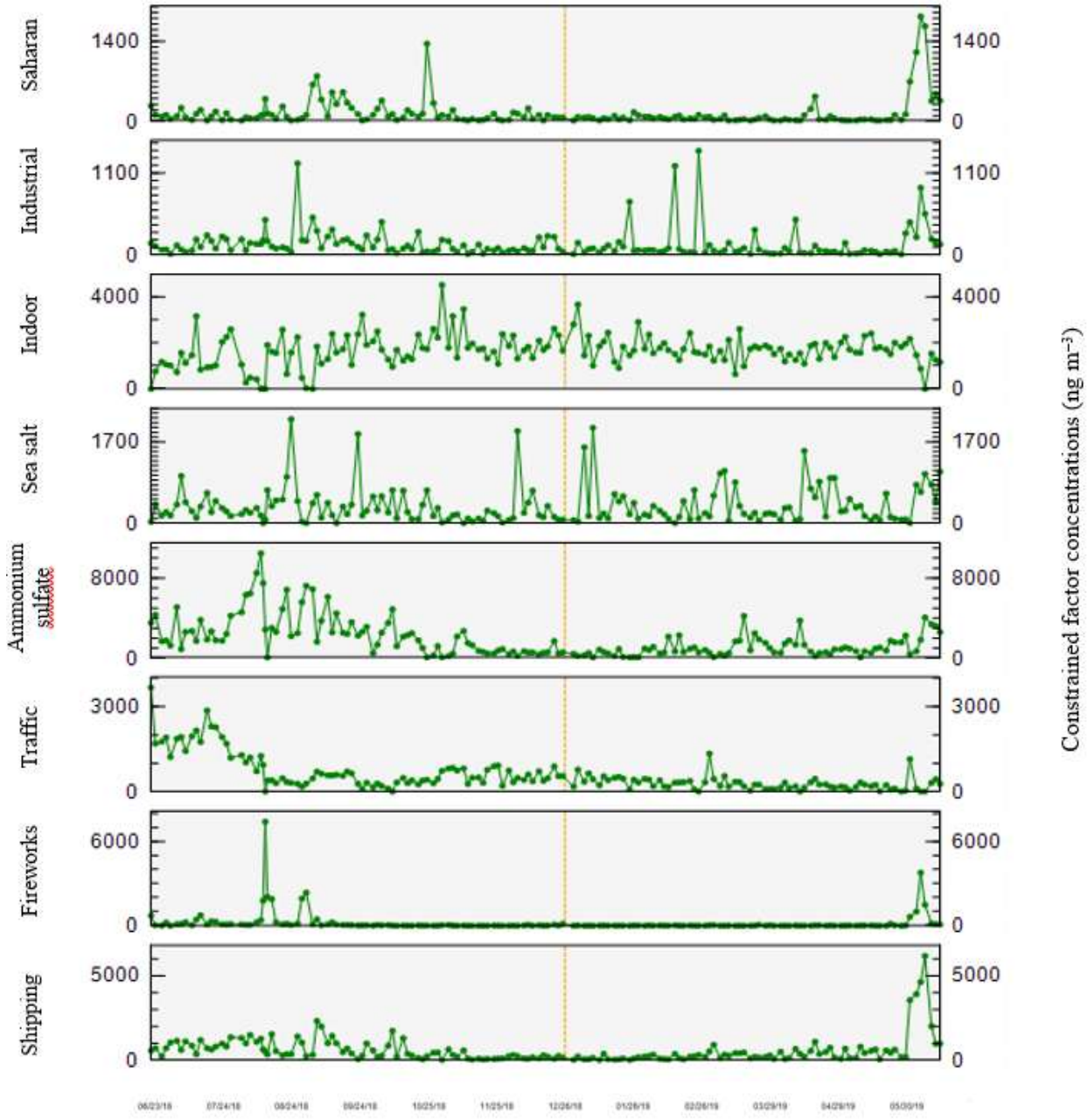


Figure B.2 Daily constrained factor contributions in ng m^{-3}.

