Development of information and optimisation solutions for real-time monitoring of indoor air quality

Thi Hao Nguyen

THESE DE DOCTORAT

Spécialité : Sciences de l'Ingénieur
Ecole doctorale : S.I.E - Sciences, Ingénierie et Environnement

Thi Hao NGUYEN

Development of Information and Optimization Solutions for Real-Time Monitoring of Indoor Air Quality

Développement de solutions d'information et d'optimisation de la qualité de l'air intérieur renseignée en temps réel

Thèse dirigée par Mme Evelyne GEHIN – Professeure

Soutenue le 14 Décembre 2022

Devant le jury composé de:

M^{me}	OUKHELLOU Latifa	Directrice de Recherche, Université Gustave Eiffel	Rapportrice
M	ROUSSEL Gilles	Professeur, Université du Littoral-Côte-d'Opale	Rapporteur
M^{me}	GONZE Evelyne	Professeure, Université Savoie MontBlanc	Examinatrice
M^{me}	GEHIN Evelyne	Professeure, Université Paris-Est Créteil	Directrice
M^{me}	IONESCU Anda	Maîtresse de conférences, Université Paris-Est Créteil	Co-encadrante
M	RAMALHO Olivier	Docteur, Centre Scientifique et Technique du Bâtiment	Co-encadrant

Abstract

The global objective of this research is to develop a system that provides information about sources and events influencing the temporal variation of indoor air pollutants, in order to optimize the action to be taken to reduce the exposure of the occupants. This study investigated a typical indoor air database obtained via a monitoring campaign performed in a real open-plan office. Indoor and outdoor pollutant concentrations and climatic parameters, occupancy and openings' status were recorded over a long period and with a fine time step. Inverse modeling based mainly on statistical analysis and machine learning has been performed in order to achieve the two main objectives: (i) the identification of indoor sources (processes) explaining the variation of indoor particulate matter concentrations, and (ii) the development of a predicting model for window opening action in the open-plan office.

In the first part, the identification of the pollutant sources and their relative contributions to the levels of indoor air particle concentrations has been achieved by a tensor decomposition method called PARAFAC (Harshman, 1970). This method can cope with data arrays of a high number of dimensions. The analyzed tensors corresponded to different combinations of parameters monitored in the open-plan office or outdoors. The different configurations always included size-resolved particle data and, sometimes, other environmental parameters in two different cases: monitored indoors or, indoors and outdoors simultaneously; in addition, the tensor structures were arranged according to daily and hourly profiles. PARAFAC outputs were analyzed in terms of sources using complementary data analysis and signal treatment methods. The method allowed to determine the relative contributions of the identified sources and the attributable concentration at a given time. The identification model created by PARAFAC can be integrated in a real-time system to provide information about the pollutant sources at a given moment, helping to take decisions in order to avoid high pollution levels.

The second part of this thesis is dedicated to the prediction of the opening state of a group of windows in the open-plan office. Three machine learning methods: Decision Trees, k-Nearest Neighbors and Kernel Approximation have been implemented. To select the appropriate set of features for the model's input, the autocorrelation functions of the different variables and the predictor importance estimates were calculated. Validation tests were performed to compare the outputs of the models and the measured windows states monitored during 18 months in the office. According to the different evaluation indicators, the results show that all the three models perform well with the testing sets. The developed methods can be helpful for understanding occupants' behavior and also for controlling indoor air pollutant levels in buildings, either as a standalone model or a part of a real-time indoor air quality monitoring system.

Résumé

L'objectif global de cette recherche est de développer un système qui fournit des informations sur les sources et les événements influençant la variation temporelle des polluants de l'air intérieur, afin d'optimiser les actions à entreprendre pour réduire l'exposition des occupants. Cette étude a exploré une base de données typique pour la qualité de l'air intérieur dans un bureau paysager, obtenue via une campagne de mesure. Les concentrations de polluants intérieurs et extérieurs et les paramètres climatiques, l'occupation et l'état des ouvrants ont été enregistrés sur une longue période et avec un pas de temps fin. Une modélisation inverse basée principalement sur l'analyse statistique et l'apprentissage automatique a été réalisée afin d'atteindre les deux objectifs principaux : (i) l'identification des sources intérieures (processus) expliquant la variation des concentrations de particules à l'intérieur, et (ii) le développement d'un modèle de prédiction de l'action d'ouverture des fenêtres dans le bureau paysager.

Dans la première partie, l'identification des sources de polluants et de leurs contributions relatives aux niveaux de concentrations de particules dans l'air intérieur a été réalisée par une méthode de décomposition tensorielle appelée PARAFAC (Harshman, 1970). Cette méthode permet de traiter des tableaux de données de grandes dimensions. Les tenseurs analysés correspondent à différentes combinaisons des paramètres mesurés à l'intérieur du bureau paysager ou à l'extérieur de l'immeuble. Les différentes configurations comprennent toujours des données sur la granulométrie des particules et, parfois, d'autres paramètres environnementaux dans deux cas : mesurés à l'intérieur ou à l'intérieur et à l'extérieur du bureau simultanément ; de plus, les structures tensorielles sont organisées selon des profils journaliers et horaires également. Les sorties de PARAFAC ont été analysées en termes de sources en utilisant des méthodes complémentaires d'analyse des données et de traitement du signal. Cette méthode a permis de déterminer les contributions relatives des sources identifiées et leur concentration attribuable à un moment donné. Le modèle d'identification créé par PARAFAC peut être intégré dans un système en temps réel pour fournir des informations sur les sources de polluants à un moment donné, aidant ainsi à la prise de décision pour éviter des niveaux élevés de pollution.

La deuxième partie de la thèse est dédiée à la prédiction de l'état d'ouverture d'un groupe de fenêtres dans le bureau paysager. Trois méthodes d'apprentissage automatique : Decision Trees, k-Nearest Neighbors et Kernel Approximation ont été mises en œuvre. Pour sélectionner l'ensemble le plus approprié de caractéristiques à utiliser comme entrées du modèle, les fonctions d'autocorrélation des différentes variables et les estimations de l'importance des prédicteurs ont été calculées. Des tests de validation ont été effectués pour comparer les sorties des modèles et les états mesurés des fenêtres mesurés pendant 18 mois dans le bureau. Selon les différents indicateurs d'évaluation, les résultats montrent que les trois modèles sont performants sur les ensembles de test. Les méthodes développées peuvent être utiles pour comprendre le comportement des occupants et aussi pour contrôler les niveaux de polluants de l'air intérieur dans les bâtiments, soit en tant que modèle autonome, soit comme partie intégrante d'un système de contrôle de la qualité de l'air intérieur en temps réel.

Acknowledgements

Personally, I felt very lucky when I was assigned this project. During the last four years, there have been moments of everything, good and bad, but at all times I have received helps and good advises from those who surround me (my professors, my family, my friends). In this sense, I feel fortunate.

First of all, I would like to express my special thanks to my supervisor, Mme. Evelyne GEHIN for her helpful advices, comments, and contributions to my study. She helps me to point out my mistakes and give me comments to deal with many arisen problems during the time doing my thesis.

I would like to express my appreciation to my co-supervisor, Mme. Anda IONESCU. During the time doing my doctoral study at Univeristy of Paris Est Créteil (UPEC), she has given me strong support and has guided me with her academic knowledge. She always made me believe that I could succeed by encouraging me to foster my ideas and research motivation. Working with Madame IONESCU, I have learned the value of research and, above all, how to become a good researcher.

I would like to express my sincere thanks to M. Olivier RAMALHO for his supports and guidance. With his enthusiastic advice, precise comments and encouragement, I can achieved the good results from this study. In addition, his contributive revisions for my weekly reports and presentations that help me to improve my writing and presenting skill significantly. Without his help, I would had have many difficulties to finish this thesis.

Besides my supervisors, I would like to thank all the member of my thesis committee for accepting to review my work. A special thank you for M. Gilles ROUSSEL and Mme. Latifa OUKHELLOU for being the reviewers (“rapporteurs”) of my PhD thesis, as well as for Mme. Evelyne GONZE for being the examiner ("examinatrice") of my thesis dissertation.

I am grateful to my "Suivi Individuel de Thèse" members: M. Vincent FEUILLET and M. Mario MARCHETTI for their enthusiasm to guide me when I doing my interviews with them.

I would like to thanks CERTES’s members, who have helped me a lot in answering many of my research questions and sharing their experiences with me.

I greatly appreciate the following organizations: Centre d’Études et de Recherche en Thermique, Environnement et Systèmes (CERTES); Centre scientifique et technique du bâtiment (CSTB); Sciences, Ingénierie et Environnement (SIE) Ecole doctorale.

Finally, I want to give the best thank to my family, my relatives and my friends who always encourage, take care of me and raise me up during the studying and researching period.

Sincerely,

Nguyen Thi Hao…

To Cuong and Emily...

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	Autocorrelation Function
ASHRAE	American Society of Heating, Refrigeration, and Air Conditioning Engineers
CMB	Chemical Mass Balance
CORCONDIA	CORe CONsistency DIAgnostic
CSTB	Scientific and Technical Center for Building
DIY	Do It Yourself
ETS	Environmental Tobacco Smoking
Hs	Specific Humidity
Habs	Absolute Humidity
IAQ	Indoor Air Quality
ICA	Independent Component Analysis
k-NN	k Nearest Neighbor
ML	Machine learning
NDIR	Non-Dispersive InfraRed
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PMF	Positive Matrix Factorization
PARAFAC	PARAllel Factor
PANDORA	comPilAtioN of inDOor aiR pollutAnt emissions
PN	Particle in number
PM	Particle matter
RH	Relative Humidity
TSD	Trend and Seasonal Detection
T	Temperature
UV	Ultraviolet
VOCs	Volatile organic compounds
WHO	World Health Oganization

Introduction

According to the World Health Organization (WHO), humans spend more than 90 percent of their time indoors (U.S. Environmental Protection Agency, 1989). People work in offices, study at schools, stay at home or inside shopping malls, or travel inside vehicules. They are constantly exposed to contaminants from both outdoor and indoor sources when they are indoors. Due to the lock-down during the COVID-19 health crisis leading to a restricted movement, working from home, people spent even more time indoors.

Indoor air pollutants can have a variety of consequences on people's health, such as: skin irritation, nausea, headache, respiratory disorders, neurological problems, the development of some tumors, *etc* (Catelinois *et al.*, 2006; Fisk *et al.*, 2010; Jones, 1999; Sundell, 2004). Indoor air quality has also an effect on school and work performance, and hence productivity. According to the WHO's global statistics, the costs to the economy from poor IAQ in France are predicted to be between 12.8 and 38.4 billion euros per year (Kirchner *et al.*, 2011). Indoor Air Quality (IAQ) has now become a public health concern and an essential research topic.

IAQ modeling is a big challenge, due to the high complexity of the underlying phenomena. An indoor environment is a microenvironment consisting in a complex volume, due to its distinctive form, its numerous and fluctuating through time, the diverse nature of pollutant transfers, and the physicochemical reactions that might occur inside. Modeling such a system needs to have an inventory of sources, characterize their emissions, then, highlight the transformation of emissions, taking into account the influence of a variety of factors (climate, building specificities, occupation and activity of the inhabitants) in order to retrieve the concentrations that occur at a particular time and location.

These phenomena, starting from emissions, up to concentrations, should be modeled by partial differential equations which do not admit an analytical solution, or by a simplistic model if different assumptions are made. In this case, the transformations are addressed following the natural path of emissions to pollutant concentrations. There is a second way of modeling, using the concentrations (the effect) as the starting point, called inverse modeling, by contrast with the direct modeling starting from the cause (sources) to effects (pollutant concentrations in the air) .

According to the research aim, direct or inverse modeling can be required. The global objective of this research is to develop a system that provides information about sources and events influencing the temporal variation of indoor air pollutants, in order to optimize the action to be taken to reduce the exposure of the

occupants. This study investigated a typical indoor air database obtained via a monitoring campaign performed in a real open-plan office. Indoor and outdoor pollutant concentrations and climatic parameters, occupancy and openings' status were recorded over a long period and with a fine time step. Inverse modeling based mainly on statistical analysis and machine learning has been performed in order to achieve the two main objectives: (*i*) the identification of indoor sources (processes) explaining the variation of indoor particulate matter concentrations, and (*ii*) the development of a predicting model for window opening action in the open-plan office.

In the first part, the identification of the pollutant sources and their relative contributions to the levels of indoor air particle concentrations has been achieved by a tensor decomposition method called PARAFAC. This method can cope with data arrays of a high number of dimensions. The analyzed tensors corresponded to different combinations of parameters monitored in the open-plan office or outdoors. The different configurations always included size-resolved particle data and, sometimes, other environmental parameters in two different cases: monitored indoors or, indoors and outdoors simultaneously; in addition, the tensor structures were arranged according to daily and hourly profiles. PARAFAC outputs were analyzed in terms of sources using complementary data analysis and signal treatment methods. The method allowed to determine the relative contributions of the identified sources and the attributable concentration at a given time. The identification model created by PARAFAC can be integrated in a real-time system to provide information about the pollutant sources at a given moment, helping to take decisions in order to avoid high pollution levels.

The second part of this thesis is dedicated to the prediction of the opening state of a group of windows in the open-plan office. Three machine learning methods: Decision Trees, k-Nearest Neighbors and Kernel Approximation have been implemented. To select the appropriate set of features for the model's input, the autocorrelation functions of the different variables and the predictor importance estimates were calculated. Validation tests were performed to compare the outputs of the models and the measured windows states monitored during 18 months in the office. According to the different evaluation indicators, the results show that all the three models perform well with the testing sets. The developed methods can be helpful for understanding occupants' behavior and also for controlling indoor air pollutant levels in buildings, either as a standalone model or a part of a real-time indoor air quality monitoring system.

In order to present the approach developped to reach these objectives, the thesis is organized in eight chapters. The **first chapter** introduces the problem of indoor air quality, the phenomena that govern it, the impact on the humans' health and present the characteristic of the indoor environments.

Then, the **second chapter** describes the study case (the open-plan office) and the available data (pollutants' concentration, climatic, occupancy and opening status); the influence of occupancy and openings is highlighted.

After these two chapters containing general information about our study, the rest

of the thesis is divided in two main parts: (1) Source Identification and (2) Forecasting of the window opening state. Each part includes three chapters, which are organized, in order: Literature review, Method, Results and Discussion.

The first original contribution of this thesis lies precisely in chapter 3. This chapter begins with a literature review about blind source separation (BSS) in general, followed by a short presentation of source separations methods in the environmental field. A brief information on several common BSS methods is also presented, followed by some studies concerning the application of source separation in outdoor and indoor environments. The application of BSS methods to the Particulate Matter (PM) source apportionment in environments and especially in an open-plan office is also included in this subsection. Finally, discussions about the advantages and disadvantages of different BSS methods and the reason for selecting PARAFAC is indicated. The selection PARAFAC to determine the sources of variability in the pollutant time series is a first original point of this thesis.

Chapter 4 presents the selected method, PARAFAC, for our blind source separation process. The general outline of this chapter is: data pre-processing for PARAFAC, the mathematical equations for calculating the final sources profiles and contributions and the method implementation. Some complementary data analysis is performed in order to better understand the database. Different analyses on the impact of the presence of occupants and windows opening on the measured concentrations of fine and coarse particles are provided. Then, the detailed PARAFAC implementation procedure is introduced with more detailed information about input data preprocessing, structuring and choosing the number of components. Data structuring in a tensor, in different combinations, consists also an originality of this study.

Chapter 5 presents four study cases with different structures of input data and the corresponding output results using PARAFAC. Other data analysis or signal processing methods were used to help the source identification and to explain the variation of each source obtained by PARAFAC. In the end of the chapter, the results are discussed and some elements are given to conclude this analysis.

Regarding the second part of the thesis, Chapter 6 presents a literature review of different models predicting the windows opening state. The reason why we chose three machine learning models: Decision Trees, k-NN classification and kernel approximation (SVM kernel) is presented.

Detailed information about three chosen ML classification methods is introduced in the chapter 7. The methods' implementation and the results are presented in the next chapter (chapter 8) where supplementary information about Autocorrelation functions (ACF) of different environment parameters is also addressed. In this chapter, the results about the rank of the important scores of predictors and decision boundaries are provided, followed by the prediction results using different ML classification models. Evaluation methods and discussion are also included in this chapter. The developped approach for predicting the opening state of a groups of windows in the open-plan office is another original point of this thesis.

Finally, the conclusions recalls the contributions introduced in the work of this thesis and presents several perspectives.

Chapter 1

Indoor Air Quality

The first chapter introduces some generalities about Indoor air quality (IAQ) and some scientific aspects of the thesis. The first section (section 1.1) presents some general information on IAQ, especially certain IAQ-specific definitions and the phenomena that govern it. Next, the typical indoor air contaminants are introduced: their properties and some measurement methods; their sources as well as their impact on human health and on the economy. Some more detailed information about Particulate Matter is presented in the section 1.2, because this thesis focuses on this pollutant. Finally, the third section (section 1.3) presents different aspects of IAQ modeling from the literature.

1.1 General information on Indoor Air Quality

Outdoor air pollution levels are very often recorded by air quality monitoring networks in the main urban areas. By contrast, the levels and the health impact of indoor air contaminants has been recently taken into account. Indeed, indoor air quality was not a major concern until the mid-1970s. Following the oil shocks of the 1970s, the debates were mostly about energy issues. The investigations then focused more on thermal concerns in order to maximize energy performance. This resulted in the increasing of confinement and changing the occupant behaviors. All of these changes in building design have led to the degradation of indoor air quality, with a potential impact on people's health, expressed as a series of symptoms (Stolwik, 1992). This was a 'warning' for the scientific community to take a greater interest about IAQ. Consequently, scientists started to study the containment and ventilation and developed investigations of the cause-and-effect correlations between poor air quality and occupant health.

Historically, indoor air pollution problems were certainly much more apparent than they are today, with soot discovered on the ceilings of prehistoric caves providing abundant evidence of high levels of pollution produced by poor ventilation of open fires (Spengler and Sexton, 1983). The link between public health and pollution in confined areas has been more obvious in recent decades (Hoskins, 2003; Jones, 1999), and the importance of emissions from various sources has been underlined (Nazaroff *et al.*, 2003; Wallace *et al.*, 2004). Indoor air pollutants can have a variety of consequences on people's health, such as: skin irritation, nausea, headache, respiratory disorders, neurological problems, the development of

some tumors, *etc* (Catelinois *et al.*, 2006; Fisk *et al.*, 2010; Jones, 1999; Sundell, 2004). They will be presented more in detail in the subsection 1.2.4: Impact of indoor air pollutants.

1.1.1 Definitions related to Indor Air Quality

- An interior environment is a volume that is enclosed and secluded from the outside world.
- The term "indoor air" refers to non-industrial spaces such as those found in private homes, public buildings and in particular office buildings, as well as in the modes of transportation such as trains or airport terminals, according to NMRHC[1] (Brown, 1997).

There are many definitions of IAQ. According to Kubba (2017), IAQ represents the quality of the air inside the buildings expressed in terms of air contaminants concentrations and thermal conditions that affect health, comfort and performance of the building's occupants.

The quality of indoor air is the composition or state of the air at a given time. It is classified into different categories: good, bad, acceptable, or unacceptable. According to ASHRAE[2], Standard 62.1 it is acceptable when the air does not contain pollutants at risky concentrations as determined by competent authorities and at least 80% of those exposed do not express dissatisfaction.

The quality of the indoor air is decided by:

- the pollution of the outside air, which is transferred inside by ventilation, combined with
- the presence of internal sources of specific pollution related to equipment (heating and combustion devices, construction products, furniture, *etc.*), and
- human activities (smoking, cooking, cleaning, *etc.*) (U.S. Environmental Protection Agency, 1989)

1.1.2 Specificities of indoor environments

Indoor environments are characterized by different specific parameters as below:

- **Occupancy density** (Environmental Protection Agency, 2014): the presence and number of people per area or volume (people/m^2 or people/m^3). It varies according to the different enclosed spaces (dwellings, schools, offices, places of leisure, *etc.*). The occupancy density inside is usually more important than outside. A high density of occupancy modifies the thermal environment, and the confinement of the air, which implies the need for ventilation or air conditioning.

[1] The National Health and Medical Research Council
[2] American Society of Heating, Refrigeration, and Air Conditioning Engineers

- **Frequency of occupancy**. The results based on the data on time spent and activities carried out by individuals (the so called Space-Time-Activities Budget) agree that city dwellers spend more than 80% of their interior in confined spaces (Derbez *et al.*, 2006; Klepeis *et al.*, 2001).

  A study conducted in the United States shows that, on average, an individual spends 88% of his day inside buildings (homes, offices and schools), 7% in a vehicle and only 5% outside (Jones, 1999; Robinson and Nelson, 1995). In France, the representative campaign at the national level conducted by the Observatory of Indoor Air Quality (OQAI) showed that the average time spent at home is 16 hours per day and for 25% of the population, this amount is greater than 20 hours (Zeghnoun *et al.*, 2010).

- **The surface-to-volume ratio (S/V)**. The interior environments are characterized by numerous surfaces available with regard to their limited volume. These surfaces represent as many possibilities of interactions with substances and particles present in the air. This surface/volume ratio varies according to some criteria as below:

  – the dimensions of the room;

  – the proportion of surfaces covered by construction or decoration products;

  – the present furniture;

  – the number of occupants and their body surface;

  – the particles suspended in the air.

  A small room will have a higher S/V ratio than a larger room. In general, the area/volume ratio in the premises is $\geq$ 2 $m^2.m^{-3}$ (even $\geq$ 3 $m^2.m^{-3}$ in the highly furnished premises). This ratio is estimated at about 3 $m^2.m^{-3}$ in indoor atmosphere (Nazaroff *et al.*, 2003).

  The study highlights the particularly important role of the surfaces as sources and sinks of indoor air pollutants, their role as reservoirs of semi-volatile organic compounds and their role in the chemical reactivity of indoor air.

- **The presence of specific pollutants**. The composition of the indoor air can be different from the composition of the outside air due to the origin of the sources involved. Most importantly, some contaminants (mainly volatile organic compounds) are found at higher indoor levels such as formaldehyde with a national median of 19.6 $\mu g.m^{-3}$ indoors versus 1.9 $\mu g.m^{-3}$ outdoors (Kirchner *et al.*, 2007).

- **Ultraviolet (UV) radiation**. The attenuation of UV radiation in an indoor environment is much greater compared to outside when the windows are closed. The absorption of UV is variable depending on the nature of the glass and the type of glazing, but it is of the order of 90%. Indoor light sources emit very little or not at all in the ultraviolet spectrum. As a result, the photolysis of the substances in the air is negligible inside with respect to the outside. This explains why some substances can accumulate more easily

like nitrogen dioxide or formaldehyde. In an open window situation, the radiation conditions tend to approach the external conditions. The radiation energy entering the indoor environment is generally of the order of $\sim$ 1 W/m^2 (day).

- **Climatic parameters**. The absence of precipitation and the generally lower amplitude of temperature and humidity variation in indoor environments lead to relative variations in air concentrations that are lower than outdoors.

1.1.3 Factors influencing Indoor Air Quality

Indoor air quality depends on various factors such as:

- **The external environment (also called macro environment)**. The external environment including: (*i*) sources of outdoor pollution, (*ii*) the nature of the soil and its level of contamination, and (*iii*) climatic and meteorological conditions, is in perpetual interaction with the indoor environment (Institute of Medicine, 2011). The air circulates from one to the other either freely (open windows) or according to constraints (air inlet mouths, infiltrations).

  The outside conditions of temperature, humidity and pressure have repercussions on the casing of the frame, which restores all or part of these conditions inside. Thermal exchanges between the exterior and the interior of the building play an important role in the dispersion of pollutants. In addition, the sun causes a warming of surfaces inside and outside the building. The external environment therefore influences the quality of indoor air at all times. Its action will be modulated by the possible presence of a specific system of ventilation and the intervention of the occupants in the openings (National Research Council (US) Committee on Indoor Pollutants, 1981).

- **Indoor climate conditions**. The indoor climate conditions are mostly set by the occupants or the building manager, through the presence and operating parameters of a heating system and sometimes a specific ventilation system. These systems will counterbalance or mitigate the impact of external conditions in favor of a better thermal comfort of the occupants. These conditions will influence the emission parameters of the sources and the air movements between the different volumes of the interior space and consequently on the levels and the distribution of the concentrations of substances and particles in the air. They can also lead to favorable conditions for the proliferation of bio-contaminants, microorganisms that can in turn emit substances and toxins.

- **The building**. Building systems and components may have a direct and/or indirect influence on indoor air quality, *e.g.* the enclosure shell at the exterior/interior interface, the building materials and interior finishes (floor, walls, ceiling, joints, and glues). In addition, the operation and nature of the heating system, ventilation and air conditioning, the existing infiltrations and ducts (piping, electrical wires) and all the relationships between these elements also need to be taken into account.

- **The furniture**. The nature of furnishing materials (wood species, wood chips, foams and fabrics), the decoration, maintenance and cleaning products, affect the quality of the indoor air. In addition, office equipment (computers, photocopiers, printers, *etc*.) especially in office spaces play an important role (Bako-Biro *et al.*, 2004).
- **The occupant**. The occupants have a determining role in the pollution levels to which they are exposed. Their activities and behavior are very influential factors in the quality of indoor air. They can activate sources of pollution *via*: smoking, cleaning, cooking, use of combustion devices, presence of domestic animals, cleaning products, *etc*. In addition, their actions on the opening or starting of a system of ventilation or air treatment can lead to dilution of the concentration of indoor pollutants or to increase the contribution of pollutants from outside, especially oxidants (ozone, free radicals). The individual or collective perception of the occupants in terms of health and comfort will condition their behavior and ultimately the quality of indoor air.

1.2 Indoor air pollution

Air pollution in indoor environments is a dynamic phenomenon characterized by the variability of pollutant emissions from various sources (Seifert and Ullrich, 1987). These can be classified into two broad categories:

- **Continuous emission sources** (*e.g.* a particleboard, pressed-wood products). These are often influenced by environmental factors such as temperature, air velocity, relative humidity, but also by the actions of the inhabitants, whose acts on the apertures change the ambient conditions. Streaming sources vary on a time scale from one day to one week or more.
- **Intermittent emission sources**. This type of emission changes significantly faster and can alter in less than an hour, if not minutes. In general, the highest amounts of pollution (peaks) are observed during these short intervals *e.g.* cigarette smoke, incense stick burning, usage of a home product, *etc*.

These sources characteristics can be used to define a typology of environments. As a result, an office building may have distinct sources that are not present in a private dwelling, and the sources and pollutants of a residential structure are also specific. For example, an office is distinguished by the absence of a combustion process. On the other hand, the operation of photocopiers and printers in the latter fosters the creation of a specific type of pollutant, such as suspended particles, in levels that identify these settings.

This section, firstly, will be introduced some specific pollutants in the indoor environment (section 1.2.1). The information about their characteristics, limitation guidelines and health effects will be briefly presented. As this study focuses on Particulate Matter, this pollutant will be presented more in detail in a dedicated subsection 1.2.2. Next, the sources of indoor air pollution, specific to an office environment are highlighted in the subsection 1.2.3. Finally, the overall impact of

indoor air pollution on both the human health and economy is discussed in the subsection 1.2.4.

1.2.1 Significant pollutants of an indoor environment

Pollutants emitted in different indoor environments are very numerous and varied. Regarding their characteristics, they can be classified by:

1. **Gaseous pollutants:** carbon monoxide (CO), oxides of nitrogen (NOx), ozone (O_3), heavy metals, formaldehyde (HCHO), VOCs, radon, *etc*;
2. **Aerosols & bioaerosols:** particulate matter, pollen, molds, bacteria, viruses, *etc*;

General information about some significant pollutants of indoor air will be presented hereafter. Particulate matter is one of them, but it will be presented in a dedicated subsection because it is in the center of this study.

- **Ozone (O_3)**. Ozone is present in the troposphere as a secondary pollutant. Under the effect of solar radiation, nitrogen oxides are produced by the oxidation of nitrogen in the air during the fuel combustion; they can react with compounds resulting from car traffic, industries, and lead to the formation of ozone. The amount of ozone present in the troposphere is thus an indicator of significant ambient air pollution. Outside, ozone pollution rises mostly in summer, particularly in the middle of the afternoon, when meteorological conditions are most favorable (high temperature, high UV radiation, long insolation, low wind, and the presence of major pollutants). The WHO ozone guideline limit for ambient outdoor pollution is 100 $\mu g/m^3$ for 8-hour daily (WHO, 2021). Indoors, some equipment such as laser printers or copiers (during operation) can emit ozone (Destaillats *et al.*, 2008; Wensing *et al.*, 2006).
- **Carbon monoxide (CO)**. Carbon monoxide is an odorless and colorless gas, highly toxic even at low concentrations (Austin *et al.*, 2002). Because it is impossible to see, taste or smell the toxic fumes, CO can kill people living in a house before they are aware of it. The effects of CO vary from person to person depending on their age, overall health, the concentration and duration of exposure. CO comes from the incomplete combustion of fuels, including: natural gas, petroleum derivatives or wood (the combustion of any carbon product). Poorly maintained heating systems and gas stoves are the most usually implicated sources of CO in indoor air. The WHO guidelines are 5 ppm for 24 hours, 10 ppm for 8 hours, and 90 ppm for 15 minutes of exposure both indoors and outdoors. Several researches have revealed that CO is likely to have a harmful effect on the health of cardiac patients at doses sufficient to create a concentration of carboxyhemoglobin (CO-associated hemoglobin molecule) of more than 2 to 3% (Brook *et al.*, 2004).
- **Carbon dioxide (CO_2)**. The two major sources of CO_2 in confined areas are metabolism and combustion. The human metabolism generates CO_2, which is emitted into the ambient upon expiration; the concentration varies depending on the number of individuals present, their physical activity, and

the ventilation of the occupied space. In general, a human emits nearly 15 L/h of carbon dioxide at rest and between 20 and 40 L/h when active. This production is expected to be 20 L/h in an office. Because CO_2 concentration is utilized as a containment indication, it is an excellent bio-effluent marker. CO_2 can have negative physiological consequences on the central nervous, cardiovascular, and respiratory systems at very high concentrations (Institute of Medicine, 2011).

- **Nitrogen oxides (NO_x)**. There are different types of nitrogen oxides (NO_2, NO, N_2O) among which nitrogen dioxide NO_2 is the most prevalent in indoor pollution studies (Maroni *et al.*, 1995). It is also emitted during combustion (heaters or hot water production, tobacco smoke, or by transferring from the outside, coming from automobile pollution). For example, the rate of NO_2 range in a kitchen using gas, can be 8 to 10 times higher than outdoors, with peaks above 1000 $\mu g.m^{-3}$. NO_2 is a pulmonary irritant.

  The WHO Regional Office for Europe suggests for N_2O a limit of 200 $\mu g.m^{-3}$ (0.11 ppm) for one hour, 120 $\mu g.m^{-3}$ (0.06 ppm) for eight hours and a maximum of 40 $\mu g.m^{-3}$ for an annual exposure (WHO-Europe, 2000). It should be emphasized that the guidelines make no distinction between indoor and outdoor air exposure, because the location of exposure only affects the composition of the air and the quantity of certain pollutants, it has no direct effect on the exposure-response connection (WHO-Europe, 2000).

- **Volatile Organic Compounds (VOCs)**. VOCs are chemical families that include alkenes, alkanes, aldehydes, ketones, esters, alcohols, and others. They are emitted as gases from certain solids or liquids. Some VOCs may have short- and long-term adverse health effects. Concentrations of many VOCs are higher indoors (up to ten times) than outdoors. VOCs are released from a variety of sources, including construction materials, glues, cleansers, household items, deodorants, photocopiers, and solvents (Destaillats *et al.*, 2008; Fenech *et al.*, 2010). Some occupant activities, such as smoking, cleaning and burning, are also producers of VOCs. Dominant permanent sources are related to construction and insulating materials; aldehydes, especially formaldehyde and acetaldehyde, are frequently in the majority. VOC concentrations are often less than 1 $\mu g.m^{-3}$ (Lévesque *et al.*, 2003).

- **Formaldehyde (HCHO)**. Formaldehyde is a member of the VOC family and is a colorless gas with a distinctive odor. It is irritating for the upper respiratory system. In recent years, research on formaldehyde exposure has received considerable attention in the IAQ field for four main reasons (Wolkoff and Nielsen, 2010):

  – The IARC (International Agency for Research on Cancer, 2006) classifies HCHO as carcinogen;

  – The study conducted by Nazaroff *et al.* (2003) stated that: the reactions between ozone and monoterpenes form formaldehyde, and

  – Epidemiological studies discovered the effects of HCHO exposure on lung problems, and

  - Studies on the exposure of vulnerable people show that children, the elderly and people with asthma and other breathing problems are more sensitive to the effects of formaldehyde; the HCHO effects seen in vulnerable people may potentially be more severe (Agency for Toxic Substances and Disease Registry, 2014).

  WHO recommends a guideline value of 0.1 μg.m^{-3} (0.08 ppm) to protect the population against irritant effects.

- **Biological pollutants**. These contaminants come from a variety of sources. The growth of some biological contaminants can be reduced by controlling the relative humidity level indoors at home. In general, a relative humidity of 30-50 percent is suggested for dwellings. Molds, mildews, germs and insects develop in standing water, water-damaged materials and moist surfaces. House dust and mites, which grow in moist, warm conditions, are a source of one of the most potent biological allergens (Radford, 1976). Bacterial cells and spores, viruses, pollen, fungus, algae, detritus, and cell fragments are all examples of bioaerosols. Bioaerosol particles are typically a small fraction of aerosol particles in our environment, but their influence can be significant. They are a source of disease transmissio and produce allergic responses (Löndahl, 2014). In addition, researchers also indicated that bioaerosols can have an impact on the global temperature, ecology, and biodiversity; some scientific studies showed that bioaerosols may have a significant influence on clouds and precipitation (Després *et al.*, 2012; Hamilton and Lenton, 1998). The bioaerosols are a type of the large category of aerosols, which will be presented next.

- **Aerosols**. Aerosols contain solid or liquid droplets in air or gas, having a negligible rate of fall. The suspended particles cover a very wide spectrum, ranging from a few fractions of nanometers to 100 microns (Seinfeld and Pandis, 2012). A specific type of aerosols are the bioaerosols, which have been presented earlier. Several classifications have been developed based on their health effects or their physico-chemical characteristics. Indeed, according to the particle size mass distribution there are:

  - ultrafine particles ($d_a < 0.1$ μm: where d_a is the aerodynamic diameter[3])
  - fine particles ($0.1 < d_a < 2.5$ μm) and
  - coarse particles ($d_a > 2.5$ μm).

  Thus, the terms PM10 and PM2.5 represent the fraction of the atmospheric aerosol which contains particles having an aerodynamic diameter less than or equal to 10 μm and 2.5 μm respectively (Chow and Watson, 1998). The PM will be developped more in detail in the next subsection 1.2.2.

[3]The aerodynamic diameter is defined as the diameter of a spherical particle with a density of 1 g/cm^3, which has the same sinking speed as the particle under consideration (Vincent, 2007).

1.2.2 Indoor Particulate Matter

The transport, resuspension and deposition of particles in indoor environments are fundamentally influenced by a series of transformations and different physicochemical processes (Gundel *et al.*, 2005).

1.2.2.1 Sources of particles

The mechanisms of formation and transformation are intrinsically linked to the different sources, climatic parameters and occupation. Figure 1.1 shows the main processes involved in determining the concentration of the indoor particles. These factors could lead to considerable changes in the chemical composition of the particles, their physical characteristics and particle size distributions. To these effects, the concentrations (in mass or in number) of the particles and the contribution of their sources would vary in different ways depending on the extent of these processes.

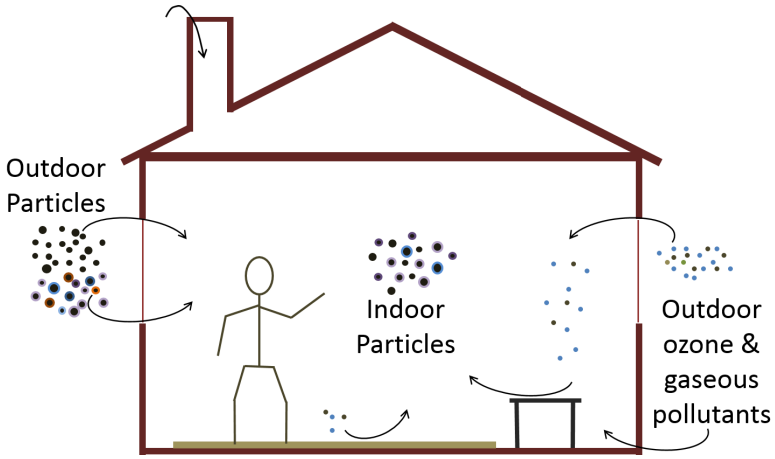


Figure 1.1: Particle sources and its flow of movement.

More information about these sources is presented below, according to their location: outdoors or indoors.

- **Outdoor environment sources**. Particles are constantly moving between indoor and outdoor environments *via* the ventilation systems, the building shell and openings (doors and windows). At least half of the particles inhaled inside are from outside (Wallace *et al.*, 2003). This finding highlights the importance of outdoor particulate pollution level affecting particle concentrations indoors.

  The infiltration capacity of the particles is variable depending on the particle size; it ranges from 0.38 to 0.94 for particles between 0.02-0.5 μm (ultrafine) and it decreases, ranging from 0.12 to 0.53 for particles which are coarser 0.7-10 μm (Abt *et al.*, 2000). Another research study has revealed that up to 33% of indoor/outdoor airborne particles can occur even with closed doors and windows (Alzona *et al.*, 1979).

According to Han *et al.* (2015), external fluctuations can explain about 81% to 90% of the variance in indoor PM2.5 concentration. To evaluate the contribution of outdoor levels to indoor concentrations, a simple indicator such as the indoor/outdoor ratio (Indoor/Outdoor - (I/O)) can be utilized. This indication is dependent on a large variety of characteristics, but particularly on the building's internal sources and air exchanges with the outside. The opening of windows, the ventilation equipment, and the permeability connected to the construction of the building, in particular, play a key role in these exchanges. As a result, extrapolating measures seen in certain types of frames to all interior situations is extremely challenging. A review of the literature on the relationship between indoor and outdoor concentrations by the ratios I/O, penetration factor, and infiltration was conducted by Chen & Zhao (2011).

- **Indoor environment**. In recent years, many studies have sought to examine in detail the contribution of occupant activity to particulate concentrations. In general, the use of the vacuum cleaner and the dust removal process contribute to the increase of the mass of the coarse particles (6 - 10 μm). A multiple linear regression analysis was applied to estimate the contribution of the various domestic activities: cooking (cooking, grilling, blowing, *etc*); cleaning (vacuum cleaner, dust removal, *etc*); occupancy (characterized by occupant movement) and washing (Abt *et al.*, 2000). The results of this study show that these variables contribute in a distributive way according to the size of the particles. Thus, according to the regression model, cooking and occupant movement have more impact in determining the fraction of particles larger than 2 μm, washing is the least important except for particles in size between 0.02 - 0.5 μm.

  The emission of fine and ultrafine particles varies according to domestic activities (Géhin *et al.*, 2008). The combustion processes mainly cooking food and the operation of a petroleum auxiliary heater represent the main sources of particles in indoor environments with a high proportion of ultrafine particles.

1.2.2.2 Indoor particle concentrations variability

In France, the main source of data on particulate pollution in the various interior spaces remains the national campaigns conducted by the OQAI [4]. For the housing campaign (2003-2005), for example, different measurements were made in the dwellings: in 297 dwellings, PM10 concentrations were monitored, while in other 290 dwellings, PM2.5 were measured. The dwelling were chosen to be representative, from a spatial distribution point of view, of the whole National Park. The medians for the monitored PM10 and PM2.5 concentrations values were respectively 31.3 $\mu g.m^{-3}$ (max = 523 $\mu g.m^{-3}$) and 19.1 $\mu g.m^{-3}$ (max = 568 $\mu g.m^{-3}$) (Kirchner *et al.*, 2007).

[4] Observatoire de la qualité de l'air intérieur - Indoor Air Quality Observatory

A measurement campaign in 133 Paris offices with different ventilation systems showed that, on average, the concentrations of particles smaller than 8 μm in offices equipped with controlled mechanical ventilation are the lowest, with 93.5 $\mu g.m^{-3}$, compared to 148 $\mu g.m^{-3}$ for offices equipped with an air conditioner and 136 $\mu g.m^{-3}$ in naturally ventilated offices (Vincent *et al.*, 1997). PM2.5 concentrations can reach 265 $\mu g.m^{-3}$ in the presence of smokers in offices (Mosqueron *et al.*, 2002) and on average, they were around 100 $\mu g.m^{-3}$ in the presence of at least two smokers in a dwelling (Ramalho *et al.*, 2012).

Several studies have attempted to quantify the contribution of computing devices and office equipment to particle concentrations (Koivisto *et al.*, 2010; Uhde *et al.*, 2006; Wensing *et al.*, 2006). Most of these studies were performed in a simulation chamber, so extrapolation of these results to real environments is still difficult.

1.2.3 Sources of indoor air pollution in the case of office environments

Indoor air pollution is caused by the complex interactions of various substances present at different levels, depending on the location and the source emission. Each pollutant is dependent of a variety of sources, and each source can produce a variety of pollutants.

The pollutant concentration is generally determined by:

- the relationship between the volume of air contained in the confined space,
- the rate of pollutant production (or emission),
- the rate of pollutant elimination by reaction or deposition on surfaces,
- the pollutant's external concentration, and
- the air transfer parameters , in particular airflow exchanged with the outside (Maroni *et al.*, 1995).

Furthermore, the emissions of materials present in the interior environment are affected by their age, environmental factors, and physicochemical properties (porosity, *etc*.). The diluting function of air renewal is affected by the concentrations of the outer and inner compartments, as well as the ventilation method. The effect of the air change rate on indoor air quality reflects the ambiguity of its function as both a source of pollution from the outside environment and a substantial sink for pollutants in the indoor air.

The office is one of the places where a lot of people spend the majority of their time per day, around 35 hours per week. Wolkoff (2013) sustained that the research related to the office environments has been carried out in order to evaluate the health impact and/or the performance of the occupants. The author concluded that the impact of indoor pollutants in offices should not be ignored. Despite the fact that several studies in the existing literature have indicated the presence of different pollutants (PM, VOC, bio contaminants, *etc*) in both homes and workplaces, significant disparities must be underlined. Furthermore, because smoking

is no longer legal in public places starting from the 1st February 2007 [5], the presence of certain contaminants (like nicotine, carbon monoxide, and ammonia) in office environments is no longer of concern.

On the other hand, other contaminants, such as VOCs, ozone, particles (e.g. toner dust) and formaldehyde, may exist in sufficient quantity to characterize these atmospheres (Salthammer *et al.*, 2010; Saraga *et al.*, 2011). Laser printers, copiers, and computers may all emit these substances into the air (Schripp *et al.*, 2009).

Some recent research attempted to measure the emissions of particles emitted by printers-photocopiers, focusing on the ultra-fine particles and VOCs in general (Kagi *et al.*, 2007; Lee and Hsu, 2007). Fine particle emissions from laser printers and copiers are affected by several factors, including the age and the type of the printer or photocopiers utilized, as well as the age and toner charge (Lee and Hsu, 2007; Uhde *et al.*, 2006; Wensing *et al.*, 2006). It is the commissioning of the equipment, in particular, that would be the source of particle pollution peaks in office environments. In other words, while measuring occupant exposure to indoor pollution in this sort of workplace, the number of prints and the start-up of the copier should be taken into account.

Additionally, the usage of a vacuum cleaner is common in offices. It is hypothesized that the symptoms of the sick-building syndrome are caused in part by maintenance products that emit VOCs (Wolkoff, 2013). A review of the literature on pollutants (specifically on semi-VOCs - SVOCs) emitted by different sources specific to the office environment was conducted by (Destaillats *et al.*, 2008). According to this research, the chamber concentration of ozone (O_3) emitted by a laser printer when it is in operation was about 9-10 ppbv[6], meanwhile it was 6 ppbv for the all-in-one office machines. The PM10 chamber concentration was 65 ppbv emitted by the laser printers in operation and about 41 ppbv for the all-in-one office machines. In addition, the emission rate of the desktop PCs of formaldehyde was detected as 5.2-12.8 $\mu g.h^{-1}.unit^{-1}$ and the chamber concentration was 0.1 $\mu g.m^{-3}$. Based on the different researches, the study included although the re-emission of ambient particles deposited in the units has been demonstrated, computers are typically not a source of ozone or particulate matter. For ozone, although the emission rates are unclear, even low levels of ozone emitted by printers and copier machines can react with other indoor pollutants, resulting in secondary pollutants and generation of ultrafine aerosol particles (Destaillats *et al.*, 2006; Singer *et al.*, 2006). Furthermore, significant amount of particulate matter are generally detected when printers, copiers and multi-functional devices are used, therefore, it is needed to investigate both the physical and chemical characterization of aerosol particle emissions during the printing process, especially for ultrafine and nanoparticles.

[5] Ministère de la Santé et de la Prévention, 2014

[6] part per billion volume

1.2.4 Impact of indoor air pollution

The influence of indoor pollution on inhabitants has turned into a major public health concern. Some pollutants, most commonly found in indoor air (formaldehyde, benzene, CO, NO_2, particulates, *etc*), but also in consumer items and food (phthalates, pesticides, heavy metals, *etc*), are reported to induce predominantly long-term health impacts (Zhang *et al.*, 2010). These chemical and physical pollutants have a cumulative effect on the body and might first express themselves as symptoms (mucous irritation, dyspnea, dry skin, *etc*.) associated with poor air quality (Hoskins, 2003).

Many recent researches have emphasized respiratory problems (Harley, 2020; Nam and Ryu, 2018). In metropolitan France, the yearly number of lung cancer deaths due to home radon exposure varies from 1 200 to 2 900 (Kirchner *et al.*, 2011).

The International Agency for Research on Cancer has re-evaluated formaldehyde in 2005 and concluded that formaldehyde is carcinogenic to humans, based on sufficient evidence on humans and on experimental animals (Cogliano *et al.*, 2005).

Many studies have been conducted to establish the root cause of IAQ problems, and significant progress has been made in recent years in identifying pollutants and the variables that lead to their existence. An up-to-date state of knowledge can be found in the researches of Ilacqua *et al.* (2017) and Nadadur (2015), and the role of different building components on the adverse effects of pollutant exposures from various sources, in a study performed by Spengler (2001).

Besides the indoor air contaminants, temperature, humidity, air movement and the quality of ventilation systems also affect IAQ. When someone suffers from influenza, in a room with high humidity and poor ventilation, the other people living there can easily be infected. The situation is worse when staying with someone who suffers from serious infectious disease like Severe Acute Respiratory Syndrome (SARS) and the result could be fatal. Looking back to the year 2003, the outbreak of SARS has made more than 8000 people become sick with severe acute respiratory syndrome that was accompanied by either pneumonia or respiratory distress syndrome (Centers for Disease Control and Prevention, 2004). Nearly ten percent of infected has been killed (774 people died in more than 23 countries). SARS spread rapidly around the world because some infected people traveled by aircrafts, one of the typical public indoor environments.

These examples, and there are very many others which could not be mentioned here, show why IAQ is so important.

In an indoor environment with good IAQ, human health can then be protected. Besides, the indoor temperature and humidity plus a good ventilation system can also bring us comfort. Evidence shows that the productivity is improved in a good indoor air quality environment. On the other hand, an environment with poor IAQ can lead to different issues. People simply do not feel comfortable. Besides, poor IAQ also brings negative health impacts to us. Two common building related health issues, the Sick Building Syndrome (SBS) and the Building Related Illness

(BRI) are caused by poor IAQ. Working in an environment with poor IAQ also leads to lower productivity and high absenteeism.

While symptoms of SBS include eye, nose or throat irritation, dry cough, dry or itchy skin, headache, dizziness and nausea, poor concentration and fatigue, building related illness (BRI) is a clinically diagnosed illness that is directly related to environments with poor indoor air quality. BRI is different from SBS, as BRI requires prolonged recovery times after leaving the building. Examples of BRI symptoms are allergic reactions, infectious disease or even cancer (Crook and Burton, 2010; Seltzer, 1994).

For more information, the "Sick Building Syndrome" (SBS) is associated with the 1970s energy crisis, which triggered modifications in the design, building materials, building equipment, and building systems such as, air conditioning, in order to save energy. As a result, many individuals complain of discomfort, even pathology: the quality of indoor air is frequently called into consideration. According to the study of Ezzati (2005), the mix of pollutants in inhaled air indoors influences symptoms or a combination of symptoms. VOCs levels have been linked to the prevalence of SBS in certain research (Nakaoka *et al.*, 2014; Suzuki *et al.*, 2020). Recently, a link between indoor particulate matter and black carbon with SBS symptoms in a public office building is addressed by Nezis and colleagues (2022). However, the association between SBS and indoor air quality is not always systematic, as many other psycho-sociological factors need to be taken into account (Dorothée *et al.*, 2013).

Indoor air quality has also an effect on school and work performance, and hence productivity. Several worldwide, largely American, studies, have looked at the costs of poor air quality (Fisk and Rosenfeld, 2004; Mendell *et al.*, 2002). According to the WHO's global statistics, the costs to the economy from poor IAQ in France are predicted to be between 12.8 and 38.4 billion euros per year (Kirchner *et al.*, 2011).

1.3 Modeling of Indoor Air Quality

This section discusses indoor air quality modeling, a big challenge, due to the high complexity of the underlying phenomena. An indoor environment is a microenvironment consisting in a complex volume, due to its distinctive form, its numerous and fluctuating through time, the diverse nature of pollutant transfers, and the physicochemical reactions that might occur inside. Modeling such a system needs to have an inventory of sources, characterize their emissions, then, highlight the transformation of emissions, taking into account the influence of a variety of factors (climate, building specificities, occupation and activity of the inhabitants) in order to retrieve the concentrations that occur at a particular time and location.

These phenomena, starting from emissions, up to concentrations, should be modeled by partial differential equations which do not admit an analytical solution, or

by a simplistic model if different assumptions are made. In this case, the transformations are addressed following the natural path of emissions to pollutant concentrations. There is a second way of modeling, using the concentrations (the effect) as the starting point, called inverse modeling, by contrast with the direct modeling starting from the cause (sources) to effects (pollutant concentrations in the air) .

Whether *via* direct or inverse modeling, the two visions aim to comprehend and/or explain the quality of the ambient air in interior spaces. When confronted with this problem, there are two possibilities depending on the viewpoint of analysis: physical modeling (direct cause-effect) or inverse modeling (usually statistics or signal processing). The usage of one or the other necessitates the utilization of their own sources of information and knowledge. Whereas models in the first scenario require knowledge of the causes and mechanisms, resulting in determinism that understands only the necessity or impossibility, statistical models represent a world made up of events stated assets that can be realized or not, based on degrees. In statistical models, knowledge of the causes is not required to create an explanation of the effects, but the effects are required to infer on the causes.

To clarify the framework within which indoor air quality modeling might be built, the deterministic method should be discussed first, followed by the "obstacles" it confronts. Then, from another point of view and according to the "need" to understand, analyze or solve an IAQ particular aspect, statistical or signal processing techniques should complete the previous discussion. For more detailed information, many works use deterministic physical models to simulate the characteristics of the indoor environment. A system of differential equations is typically used to create the model for example when one is interested in the variation of a time series data (e.g. indoor air pollutant concentrations). These deterministic physical models are used in the context of mass conservation equations, CMB (Chemical Mass Balance) models (Christensen and Gunst, 2004).

Deterministic modeling often splits the domain into homogenous zones and then applies a mass balance to each zone to calculate the various parameters. This idea has been extensively studied and produces good results in simple experimental circumstances when the variables fluctuate slightly or not at all.

For forecasting, one of the difficulties encountered by deterministic model is methodological: the physico-chemical parametrization is usually based on a static formulation. A significant difficulty lies in the practical difficulty of implementing this type of model in a real environment. For example, despite the importance of the information listed in the PANDORE [7] database (Abadie and Blondeau, 2011), which includes around 500 sources of pollutants representing nearly 7000 pollutant emissions, the forecast remains difficult to implement. The elaboration of a target volatile organic compounds (VOC) list based on the emission rates implemented in the database is also described. However, there exits some constraints for using this database as the emission rates are affected by the experimental conditions. The application of this data for another environment has to be handled

[7] a comPilAtioN of inDOor aiR pollutAnt emissions

with care. In addition, each pollutant was considered independently, meanwhile, the cumulative risk of several components should be addressed instead.

Indeed, the range of possible emission rates for a given material is not known. In addition, the coatings, form with the adhesive and the substrate, a composite material for which it is difficult to determine the final emission rate which does not result from an additivity hypothesis.

Moreover, the emission rates change over time not only due to climatic conditions, but also by the action of certain oxidants capable of modifying the nature of the emissions. Furthermore, the model does not take into account the presence and behavior of the occupant who, through his/her actions, modifies the terms associated with the emission, the renewal of air and, to a lesser extent, the surfaces available for sorption of species present in the air. These parameters are crucial to be able to predict the evolution of concentrations in a real occupied environment.

Regarding the statistical approach, this method is generally used when the *a priori* knowledge of a system is insufficient or when the parameters resulting from the physical models cannot be completely specified. In this thesis, only the outputs (temporal variables) of the system could be collected by measurement (e.g. indoor air pollutant concentrations and factors influencing them, but no information about the source emissions). Statistical models then aim to use all available information in order to better reproduce the behavior of the real system on the basis of these data. In particular, inverse modeling makes it possible to infer the nature of the system or to provide forecasts on the future state of the system.

Modeling of indoor air quality can also be motivated by two main practical objective: (i) to highlight the variability of the sources of fluctuation and their contributions; (ii) for forecasting purposes: either by the single time series of pollutant concentrations taken individually, or *via* a set of state variables and factors. More details about these approaches and some examples from the scientific literature will be presented in the next chapter.

Chapter 2

Study Case and Data Presentation

Air quality in an indoor environment is often characterized by analyzing the air composition using a technology designed to measure the concentrations of target contaminants in that environment. IAQ assessment can be more precise if measurements are performed with a shorter time step over a longer period. A sensor system providing the concentration of some target indoor air pollutants, with a very short time step of 1 minute, in an open plan office, during a 4-year (2012-2015) monitoring campaign has been analysed in this study and it will be introduced in this chapter. This presentation will be completed by some overall statistics and analysis of the monitored data. A brief presentation will be given here, but an exhaustive one is available *via* the report TRIBU (Ramalho *et al.*, 2016).

2.1 The open-plan environments

The studied office is located in a building called ARIA, at the Scientific and Technical Center for Building - Centre Scientifique et Technique du Bâtiment (CSTB), situated as 84 Avenue Jean Jaurès, 77420 Champs-sur-Marne. It is situated in Greater Paris, in a suburban area bordered by many departmental routes (D199, D104, D226...) as well as one highway (A4 – autoroute de l'Est), as presented on the Figure 2.1. In the north-east direction, there is a lake called "Lac de Vaire sur Marne" and a river called "La Marne". Aside from that, two parks with lots of trees (Parc de Champs sur Marne and Parc departmental de la Haute-Ile) are located nearby.

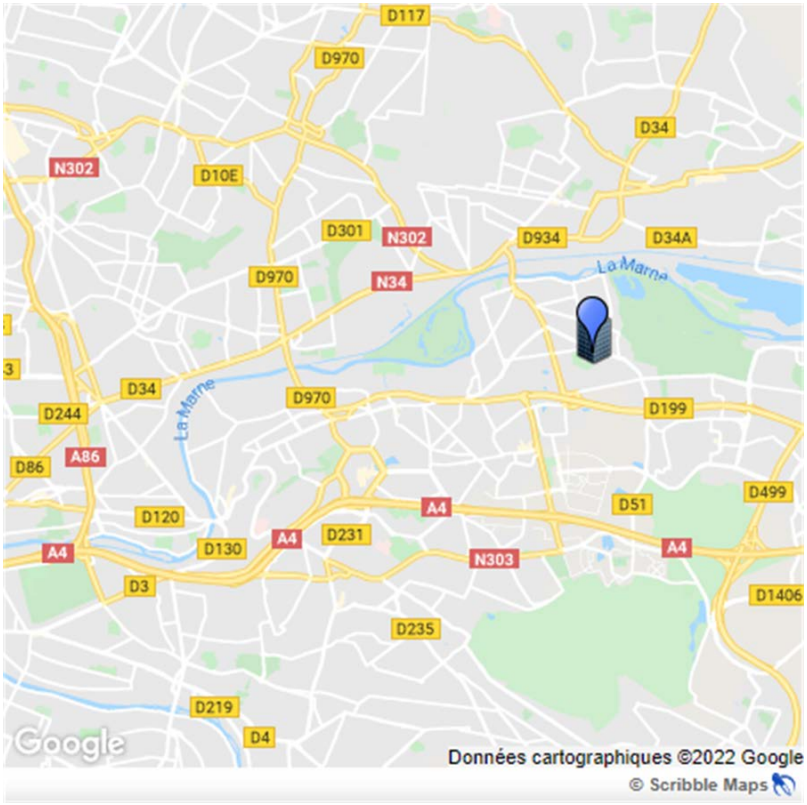


Figure 2.1: Map of the location of the studied office - CSTB (in blue).

The ARIA building is a relatively modern one in CSTB, with walls that are around 20 cm thick. It has two floors and several offices, conference rooms, experimental laboratories, *etc*. The open-plan office is situated on the second floor, where there are also many individual offices. The studied office is at the end of the entry lobby (on the left side). As it is situated at the second floor and it is not covered by the other buildings or trees, it can absorb a lot of sunshine and wind *via* a large number of sliding windows (5 windows).

The working space of the open-plan office is occupied by 6-15 persons, depending on the period of the year. According to the working hours, the office is normally occupied from 8:00 to 18:00, from Monday to Friday. It has a total area of 132 m^2 and a total volume of 364 m^3. On the Figure 2.2, it is limited by a red border. Glass walls or wood walls are used to separate the individual offices from the center space.

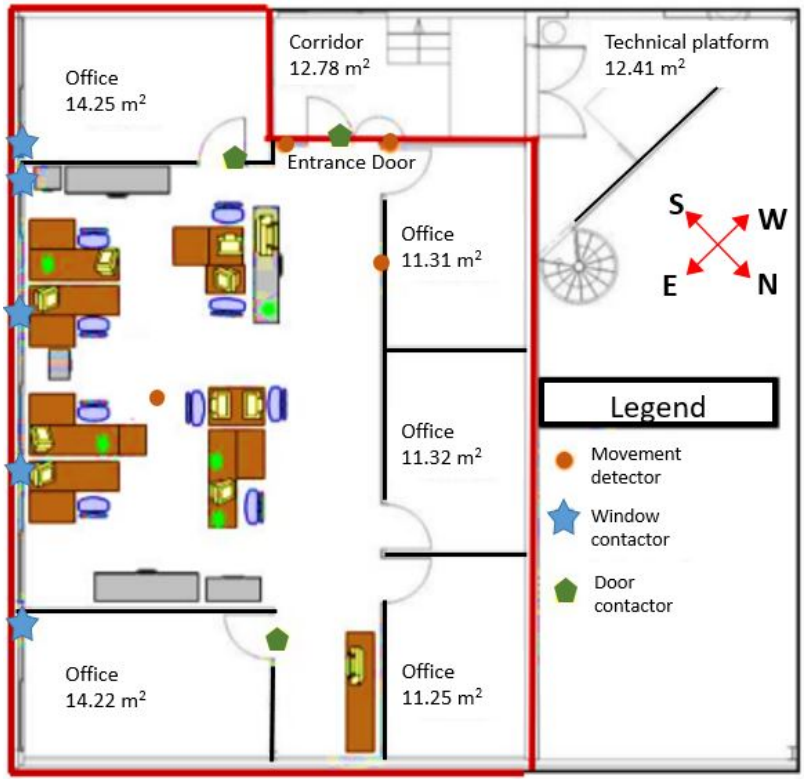


Figure 2.2: Plan of the studied office in CSTB (delimited in red).

On the plan presented on Figure 2.2, there is a common space (Corridor) right in front of the entrance door where people use to have a break. In this area, an oven is frequently used during lunchtime. In addition, there is a kettle for boiling water and a coffee machine.

The walls of the workplace are painted a bright yellow tone. The floor is covered with felt carpet, and there are artificial ceiling pieces at the height of about 3 meters. The furniture is mostly made of compact melamine wood offices which consists of: wood tables, polypropylene chairs, aluminum cabinets and around 18 computers. The number of active computers varies with occupancy, but generally, at least seven computers are permanently active during the working hours. Two printers were active in 2012 and a multi-function printer-copier has been in operation since 2013 (Figure 2.3). Other than this laser copier, there is no specific source of particles in that open office. Besides, many papers and books are placed around the room.

Figure 2.3: The active printer in studied office

In addition to the natural ventilation (opening of windows and doors controlled by the occupants or infiltration), a simple flow ventilation system without sweeping is provided for the entire open space and individual offices. It provided a constant air extraction rate of 252 m^3/h in 2012 and 228 m^3/h in 2014. The studied space communicates through the rest of the building *via* a single door that leads to a circulation space, leading to the experimental hall, or a small outdoor courtyard to reach the roof, or the left wing of the building.

Concerning the outdoor environment, a permanent weather station located on the roof of the target building automatically recorded the temperature, relative humidity, atmospheric pressure, solar irradiance, speed and direction values of wind. It also detected rainy events.

2.2 Case study database

In this section, we introduce the different available data of the database used for the open-plan office study. The active instrumentation of the environments described above has provided more or less complete information on the ambient air quality of these spaces: (*i*) measurements of pollutant concentrations; (*ii*) measurements of climatic conditions and (*iii*) the influence of the occupation and the state of the openings.

For each environment (indoors/outdoors), the active instrumentation monitored time series with different time steps (from 1 minute to 1 hour) over periods ranging from a few days to one year. The different measured parameters are listed in Table 2.1.

Each time series is recorded according to the type of the variable considered (last column of Table 2.1) and by different instruments or sensors.

Table 2.1: The parameters measured in different environments (indoors/outdoors).