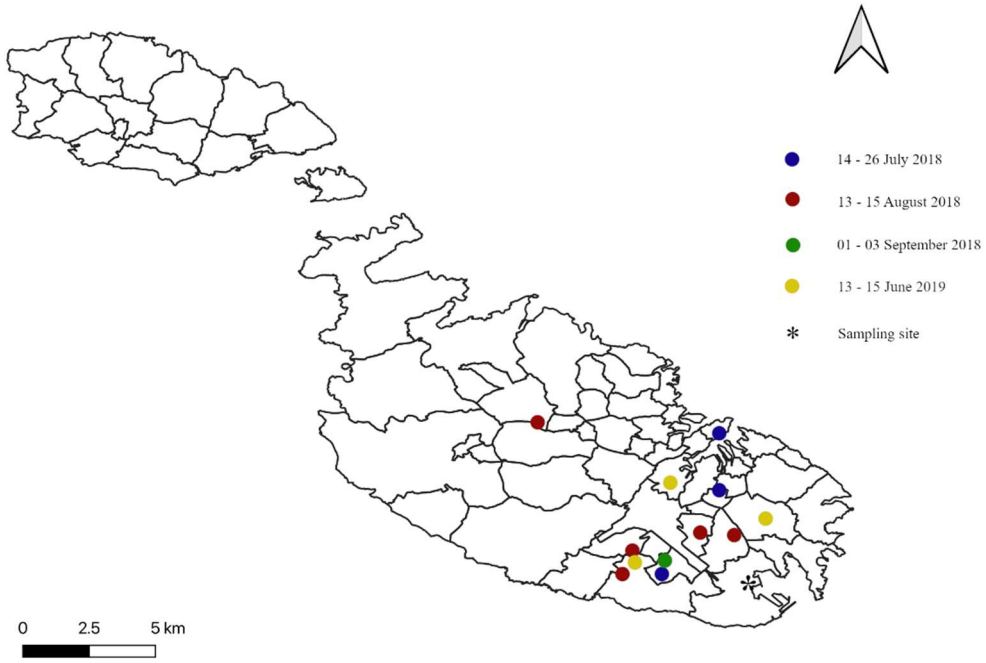


Figure B.3 Map showing the village feasts celebrated on the dates when high contributions were recorded from the indoor fireworks factor.

Appendix C: Chapter 4

Table C.1 Number of imputed hourly PM values per site due to instrument malfunction.

Site	2	3	4	5	6	7
Indoor PM	0	0	11	0	0	0
Outdoor PM	0	0	166	17	172	181

The "na.interp " function from the forecast package in RStudio was used to interpolate missing values in the time series data. The gaps were filled using the existing data structure, thus maintaining the inherent trend found in the data. This function uses a weighted average of the surrounding points, based on distance, to fill in the missing values. Except for Site 4, which had four days of missing outdoor data, all other imputations were carried out for 6 to 24-hour hours periods, translating into one to four 6-hour data points.

Table C.2 Details of the predictor variables for the models

Type	Predictor Variable	Abbreviation
Outdoor Particulate Matter	PM_1	pm1O
	$PM_{2.5}$	pm2.5O
	PM_{10}	pm10O
Meteorological parameters	Wind speed	ws
	Wind direction	wd
	Outdoor temperature	Temp.O
	Indoor temperature	Temp.I
	Outdoor humidity	Humidity.O
	Indoor humidity	Humidity.I
Lagged variables	Outdoor PM_1	pm1.L
	Outdoor $PM_{2.5}$	pm2.5.L
	Outdoor PM_{10}	pm10.L
	Wind speed	ws.L
	Wind direction	wd.L
	Outdoor temperature	Temp.O.L
	Indoor temperature	Temp.I.L
	Outdoor humidity	Humidity.O.L
	Indoor humidity	Humidity.I.L
Sampling information and time-based variables	Sampling sites	Each site as a categorical variable
	Sampling time	Time
	Sampling day	Day