Parameter	Device	Variable Name	Unit	Type
Window Opening **Door Opening**	Opening detector (CSTBox)	OF OP	0/1	binary
Movement	Passive infrared sensor (CSTBox)	Occ	0/1	binary
Irradiance	Solarimeter	Irr	$W.m^{-2}$	double
Nitrogen oxides (NOx) concentration **Ozone (O_3) concentration**	Microstation Environnement SA	NO_2, NO O_3	ppb (parts per billion)	double double
Pressure	Weather station	Press	hPa	double
Temperature **Relative Humidity** **Carbon oxides concentration**	Q-Track Probe (indoors) or Weather Station (outdoors)	T RH CO_2, CO	°C % ppm (parts per million)	double
Rain	Raining detector	Rain	0/1	binary
Formaldehyde concentration	AL1021 Aerolaser (Hantzsch reaction)	HCHO	ppb (parts per billion)	double
Aerosols concentration	Optical measurement Grimm Dust Monitor 1.108	PN	Number of particles/litter	double

2.2.1 Pollution and climatic data

The concentration in number of airborne particle per liter of sampled air, of 15 size ranges (called also fractions), varying from 0.3 to 20 μm, is continuously measured (every minute) by an optical particle counter (Grimm Dust Monitor 1.108 - see also Figure 2.4).

It is possible to calculate the specific humidity (H_s) by evaluating before the absolute humidity (H_{abs}), which is based on the relative humidity (RH), the air temperature (T) and the molar mass of the water (M_{water}) and of the air (M_{air}) by using Rankine's formula to approximate the saturated vapor pressure required for the calculation (see equations (2.1) and (2.2)).

$$H_{abs}(\frac{g}{kg\,humid\,Air}) = \frac{RH}{100} \times \frac{M_{water}}{M_{air}} \times e^{(13.7-\frac{5120}{T+273.15})} \times 1000 \tag{2.1}$$

$$H_s(\frac{g}{kg\,dry\,Air}) = \frac{H_{abs}}{(1000 - H_{abs})} \times 1000 \tag{2.2}$$

The mean daily temperature and the prevailing mean outdoor air temperature (PMA) were calculated using the seven-day weighted running mean outdoor air

temperature. Equation (2.3) gives the preferred expression for PMA with "an exponentially weighted, running mean of a sequence of mean daily outdoor temperatures prior to the day in question", according to ASHRAE (American Society of Heating Refrigerating and Air-Conditioning Engineers, 2017).

$$\text{PMA} = (1-\alpha)[t_{e(d-1)} + \alpha t_{e(d-2)} + ... + \alpha^6 t_{e(d-7)}] \tag{2.3}$$

For midlatitude climates, where people are more familiar with synoptic-scale weather variability, a lower value of α could be more appropriate so we chose $\alpha = 0.6$. In equation (2.3), $t_{e(d-1)}$ represents the mean daily outdoor temperature for the previous day, $t_{e(d-2)}$ is the mean daily outdoor temperature for two days before, and so on.

2.2.1.1 Measurement of Particulate Matter

The optical counter of particles, the Dust Monitor model 1.108 (GRIMM) continually counts (every minute) the number of particles per size band (aeodynamic diameter) for every liter of air sampled. The device counts the particles in the air sample that passes through the light beam based on light diffraction (monochromatic laser diode). The size distribution of the collected particles (15 size classes between 0.3 and 20 μm) is determined by measuring the angular dispersion generated by the passage of particles of different sizes through a light ray produced by a monochromatic laser diode at an angle of 60°-120°. The equipment has a flow rate of 1.2 L/min. With a sensitivity of 0.001 particles/cm^3 and a reproducibility of 2%, the device can count particles up to 2000 particles/cm^3 without coincidence effects.

To perform the measurements, the optical counter uses two laser powers. High laser power is used between 0.3 and 2 μm. The low laser power is utilized between 2 and 20 μm. The measurement at 2 μm is repeated twice, once with high laser power and once with low laser power, and the result is the average of the two obtained values. When calibrating each device, the maximum allowable error is 10% for the size range between 0.3 and 2 μm and 20% for the size range between 2 and 20 μm. Every minute, three measurements are obtained using three instruments: two in the office area (one near the multifunction printer, the second one directly opposite) and the third one, outside, on the building's roof (Figure 2.4).

Particle concentrations are recorded every minute in a memory card with an autonomy of more than 45 days. Data is retrieved approximately every 15 days.

It is much easier to obtain the PM (particulate matter in mass concentration) value than the PN values in order to use them in a real-time model, because this type of measurement is more in common (PM2.5 or PM10). From the PN concentrations, it is possible to calculate the mass fractions of PM2.5 and PM10 according to the method of Cheng and Lin (2010). The equations (2.4) and (2.5) explain how to convert the particle concentrations obtained into mass concentration (μg.m^{-3}) and then calculate the PM2.5 and PM10 fractions.

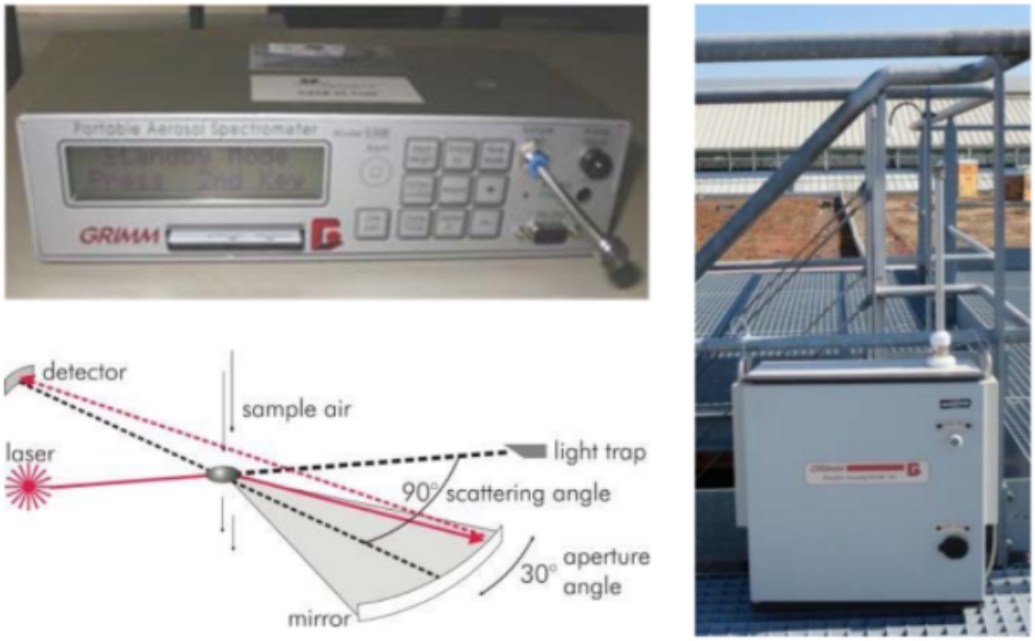


Figure 2.4: The Dust Monitor 1.108 optical counter, its principle and the outdoor measurement box (GRIMM).

According to Cheng and Lin (2010), the concentration in number should be transformed into mass concentration:

$$m(d_{pi}) = C_f \frac{\pi}{6} d_{pi}^3 n(d_{pi}) \tag{2.4}$$

where i corresponds to the channel number of the optical particle counter, d_{pi} corresponds to the average diameter between the lower and upper limit of the channel, $m(d_{pi})$ is the mass concentration of particles having an average diameter d_{pi}. C_f is called by the author as the correction factor; it represents an effective density of the particles. By default, it is fixed at 1 μg.cm^{-3} (Cheng and Lin, 2010), but a correction can be made if the nature of the particles and their density are known.

The value of C_f is not clearly determined; different researches tried to get an estimate considering it as an effective density of the particles. In order to deepen this subject, a litterature review has been performed.

Based on the measurements of ambient aerosol conducted in Beijing during winter 2007, the material density and effective density of ambient particles were estimated to be 1.61 ± 0.13 g.cm^{-3} and 1.62 ± 0.38 g.cm^{-3} for PM1.8 and 1.73 ± 0.14 g.cm^{-3} and 1.67 ± 0.37 g.cm^{-3} for PM10 (Hu *et al.*, 2012).

According to another research, the monthly mean effective densities for ambient submicron particles were found to vary from 1.3 g.cm^{-3} to 1.6 g.cm^{-3} depending on the month of year with the lowest and highest densities 1.31 g.cm^{-3} and 1.62 g.cm^{-3} in November 2012 and August 2013, respectively (Zhao *et al.*, 2017).

The mean apparent particle density, determined from a Harvard-Marple impactor and number size distribution on a daily basis was 1.6 ± 0.5 g.cm^{-3} (Pitz *et al.*, 2003).

Interestingly, during a research conducted in New Delhi, India, the authors calculated the aerosol effective density by using scanning mobility particle sizer and quartz crystal microbalance (QCM) with the estimation of involved uncertainty (Sarangi *et al.*, 2016). The aerosol stream was subdivided into two parts. One was sent to a condensation particle counter (CPC) to measure particle number concentration, whereas the other one was sent to the QCM to measure the particle mass concentration simultaneously. Based on these two parts, the total volume of particles was estimated and used to calculate the uncertainty and then the effective density. Finally, this research indicated that effective density for ambient particles at the beginning of the winter period was 1.28± 0.12 g.cm^{-3}.

After the calculation performed using equation (2.4) for each size, in order to obtain PM2.5 and PM10 fractions, the previously calculated values, weighted by the collection efficiency function specific to the Grimm counter used to determine the number of particles for each size range, should be summed, as presented in equation (2.5):

$$PM = \sum_{i=1}^{15} m(d_{pi}) f(d_{pi}) \tag{2.5}$$

where PM corresponds to PM2.5 or PM10 and $f(d_{pi})$ is the fraction of d_{pi} taking into account the collection efficiency of the reference instruments (Hinds, 1999). The collection efficiency $f(d_p i)$ is a continuous function covering all the range of the monitoring device. It can be expressed differently if it is used for PM10 or PM2.5 estimation.

These contributions can be estimated for each fraction of particles by the equations below:

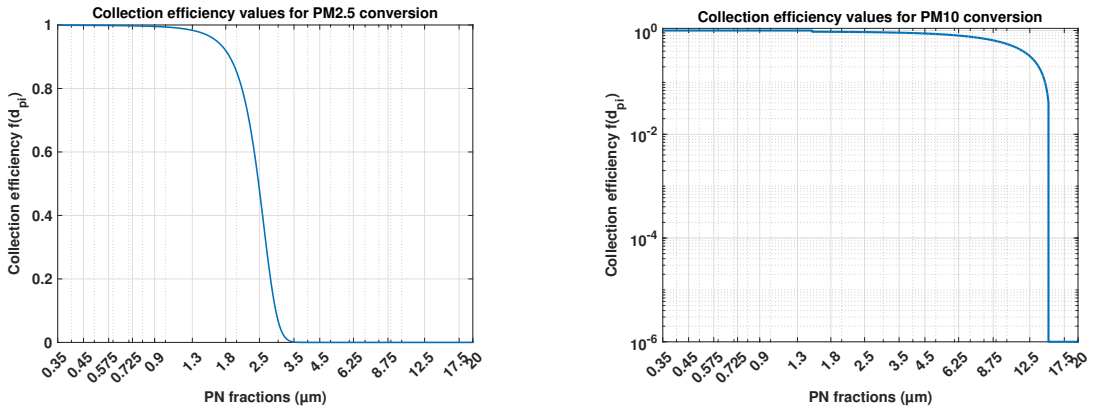
$$f_{PM10}(d_{pi}) = 1 \quad \text{for} \quad d_{pi} < 1.5\mu m \tag{2.6}$$

$$f_{PM10}(d_{pi}) = 0.9585 - 0.00408 d_{pi}^2 \quad \text{for} \quad 1.5 < d_{pi} < 15\mu m \tag{2.7}$$

$$f_{PM10}(d_{pi}) = 0 \quad \text{for} \quad d_{pi} > 15\mu m \tag{2.8}$$

$$f_{PM2.5}(d_{pi}) = [1 + exp(3.233 d_{pi} - 9.495)]^{-3.368} \tag{2.9}$$

The collection efficiency functions $f(d_p i)$ used to calculate PM2.5 and PM10 according to different particle fractions are represented in Figure 2.5:



(a) (b)

Figure 2.5: The collection efficiency functions for (a) PM2.5 and (b) PM10, according to different particle fractions.

2.2.1.2 Measurement of CO_2

Carbon dioxide (CO_2) is a tracer of both metabolic and combustion activity. In the absence of combustion sources, the only source of CO_2 in the office area is the metabolic activity associated with human presence. CO_2 monitoring is thus related to the time of occupation, during which the amount of CO_2 tends to rise due to the occupants' breath, in the absence of ventilation. The information provided by the presence detectors indicate the presence of people in the room, but it can also indicate the existence of other sources (other occupants) in the room, which the detectors cannot obtain (out of the measurement range).

In the real conditions of this office, the level of CO_2 in the room is affected by the rate at which the air in the room is renewed. Consequently, the CO_2 measurement does not depend only of the people's presence in the room. However, by analyzing the variations in CO_2 concentrations based on the occupation, it is possible to explain this variable. This continuous monitoring allows us to estimate the air renewal at various periods of the day.

In general, air exchange is determined experimentally at periodic intervals by injecting a tracer gas. However, the variability of air renewal is practically never provided. In addition, the occupants' behavior, particularly the opening and closing of windows and doors, can significantly alter this variable over the day. Its modification will have a direct impact on the oscillations of CO_2 and other pollutant concentrations. It is consequently very difficult to characterize the variability of air renewal.

The measurement of CO_2 in the office space is ensured by non-dispersive infrared (NDIR) absorption method with a Q-Track probe model 8550 (TSI Inc.) which records every minute the average values of CO_2, temperature and relative humidity.

Outdoors, the CO_2 measurement is performed by an instrument developed at CSTB on the same principle (NDIR sensor). Every 10 minutes, the instrument (Lum'Air prototype) records the average CO_2 value. It is placed inside the particle

counter's environmental box (see Figure 2.4). The box is ventilated, allowing CO_2 monitoring and protecting the device from the weather damage. The coefficient of variation of the instrument is 6% at 400 ppm and around 1% at 1000 and 2500 ppm.

The instruments have a range from 0 to 5000 ppm (up to 6000 ppm for the Q-Track probe) and share the same uncertainty of ± (50 ppm + 3% of reading). The instruments are equipped with a drift correction mechanism, but over the long term, in the absence of interventions, it is still essential to correct a zero drift of the order of +3 ppm per month for the Lum'Air prototypes.

2.2.2 The state of occupation and windows and door opening

In addition to the continuous measurement of these target pollutants, particulate matter and CO_2, more information has been continuously collected on the state of the openings and the open-plan occupation. These two parameters make it possible to understand the main causes of the real fluctuations of the IAQ.

As the concentration of CO_2 largely reflects the metabolic activity of the occupant and it is frequently considered a good tracer of human bio-effluents, CO_2 and occupancy status are therefore the most reliable indicative pair for presence.

The presence detection modules communicate with a device (CSTBox) that gathers and handles data from a building's network of sensors, contactors, or detectors. When no motion is detected, no information is transmitted to the CSTBox, and when one of the modules detects movement, the recorded value (during 10 seconds) is returned. The motion data is translated into binary data with a 1-minute time step.

The doors and windows were equipped with contactor sensors, and the data on the state of the openings (windows and doors) were recorded by the CSTBox (see Figure 2.6). These opening detection modules have also been included in the CSTBox. The data recorded by the CSTBox are time series with irregular time steps: the detection modules send back information as soon as a change of state occurs.

(a) (b)

Figure 2.6: Window opening detector: (a) closed window, (b) opened window (opened zone is indicted by the red arrow)

To synchronize all the time series at the same time step, a data preprocessing has been performed. It should be noted that only the open-plan's office has been instrumented with opening detectors. Motion detection was measured in both the individual office and in the open-space. By converting the motion into a binary variable and the state of the windows into a categorical variable, we can create two exogenous variables that can explain the characteristics of the variability of the IAQ parameters.

2.3 Data Description

After introducing general information on available data in the dataset, some simple statistics and autocorrelation function analysis were applied and the results will be presented in this section.

2.3.1 Simple statistics

2.3.1.1 Pollution and climatic parameters

The monitoring campaign in the open-plan office took place from late January 2012 to early July 2015, with a variable number of measured parameters. The summary of the parameters measured by year is presented in Tables 2.2

The percentage represents the number of minutes during the year when the parameter has been recorded (Tables 2.2). Several parameters have not been entered on a minute scale, *i.e.* the meteorological data in 2012 (most of them has been measured at a 10-minute time step), outdoor CO_2 measurement (10-minute time step) and formaldehyde in 2014 and 2015 (20-minute time step). Therefore, the 3% of the 20-minutes reported for formaldehyde in 2015 are spread over 60% of the year.

Long-term monitoring of such a huge number of parameters is subject to significant incidents, but it still needs regular check of the equipment by skilled individuals who are not always available during all-over the year. This explains the year-to-year variation in reliable measurements available. In addition, the conditions of each year are also different, some parameters, such as the counting of electrical pulses (to characterize the printer activity), were just recently set up. Depending on the purpose of the study, these data can be used at the measurement acquisition time step or at a greater time step, which would minimize the amount of missing data points.

Some general statistics for pollutant data indoors and outdoors during the year 2014 are presented in Table 2.3 and Table 2.4, respectively. One can notice that in general, the average levels of O_3 concentration (6388 hours of indoors measurement corresponding to 73% of the year) and irradiance outdoors (8760 hours of measurement corresponding to the whole year) were much higher in comparison with the indoor ones, which is often the case. These high values can be explained by the fact that the pollutants coming from the traffic system (vehicles emit nitrogen oxides are transformed in ozone with the help of the sun light (irradiance)). The maximum ozone concentration monitored indoors was 57 ppb (hourly averaged), but its average was rather low (just 5.7 ppb). Interestingly, there was a high difference between the mean and the median value of O_3 concentration indoors (mean value is 2 time higher than median value), in comparison with the concentration outdoors (mean value is 10% higher than median value), as the variability of ozone is the one of the consequences of the opening factor. Therefore, even with only one window opened, the mean concentration could be higher.

Table 2.2: Summary of the available data for the parameters monitored outdoors and indoors 2012 to 2015 expressed in % of the whole measurement period. Grey cells of the table: measurement time step = 10 mins or 20 mins, for the other ones = 1 minute (Ramalho *et al.*, 2016).

Parameter	Unit	2012	2013	2014	2015 [a]
Outdoors					
Wind speed	m/s	13%	79%	100%	94%
Wind direction		13%	79%	100%	71%
Temperature	°C	13%	83%	100%	94%
Relative Humidity	%	13%	83%	100%	94%
Specific Humidity	g/kg dryAir	13%	83%	100%	94%
Irradiance	W/m^2	13%	83%	100%	94%
Pressure	hPa	13%	83%	100%	94%
Rain	0/1	13%	83%	100%	94%
CO_2	ppm	–	–	8%	10%
Particulates [0.35 – 20 µm]	#part/L	–	36%	95%	34%
Formaldehyde	ppb	–	–	–	3%
Parameter	Unit	2012	2013	2014	2015
Indoors					
Presence	0/1	78%	78%	99%	100%
Window opening	0/1	41%	78%	99%	100%
CO_2	ppm	79%	65%	97%	96%
Temperature	°C	79%	65%	96%	83%
Relative Humidity	%	79%	65%	96%	93%
Specific Humidity	g/kg dryAir	79%	65%	96%	93%
Irradiance	W/m^2	63%	68%	91%	84%
Printer Pulses		6%	90%	80%	61%
CO	ppm	–	41%	37%	28%
NO	ppb	–	42%	–	–
NO_2	ppb	–	33%	–	–
Ozone	ppb	–	42%	47%	92%
Particulates [0.35 – 20 µm] – near the windows	#part/L	77%	43%	96%	86%
Formaldehyde	ppb	–	25%	3%	3%

[a] The data has been recorded only from January to July 2015.

Table 2.3: Some simple statistics of pollutants indoors for hourly data of the year 2014.

	O_3 indoors (ppb)	CO_2 indoors (ppm)	Irradiance indoors (W/m^2)	HCHO indoors (ppb)	CO indoors (ppm)	Printer Pulses (counts/min)
No. of samples	6388	8760	8760	5033	4383	8171
Max value	56.33	863.80	402.23	50.81	0.79	5.07
Min value	0.00	424.50	0.00	2.95	0.00	0.00
Mean value	5.71	502.22	3.00	18.70	0.20	1.42
Median value	2.84	485.03	0.00	16.50	0.18	0.62
Std value	7.13	58.95	15.65	8.26	0.11	1.02

Table 2.4: Some simple statistics of pollutants outdoors, opening factor and occupancy for hourly data of the year 2014.

	O_3 outdoors (ppb)	CO_2 outdoors (ppm)	Irradiance outdoors (W/m^2)	Number of windows opened (windows)	Presence (0/1)
No. of samples	8541	7045	8760	8760	8760
Max value	158.00	553.50	913.99	5.00	1.00
Min value	0.00	333.83	0.00	0.00	0.00
Mean value	40.08	415.45	125.49	0.87	0.35
Median value	36.50	410.50	6.57	0.00	0.00
Std value	30.67	28.04	199.66	1.27	0.48

By contrast with the ozone, the CO_2 concentration outdoors (7 045 hours of measured values corresponding to 80.4% of the year), was lower than in the indoor environment (full year available data), which can be expected because indoor environment is a confined one and it is due to the presence of people in the office.

Regarding formaldehyde concentration indoors, the office has a good air quality, while the average value of HCHO indoors was only 18.7 ppb. The maximum concentration of HCHO was only 50.81 ppb, still very low in comparison with the WHO limit (guideline) which is 80 ppb.

Fortunately, the CO concentration was maintained at a fair level with the average value of only 0.2 ppm and the highest one was only of 0.79 ppm (the limitation is 5 ppm).

For the same period of the year (from January to June), Table 8.1 and Table 8.2 represent the main statistics of the environmental parameters of the years 2014 and 2015, respectively. We can see that there are no significant differences between the averaged values of these two years during the 6 months (from January to June). One can notice that the maximum values of PM2.5 and PM10 concentrations in 2014 are quite higher than those monitored during 2015 (91.87 #/L and 106.78 #/L in 2014 in comparison with 21.3 #/L and 43.71 #/L in 2015).

Table 2.5: The statistics for environmental parameters of 2014

Features	**Indoor CO_2** (ppm)	**Indoor $PM_{2.5}$** ($\mu g.m^{-3}$)	**Indoor PM_{10}** ($\mu g.m^{-3}$)	**Indoor T** (°C)	**Outdoor T** (°C)	**Indoor Hs** (g/kg)	**Outdoor Hs** (g/kg)
Max value	1144.00	91.87	106.78	31.30	35.60	15.11	17.30
Min value	416.80	0.26	0.31	15.00	-4.30	4.28	3.98
Mean value	501.10	2.47	4.32	23.00	13.50	8.88	9.65
Median value	480.50	1.76	3.15	22.40	13.50	8.95	9.66
Std value	64.30	2.87	4.18	2.30	6.00	1.91	2.47

Table 2.6: The statistics for environmental parameters of 2015

Features	**Indoor CO_2** (ppm)	**Indoor $PM_{2.5}$** ($\mu g.m^{-3}$)	**Indoor PM_{10}** ($\mu g.m^{-3}$)	**Indoor T** (°C)	**Outdoor T** (°C)	**Indoor Hs** (g/kg)	**Outdoor Hs** (g/kg)
Max value	1038.82	21.30	43.71	33.33	39.22	13.33	14.94
Min value	421.48	0.13	0.16	18.24	-1.80	3.55	3.48
Mean value	498.45	2.50	4.45	23.10	11.28	6.44	7.11
Median value	477.02	1.93	3.40	22.30	10.30	6.22	6.69
Std value	61.38	2.11	3.70	2.43	7.01	1.55	2.09

During these six-month of measurement, CO_2 concentrations ranged from 420 to 1100 ppm with a median of around 475 ppm and a standard deviation of 65 ppm. The average profile of the CO_2 concentration remains stable at a level of about 500 ppm indoors. Similarly, these general statistics for meteorological variables like temperature and specific humidity, indoors and outdoors, do not show significant difference between these two years. Based on this comparison, we may infer that the variations from year to year of: PM concentrations, CO_2 concentration and climatic parameters values, were not significant.

Regarding the particle concentration in number (1-minute step), Table 2.7 and Table 2.8 show the general statistics for different fractions for the outdoor and indoor environment, respectively. For PN2.5, it shows that indoor PN2.5 concentration in number varied from 0 to 2866 particles/liter (average: 45 particles/liter; median: 35 particles/liter), and the range of outdoor concentrations was from 0 to 2777 particles/liter (average: 60 particles/liter; median: 50 particles/liter). The maximum, minimum, average and median values of outdoor concentrations were almost significantly higher than the values indoors, which indicates a low indoor-outdoor air exchange rate. However, in some case, the indoor concentration was temporarily higher than the outdoor one (max values of PN 1.8 and PN 2.5). This could happen when the doors/windows of the office were closed after a peak of outdoor concentration, as the required decay times for indoor concentrations is longer than for outdoors.

Table 2.7: Statistics of 1-minute step data of outdoor particle concentration in number (# particles/L) in 2014 (January - December).

PN fraction (µm)	0.35	0.45	0.575	0.725	0.9	1.3	1.8	2.5	3.5	4.5	6.25	8.75	12.5	17.5	20
Max value	8486417	1178992	354342	88879	26665	6590	2742	2777	1626	1454	3234	890	192	82	103
Min value	1436	105	30	5	0	0	0	0	0	0	0	0	0	0	0
Mean value	50069	14365	4219	996	449	200	99	60	17	7	4	1	0	0	0
Median value	26436	5648	1861	590	320	155	81	50	13	5	2	0	0	0	0
Std value	83841	24621	8259	1660	516	162	72	47	19	11	23	7	2	1	1

Table 2.8: Statistics of 1-minute step data of indoor particle concentration in number (# particles/L) in 2014 (January - December).

PN fraction (µm)	0.35	0.45	0.575	0.725	0.9	1.3	1.8	2.5	3.5	4.5	6.25	8.75	12.5	17.5	20
Max value	1588791	647162	139129	34040	12235	3350	2781	2866	1296	601	336	108	95	19	28
Min value	1315	236	78	15	5	0	0	0	0	0	0	0	0	0	0
Mean value	23722	6554	2176	794	378	140	125	45	13	5	3	1	1	0	0
Median value	14788	3338	1072	422	231	98	93	35	8	3	1	0	0	0	0
Std value	31242	11119	4019	1499	624	165	111	39	15	7	6	2	2	0	0

2.3.1.2 The state of occupation and opening of the windows

In this sub-section, some simple statistics on the data describing the state of the openings and occupancy are presented. The variable "Occupancy" designates the state of occupation of the open-plan office and it is classified as follows: it takes the value of 1 if at least one of the sensors has detected a movement, and the value of 0 otherwise.

In general, the office is occupied around 7.8% (41 088 minutes) of the time during the period from the 1st of January 2014 to the 31st of December 2014 (with 1-minute time step). This data does not really provide information on the occupancy of the open plan, but rather the total number of motion detections throughout the measurement period. If the number of detections per hour is summed, the variable "Occupancy" does not exceed 30 minutes per hour, with a maximum value recorded at 10 a.m. corresponding to 10% of the total time during the occupation period (from 7 a.m. to 7 p.m.). According to the motion detector, during the daytime, the hourly occupancy rate is about 20 minutes per hour, regardless weekends or holidays.

The parameter "Occupancy" is very variable due to various aspects (the number of occupants, the spatial coverage of the environment by the sensors, *etc*.), but it is less biased due to its quantification method. Indeed, if we analyze the distribution of occupation over all the hours of the day and regardless the day, we find that the space was occupied 19% of the time, and 26% if weekends are excluded. These numbers may appear to be erroneous, because in principle, we should expect a proportion of roughly 20-25% for 20 working days per month and an 8-hour daily presence. However, because the open-plan office is occupied by a different number of employers depending on the period of the year, the occupation is highly variable.

When working with hourly averaged data, the Occupancy variable is recalculated for an hourly time step; thus Occupancy for a specific hour will have the value of 1 when there are more than 20 minutes of presence detected and 0 if less (as we want to have the definitive value instead of float value (*i.e*. Occupancy = 1 for 25 minutes of occupancy per hour instead of Occupancy = 25/60 = 0.42).

The simple statistics for hourly averaged data of this variable according to: the hour of the day, the day of the week, and the month, during 2014 are displayed in Figure 2.7 - Figure 2.9, respectively.

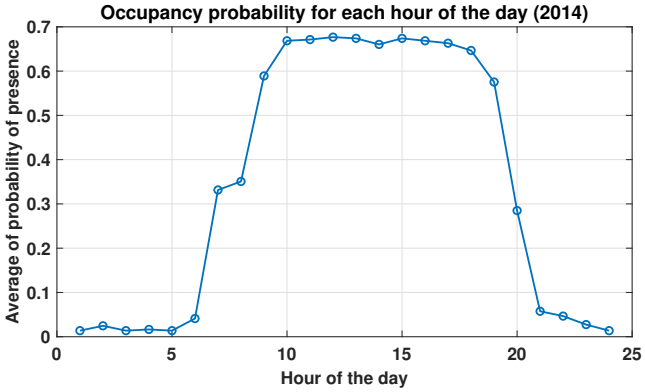


Figure 2.7: Hourly average value of Occupancy according to the hour of the day based on data during 2014.

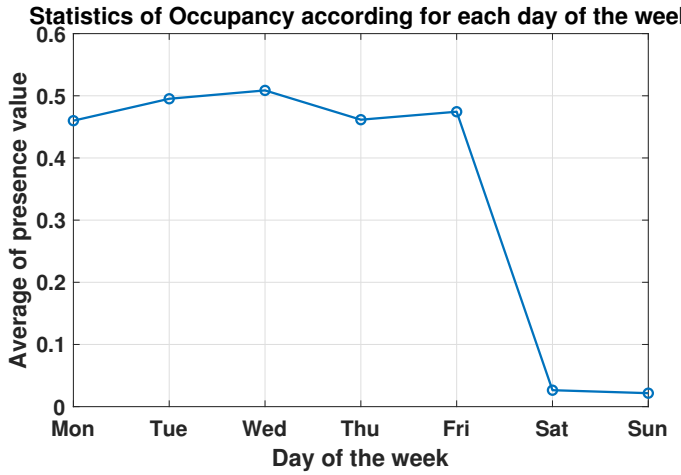


Figure 2.8: Hourly average value of Occupancy according to the day of the week based on data during 2014.

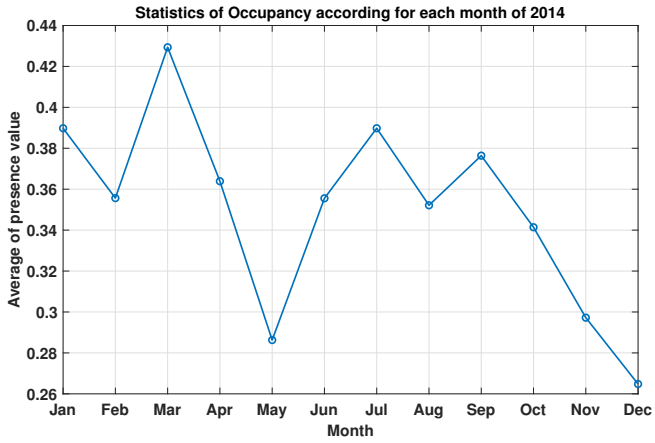


Figure 2.9: Hourly average value of Occupancy according to the month based on data during 2014.

After recalculating the hourly values of Occupancy as presented earlier, hourly data shows that there are 5693 hours (65%) when the office is non-occupied and 3067 hours (35%) when the office is occupied. During working days, the open-plan office was highly occupied between 9 a.m. and 6 p.m., with the probability of being occupied more than 60% (see Figure 2.8). Besides, the 'arriving office time' (7 - 8 a.m.) and 'leaving office time' (7 p.m.) are characterized by a lower probability of being occupied, around 35%.

Regarding the day of the week (see Figure 2.8), the working days (Monday to Friday) are characterized by the highest value of occupancy; the values varied from 45% on Monday and Thursday to 51% on Wednesday. The other working days (Tuesday and Friday) were occupied with the average probability of 48%.

The month of December got the lowest value of occupancy with only 26% (Figure 2.9). This can be explained by the fact that this month includes the Christmas and New Year holidays (around 2 weeks), when all the people are not at the office. Similarly, as May and November are the months with many national holidays, these two months also have a low percentage of occupancy (around 29%). In August, October, February and June, there are also some national holidays and vacations, leading to the value of occupancy slightly decreased – around 35%. There are also some students who are sometimes in the offices, sometimes at their university.

Regarding the opening factor, there are five windows in total, which are equipped with 5 sensors to detect their open/close status and transmit it to the CSTBox *via* a wireless network.

During the working time, the occupants tend to open at least one window, and rarely open all the five windows at the same time. In addition, the probability that no window is opened during the occupation is about 29%.

According to the daytime profile, when leaving the open-plan office, occupants tend to leave at least one window opened, especially on Thursdays and Fridays. We also notice that certain windows were opened on Saturday but not on Sunday, indicating that there was somebody present on that day who would have closed the windows (probably the guard's round or the cleaning service), even if in a punctual way. Furthermore, the presence detectors show that there are brief occupations throughout late-night and weekend hours (probably the guard's round or the cleaning service again).

2.3.2 Autocorrelation Function

In order to obtain more information about the monitored time series, the autocorrelation functions (ACF) have been calculated (using hourly averaged data). The ACF of a time series $Y(t)$ provides a measure of the correlation between y_t and y_{t+k}, where $k = 0, ..., K$ ($k \in \mathbb{Z}$, K is not larger than $T/4$, where T is the total number of observations) and y_t is assumed to be the realization of a stochastic process. According to Box *et al.* (1994), the autocorrelation r_k for lag k is:

$$r_k = c_k / c_0 \tag{2.10}$$

where:

$$c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \overline{y})(y_{t+k} - \overline{y}) \tag{2.11}$$

and c_0 is the sample variance and $\overline{y}$ is the sample mean of the time series.

The ACF result defines how data points in a time series are related, or, it measures the self-similarity of the signal over different time delay.

The examples for ACF values of hourly averaged HCHO concentration indoors and hourly averaged Occupancy values are displayed in Figure 2.10 and Figure 2.11, respectively. One can notice that the autocorrelations of HCHO persist in the positive for long delays, which means that a value at time t of the HCHO concentration can have an impact on a value of several days later. In contrast, the ACF of the Occupancy hourly values becomes negative and remains at low levels, and then switches back to positive values after a lag of around 17 hours. In general, Occupancy depict the same structures of spectral variability as CO_2 concentration: the fundamental frequency peaks at every 24 hours. The ACF of Occupancy value alternates sign every 8 hours on a lag of 24 hours. Furthermore, the 'weekly periodicity' (at the lag of 168 hours) in the ACF values of Occupancy is noteworthy.

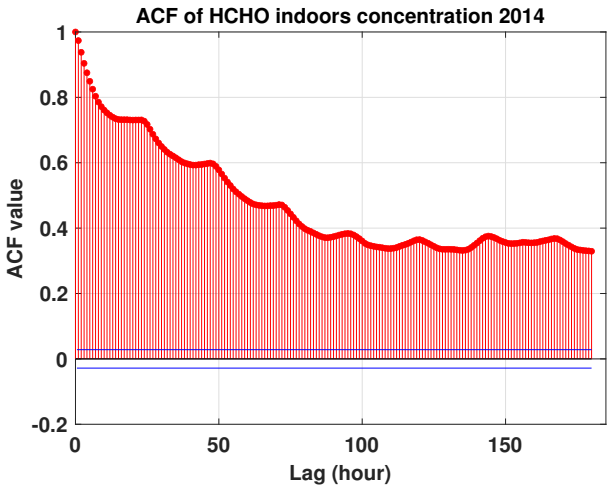


Figure 2.10: ACF value of HCHO concentration indoor 2014.

Figure 2.12 presents the autocorrelation function (ACF) values calculated from the time series of the concentration in number of particles indoors of the year 2014. The autocorrelation values corresponding to the sizes of fine particles (PN 0.35 and PN 1.3) were significantly high with a very slow decreasing. Similar to HCHO, the long persistence and high value of ACF are expressed. In fact, this persistence materialized by a slow decrease in autocorrelations (long-term correlation) represents a complex mechanism associated with the sources of fine particles. These sources present multi-frequency fluctuations, *i.e.* according to different time scales (Ramalho *et al.*, 2016).

Regarding the values of ACF of medium (PN 4.5) and coarse (PN 17.5) size particles, the similar type of behavior as ACF values of CO_2 indoors concentration and Occupancy are observed. These autocorrelation functions presented like a

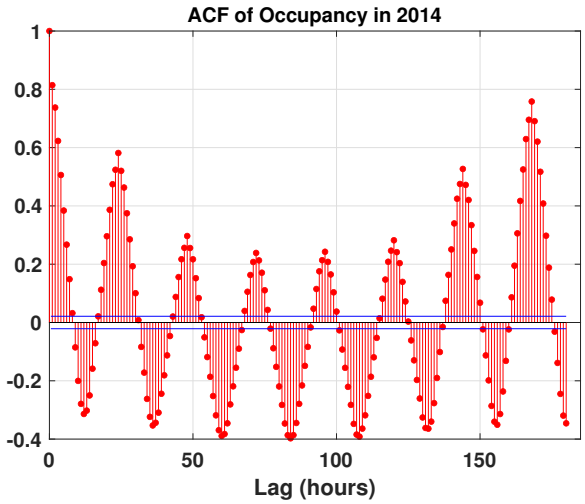


Figure 2.11: ACF value of hourly averaged value of Occupancy in 2014.

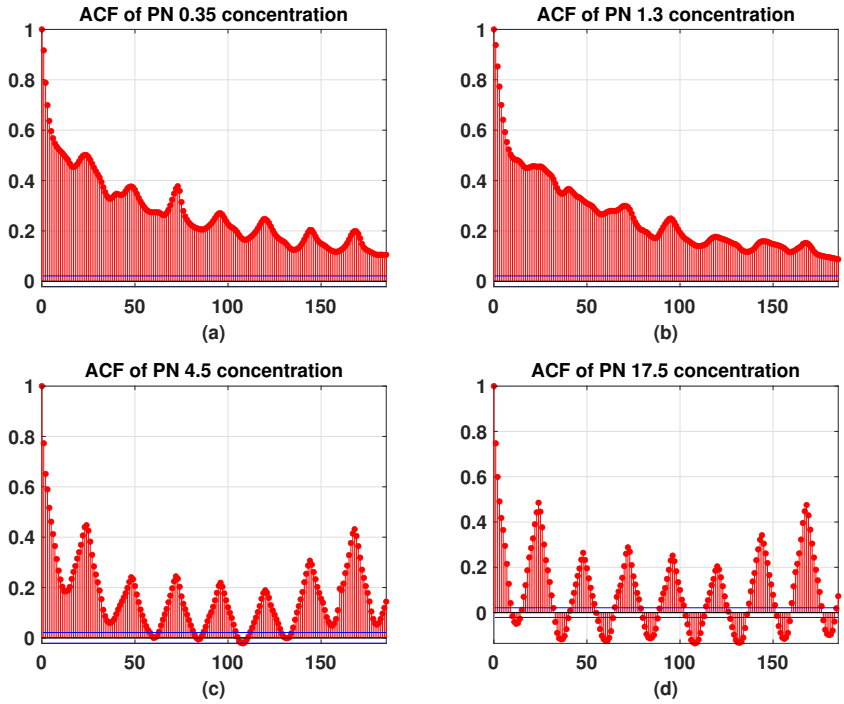


Figure 2.12: Autocorrelation function values of the time series of (a) PN 0.35, (b) PN 1.3, (c) PN 4.5 and (d) PN 17.5 concentration in number in 2014.

mixture of exponential and sinusoidal functions, the values keep switching between negative and positive level and the fundamental frequency peaks at every 24 hours are showed. The 'weekly periodicity' also easy to indicated. Indeed, the ACF of these fractions reflects the seasonal aspect of their concentration.

2.3.3 Remarks

In this section, we introduced about the studied open-plan office, where all the data in our dataset are measured. The information about available data on pollution, climatic, opening factors and occupation is then presented. We then performed some simple statistics and autocorrelation function analysis to have general information about the dataset.

According to the available data and the objective of the study, data of allover the year 2014 (almost fully recorded) was used for the first part in order to identify the pollutant sources of an open-plan office and assess their relative contributions to the general level of pollutant concentrations of particulate matter. In the second part, data of 18 months (from January 2014 to June 2015) was used for predicting the windows opening state as it is interesting to test the trained model (using data of 2014) for a completely new year data (data of 2015).

Part I

Source identification of indoor pollutants

This part of the thesis concentrates on tackling the first objective: identifying the pollutant sources of an open-plan office and assessing their relative contributions to the general level of pollutant concentrations of particulate matter. We have focused this work on pariculate matter because this pollutant is the cause of many health effects as presented in chapter 1 - section 1.2, and it is characterized, furthermore, by a complex chemical composition. This study tries to reveal the underlying factors that affect the temporal variation of particle matter in the open-plan office.

The concentration of particulate matter (PM) indoors can be affected by several factors including deposition on surfaces, resuspension, inherent variation of particle source strength, transfer from outdoors or from adjacent rooms, air exchange rate and climatic parameters such as: humidity and temperature. Considering the multitude of these factors, it is quite challenging to identify the PM sources inside the open-plan office based only on the observation data. In addition, the presence of occupants and their behavior can affect several of these factors, air exchange rate in particular through the action of windows opening. Both occupants' presence and windows opening are monitored and provided in our database, as presented in chapter 2.

Given that we have at our disposal only some measurements concerning the effect (the PM pollutant concentrations), an inverse (or receptor) modeling has to be developed in order to get information about the PM sources. In this case, there are two possibilities to get back to the sources: either by using the blind source separation methodology or by using a direct model coupled with the observation information, *via* the data assimilation. The first alternative has been chosen as it is difficult to have at our disposal a physical model for the open-plan office.

From the various methods used for blind source separations (BSS), a tensor decomposition method called PARAFAC was selected and applied to size-resolved particle data measured in the open-plan office. In order to help interpreting the identified factors which are the PARAFAC outputs, complementary data analysis and signal treatment methods were applied to them.

The general outline of this part is briefly described. Firstly, chapter 3 presents some generalities about the BSS techniques, focusing on their theoretical foundation (section 3.1), followed by a brief introduction of several source separations methods in the literature and some examples of their application for time series data of indoor and outdoor environments (section 3.2). Chapter 4 presents the detailed information about the selected BSS method - PARAFAC). The data pre-processing, the PARAFAC mathematical equations and the implementation procedure are introduced. In this chapter, different pre-processing methods are applied to PN concentration data in order to have a well-organized tensor as input for the PARAFAC decomposition. In chapter 5, different structures of input data are constructed to implement in the PARAFAC model and the results about source profiles and their contributions are presented, followed by a detailed interpretation for the source identification. Some conclusions and a brief discussion close this part.

Chapter 3

Blind source separation techniques

This section presents a literature review about blind source separation (BSS) in general (section 3.1), followed by the introduction of source separations methods in the environment field (section 3.2). A brief presentation of several common BSS methods is given in subsection 3.2.1. Some studies concerning source separation in outdoor and indoor environments are presented in subsection 3.2.2. The application of BSS methods in the particular case of Particulate Matter, the pollutant studied in this thesis, for the source apportionment in different environments and especially in an open-plan office, our main subject of interest, is also included in this subsection. Finally, section 3.3 presents a discussion about the advantages and disadvantages of different BSS methods and gives the main argument for selecting the PARAFAC method in this thesis.

3.1 Generalities on Blind Source Separation (BSS)

The problem of source separation from a mixture of signals is not a problem specific to the environment, it may also be found in other fields. The typical examples were presented in speech signal processing (Choi *et al.*, 2002; McDermott, 2009),which attempted to obtain voice separation from recorded voices of various persons speaking at the same time using several microphones (the 'Cocktail party effect'). The 'Cocktail party effect' was first defined and named "the cocktail party problem" by Colin Cherry (1953). In his study, Cherry attempted to perform attention experiments in which participants simultaneously listened to differentiate two different signals from a single loudspeaker. His research shows that the capacity to distinguish sounds from background noise is affected by a combination of factors, including the speaker's gender, the direction from where the sound is originating, the pitch and the tempo of speech.

In the both previously cited fields, an equivalent mathematical formalism can be applied. Despite the fact that the source separation problem is mathematically formalized in the same way in both signal processing and environment, the assumptions or the constraints imposed have given rise to a variety of solution approaches.

In the field of the air quality, the purpose is generally to emphasize/highlight the "signatures" of the different sources, allowing thus their identification. Assuming that the elements emitted by a source must be found grouped (statistically

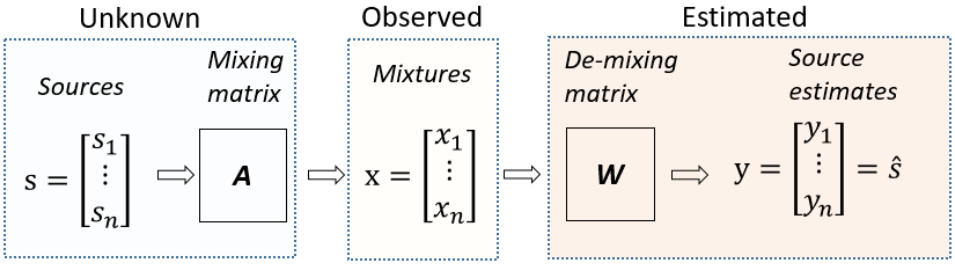


Figure 3.1: General principle of blind source separation methods (Gilbert, 2019).

correlated) in the environment, the use of classification and pattern recognition methods remains very appropriate due to their ability to highlight the groups corresponding to the strongest correlations (interpreted then in terms of "signatures" of the sources) (Ionescu, 2010).

Blind Source Separation (BSS) is a generic term for problems that involve the reconstructing of a set of time series of unknown sources from observations on mixtures of these sources. The word "blind" in the BSS term refers to the fact that there is no (or very little) information about the sources or the mixing system. In statistics, it refers to unsupervised learning approaches (without prior knowledge about the mixtures). The term "blind" is imposed in telecommunications literature and is now universally used (Comon and Jutten, 2007).

Figure 3.1 (Gilbert, 2019) shows the general principle of blind source separation methods. In this method, the available information is represented by a set of observations $(x_1, ..., x_n)$; each observation is a combination (mixture) of the different unknown sources $(s_1, ..., s_n)$, *via* an unknown mixing system $\mathcal{A}$. The purpose of BSS is to use the observation $(x_1, ..., x_n)$ in order to estimate the sources $(y_1, ..., y_n)$, typically *via* an estimated de-mixing system $\mathcal{W}$. Because the source estimating method only has access to the observations of the mixed sources with no knowledge about the sources or the mixing technique used to obtain the observations. This sort of source separation is referred to as "blind" and can comply with an infinity of solutions.

The simple mathematical explanation for Figure 3.1 is represented in equation (3.1).

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) + e(t) \tag{3.1}$$

In this equation, $x_i(t)$ represents the observed/measured data at instant t, $s_j(t)$ represents the original sources and $e(t)$ is the noise or the error measurement. The equation (3.1) could be written in a more compact form as in equation (3.2):

$$x(t) = As(t) + e(t) \tag{3.2}$$

where A is the mixing matrix ($A \in R^{nxn}$) which contains the mixture coefficients a_{ij}, $s(t) = [s_1(t), \ldots, s_n(t)]^T$ is an nx1 column vector including the sources signals at a given time t and the vector $x(t) = [x_1(t), \ldots, x_n(t)]^T$ is composed of the n observed signals at the same moment t. For the simplest BSS model with assumption of independence among the entries of the input vector $s(t)$ and possibly some a priori information about the probability distribution of the inputs, a nxn 'demixing-matrix" W ($W \in R^{nxn}$) can be used to calculate $y(t) = [y_1(t), \ldots, y_n(t)]^T$ as given in the equation (3.3):

$$y(t) = Wx(t) \tag{3.3}$$

As already said before, equation (3.1) accepts an infinity of solutions. In order to select the most suitable one according to the problem to be solved, some constraints should be imposed. Even with these constraints, the method is still called "blind".

The initial research on blind source separation started in the 1980s and originally focused on physiological signal processing, specifically decoding vertebrate motion (Roll, 1981). The biological challenge that prompted the study on source separation is described in the study of Comon and Jutten (2007). This problem entailed investigating the muscle responses emitted following various types of excitations. Since then, the resolution of source separation problems has moved to other academic disciplines and aroused the scientific community's attention. The goal is to use BSS algorithms to answer questions in a wide variety of applications dealing with various types of signals.

3.2 Source separation in the environmental field

In the environmental field, receptor models are commonly used to find information on the sources of air pollutants. "Receptor models are mathematical or statistical procedures for identifying and quantifying air pollution sources at a receptor location" (United States Environmental Protection Agency, 2022). These models use approaches for tackling the mixture problem resolution by using chemical composition data for gases and particles measured. These models are therefore a complement to other air quality models for identifying sources contributing to air quality problems. The fundamental principle of receptor modeling is based on the assumption that mass is conserved; on this basis a mass balance analysis can be used to identify and apportion sources in the atmosphere (Hopke, 2010). These methods are based on some assumptions regarding the source, chemical species and measurement methodology if this information is not known. They require a certain degree of knowledge about the sources such as: the number of sources, source profiles (which substances are emitted by which source) or source strengths. Of all considered techniques, conventional factorization and chemical mass balances represent the two extremes. Conventional factorization requires little knowledge, while chemical mass balance strategies require exact knowledge about the source(s). Other techniques, such as Positive Matrix Factorization or UNMIX, can be considered as intermediate strategies and are based

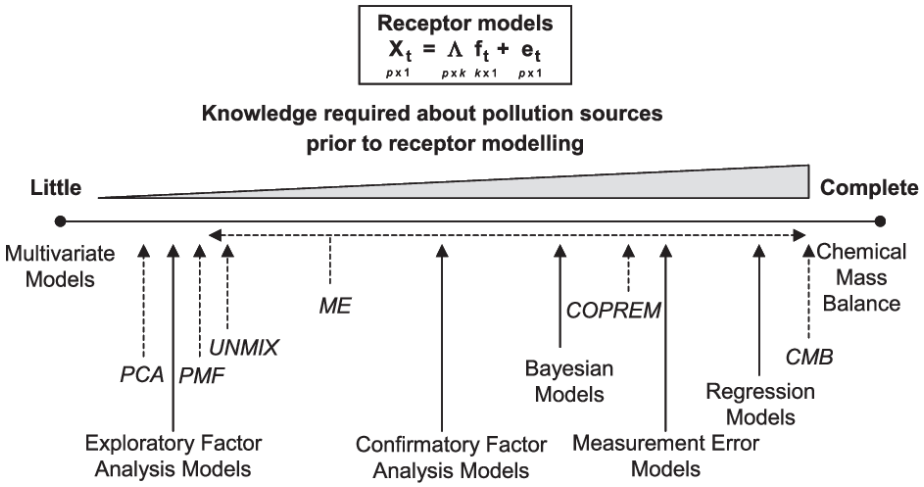


Figure 3.2: Approaches for estimating pollution source contributions using receptor models (modified from the study of Schauer *et al.* (2006)). Specific models are shown in italics and with dotted arrows (Vianaa *et al.*, 2008) .

on partly overlapping or slightly different assumptions and source knowledge requirements.

Figure 3.2 illustrates a wide spectrum of methods that address the issues of identification and contributions of pollution sources by putting into perspective the level of information required to solve the source identification problems. Clearly, the mass balance model (CMB, for Chemical Mass Balance) requires a "perfect" *a priori* knowledge of the type of the sources influencing the measurement site (and their chemical profile or strength) and seeks only their contributions.

On the other hand, the so-called "statistical" methods such as the Positive Matrix Factorization (PMF), the Non-Negative Matrix Factorization (NMF) or Principal Component Analysis (PCA) are based on the identification of sources *a posteriori* among the factors, taking into account the most probable interpretation in the context. All these methods can be included within the Exploratory Factor Analysis Models. They are all factorization techniques, under different constraints applied to the factors: PMF or NMF under the constraint of positivity (or non-negativity), PCA under the constraint of orthogonality or decorrelation, ICA under the constraint of statistical independence.

In order to identify the different source categories and estimate their respective contributions, advanced multivariate receptor models have been developed and applied successfully in many air pollution studies. The three most widely applied source apportionment techniques are: principal component analysis/absolute principal component scores (PCA/APCS), chemical mass balance (CMB) and positive matrix factorization (PMF). In this section, the brief information about these techniques is introduced in complementary with introduce about the chosen method PARAFAC in our study.

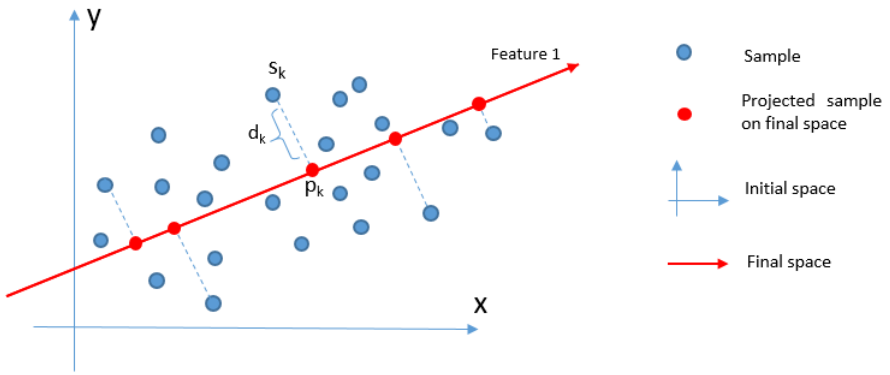


Figure 3.3: An example of PCA for projecting 2D data (cloud) into 1D (a line).

3.2.1 The most common source separation models

This section presents a state of the art of the four most widely applied source apportionment techniques above. The aim is not to give an exhaustive list of possible approaches but rather to show their diversity. These methods will be presented because their later application will have as a goal to highlight the "signatures" of the various sources, thus allowing their identification.

Indeed, by making the hypothesis that the elements emitted by a source must be found in the environment of this one grouped (statistically correlated), the use of the methods of classification and recognition of forms demeures very appropriate by their capacity to highlight the groups corresponding to the strongest correlations (interpreted then in terms of "signatures" of the sources).

3.2.1.1 Principal Component Analysis

Principal Component Analysis (PCA) was first formulated in statistics by Pearson (1901), who described this analysis as the search for lines and planes corresponding to the best fit to a system of points in space, in other words, a reduced dimensional image of the scatter plot. In 1933, Hotelling proposed it to solve the problem of statistical decorrelation (Hotelling, 1933). Since then, PCA has become a widely used tool in signal processing, especially for compression and pattern recognition.

PCA also be called "geometric data analysis" or "correlation analysis" (Wolff, 2003), this method is a descriptive technique for studying the relationships between quantitative variables, without taking into account any structure a priori. It consists in summarizing (synthesizing) the information contained in a table of numbers by replacing the initial variables by a smaller number of composite variables (the components), which are not correlated with each other. From an algebraic point of view, PCA corresponds to a representation of the cloud of points in a lower dimensional space.

The purpose of PCA is to provide the best possible visualization of multivariate data in lower dimensions to best represent the original data.

Figure 3.3 shows a simple example of how the 2D data project to 1D by PCA. There are a cloud of sample points $x_1, ..., x_n$, which are denoted as x_k on 2D space and p_k is the projected point of this sample on new 1D space. The distance between x_k and p_k is d_k. The PCA process try to select the line which minimizes the sum of d_k and best represents x_k. Then this line is represented as the new feature space.

The general process of PCA computation is as follow:

1. Compute mean vector μ and covariance matric Σ of original points
2. Compute eigenvectors ν and eigenvalues λ of Σ (eigenvectors and eigenvalues satisfy this equation: $\Sigma\nu = \lambda\nu$)
3. Select top n eigenvectors (n is the number of dimensions of the final space)
4. Project original points (samples) onto subspace:

$$y = A(x - \mu) \tag{3.4}$$

where y is the new projected point/sample, x is the original one and the rows of A are the eigenvectors.

The advantages of PCA are that it preserves the important feature of data and greatly reduces the dimensionality of the feature space.

3.2.1.2 Chemical mass balance

Chemical Mass Balance (CMB) or Mass Balance Analysis (MBA) is used to solve linear mixing problems in ambient air, provided that the compositions of the sources are known; as the name indicates, the basic principle is based on the mass balance (Hopke, 1991).

The concentration c_i of an element i measured at the receptor site is composed of the sum of the contributions of each of the surrounding sources:

$$c_i = \sum_j c_{ij} + e \tag{3.5}$$

where c_{ij} is the concentration of element i from source j; e is the error term that takes into account the background concentration and analytical uncertainties.

The term c_{ij} can be decomposed as follows:

$$c_{ij} = a_{ij} \times f_j \tag{3.6}$$

where a_{ij} is the contribution of source j for element i; and f_j is the composition profile of source j.

Each source can be described by an association of specific elements and their abundances in this source j, the profile corresponds to the concentrations of elements specifically from of source j.

The main assumptions of the CMB model are:

- there is no chemical reaction among chemical species (they add linearly);
- the compositions of the sources are linearly independent;
- the number of sources does not exceed the number of chemical species;
- the measurement uncertainties are random, uncorrelated and follow a normal distribution (Christensen and Gunst, 2004).

3.2.1.3 Positive Matrix Factorisation

Positive Matrix Factorisation (PMF) was introduced by Paatero in the 90's for the quantification of the contribution of air pollution sources (1994). This method seeks a factorization of a positive matrix into a product of positive matrices (source profiles and source contribution). One of the most common application is when chemical analyses are performed in the environment giving the concentration of different species.

This method assumes that the concentration of a chemical species j in a sample i is equal to the product of the contribution of a source k in this sample and the concentration of the chemical species in this same source k. Formally, the problem is mathematically reduced to the following formulation:

$$\mathbf{X} = \mathbf{G} \bullet \mathbf{F} + \mathbf{E} \tag{3.7}$$

The matrix $\mathbf{X} \in R^{nxm}$ represents the concentrations recorded for m species (the columns of $\mathbf{X}$) of n samples (the rows of $\mathbf{X}$). The species concentration matrix $\mathbf{X}$ can be factorized into two matrices: $\mathbf{G}$ is the factor contribution matrix and $\mathbf{F}$ is the factor profile matrix, while $\mathbf{E}$ is the residual matrix. This decomposition is not unique and according to the constraint imposed, there are different well-known approaches.

The number of sources is unknown *a priori* and has to be chosen by the user according to different criteria. The method allows talking into account the uncertainity of the measurement by introducing different weights. In the case of the time series analysis, the method can be applied to determine the main factors or sources of variability. The matrix $\mathbf{X}$ in equation (3.7) represents the observations recorded by m sensors at T different times: $\mathbf{X} \in R^{mxT}$. A column of $\mathbf{X}$ represents the sensor records at a given time. These records come from n sources. The emission intensities of the sources at a given time t are in the column t of a matrix $\mathbf{X} \in R^{mxT}$.

The PMF method tries to estimate $\mathbf{F}$ and $\mathbf{G}$ by using a least squares minimization of the error $\mathbf{E}$:

$$\min_{\mathbf{G}\geqslant 0, \mathbf{F}\geqslant 0} \|\mathbf{X} - \mathbf{G} \bullet \mathbf{F}\|^2 \tag{3.8}$$

3.2.1.4 Nonnegative Matrix Factorization

The Nonnegative Matrix Factorization (NNMF) is very similar to PMF. In both cases, PMF and NNMF are looking for positive or non-negative factors. One of the main differences between the two methods is that the NNMF does not weight systematically the measurement uncertainties.

The algorithms used to solve the NNMF have been specifically developped. To solve the least squares minimization problem, Lee and Seung (1999) use the projected gradient method: fix the matrix **F**; perform a gradient descent with respect to **G**; set all the negative components of **G** to zero; completely change the roles of variables **F** and **G** and repeat the process until convergence. The same authors have also developed other algorithms to minimize the Frobenius norm of the residue, or to minimize a divergence Kullback-Liebler divergence. The resolution of the NNMF with a second order optimization has been achieved by Zdunek and Cichocki (2007) by combining two methods: the projected gradient method and the conjugate gradient method, in order to improve the convergence of the NNMF.

NNMF has found a wide range of applications in the field of signal processing and image processing (Cichocki *et al.*, 2006; Lee *et al.*, 1999; Li *et al.*, 2001). Regarding the environmental field, this method has recently been used under constraints for PM2.5 source apportionment in Northern France (Kfoury *et al.*, 2014, 2016, or for source identification of PM10 from an industrial area (Limem *et al.*, 2014).

3.2.1.5 PARAFAC

Tensor decomposition is developed in order to be able to treat complex (multidimensional) data. This method can deal with data arrays of a higher dimension, presenting a tensor structure. Moreover, tensor decomposition-based methods avoid the ambiguity of rotation and have the advantage of the solution's uniqueness in comparison with the matrix decomposition methods.

By definition, tensors are generalizations of matrices to higher dimensions and can consequently be treated as multidimensional fields (Hitchcock, 1927). In general, the tensor decomposition tries to express a tensor as a minimum-length linear combination of rank-1 tensors. For the definition of a tensor's rank by giving some examples to illustrate it, see Table 3.1.

The Parallel Factor Analysis (PARAFAC) method is one of the several decomposition methods for multi-dimensional data. This method extends the bi-linear principal component analysis method to higher order arrays. The approach was proposed simultaneously by Harshman (1970) and by Carrol and Chang (1970), the latter being known under the name CANDECOMP (CANonical DECOMPosition) and the method PARAFAC/CANDECOMP is also known as Canonical Polyadic Decomposition (CPD).

"Factor analysis seeks the minimum number of parameters to describe the maximum amount of inter-correlation among the variables. Whenever one is fitting

Table 3.1: The illustration for the rank of a tensor definition.

Tensor	Dimensionality, rank or order of tensor	Example
Scalar	0	14
Vector	1	[14 07]
Matrix	2	$\begin{bmatrix} 14 & 07 \\ 20 & 06 \end{bmatrix}$
Cube	3	$\left[\begin{bmatrix} 14 & 07 \\ 20 & 06 \end{bmatrix} \begin{bmatrix} 21 & 12 \\ 24 & 03 \end{bmatrix} \right]$

a model to data, one seeks parameters of the model that fit the data as closely as possible - parameters that optimize some measure of fit." (Green, 1966).

Using the least-squares criterion of fitting, PARAFAC analysis procedure fits the mathematical model to the data as close as possible (Paatero, 1997). Even if its corresponding model fitting degree is not as good as for other matrix decomposition methods (PCA, ICA, PMF, etc.), it presents the advantage that it gives a unique output and it is very easy to increase the complexity of the data dealt with (just adding more dimensions to the input).

A simple PARAFAC model for a 3-dimension array is given by three loading matrices A, B and C, leading to a trilinear model (Paatero, 1999) (see equation 3.9 and Figure 3.4). With the parameters a_{if}, b_{jf}, and c_{kf} which correspond to the i,j,k mode's loading vectors, respectively, the model tries to minimize the sum of squares of the residuals e_{ijk}

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{3.9}$$

where x_{ijk} is the original pre-processed data, F is the number of factors / sources / components extracted in each mode (i,j,k).

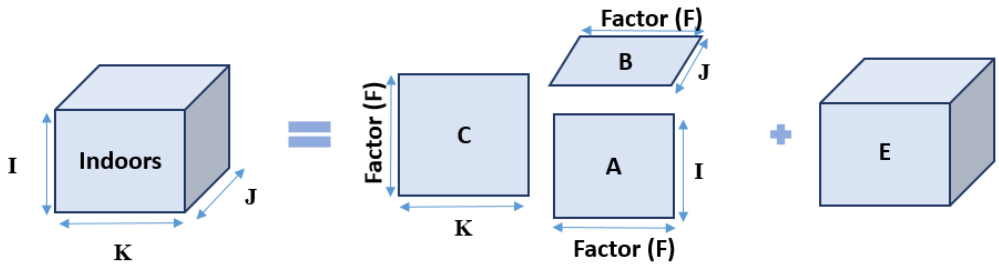


Figure 3.4: An example of PARAFAC model for 3-dimension array input.

The study of Paatero (1999) introduced a PARAFAC model in the environmental field for the chemical composition of aerosol samples, according to the equation 3.9. The 3-dimensional tensor x_{ijk} consists of the concentrations of different

chemical species, measured in different times and locations, the mode j represents the chemical species, the mode i indicates observation times (e.g. years, days, hours) and mode k corresponds to locations of measurement (e.g. indoor, outdoor, personal). For example, a layer $x_{..k}$ corresponds to the values which were measured at location k. Each column f of factor loading matrices A, B, and C represents one source of aerosol particles. For example, the column $A_{.f}$ represents the temporal development of aerosol concentration due to source f (the source contribution in time), the column $B_{.f}$ represents the chemical composition emitted by source f (the source profile). Finally, the coefficients $A_{.f}$ show the spatial distribution of the concentration due to source f among the K locations where measurements have been made.

Similarly, Hopke *et al.* (2003) use this three-way factor method to retrieve the source contribution estimates of indoor and outdoor particulate matter data. In this study, the versatile air pollutant samplers (VAPS) were considered as a three-dimensional array using a model based on the equation 3.9 where a_{if} is the f^{th} source contribution for each i^{th} sampling interval, b_{jf} contains the j^{th} concentration in the f^{th} source profile, and c_{kf} indicates the k^{th} sample type (community, central indoor, or outdoor) such that the product of a_{if} and c_{kf} provides the source contributions for each sample.

In general, the loadings can be interpreted as the strength of influence on the correlation for the respective i,j,k mode. The higher the loading value, the stronger the influence of the respective feature on the correlation.

3.2.2 Application of source separation methods in environmental sciences

The application of source separation and identification methods in environmental sciences often appears in a perspective of source recognition based on chemical signatures (association of chemical elements specific to the source). Often this knowledge exists although a little bit uncertain but is required for the interpretation of the results. Sometimes, they are not unique, because we find "partial" signatures that can correspond to several sources. Therefore, detailed knowledge of the study site can help to identify them.

3.2.2.1 Outdoors

Outdoor environmental pollution has been the subject of study in parallel with the rapid development of modern urbanization and industrialization, due to high pollution levels as well as negative health impacts from outdoor environmental exposure.

Efforts have been made for more than 50 years to link observed levels of airborne components to their sources. Measurement technology has progressed throughout this time period, enabling the gathering of exceptionally time-resolved, complete chemical compositional data. Similarly, advances in computer technology

have permitted the contemporaneous development of data analysis tools that allow information to be extracted from this data. There is currently a strong potential for delivering pertinent information on pollution sources and atmospheric processing that can help affecting air quality management policies. To generate more precise apportionments, attempts have been made to integrate receptor and chemical transport models (Hopke, 2016).

According to the review from Hopke (2016), Colucci and Begeman (1965) were the first authors to report the apportionment of pollutants (polycyclic aromatic hydrocarbons, PAHs) to a specific source type (automobile emissions) based on the concentrations of the co-emitted carbon monoxide (CO) and lead. In their research, the particulate matter samples were taken at three sites: (1) John Lodge-Edsel Ford Freeway Interchange, Detroit (Freeway), (2) Grand Circus Park, Detroit (Downtown), (3) General Motors Technical Center, Warren (Suburban). Chemical analyses were performed afterwards for these particulate matter samples. Furthermore, detailed meteorology information, especially the direction and velocity of the wind was obtained from the United States Department of Commerce Monthly Data and Data Supplements for Detroit City Airport, as they are a significant aspect in sampling operations. The apportionment was based on concentrations of lead, carbon monoxide, "tar" and polynuclear aromatic hydrocarbons, in both exhaust gas and in the atmosphere. They also made an assumption that automobiles are the exclusive sources of lead and carbon monoxide in the environment; therefore, the calculated contributions by automobiles to the polynuclear aromatic hydrocarbons in air are maxima, and the actual contributions might be lower.

In 1967, Blifford and Meeker (1967) based their research on the correlation matrix and the principal components analysis (PCA) to examine particle composition data collected by the National Air Sampling Network (NASN) at 30 urban locations in the United States for the years 1957 through 1961. The first four principal factors were tentatively assigned to (I) industrial pollution, (II) automobiles, (III) fuel burning, and (IV) petroleum refining, on the basis of their chemical composition. These four factors account for about 70% of the variance while another 20% appears to be due to the widespread use of plating materials. In this research, they also indicated that factor analysis appears to be a helpful approach for building pollution models, and additional factorial experiments including different data sets and the introduction of meteorological information would most likely lead to further separation of pollution sources. Such investigations might serve as a foundation for the reasonable design of sampling networks that use the minimum number of stations and analyses.

In general, Positive Matrix Factorization (PMF) is widely used in the atmospheric community to identify and quantify the contribution of sources to ambient concentrations in urban and rural regions (Mooibroek *et al.*, 2011) as well as in industrial locations (Kara *et al.*, 2015). In a study conducted on source apportionment of VOCs in Los Angeles, USA (Brown *et al.*, 2007), the authors succeeded to identify different factors at two sites: Azusa and Hawthorne, respectively. There, eight 3-h canister samples were collected every third day at a number of sites in the Los Angeles South Coast Air Basin during the ozone season (generally July–September) from 2001 to 2003. By using PMF as a source apportionment tool, five factors were

identified at Azusa: (1) evaporative emissions (31%), (2) liquid/unburned gasoline (27%), (3) motor vehicle exhaust (22%), (4) coatings (17%), and (5) biogenic emissions (3%). Meanwhile, at Hawthorne, six factors were identified: (1) evaporative emissions (34%), motor vehicle exhaust (24%), industrial process losses (15%), natural gas (13%), liquid/unburned gasoline (13%), and biogenic emissions (1%). In general, transportation related factors accounted for 71-80% of total reconstructed VOC mass concentration, these results are similar to previous source apportionment results using the chemical mass balance (CMB) model.

Another similar study was conducted in Houston, USA (Leuchner and Rappenglück, 2010). From the measurements located at the Moody Tower on the University of Houston campus, eight factors were identified, of which industrial sources accounted for approximately two-thirds of the reconstructed VOC mass concentration. In this study, the EPA PMF 1.1 receptor model based on the multilinear engine ME-2 (Paatero, 1999) was applied. According to this research, the PMF model provides robust results for the identification of sources in complex atmospheric environments (e.g. Houston). In this work, PCA and UNMIX models were also applied to the data set. The source profiles obtained by the PMF were the physically most reasonable results. Additionally, the modeled profile compositions show good agreement with canister samples taken within the Houston area representing crude oil handling, traffic, and an oak forest.

A study which was conducted in Shanghai, China also succeeded to use PMF to identify seven factors at the central Shanghai site, where transportation sources accounted for 40% of the reconstructed mass concentration (Cai *et al.*, 2010). VOCs were sampled from 6:00 to 9:00 using a 6 Litters silonite canister from January 2007 to March 2010. In order to study the diurnal variations, VOCs were intensively measured (8 samples a day with a 3 h interval) from August 25th to September 20th, 2009. Based on the measured VOC concentrations, a PMF model coupled with the information related to VOC sources (the distribution of major industrial complexes, meteorological conditions, etc.) was applied to identify the major VOC sources in Shanghai. The seven factors identified were: (1) vehicle-related source (25%), (2) solvent-based industrial source (17%), (3) fuel evaporation (15%), (4) paint solvent usage (15%), (5) steel-related industrial production (12%), (6) biomass/biofuel burning (9%) and (7) coal burning (7%).

A study of Polissar *et al.* (1996) also applied PMF to measurements at seven National Park Service sites in Alaska. For each site, the authors have a large number of daily samples (between 300 and 600) of suspended particulate matter (PM2.5). This application has shown that the main sources of pollution are regional sources, located at long distance. The extended study by the same authors (1998) have highlighted the possible sources, which they have classified in 4 categories, independently of the study site: transported aerosols of anthropogenic nature, sea salt, local soil dust particles and aerosols with high concentrations of black carbon from local or regional sources (forest fire).

A hybrid receptor model of constrained weighted-non-negative matrix factorization (CW-NMF) was used to investigate the impact of steelworks emission on the composition of PM2.5 in Dunkerque, Northern France (Kfoury *et al.*, 2016). In this research, the knowledge on source tracers and the relative composition of the

different sources was used as *a priori* information when running the CW-NMF model. According to the results, this study succeeded to identify eleven source profiles with varying contributions: 8 sources are characteristics of coastal urban background site profiles, and other 3 sources are related to steelmaking activities. The most significant contributors are: secondary nitrates, secondary sulfates, and combustion profiles, which account for 93% of the PM2.5 concentration. The authors also indicated that this work is the first to propose the use of *a priori* information in a hybrid receptor model based on a matrix factorization method and taking into account soft constraints on chemical profiles.

Regarding the tensor decomposition method, a multidimensional modeling of aerosol monitoring data has been obtained from the study of Astel *et al.* (2010). A three-way particulate matter (PM) data set obtained from four separate sampling locations in the Lower Austria area was used as input for the three-mode principal component analysis (Tucker3) model. Finally, the tensor decomposition model chemometric approach was successful in evaluating particulate matter chemical profiles in order to identify the main sources of pollution and analyze their spatiotemporal impact. Three latent factors determining data structure were well structured. These factors are linked to the pollution and natural source profiles in particulate matter generation in the monitoring area, such as combustion processes (indicators PM10, OC, EC), soil (indicators Si, Al), and street dust (indicators Ca, Fe). According to the study, the most significant benefit of multivariate modeling is the opportunity to analyze seasonal effects within the monitoring procedure.

PARAFAC was used to model a four-way environmental data set that comes from air quality monitoring in two industrial regions in Austria (Stanimirova and Simeonov, 2005). However, this study was more focused on the influence of chemical composition on the air quality than the variation of its sources.

3.2.2.2 Indoors

In the past, source separation was mainly addressed in the context of the outdoor environment. Some studies focus on the estimation of a particular source contribution, which is the outdoor environment, seeking to analyze the variability of the indoor-outdoor concentration ratio or transfer. Some studies stop at this level, others look for more details, separating the sources (and their contributions) in the indoor environment. The most studied environments are residential, schools or mobile environments, such as car or bus interiors.

We can find publications on the following topics in the specialized literature:

- The characterization of sources by an emission profile obtained by direct measurement, at the source;
- The identification of sources and in certain cases the source's contributions by receptor models.

The major sources of indoor pollution are now relatively well-known, although their contributions have rarely been evaluated. A significant number of research

focus on the search for a specific input: the contribution of outdoor air, by examining the association between pollutant values obtained indoors and those measured outdoors. Some studies end there, while others dig a bit deeper, trying to identify the sources (and their contributions) at the level of the indoor environment.

Guo (2011) studied the sources of VOCs in 100 selected homes in Hong Kong in winter 2002. In this study, the author applied PCA with VARIMAX, then evaluated the contributions by using the absolute principal component scores (APCS) technique combined with multiple linear regression (MLR). This technique differs from classical PCA by the positivity constraint imposed on the profiles and contributions. The choice of their method was justified by the fact that the APCS technique requires a minimum of inputs characteristics for the sources, but provides information on both the profiles and the contributions. This receptor model, like all others for that matter, may fail to separate sources if they are highly correlated (collinear). The authors analyzed samples for 15 species of VOCs and formaldehyde. They concluded that the dominant VOCs sources in Hong Kong homes were: (1) off-gassing of building materials (76.5 ± 1%), (2) room freshener (8 ± 4%), (3) household products (6 ± 2%), (4) mothballs (5 ± 3%), (5) painted wood products (4 ± 2%) and (6) consumer products. In addition, the analysis of the emission strengths of the six identified sources revealed that a small number of homes were the significant contributors to the increased concentrations of target VOCs released from these sources.

Suryawanshi *et al.* (2016) have used PMF to identify the strength of indoor air pollution sources in India. A total of 96 samples of PM0.6 were collected from different indoor microenvironments in IIT Kanpur campus, from November 2013 to September 2014. The collected samples were then subjected to chemical analysis. PMF was used for the source apportionment process. The analysis shows that five sources were responsible for the indoor pollution. These five sources were: coal combustion (21.8%), tobacco smoking (9.8%), wall dust (25.7%), soil particles (17.5%) and wooden furniture/paper products (25.2%). The study also indicated that factor contributions of the sources were not constant and they changed with time. This change in the contribution of factors with time might be due to the change in temperature, humidity and other influences.

Regarding inverse models for identifying the sources of particles, Zhao *et al.* (2007) investigated the exposure of 56 asthmatic children (aged 6-13) in schools using an expanded PMF receptor model as expressed in equations (3.10) and (3.11) below:

$$x_{ijdt} = \sum_{p=1}^{N} g_{ipdt} f_{jp} + \sum_{p=N+1}^{N+H} h_{ipdt} f_{ip} \qquad \text{t = 1/2: personal/indoor} \tag{3.10}$$

$$x_{jdt} = \sum_{p=1}^{N} g_{pdt} f_{jp} \qquad \text{t = 3: outdoor} \tag{3.11}$$

where i represents the individual (subject), j is the pollutant species (17 species: EC, NO3-, Na, Mg, Al, Si, S, Cl,K, Ca, Ti, Mn, Fe, Co, Cu, Zn, and Br), d is the sampling date, t is the type of environment (personal/indoor/outdoor), N is the number of outdoors sources and H is the number of indoor sources. Based on these explanations, x_{ijdt} means the concentration of pollutant j with type t collected on subject i on the date d. g denotes the contribution of source and f represents the relative concentration of species j in source p.

The authors searched for common sources of PM2.5 in three categories of habitats in Denver: personal house, indoors (school), and outdoors (school). Samples were collected by Teflon filters over two winter periods and then weighed and examined by X-ray fluorescence (XPF) to detect elemental concentrations from Na to Pb. For the three types of environments, they discovered four outside sources (Secondary sulfate, Soil, Secondary nitrate and Motor vehicle emissions) and three indoor sources (Chlorine-based cleaning, Cooking and Environmental tobacco smoking). Cooking was found to be the most important indoor source (30.2% contribution). Tobacco has a significant impact on the particles in personal dwellings (9.2%). The impact of high traffic flow outside the school was observed (26.5%). The authors employed an expanded model for PMF, with 4-dimensional element matrices (as equation 3.10) representing the concentration of the pollutants in a sample of a certain environment (houses, indoors, outdoors) taken on a given subject on a given day. The record of indoor activities (cleaning, swimming, cooking, etc.) on personal exposure was helpful in identifying the sources collected by the PMF. For example, the chlorine concentration exposure strength of the evening cleaning-exposed participants was twice that of the non-cleaning exposed.

Similar to the research above, in 2014, Amato and colleagues (2014) presented their study on source apportionment by using a constrained PMF model. In this research, two criteria (based on signal-to-noise ratio and detection limit) were used to select 31 strong and 2 weak species out of the total of 61 available species as the input of PMF. PM2.5 samples were collected at indoor and outdoor environments of 39 primary schools in Barcelona during 2012. After the source separation process, seven outdoor sources (Traffic, Secondary Sulfate & Organics, Secondary Nitrate, Road Dust Metallurgy, Sea Spray and Heavy Oil Combustion) and two children-activity-related sources (Mineral, Organic/Textile/Chalk - OTC) were identified. In conclusion, the research shows that children are exposed to a significantly high level of PM2.5 by the high infiltration rate of outdoor urban sources (53%) and the contribution of the OTC source (45% of indoor PM2.5). In addition, this research also indicated that "traffic contributions were significantly higher for classrooms with windows oriented directly to the street, rather than to the interior of the block or to playgrounds". This emphasizes the significance of urban design in reducing children's exposure to traffic pollutants.

A receptor model based on Non-negative Matrix Factorization (NMF) has been applied to the particle number concentrations (PNC), which were measured during the period from January 1st, 2015, to June 30th, 2015, in the same environment (the open-plan office in CSTB) as in our study (Ouaret *et al.*, 2021). This research

focuses on the time variability source characterization, namely the "temporal fingerprint" of the sources or group of sources. NMF has distinguished five major patterns obtained from the PN concentrations time series. The apportionment results were then expressed as source diurnal profiles and strengths by relating the obtained source contributions to the source information provided by the office occupancy and natural ventilation (the effect of opening windows).

Another study by Molnar *et al.* (2014) uses PMF for source identification of PM2.5, with the majority contribution being outdoor sources with 69% of all sources studied. Occupant activity accounted for 21% (2.2 $\mu g/m^3$) of personal exposure. The study took place in Gothenburg, Sweden in spring (April 2^{nd}–June 7^{th}, 2002 and March 27^{th}–June 12^{th}, 2003) and autumn (September 26^{th}–November 6^{th}, 2002 and October 7^{th}–30^{th}, 2003) seasons. 30 participants were performed in parallel with PM2.5 measurement for personal exposure, indoor, and residential outdoor. In addition, the measurement also took place at a stationary outdoor urban background station. The participants lived within 0.8–15 km from the urban background station (median distance 3.3 km). The sampling time was 24 hours. According to this study, the PMF approach with factor selection has been proven to be a valuable tool in the PMF study of different microenvironments. By integrating the records for the distinct microenvironments into a bigger dataset and utilizing the PMF with factor selection approach, the accurately estimating of the source contributions increases. In conclusion, a four-factor model (long range transport (LRT) + ship emissions (69%), local combustion (20%), traffic, and sea salt + resuspension) identified the major sources for PM2.5 at the urban background and residential outdoor sampling sites. The small contribution from traffic was due to the fact that the measurement locations were not close to any major traffic routes. Regarding the sources of PM2.5 indoors, six different sources/factors could be identified: indoor resuspension (5 $\mu g/m^3$), traffic (2.2 $\mu g/m^3$). The remaining four factors (marine, indoor Cu, soil resuspension and LRT) contribute to a lower, but similar extent. This research is interesting because not many studies performed the measurements of PM in both personal, indoor and residential outdoor simultaneously and then performed source apportionment using the PMF technique on the datasets.

A study by Yi *et al.* (1990) tried to identify the contribution of the various sources of particulate matter to that deposited on the semiconductor wafer by using factor analysis, mainly PCA. The particle concentrations in number were measured by optical particle counter in two cleanrooms: one at IBM (with four pollutant sources) and the other at the University of Cincinnati, US (with five pollutant sources). In conclusion, a receptor model for source resolution of microcontamination in these clean rooms was presented. Quantitative contributions of particulate sources to the aerosol concentration near the wafer fabricating units also be determined. For more detail, with the actual source particle size distribution is normalized ($\sum_i g_{ij} = 1$), the particles number concentration balances in a size interval i is represented as equation (3.12):

$$P_i = \sum_j N_j g_{ij} \tag{3.12}$$

where P_i is the predicted number concentration in size interval i at a receptor site, g_{ij} is the fractional number concentration size i for particles emitted from source j and N_j is the contribution of the source j to the total number concentration measured at the receptor.

Kopperud *et al.* (2004) applied a CMB-type technique to estimate the contribution of outdoor sources and indoor activities creating resuspension of particles in indoor air. For five consecutive days in April 2000 in California (US), the particle counters, nephelometers, and filter samples of integrated PM were used to measure the indoor and outdoor PM concentration, chemical composition, and air-exchange rate for the PM with an aerodynamic diameter of less than or equal to 2.5 μm (PM2.5) and PM with an aerodynamic diameter of less than or equal to 5 μm (PM5). A CMB Receptor Model Version 8 was used to determine the source contributions for each study day. In conclusion, the study revealed that indoor sources can account for up to 89% for high-activity days. Meanwhile, during the minimal-activity days, indoor sources accounted for about 30% of PM2.5 and 50% of PM5. This study also indicated that typical indoor activities can resuspend significant amounts of PM2.5. Even the very normal movement around the house can result in enough dust resuspension to account for more than 25% of the indoor PM2.5 concentration.

In 2017, Zhang *et al.* (2017) tried to extract the indoor airborne particle sources in urban office areas in Guangzhou, China by applying PCA. Regarding the studied data, measurements of indoor and outdoor PM2.5 were conducted in five types of office spaces: single-user, multi-user, photocopy room, ETS (Environmental tobacco smoking) office and fresh air office. PM2.5 was collected simultaneously by intelligent PM2.5 samplers (TH-150C) at the indoor and outdoor sites, from March 1st to 8th, 2015 (high pollution event days) and June 14th to 21st, 2015 (low pollution event days). The samplers were set at a flow rate of 100 L/min for 24 hours. The researchers investigated in the indoor-outdoor interactions between PM2.5 mass and its chemical constituents, which included water-soluble ions, carbonaceous species, and metal elements. A principle component analysis (PCA) was used to confirm the relationship between indoor and outdoor PM2.5 pollution. In conclusion, the printing and tobacco smoking were found to be the two most important sources of PM2.5 in the office. The study also suggested that improper human behavior can lead to the formation of indoor PM2.5 on a daily basis. In addition, unexpected outside pollution events might result in poor indoor air quality in urban office environments. Office workers should pay attention to their office environment because after hours of busy working, they need to maintain the human body healthily.

An online monitoring and interpretation method of IAQ using parallel factor analysis (PARAFAC) has been developed in the study of Lee *et al.* (2014)). Two types of models (global and seasonal models) were developed and their performances were also compared. The results demonstrate that there are certain differences according to the periodic pattern of the IAQ dynamics. The analysis results indicated that the seasonal models outperformed the global model in terms of model fit and the interpretation of indoor air contaminants. Furthermore, PARAFAC

helps to identify hourly fluctuations in IAQ dynamics as well as seasonal variations. The results of an experiment at a subway station shown that the proposed method provides more accurate online monitoring and a more physically relevant interpretation of IAQ than conventional univariate and multiway principal component analysis (MPCA) monitoring methods.

Martuzevicius and colleagues (2008) investigated the sources of PM2.5 in six dwellings located next to major highways (30-300 m), focusing on the impact of traffic and the link between outside and inside particle levels. The sampling campaigns were conducted from March 30th to May 14th, 2004 and from September 13th to October 22nd, 2004. The authors used a multilinear model of positive matrix factorization, specifically a trilinear model termed PARAFAC, to calculate the amount of particles originating from the traffic of inside residences. Thanks to the uniqueness of the solution advantage of PARAFAC, the paired indoor and outdoor PM concentrations can form a three-way array, assuming that these are attributed to similar sources, and only the strength of those sources varies between indoor and outdoor measurements. Chemical analyses of the samples were used to create the database (EC, OC, Si, S, Mn, Fe, Zn, Br, Pb). The authors concluded that indoor sources (activities like: smoking, cleaning, cooking and painting) contributed more to total PM2.5 levels than outdoor sources, even under conditions close to road traffic. The PM2.5 I/O ratio ranged from 0.5 ± 0.2 to 2.9 ± 1.2 in spring and from 0.7 ± 0.1 to 4.7 ± 6.9 in fall. According to the study conclusion, the structure of the house envelope and ventilation pattern appear to be the more important factors in affecting the indoor concentrations of the traffic-related aerosol, which were not quantitatively assessed in the study.

Furthermore, other researches (Hopke *et al.*, 2003; Larson *et al.*, 2004; Yakovleva *et al.*, 1999) also used the PARAFAC model in their works. Hopke *et al.* (2003) use this three-way factor method to retrieve the source contribution estimates of indoor and outdoor particulate matter data. Two sets of measurement data were analyzed: versatile air pollutant samplers (VAPS) sample particle composition collected at the community, outdoors, and central indoors of an elderly residential facility, and personal exposure monitors (PEM) sample particle composition collected from 10 elderly subjects for outdoor, central indoor, personal, and individual apartment environments. For the VAPS data set, secondary sulfate, secondary nitrate, motor vehicles, and organic carbon (OC) were identified as the sources. Meanwhile, for the PEM sets, sulfate, soil, and an unknown factor were identified as outdoor sources. Regarding the indoor environment, gypsum or wallboard, personal care products, and activity related were identified as internal factors. External factors contributed 63% to personal exposure with the most significant contribution from sulfate (48%). In this model, whereas the impact of both outdoor and indoor sources on indoor concentrations was successfully assessed, its data set was limited to a maximum of 24 samples and species above the detection limit.

3.3 Discussion

The reviewed papers in the literature show the similarity of the different methods to blindly separate the sources contributing to the observed level of particles in indoor and outdoor environments. However, separation is not identification. Without external information, the extracted factors remain difficult to identify. After their extraction, a second phase of exploitation must therefore be carried out to search for associations between the extracted factors and other observed phenomena leading to their identification.

Regarding the separation methods, each method has her own advantages and limitations. In general, the PMF method showed good agreement with the UNMIX model (Anderson *et al.*, 2002) and performed very well in comparison with CMB and PCA (Willis, 2000). Miller *et al.* (2002) found out that in comparison with CMB and PCA model approaches, the extracted factors from the PMF analysis represented the major sources that were used to generate the simulated data most closely. The lack of the non-negativity constraint is another significant limitation of PCA and CMB (Anderson *et al.*, 2002). Without this constraint, the values of the factors profiles could be large negative, leading to the result that less of the variability in the data was explained. It should be take into account also the weighted NMF algorithm (Delmaire *et al.*, 2010), which has been modified by applying constraints, a new version of the NMF in order to take into account the a-priori knowledge on the source chemical composition, considering the individual variances on the data input.

Therefore, the application of these techniques to indoor pollutant time series requires making some additional assumptions about the nature of the sources and their mixtures: linear or nonlinear, convoluted or instantaneous, time-varying or time-invariant.

Regarding the extracted factors from different studies in the literature, there were several factors of similar type/origin in the different environments. Typical ambient/outdoor sources were: region-related sources, traffic-related sources, crustal material, and marine influence (when relevant). Several of these sources also contributed in various degrees to indoor and personal exposure. Common indoor and personal sources were: air resuspension of particulate matter, indoor activities such as cooking, and other personal activities. The absolute (or relative) contribution of these sources or other ones, may differ from study to study due to local and regional conditions, for example, the size of the city, vehicle fleet composition, building types and ventilation, climate, season, and industries nearby the sampling stations. It also depends on which substances were measured and used in the models. It is therefore hard to make quantitative comparisons between studies from different locations, but qualitative comparisons can be helpful, especially during the factor identification process.

The matrix factorization methods can create a model that is well-suited to the data. However, if the dimension of the data structure is increased to more than two, these techniques must deal with the problem of rotation, uniqueness of the solution, and data complexity.

Meanwhile, PARAFAC can cope with data array of a greater complexity-higher number of dimension. Even if its corresponding model fitting degree is not as excellent as that of other matrix decomposition methods, it produces a unique output and permits to easily expand the complexity of the data.

It is important to have a dataset which can analyze both the hourly and daily variations of indoor air quality, caused mainly by occupancy and daily weather changes. This study, fortunately, has a chance to work with an extremely detailed database (see chapter 2 for detailed information). Therefore, in this study other kind of information, such as measured locations (indoors/outdoors) was used to introduce as two different layers of the same variables, and other measured pollutants concentration (HCHO, O_3, CO_2, *etc.*), the other environmental parameters such as: climatic, opening state, *etc.* were also used as input. Based on the advantages of PARAFAC, we decided to use this method to interpret such time correlations and then, to identify the sources of indoor air.

The detailed information about PARAFAC method and its implementation is presented in the next chapter.

Chapter 4

Tensor Decomposition method – PARAFAC

This chapter presents the selected BSS method, PARAFAC, in particular: data pre-processing to prepare PARAFAC inputs (section 4.1), the mathematical equations for calculating the final sources profiles and contributions (section 4.2) and, the method implementation (section 4.3).

4.1 Data pre-processing for PARAFAC

Preprocessing n-way ($n \geq 3$) arrays is more difficult than preprocessing two-way arrays; this is comprehensible given the multilinear variation assumed to be an appropriate model of the data (Bro, 1997). For the three-way array pre-processing, centering the first mode can be done by unfolding the calibration array to an I x JK matrix, and then centering this matrix as in ordinary PCA (see Figure 4.1).

$$x_{ijk}^{cent} = x_{ijk} - \overline{x_{jk}} \tag{4.1}$$

where

$$\overline{x_{jk}} = \frac{\sum_{i=1}^{I} x_{ijk}}{I} \tag{4.2}$$

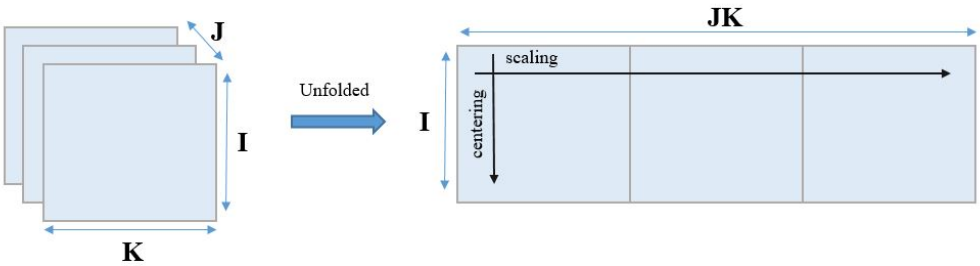


Figure 4.1: An example of three-way unfolded array. Centering must be done across the columns of this matrix and scaling has to be done on the rows.

This is commonly known as single-centering. The centering represented above is known as centering across the first mode, according to the terminology proposed by Berge (1989).

Depending on the situation, the centering can be applied to any mode. If centering is to be conducted across several modes, it must be done by first centering one mode and then centering the outcome of this centering.

When 2-centering is accomplished in this way, it is commonly referred to as double centering. Triple-centering involves focusing on each of the three modes one at a time. The effect of scaling and centering on the trilinear behavior of the data is discussed in several studies (Berge, 1989; Harshman, 1970; Paatero, 1999).

The scaling presented above is known as scaling within the first mode. When scaling across several modes is needed, the problem becomes more difficult since scaling one mode influences the scale of the other modes. If scaling to norm one is needed within many modes, this must be done iteratively until convergence is achieved (Berge, 1989). Another complicated matter is the relationship between centering and scaling. Scaling within one mode, in general, affects prior centering within that mode but not across other modes. Centering across one mode interferes with scaling across all modes (Harshman, 1970). As a result, only centering across arbitrary modes or scaling within one mode is simple, and not all iterative scaling and centering combinations will converge.

Centering can then be performed after scaling and thereby it is assured that the modes to be centered are indeed centered (Bro and Kiers, 2003).

An M-file is included in the Matlab code available on the Internet[1] to run the iterative scaling and centering procedures. Centering across the mode of interest is a typical guideline, however, the aim of centering is to remove constant levels, therefore, data understanding can determine the appropriate preprocessing. The required centering and scaling processes are presented in Figure 4.1 which shows the array unfolded to a matrix . Centering must be performed across the columns of this matrix, whereas scaling must be performed across the rows of this matrix.

[1] https://www.mathworks.com/matlabcentral/fileexchange/1088-the-n-way-toolbox

4.2 Source profiles and contributions

Based on the mathematic equation of PARAFAC (equation 3.9), the source concentration profiles and contribution values were calculated according to the equations below.

As the data needed to be scaled for the PARAFAC input preparation, the scaled 3-dimensions input data is calculated by the equation 4.3:

$$\mathbf{X_{ijk_{scaled}}} = \sum_{r=1}^{F} A_{ir}B_{jr}C_{kr} + \mathbf{E_{ijk}} = \mathbf{F_1} + \mathbf{F_2} + \mathbf{F_3} + \mathbf{E_{ijk}} \tag{4.3}$$

Then, the approximated value (with hat) of each extracted source is obtained by the equations 4.4- 4.6 below:

$$\hat{\mathbf{S_1}} = \mathbf{F_1}. * std(x_{ijk})_J; \tag{4.4}$$

$$\hat{\mathbf{S_2}} = \mathbf{F_2}. * std(x_{ijk})_J; \tag{4.5}$$

$$\hat{\mathbf{S_3}} = \mathbf{F_3}. * std(x_{ijk})_J; \tag{4.6}$$

where $std(x_{ijk})$ is calculated according to equation 4.7:

$$\mathbf{X_{ijk_{scaled}}} = \frac{\mathbf{X_{ijk}}}{\sqrt{\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{x_{ijk}^2}{IK}}} = \frac{\mathbf{X_{ijk}}}{std(x_{ijk})_J} = \frac{\mathbf{S_1} + \mathbf{S_2} + \mathbf{S_3}}{std(x_{ijk})_J} \tag{4.7}$$

The equation 4.7 can be expressed as in equation 4.8:

$$\frac{\mathbf{S_1}}{std(x_{ijk})_J} + \frac{\mathbf{S_2}}{std(x_{ijk})_J} + \frac{\mathbf{S_3}}{std(x_{ijk})_J} = \hat{\mathbf{F_1}} + \hat{\mathbf{F_2}} + \hat{\mathbf{F_3}} \tag{4.8}$$

The notations used here are the following:

A (capitalized, bold)	3D array
A (capitalized, bold, italic)	3D array unfold to a matrix
A (capitalized, italic)	matrix
a (italic)	vector
a	scalar
.*	multiple with correspondence index (element by element multiplication)

Based on these equations, it is possible to estimate the source's profile and its attributable concentration at a given moment. The detailed results are presented in the next section.

4.3 PARAFAC Implementation

This section presents the implementation process of PARAFAC. Firstly, the subsection 4.3.1 briefly reminds the information about input data and some simple statistics. In addition, different analyses on the impact of the presence of occupants, windows opening on the measured concentrations of fine and coarse particles are presented. In the subsection 4.3.2, the PARAFAC implementation procedure is introduced with more detailed information about input data structuring and choosing the number of components/factors.

4.3.1 Input data

Different inputs configuration have been considered: combined PN data (indoors/outdoors) with/without other environmental parameters into different structures (with 3, 4, 5 dimensions, *etc*).

The most important data is the number concentration of particles for different fractions measured every minute during one year (2014). The concentration is presented as a number of particles of a given size range per liter of air (so called PN - Particle Number concentration). There are 15 fractions in total, named as: PN0.35, PN0.45, PN0.575, PN0.725, PN0.9, PN1.3, PN1.8, PN2.5, PN3.5, PN4.5, PN6.25, PN8.75, PN12.5, PN17.5 and PN20 according to their sizes (optical diameter in μm). For each size fraction, the concentration was scaled by dividing it, by the standard deviation of the concentrations of this size fraction (values of the standard deviation is given in Table 4.1 and Table 4.2).

The general statistics parameters (min, max, mean, median, standard deviation (std) and amplitude range) of these PN data according to each size fraction are displayed in the Table 4.1 and Table 4.2 for Outdoor and Indoor environments, respectively.

From September 21st to September 29th, outdoor recorded data are missing due to the dysfunction of the measuring instrument. As a consequence, 1.6% of the full year data (505 571 minutes) are missing. Most of the time, outdoor PN parameters (min, max, mean, median, standard deviation and amplitude range) are higher than indoors.

Table 4.1: Statistics of 1-minute step data of outdoor particle concentration in number (PN -# particles/liter) according to different size fractions measured by a Grimm optical counter in 2014 (January - December: 497586 samples)

PN size (µm)	**0.35**	**0.45**	**0.575**	**0.725**	**0.9**	**1.3**	**1.8**	**2.5**	**3.5**	**4.5**	**6.25**	**8.75**	**12.5**	**17.5**	**20**
Max value	8486417	1178992	354342	88879	26665	6590	2742	2777	1626	1454	3234	890	192	82	103
Min value	1436	105	30	5	0	0	0	0	0	0	0	0	0	0	0
Mean value	50069	14365	4219	996	449	200	99	60	17	7	4	1	0	0	0
Median value	26436	5648	1861	590	320	155	81	50	13	5	2	0	0	0	0
Std value	83841	24621	8259	1660	516	162	72	47	19	11	23	7	2	1	1

Table 4.2: Statistics of 1-minute step data of indoor particle concentration in number (# particles/L) according to different size fractions measured by a Grimm optical counter in 2014 (January - December: 505571 samples)

PN size (µm)	**0.35**	**0.45**	**0.575**	**0.725**	**0.9**	**1.3**	**1.8**	**2.5**	**3.5**	**4.5**	**6.25**	**8.75**	**12.5**	**17.5**	**20**
Max value	1588791	647162	139129	34040	12235	3350	2781	2866	1296	601	336	108	95	19	28
Min value	1315	236	78	15	5	0	0	0	0	0	0	0	0	0	0
Mean value	23722	6554	2176	794	378	140	125	45	13	5	3	1	1	0	0
Median value	14788	3338	1072	422	231	98	93	35	8	3	1	0	0	0	0
Std value	31242	11119	4019	1499	624	165	111	39	15	7	6	2	2	0	0

Table 4.3: Statistics of other parameters monitored during 2014 (other pollutant concentrations, and printer's pulse). *the short name will be used for legending the figures.

	O_3 indoors	CO_2 outdoors	CO_2 indoors	CO indoors	HCHO indoors	Printer Pulse
Unit	ppb	ppm	ppm	ppm	ppb	counts/min
Short named*	O3	C2o	C2i	CO	HC	Pls
No of sample	6388	7045	8759	4383	5033	8171
Max	56.33	553	864	0.79	51	5.1
Min	0	334	424	0	3.0	0
Mean	5.7	415	502	0.20	19	1.4
Median	2.8	410	485	0.18	17	0.6
Std	7.1	28	59	0.11	8.3	1.0

Regarding other environment parameters, Table 4.3 and Table 4.4 present the general statistics of other measured pollutants and meteorological parameters, respectively. It is worth noting that several pollutant levels were not fully recorded for the entire year 2014, particularly wind directions and CO concentrations inside (about 4300 hours per year).

Figure 4.2 presents the hourly values of temperature outdoors and indoors monitored during 2014. One can notice that temperature indoors is more stable, having a smaller amplitude range than the temperature outdoor and the mean value indoors is much higher than outdoors (23.41 °C avaraged indoors in comparison with 15.79 °C avaraged outdoors).

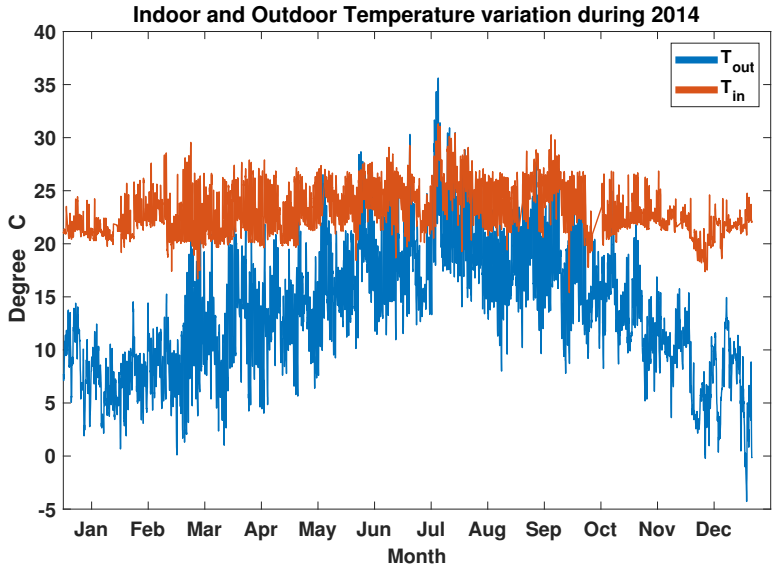


Figure 4.2: Hourly average value of temperature outdoors (in blue) and indoors (in orange) during 2014.

Similarly, the variation of hourly values of specific humidity indoors and outdoors during 2014 is displayed in Figure 4.3. There is no significant difference between the specific humidity values of these two environments.

Table 4.4: Statistics of meteorological parameters in 2014.

	Wind velocity outdoors according to direction (m/s)							
	Est	**Nord**	**South**	**West**	**Nord Est**	**Nord West**	**South East**	**South West**
Short named	E	N	S	W	NE	NW	SE	SW
No of sample	4398	4392	6541	4948	3902	4716	6227	6203
Max	11.30	11.30	16.99	11.66	11.20	12.56	13.21	12.87
Min	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40
Mean	2.44	2.56	2.77	2.74	2.50	2.71	2.61	2.75
Median	2.12	2.28	2.54	2.49	2.10	2.44	2.38	2.51
Std	1.17	1.17	1.32	1.27	1.23	1.34	1.17	1.23

Table 4.4: Statistics of meteorological parameters in 2014 (continue).

	Specific humidity (g/kg)		Temperature (°C)		Irradiance (W/m^2)	
	Hs outdoors	**Hs indoors**	**T outdoors**	**T indoors**	**Ir outdoors**	**Ir indoors**
Short named	Hso	Hsi	To	Ti	Io	Ii
No of sample	8759	8759	8759	8759	8759	8759
Max	17.30	15.11	35.61	31.30	914	402
Min	3.98	4.28	-0.22	14.98	0	0
Mean	9.65	8.88	15.79	23.41	125	3
Median	9.66	8.95	15.83	22.94	6.6	0
Std	2.47	1.91	5.91	2.43	200	16

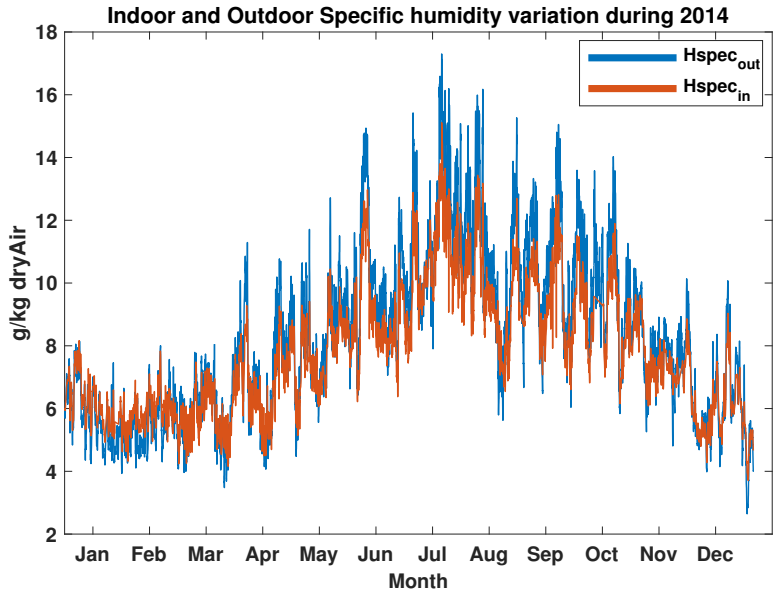


Figure 4.3: Hourly averaged values of specific humidity outdoors (in blue) and indoors (in orange) during 2014.

PN concentration values were categorized according to the conditions of Occupancy and Opening. For comparison, firstly, the averaged number concentration measured values of the two representative size fractions of PN (PN0.725 for the fine particles and PN8.75 for the coarse ones) are presented in the Figure 4.4. Next, Figure 4.5 and Figure 4.6 show the same averaged concentration for these two representative size fractions, but under the different conditions of Occupancy and Windows Opening. All of the averaged concentration values are calculated according to: (*i*) the day of the week, (*ii*) the hour of the day and (*iii*) the month.

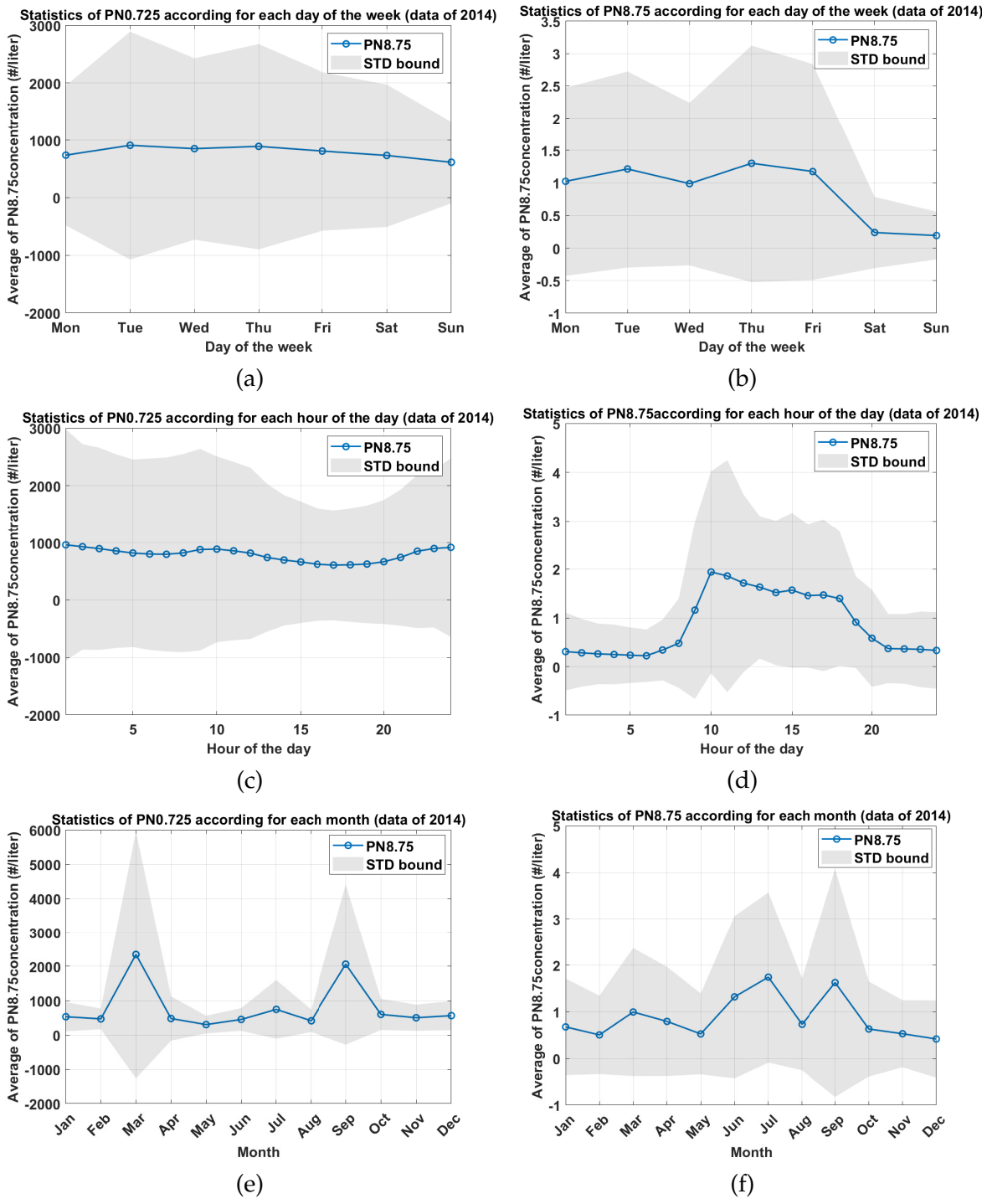


Figure 4.4: Averaged number concentration of PN0.725 for fine particles (left side) and PN8.75 for coarse particles (right side) according to the day of the week, the hour of the day and the month (year 2014).

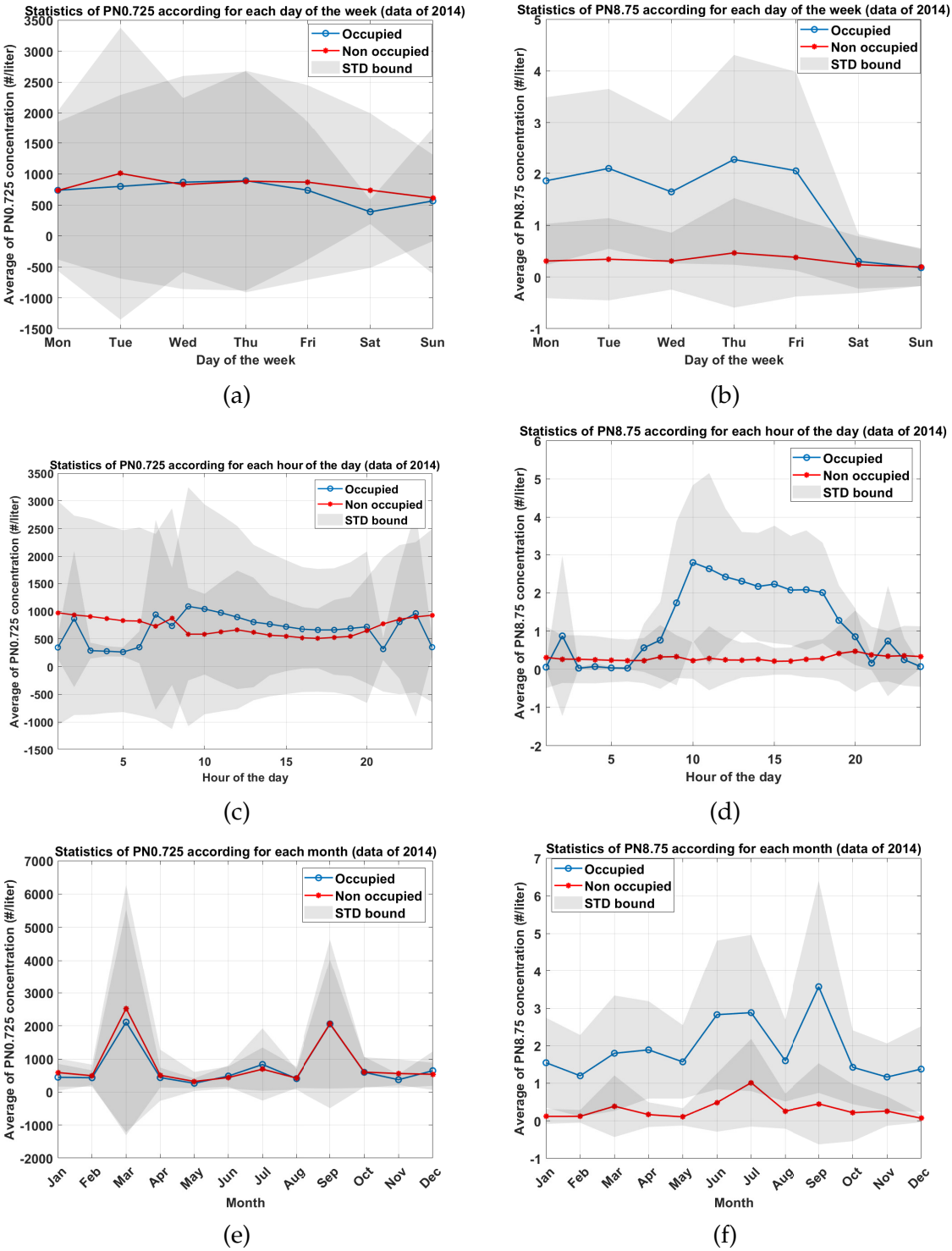


Figure 4.5: Averaged number concentration of PN0.725 for fine particles (left side) and PN8.75 for coarse particles (right side) when the office is occupied and non-occupied, according to the day of the week, the hour of the day and the month (year 2014).

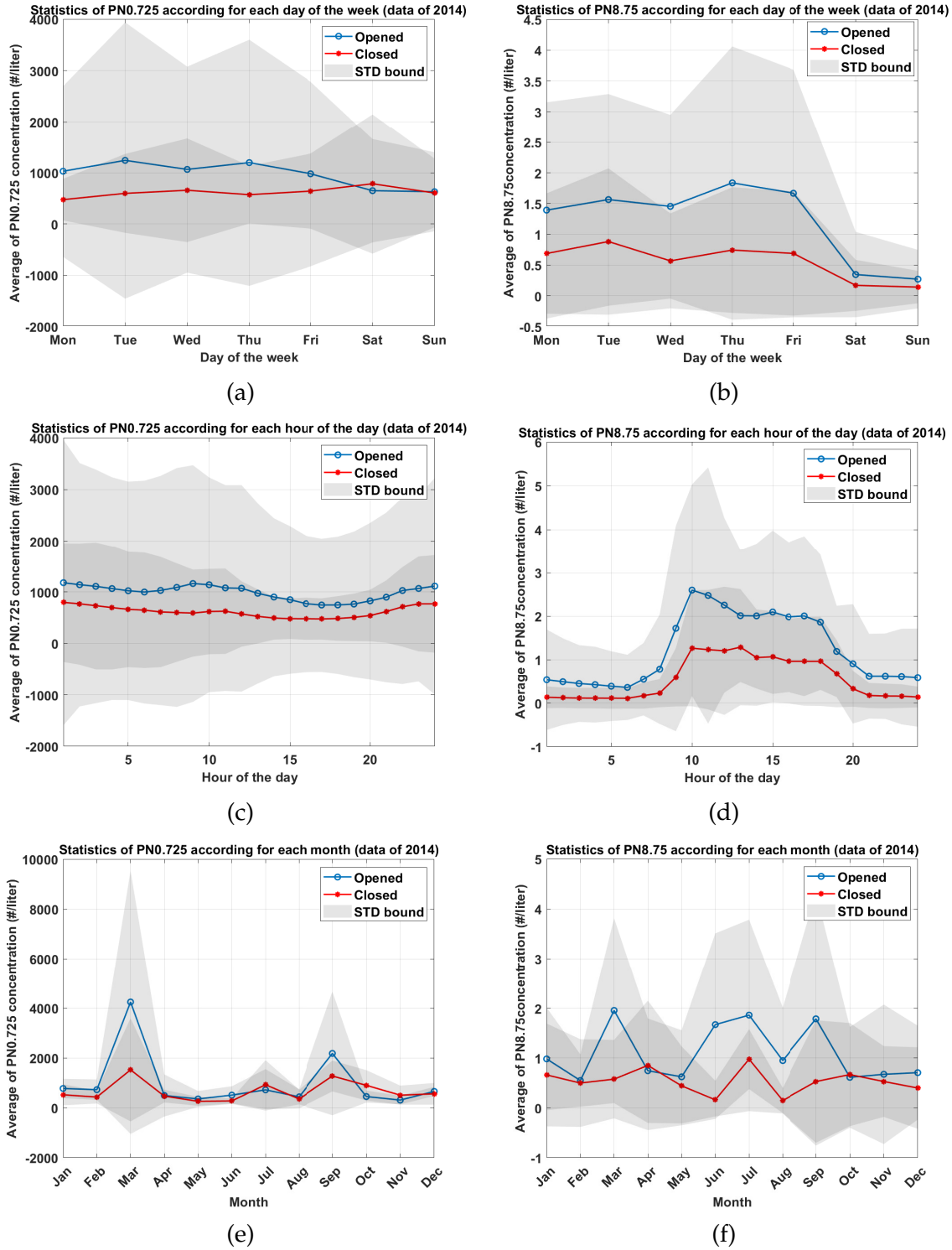


Figure 4.6: Averaged number concentration of PN0.725 for fine particles (left side) and PN8.75 for coarse ones (right side) when windows are opened (at least 1 window is opened) or closed (all of the windows are closed), according to the day of the week, the hour of the day and the month (year 2014).

Several remarks can be made from the figures:

- **The influence of the occupancy**
  - The status of occupancy does not seem to have a significant impact on the fine PN concentration according to the day of the week, hour of the day or month (see Figure 4.5a, 4.5c and 4.5e). Similar situations are obtained for these other fine PN fractions (up to PN1.8 – see the detailed results in the Appendix section).
  - In contrast, coarse size PN is much more affected by the occupancy. During the working day (Monday – Friday), the hourly averaged number concentration of PN8.75 is approximately 4 times higher when the office is occupied than when non-occupied. Similarly, the working hour period (9 a.m. – 7 p.m.) with occupancy shows a much higher value of coarse size PN than without occupancy. The monthly averaged number concentration of PN8.75 (coarse particles) is also higher considering the occupancy (around 2 particles/L) or the non-occupancy period (around 0.4 particles/L).
- **The influence of the windows opening state**
  - There are differences in the overall trend of fine PN (PN0.725)and coarse PN (PN8.75) according to the day of the week (Figure 4.6a and 4.6b) and the hour of the day (Figure 4.6c and 4.6d), in both cases Opened and Closed.
  - Both fine PN and coarse size PN have higher concentrations when windows are opened, excepting during the weekends and from October to December.
  - Similar to the impact of occupancy, when the windows opening status is changed from Closed to Opened, the number of coarse size PN is much higher during the working hour period and working days.
  - The concentration of fine particles according to the hour of the day is higher during nighttime and lower during the daytime in both cases (opened and closed windows). This is very similar to the trend of fine particulate when the office is non-occupied.
- **The combined influence of occupancy and windows opening state**

  In addition, the averaged number concentration of these two size fractions for the combined situations of Opening state and Occupancy status are displayed in Figure 4.7. The total of four cases: opened/occupied, opened/non-occupied, closed/occupied and closed/non-occupied were studied in order to see which factor has the most important impact on the concentration of particles. Similar remarks could be obtained from this figure where higher levels of coarse particles were correlated with the Occupancy and both fine and coarse particles were affected by the opening of the windows.

  Morever, one can notice that there were two clear peaks of fine particles on Sunday (Figure 4.7a) and at 11 p.m. (Figure 4.7c) for closed/occupied case

(green color). In fact, these results were not representative as there were only two samples that met the conditions of: Sunday, occupied (maybe the guard round) and closed windows. One sample took place at 4 a.m. of January 5th (321 particles/L) and another one at 11 p.m. of 9th March (6250 particles/L – during pollution episode). The latter sample also was the reason for a peak at 11 p.m. on Figure 4.7c.

Based on these observations, one can conclude that:

- Both fine and coarse particles are highly affected by windows opening (higher infiltration of outdoor particles);
- Meanwhile, coarse size particles are mainly correlated with occupancy.

This information could be useful to explain the time variation in the sources' profiles which are extracted by PARAFAC.

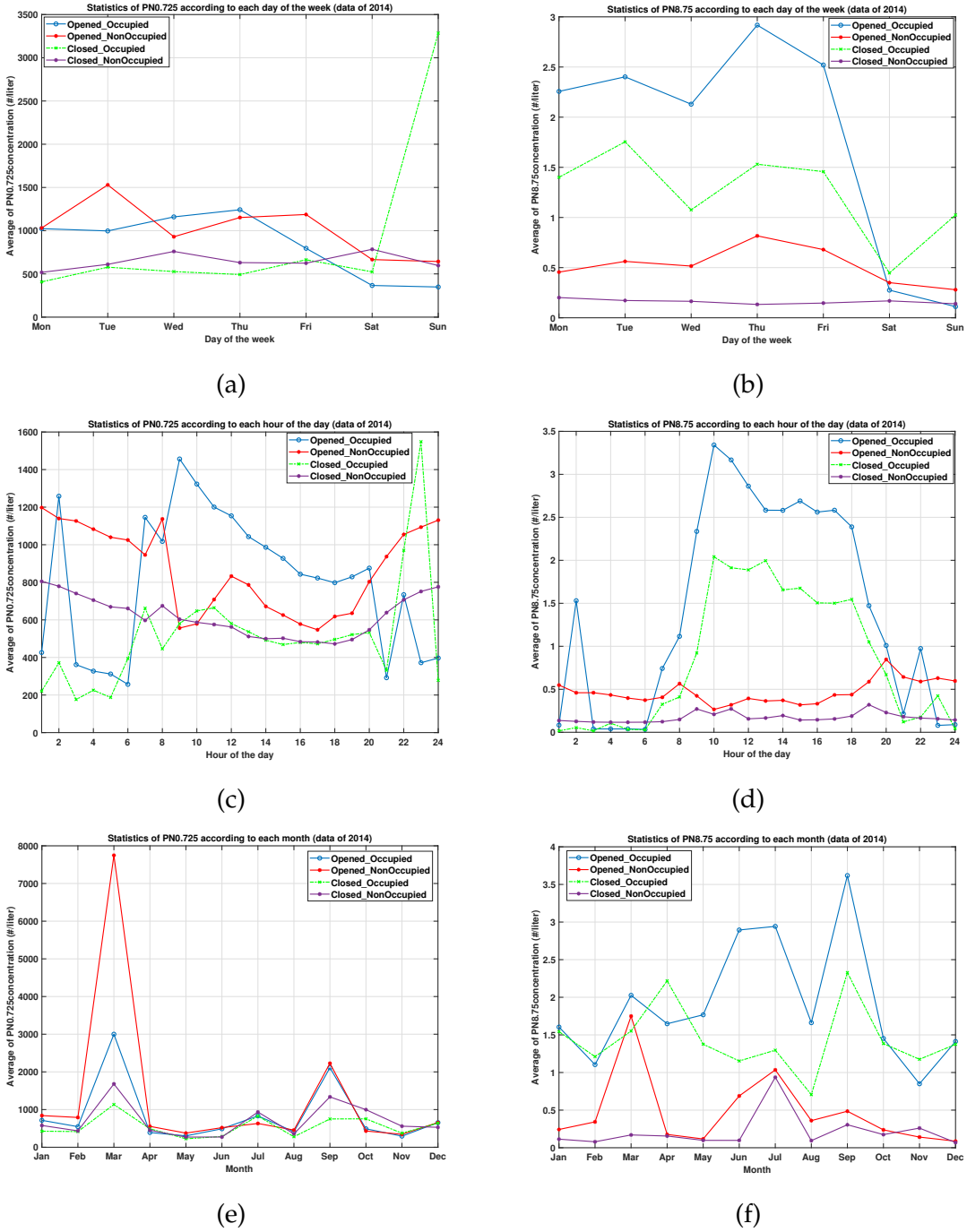


Figure 4.7: Averaged number concentration of PN0.725 for fine particles (left side) and PN8.75 for coarse particles (right side) in the different cases of window and occupancy status, according to the day of the week, the hour of the day and the month (year 2014).