Appendix D: Chapter 5

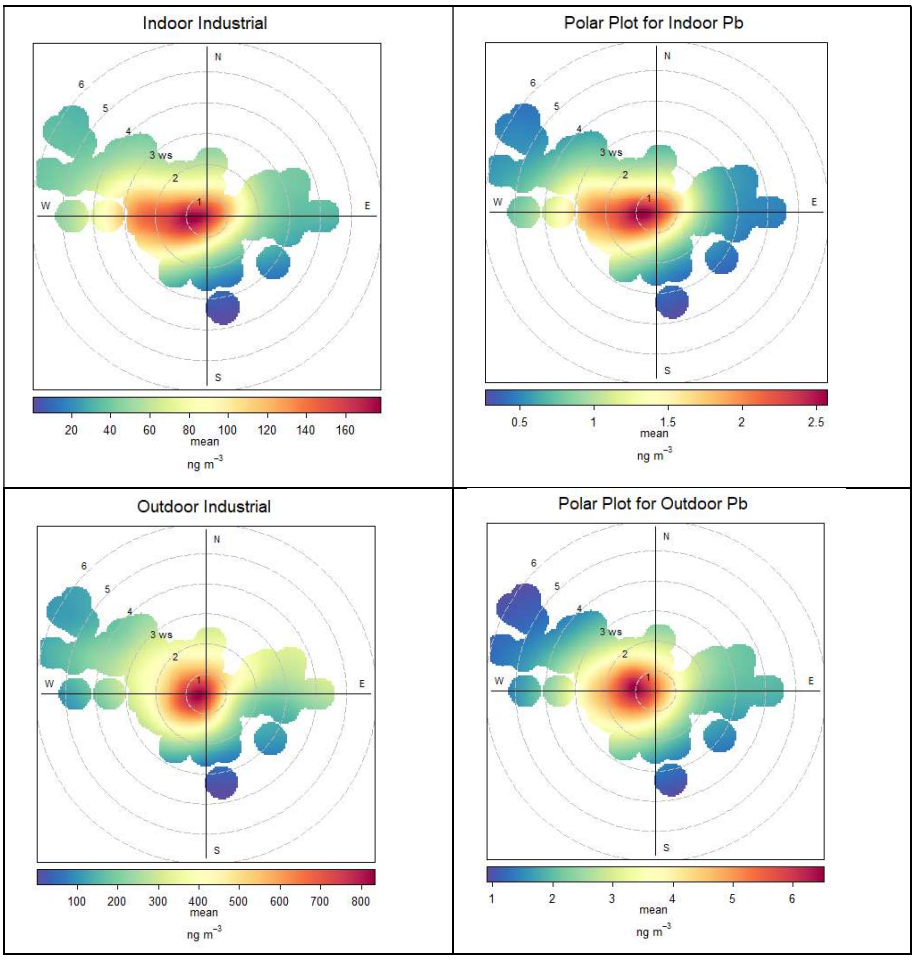


Figure D.1 Polar plots comparing indoor and outdoor industrial source profiles with indoor and outdoor Pb.