4.3.2 Implementation

For the PARAFAC implementation, the following steps were performed: data pre-processing (1), data structuring (2) and then, choosing the optimal number of components (3).

1. **Data pre-processing**.
   - Hourly averaged values were used.
   - In order to replace missing values, a linear interpolation was applied.
   - Then, the scaling within the variable mode (PN fractions, meteorological parameters and pollutant concentrations) was performed according to equation 4.9. Scaling offers the same possibility or weight to each variable to contribute to the model, avoiding the fact that the variables have different units and different amplitude ranges:

$$x_{ijk_{scaled}} = \frac{x_{ijk}}{\sqrt{\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{x_{ijk}^2}{IK}}} \tag{4.9}$$

   where x_{ijk} represents the concentration of particles expressed in number/L for all the size of PN and climatic parameters and other measured pollutants concentrations (j up to J) for the I day sample (i up to I) and for the hour k from 1 to 24 (K).

2. **Data structuring**.

   Depending on the aim of the analysis, a n-layer tensor is built containing the values of the PN fractions and additionally some other data. After this step, the n-dimensional array is obtained as input data of PARAFAC.

3. **Determining the number of components**.

   To select the most suitable number of components or factors for the PARAFAC decomposition, many criteria can be used, such as: the variance explained by the model, the visual appearance of loadings, the number of algorithm iterations, and the core consistency diagnostic (CORCONDIA) (Bro and Kiers, 2003). CORCONDIA is one of the most frequently applied techniques in the literature (Andersen and Bro, 2003). CORCONDIA's estimate of the number of components, however, remains challenging in the context of complicated data. As a result, rather than being dependent on a single diagnostic tool, it is generally advised to use several diagnostic techniques be used in combination (Harshman, 1970).

   The determination of the number F of components (or factors) is challenging, and no method that provides clear values has yet been discovered. When F is too small, not all of the effects in the input data are identified. However, if F is too large, noise is more modeled, and the observed effects in the data are characterized by coupled components (Bro, 1997). The idea is to increase the number of components F until the decline in the residual error diminished sufficiently and there is no need to incease the number of components because the error decrease is not significant. The model with the fewest components was then picked as the one capable of explaining the most variance without correlation among the components.

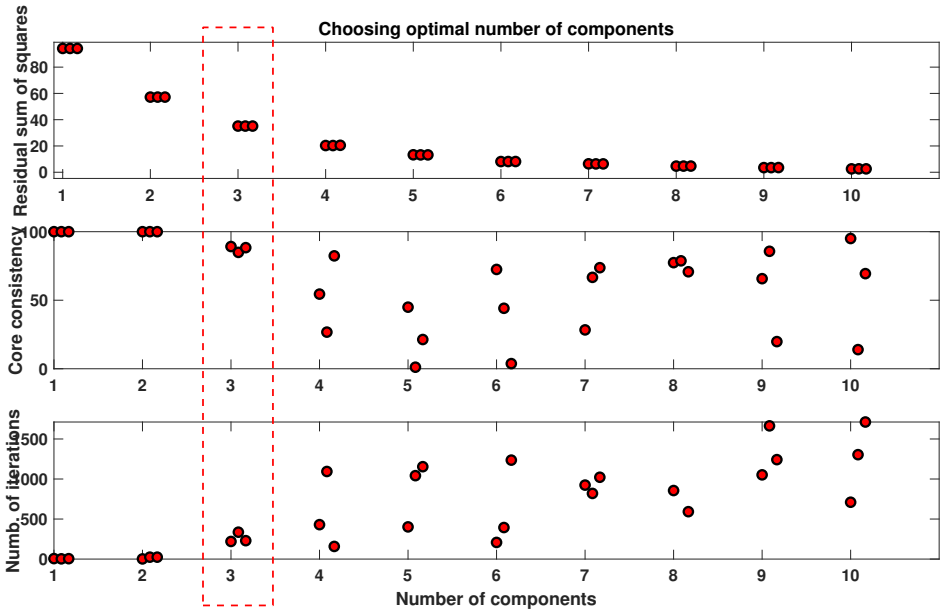


Figure 4.8: An example of PARAFAC diagnostic for a 3-dimensions PN fractions. Each number of components is fitted 3 times.

In this study, the core consistency diagnostic method was applied to choose the optimal number of components. The most suitable number of components is chosen as the highest number with a valid core consistency value (from 80% to 100%). An example of the core consistency diagnostic is displayed in Figure 4.8 where not only the information about core consistency but also the sum of the squares of the residuals and the number of iterations are taken into account.

4. **Results post-processing**. After the data construction and the diagnostic step to choose the suitable number of components, the obtained PARAFAC outputs were analyzed by different data analysis methods or signal processing techniques, in order to help the interpretation and the identification of the determined factors as possible sources (or cluster of sources) of the particle concentration variability. For example, autocorrelation functions (ACF) were applied to the time profiles of the retrieved factors. The contributions of the obtained "sources" to the monitored concentration values (data input) were furthermore calculated.

Chapter 5

Different data cases: Implementation, Results and Discussion

Four study cases with different structures of input data were constructed and the corresponding output results using PARAFAC are presented and analyzed in this chapter. Other data analysis or signal processing methods were used to help the source identification and to explain the variation of each source obtained by PARAFAC. In the end of the chapter, the results are discussed and some elements are given to conclude this analysis.

5.1 Indoor data

This section presents the results of the two first input structures: (*i*) the 3D-array (tensor) containing the different size of PN concentrations indoors and (*ii*) the 3D-array containing, in addition, some other supplementary variables indoors, too.

5.1.1 Case 1: Only particulate matter data

We structured the variation of the measured particles indoors according to daily variations taken as samples and hourly variations taken as events. Therefore, a 15-layer tensor containing the values of PN for the 15 size fractions structured according to 365 days and 24 hours events, has been structured in a 3D-array of $365_{days} * 15_{PNfractions} * 24_{hours}$, the input data of PARAFAC.

In this case, a PARAFAC model for this three-dimensional data of measured particle indoors is illustrated in Figure 5.1. The three output matrices represent the loading vectors of 3 modes: (A) day of the year, (B) PN fraction and (C) hour of the day while the tensor E contains the modeling residuals.

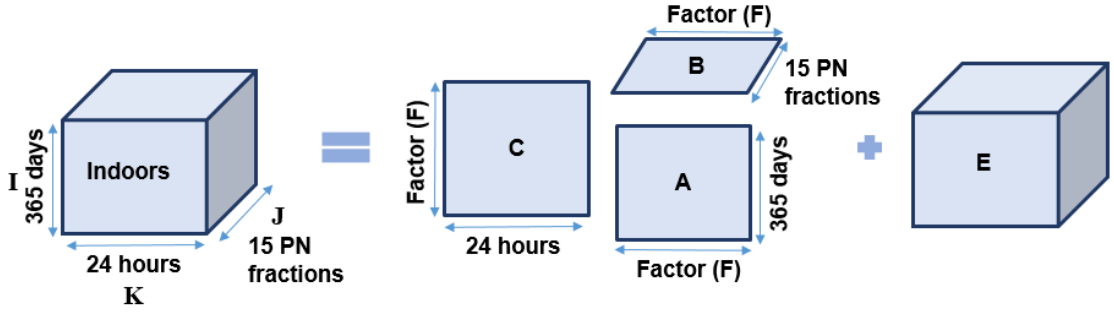


Figure 5.1: The PARAFAC model for three-dimensional data of indoor particle measurements (PN indoors)

Regarding the most suitable number of components, the diagnostic lead to a choice of 3 factors, with a value of the core consistency of 93% (see Figure 5.2).

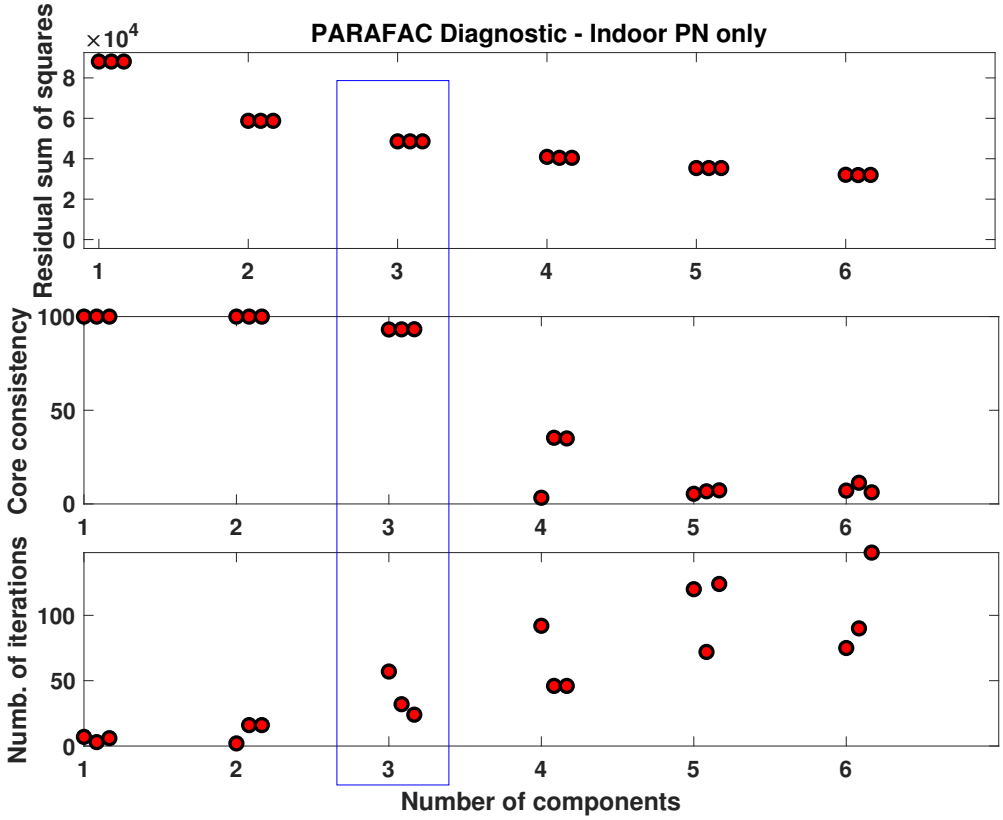


Figure 5.2: PARAFAC diagnostic for indoor particulate matter input.

Based on the three loading matrices which are the output of PARAFAC, it was possible to identify the potential sources that jointly contribute to the indoor air particle levels. The summary and the detailed loading matrices are displayed in Figure 5.3 and Figure 5.4, respectively.

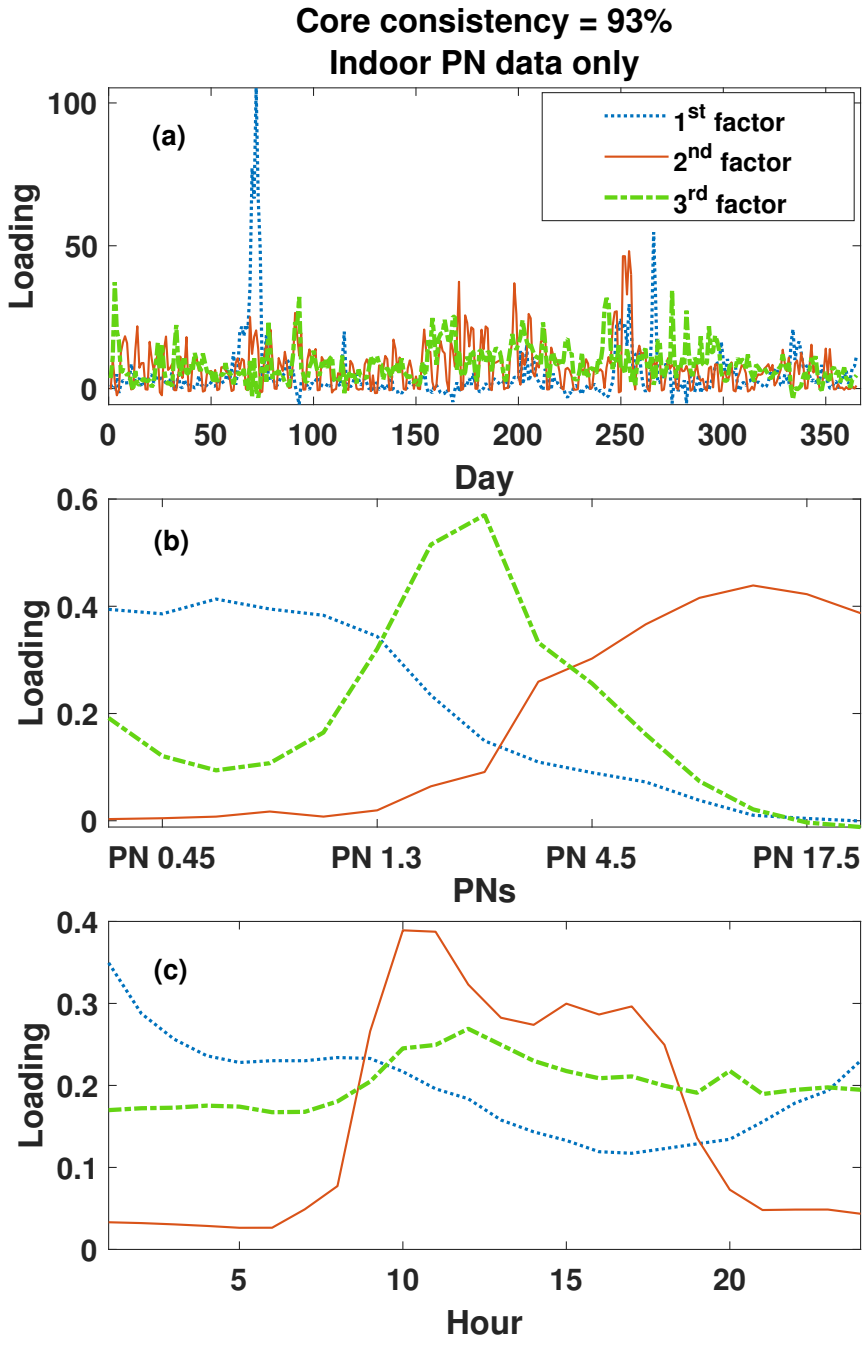


Figure 5.3: The PARAFAC outputs for indoor particulate matter input.

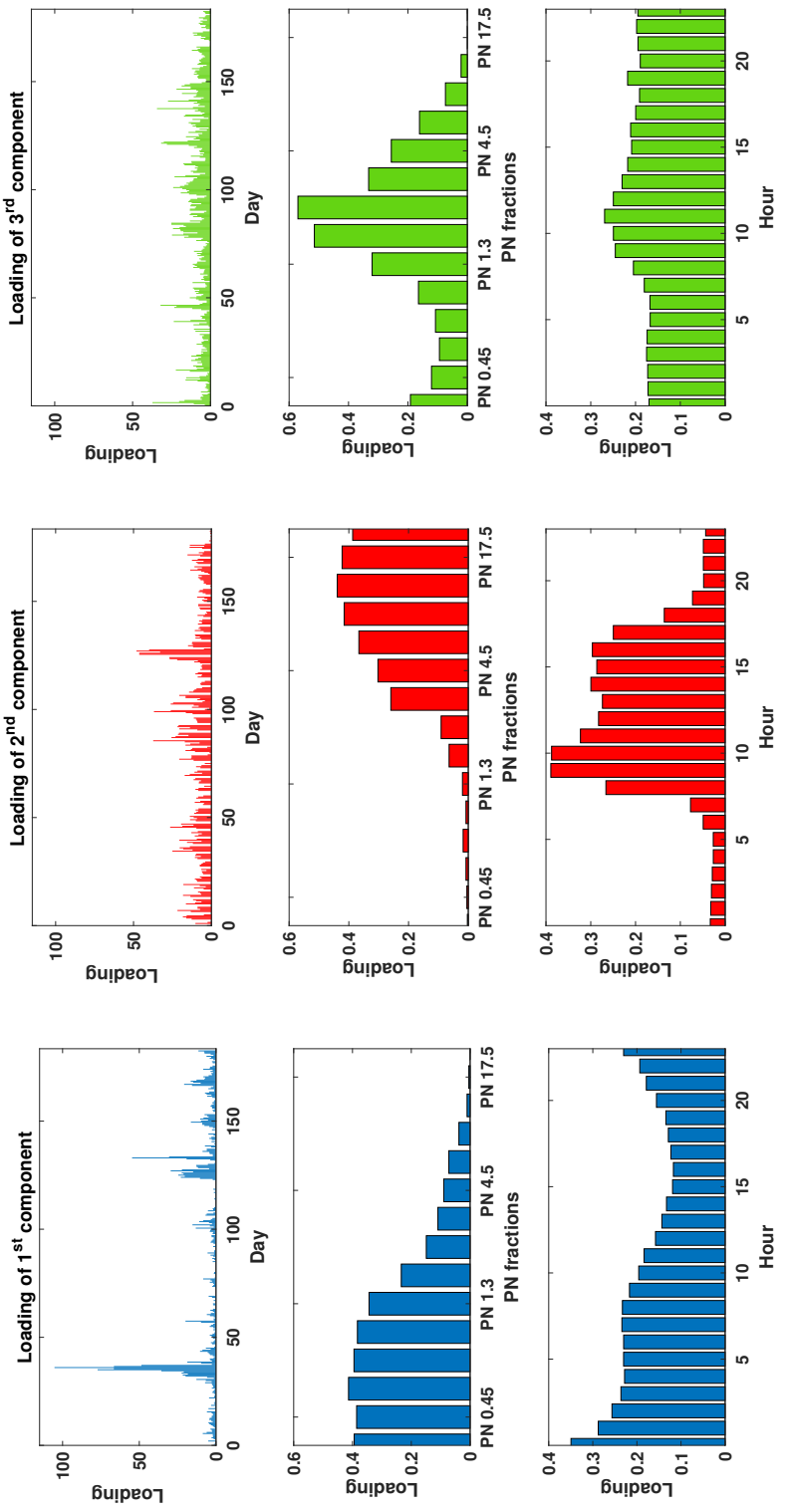


Figure 5.4: Detailed loadings of the three output matrices for indoor particulate matter input (first line: matrix A, daily profiles, second line: matrix B, contributions according to the size fraction, third line: matrix C, hourly profiles). Each column corresponds to a factor or component.

In addition, the autocorrelation functions for the three daily profiles obtained in the first loading matrix A and corresponding to Figure 5.3a and to the first line of the Figure 5.4 are displayed in Figure 5.5.

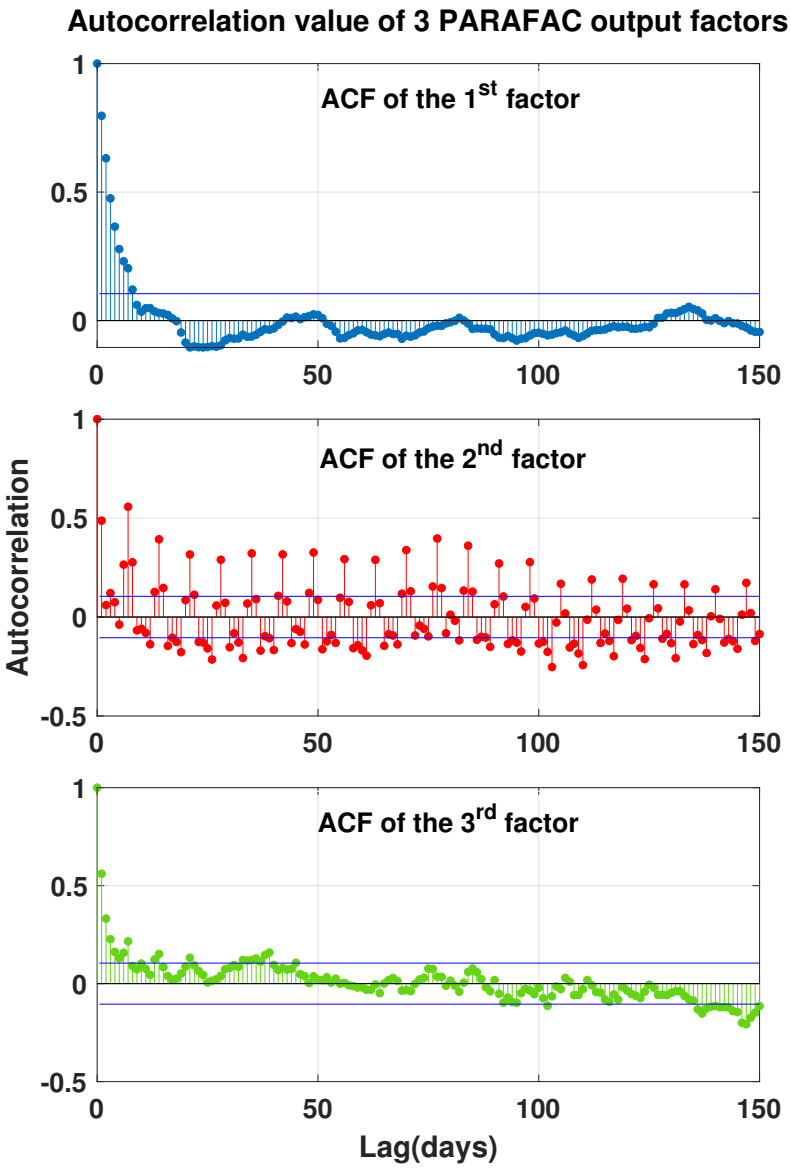


Figure 5.5: Auto-correlation functions for the three daily profiles obtained in the first loading matrix A, for the first (top of the figure), second (middle), and the third (bottom) factors.

Figure 5.5 shows that the first daily profile corresponding to the 1st factor has a high autocorrelation value at one-day lag (daily periodicity). In addition, according to PN fractions size (see Figure 5.3b and second line, first column of the Figure 5.4), the loading profile of this factor shows that it is associated mainly with the small size fractions (mainly PN 0.35 - 1.3 μm). Figure 5.6 shows the correlation between the previous day outdoor concentration and the present indoor values of concentrations as daily means. The determination coefficient R^2=0.61

suggests that 61% of the variance of the first factor can be explained by the fine particles coming from outdoors, which are found, some of them (≈61%), indoors one day later.

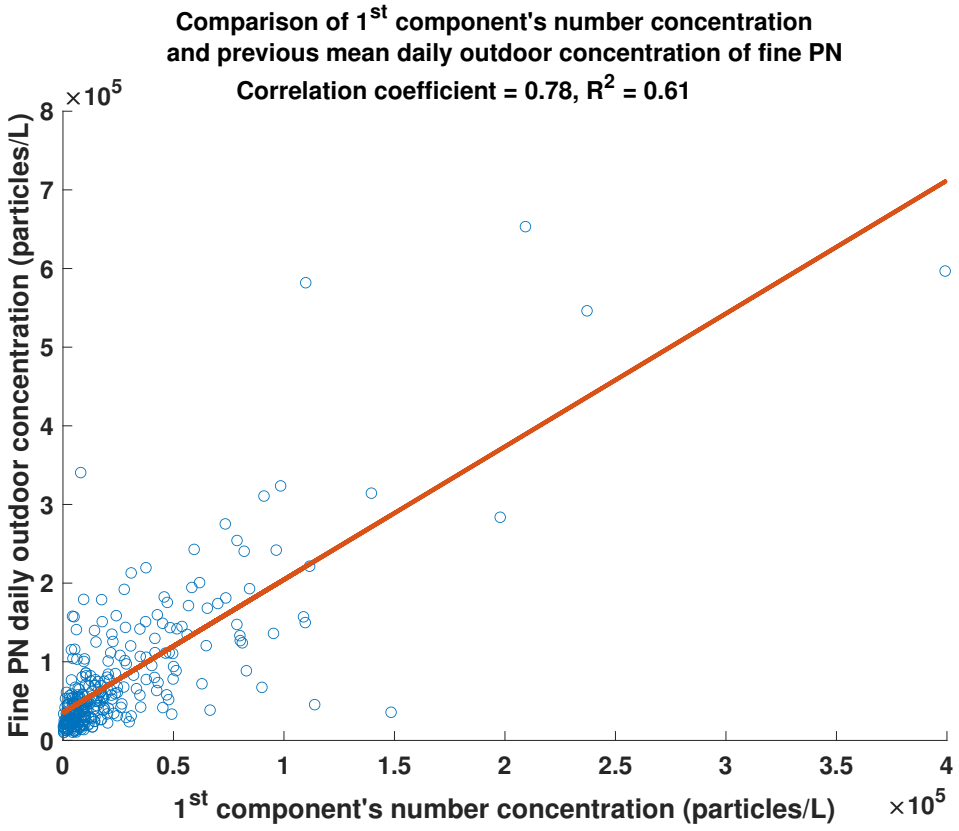


Figure 5.6: Correlation between the previous day concentration of fine PN daily averaged values monitored outdoors and attributable number concentration (daily averaged) of the 1st component.

Moreover, in the monthly profile associated with the first factor, the high values observed in March are related to a specific outdoor pollution episode that occurred within all the area. According to the report from the European Environment Agency, "The factors leading to such high concentration levels were a combination of meteorological conditions (stable and calm weather, which prevents air pollution from dispersing; and relatively high temperatures during the daytime for the period) and various emissions sources" (2014). This allows us to associate the first factor mainly with outdoor particle sources. High indoor pollution concentrations were observed when this outdoor pollution episode occurred, showing the impact of the **outdoor environment (sources)** on the indoor environment. This impact is higher for finer particles bacause they can penetrate also by infiltration even if windows are closed. The penetration seems to be slightly higher during the nighttime (see Figure 5.4a bottom), but it is rather uniform according to the hour.

A 7 days periodicity is detected for the second factor loading, according to the autocorrelation value (see Figure 5.5 middle). In addition, its loading is very high during the daytime (8 a.m. – 8 p.m.) in comparison with nighttime (8 p.m. –

8 a.m.), as shown in Figure 5.3c and Figure 5.4b. This is similar to the trend and periodicity of CO_2 indoor concentration, which traces the presence of **occupants indoors**. The PARAFAC loading output also shows that this component concerns the coarse size particles, see Figure 5.3b and Figure 5.4b, especially sizes higher than 4.5 μm. The second factor can thus be attributed to indoor sources in particular those related to the occupants and their activities. It is known that the occupants produce coarse particles indoors by walking and cleaning, *etc*.

The third factor includes medium size particles but it does not have an obvious identification. This factor has no specific trend or periodicity, but it is not random because it is structured in a way. According to the model's loading output, it is slightly increasing during daytime and is associated with middle range particle sizes between 2 and 4 μm. This factor can be associated with a group of particle sources which could not be unmixed. It thus corresponds to **unexplained variations**.

Figures 5.7 represents the time profiles of the attributable hourly concentration of each source, based on the calculations given in Section 4.2 (regression by multiple with standard deviation values). The time profiles help in improving the identification results, while the first component presents the outdoor-peak event during March, the second component clearly shows a weekly profile and the third component does not present any specific trend.

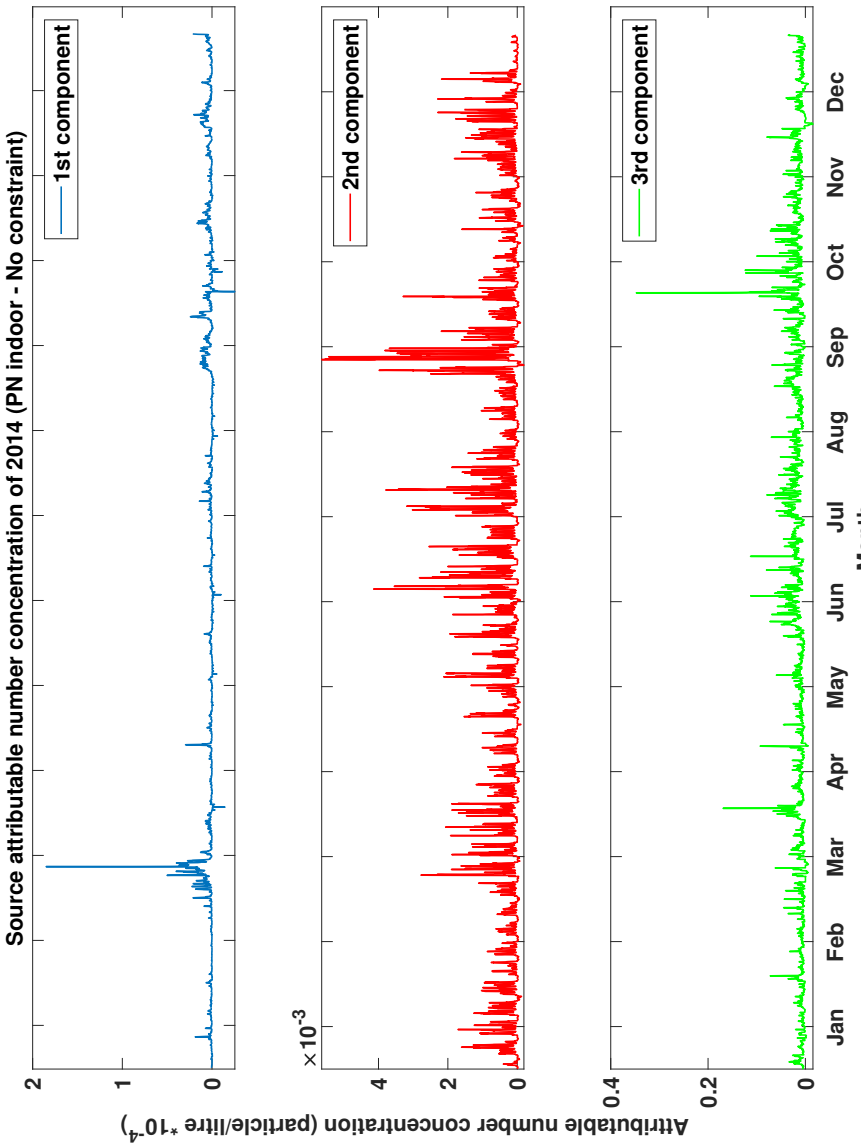


Figure 5.7: Time profile of the attributable hourly concentration of each source in number of particles/liter

The PN concentration attributable of the second factor ('occupants') was compared with the concentration of CO_2 indoors, as the CO_2 concentration is the most appropriate parameter related to the human presence indoors. The comparisons of the variations of these two concentrations allover the year 2014 and specifically in September 2014 are displayed in Figure 5.8 and Figure 5.9, respectively. One can notice the synchronization of these two profiles. This result makes the identification of the second factor more reliable.

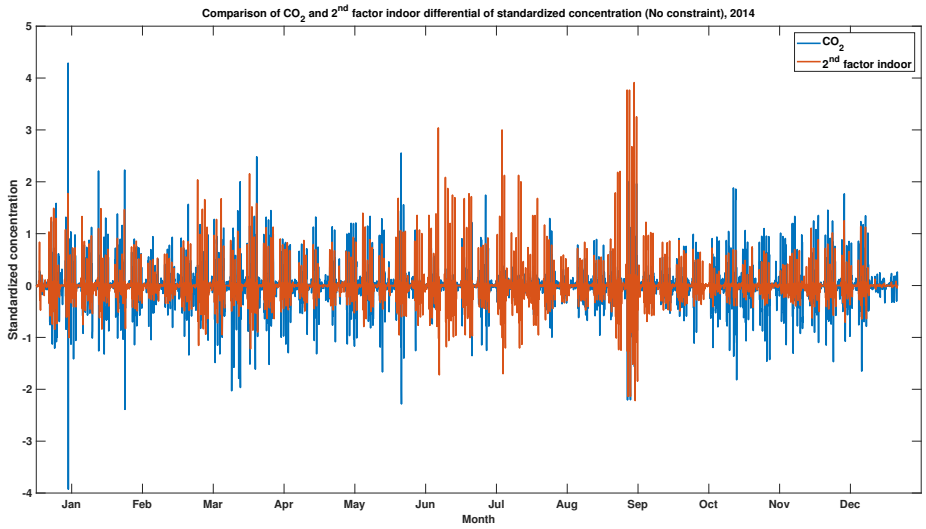


Figure 5.8: Comparison of differential of concentration of CO_2 indoor and 2^{nd} factor extracted by PARAFAC during 2014.

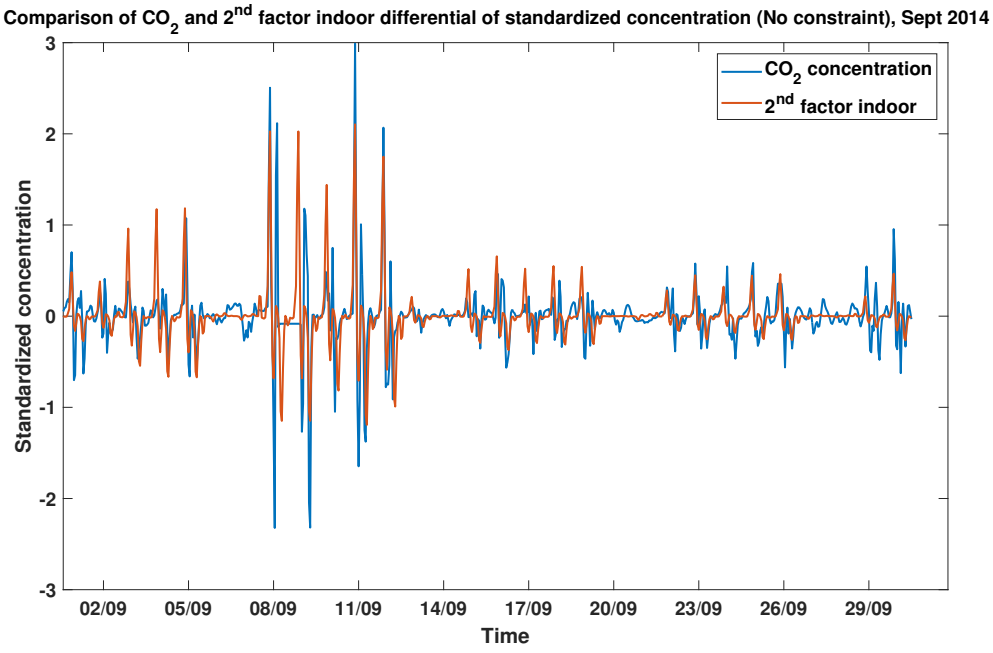


Figure 5.9: Comparison of differential of concentration of CO_2 indoor and 2^{nd} factor extracted by PARAFAC, September 2014.

In order to estimate the source contributions, the attributable concentration in number profile should be transform as an attributable mass concentration (using the equations in the section 2.2). Figure 5.10 presents the attributable PM10 concentration of the three extracted components, obtained by the conversion from concentration in number to mass concentration. The original measured concentration in number has also to be converted in mass concentration (PM10).

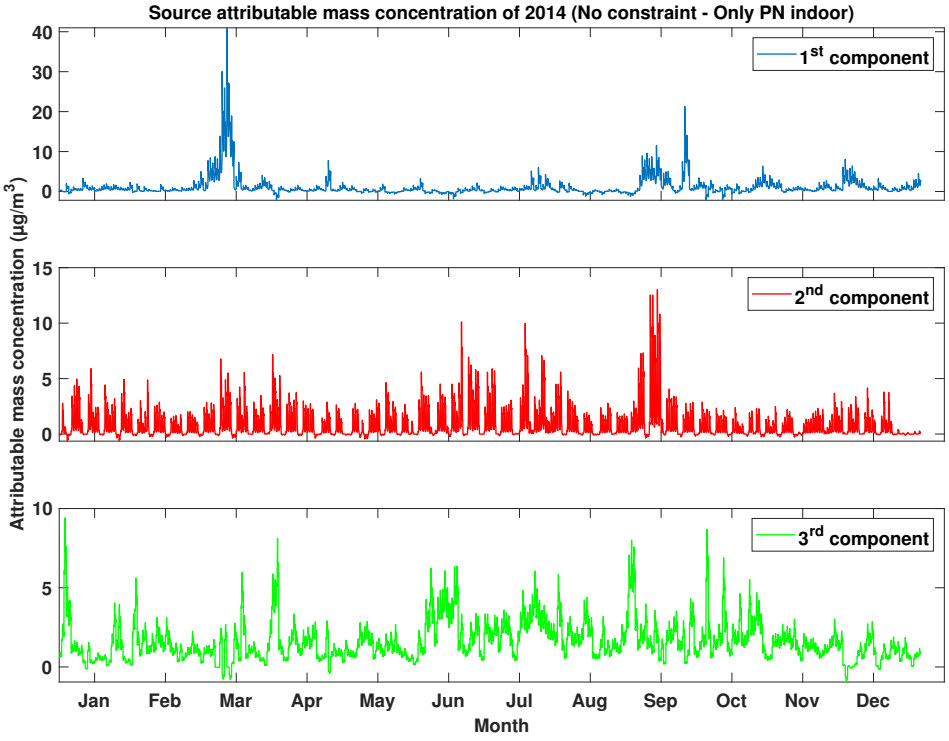


Figure 5.10: Time profile of the attributable hourly mass concentration of each source during 2014.

The regression between the 1st attributable PM10 mas concentration and the original PM10 showed that 40% of the total variance is explained by the 1st component. Similarly, the 2nd component explained 26% of the total variance. Meanwhile, the 3rd component accounted for only 6%.

In conclusion, based on the autocorrelation value of the factor attributable concentration, we were able to identify two major sources. The result of PARAFAC and its factor concentration gives us an acceptable result of identification. The three factors were tentatively identified as the following sources: outdoor inputs (1st component) contributing to 40% of the variability of the indoor air pollution and more specifically concenrning the fine particles; indoor occupancy and related activities (2nd component) contributing to 26.3% of the variability of the indoor air pollution and more specifically concenrning the coarse particles; and finally, a source or a group of non-identified sources with a very specific profiles or submitted to a non-linear mixture (3rd component) which contributes only 5.8% of the variability of the indoor air pollution.

5.1.2 Case 2: All indoor data

Similar to the structure of PARAFAC input in the previous subsection, with the input data including all of the available indoor data, a 3-dimensional array of $127_{days} * 31_{variables} * 24_{hours}$ is obtained. As the data of other pollutants are not fully recorded during the whole year, it was possible to examine only the period of June – October 2014 (127 days), which is the period when all the variables are fully measured.

An illutration of the PARAFAC model for this three-dimensional data is presented in Figure 5.11. The three output matrices containt the loading vectors of the three modes: (A) day of the year, (B) variables (15 PN fractions; CO, O_3, CO_2 and HCHO concentrations; 8 wind direction and speed, Printer Pulse, Irradiance, Temperature and Specific humidity) and (C) hour of the day.

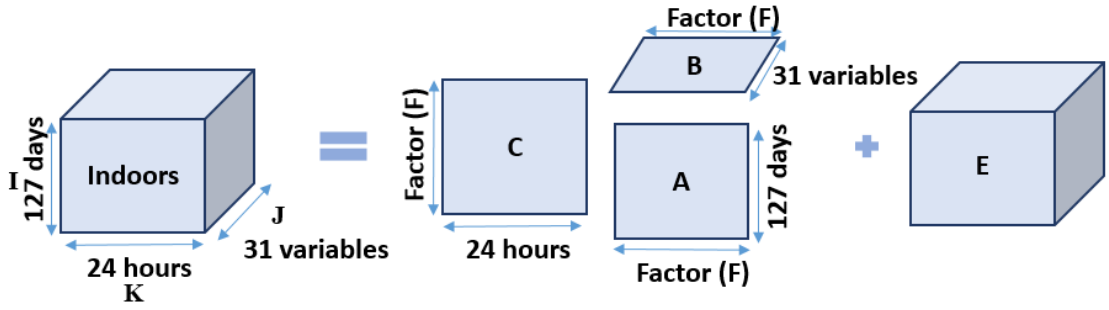


Figure 5.11: The PARAFAC model for three-dimensional data of all indoors measurements (PN, other poluttants concentrations indoors and climatic parameters).

Once again, the most suitable number of components was determined to be three, with the value of core consistency of 84.4% (see Figure 5.12).

The three loading matrices output of PARAFAC are displayed in Figure 5.13 and detailed in Figure 5.16. It can be observed that when including all of the standardized measured data indoors as inputs, the outputs of PARAFAC for PN (considering second and third factor) are quite similar to the previous model outputs (when using only PN indoors data as input) but with some additional information. These new results concerning the consistency of PARAFAC indicate again a best selection for three factors, as in the case when using only PN data indoors. Again, the component related to **occupants indoors** is easily detected by its high loading during working hours (8 a.m. – 8 p.m.) in the hourly mode and weekly periodicity in the daily mode.

In addition, the second component shows high loadings for coarse size particulate matter, printer pulse value and CO_2 concentration. These results are expected, as printer pulse is related to the use of the printer that occurs only when occupants are present and CO_2 is directly emitted by humans in the office.

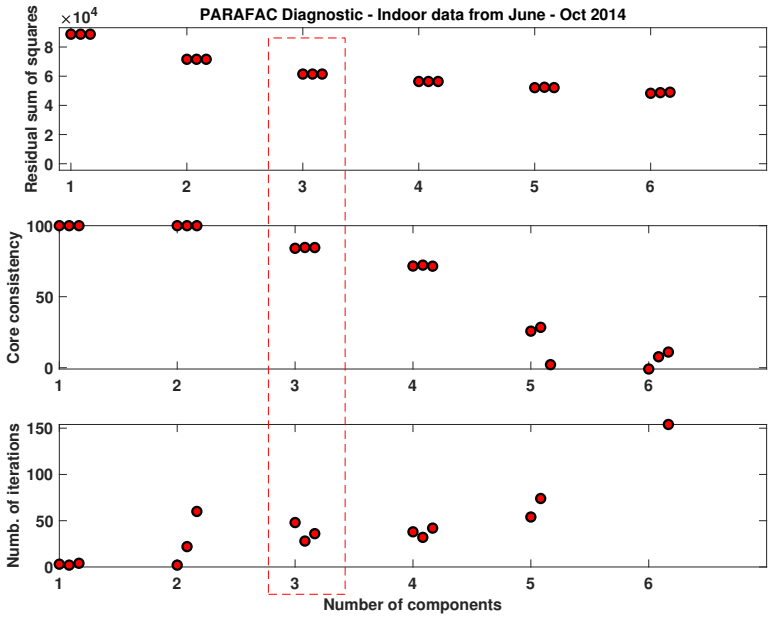


Figure 5.12: The core consistency diagnostic for all indoor data from June to October 2014.

The first component has a high loading profile for temperature and humidity. Besides, the significant high loadings for CO_2 and formaldehyde (HCHO) concentrations are also noteworthy. This means that these parameters (temperature, humidity, CO_2, HCHO) have an important impact on it or are significantly correlated to it. However, this component shows very low loadings for all the PN fractions and thus seems not being related to particulate matter. Similarly, this component has a constant loading around 0.2 across the whole day and around 80 during the whole period, so it does not present any time variability. This factor seems related to indoor data other than PN characterizing the thermal comfort inside the office (temperature, humidity) and other pollutant such as CO_2 and HCHO.

The third component is correlated with fine particles and has a slightly higher loading in the morning than in the afternoon. This trend is similar to the first component - **'outdoor environment' sources** of the previous section when only PN indoors concentration were taken into account.

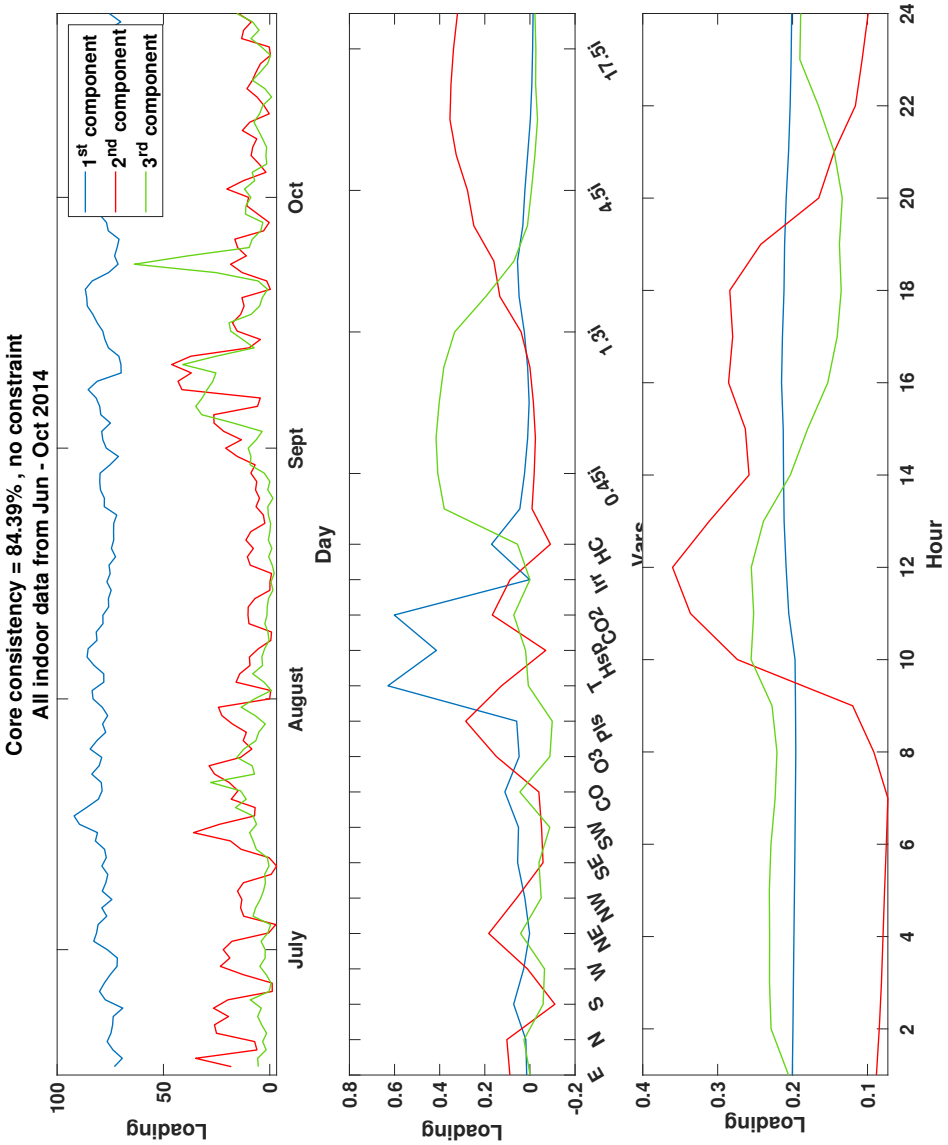


Figure 5.13: The PARAFAC loading outputs for all quantitative indoor data from June to October 2014.

As no constraint has been imposed in the PARAFAC model, negative values were obtained for the 2nd component as presented in Figure 5.14. Due to this reason, the model was re-implemented by applying the non-negativity constraint, the summary and detailed loadings and the attributable number concentration of the three new components are displayed in Figure 5.15, Figure 5.16 and Figure 5.17, respectively (the 2nd and 3rd component order are different in comparison with the 'no constraint' results).

According to the loading results of PARAFAC with non-negativity constraint, one can easily notice that the 3rd component significant by related to **'occupants' indoors**. This component shows high loadings during working hours and it is mainly correlated with coarse particles. In addition, the time profile of the attributable number concentration of 3rd component is also lowest during August as this month is the time of holiday, and so not many occupants were in the office.

Again, the 1st component remains unexplained and the 2nd component can be associated with **'outdoor environment' sources**.

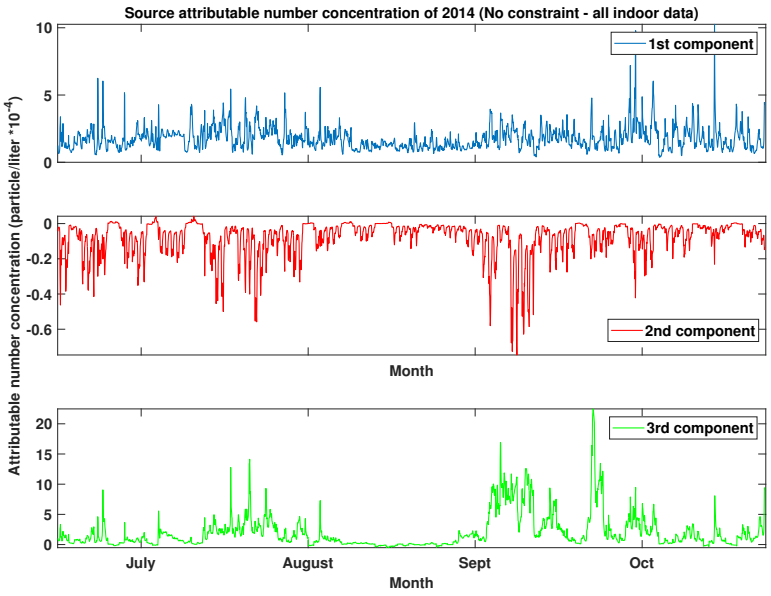


Figure 5.14: Time profile of the attributable concentration to each source in number of particles/liter for all indoor data input. No constraint was applied for the PARAFAC model.

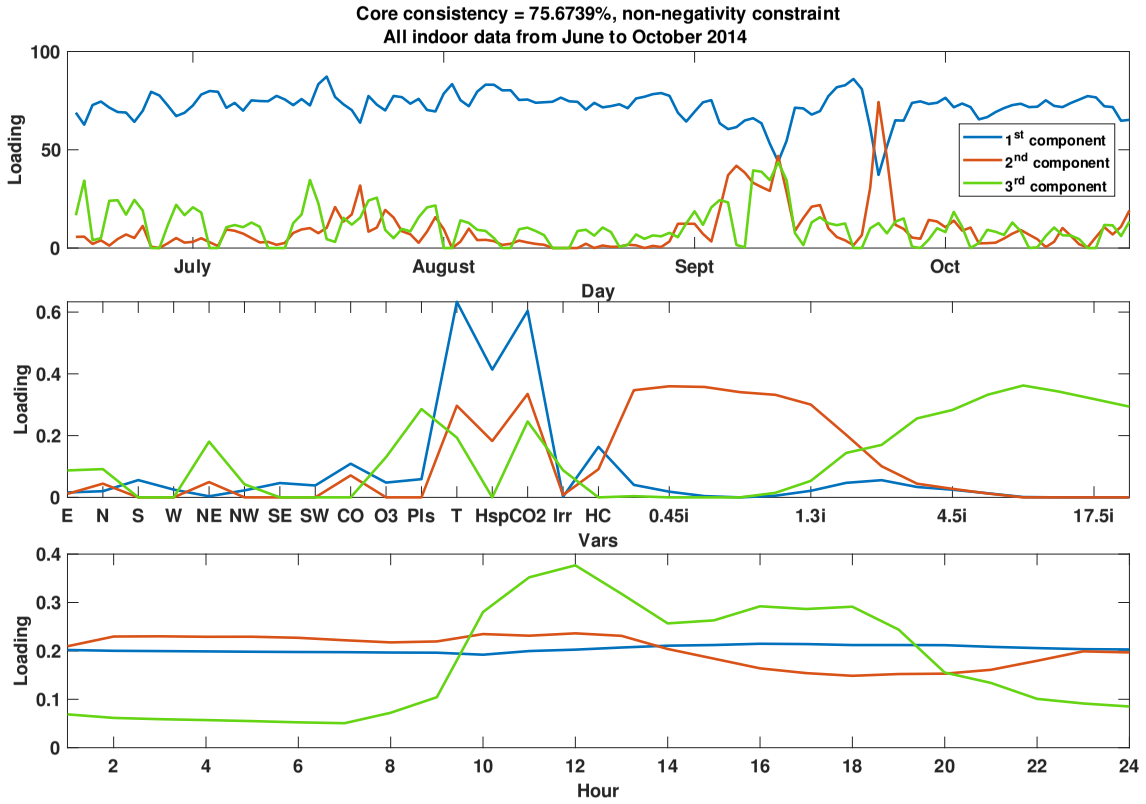
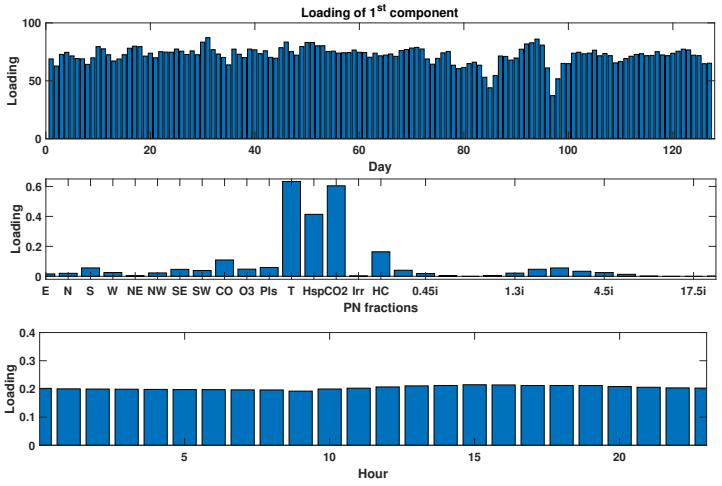
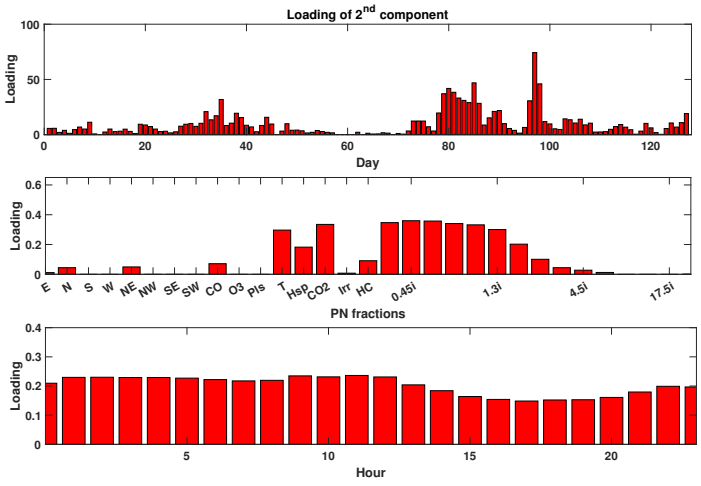


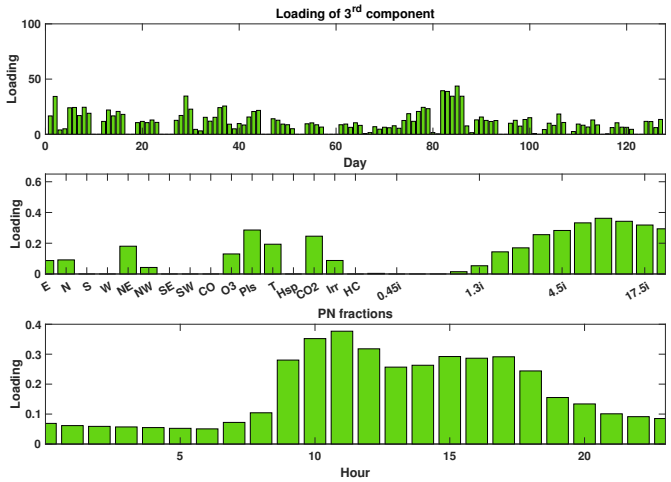
Figure 5.15: The PARAFAC loading outputs for all quantitative indoor data from June to October 2014. The non-negativity constraint was applied to avoid negative values.



(a)



(b)



(c)

Figure 5.16: Detailed loadings of the three output matrices for all quantitative indoor data from June to October 2014.

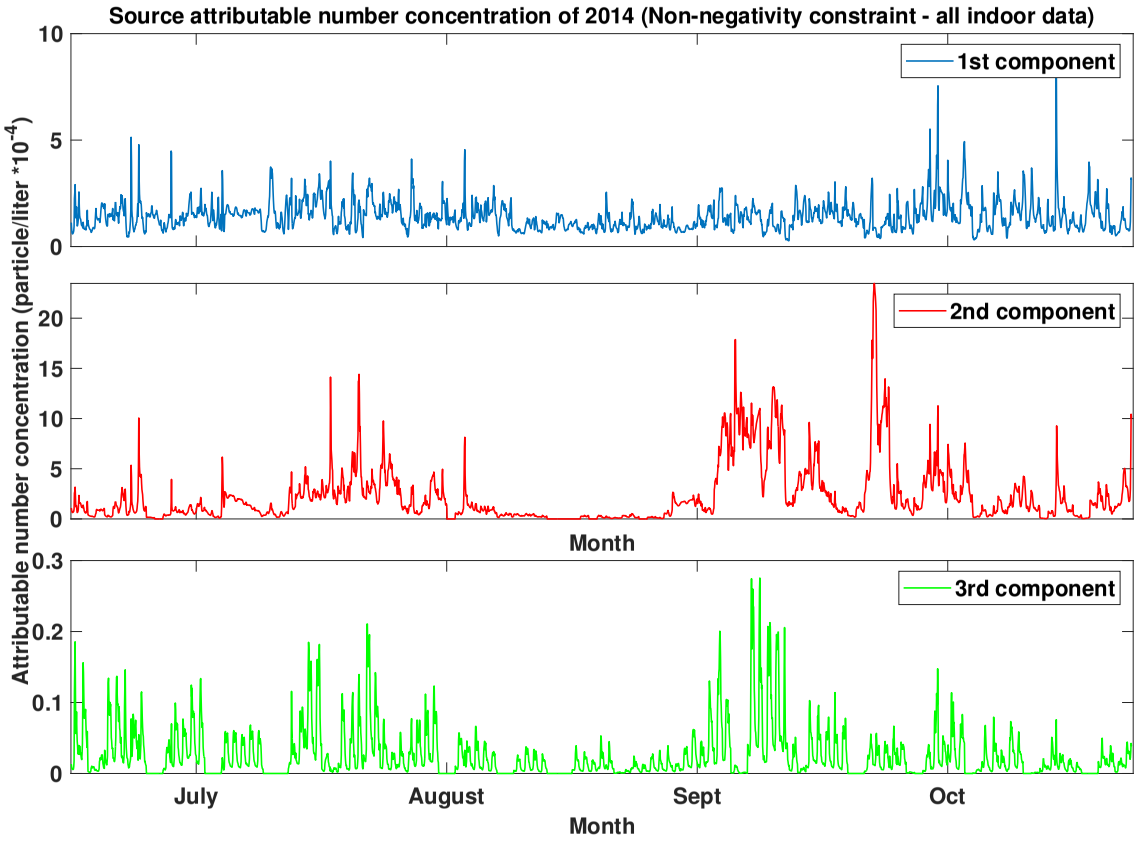


Figure 5.17: Time profile of the attributable hourly concentration to each source in number of particles/liter for all indoor data input. The non-negativity constraint was applied to avoid negative values.

5.2 Both Indoor and Outdoor data

This section presents the results of two other input structures: (*i*) a 4D-array which contains PN concentrations indoors and outdoors and (*ii*) a 3D-array which contains all the measured variables indoors and outdoors.

5.2.1 Case 3: Only particle matter data

As stated in the Introduction part, outdoor data has been introduced as another layer of the input array of PARAFAC. Therefore, in this subsection, the variation of indoor environment data according to different times and locations of measurement (I/O) is analyzed. A 4-dimensional array is constructed: $365_{days} * 15_{PNfractions} * 24_{hours} * 2_{locations}$.

An illutration of the PARAFAC model for this 4-dimensional data is presented in Figure 5.18. The four output matrices contain the loading vectors of 4 modes: (A) day of the year, (B) PN fractions, (C) hour of the day and (D) measurement location.

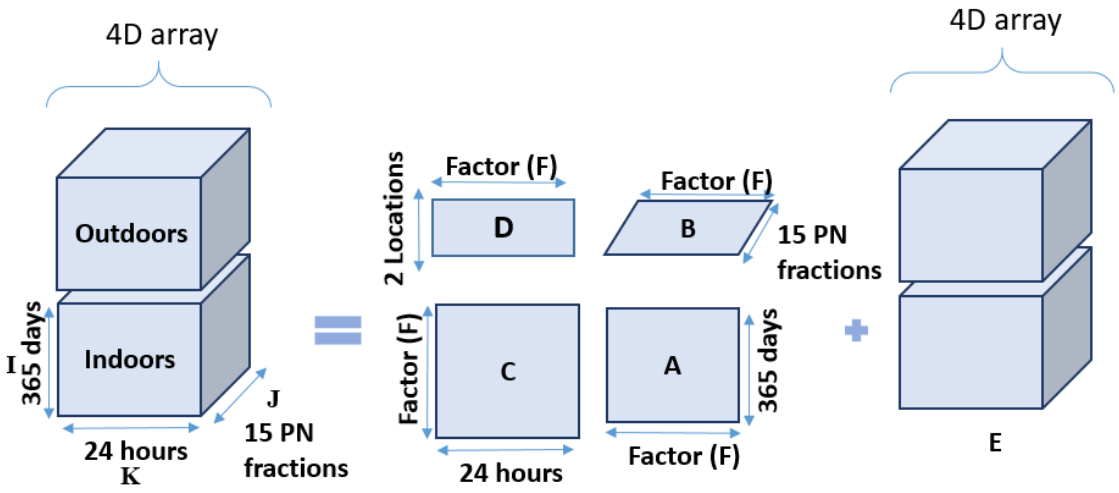


Figure 5.18: The PARAFAC model for 4-dimensional data of PN measurements indoors and outdoors.

In this case, the most suitable number of components is also determined to be three, with the value of core consistency of 70.4% (see Figure 5.19).

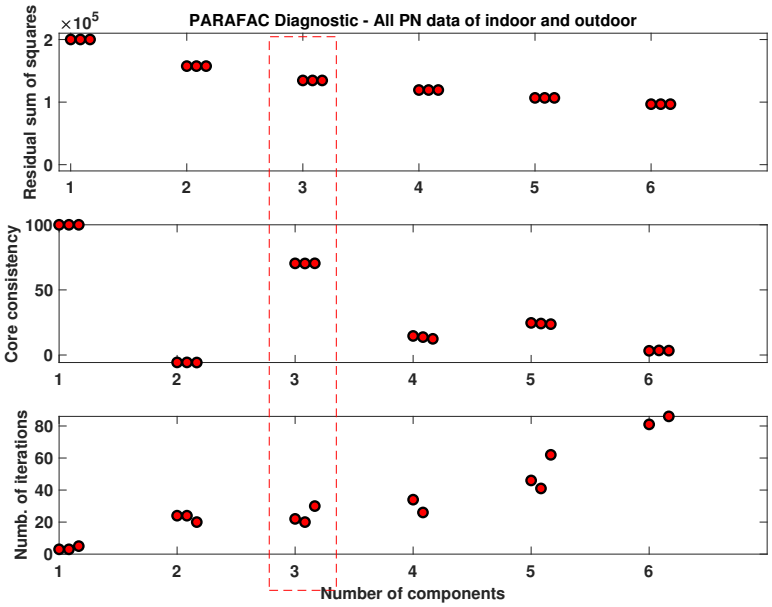


Figure 5.19: The core consistency diagnostic for both indoor and outdoor particulate matter (4D-structure).