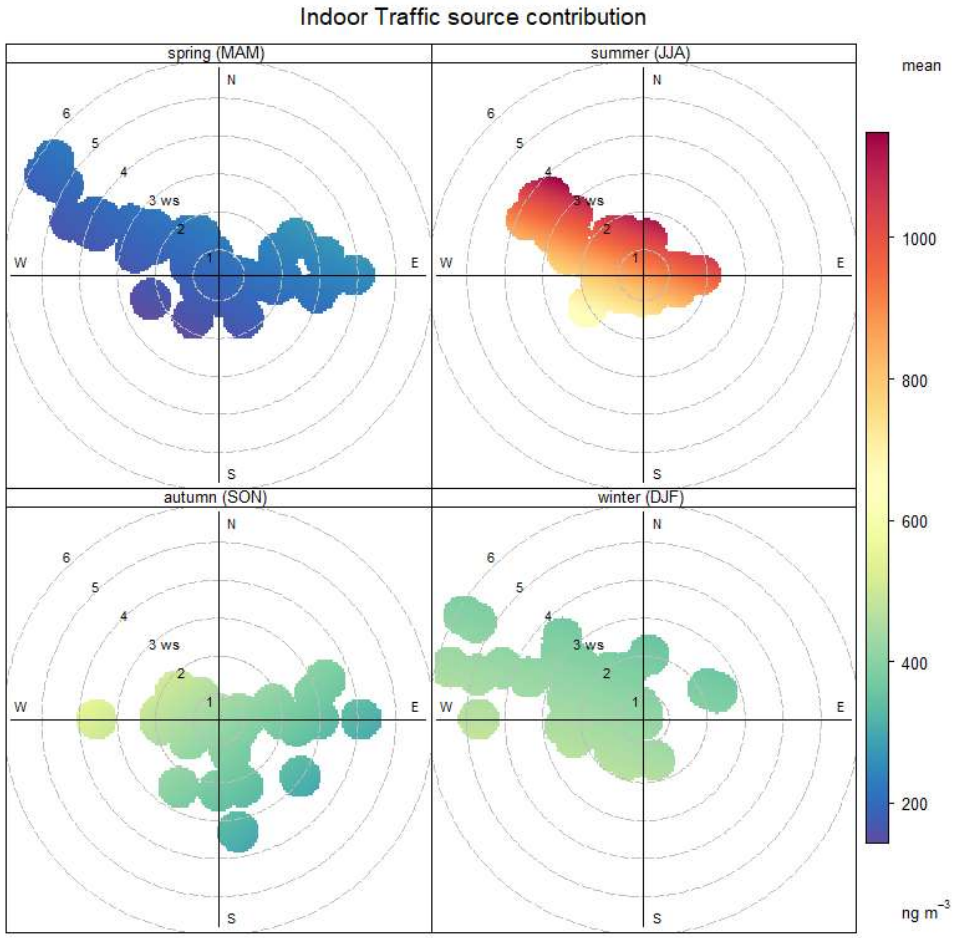


Figure D.2 Seasonal indoor traffic source contributions.