The loadings of PARAFAC are displayed in Figure 5.20 and detailed in Figure 5.21. The second component mainly depends on the concentration of small size particles with a higher loading for PN 0.35 - PN 1.3 (fine particles). The specific outdoor pollution episode on March was also detected for this component. In addition, this component is not affected by the location (the same loading for indoors and outdoors), meaning that they have an equivalent influence on the variation of the fine particles. Therefore, this component could be associated with outdoor fine particles that infiltrate indoors with few or no deposition on surfaces, and without the contribution of potential indoor sources of fine particles.

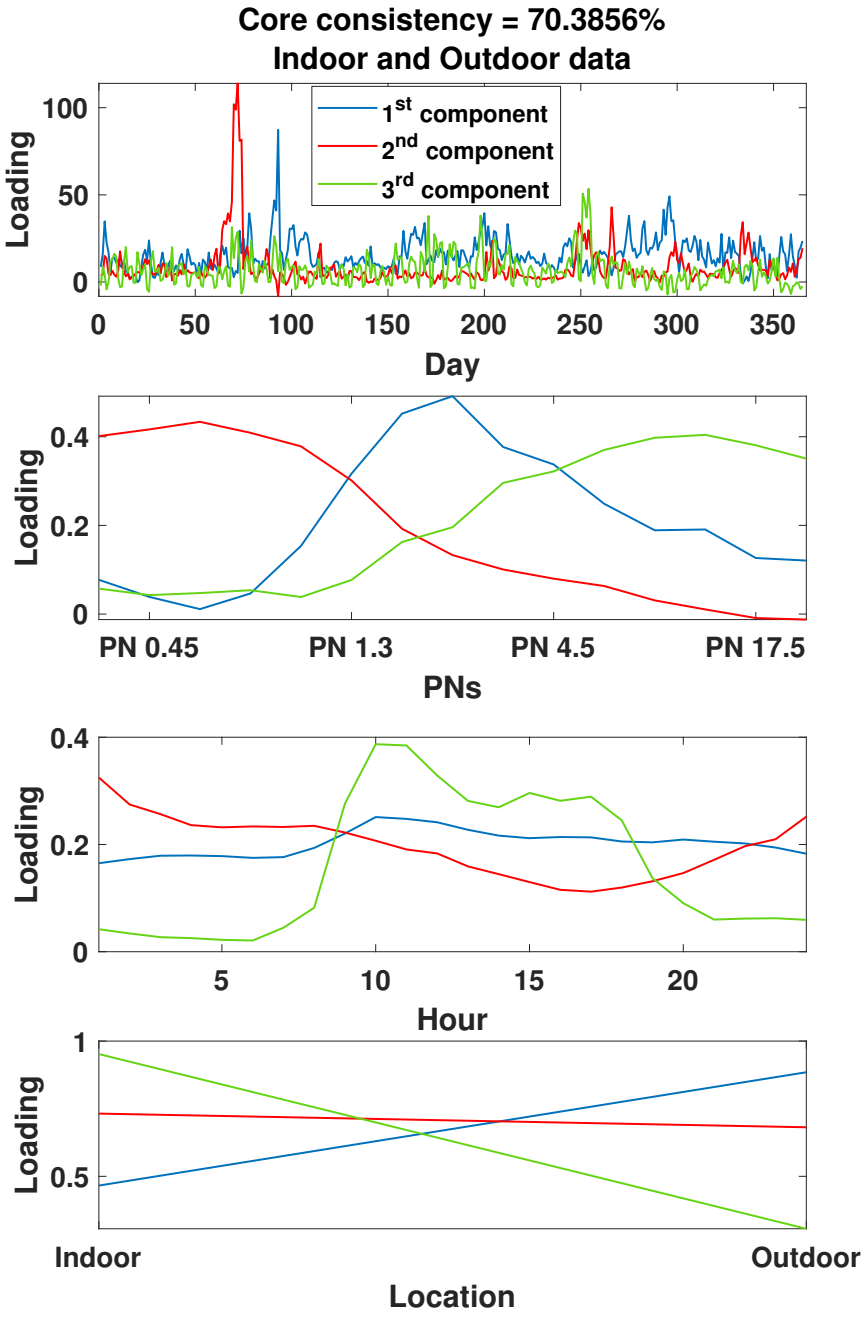


Figure 5.20: The PARAFAC outputs for both indoor and outdoor particulate matter (4D-structure).

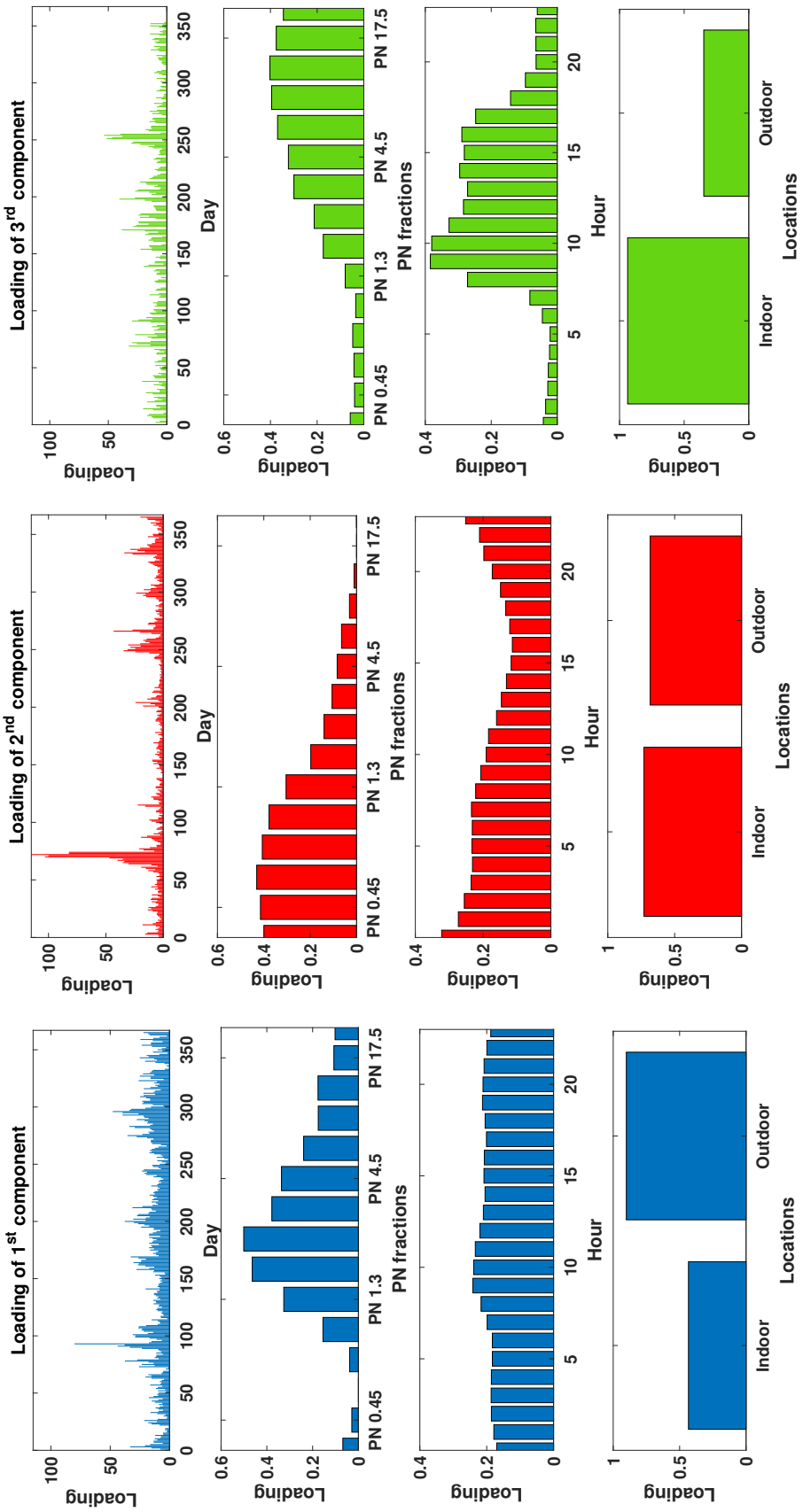


Figure 5.21: Detailed loadings of the three output matrices for both indoor and outdoor particulate matter (4D-structure).

The third component (coarse particles) is strongly correlated to the indoor environment (loading value = 0.92). In addition, a weekly periodicity is detected and a typical daily profile is observed (see Figure 5.22). This is the same as the periodicity of CO_2 indoor concentration, which corresponds to the presence of **occupants indoors**.

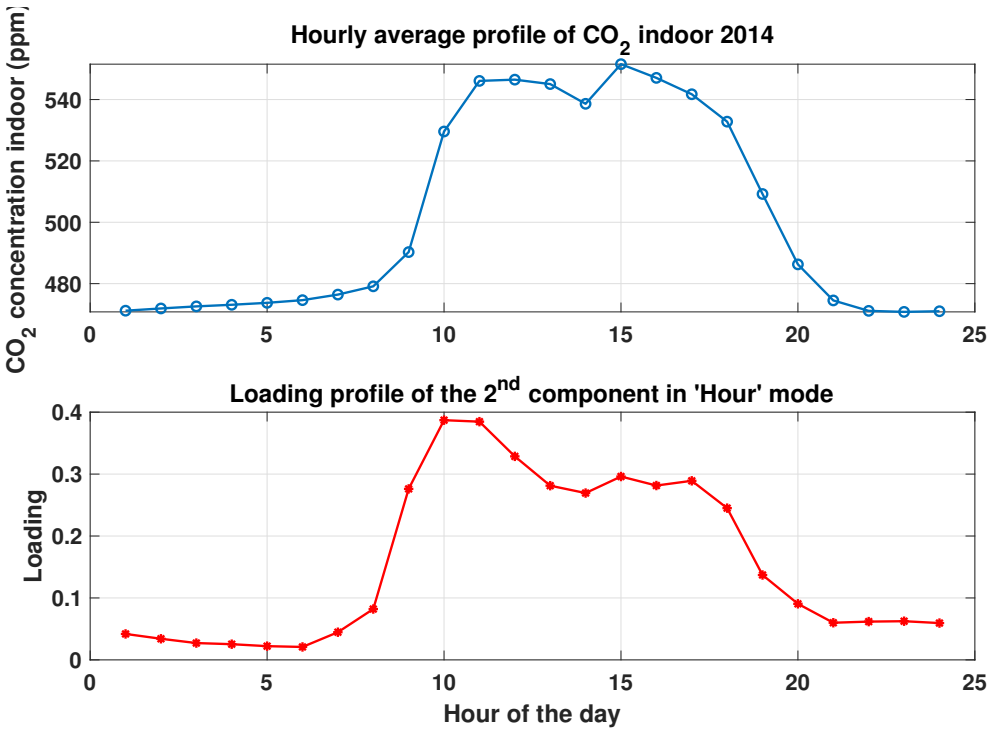


Figure 5.22: Comparison of the daily profile of CO_2 indoor concentration and the third PARAFAC component.

The first component (medium and coarse size particles) is more correlated with the outdoor environment. The coarse particles from outdoors can be caused by the **resuspension** of coarse outdoor dust or by **crustal erosion**, caused by the wind. These higher size particles are also more importantly affected by deposition on surfaces when they infiltrate indoors, hence the outdoor prevalent influence. Based on these results, we can then calculate the source's attributable number concentration, by using the calculation from section 4.2 (regression by multiple with standard deviation values), as in Figure 5.24, respectively. In this case, the non-negativity constraint was imposed for the PARAFAC algorithm in order to obtain positive values. The loading outputs with non-negativity constraint are very similar to the ones obtained in the "no constraint" case (see Figure 5.23). The order of components also remains the same as in Figure 5.20.

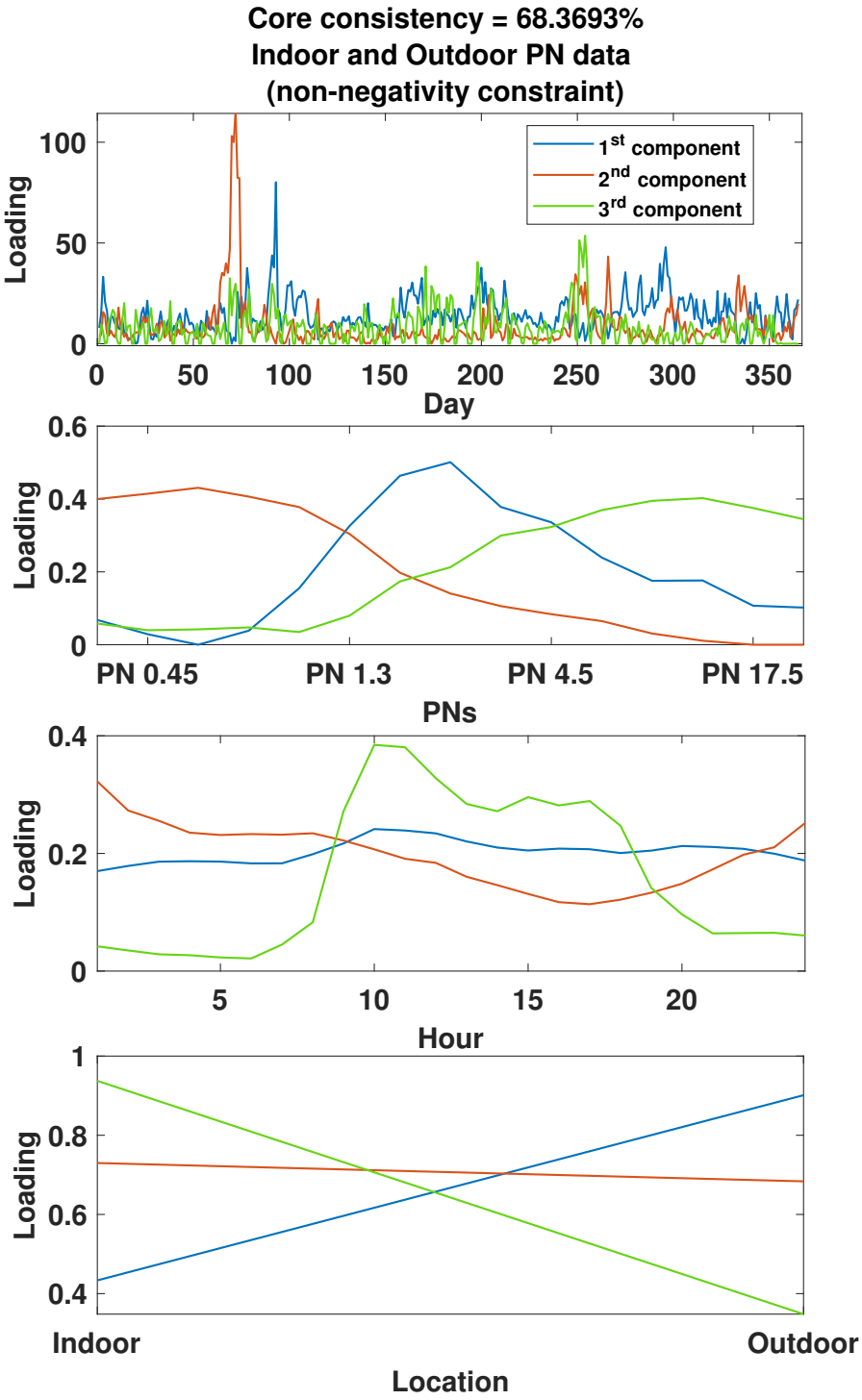


Figure 5.23: The PARAFAC outputs for both indoor and outdoor particulate matter (non-negativity constraint applied).

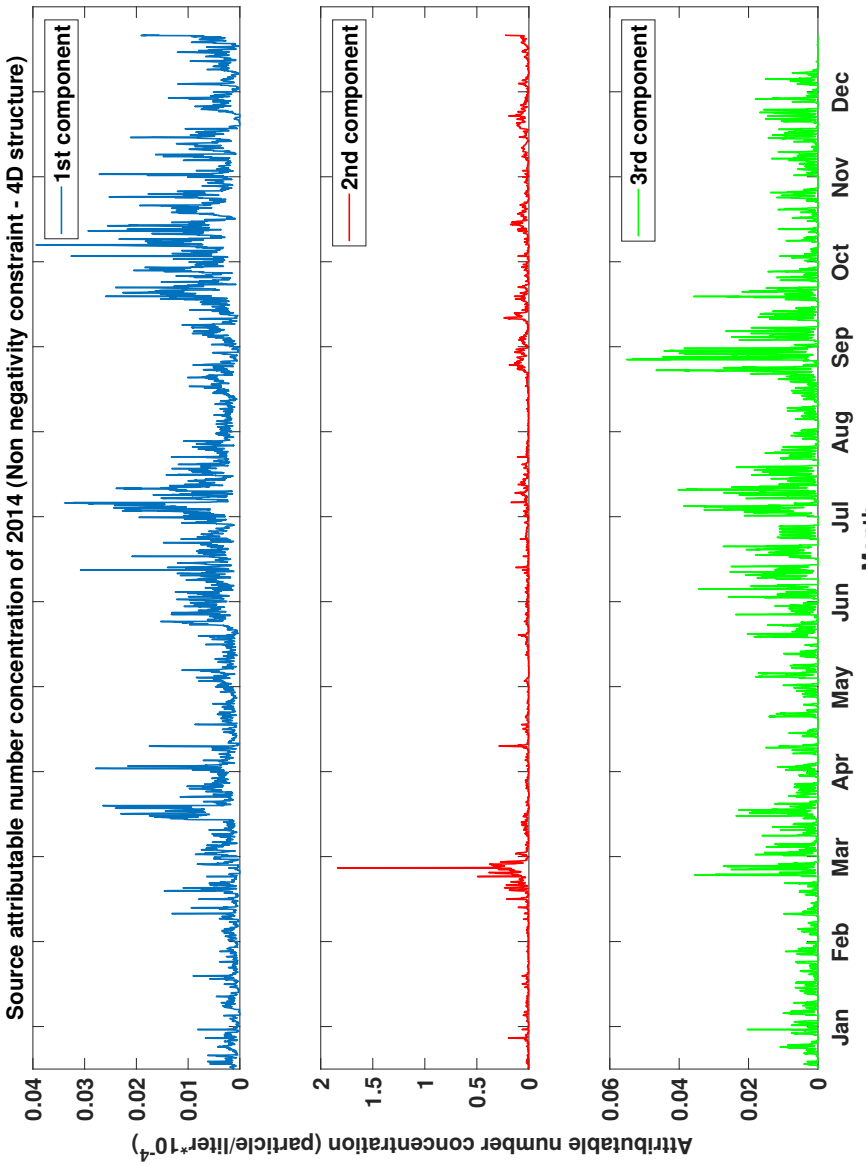


Figure 5.24: Time profile of the attributable hourly concentration of each source in number of particles/liter. Non-negativity constraint was applied to avoid negative results

In addition, to corroborate these assumptions, a PCA was applied to the matrix of daily averaged values of indoor and outdoor PN fractions as active variables (PARAFAC components are used as passive or supplementary variables). The result of the PCA shows the correlation of each component with locations and particle size fractions (Figure 5.25). The first component is clearly correlated with medium and coarse outdoor particles. The second component is associated with fine particles both indoors and outdoors. The third component presents similar variations with coarse indoor particles.

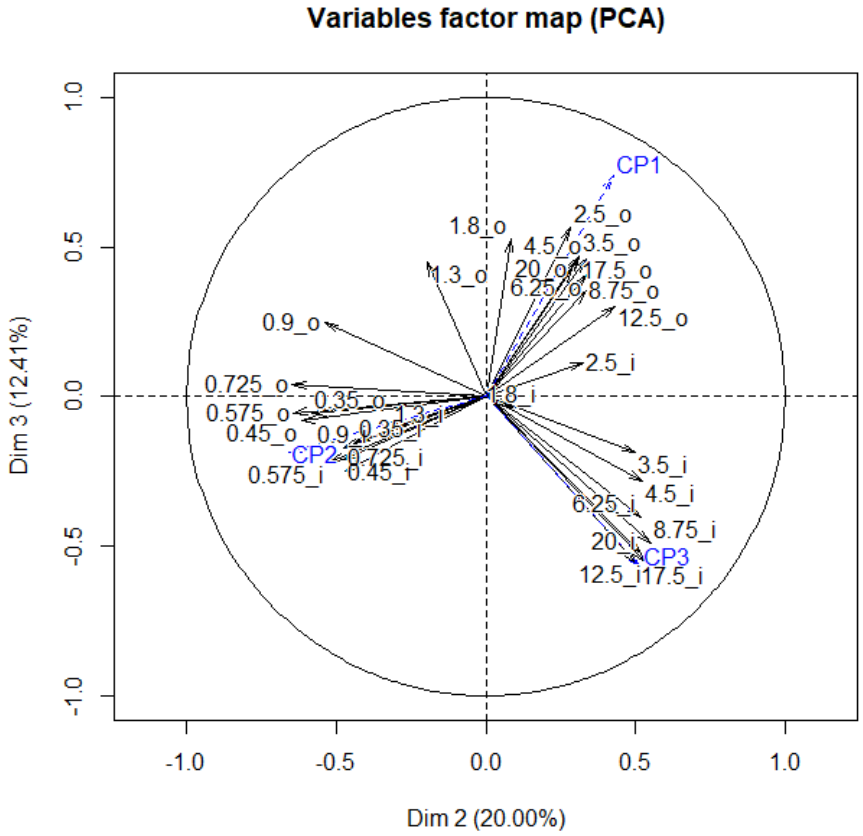


Figure 5.25: PCA for 15 fractions of PN indoors and outdoors (2^{nd} and 3^{rd} component explaining 32% of the variance) and the 3 PARAFAC extracted components CP1, CP2 and CP3 (as passive variables - blue color). The name convention for PN fractions is: fractions size_i for PN indoors and fractions size_o for PN outdoors).

Figure 5.26 presents the attributable PM10 concentration of the three non-negativity extracted components, obtained by the conversion from concentration in number to mass concentration.

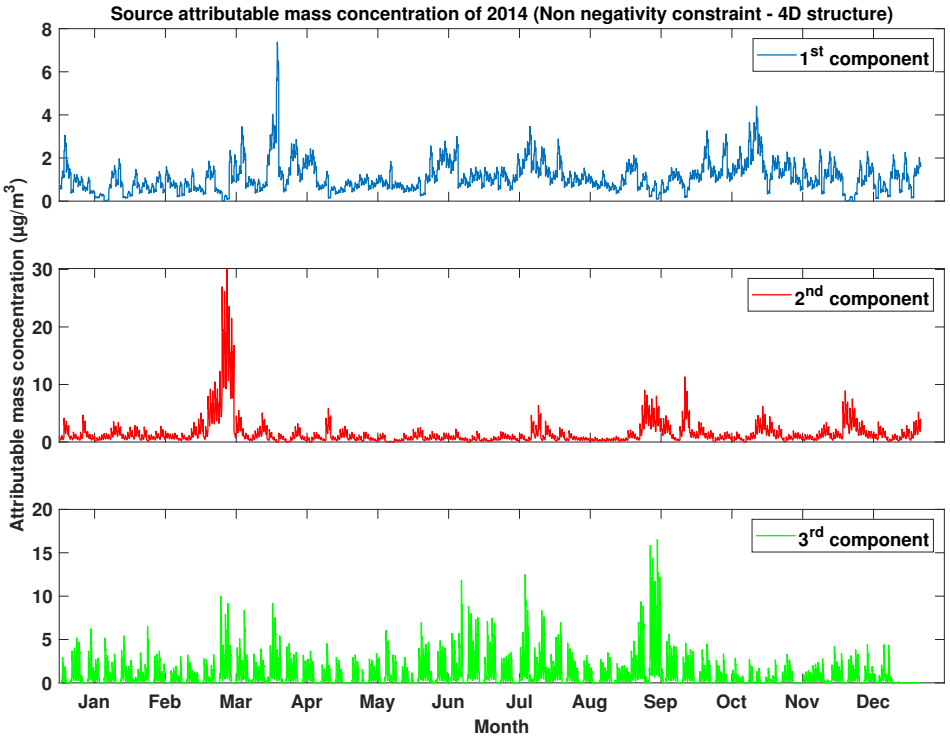


Figure 5.26: Time profile of the attributable hourly mass concentration of each source during 2014 (Non negativity constraint was applied to avoid negative results).

The regression between the 2nd attributable PM10 mas concentration and the original PM10 showed that 37% of the total variance is explained by the 1st component. Similarly, the 3rd component explained 30% of the total variance. Meanwhile, the 1st component accounted for only 2%.

5.2.2 Case 4: All indoor and outdoor data

In this subsection, all of the monitored data both indoors and outdoors was used as input for PARAFAC (period with full measured data - from June to October 2014). Therefore, a 3-dimensional array is constructed: $127_{days} * 50_{variables} * 24_{hours}$.

A PARAFAC model for this 3-dimensional data is illutrated in Figure 5.27. The four output matrices contain the loading vectors of 3 modes: (A) day of the year, (B) variables (15 PN fractions indoors and outdoors; CO, O_3 and HCHO concentrations indoors; CO_2 concentrations indoors and outdoors; 8 wind directions and speeds, Printer Pulse; Irradiance, Temperature and Specific humidity indoors and outdoors), and (C) hour of the day.

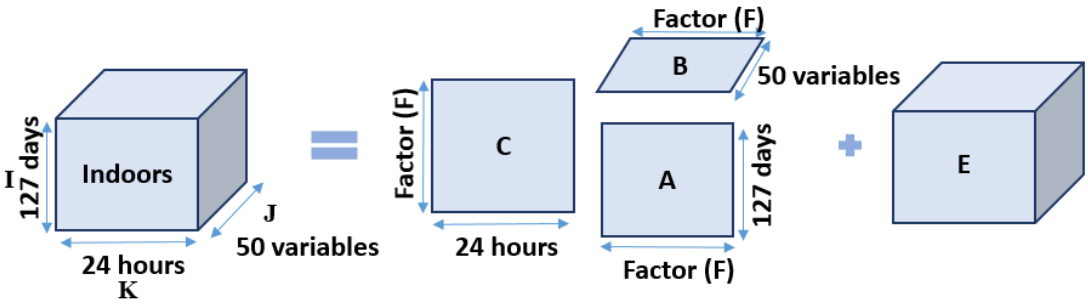


Figure 5.27: The PARAFAC model for 3-dimensional data of all measurements variables (PN frations, other pollutants and climatic parameters) indoors and outdoors.

With all indoor and outdoor data used as inputs, the most suitable number of components was determined to be two, the value of core consistency being 100% (Figure 5.28). The two loading matrices output of PARAFAC are displayed in Figure 5.29 and detailed in Figure 5.30.

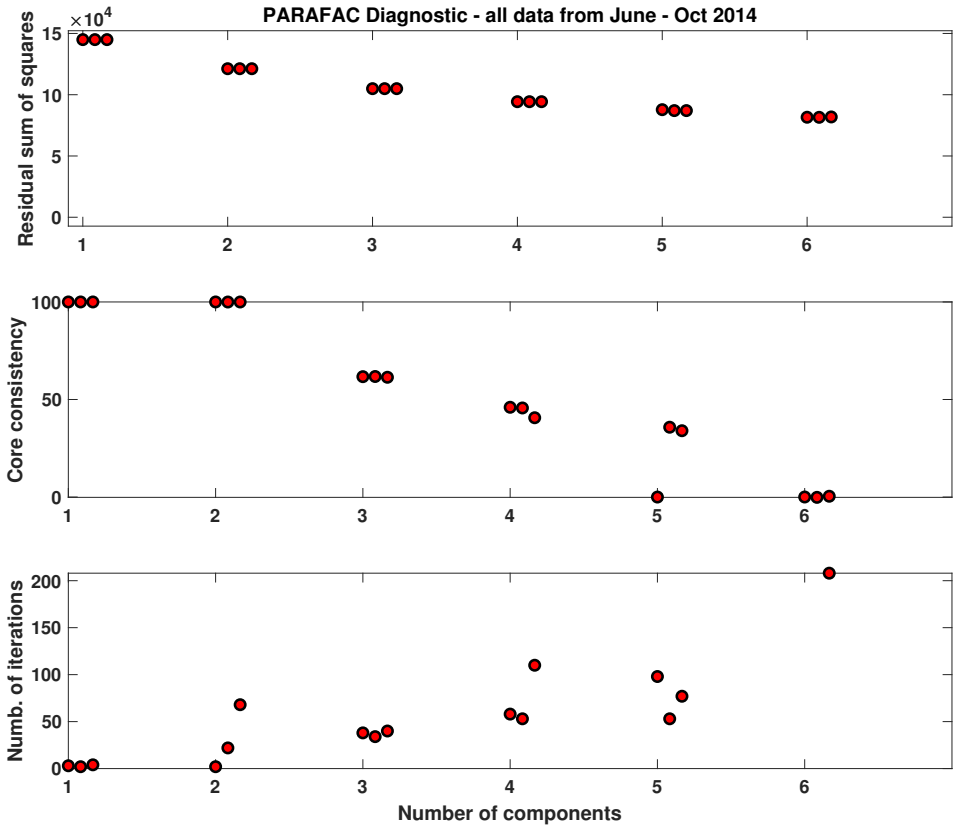


Figure 5.28: The core consistency diagnostic of PARAFAC for the input data including all the recorded data indoors and outdoors from June to October 2014.

It is interesting to note that only two components were extracted for an acceptable core consistency value. With three components, the core consistency drops to 60%.

The second component is related to the "working hours" and so it can correspond to **indoor occupants and related activities** as shown especially by the hourly profile (see Figure 5.29 – Hour mode and Figure 5.30(b)). It also displays high loading values for all the sizes of indoor and outdoor particles.

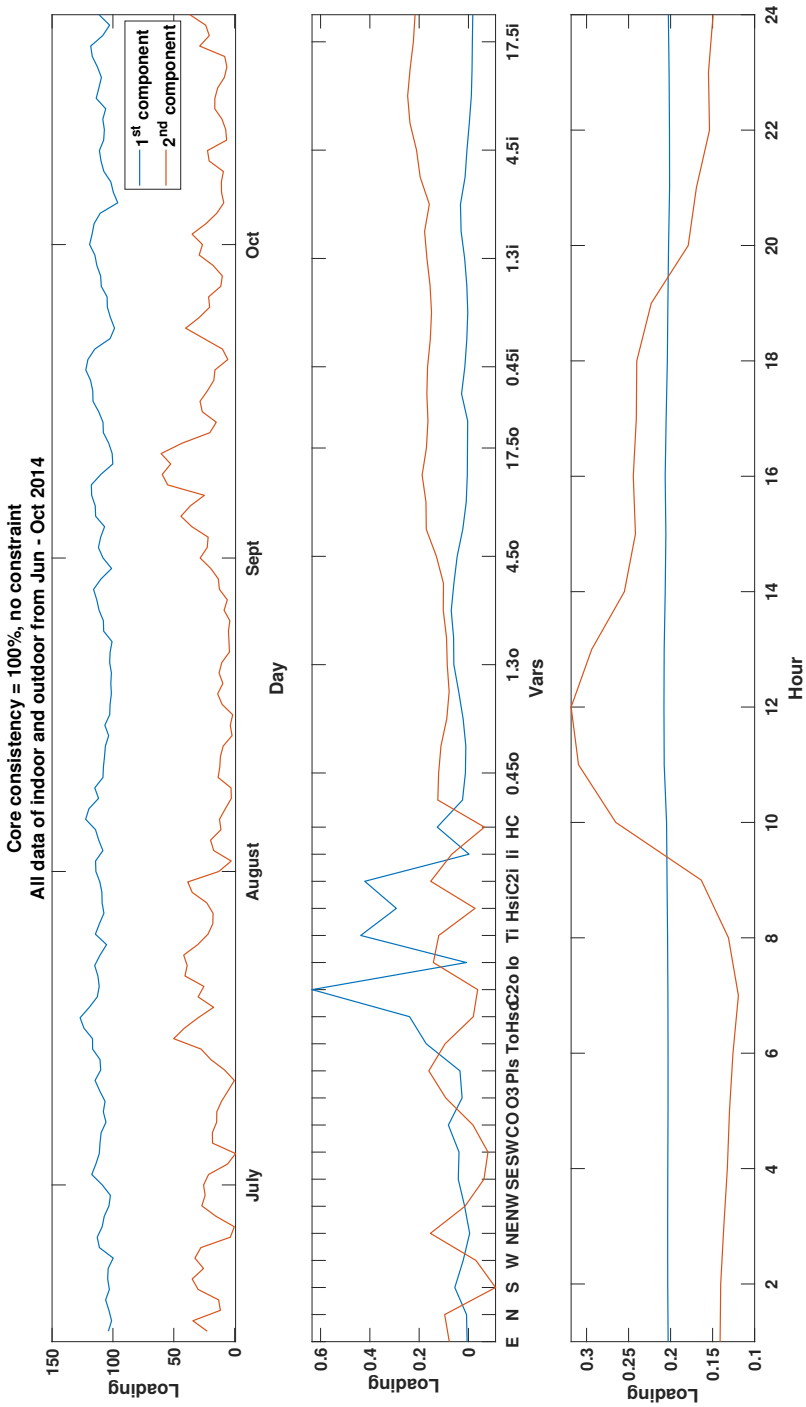
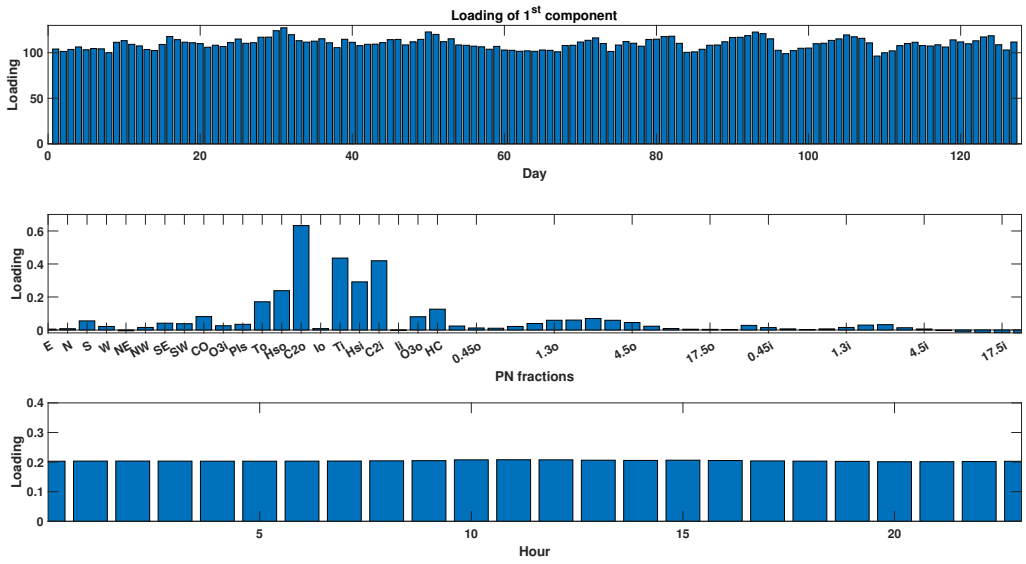
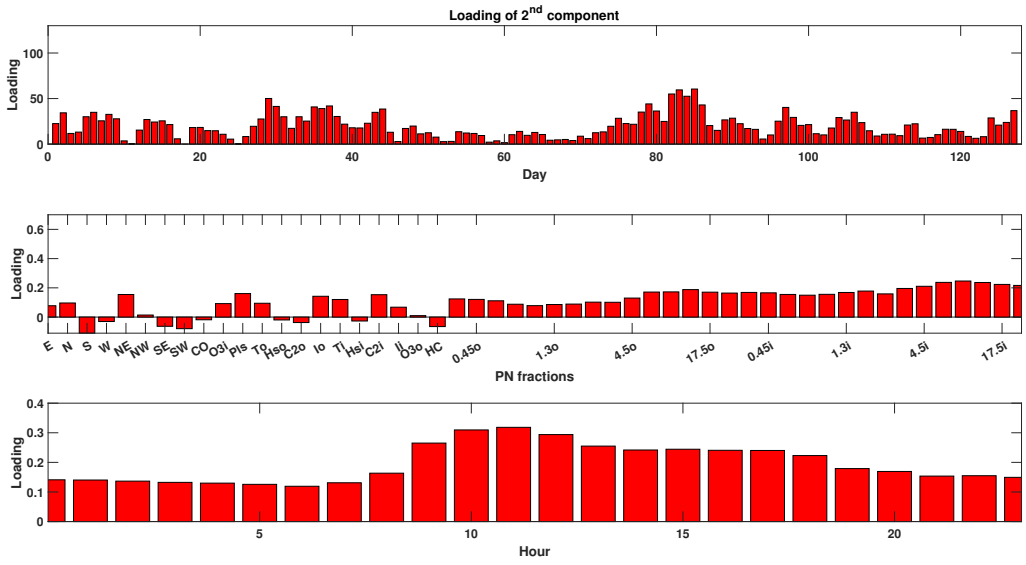


Figure 5.29: The PARAFAC outputs for the input data including all recorded data indoors and outdoors.



(a)



(b)

Figure 5.30: Detailed loadings of the three output matrices for the input data including all the recorded data indoors and outdoors.

Regarding the first component, it looks very similar to the first component retrieved from PARAFAC when applied on all indoor data (Case 2: All indoor data). It displays high loading values for temperature and humidity both indoors and outdoors, and for CO_2 and HCHO concentration. In addition, one can notice the very low values of loading for all the PN fractions (Figure 5.30a middle) and the loading value is constant for both the daily profile (0.2, see Figure 5.30a bottom) and for the whole period (100, see Figure 5.30a top). This component is not related to particulate matter and represents the other ambient data that vary much less than particles during the studied period.

Regarding the attributable concentration to each source in number of particles/liter for this case, Figure 5.31 represents the time profile of the two extracted components. Similar to the other results (Case 1-3), the low concentration in August is also detected by the 2nd component, the one related to indoor occupants and related activities.

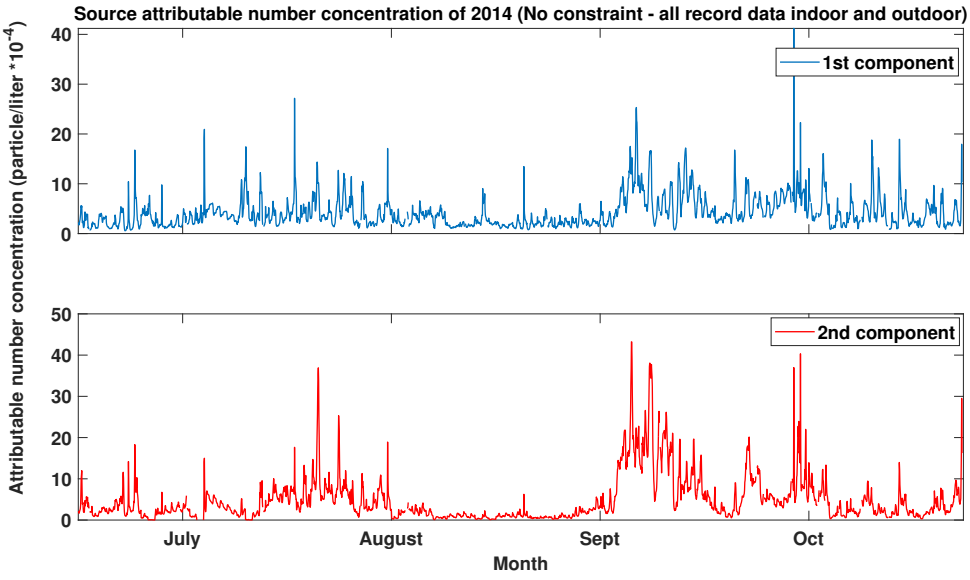


Figure 5.31: Time profile of the attributable hourly concentration for each source in number of particles/liter for all recorded data indoor and outdoor input.

5.3 Conclusion and Discussion

In this research, we have successfully applied PARAFAC as a tensor decomposition algorithm to decompose the n-dimensional data, which contains the number concentration of airborne particles of different sizes with or without other measured data (climatic and other pollutants). From the loadings, we are able to analyze the variation in the data, identify the main sources of pollutants and assess their relative contributions. The particularity of this method (PARAFAC) is that indoor and outdoor particles of given sizes are considered in parallel layers and not as different variables of the same layer (matrix-based methods such as PMF). This structure allowed to determine the relative contribution of outdoor sources to the indoor concentration of particles, which is a topic of utmost interest in Indoor Air Quality studies nowadays.

When applied to PN concentrations, three factors were retrieved from PARAFAC and identified: outdoor sources, indoor sources caused by occupants' presence and activities, and an unexplained factor that may include other random events. The method allowed to determine the relative contributions of the sources and the attributable concentration at a given time. A system based on PARAFAC that provides information about pollutant sources at a given moment, could be created in the future.

The addition of other data did not improve the separation and identification of particle sources. The other data were identified as a single factor with no relation to particulate matter. The addition of outdoor PN data allowed to retrieve three explainable factors with the addition of a new dimension. In particular, it allowed to dissociate fine particle from outdoors that behave the same indoors and medium/coarse particles from outdoors that vary differently compared to indoors. The added dimension also increases the complexity of interpretation. The use of PCA afterwards helped to better understand the retrieved factors.

Part II

Forecasting of the window opening state

The opening state of the windows has an important influence on the IAQ, as it can modify the air exchange rate and as such the transfer between indoor and outdoor environments. Opening a window may lead to a sudden increase in the air exchange rate and to both (*i*) a quick decrease of the concentration of indoors generated pollutant like CO_2 and (*ii*) a possible increase of the indoor concentration of pollutants coming from outdoors as PM. The thermal comfort and indoor air quality can be improved by window opening/closing. It is therefore necessary to understand and model the influence of this factor on IAQ.

In this second part of the thesis, we tried to model the windows opening state in a real open-plan office with five windows. From the various approaches three machine learning models: Decision Tree, kNN and Kernel Approximation, were selected to be tested in our study case.

The general outline of this part is organized as follows:

- Chapter 6 - literature review concerning the models employed to predict the windows opening state.
- Chapter 7 - presentation (description) of the three selected ML models.
- Chapter 8 - model implementation process (parameters selection and hyperparameter setting); the results concenrning the performance of different ML prediction models and discussion.

Chapter 6

Modeling of the windows opening state in the literature

The opening state of the windows has an important influence on IAQ, as it can modify the air exchange rate and as such the transfer between indoor and outdoor environments (Godish and Spengler, 1996). Opening a window may lead to a sudden increase in the air exchange rate and to both (*i*) a quick decrease of the concentration of indoors generated pollutant like CO_2 and (*ii*) a possible increase of the indoor concentration of pollutants coming from outdoors as PM. A research in a mock-up building revealed that the thermal comfort and indoor air quality can be improved by window opening/closing (Park, 2013). It is therefore necessary to understand and model the influence of this factor on IAQ.

Window-opening activity is affected by a variety of parameters, such as outdoor temperature, air quality, human presence and season (Park and Choi, 2018; Park *et al.*, 2020; Raja *et al.*, 2001). Occupant's behavior is an important factor but it can vary among individuals (Park and Choi, 2018), leading to different impacts on the indoor environment (Park *et al.*, 2020).

On the one hand, theoretical physics-based models (models based on physics rules) struggle to explain the changes in window-opening behavior (Dai *et al.*, 2020), in the perspective of direct modeling. On the other hand, machine learning models develop computational algorithms designed to simulate human intelligence by learning from their surroundings (El Naqa and Murphy, 2015), in the perspective of inverse modeling. Considering the complexity of the underlying relationships, a machine learning model could be a good alternative to a physics-based model and a powerful tool for predicting or forecasting window-opening behavior.

In the last decades, Machine Learning (ML) models have been effectively used in the prediction of indoor air quality (Chen *et al.*, 2018; Martínez-Comesaña *et al.*, 2022; Wei *et al.*, 2019) and energy consumption (Amasyali and El-Gohary, 2018; Edwards *et al.*, 2012), proving the potential of using machine learning models in indoor environments. Regarding windows opening modeling, a recent study (Tien *et al.*, 2021) has used the Deep Learning technique for Neural Networks (a specific type of ML) for the detection and recognition of the opening state of the windows by using a camera in order to propose frameworks for energy saving.

According to the review paper of Dai and his colleagues (2020), the common ML models for predicting window-opening behavior include: logistic regression, artificial neural networks (ANN), the Markov chain model, and support vector machines (SVM). Figure 6.1 presents different types of ML algorithms (Atul, 2022) and some of them will be presented more in detail in chapter 7.

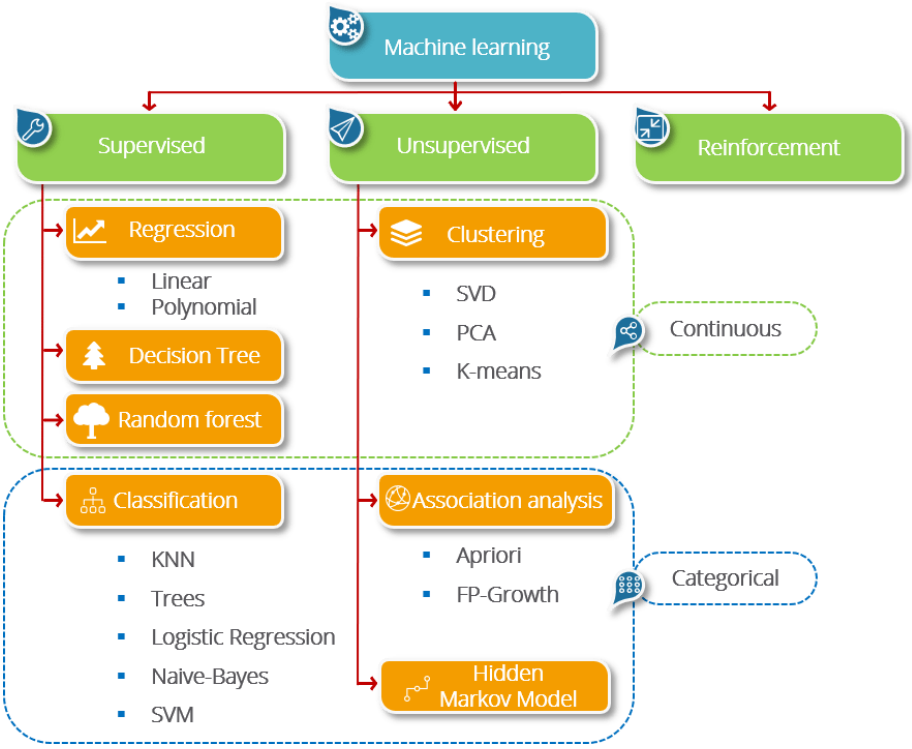


Figure 6.1: Different types of Machine Learning algorithms (Atul, 2022).

It is important to notice that we decided to present all these models within the frame of ML. They can be presented as well as statistical models. This aspect is not of utmost importance, because we refer to the same model. Their most important common feature is that they are data-driven models and that an inverse modeling is performed each time. Figure 6.1 gives some examples: the well-known statistical technique of regression can be equivalently used as a supervised ML model. Multivariate statistical analysis performed by using PCA or K-means clustering can be considered as unsupervised ML techniques.

A stochastic window status profile generator (WinProGen) has been introduced by Cali and colleagues (2018). For the development of window state profiles, three models have been established in WinProGen; they depend on the hour of the day, the day of the week, and/or on the daily average ambient temperature. This model uses a database with transition probability matrices obtained from 300 windows in 60 apartments in southern Germany, monitored during 2012 with 1-minute time step.

Reliable predictions of buildings' energy performance are obtained when applying these generated window state profiles to the dynamic simulation of two demonstrator buildings. The implemented stochastic models are Markov chains. The

Markov chain describes a sequence of possible events in which the next state X_{k+1} is conditionally dependent on the past $X_0, \ldots, X_{k-1}$) given the present state X_k (Serfozo, 2009). With this assumption, we have that:

$$p_k^{ij} = p(X_{k+1} = s_j | X_k = s_i) \tag{6.1}$$

$$p(X_{k+1} = s_j | X_1, X_2, ...X_k) = p(X_{k+1} = s_j | X_k) \tag{6.2}$$

where p_k^{ij} represents the transition probability and it denotes the probability of the system to change from state s_i (open/close) to state s_j (close/open) at time step k.

This model has the advantage of appropriately accounting for the process's time dependency. However, according to the authors, this model struggled to deal with a large number of input variables in comparison with the logistic regression method. Therefore, they proposed, as future work, to develop a hybrid model, combining both the Markov chain technique and the logistic regression analysis (Cali *et al.*, 2018).

Most of the research used logistic regression to compute the correlation between the probability of a window opening and the variables of influence (Andersen *et al.*, 2013; Yao and Zhao, 2017). Logistic regression (Hosmer and Lemeshow, 2000) is a statistical approach that determines the likelihood of a given event (e.g., opening a window) occurrence based on relevant factor elements (e.g., outdoor/indoor air temperature or PM2.5 concentrations). The Wald statistic test, which has a χ-squared distribution, is an useful approach to identify the contribution of various components to the event occurrence when using logistic regression. Thus, a significant 2-tailed P-value for a certain predictor indicates if this predictor is essential in the logistic regression model (Pan *et al.*, 2018), as given in equation (6.3):

$$P = \frac{e^{(a+bx)}}{[1 + e^{(a+bx)})]} \tag{6.3}$$

where P is the probability of the window-opening, x is an influential factor, a and b are constants, which represent the regression coefficients. These constants are estimated by regression analysis using a maximum likelihood estimation.

Andersen and Yao used logistic regression to compute the correlation between window opening and the variables of influence in order to predict the probability of a window opening/closing event. In their researches, 19 dwellings in Beijing (Yao and Zhao, 2017) and 15 residences in Denmark (Andersen *et al.*, 2013) are studied. Predictive models of the occupants' window opening behavior were established based on multivariate linear logistic regression. Their results indicated that outdoor air temperature was the most influential variables in determining the window opening and closing probability, followed by indoor CO_2 concentration, indoor air temperature, outdoor and indoor relative humidity, ambient PM2.5 concentrations, and outdoor wind direction and wind speed.

This method has the advantage of providing interpretative parameters and could be regularized to minimize over-fitting. However, the model struggles to address the complicated relationships, due to its low flexibility (Dreiseitl and Ohno-Machado, 2002).

Other researchers attempted to apply a data-mining approach to discover the effects of the window opening and closing behavior on energy consumption in buildings (D'Oca and Hong, 2014). This paper proposes a framework for identifying valid window operating patterns, in measured data, by combining logistic regression analysis with two data-mining approaches: (i) cluster analysis and (ii) association rules mining. Analyses were performed on the data set obtained by monitoring 16 offices in Frankfurt am Main (Germany). The dataset contains indoor and outdoor physical factors as well as human interaction with operable windows, which was measured in 10-min interval data over two complete years (2006 and 2007). In their study, 8 non-numerical and 7 numerical variables are used for calculating the probability of opening and closing for a window. In total, a huge quantity of detailed data was used. According to the four aims of the research, (i) three motivational (thermal-driven, thermal/time-driven, time-driven), (ii) three opening duration (long, medium, short), (iii) three interactivity (active, neutral, passive) and (iv) three degree of opening position (small, intermediate, big) behavioral patterns were achieved. The authors succeeded to obtain distinct behavioral patterns to serve as a basis for 12 association rules, which classified two typical window opening office user profiles: (i) physical environmental driven and (ii) contextual driven. Based on that, appropriate recommendations for different natural ventilation strategies as well as robust building design could be achieved.

A similar study (Markovic *et al.*, 2018) suggested a generic model that identifies window states using a fully connected feed-forward neural network. The network consists of 25 neurons in the input layer, corresponding to 22 variables from the current time step and 3 variables from 10 minutes before the current time step, as input features. An optimal performance was achieved by a five hidden layers neural network. For both training and testing processes, this model used around 20 million data samples. Data from Aachen University's offices was used for the training step. The data were logged in a minute-wise frequency from January 1st, 2014 to October 1st, 2015, including detailed indoor climate, air quality and occupant behavior information from 52 single or double occupied offices. After that, the proposed model was evaluated on other additional data sets, which were collected from offices in Frankfurt (Germany) and Philadelphia (USA). The additional data set was divided into adaptation set and evaluation set. During the adaptation process, the pre-trained weights were adapted by running several tuning iterations, while no hyperparameter tuning or further calibration was required. Based on this procedure, the only required step is the weight adaptation when applied to the other buildings, other while, this model did not require any parameter search or calibration. The resulting evaluation accuracy and F1 scores on the office buildings ranged between 86 and 89% and 0.53–0.65 respectively. The resulted model could be used by the engineers and designers as a standalone, or as a part of a thermal building simulation.

Six machine learning algorithms were trained in the research of Park *et al.* (2020). The authors have used monitoring data of 23 sample homes located in Seoul and suburban areas for predicting the occupant's behaviour in the manual control of windows. According to the analysed predictive performance, the k-NN model

shows the best fitness with the monitored data set. Regarding the input parameters, the Gini importance score indicated that there are five main driving parameters: (i) prevailing mean outdoor air temperature (PMA), (ii) mean daily temperature, (ii) CO_2 indoor concentration, (iv) relative humidity indoors and (v) the difference between outdoor temperature and the operative temperature indoors.

The Kernel Approximation method has been mainly applied in speech enhancement methods (Zhao *et al.*, 2016). Regarding the Decision Tree, this method has been used to classify the most important parameters among a large range of variables such as: sociodemographic data, health and lifestyle habits, ergonomic and psychological factors for the Sick Building Syndrome (SBS) (Sarkhosh *et al.*, 2021).

For our research situation, many supervised ML methods such as Decision Trees, Support Vector Machines, k-Nearest Neighbor, and Ensemble classification can be used. We decided to study the ability of different ML classifiers including: decision trees, k-NN classification and kernel approximation (SVM kernel), to predict the state of the window opening in an open-plan office, as presented hereafter. The reason for selecting the Decision Trees is that this method offers the possibility to get the extracted rules and apply them for other study cases. Regarding k-NN, this method is recommended as 'a theoretically optimal method of classification' (Hastie *et al.*, 2001). Finally, we chose Kernel Approximation as it can take into account the non-linearity relationship among the variables. The detailed information about these three methods is presented in the chapter 7.

Chapter 7

Description of the models used for predicting the windows opening state

Before introducing the different models that we have used in our study, we want to provide some terminologies frequently used in ML classification method as below:

- **Model-based algorithm**: an alternative methodology for applying machine learning, which seeks to create a bespoke solution tailored to each new problem. The solution is expressed through a compact modeling language, and the corresponding custom machine learning code is then generated automatically (Bishop, 2012).
- **Instance-based algorithm** (sometimes called memory-based learning): an opposite of model-based algorithm, this methodology generates classification predictions using only specific instances (Aha *et al.*, 1991), or in other words, "it constructs hypotheses directly from the training instances themselves" (Norvig and Peter, 1995).
- **Supervised**: in supervised learning, labeled data are used. They represent a data set that has been categorized, to infer a learning algorithm. The data set is used as the basis for predicting the classification of other unlabeled data using machine learning algorithms (Mark *et al.*, 2015).
- **Unsupervised**: unsupervised learning algorithms are used to group cases based on similar attributes, or naturally occurring trends, patterns, or relationships in the data. Unsupervised models include clustering techniques and self-organizing maps (Colleen, 2015).
- **Reinforcement**: a machine learning training method that rewards desired behaviors and/or punishing undesired ones. In general, it is capable of perceiving and interpreting its environment, taking actions and learning through trial and error (Kaelbling *et al.*, 1996).
- **NP-hard**: any solving algorithm can be translated into an algorithm for solving an NP-problem in order to appreciate its computing time (Nondeterministic Polynomial time problem). NP-hard therefore means "at least as hard as any NP-problem" in terms of computing time.

- **Feature**: an individual measurable property or characteristic of a phenomenon (Bishop, 2006). Choosing informative, discriminating and independent features is a crucial element of effective algorithms in pattern recognition, classification and regression.
- **Samples**: a smaller, manageable representation of a larger group; it is a subset of a bigger population that contains the features of that larger group (Keton, 2021).

  An example about features and samples in a dataset is displayed in Figure 7.1 where the lines represent the samples and columns represent the features.

Features **Label**

	PM 2.5 ($\mu g/m^3$)	PM 10 ($\mu g/m^3$)	VOC (ppm)	CO_2 (ppm)	HCHO (ppm)	IAQ
1	3	30	10	410	0.05	Good
2	13	66	19	900	0.3	Moderate
3	58	300	60	2100	0.7	Unhealthy
4	11	22	5	500	0.1	Good
5	90	257	55	2200	0.75	Unhealthy
6	20	120	22	670	0.35	Moderate

Samples

Figure 7.1: An example database about Indoor Air Quality classification.

- **Overfitting**: "the production of an analysis that matches too closely or perfectly to a specific collection of data, and may thus fail to fit to new data or predict future observations accurately" (Oxford Dictionaries, 1930). An overfitted model is a statistical model that has more parameters than can be justified by the data (Everitt and Skrondal, 2010). In this case, we can say that the algorithms "learns by heart" the samples and it is not able to generalize when applied to unseen ones.
- **Underfitting**: a data science scenario in which a data model is unable to accurately capture the connection between the input and output variables, resulting in a high error rate on both the training set and unseen data (Education, 2021).

  Figure 7.2 illustrates an example of over-fitting and under-fitting.
- **Decision boundary**: a hypersurface that separates the data points into specific classes, where the algorithm switches from one class to another (Sahu, 2021).
- **Bias**: a systematic error in science and engineering. Statistical bias is caused by an unfair sample of a population or by an estimating procedure, that does not produce accurate findings on average (Welsh and Begg, 2016).

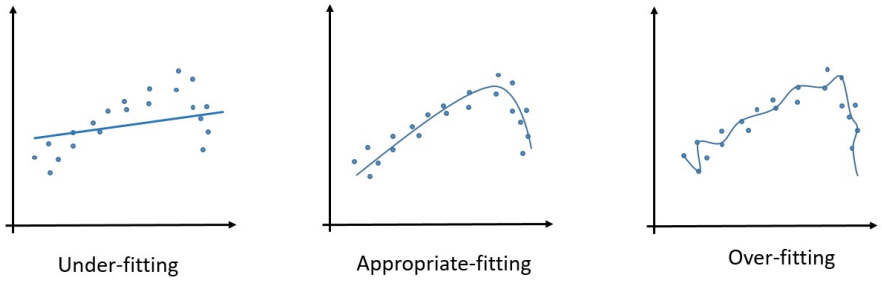


Figure 7.2: An example of over-fitting and under-fitting.

- **Training set**: a data set of samples is used during the learning process and is used to fit the parameters of a classifier (Ripley, 1996). For classification tasks, a supervised learning algorithm examines the training data set to determine, or learn, the appropriate variable combinations that will provide a strong prediction model.
- **Testing set**: a data set that is independent of the training data set but has the same probability distribution (Ripley, 1996).
- **Validation set**: a collection of samples used to fine-tune a classifier's hyper-parameters.
- **Cross-validation**: a resampling approach that tests and trains a model on different iterations using different subsets of the data (Stone, 1974). The purpose of cross-validation is to evaluate the model's ability to predict new data, which was not used in the estimation process. Based on that, it is possible to identify errors such as overfitting or bias selection. In addition, this method provides insight into how the model would generalize to an independent dataset (Cawley and Talbot, 2010).

7.1 k- Nearest Neighbor Classification

k-Nearest Neighbors models (Fix and Hodges, 1951) are a type of instance-based model that is used mainly for classification in the Machine Learning field. Its fundamental is as follows: similar objects exist in close proximity.

While model-based algorithms (the opposite of instance-based models) use the training data to create a model with input parameters, the instance-based models, such as k-Nearest Neighbors, use the entire training data set to determine the model, without learning any parameters to assign a class or category to a specific new data point.

Because of their non-parametric nature, the k-NN models can quickly assess the viability of a multi-class classification problem. k-Nearest Neighbors is one of the most simple and easy-to-use model to classify data.

One of the main advantages of the k-NN of models is that they are able to quickly adapt to new samples since they do not need to recalculate any weights. The

downside is that all the dataset even very large is kept in memory and not a reduced or compact set of weights, which might be **computationally very costly**.

7.1.1 The k-NN algorithm

The basic steps of the k-NN algorithm for classification are described below:

1. Load the data
2. Initialize k to the chosen number of neighbors. In the k-NN model, k is defined as the number of nearest neighbors. This parameter is the core-deciding factor. When k=1, then the algorithm is known as the nearest neighbor algorithm. The more detailed information about how to select the suitable k-value is presented in the subsection 7.1.2.
3. For each data sample:
   (a) calculate the distance between the query sample and the current sample from the dataset by using distance measures such as Euclidean, Chebyshev, City Block, Cosine, etc. The main intuitions of some distance metrics are displayed in the Figure 7.3.

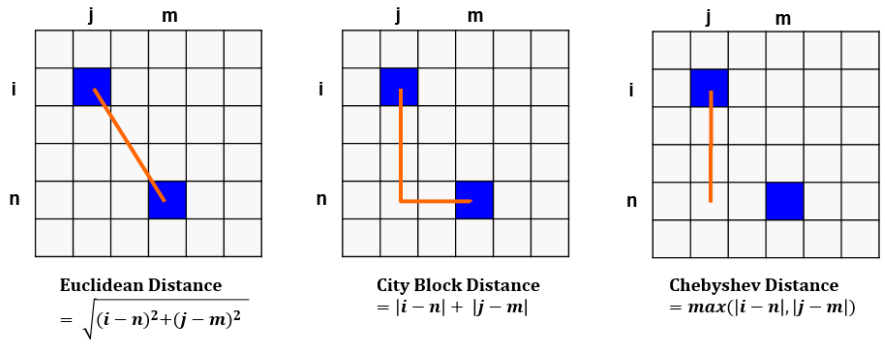


Figure 7.3: Different distance measures used in k-NN classification.

   (b) Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by distances
5. Pick the first k entries from the sorted collection
6. Get the labels of the selected k entries
7. Return the mode (the value that appears the most often) of the k labels.