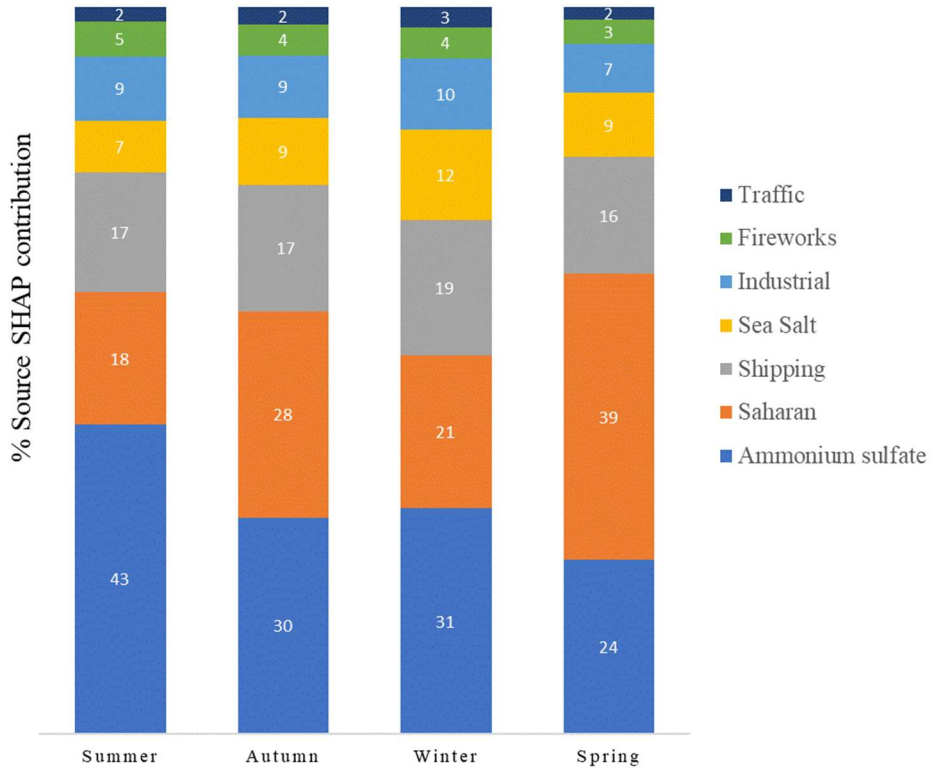


Figure D.3 Seasonal percentage SHAP source contributions to outdoor $PM_{2.5}$

Appendix E: Machine Learning R script

```
library(ggplot2)
library(ggpubr)
library(devtools)
library(caret)
library(randomForest)
library(Metrics)
library(rpart)
library(rfUtilities)
library(party)
library(psych)
library(xgboost)
library(dplyr)
library(mgcv)
library(pdp)
library(forecast)
library(openxlsx)

load(".RData")

# Create excel file
ML_data = createWorkbook() #creates xlsx workbook with openxlsx package
addWorksheet(ML_data, "predicted_6_Hourly")

combined_6hr <- read.csv(paste0(path, "/combined_6_hourly_6.csv"), header=T,sep=",")
combined_6hr$Date <-  as.POSIXct(combined_6hr$Date, format = "%d/%m/%Y %H")
combined_6hr$Site <- as.factor(combined_6hr$Site)  # converts site to factor


#Split a data set into training and testing
set.seed(123)
tolerance = 1
var_diff = Inf
while (abs(var_diff) > tolerance) {
  # Create train-test split
  trainIndex <- createDataPartition(combined_6hr$Site, p = 0.8, list = FALSE)
  training <- combined_6hr[trainIndex, ]
  testing <- combined_6hr[-trainIndex, ]
  var_train <- var(training$pm2.5I, na.rm = TRUE)
  var_test <- var(testing$pm2.5I, na.rm = TRUE)
  var_diff <- var_train - var_test
}



# Check proportions in the training set
prop.table(table(training$Site))
# Check proportions in the testing set
prop.table(table(testing$Site))
```

RF Model

```
# Train model
train_control_RF <- trainControl(
  method = "repeatedcv",              # Repeated Cross-Validation
  number = 10,                        # Number of folds
  repeats = 3,                        # Number of repeats
  search='grid',                      # Grid search
  verboseIter = TRUE,                 # Print progress
  allowParallel = TRUE,               # Allow parallel processing
  summaryFunction = defaultSummary  # Use default summary function for regression
)

tune_grid <- expand.grid(.mtry = (1:21))

modfit_RF <- train(pm2.5I ~ pm1O + pm10O +pm2.5O  + Temp.O +Temp.I + Humidity.I + Humidity.O + wd +
ws + Day + Time + Site + pm1O.L + pm10O.L +pm2.5O.L  + Temp.O.L +Temp.I.L + Humidity.I.L +
Humidity.O.L + wd.L + ws.L,
                   data = training,
                   method = "rf",
                   trControl = train_control_RF,              # Training control
                   tuneGrid = tune_grid,                      # Tuning grid
                   metric = "RMSE"                            # Evaluation metric (Root Mean Squared Error)
)

#RF model summary
print(modfit_RF)
plot(modfit_RF)

varImp(modfit_RF) #Importance of different predictors
plot(varImp(modfit_RF))

modfit_RF$results
summary(modfit_RF$finalModel)
```