The obtained label is assigned to the query sample in the classification task.

Examples of k-nearest neighbor classifiers are displayed in Figure 7.4 and Figure 7.5. In the case of the 'nearest neighbor' classification, the classifier searches for just one nearest neighbor and the query sample is assigned to Group 2 (star shape). Meanwhile, using the same set of initial data, in the case of the 3-nearest

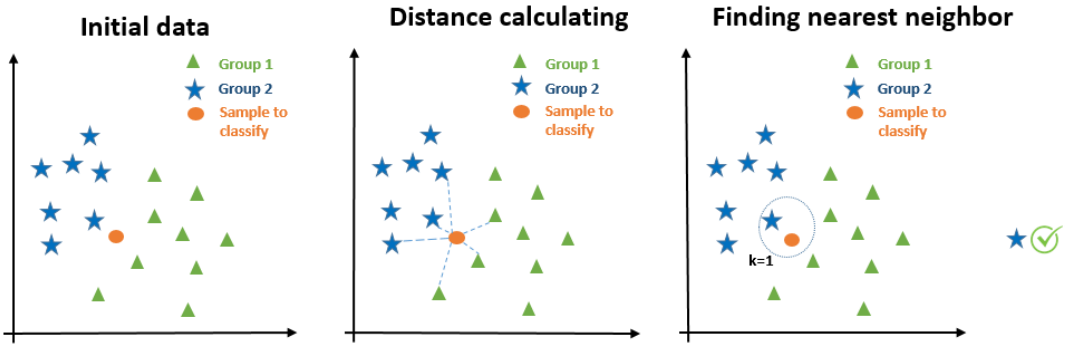


Figure 7.4: An example of the nearest neighbor classification (k=1). The sample is finally classified as belonging to group 2 ('star shape').

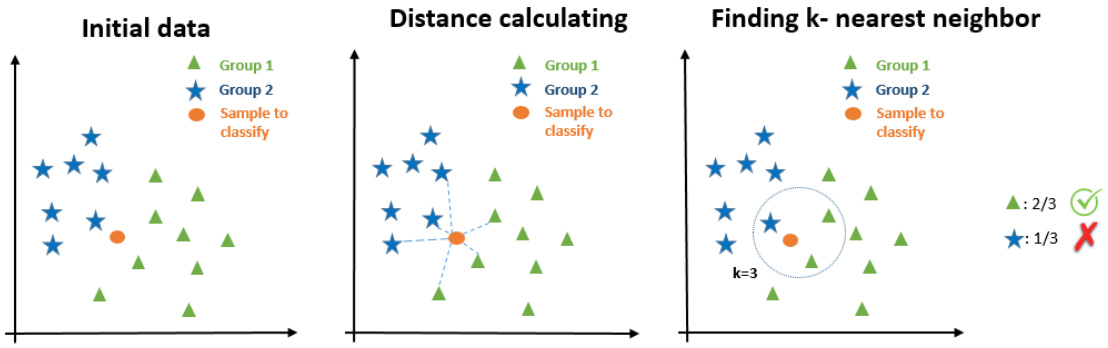


Figure 7.5: An example of the k-nearest neighbor classification (k=3). The sample is finally classified as belonging to group 1 ('triangle shape').

neighbor, the query sample is categorized as Group 1 (triangle shape), after checking the labels of the three nearest neighbors. This is interesting because the decision of choosing the number of neighbors has an impact on the final outcome.

7.1.2 Choosing the most adapted value for k

To find the most adapted k-value for the data, it is necessary to run the k-NN algorithm many times with different values of k and pick the k-value that decreases the amount of errors while maintaining the algorithm's capacity to make accurate predictions when presenting data that it has never seen before.

When the value of k equals to one, the predictions become less stable. Inversely, when k increases, the forecasts become more stable owing to majority voting/averaging and, as a result, more likely to be correct (up to a certain point). When one begins to notice an increase in the amount of errors, it means that the value of k has been pushed too much.

In order to get a majority vote among labels (e.g., determining the mode in a classification issue), k is normally chosen as an odd number to have a tiebreaker.

So, how to choose the right k value? There are no pre-defined statistical procedures for determining the optimal value of k. Choose a random k-value and begin calculating. A small value of k results in unstable decision boundaries. A high k-value is preferable for classification since it smooths out the decision boundaries. One needs to create a visualization of the error rate versus k for values within a given range. Then select the k-value with the lowest error rate. Typically, the ideal k-value is determined to be the square root of N, where N is the total number of samples (Jirina, 2011).

The k-NN classification is recommended as 'a theoretically optimal method of classification' (Hastie *et al.*, 2001). However, this method is not easy to interpret and it does not offer the possibility to extract a rules set in order to apply it for another dataset. In addition, the k-NN classification cannot deal with both numerical and categorical data at the same time. It is required to convert numerical data to categorical.

7.2 Decision Tree

Decision Tree (Quinlan, 1986) is a Supervised ML Algorithm that employs a set of rules to make decisions in the same way that people do. Some classification methods, such as Naïve Bayes, are probabilistic, although a rule-based technique is also available.

7.2.1 A tree that makes decisions

The idea behind Decision Trees is to use dataset attributes to create binary yes/no questions, and then segment the dataset until all the data points from each class become isolated. With this strategy, one can organize the data in a tree structure. A node is added to the tree when a question is asked. Furthermore, the first node is known as the root node. The answer to a question separates the dataset and creates new nodes based on the value of a characteristic. If the process is stopped after a split by some rules (for example: stop splitting if more than 95% belong to a single class, stop splitting if less than 5 individuals, do not split if the new node has less than 5 individuals, ...), the final nodes are known as leaf nodes.

A basic Decision Tree structure and its terminologies are introduced in the Figure 7.6:

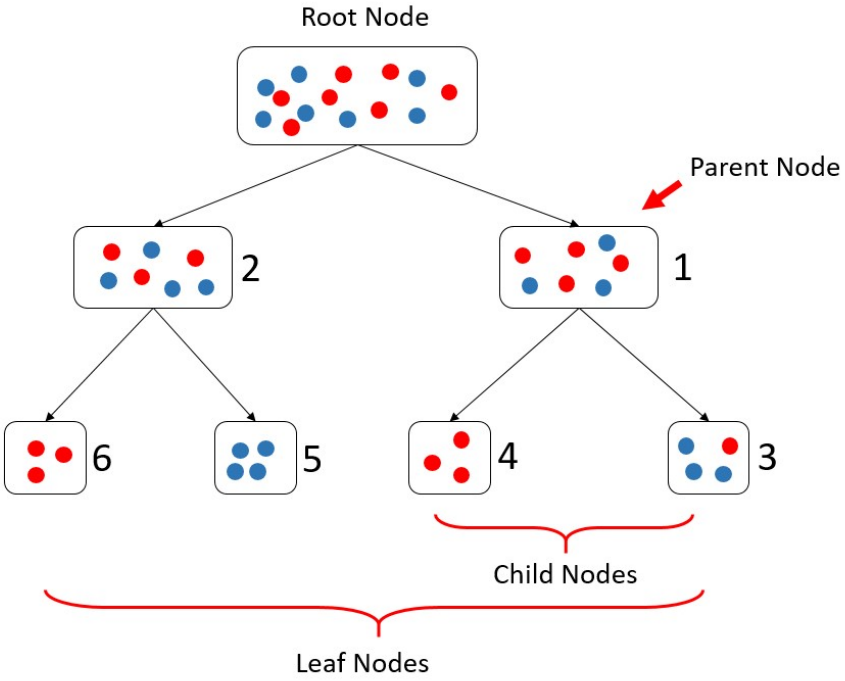


Figure 7.6: A basic structure of a Decision Tree.

Parent and Child Node: A Parent Node (nodes 1 and 2 in the figure) is a node that is divided into sub-nodes, and these sub-nodes are known as Child Nodes (nodes 3-6 in the figure). Because a node may be split into several sub-nodes, it can act as a Parent node for a large number of Child Nodes.

Root Node: The decision tree's highest node. There is no Parent node for this node.

Leaf / Terminal Nodes: Nodes that do not have any Child Node (nodes 3-6 in the figure).

The technique attempts to have all the leaf nodes belonging to a single class. These are referred to as pure leaf nodes (like nodes 4-6 in Figure 7.6). However, most of the time the final result consists in mixed leaf nodes, which means that not all data points belong to the same class (like node 3 in the figure). In the end, the algorithm can only assign one class to each leaf node's data points. With pure leaf nodes, there is no further ambiguity because all the data points in that node have the same class. However, in the case of mixed leaf nodes, the method assigns the most frequent class among all the data points in that node. For example in Figure 7.6, node 3 would be assigned to the class of blue color.

The ideal tree is the smallest tree with the fewest splits that can correctly categorize all the data points. This appears to be an easy issue; however, it is a nondeterministic polynomial (NP)-hard problem (see NP-hard problem's definition at the beginning of this Chapter 7). Building the optimal tree would require a polynomial time, which rises exponentially with the size of the dataset. For example, if a dataset contains only 10 data points and the algorithm is of quadratic complexity,

$O(n^2)$, the tree is built in 10*10 = 100 iterations. Increasing the size of the dataset to 100 data samples, the number of iterations of the algorithm will increase to 10.000.

To convert the NP-hard task into a computationally viable one, the solution employs a greedy strategy to create the next best tree. Instead of attempting to make the best overall decision, this method makes locally optimum judgments to select the feature utilized in each split. Because it optimizes for local decisions, it is solely concerned with the node at hand and what is optimal for that node in particular. As a result, it is not necessary to investigate all possible splits for that node and beyond (Bento, 2021).

Picking the best split

The algorithm attempts to partition the dataset into the lowest subset feasible at each split. The aim, like with any other Machine Learning method, is to minimize the loss function as much as feasible (Tan *et al.*, 2005). Stochastic Gradient Descent is a popular loss function for classification algorithms. Given that the loss function should be differentiable, it is not possible to use in this circumstance. However, because data points from distinct classes have to be separated, the loss function should assess a split based on the proportion of data points from each class before and after the split. In other words, a loss function that assesses the split based on the cleanliness of the resultant nodes is desirable. Examples of loss functions that compare the class distribution before and after a split are Gini Impurity and Entropy (Tan *et al.*, 2005).

- *Gini Impurity*

  Gini Impurity is a measure of the variance across the different classes (James *et al.*, 2013):

  $$G(node) = \sum_{k=1}^{c} p_k(1 - p_k) \tag{7.1}$$

  where p_k is the probability of picking a data point from class k, c is the total number of classes (or labels) of data.

- *Entropy*

  Similar to Gini Impurity, Entropy is a measure of chaos within the node. In addition, chaos, in the context of decision trees, means having a node where all the classes are equally present in the data.

  $$Entropy(node) = -\sum_{k=1}^{c} p_k \log(p_k) \tag{7.2}$$

  When using Entropy as a loss function, a split is performed only if the Entropy of each of the resulting nodes is lower than the Entropy of the parent node. Otherwise, the split is not locally optimal.

7.2.2 Decision Tree's advantages

Decision trees are based on a simple algorithm and present several advantages (Bento, 2021):

- **Interpretability**: The decision tree can be visualized. One of the most significant advantages of tree-based algorithms is the ability to visualize the model. You can see the algorithm's decisions and how it categorized the various data pieces. This is a significant benefit because most algorithms operate as black boxes, making it difficult to determine what led the algorithm to predict a specific result.
- **No preprocessing required**: There is no need to prepare the data before generating the model. Instead of examining the full feature set, the rules in tree-based algorithms are developed around each particular feature. Because each choice is made by examining one characteristic at a time, their values do not need to be normalized.
- **Data robustness**: The algorithm works well with all types of data. Tree-based algorithms are excellent when dealing with different data types. The dataset can contain both numerical and categorical data, and none of the categorical characteristics must be encoded.

Despite their benefits, Decision Trees are not as accurate as other classification and regression techniques. Overfitting is a disadvantage of decision trees. Overfitting the training set is common when designing a very long tree, partitioning the feature set until achieved pure leaf nodes. The resultant tree is not only complicated, but also difficult to read and interpret. Optimal trees can also be spruced to avoid overfitting, but it requires another data set. However, if the decision tree is too tiny, it will underfit the data, resulting in excessive bias.

In addition, Decision Trees are robust in terms of the data types they can handle, but the algorithm itself is not very robust. A slight change in the data can drastically change the tree and, consequently the final results (James *et al.*, 2013).

To summarize, Decision Trees are a rule-based method for solving classification and regression tasks. They divide the dataset using the values in each feature to group all data points with the same class together. However, there is an obvious trade-off between interpretability and performance. A small tree is simple to perceive and comprehend, but it contains a lot of variation. A little modification in the training set can result in an entirely different tree and predictions. A large tree with several splits, on the other hand, produces better classifications. However, it is most likely, to remember the training dataset (overfitting).

7.3 Kernel Approximation

When data is not linearly separable in the current feature space, kernel techniques project input data points into a high-dimensional feature space *via* nonlinear mapping and determine the appropriate hyperplane in that feature space. However, if the number of data points is enormous, the size of the kernel matrix will be large, which may influence the efficiency of the algorithms. For example, computing the kernel matrix when the size of observation matrix is mxn, leads to an algorithm with at least O(n^2) storage space and O(n^2m) running time. In addition, the inverse of a kernel matrix calculation is the main computational burden with O(n^3)

running time. Facing with this problem, the study of low-rank approximation of the kernel matrix to reduce the algorithm complexity is introduced (Zhao *et al.*, 2016).

7.3.1 Kernel-based method

Kernel methods like Support Vector Machines (SVM) or kernelized PCA rely on the property of reproducing kernel Hilbert spaces (Schölkopf and Smola, 2002). For any positive definite kernel function K (a so-called Mercer kernel), it is guaranteed that there exists a mapping into a Hilbert space H, such that

$$K(x_i, x_j) = <\phi(x_i), \phi(x_j)> \tag{7.3}$$

where $<.,.>$ denotes the inner product in the Hilbert space. x_i, $x_j \in X$ are two samples. The nonlinear feature mapping $\phi : X \to F$ maps each element of the observation space X into a high-dimensional feature space F. By using this mapping method, it is not necessary to represent explicitly as long as kernel algorithms have access to K. That means $\phi(x)$ can be high-dimensional or even infinite-dimensional, the inner products can be evaluated in an inexpensive manner by K (Le *et al.*, 2013). This is referred to as the "kernel trick".

An example illustration of this mapping method for SVM is displayed in the Figure 7.7.

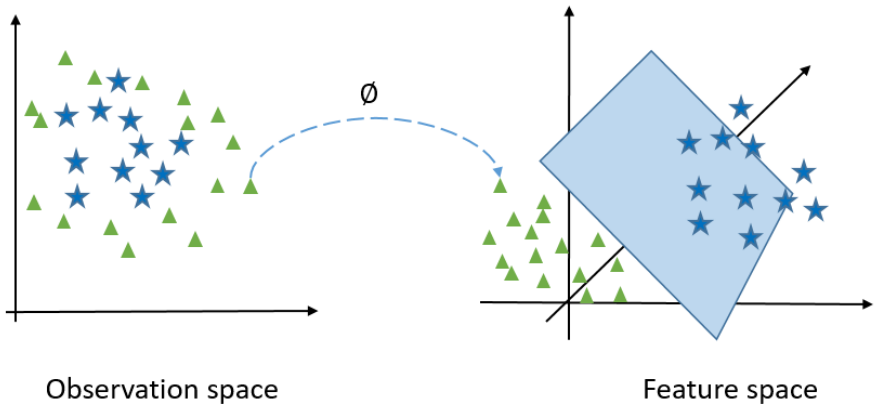


Figure 7.7: Kernel trick in Support Vector Machine.

The kernel function $K(x_i, x_j)$ is intended to measure the "similarity" between x_i and x_j (the larger, the more similar). The two most widely-used such functions are

- Linear kernel: $K(x_i, x_j) = <x_i, x_j>$
- Gaussian kernel: $K(x_i, x_j) = -\frac{\exp\left\|x_i - x_j\right\|^2}{2\sigma^2}$

The linear kernel gives a signed measure of the similarity between x_i and x_j, in the sense that the angle between the two points plays a role in determining how similar they are, and can lead to negative values of K. On the contrary, the Gaussian

kernel depends only on the Euclidean distance between x_i and x_j, and is based on the assumption that similar points are close one to each other in the feature space (in terms of Euclidean distance). This latter assumption is very reasonable in many cases, hence the Gaussian kernel is often used in practice.

The advantage of using the Kernel function K is that the mapping ϕ never has to be calculated explicitly, allowing for arbitrary large features (even infinite). Meanwhile, one drawback of the kernel methods is that it might be necessary to store many kernel values K(x_i,x_j) during optimization. If a kernelized classifier is applied to new data x_k, K(x_i,x_k) needs to be computed to make predictions, possibly for many different x_i in the training set (Pedregosa *et al.*, 2022).

7.3.2 Kernel Approximation and its advantages

Kernel approximation is an effective technique for overcoming the low scalability of kernel-based techniques by establishing an explicit mapping ψ: $R^d \rightarrow R^s$ such that $K(x,y) \approx \psi(x)^T\psi(y)$. By doing so, an efficient linear model can be well learned in the transformed space with $O(ns^2)$ time and $O(ns)$ memory while retaining the expressive power of nonlinear methods, where n is the number of samples in the original d-dimensional space and s is the number of features, which is normally a very high number. According to the review paper of Liu and colleagues (2021), in recent years, a number of kernel approximation algorithms, such as divide-and-conquer approaches (Hsieh *et al.*, 2014), greedy basis selection techniques (Alex and Bernhard, 2000) and Nyström methods (Williams and Seeger, 2001), have been developed.

While these approaches provide a data-dependent vector representation of the kernel, the Random Fourier features (RFF) (Rahimi and Recht, 2008), on the other hand, is a typical data-independent technique to approximate the kernel function. The Random Features is one of the most popular techniques to speed up kernel methods in large-scale problems. For further information, RFF applies in particular to shift-invariant (also called "stationary") kernels that satisfy $K(x_1, x_2) = K(x_1 - x_2)$. By virtue of the correspondence between a shift-invariant kernel and its Fourier spectral density, the kernel can be approximated by $K(x_1, x_2) \approx \phi(x_1), \phi(x_2)$, where the explicit mapping $\phi : R^d \rightarrow R^s$ is obtained by sampling from a distribution defined by the inverse Fourier transform of K. To scale kernel methods in the large sample case (e.g., $n \gg d$), the number of random features s is often taken to be larger than the original sample dimension d but much smaller than the sample size n to achieve computational efficiency in practice.

The Random Kitchen Sinks (Rahimi and Recht, 2008) and Fastfood (Le *et al.*, 2013) are two examples of random feature expansions; these schemes tried to approximate Gaussian kernels of the kernel classification algorithm for big data in a computationally efficient way. Firstly, they find a random transformation so that its dot product approximates the Gaussian kernel:

$$K(x_1, x_2) = <\phi(x_1), \phi(x_2)> \approx T(x_1)T(x_2)^T \tag{7.4}$$

where $T(x)$ maps x in R^p (p is the number of input features) to a high-dimensional space (R^m). The Random Kitchen Sinks scheme uses the random transformation

$$T(x) = m^{-1/2} exp(iZx^T)^T \tag{7.5}$$

where $Z \in R^{mxp}$ is a sample drawn from $N(0, \sigma^{-2})$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage.

The Fastfood scheme introduces another random basis V instead of Z using Hadamard matrices combined with diagonal Gaussian scaling matrices.

$$V = \frac{1}{\sigma\sqrt{d}} SHG\Pi HB \tag{7.6}$$

where $\Pi \in \{0,1\}^{dxd}$ is a permutation matrix and H is the Walsh-Hadamard matrix. S,G and B are all diagonal random matrices.

When the implemented function uses the Fastfood scheme for random feature expansion and uses linear classification to train a Gaussian kernel classification, the model needs only to form a matrix of size nxm, with m typically much less than n for big data, in comparison with support vector machine that requires computation of the nxn Gram matrix. This random basis reduces the computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

Kernel approximation is used in a variety of contexts and its use is crucial for scaling many learning algorithms to a very large task (Cortes *et al.*, 2010). Approximate subspaces can be built adaptively using streaming data acquisition. After that, explicit feature vectors are obtained using a transformation onto the estimated subspace, and linear learning approaches can be used. In terms of computation, processes in kernel techniques can be easily parallelized, and advanced infrastructures may be used to create efficient computing. Furthermore, the produced explicit feature vectors may be simply integrated with different learning approaches (Yu *et al.*, 2018).

Chapter 8

Models Implementation, Results and Discussion

This chapter presents the implementation procedure for three prediction models. Firstly, the parameters selection based on autocorrelation function values is presented. In the second section, the detailed information for data preparation and hyperparameter values setting are introduced.

8.1 Parameters selection

The data quality and quantity have an influence on the majority of data-driven techniques, including data mining and machine learning. Furthermore, it is important to determine which factors impact the target value (the model output) and how many features (model inputs) should be used to build predictive models. In practice, several environmental factors may influence the accuracy of window opening prediction. However, due to realistic limits, it is impossible to search for all of these features. According to some previous studies, the outdoor temperature, indoor CO_2 concentration and the prevailing mean air temperature (see section 2.2 for the definition) are the most important variables in determining the probability of opening/closing the windows, followed by indoor air temperature, outdoor and indoor humidity (Andersen *et al.*, 2013; Fabi *et al.*, 2012; Park *et al.*, 2020; Yao and Zhao, 2017).

In addition, non-environmental factors, such as: seasonal change, time of the day and personal preference, also affect the window-opening probability (Pan *et al.*, 2018). Thus, in our model, the following variables were selected:

- temperature (T) and specific humidity (Hs) of both indoor and outdoor environments and the prevailing mean outdoor air temperature (PMA);
- indoor CO_2 and indoor particulate matter concentrations (PM2.5 and PM10);
- wind direction, raining condition, door status, occupancy status;
- month, day of the week, hour of the day.

The studied office has a permanent mechanical exhaust ventilation. The single flow ventilation system provides a constant air extraction rate of 228 m^3. h^{-1} (measured in 2014 at ± 6%). Ten air inlets are attached to the joinery of the five sliding

windows. These five windows were equipped with sensors that detected each opening or closing event and recorded by the CSTBox-DIN4 through a wireless motion detector (see Figure 2.6). The collected data is transmitted to and stored on a central server, which enabled viewing and downloading. The opening factor data are time series with irregular time steps as the detection modules send back information as soon as a change of state occurs. Therefore, a pre-processing step was performed to synchronize all the time series at the same time step (1 minute) (Ramalho *et al.*, 2016).

The main statistics of the monitored environmental parameters for the years 2014 and 2015 are displayed in Table 8.1 and Table 8.2, respectively. It should be noted that the comparison of these two years is not very representative as 2014 data covered the whole year, and the 2015 monitoring set covered only the first 6 months. However, there are no significant differences between the averaged values of these two years. One can notice that the maximum values of PM2.5 and PM10 concentrations in 2014 are quite higher than those monitored during 2015 (91.87 μg.m^{-3} and 106.78 μg.m^{-3} in 2014 in comparison with 21.3 μg.m^{-3} and 43.71 μg.m^{-3} in 2015). This can be explained by the outdoor pollution episode of particulate matter that happened in March 2014, a quite remarkable event. In addition, higher specific humidity is observed in 2014 compared to 2015, but the monitored data of 2015 does not include July to December.

Table 8.1: The statistics for environmental parameters of 2014

Features	**Indoor CO_2** (ppm)	**Indoor $PM_{2.5}$** ($\mu g.m^{-3}$)	**Indoor PM_{10}** ($\mu g.m^{-3}$)	**Indoor T** (°C)	**Outdoor T** (°C)	**Indoor Hs** (g/kg)	**Outdoor Hs** (g/kg)
Max value	1144.00	91.87	106.78	31.30	35.60	15.11	17.30
Min value	416.80	0.26	0.31	15.00	-4.30	4.28	3.98
Mean value	501.10	2.47	4.32	23.00	13.50	8.88	9.65
Median value	480.50	1.76	3.15	22.40	13.50	8.95	9.66
Std value	64.30	2.87	4.18	2.30	6.00	1.91	2.47

Table 8.2: The statistics for environmental parameters of 2015

Features	**Indoor CO_2** (ppm)	**Indoor $PM_{2.5}$** ($\mu g.m^{-3}$)	**Indoor PM_{10}** ($\mu g.m^{-3}$)	**Indoor T** (°C)	**Outdoor T** (°C)	**Indoor Hs** (g/kg)	**Outdoor Hs** (g/kg)
Max value	1038.82	21.30	43.71	33.33	39.22	13.33	14.94
Min value	421.48	0.13	0.16	18.24	-1.80	3.55	3.48
Mean value	498.45	2.50	4.45	23.10	11.28	6.44	7.11
Median value	477.02	1.93	3.40	22.30	10.30	6.22	6.69
Std value	61.38	2.11	3.70	2.43	7.01	1.55	2.09

In reality, the windows opening status does not change much within a given hour, hence using such a detailed database with a 1-minute time step is not necessary. In addition, some monitored data were missing, therefore we decided to use the hourly average data in this study.

Based on the 1-minute time step data, the hourly average values of the selected parameters were calculated as in equation 8.1. A linear interpolation was applied in order to replace missing values.

$$x_{hourly} = \frac{1}{60} \sum_{i=1}^{60} x_{minute_i} \tag{8.1}$$

The window opening status for a specific hour was calculated as the mode value (most frequent) of the number of opened windows, according to the equation 8.2.

$$x_{hourly} = mode(x_{minute_i}) \qquad (0 < i \leq 60) \tag{8.2}$$

As the value of opened windows is not varied much during 60 minutes, the mean (averaged) and mode value are almost the same, so we could use either of them.

The ACF results of the environmental data monitored during 2014 are represented in the Figure 8.1. Very similar results were obtained for data of the year 2015 so they are not presented here. One can notice the persistence of the temperature (T) and specific humidity (Hs) indoors and outdoors, which means that a value at time t of the temperature or specific humidity is correlated to a value one day later (t+24), two days later (t+48), or even three days later (t+72). In addition, the ACF of the CO_2 concentrations becomes negative and remains at low levels, and then switches back to positive values after a lag of 17 hours. While for outdoor T and Hs (indoors and outdoors), the autocorrelations persist in the positive domain for long delays. In general, temperatures depict the same structures of spectral variability as CO_2: the fundamental frequency peaks at every 24 hours. The ACF of CO_2 alternates sign every 8 hours on a lag of 24 hours. This implies that, instead of using the information of the 'previous hour', in the real-time system, we could use the value of 'the previous 24^{th} hour' (t-24) environmental data as input for this model, which is easier to access.

Furthermore, the 'weekly periodicity' (at the lag of 168 hours) in the ACF values of CO_2 and PM10 concentration is noteworthy. The information of the 'previous 168^{th} hour' data could be then used as input for the model when the 'previous 24^{th} hour' data is not available. Besides, it can be also noticed that the ACFs of PM concentrations and number of opened windows present high values at a lag of 24 hours (see Figure 8.1d and 8.1c). We decided to use also the PM concentrations and the number of opened windows, corresponding to the 24 hours lag, as inputs of the prediction model.

In conclusion, non-environmental, environmental features and window status of **the previous 24^{th} hour** moment, were selected as initial inputs of a model built in order to predict the opening status of windows at the **current hour** as presented in the section 8.2.

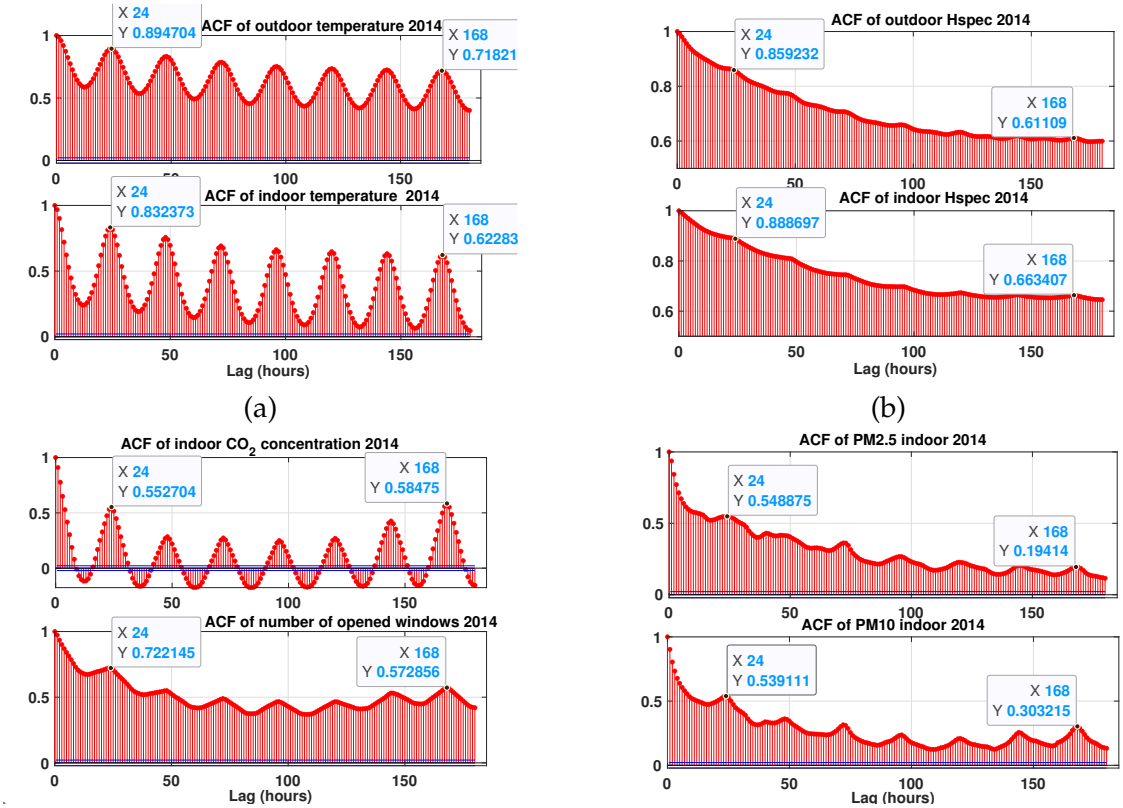


Figure 8.1: Autocorrelation values of environmental variables in 2014: (a) Indoor and outdoor temperature, (b) indoor and outdoor humidity, (c) indoor CO_2 and number of opened windows, and (d) indoor PM2.5 and PM10. The 24-hour and 7-day peaks are indicated on the plot of each ACF (X represents the lag and Y represents the ACF value).

8.2 Classification model Implementation

In this section, the data pre-processing is introduced, followed by the model's parameterization.

8.2.1 Data pre-processing

After recalculating the number of opened windows for a specific hour using the mode value (equation 8.2), these values were then categorized into four different groups, labeled as follows:

- ALL CLOSED: all of the windows are closed ($x_{hourly} = 0$)
- MOSTLY CLOSED: 1 window is opened ($x_{hourly} = 1$)
- MOSTLY OPENED: 2 or 3 windows are opened ($2 \leq x_{hourly} < 4$)
- ALL OPENED: 4 windows or more are opened ($x_{hourly} \geq 4$)

We remind that the office is equipped with five windows. In 2015, one window sensor was out of order, thus the respective window remained closed all the time. Therefore, the maximum number of opened windows is five in 2014 and four in 2015.

The distribution profiles according to the non-environmental parameters (month, day of the week and hour of the day) and the initial statistics of these four groups during the years 2014 and 2015 are displayed in Figure 8.2 and Figure 8.3, respectively.

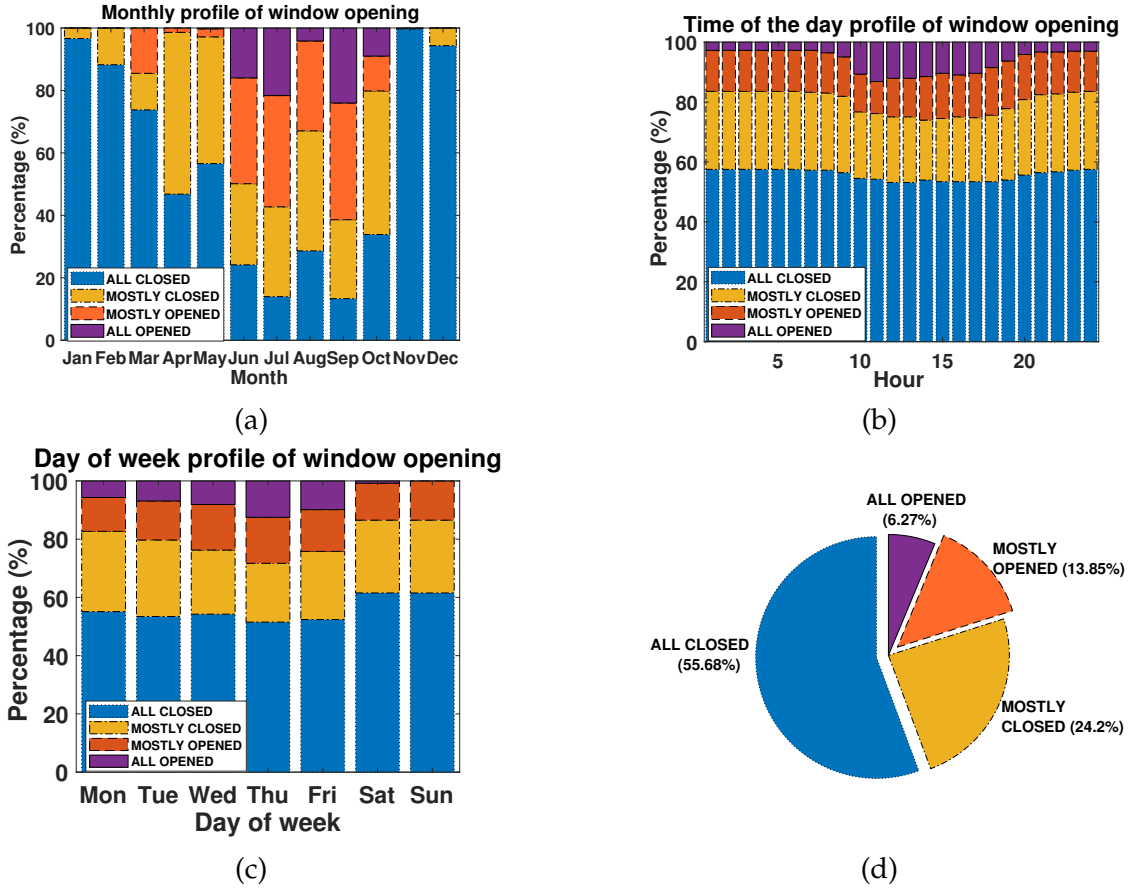


Figure 8.2: Distribution profile of window opening of 2014 according to the (a) Month, (b) Hour of the day and (c) Day of the week.(d) Statistics for window opening categories.

Figure 8.2d shows that in 2014, for more than half of the time (55.68%), the status of this group of windows is 'ALL CLOSED'. This label is dominant during the winter period (November – March). 'MOSTLY CLOSED' and 'MOSTLY OPENED' labels are quite equally distributed with 24% and 14%, respectively. The fourth label 'ALL OPENED' accounts for just 6.3% of the total time and it appears only in summer and the beginning of autumn (June – October) and during the working time (9 a.m. – 6 p.m.) only. This is expected because "during the working time, the occupants tend to open at least one window, and rarely open the full five windows at the same time" (Ramalho *et al.*, 2016).

The statistics for the window opening state according to the previous defined categories show in 2015 even a higher percentage (88.9%) of the "ALL CLOSED" label.

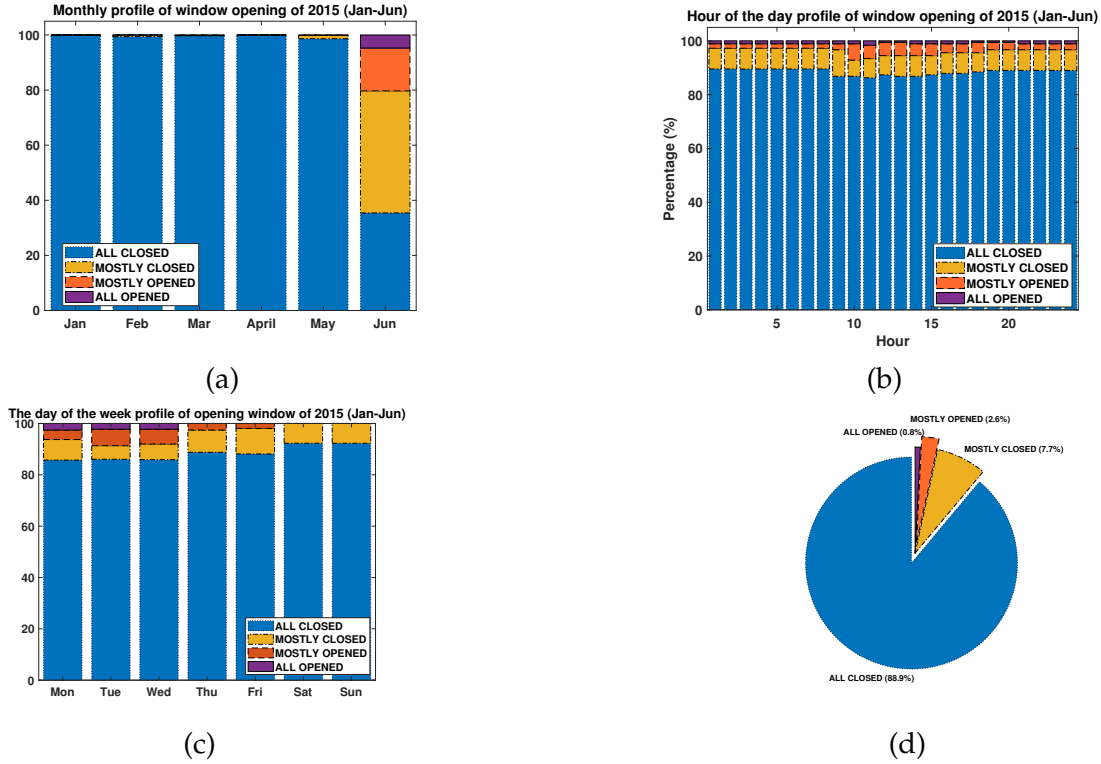


Figure 8.3: Statistic profile of 4 groups of window opening from January to June of 2015 according to (a) Temperature, (b) Specific humidity (c) CO2 concentration and (d) PM concentration.

The "ALL OPENED" label is obtained only in June with 0.8% for the 6-month period. The "ALL CLOSED" profile can be observed almost all the time from January to April (Figure 8.3a). This is quite different in comparison with the distribution profile of the year 2014 without an obvious reason.

Regarding the environmental parameters, Figure 8.4 presents the mean values and the standard deviations of these variables according to the groups. Differences in the mean values of the outdoor temperature, specific humidity (indoors and outdoors) and PM10 indoors can be observed for the four windows categories (Figures 8.4a, 8.4b and 8.4d). For these parameters, the higher the value, the greater number of windows are opened. For indoor temperature and PM2.5 the differences among groups are small. The indoor mean CO_2 concentration keeps a stable value among these four groups (Figure 8.4c). Given that the measurement uncertainty is 50 ppm ± 3% for reading, the range of variation 480-520 ppm is less than the uncertainty. So, one can consider that the CO_2 value does not vary significantly, which means that the office is "well ventilated".

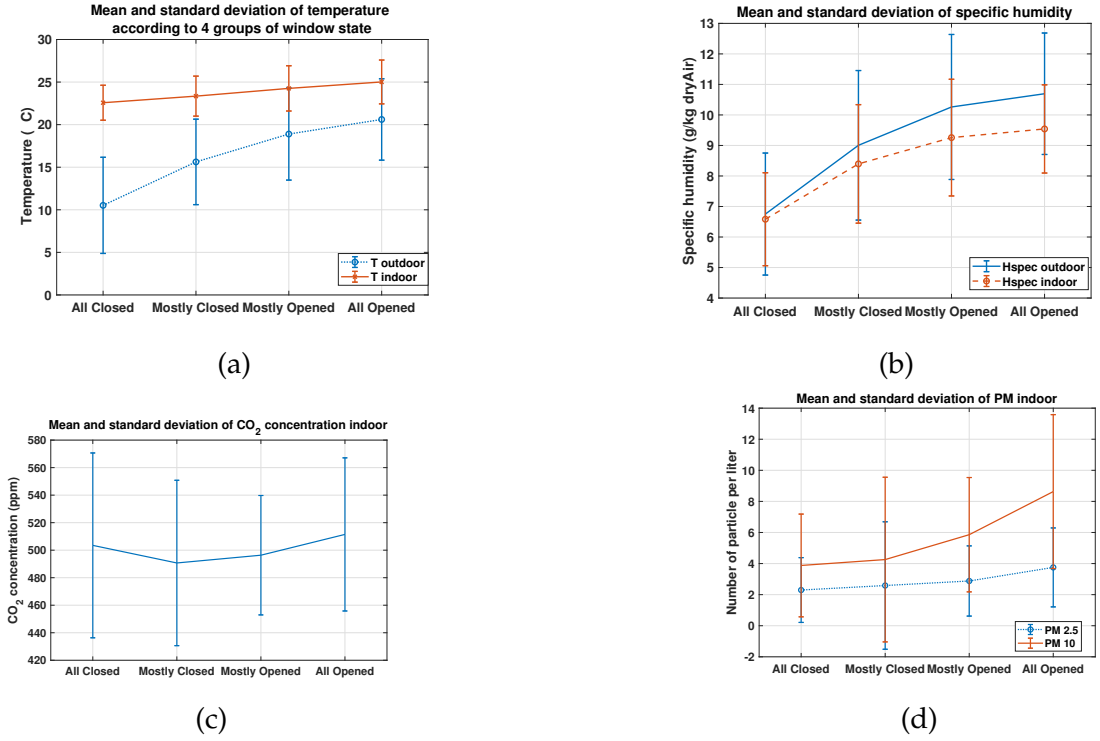


Figure 8.4: Statistics (mean and standard deviation) of (a) Temperature, (b) Specific humidity (c) CO_2 concentration and (d) PM concentration, for each opening label, from January to June 2015.

For the model implementation, different data sets are required: training, validation, testing, etc. We decided to divide the time series data into sets of 25 hours and use the 20 first hours for training and validation, and the remaining 5 hours for testing (ratio 80:20 – see Figure 8.5). The reason why we did not use the day 365^{th} for training is that we need the windows status of this day to evaluate the testing set of the 364^{th} day ('previous 24^{th} hour'). In total, 6980 hours were used for training.

A 10-fold Cross-Validation (CV) has been applied to the training dataset. The purpose of Cross-Validation (CV) is to detect possible over-trained models with high internal accuracy and low external prediction power (overfitting problem). The diagram showing how the training dataset has been split for the 10-fold CV method is displayed in Figure 8.6. The CV divides the training data into ten groups. The model trains on 9 sets and is evaluated on the 10^{th} set, not used for training, during each iteration. To decrease variability, several CV iterations should be used (normally 10 iterations in the case of 10-fold). The model's performance is evaluated using the average error over all iterations.

As the k-NN method can not deal with both numerical and categorical data at the same time, quantitative data had to be recoded to generate qualitative (categorical) data. Numerical data were obtained from environmental parameters monitoring; in order to be transformed into categorical data, the values of each variable were divided into 10 groups (or categories) based on their percentiles in order to equally represent the groups. The first 10 percentiles belong to the first group, the data of percentiles from 11 to 20 belong to the second group, and so on.

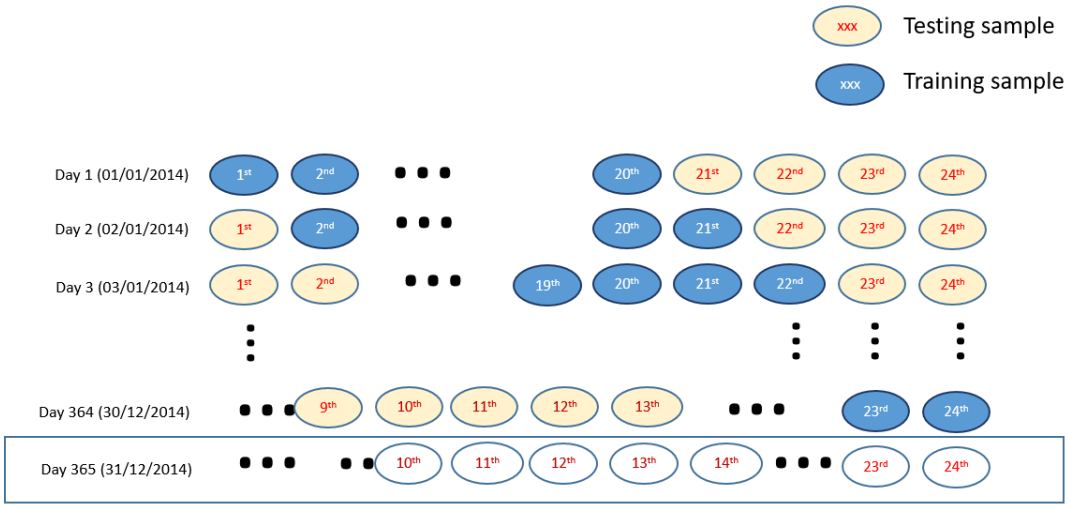


Figure 8.5: Figure explaining how the data has been split into training and testing sets (sets of every 25 hours).

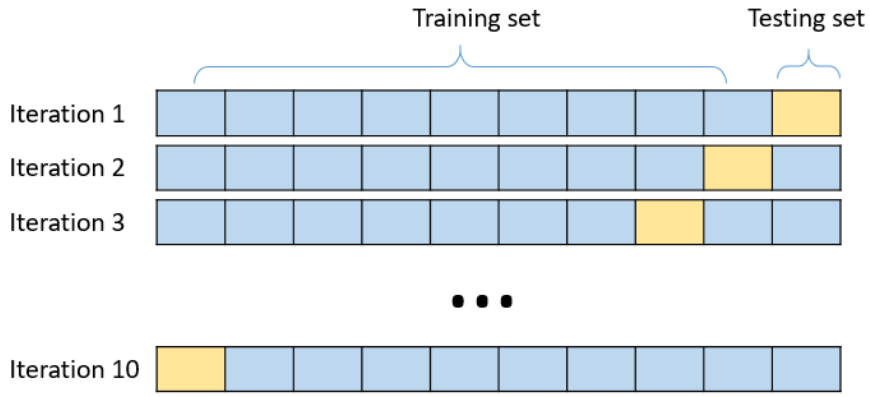


Figure 8.6: The scheme for the 10-fold cross validation method.

8.2.2 Model's parameterizations

The Classification Learner application of Matlab® was used for the model development. The 'OptimizeHyperparameters' option for 'all' the input parameters was used to obtain the best values for the hyperparameters of the models and to avoid overfitting. This optimization attempts to minimize the cross-validation loss (error) by varying the parameters. The summary of the obtained values of the different hyperparameters for the three models are presented in Table 8.3.

Table 8.3: Summary of the different hyperparameters for the three models.

Algorithm	Hyperparameter	Value
Decision Tree	Maximum number of Splits Split Criterion Minimum leaf size	4454 deviance 1
k Nearest Neighbor	Number of neighbor (k) Distance metric function Standardize	3 hamming true
Kernel Approximation	Kernel function Polynomial Order Standardize	polynomial 3 true

The other general parameters of the models are listed below:

- Number of data – training set: 6980 samples (80% data of 2014)
- Number of data – testing set:
  - Testing set of 2014 (which will be called 'test set 2014'): 1745 samples (the rest of 20% data of 2014)
  - Testing set of 2015 (which will be called 'test set 2015'): 4345 samples (data from January to June 2015)
- Data type: hourly averaged data
- Validation method: 10-fold cross validation
- Initial number of input variables: 16 variables as in Table 8.4:

Table 8.4: Summary of the input variables for the predicting model.

Index	Name	Variable	Type of data	Moment
1	Month	month	categorical	Current moment
2	DoW	day of the week	categorical	Current moment
3	HoD	hour of the day	categorical	Current moment
4	T_out	outdoor temperature	numerical/categorical[a]	previous 24th hour
5	T_in	indoor temperature	numerical/categorical	previous 24th hour
6	Hs_out	outdoor specific humidity	numerical/categorical	previous 24th hour
7	Hs_in	indoor specific humidity	numerical/categorical	previous 24th hour
8	CO2_in	indoor CO_2 concentration	numerical/categorical	previous 24th hour
9	PM2.5in	indoor PM2.5 concentration	numerical/categorical	previous 24th hour
10	PM10in	indoor PM10 concentration	numerical/categorical	previous 24th hour
11	Prv_Wd	state of group of windows	categorical	previous 24th hour
12	PMA	prevailing mean outdoor air temperature	numerical/categorical	previous 24th hour
13	WindD	wind direction	categorical	previous 24th hour
14	Rain	raining status	categorical	previous 24th hour
15	Occ	occupancy status	categorical	previous 24th hour
16	Door	entrance door status	categorical	previous 24th hour

[a]This variable is coded in 10 categories for the k-NN classification model. For Kernel Approximation and Decision Tree, the monitored numerical data is kept as original.

8.3 Results

In this section, some results concerning the decision boundary and the rank of the importance scores of the predictors are firstly presented. Than, the performance of different ML prediction models is explored. Finally, several evaluation methods are provided, completed by a discussion in the last section of the chapter.

8.3.1 Decision boundaries

Understanding decision boundaries can provide us a visualization of how the training data we choose influence our model's performance and capacity to generalize. By observing the decision boundaries, one can see how sensitive models are to each dataset, which is a great technique to understand about how various algorithms perform, and their limits for specific datasets.

Figure 8.7 illustrates an example of decision boundary for the nearest neighbor classification on the Iris dataset (Fisher, 1936). According to this figure, three Iris species are classified based on the value of sepal length and width. For example, if a sample's sepal length is less than 4.5 cm, this sample is categorized to the red color specie. Besides, if the sample's sepal length is in the range of 5.0 and 6.5 cm, the assigned specie group is then determined by the sepal width of the sample.

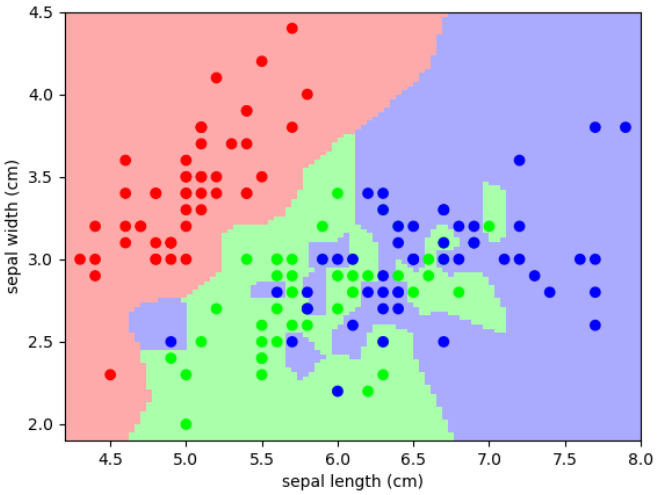
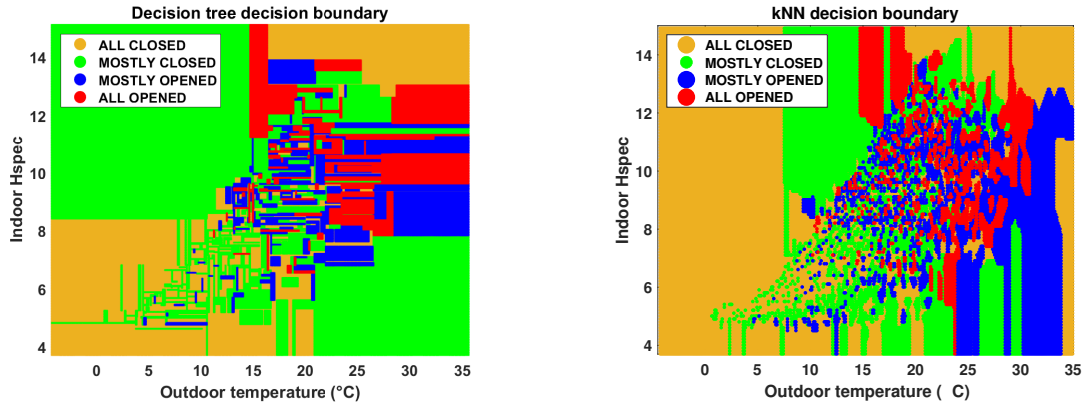


Figure 8.7: An example of decision boundary of nearest neighbor classification on iris dataset (scipy lectures.org, 2022).

The results of the decision boundaries of the Decision Tree and k-NN classification models based on the values of outdoor temperature and indoor specific humidity in our dataset are displayed in Figure 8.8.



(a) (b)

Figure 8.8: Decision boundary for a window status prediction model based on outdoor temperature and indoor specific humidity when using (a) Decision Tree model or (b) k-NN model.

It is obvious that there are differences in boundaries between these two classification methods. While Decision Tree uses straight lines to separate the window's states, the k-NN algorithm categorized the samples into different 'dot groups'. Both methods seem to be very sensitive since they have shown some extreme classification probabilities that are influenced by single points.

Interestingly, the color zones of these two figures are somehow similar for the 'ALL CLOSED' class. For example, it is 'ALL CLOSED' when the temperature outdoors is smaller than 16°C and the specific humidity indoors is smaller than 8.5 (g/kg), or, when the temperature outdoors is greater than 20°C and the specific humidity indoors is greater than 12.5 (g/kg).

In fact, determining the exact boundary line to separate the four groups is quite difficult. The state of window opening is determined by several factors, not just by these two features (outdoor T and indoor Hs). In addition, the features does not have the same influence on the classification. Some can effectively separate all classes, while others might work with only a subset, and some might not be helpful for class separation at all. In this case, an n-dimensional hyperplan is needed to display the decision boundary.

8.3.2 Rank of the importance scores of predictors

Because input variables have a direct influence on the model predictive performance, it is essential to determine which variables are the most important for the model development. The input selection is based on the relevance of the different predictors by evaluating the relative contribution of a given input to the performance of a particular model. This approach is called model-dependent and the advantage of this method is that the input selection is strongly related to the model performance, giving useful information for building predictive models.

Tree-based models have the advantage of being able to deal with massive volumes of data and a wide range of features while being simple to comprehend.

Therefore, in this study, a decision tree model was used to assess the predictor importance for the different factors on the window opening status. Similar results were obtained for the other two methods (k-NN and Kernel Approximation) and will not be presented here.

The "predictorImportance" Matlab function computes the importance measures of the predictors (model inputs) in a tree by summing changes in the node risk due to the splits on every predictor, and then by dividing the calculated sum by the total number of branch nodes (MathWorks, 2021). The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. If a tree splits a parent node (for example, node 1 as in Figure 7.6) into two child nodes (nodes 2 and 3 in Figure 7.6), then the "predictorImportance" increases the importance of the split predictor by

$$(R_1 - R_2 - R_3)/N_{branch} \tag{8.3}$$

where R_i is the node risk of node i, and N_{branch} is the total number of branch nodes. A node risk is defined as a node error or node impurity weighted by the node probability:

$$R_i = P_i E_i \tag{8.4}$$

where P_i is the node probability of node i, and E_i is either the node error (for a tree grown by minimizing a non-impurity criterion such as the Mean Squared Error (MSE) or the Twoing criterion) or the node impurity estimated *via* different criteria such as Gini Impurity or Entropy (see section 7.2 for more details) of node i.

- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $$1 - p_j \tag{8.5}$$

  where p_j is the probability of picking a data point from class j.

- The Twoing rule is not a purity measure of a node, but is a different measure for deciding how to split a node. Let L(i) denote the fraction of the members of class i in the left child node after a split, and R(i) denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

  $$P(L)P(R)(\sum_i |L(i) - R(i)|)^2 \tag{8.6}$$

where P(L) and P(R) are the fractions of observations that split to the left and to the right respectively. If the value is large, the split made each child node purer. Similarly, if the value is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.

Figure 8.9 shows the relative importance of the factors for the window opening status prediction by using the Decision Tree model. Similar results were obtained

for the other two methods (k-NN and Kernel Approximation) and will not be presented here. This figure shows the relative significance of the categorical variables (month, day of the week, hour of the day, and the previous 24th hour windows state), as well as the previous 24th hour value of the prevailing mean outdoor temperature outdoors (PMA). According to this observation, these parameters are the most important ones for this modeling. Surprisingly, an important influencing factor - the outdoor temperature, has a small effect on the model's performance. This can be explained by the substantial impact of the specific humidity and PMA, which are calculated using the outside temperature value as in the equations (2.1) and (2.3). The rain condition and the status of occupancy show very low importance. Based on this result, we decided to implement the models without these two parameters (Rain and Occupancy). In conclusion, 14 parameters were selected as inputs for our predicting models: Month, DoW, HoD, T_out, T_in, Hs_out, Hs_in, CO2_in, PM2.5in, PM10in, Prv_Wd, PMA, WinD, Door.

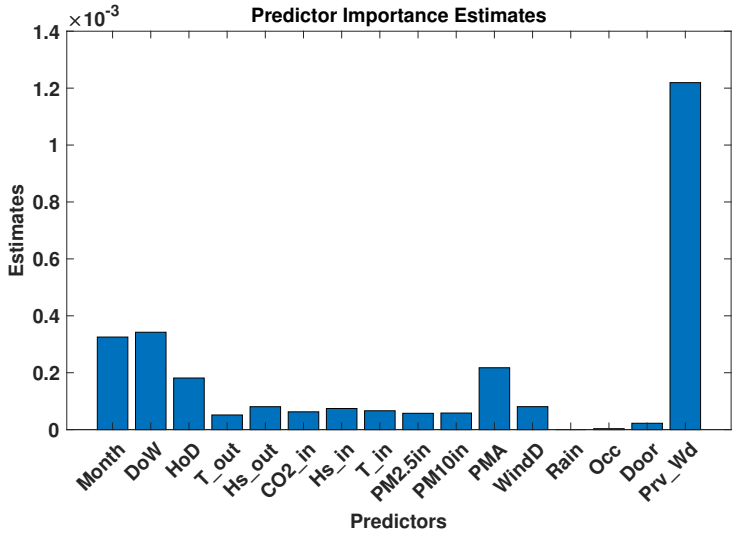


Figure 8.9: Predictors importance for predicting window opening status for a Decision Tree with the input containing all the available parameters. The Month, DoW and HoD correspond to the current moment, all the other variables correspond to the previous 24th hour (see table 8.4).

8.3.3 Performance of the window opening state model

Data monitoring starts on the 1st of January 2014 and ends on the 30th of June 2015 (13104 samples-hours). We have decided to use 80% of the data measured during the year 2014 for the training and validation sets (6980 samples). The remained data was divided into 2 sets for testing: (i) the rest of 20% of the data of the year 2014 (1745 samples) and (ii) data from January 2015 - June 2015 (4345 samples), because we want to observe the different behaviors of the built model when it has to deal with data of the same period (the same year 2014) and with data from a completely new period (data of 2015).

8.3.3.1 Performance of the Decision Tree classifier

Based on the results of the hyperparameters optimization presented in the Table 8.3, a Decision Tree of 541 nodes (Tree Depth = 16) has been obtained after using 80% data of the year 2014 for training and validation, with accuracies of 98.09% and 89.81%, respectively. Using this trained decision tree, we predicted the testing set containing the rest of 20% of the 2014 data and then we compared it to the monitored values. A value of 86.36% for accuracy (% of well-classified data) was achieved for this test.

A confusion matrix of the Decision Tree method for this testing set is displayed in Figure 8.10. The confusion matrix, usually known as an error matrix (Stehman, 1997), is a specific table that provides visualization of the performance of an algorithm, most commonly in a supervised learning algorithm (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents samples from an actual class, whereas each column represents samples from a predicted class. The name of this table is derived from the fact that it is simple to determine whether the system is confusing two classes (i.e. commonly mislabeling one as another).

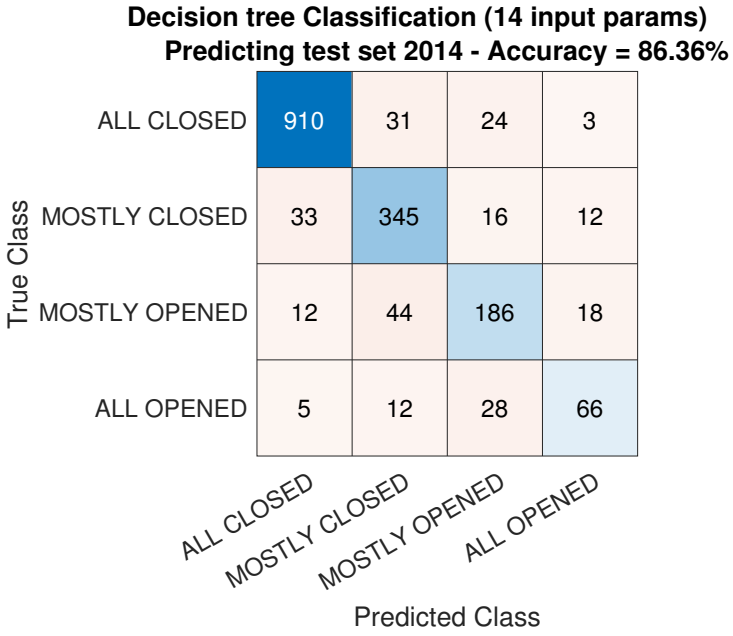


Figure 8.10: Confusion matrix of the Decision Tree classification for test set including the remaining 20% of 2014 data.

As we can see from the Figure 8.10, the model has a tendency of mislabeling one sample as a 'neighbor label'. The explanation for this could be that the environmental factors change gradually, the 'ALL OPENED' and 'ALL CLOSED' states are easily identifiable, but the 'ALL CLOSED' and 'MOSTLY CLOSED' ones can be ambiguous. The decision tree achieves 910 correct predictions and misses 58 (31+24+3) when the true label is 'ALL CLOSED'; 31 samples were incorrectly predicted being labeled as 'MOSTLY CLOSED' state, 24 samples were wrongly labeled as 'MOSTLY OPENED,' and 3 samples were misclassified as

'ALL OPENED.' Similarly, when the true label is 'MOSTLY CLOSED,' 345 samples are properly predicted whereas 61 are incorrectly classified (33+16+12). The labels 'MOSTLY OPENED' and 'ALL OPENED' are accurately predicted in 186 and 66 examples, respectively.

Using the same trained Decision Tree classifier, the window status of the first 6 months from January to June, 2015 were predicted, and compared to the monitored values. A value of 84.14% for accuracy was achieved.