XGBoost Model

```
# train model
control_XGboost <- trainControl(
  method='repeatedcv',
  number=10,
  repeats=3,
  search='grid',
  verboseIter = TRUE,
  allowParallel = TRUE,                    # Allow parallel processing
  summaryFunction = defaultSummary         # Use default summary function for regression
)
```

```
# Define the tuning grid
```

```
tune_grid <- expand.grid(
 nrounds = c(100, 150, 250, 500),            # Number of boosting rounds
  eta = c(0.01, 0.1, 0.25, 0.5),             # Learning rate
  max_depth = c(2, 3, 4, 5, 6),              # Maximum tree depth
  gamma = c(0, 2, 5),                        # Used for tuning of Regularization, default: 0
  colsample_bytree = 1,                      # Column sampling, default: 1
  min_child_weight = 1,                      # Minimum leaf weight, default: 1
  subsample = 1                              # Row sampling, default: 1
)


modfit_XGboost <- train(pm2.5I ~ pm1O + pm10O +pm2.5O  + Temp.O +Temp.I + Humidity.I + Humidity.O +
wd + ws + Day + Time + Site + pm1O.L + pm10O.L +pm2.5O.L  + Temp.O.L +Temp.I.L + Humidity.I.L +
Humidity.O.L + wd.L + ws.L,
                data = training,
                method = "xgbTree",                 # XGBoost with xgbTree
                trControl = control_XGboost,        # Training control
                tuneGrid = tune_grid,               # Tuning grid
                metric = "RMSE"                      # Evaluation metric (Root Mean Squared Error)
)

#XTboost model summary
print(modfit_XGboost)
plot(modfit_XGboost)
modfit_XGboost$bestTune
varImp(modfit_XGboost)
plot(varImp(modfit_XGboost))
modfit_XGboost$results
summary(modfit_XGboost$finalModel)
```

GLM Model

```
training_data <- training
testing_data <-  testing

# Scaling data, standardization using caret
preProcValues <- preProcess(training_data, method = c("scale"))
standardized_train <- predict(preProcValues, training_data)
# Apply the same transformation to testing data
standardized_test <- predict(preProcValues, testing_data)

train_control_LM <- trainControl(
  method = "repeatedcv",                    # Repeated Cross-Validation
  number = 10,                              # Number of folds
  repeats = 3,                              # Number of repeats
  verboseIter = TRUE,                        # Print progress
  allowParallel = TRUE,                     # Allow parallel processing
  summaryFunction = defaultSummary          # Use default summary function for regression
)
```

```
tune_grid <- expand.grid(
  alpha = 1,                                  # Elastic Net mixing parameter (0 for Ridge, 1 for Lasso)
  lambda = c(0.001, 0.01, 0.1,0.25, 0.5, 1)     # Regularization parameter
)

modfit_LM <- train(pm2.5I ~ pm1O + pm10O +pm2.5O  + Temp.O +Temp.I + Humidity.I + Humidity.O + wd
+ ws + Day + Time + Site + pm1O.L + pm10O.L +pm2.5O.L  + Temp.O.L +Temp.I.L + Humidity.I.L +
Humidity.O.L + wd.L + ws.L,
                data = standardized_train,
                method = "glmnet",                  # Generalized Linear Model with regularization
                family = "gaussian",
                trControl = train_control_LM,       # Training control
                tuneGrid = tune_grid,               # Tuning grid
                metric = "RMSE"                     # Evaluation metric (Root Mean Squared Error)
)

print(modfit_LM)
plot(modfit_LM)
varImp(modfit_LM) #Importance of different predictors
plot(varImp(modfit_LM))
modfit_LM$results
summary(modfit_LM$finalModel)
coefficients <- coef(modfit_LM$finalModel, modfit_LM$bestTune$lambda)
print(coefficients)
```