The confusion matrix for this testing set (data of 2015) is displayed in Figure 8.11.

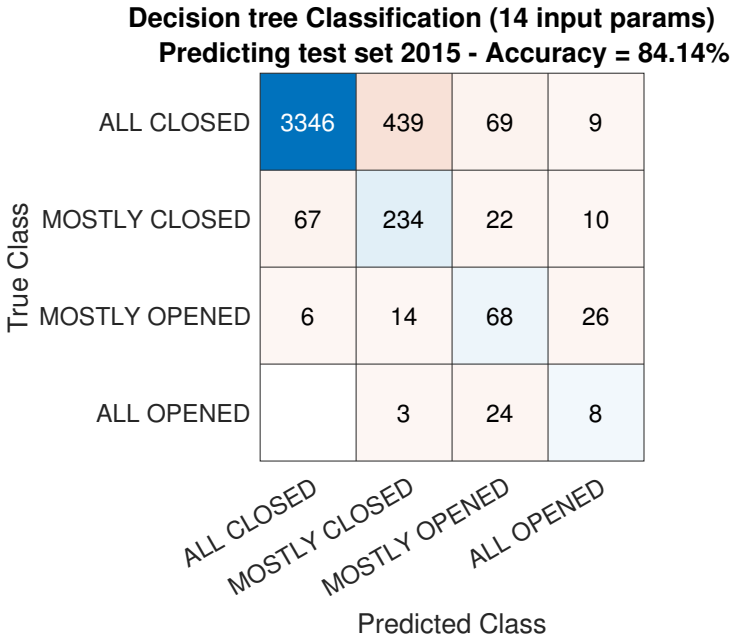


Figure 8.11: Confusion matrix of the Decision Tree classification for the test set including data from January to June, 2015.

Similar to the test set 2014, the true label 'ALL CLOSED' has the highest number of right predictions; in this casen the model successfully labeled 3346 samples and mislabeled 517. The label 'MOSTLY CLOSED' ranks second with 234 accurate samples from 333 samples, and 'MOSTLY OPENED' follows in the third position with 68 correctly classified samples from 114 samples. Specifically, the model can properly identify only 8 samples of the 'ALL OPENED' label and misclassifies up to 24 samples as 'MOSTLY OPENED'.

The more detailed evaluation of these confusion matrices will be discussed in the subsection 8.3.4.

8.3.3.2 Performance of the kNN classifier

Regarding the k-NN classification model, k=3 was obtained after the hyperparameters optimization (see Table 8.3). The achieved accuracies were 99% for training and 92.3% for validation.

The confusion matrix obtained on the test set 2014 is displayed in Figure 8.12. This model obtained a value of overall acurracy of 86.53%. From the figure, the highest

number of wrong classified belongs to the "MOSTLY OPENED" label, while 40 samples are wrongly predicted as "ALL OPENED". Similar to the Decision Tree model, 'ALL CLOSED' label achieved the highest performance, 96.2% samples of this label were correctly predicted (931 correct predictions from a total of 968 samples). The 'MOSTLY CLOSED' label got the second rank with 84.7% correctly predicted samples (344 correct from a total of 406 samples). Finally, the 'MOSTLY OPENED' and 'ALL OPENED' labels rank the last as they have only 66.2% and 56.8% correct predictions, respectively.

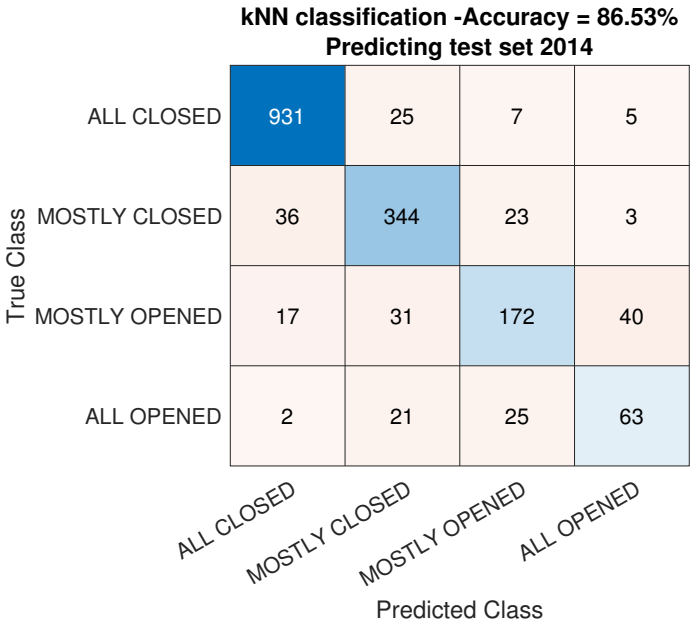


Figure 8.12: Confusion matrix of the k-NN classification for the test set including the remaining 20% of 2014 data.

Similarly, the confusion matrix for the same trained k-NN model applied for the testing set 2015 data is represented in Figure 8.13. Same as the Decision Tree model results for the test set 2015, one can observe that a significant number of "ALL CLOSED" labels are misclassified as "MOSTLY CLOSED" (365 samples - 9.4%). Eventhough, "ALL CLOSED" label still achieved the highest number of correct classifications (88.4% - 3414 correct predictions out of 3863 total samples). The "MOSTLY CLOSED" and "MOSTLY OPENED" achieved the second and third ranks with 46.2% and 32.5%, respectively. The 'ALL OPENED' label, again, got the last position with only 5 correct predictions (14.3%).

8.3.3.3 Performance of the Kernel Approximation classifier

A polynomial kernel function of order 3 has been obtained after the hyperparameter optimization. In comparison with the two other classification models, when using the Kernel Approximation classifier, the training accuracy results were even lower: only 81.7% for training and 80.6% for validation.

The confusion matrices for Kernel Approximation classifications for the years 2014 and 2015 are displayed in Figure 8.14 and Figure 8.15, respectively. While the accuracy was only 79.3% for the test set 2014, this method achieved up to 92.9% for

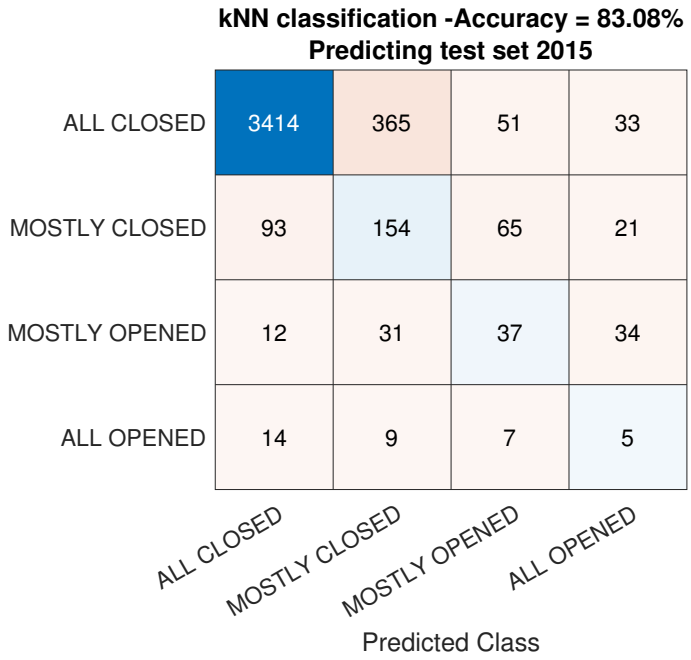


Figure 8.13: Confusion matrix of the k-NN classification for the test set including data from January to June, 2015.

the test set 2015. Similar to the two other models, this model also has a tendency of mislabeling one sample as a 'neighbor label'. According to the Figure 8.14, Kernel Approximation misclassified the 'MOSTLY CLOSED' as 'MOSTLY OPENED' quite a lot (60 samples) and vice-versa (58 samples). For the testing set of 2015, the same mistake was made when 75 samples were mislabeled as 'MOSTLY CLOSED' and up to 82 samples were wrongly classified as 'MOSTLY CLOSED' instead of 'ALL CLOSED'.

It is interesting to note that the Kernel Approximation method has a different rank of correct predictions among labels in comparison with the two other models for the test set 2015. For the test set 2015, the true label 'ALL CLOSED' still has the highest number of right predictions (97.3%), however, the 'MOSTLY OPENED' (42.9%) and 'MOSTLY CLOSED' (36%) labels switched their ranks as second and third, respectively. The 'ALL OPENED' label, again, is in the last position.

Specifically, this method has the highest correct predictions for the label "ALL OPENED" for the test set 2015 with 15 samples on a total of 35.

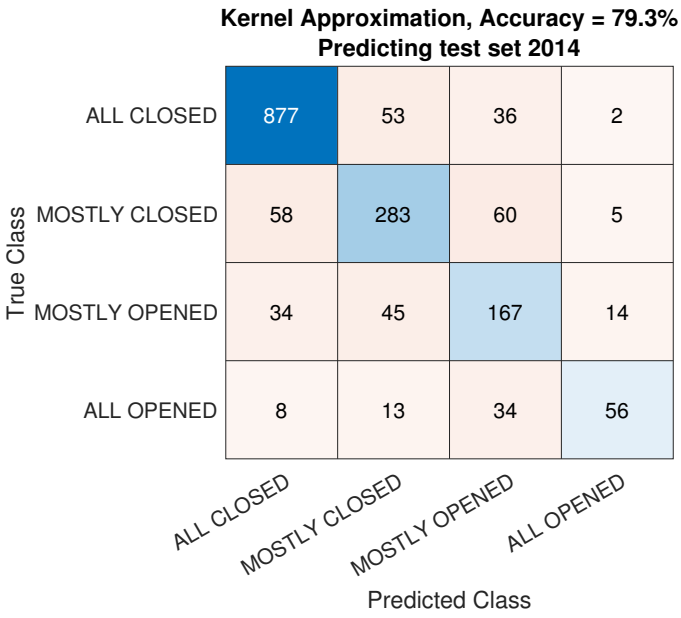


Figure 8.14: Confusion matrix of the Kernel Approximation classification for the test set including the remaining 20% of 2014 data.

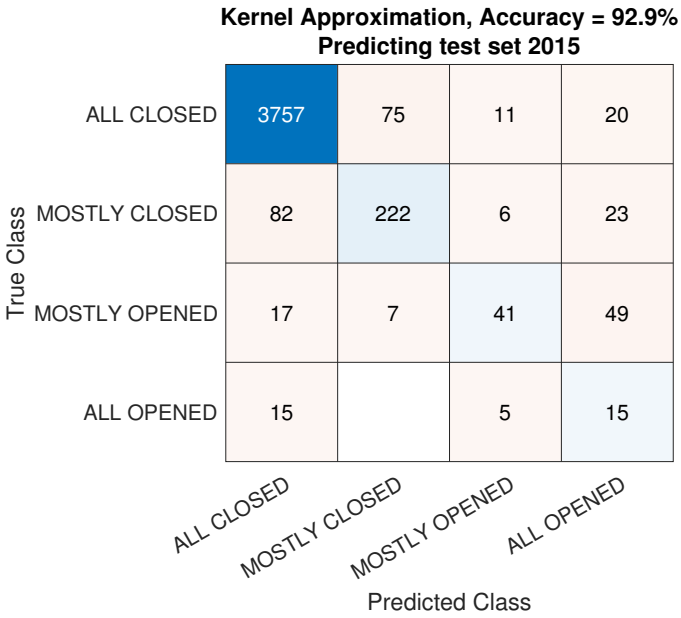


Figure 8.15: Confusion matrix of the Kernel Approximation classification for the test set including data from January to June, 2015.

8.3.3.4 Accuracy statistics for the Decision Tree model

For a deeper analysis of the results, it is necessary to explore the detailed statistics of the accuracy according to the day of the week, the hour of the day, and the month. We decided to present in this subsection only the results obtained for the Decision Tree model because for the other two models, they keep the same global trend.

The statistics for each month of the test set 2014 are showed in the Figure 8.16. The highest accuracies were obtained when predicting the windows state for the winter season (October – February, more than 90%). This is expectable because the windows are mainly closed during this time. Besides, the lower accuracy values correspond to the months of the summer season (June – September, around 70%, except for August 79% - the month of vacation), which mostly contains the labels 'ALL OPENED' and 'MOSTLY OPENED'.

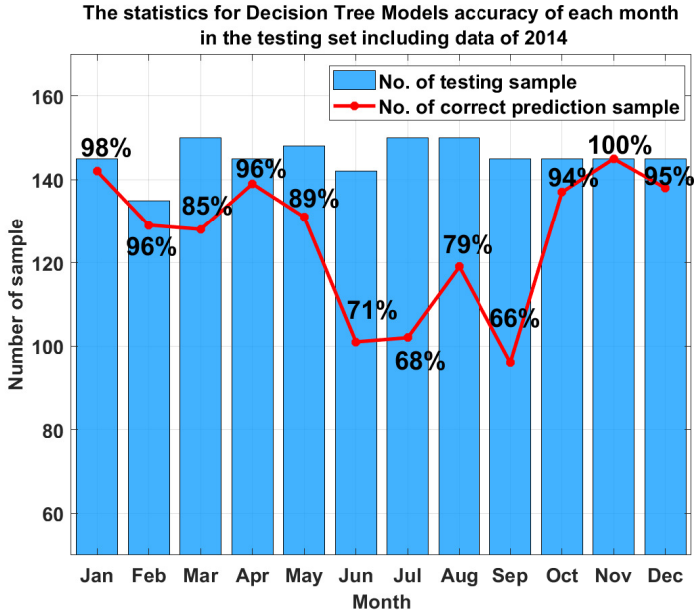


Figure 8.16: The statistics for Decision Tree Models accuracy of each month in the testing set including data of 2014.

Similarly, Figure 8.17 presents the statistics of the same classification model according to each month for the testing set of 2015. Once again, the winter period (January – February) has achieved the highest value of accuracy (more than 92%). Especially, the month of May obtained a quite high accuracy with up to 95%; this could be explained by the fact that May is the month with many holidays, thus not many people were in the office, and the windows were mainly closed. The lowest accuracy is obtained for the month of June (summer season) with value of only 59%.

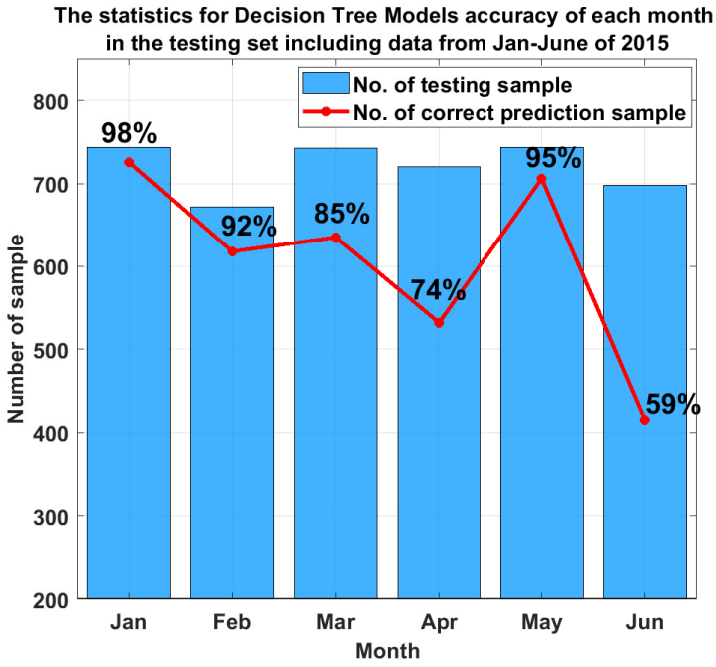
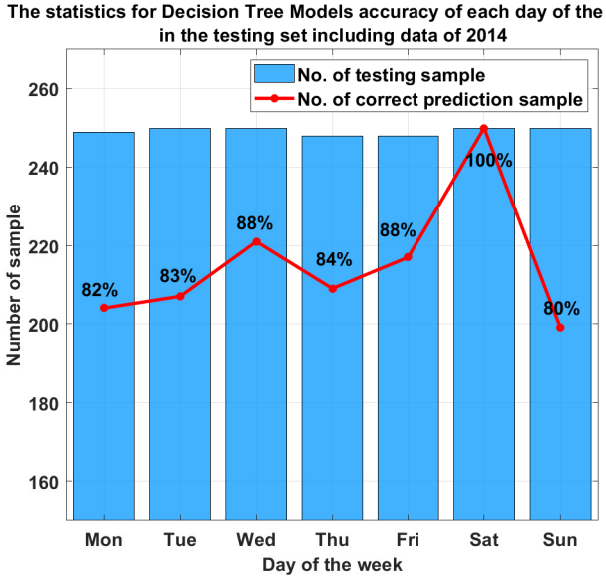
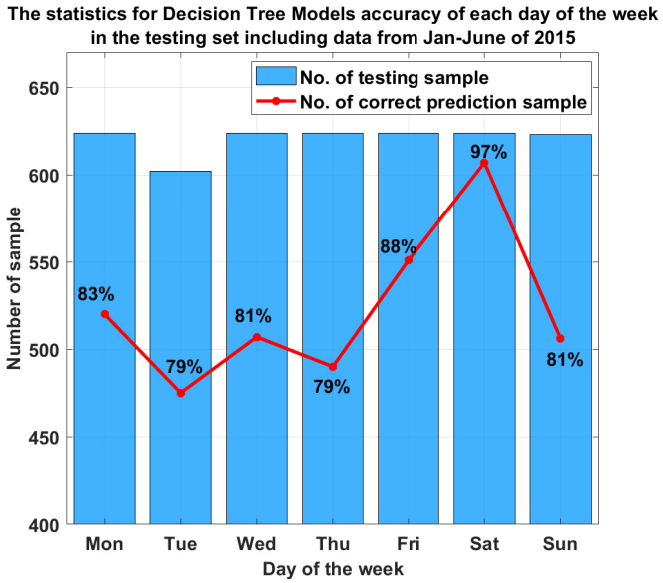


Figure 8.17: The statistics for Decision Tree Models accuracy of each month in the testing set including data from January to June, 2015.

Regarding the statistics according to the day of the week, Figure 8.18 shows the detailed accuracy for each day of the week for the two sets of testing data. One can notice that the overall trend of the accuracy is similar for the two datasets: the lower accuracy values were obtained for the working days (Monday to Friday) while the highest ones were achieved when predicting the windows state for Saturday (more than 97%). Interestingly, the lowest values of accuracy (80%) was obtained for Sunday for the data of 2014; meanwhile, for the data of 2015, the lowest values of accuracy were obtained for Tuesday and for Thursday (only 79%).



(a)



(b)

Figure 8.18: The statistics for the Decision Tree accuracy according to each day of the week for the testing set including: (a) data of 2014 and (b) data from January to June, 2015.

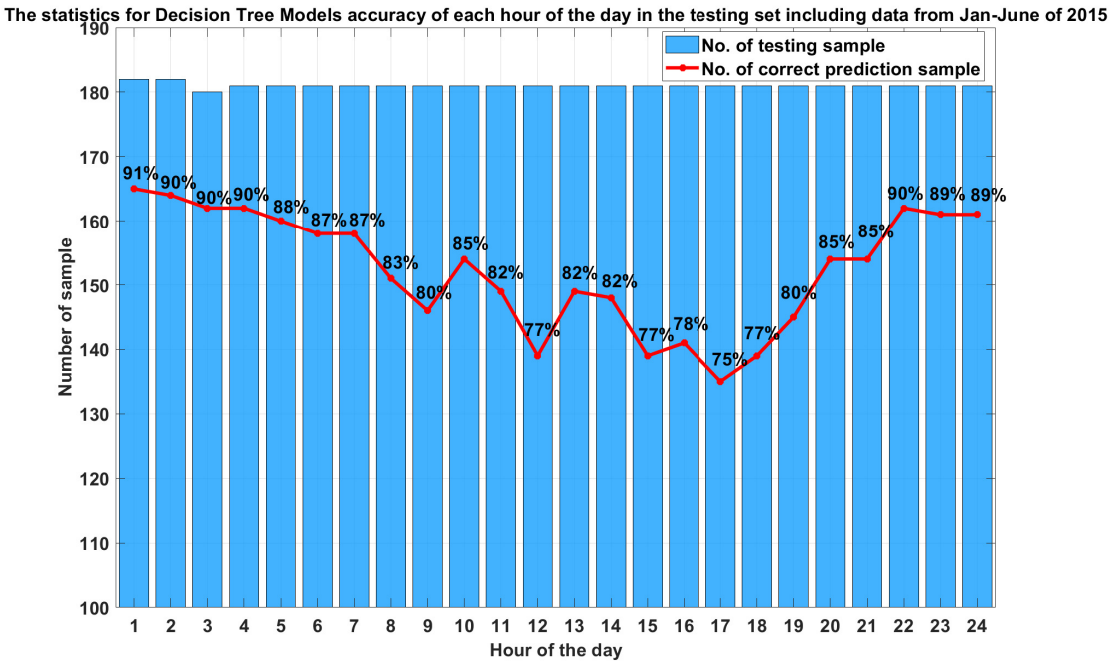


Figure 8.19: The statistics for Decision Tree Models accuracy for each month in the testing set including data from Jan-June of 2015.

According to the Figure 8.19, the highest accuracies are achieved when predicting the windows state for night-time periods (8 p.m. – 7 a.m., more than 88% excepting at 11 p.m. when maybe the guard round took place). In contrast, the lowest accuracy values correspond to the lunch-time periods (12 a.m.– 2 p.m. around 76%) and the 'office leaving' hour (5 p.m. - 73%). In all these periods, there are more changes in the status of the windows and they mostly contain the labels 'ALL OPENED' and 'MOSTLY OPENED'.

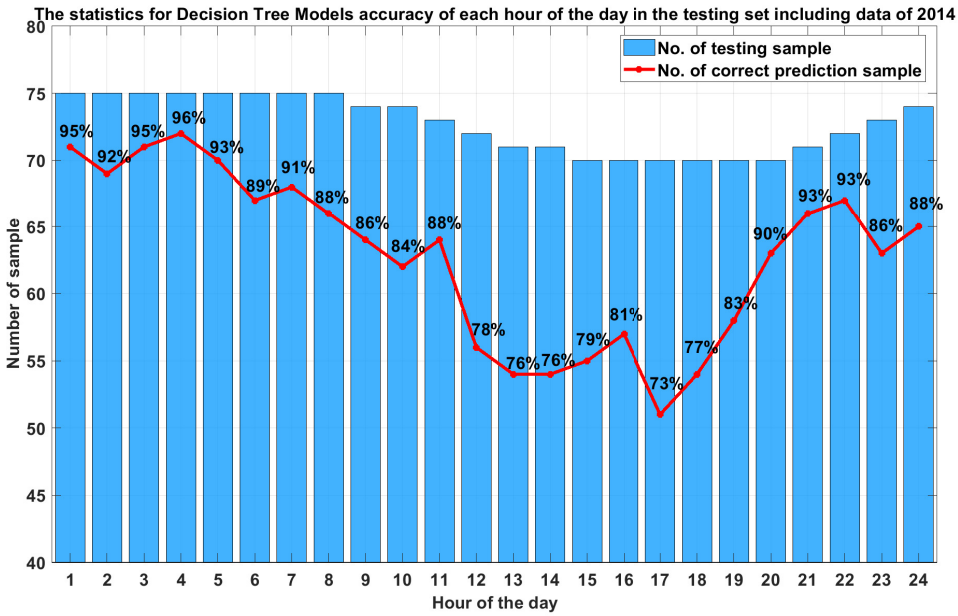


Figure 8.20: The statistics for Decision Tree Models accuracy according to each hour of the day for the testing set of data from January to June, 2015.

Similar to the testing set of 2014, the highest values of accuracy (around 88%) were obtained during the night-time periods (10 p.m. – 7 a.m.). Meanwhile, the lowest accuracy values correspond to day-time periods (9 a.m.–6 p.m.). According to the hour of the day, the prediction accuracy at 5 p.m. is still the lowest (75%), probably because it corresponds to the "office leaving" hour. Some people tend to close the windows before leaving while others leave them opened.

8.3.4 Evaluation

Firstly, some definitions of ML terminology for evaluating the model's performance are introduced in the Figure 8.21. A true positive outcome is obtained when the model correctly predicts the positive class (predicts 1 when the actual class is 1). Similarly, a true negative is a result in which the model correctly predicts the negative class (predicts 0 when the actual class is 0). In contrast, a false positive is an outcome where the model incorrectly predicts the positive class (predicts 1 when the actual class is 0). Finally, a false negative is an outcome in which the model predicts the negative class incorrectly (predicts 0 when the actual class is 1).

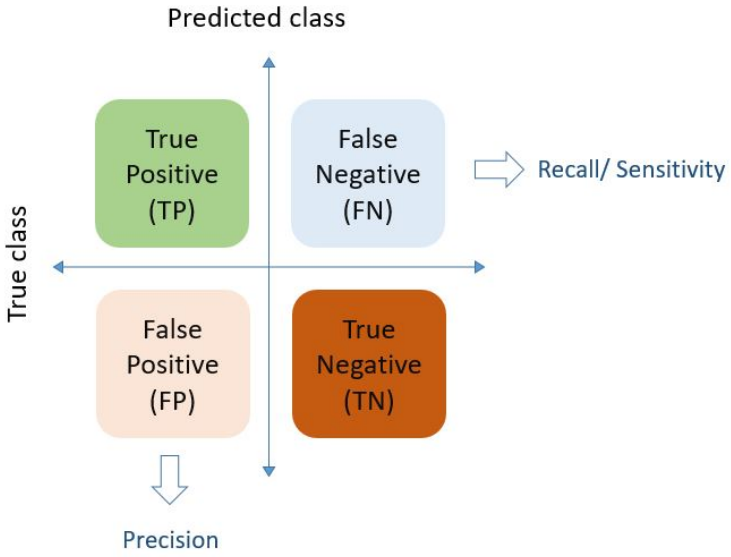


Figure 8.21: Definition of ML terms for evaluation the model's performance.

The quality of a classifier can be evaluated by equations 8.7 - 8.10, which introduce four perfomance indicators based on the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

While the Accuracy can be used to evaluate the model's percentage of well-classified data, Recall and Precision are two other important indicators to evaluate the performance of classification. The value of accuracy is calculated as given in the equation 8.7:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{8.7}$$

Recall is also denoted as sensitivity and it is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$Sensitivity(Recall) = TP/(TP + FN) \tag{8.8}$$

Precision is also known as positive predictive value and has been defined as the fraction of relevant instances among the retrieved instances.

$$Precision(F_{rate}) = TP/(TP + FP) \tag{8.9}$$

Equations 8.8 and 8.9 were then used to calculate both Recall and Precision in this study. The F1-coefficient has been used for evaluating the model's predictive performance by combining the results from both Recall and Precision:

$$F1 = 2(Recall)(Precision)/(Recall + Precision) \tag{8.10}$$

Table 8.5 summarises the general accuracy values of the three methods: Decision Tree, k-NN and Kernel approximation, when predicting the test set 2014 and the test set 2015. When using Decision Tree and k-NN almost the same performance

has been achieved for the two testing sets, an accuracy value around 84%. A significantly higher accuracy when predicting the data of 2015 was obtained by means of the Kernel Approximation (around 93%). The fact that Kernel Approximation model's accuracy when predicting the test set 2014 is lower than predicting the test set 2015 (79% versus 93%) can be explained by the particular distribution of labels in 2015 and by the high perfomance of this method for separation in the case of nonlinear problems.

Table 8.5: Overall accuracy for the three models

Algorithm	Test set 2014	Test set 2015
Decision Tree	86.36%	84.14%
k-Nearest Neighbor	86.53%	83.08%
Kernel Approximation	79.30%	92.90%

Figure 8.22 presents the calculated Recall (Sensitivity) values for each state of window opening. For the test set 2014, one can notice that the three models give quite similar results, slightly lower for the Kernel Approximation method. While the highest Recall value is obtained when predicting the 'ALL CLOSED' state of the group of windows ($\approx$ 90%), the lowest value corresponds to the 'ALL OPENED' label ($\approx$ 60%). Similarly, for the test set 2015, the highest Recall value is still obtained when predicting the 'ALL CLOSED' label (90%) while the lowest belongs to the 'ALL OPENED' label (excepting the Recall value obtained by the Kernel Approximation method for test set 2015, where the lowest value corresponds to the 'MOSTLY OPENED' label).

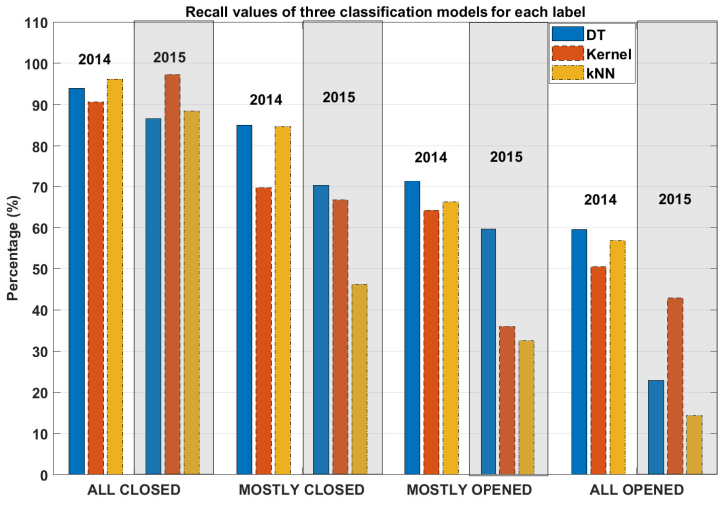


Figure 8.22: Recall values of the three classification models: Decision Tree, k-NN and Kernel approximation. The Recal values corresponding to the testing data from January to June 2015 are displayed on a grey background.

Figure 8.23 and Figure 8.24 present the Precision values and the F1 scores, respectively. The same situation is obtained for both testing sets. While the highest

values are obtained when predicting the 'ALL CLOSED' state, the lowest values correspond to the 'ALL OPENED' label (excepting the Precision value obtained by the Kernel Approximation method for the test set 2014, where the lowest value correspond to the 'MOSTLY OPENED' label). For the test set 2015, regarding the Precision values, an even lower value of 5.4% is observed for the 'ALL OPENED' label, when using the k-NN model. The reason for which the model's accuracy when predicting the 'ALL OPENED' label was much lower than for the 'ALL CLOSED' label is the particular distribution of labels during the two years. The windows are mainly 'ALL CLOSED' and this label is "well learned" by the model. Window opening models are often biased towards the over-represented class where windows remained closed (Markovic *et al.*, 2018).

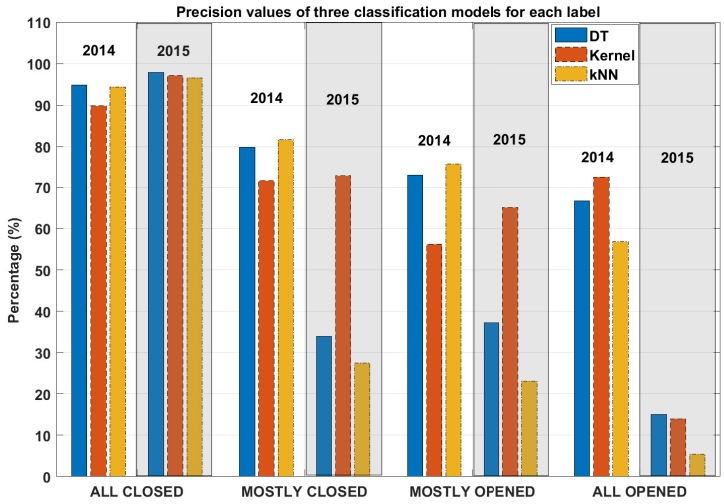


Figure 8.23: Precision values of the three classification models: Decision Tree, k-NN and Kernel approximation. The Precision values corresponding to the testing data from January to June 2015 are displayed on a grey background.

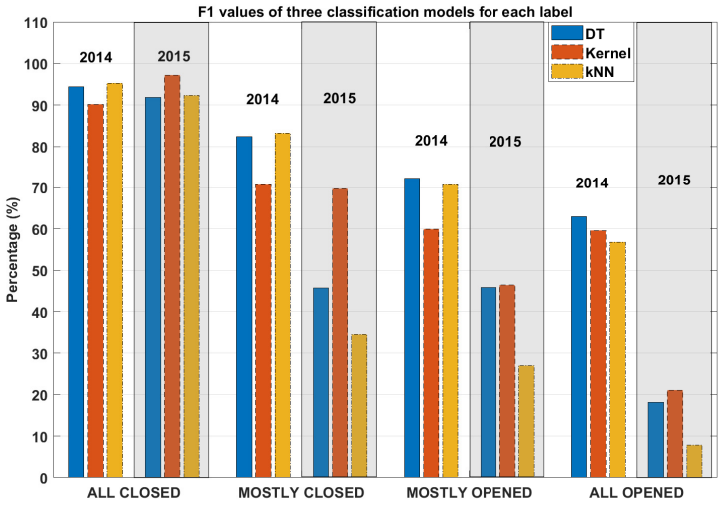


Figure 8.24: F1 values of the three classification models: Decision Tree, k-NN and Kernel approximation. The F1 values corresponding to the testing data from January to June 2015 are displayed on a grey background.

It is interesting to note that, the Accuracy gives an overall result without information about a specific label. Meanwhile, in the case of the Recall and Precision indicators, a detailed accuracy is obtained for each label, in different perspectives: Precision - How many predicted samples of this label are correct? Recall - How many samples of this label are correctly predicted? From the Figure 8.22, one can observe the significant differences for the Recall values when using the Kernel Approximation for 'ALL OPENED' label (significantly more performant) and in the case of the Decision Tree for the 'MOSTLY OPENED' label, both of them on the test set 2015. Similarly, Figure 8.23 reveals the high differences in Precision values for the 'MOSTLY OPENED' and 'MOSTLY CLOSED' label for the test set 2015 when using the Kernel Approximation. However, when the F1 values were calculated, these differences were smaller.

Overall, the Decision Tree method appears to be the best classification model, with the best balance of Recall, Precision and F1 values regarding the four labels. Kernel Approximation occasionally achieved the highest evaluation values (particularly for the test set 2015 for 'ALL CLOSED' and 'ALL OPENED' labels). This can be explained by its high perfomance in separation in the case of nonlinear problems. However, the overall accuracy for the test set 2014 of this method is slightly lower in comparison with the two other methods. In addition, Decision Tree also provides the list of classification rules (export in .txt file), which can easily be used to apply for new data. Regarding the k-NN model, the low values of these evaluation indicators could be explained by the fact that this method has been applied on categorical data for all the parameters, by contrast to the other methods, which allow the both types of inputs (numerical and categorical). This decoding operation probably leads to a loss of information.

8.3.5 Conclusion and Dicussion

In conclusion, in this study, we have obtained three ML classification models to predict the opening state for a group of windows in an open-plan office. To select the appropriate set of features, the ACF values and predictor importance estimates were calculated. In our case, the most pertinent inputs were: the previous 24^{th} hour state of the windows (which can be related to the personal preferences of the occupants), the day of the week, the month, the hour of the day (which can be related to the occupancy and the personal preferences) and the previous 24^{th} hour of the prevailing mean outdoor air temperature (outdoor environment condition). The models were then established by using these important parameters completed with the 'previous 24^{th} hour' of the following variables: the wind direction, entrance door status, indoor CO_2 and particle matter (PM2.5 and PM10) concentrations, as well as both indoor and outdoor temperatures and specific humidity. Validation tests have been used to compare the outputs of the models and the measured windows states obtained in the years 2014 and 2015 in the open-plan office. According to the different evaluation indicators, the results show that all the three models perform well with the testing sets.

In the future, we can improve the over-represented 'ALL CLOSED' label by resampling in order to have an unbiased data set or by providing different weights for each label to penalize misclassification. In addition, with an algorithm that combines multiple trees and control for bias or variance, like Random Forests (Ho, 2016) or Gradient boosted trees (Natekin and Knoll, 2013), the Decision Tree model could have a better performance. For the k-NN model, an efficient method to deal with both the numerical and categorical data in order to avoid the loss of information needs to be further investigated. Furthemore, the high performance of the Kernel Approximation approach - a good nonlinear separator, is also noteworthy.

We could then use one of the three developed models as a standalone, or as a part of a real-time IAQ monitoring system, in order to optimize the action to be taken to reduce the exposure of the occupants.

Conclusions and Future Works

After the first chapter introducing some generalities on IAQ and the second one presenting the case study and the database, the main results, contributions and some corresponding possible extensions of the thesis are summarized as follows.

Conclusions

The global objective of our research was to develop a system providing information about pollutant sources and events influencing the temporal variation of indoor air pollutants, helping to optimize the action to be taken to reduce the exposure of the occupants. The study conducted in this thesis had a dual purpose. The first objective concerns the determination of some factors to understand and analyze the structures of temporal variability of particle matter concentrations in the indoor air of the studied open-plan, the variability of their sources, as well as the source contributions. The second part of this thesis aimed to develop a predictive model for IAQ management, in particular, a model to predict the opening state of a group of windows in the same open-plan office.

For the first objective, concerning the sources identification, a tensor decomposition method named PARAFAC has been chosen among the different BSS methods as it can produce an unique output by contrast with the other ones and it allows also to easily expand the complexity of the input data by using multi-dimensional structures. In addition, the particularity of this method (PARAFAC) is that indoor and outdoor particles of given sizes can be considered in parallel layers and not as different variables of the same layer (matrix-based methods such as PMF), corresponding better to the reality. By using different combinations of different parameters, structures were generated and they allowed to determine the relative contribution of the office occupants and their activities and of the outdoor sources to the indoor concentration of particles, which is a topic of utmost interest in Indoor Air Quality studies nowadays.

For the purpose of prediction of the opening state for a group of windows in an open-plan office, three machine learning classification methods permitted to obtain good results. Validation tests have been used to compare the outputs of the models and the measured windows states obtained during 18 months in the open-plan office. According to the different evaluation indicators, the three models performed well with the testing sets. These models can be considered as one of the first models to be included in more complex exploration on prediction models of IAQ in order to improve it; meanwhile it is already possible to use them as a new function of a sensor or for the anticipation of opening/ventilation management.

- **Source identification of indoor pollutants**

  This part of the thesis focuses on the identification of air pollutant sources of an open-plan office and assessing their relative contribution. The target pollutant studied in this work was the particulate matter, because it causes many health effects (and because of its complicated chemical composition makes it more difficult to estimate the concentration and contribution of their sources). Our study chalenges to reveal the underlying factors that affect the temporal variation of particle matter.

  An important contribution of this thesis is the use of BSS methods for searching for pollution sources, where the original information does not concern a chemical speciation, but a size-resolved information about the airborne particles. Using the PARAFAC method has specific advantages like the uniqueness of the solution and the possibility to create multi-dimensional data strutures. For example, the same fractions of particulate matter indoors and outdoors can be considered in parallel layers. Different strutures of variables consisting in 3D or 4D array layers could be built and factorized in order to obtain source profiles and their contributions.

  The major sources revealed by this decomposition are: (*i*) human presence and the related activities indoors, contributing to about 25-30% of the global level of indoor particles and concerning mostly coarse particles (4.5-20 μm); (*ii*) outdoor air pollution coming by infiltration natural or mechanical ventilation, contributing to about 40% of the global level of indoor air particles, and concerning mostly fine particles (0.35-1.3 μm), which are the product of different combustion processes like traffic, industry, heating, *etc*. In order to identify these factors, complementary statistics analysis has been performed (*e.g.* PCA) or signal treatment (*e.g.* ACF)

  One of the drawback of PARAFAC is that other sources probably with a minor effect, could not be retrieved, as in the case of the 2D-array methods such as PMF, NMF, PCA, *etc*. A combined study using on the one hand tensors decomposition and on the other hand matrix factorization could be conducted in order to reveal the most robust, major sources (given by PARAFAC) and to complete with additional, minor sources given by methods such as PMF.

  Although climatic parameters indoors and outdoors were available, the results obtained including them in the analysis seemed to be, for the moment, difficult to be interpreted. Further reflection should be considered in order to be able to take into account the potential of the information given by all the other factors, especially climatic.

- **Window prediction for the opening state of the windows**

  The opening state of the windows has an important influence on IAQ, as it can modify the air exchange rate and as such the transfer between indoor and outdoor environments. In this second part of the thesis, we tried to model the windows opening state in the same real open-plan office with five windows. The three ML models: Decision Tree, k-NN and Kernel Approximation have been implemented.

The ACF values and predictor importance estimates were calculated to select the appropriate set of input features. In our case, the most pertinent inputs were: the previous 24^{th} hour state of the windows (which can be related to the personal preferences of the occupants), the day of the week, the month, the hour of the day (which can be related to the occupancy and the personal preferences) and the previous 24^{th} hour of the prevailing mean outdoor air temperature (outdoor environment condition). The models were then established by using these important parameters completed with the 'previous 24^{th} hour' of the following variables: the wind direction, entrance door status, indoor CO_2 and particle matter (PM2.5 and PM10) concentrations, as well as both indoor and outdoor temperatures and specific humidity. Validation tests have been used to compare the outputs of the models and the measured windows states during 18 months in the open-plan office. According to the different evaluation indicators, the results show that all three models perform well with the testing sets.

Future Works

There are still numerous unresolved difficulties and undiscovered leads that raise fascinating research prospects and possible enhancements to source identification techniques, their contributions, and, especially, our prediction methods. Indeed, while this thesis aims to provide basic improvements in our knowledge of the variability of IAQ, it is far from presenting a comprehensive solution to all difficulties related to the indoor environment.

In this section, we provide some perspectives on the work have been done during this thesis.

- **The questions concerning source identification**.

  The source separation techniques (PARAFAC) discussed in this manuscript have provided enriching feedback. This type of model would make it possible to answer the fundamental question about multi-exposure to microenvironmental contaminants of indoor air quality. From this study, several perspectives on this work can be envisaged. On the theoretical level, an open problem is that of estimating the number of sources. So, a fundamental question is to study if the estimate of the source number is a "demonstrable" problem; if yes determine an algorithm to estimate it.

  Regarding the amount of information and the complexity of the data, the active instrumentation of the open-plan office during about three years of measurement generated a very considerable flow of information, thus exceeding 5 million (or even more) samples. This information continues to flow into existing databases and is now very bulky. The question of the time step plays a very important role, both at the theoretical and practical levels, then "how far to go in the temporal scale"?

  This question may not seem to be of prime importance. However, it remains, in our eyes very intriguing, because it is the choice of the class of models to use in the forecasting stage. So, the time step and the "mass" of available data raise a question of statistical methodology.

So, do we just have to borrow the methods already applied in the other domains and transpose them to the current databases, or it is necessary to build the specific models according to the characteristics of the data related to the IAQ?

In terms of the processing of real data, an interesting study to be carried out concerns the processing of time series resulting from the analysis of several environments, so a spatial dimension is added to both temporal and individual dimensions. The question of data fusion is immediate.

Finally, it would be interesting to create an automation process for sources separation and identifications. This can be used for a real-time monitoring system in the future.

- **Prediction of forecast models**.

  In the future, we can improve the over-represented 'ALL CLOSED' label by resampling it in order to have an unbiased data set or by providing different weights for each label to penalize misclassification. In addition, with an algorithm that combines multiple trees and control for bias or variance, like Random Forests or Gradient boosted trees, the Decision Tree model could have a better performance. For the k-NN model, an efficient method to deal with both the numerical and categorical data in order to avoid the loss of information needs to be further investigated. Furthermore, the high performance of Kernel Approximation approach - a good nonlinear separator, is also noteworthy.

  Regarding the study case, it would be interesting to explore other indoor environments (dwellings, schools, private rooms, etc.). In addition, a deepen knowledge about the temporal structure of data is need in order to help choosing potential input parameters. For the data validation and predicting, an automation process should be studied in the future.

  We could then use one of the three developed models as a standalone, or as a part of a real-time IAQ monitoring system, in order to optimize the action to be taken to reduce the exposure of the occupants.

Publications

JOURNAL

1. T.H Nguyen, A. Ionescu, E. Gehin, O. Ramalho, **"Predicting the Opening State of a Group of Windows in an Open-Plan Office by using Machine Learning Models,"** (*Building and Environment*, 2022(225))

   https://doi.org/10.1016/j.buildenv.2022.109636

2. T.H Nguyen, A. Ionescu, O. Ramalho, E. Gehin, **"Predicting the Window Opening State in an Office to Improve Indoor Air Quality,"** (*Engineering Proceedings*, 2021(5))

   https://doi.org/10.3390/engproc2021005024

INTERNATIONAL CONFERENCES (oral communications)

1. T.H Nguyen, A. Ionescu, O. Ramalho, E. Gehin, **"Application of PARAllel FActor (PARAFAC) Analysis for Indoor Air Quality of an Open-space Office, France,"** (*Healthy, Energy Efficiency and Intelligent Building Systems HEIBS2021, 2021*, Paris, France)

2. T.H Nguyen, A. Ionescu, O. Ramalho, E. Gehin, **"Predicting the Window Opening State in an Office to Improve Indoor Air Quality,"** (*International conference on Time Series and Forecasting ITISE 2021*, Spain)

3. T.H Nguyen, A. Ionescu, O. Ramalho, E. Gehin, **"Identification of Indoor Aerosol Sources in an Open-Plan Office using Factor Analysis,"** (*European Aerosol Conference EAC2021*, virtual conference)

NATIONAL CONFERENCE (oral communication)

1. T.H Nguyen, A. Ionescu, O. Ramalho, M. Mathis, E. Gehin, **"Parallel factor (PARAFAC) analysis for interpretation of variation of indoor air pollutants in an open-plan office,"** (*Congrès Français sur les Aérosols (CFA2020)*, Paris, France)

APPENDIX

PN time profiles

This appendix present different time profiles for each size fraction monitored indoors. Figure A.1 illustrates the wekkly profiles for the 15 fractions measured indoors during 2014: the yearly averaged for each day of the week (and its standard deviation).

Similarly, Figure A.2 illustrates the hourly profiles for the 15 fractions measured indoors during 2014: the yearly averaged for each hour of the day (and its standard deviation).

The monthly profiles for the 15 fractions monitored indoors during 2014 are presented in the Figure A.3.

In the Figure A.4 - Figure A.6, the same profiles are plotted, but making the difference when the office is occupied or non-occupied, while the Figure A.7 - Figure A.9 correspond to two cases concerning the windows status.

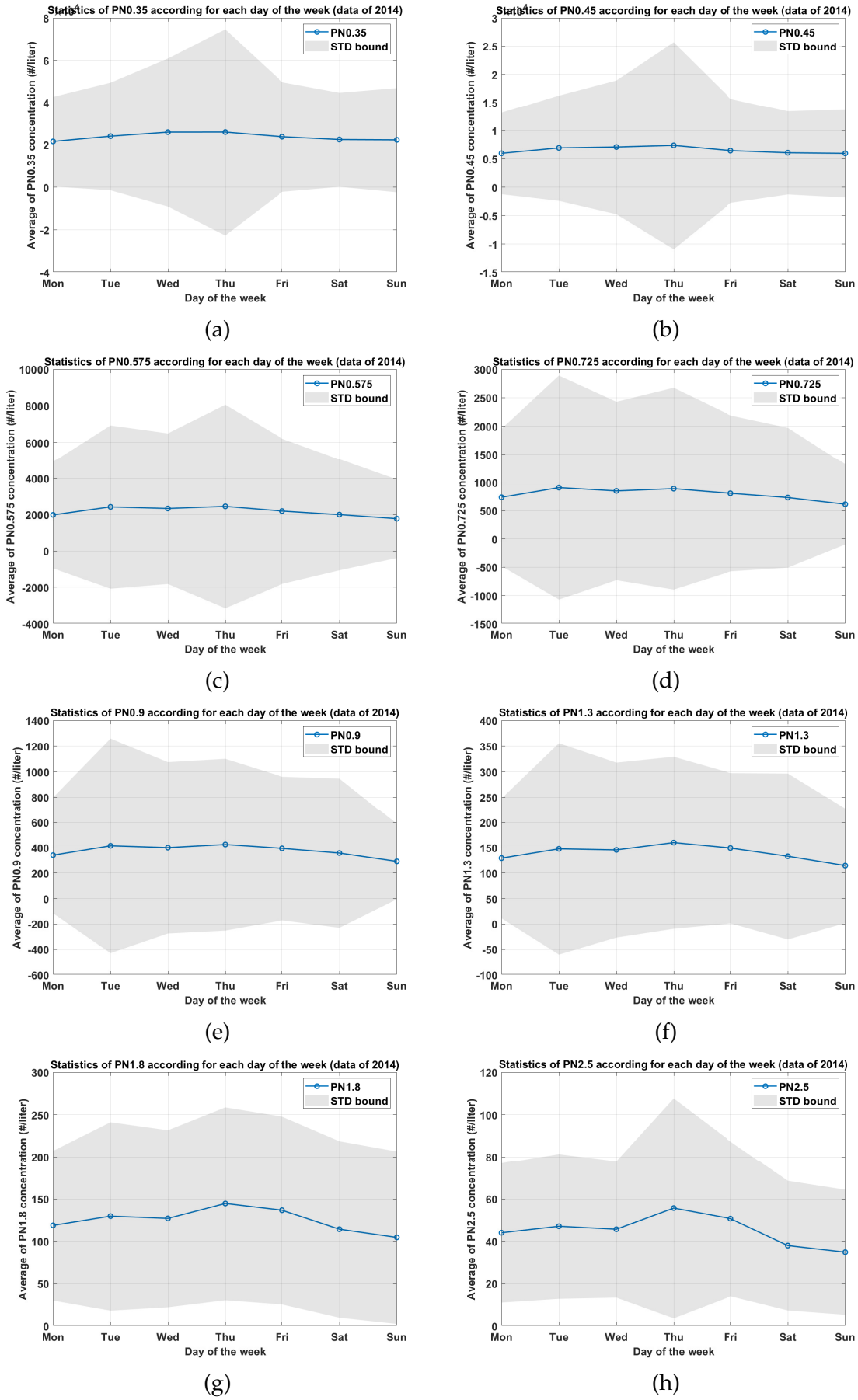
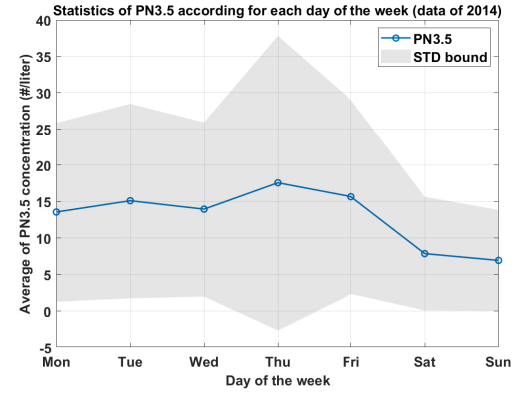
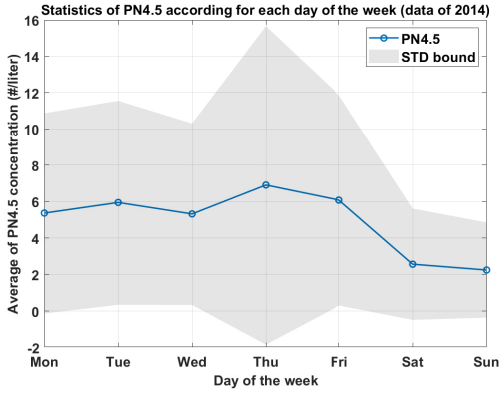


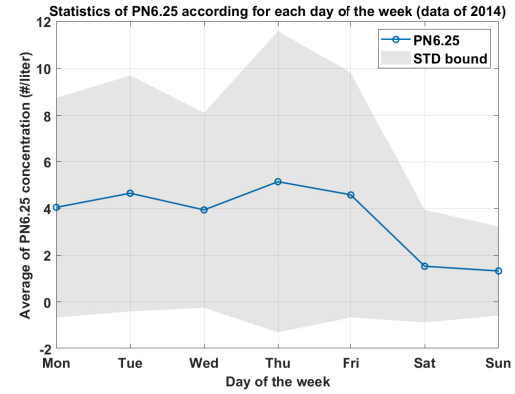
Figure A.1: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week (to be continued).



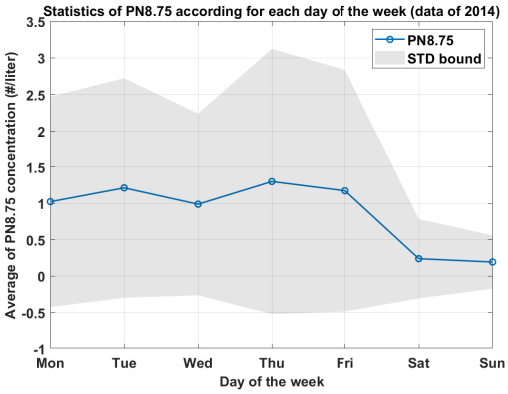
(i)



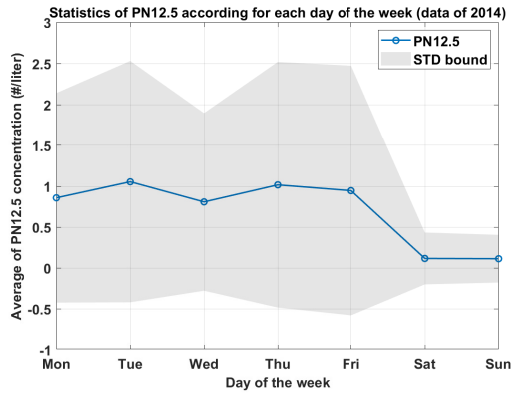
(j)



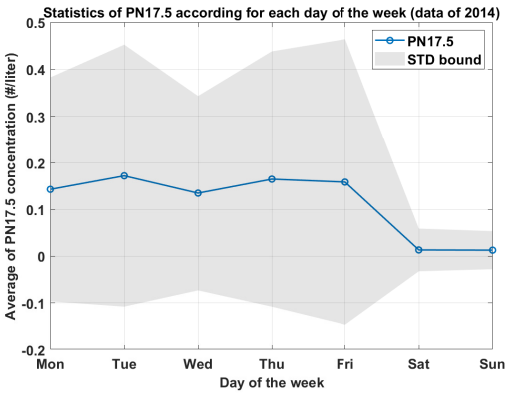
(k)



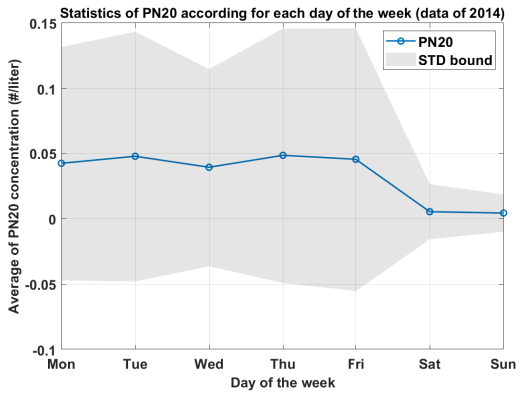
(l)



(m)



(n)



(o)

Figure A.1: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week (continued).

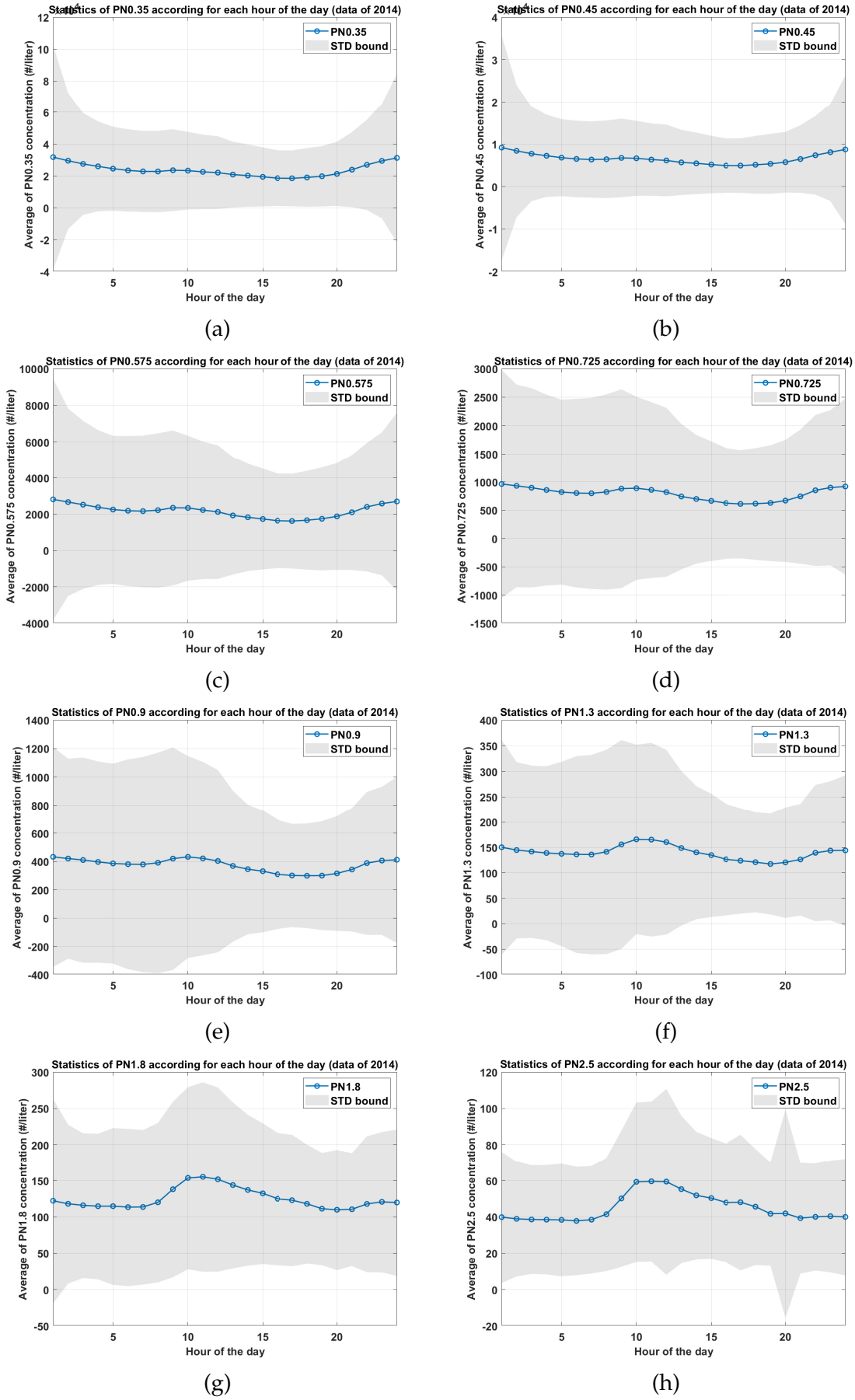
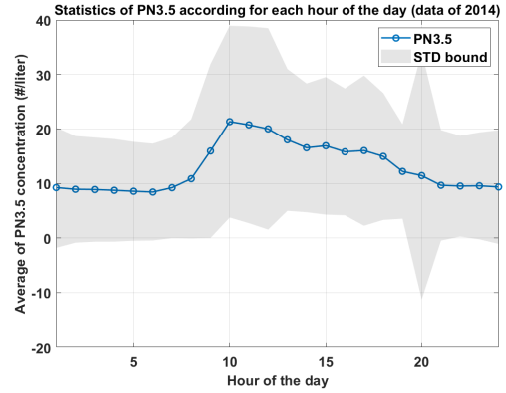
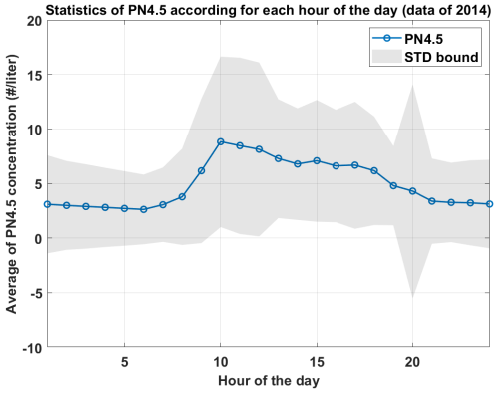


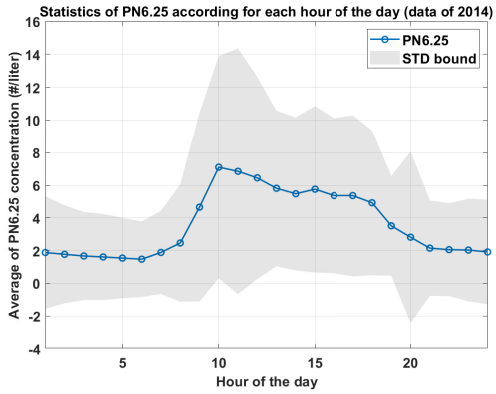
Figure A.2: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day (to be continued).



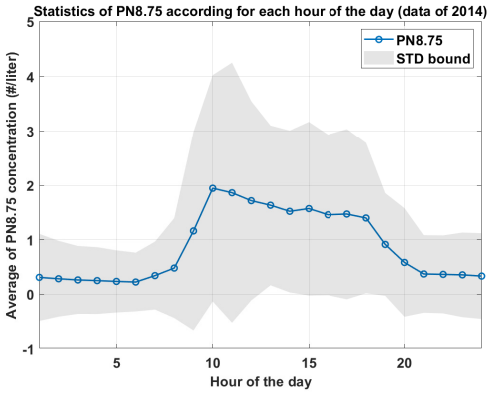
(i)



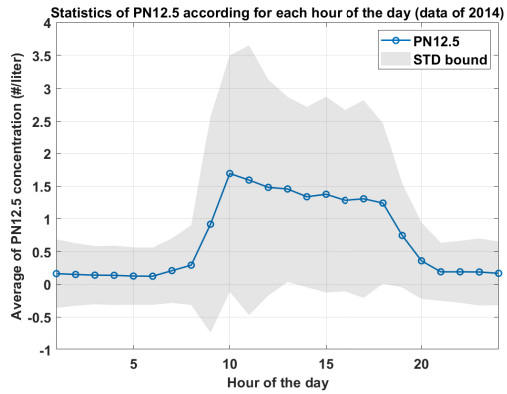
(j)



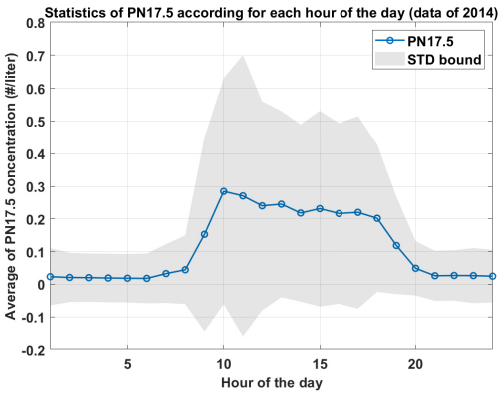
(k)



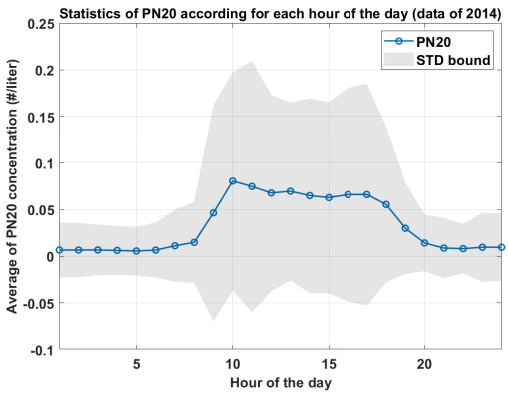
(l)



(m)



(n)



(o)

Figure A.2: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day (continued).