Appendix F: Indoor RF-SHAP R Script

```
library(doParallel)
library(ggplot2)
library(ggpubr)
library(devtools)
library(caret)
library(randomForest)
library(Metrics)
library(rpart)
library(rfUtilities)
library(party)
library(psych)
library(dplyr)
library(glmnet)
library(mgcv)
library(pdp)
library(ranger)
library(kernelshap)
library(shapviz)
library(openxlsx)

indoor <- read.csv(paste0(path, "/mydata.csv"), header=T,sep=",")
indoor$date <-  as.POSIXct(indoor$date, format = "%d/%m/%Y)
indoor$windows <- as.factor(indoor$windows)
indoor$frying <- as.factor(indoor$frying)
indoor$washing <- as.factor(indoor$washing)
indoor$Sahara <- as.factor(indoor$Sahara)

# train model
set.seed(123)
train_control_RF_PMF <- trainControl(
  method = "repeatedcv",                    # Repeated Cross-Validation
  number = 10,                              # Number of folds
  repeats = 3,                              # Number of repeats
  search='grid',                            # Grid search
  verboseIter = TRUE,                       # Print progress
  allowParallel = TRUE,                     # Allow parallel processing
  summaryFunction = defaultSummary          # Use default summary function for regression
)
tune_grid <- expand.grid(.mtry = (1:15))

modfit_RF_PMF <- train(pm2.5I ~ ws + wd + Indoor + Saharan + Sea.Salt + Ammonium.Sulfate + Traffic +
                         Fireworks + Industrial + Shipping + Temp.O + Temp.I + Humidity.O + Humidity.I +
                         Pressure + windows + frying,
                       data = indoor,
                       method = "rf",
                       trControl = train_control_RF,            # Training control
                       tuneGrid = tune_grid,                    # Tuning grid
                       metric = "RMSE"                          # Evaluation metric (Root Mean Squared Error)
)
```

```
# Testing model
p_RF_PMF<-predict(modfit_RF_PMF, newdata =indoor)
cor(indoor$pm2.5I,p_RF_PMF)
R2 <- (cor(indoor$pm2.5I,p_RF_PMF))^2
rmse <- round((RMSE(p_RF_PMF, indoor$pm2.5I)), digits =2)
p_RF_PMF<-as.data.frame(p_RF_PMF) #IMP to plot ggplot
{observed <- indoor$pm2.5I
mean_observed <- mean(observed)
predicted <- p_RF_PMF
ioa <- round(1 - (sum((predicted - observed)^2) / sum((abs(predicted - mean_observed) + abs(observed -
mean_observed))^2)), digits =2)}


#SHAP
indoor_predictors <- select(indoor, ws, wd, Indoor , Saharan ,Sea.Salt , Ammonium.Sulfate , Traffic ,
                      Fireworks , Industrial , Shipping , Temp.O , Temp.I ,Humidity.O , Humidity.I ,
                      Pressure, windows, frying)  # Selects predictor columns from indoor

s <- kernelshap(modfit_RF_PMF, indoor_predictors, bg_X = indoor_predictors)
sv <- shapviz(s)

sv_force(sv, row_id = 57) # for individual   29/10   Sahran, closed windows
sv_force(sv, row_id = 156) # for individual  17/06 Sahran, open windows

{
 importance_plot <- sv_importance(sv, max_display = 17L,
                           fill = "#2a477f",
                           show_numbers = TRUE,
                           number_size = 5,
                           bar_width = 0.5)

 # Customize the text size using ggplot2
 importance_plot <- importance_plot +
   theme_bw() +  # Set white background
   theme(
     axis.title = element_text(size = 16),      # Adjust axis title size
     axis.text = element_text(size = 16),       # Adjust axis text size
     legend.title = element_text(size = 16),   # Adjust legend title size
     legend.text = element_text(size = 16),    # Adjust legend text size
     plot.title = element_text(size = 18),       # Adjust plot title size
     panel.grid.major = element_blank(),       # Optional: remove major grid lines
     panel.grid.minor = element_blank(),        # Optional: remove minor grid lines
     panel.grid.major.x = element_line(color = "grey80") # Keep major grid lines on y-axis
   ) +
   ggtitle("SHAP Importance Plot")  # Optionally, add a title

 # Print the plot
 print(importance_plot)
}
```