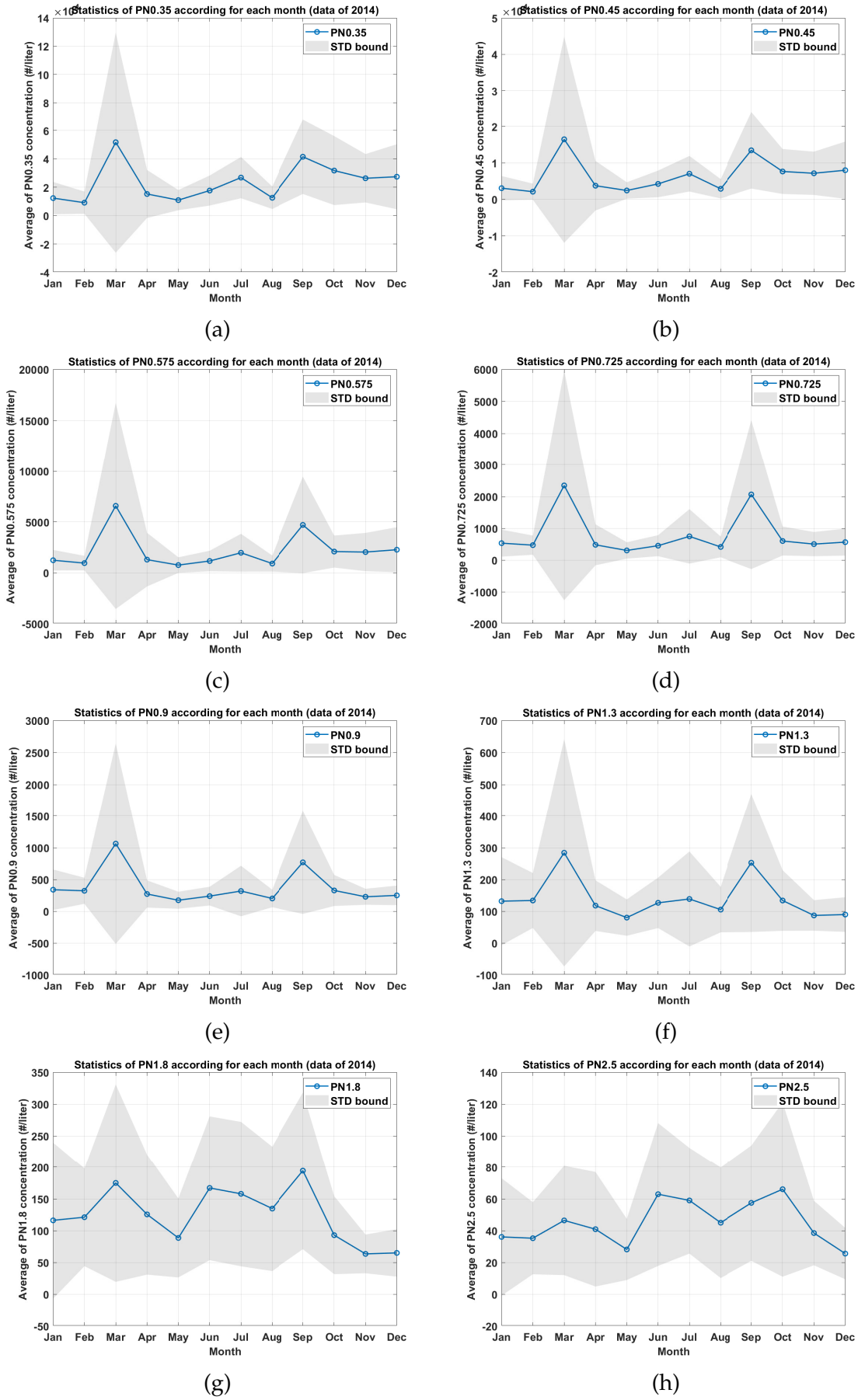
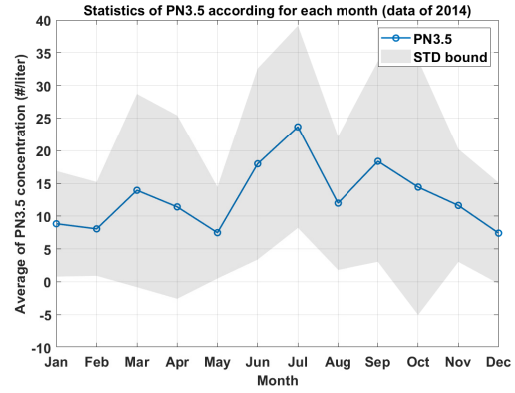
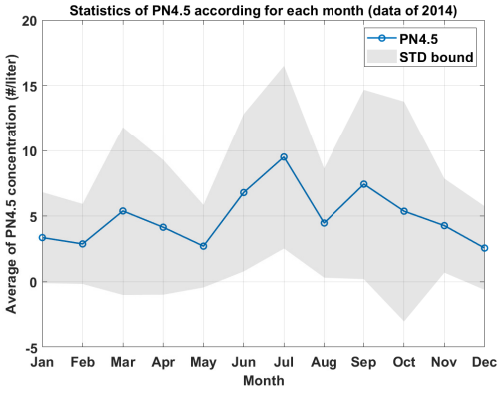


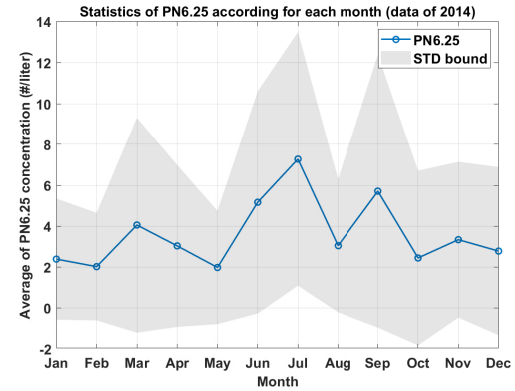
Figure A.3: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month (to be continued).



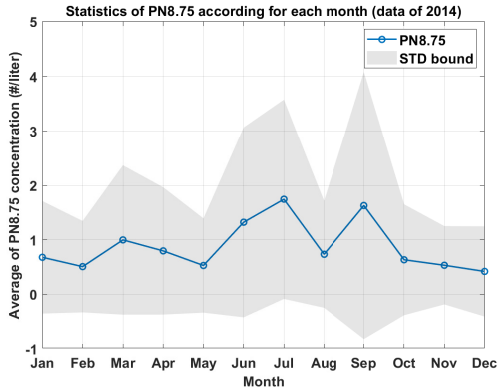
(i)



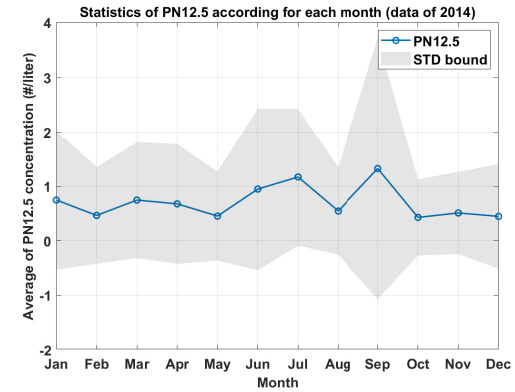
(j)



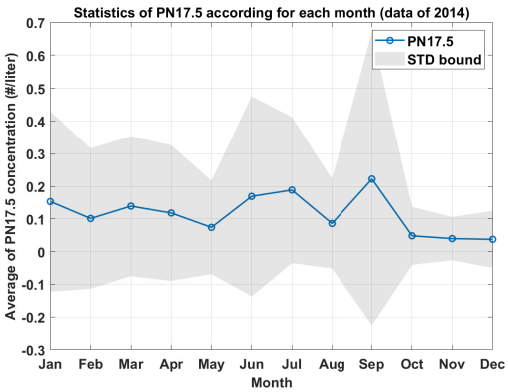
(k)



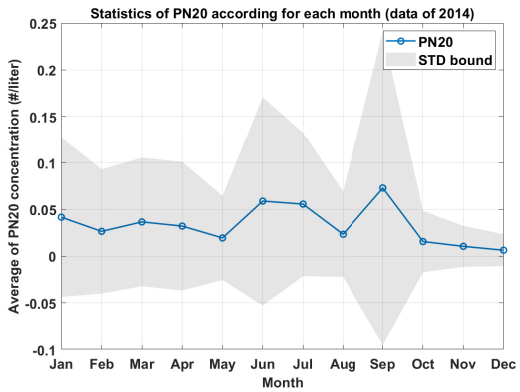
(l)



(m)



(n)



(o)

Figure A.3: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month (continued).

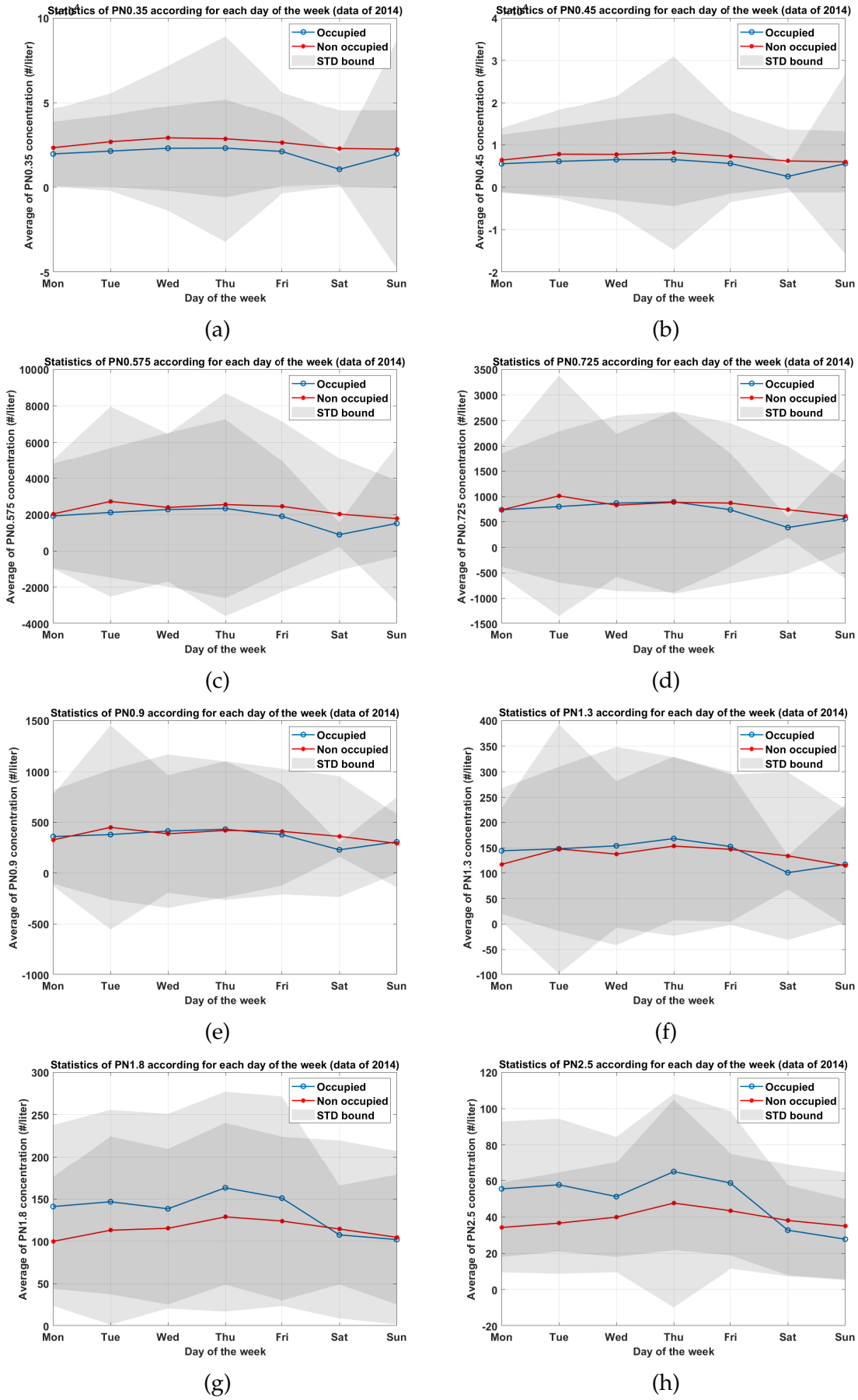
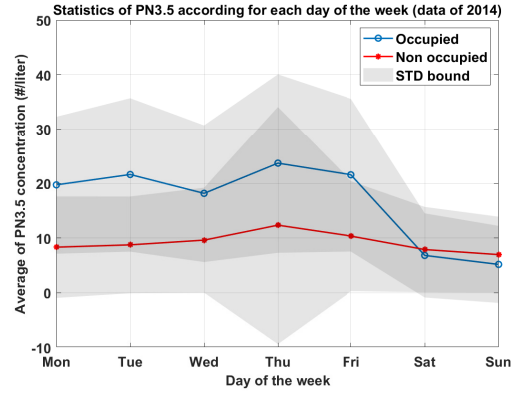
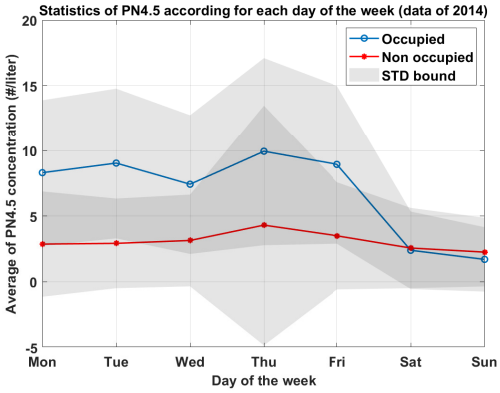


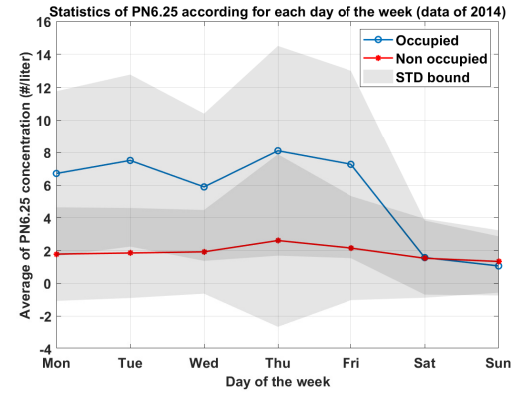
Figure A.4: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week in two cases: the office is occupied or non-occupied (to be continue).



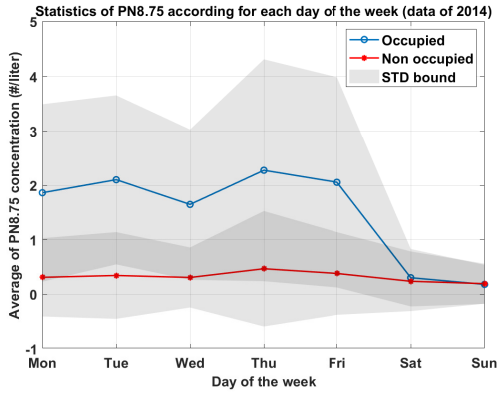
(i)



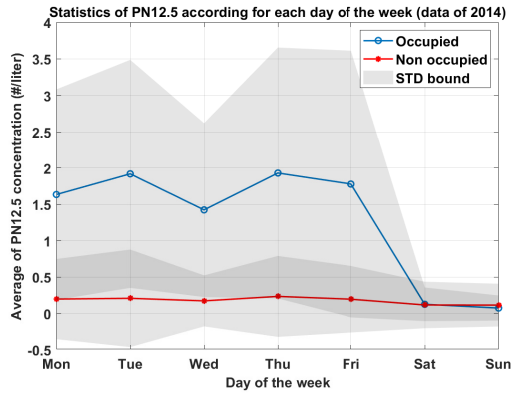
(j)



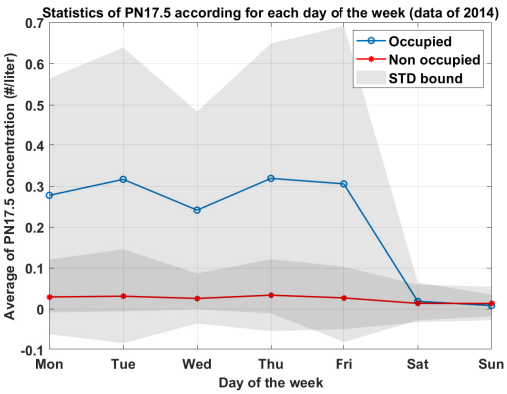
(k)



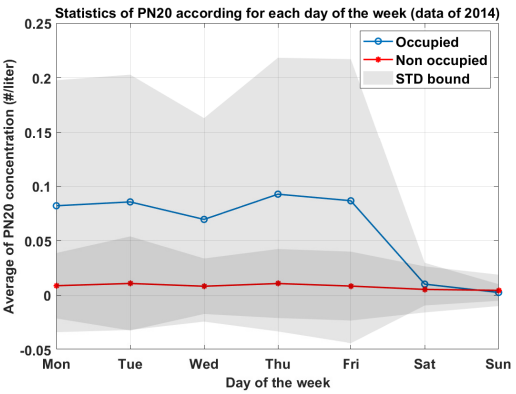
(l)



(m)



(n)



(o)

Figure A.4: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week in two cases: the office is occupied or non-occupied (continued).

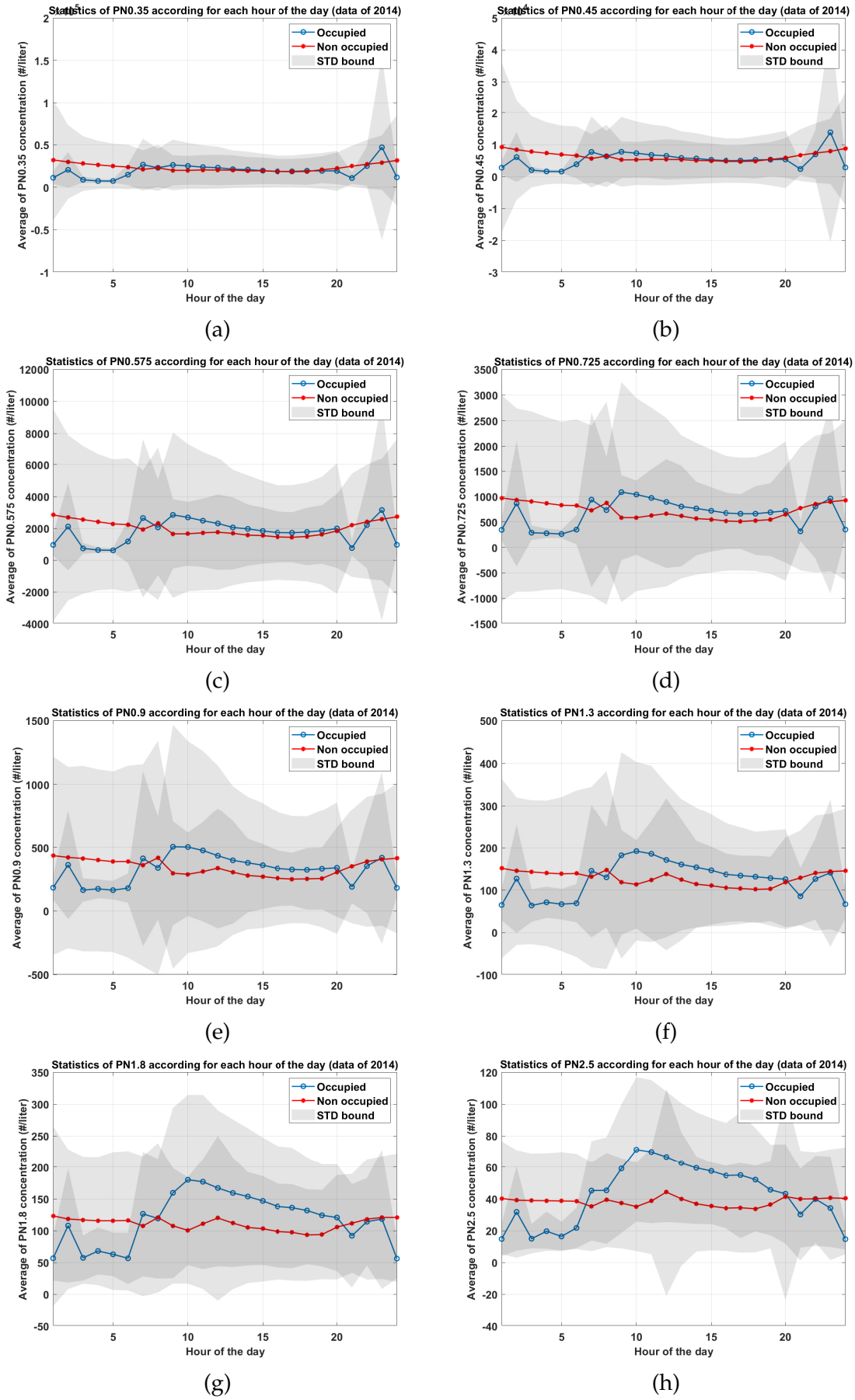
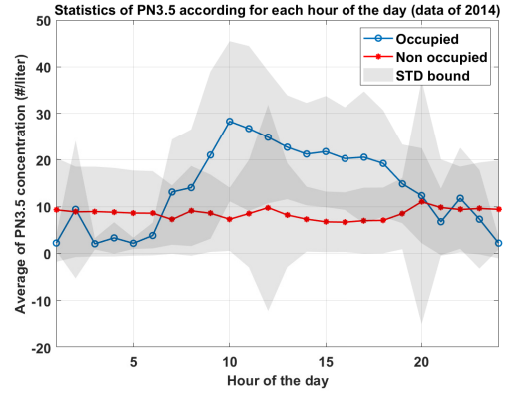
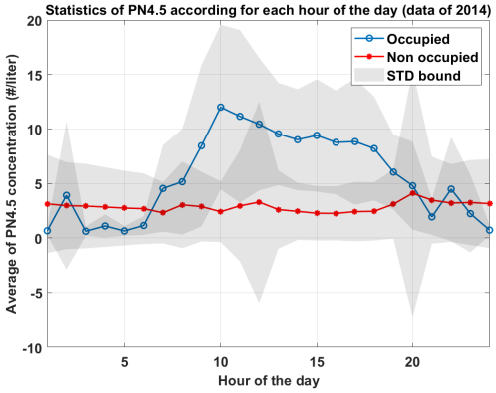


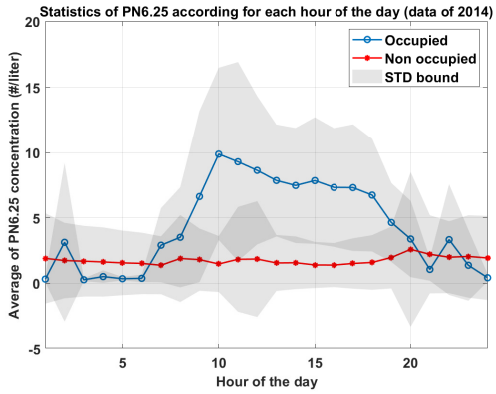
Figure A.5: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day in two cases: the office is occupied or non-occupied (to be continue).



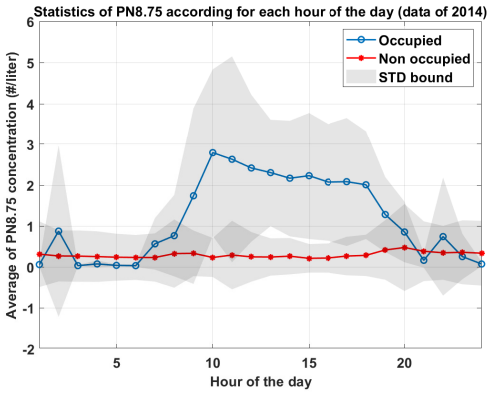
(i)



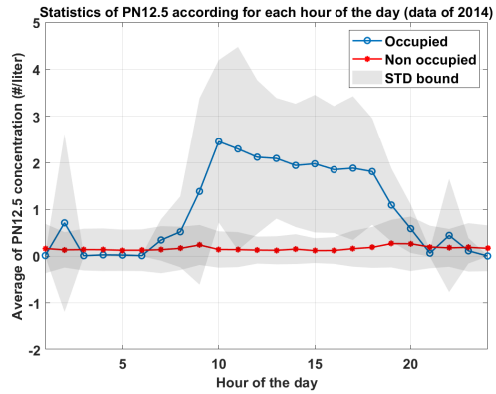
(j)



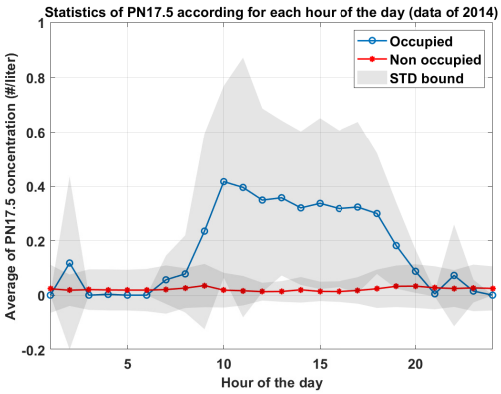
(k)



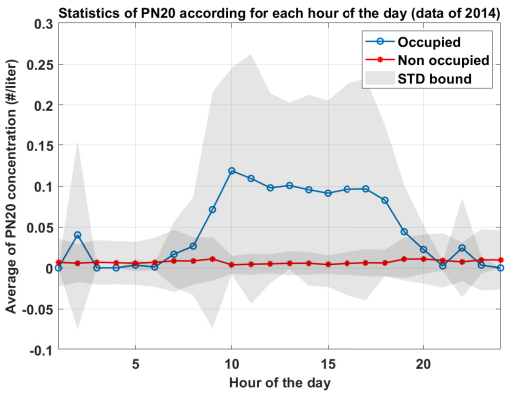
(l)



(m)



(n)



(o)

Figure A.5: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day in two cases: the office is occupied or non-occupied (continued).

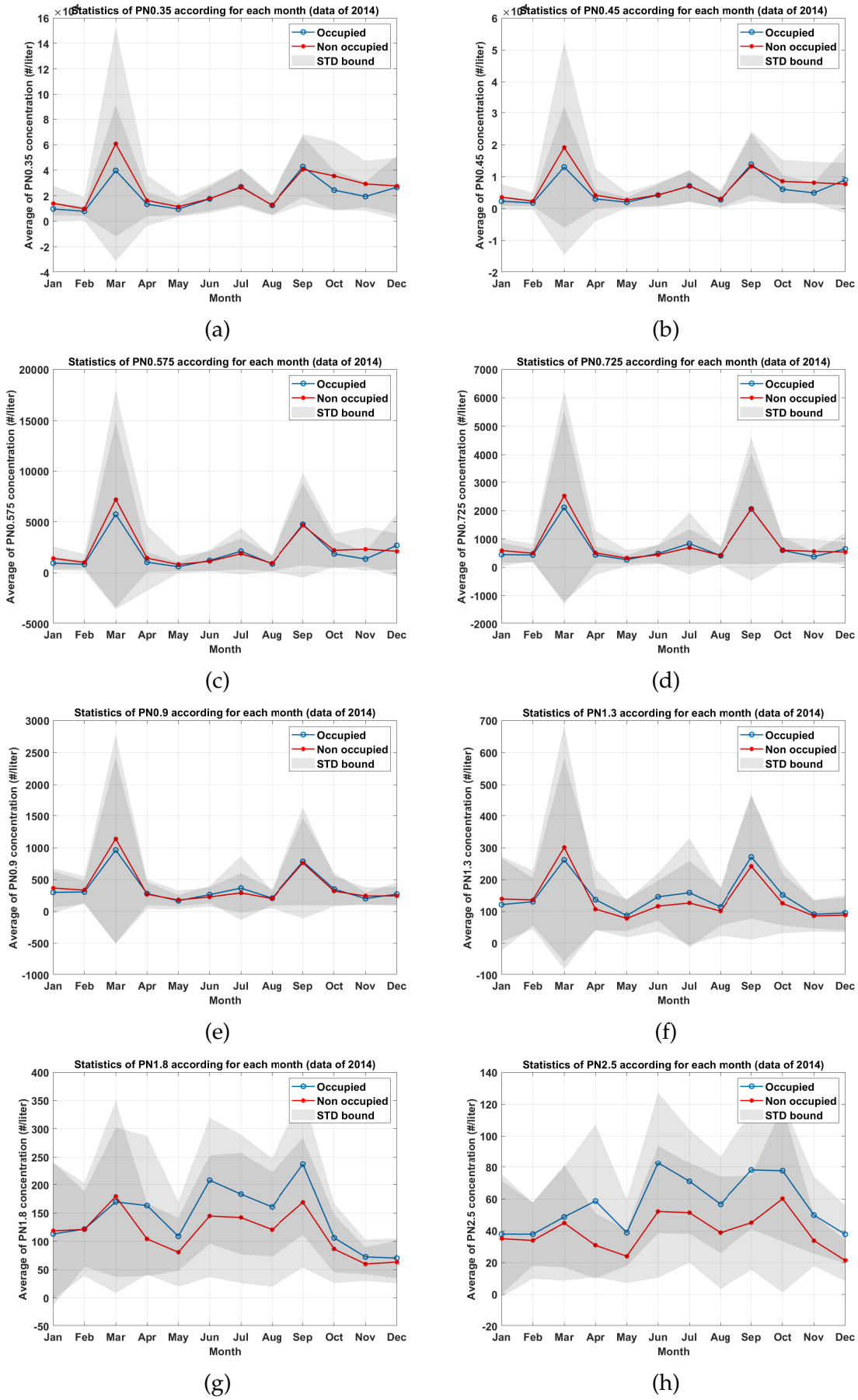
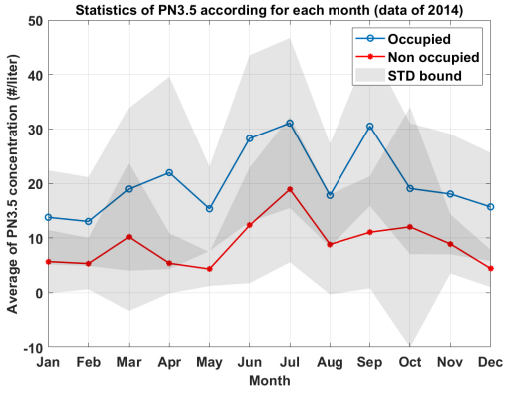
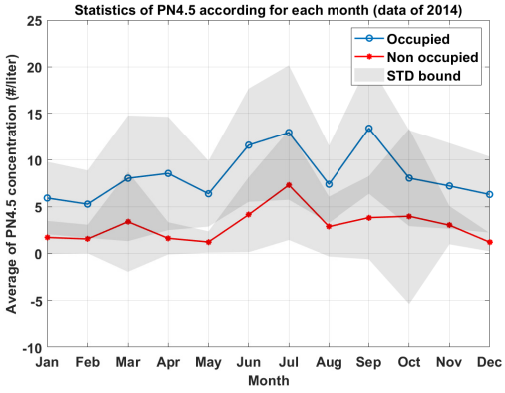


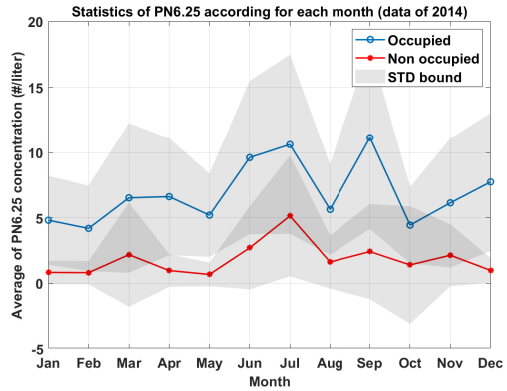
Figure A.6: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month in two cases: the office is occupied or non-occupied (to be continue).



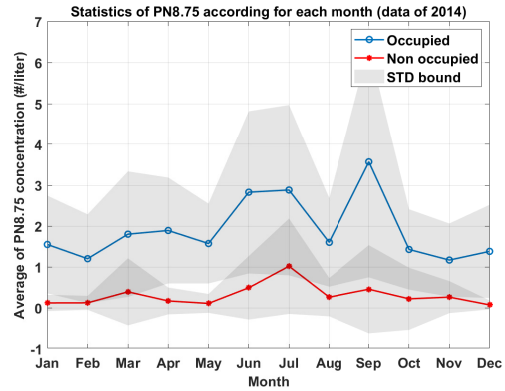
(i)



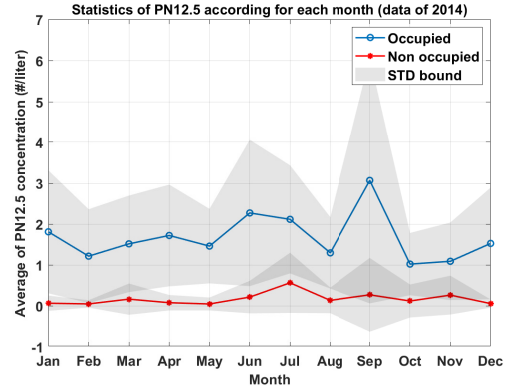
(j)



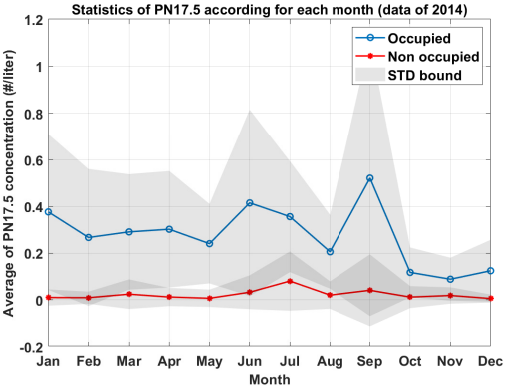
(k)



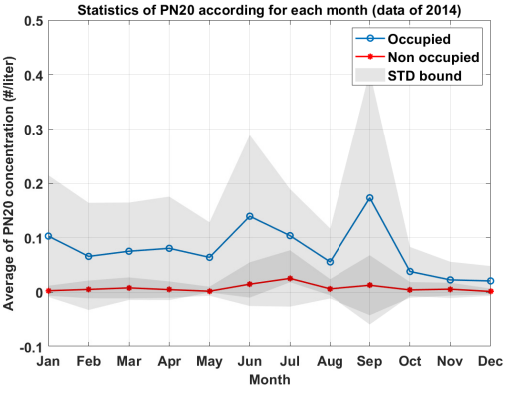
(l)



(m)



(n)



(o)

Figure A.6: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month in two cases: the office is occupied or non-occupied (continued).

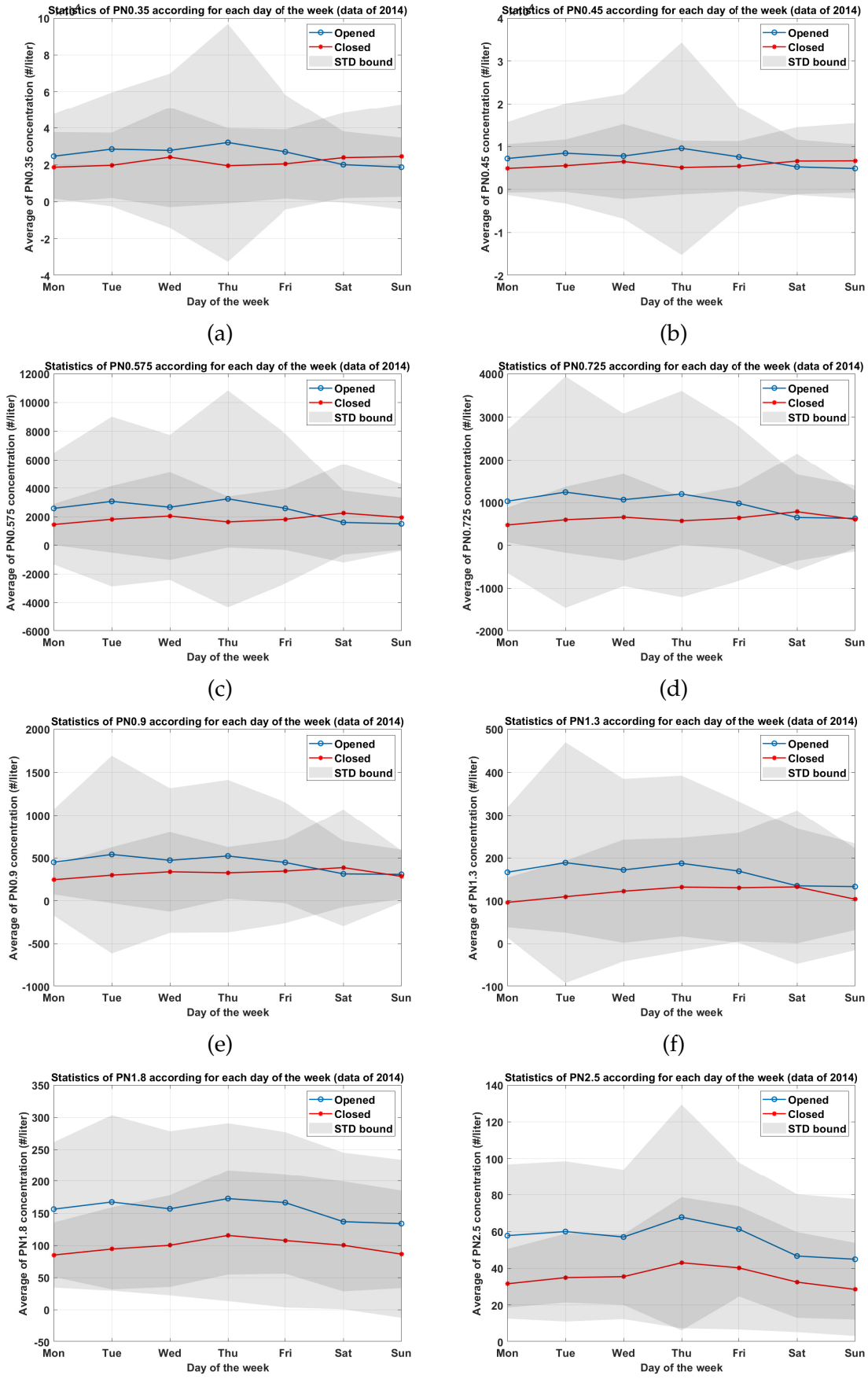
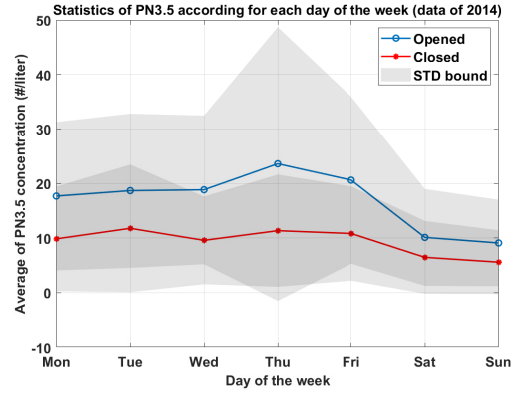
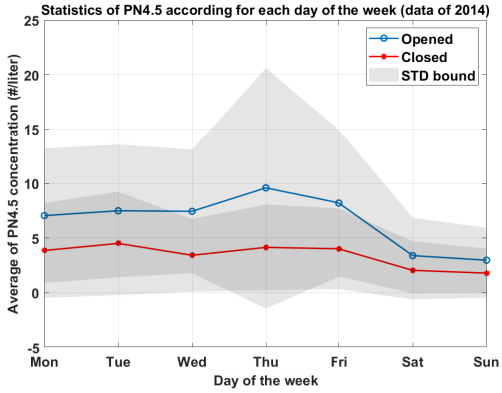


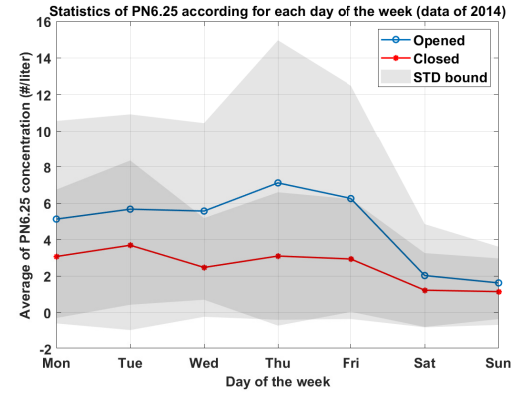
Figure A.7: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week in two cases: when the windows are opened (at least 1 window is opened) or closed (to be continue).



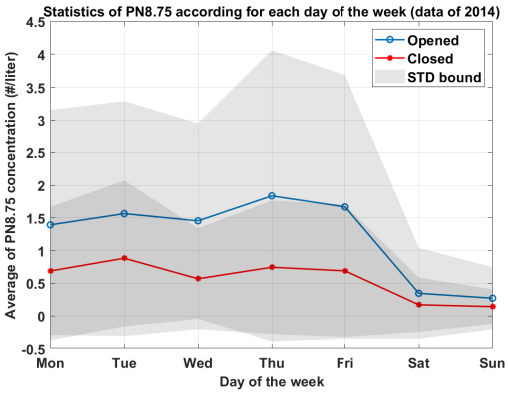
(i)



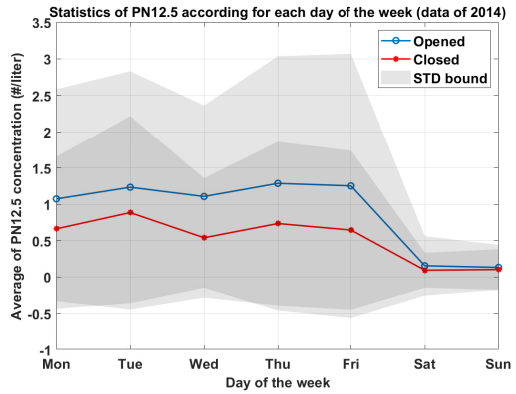
(j)



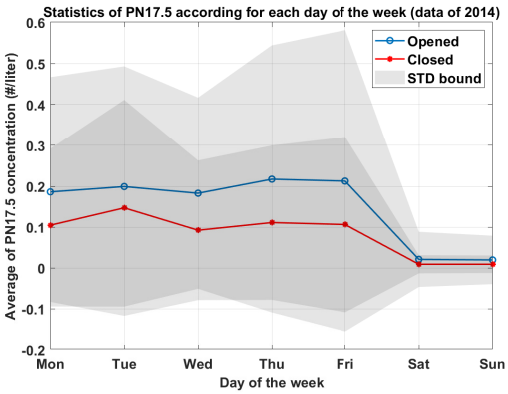
(k)



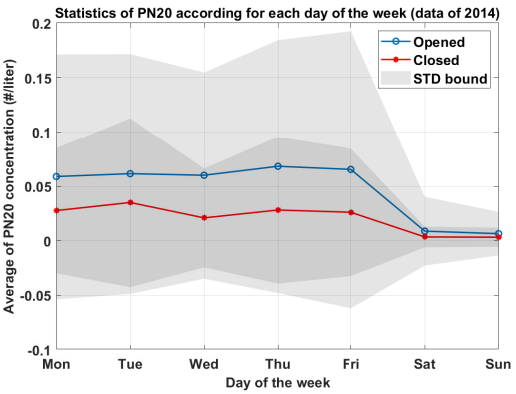
(l)



(m)



(n)



(o)

Figure A.7: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the day of the week in two cases: when the windows are opened (at least 1 window is opened) or closed (continued).

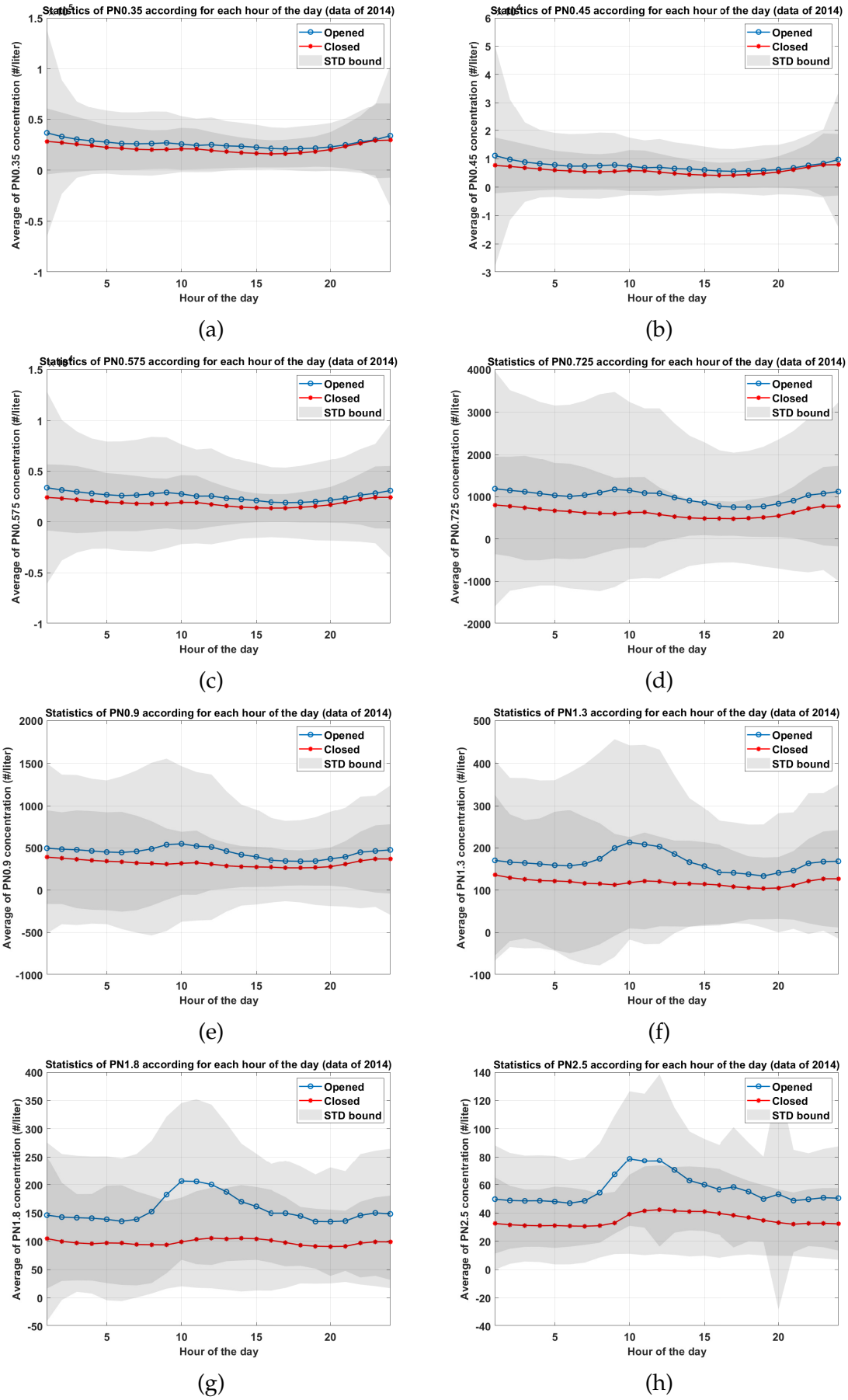
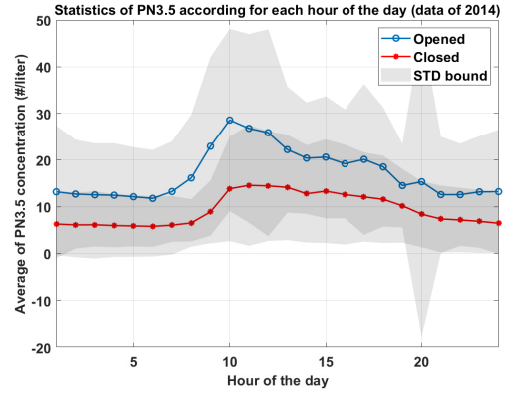
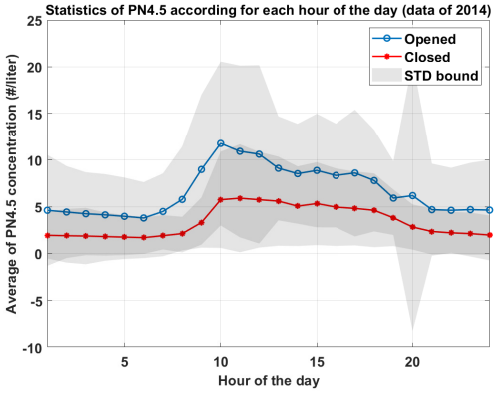


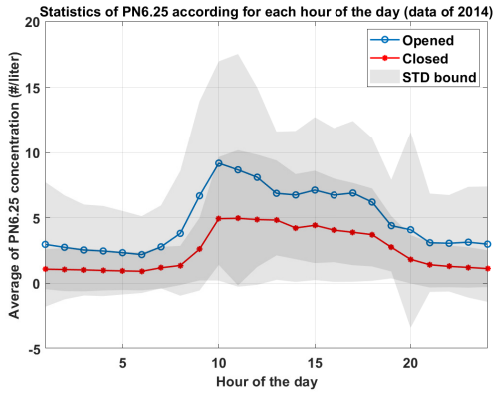
Figure A.8: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day in two cases: when the windows are opened (at least 1 window is opened) or closed (to be continue).



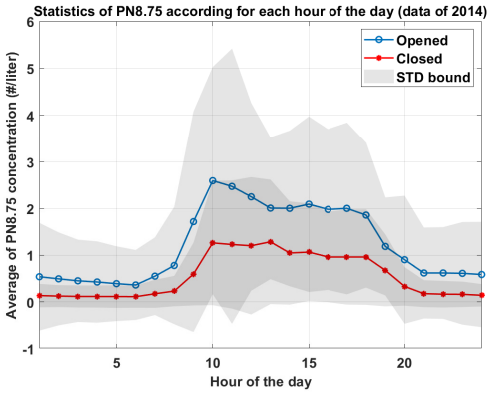
(i)



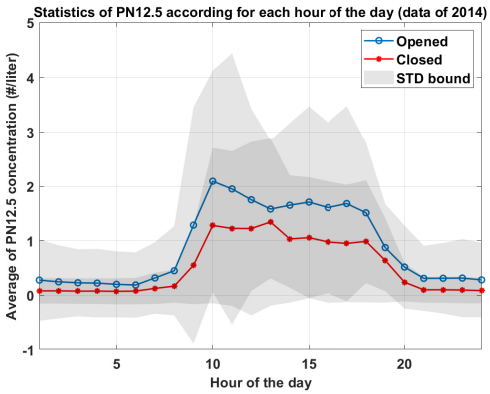
(j)



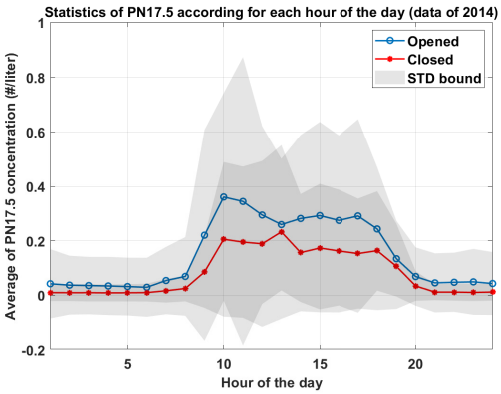
(k)



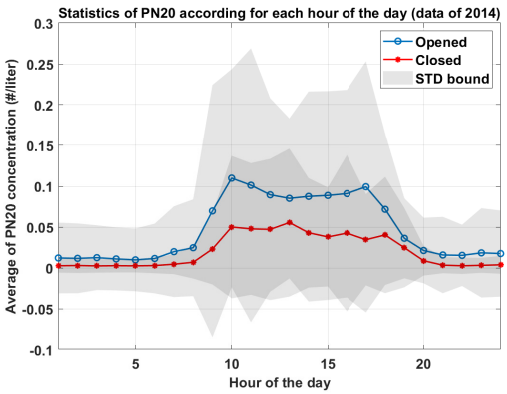
(l)



(m)



(n)



(o)

Figure A.8: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the hour of the day in two cases: when the windows are opened (at least 1 window is opened) or closed (continued).

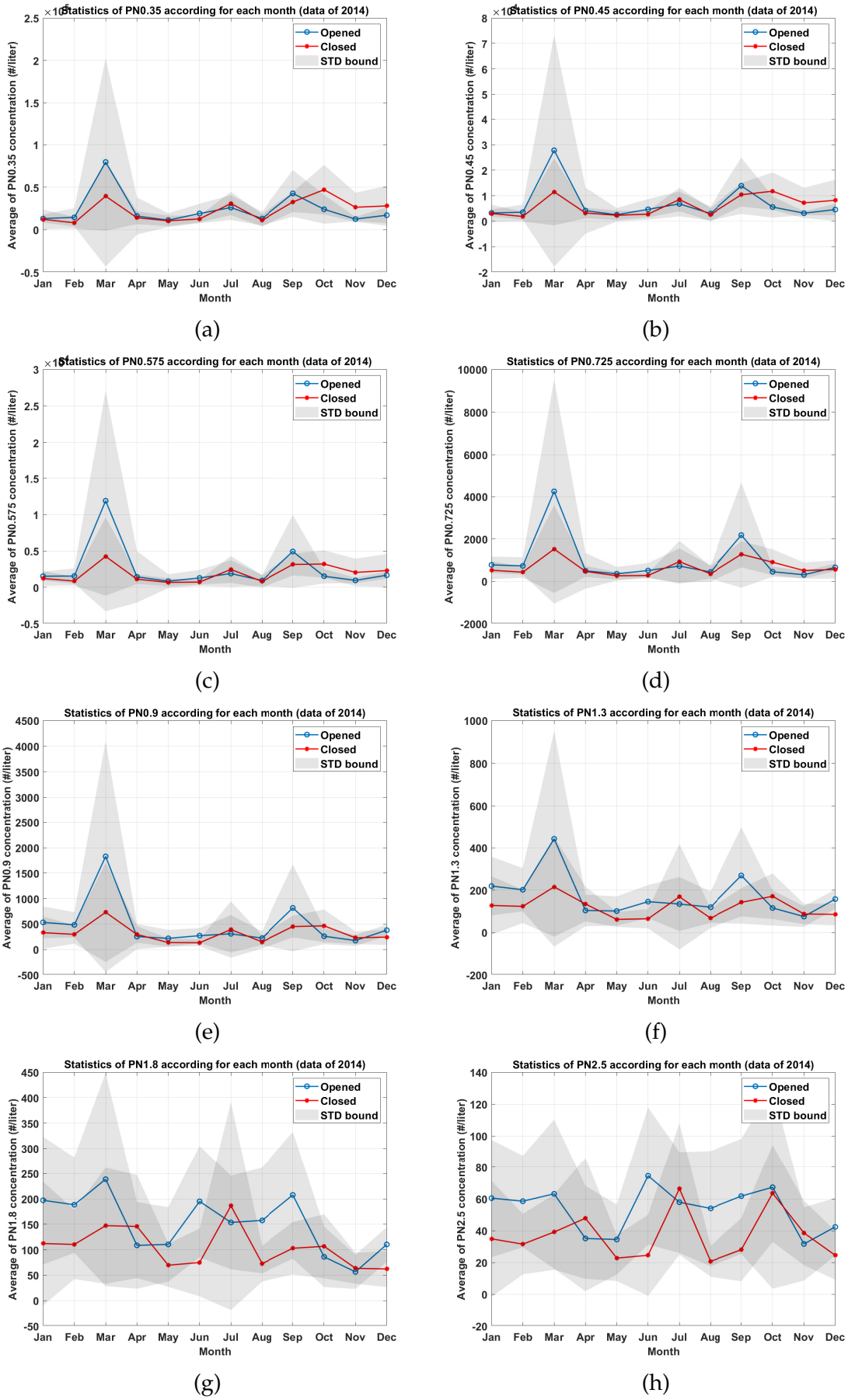
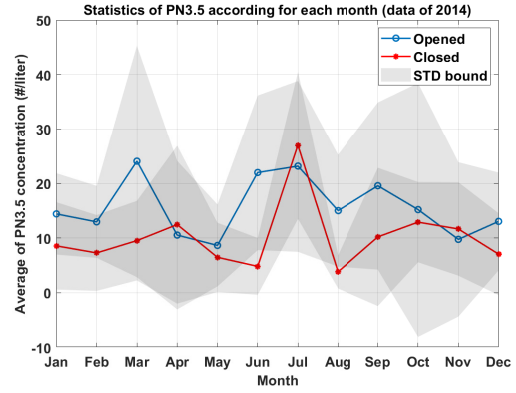
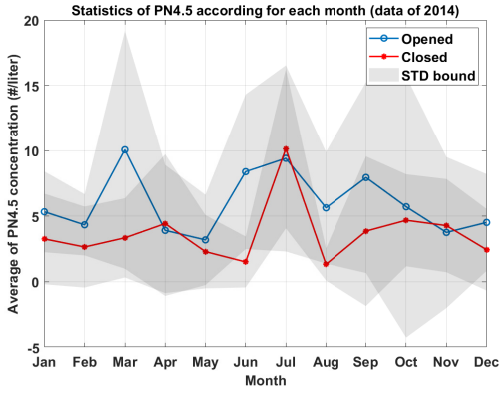


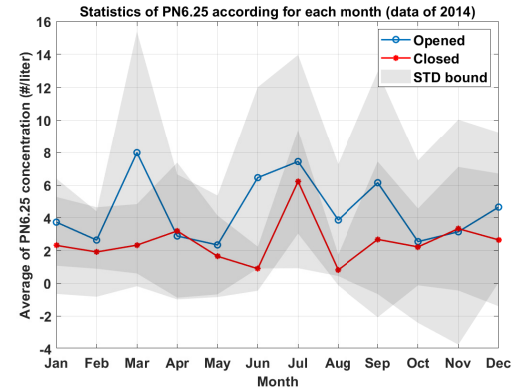
Figure A.9: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month in two cases: when the windows are opened (at least 1 window is opened) or closed (to be continue).



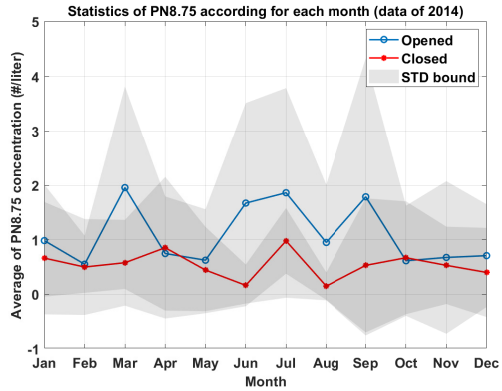
(i)



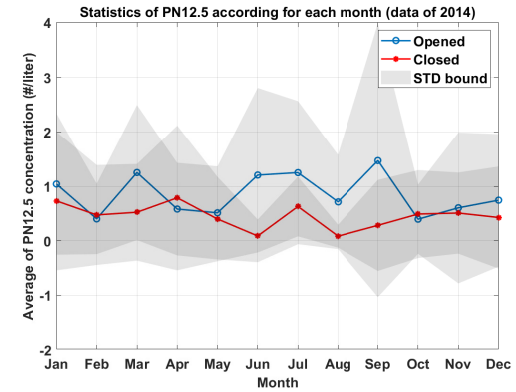
(j)



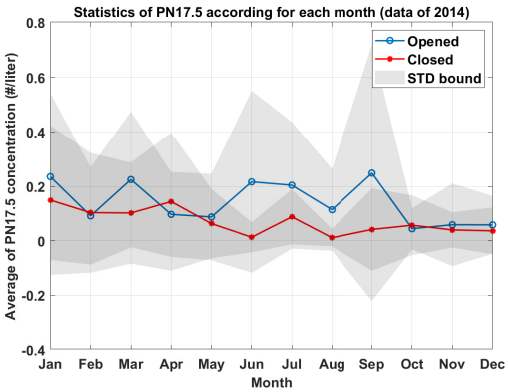
(k)



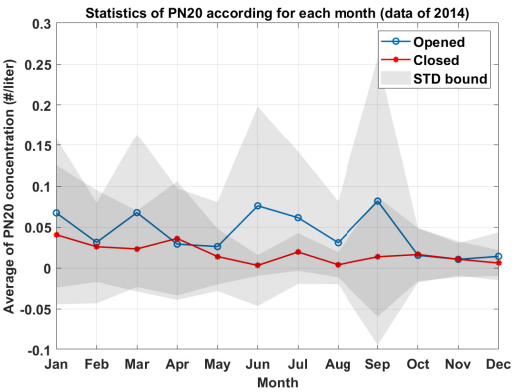
(l)



(m)



(n)



(o)

Figure A.9: Yearly averaged number concentration for the 15 PN fractions monitored indoors during 2014, according to the month in two cases: when the windows are opened (at least 1 window is opened) or closed (continued).

Bibliography

Abadie, M. and P. Blondeau (2011), "PANDORA database: A compilation of indoor air pollutant emissions", *HVAC & R Research*, 17 (Aug. 2011), pp. 602-613, DOI: 10.1080/10789669.2011.579877.

Abt, E., H. Suh, P. Catalano, and P. Koutrakis (2000), "Relative Contribution of Outdoor and Indoor Particle Sources to Indoor Concentrations", *Environmental Science & Technology*, 34 (Aug. 2000), 3579–3587, DOI: 10.1021/es990348y.

Agency for Toxic Substances and Disease Registry (2014), *Medical Management Guidelines for Formaldehyde*, {https://wwwn.cdc.gov/TSP/MMG/MMGDetails.aspx?mmgid=216\&toxid=39}.

Aha, W., D. Kibler, and M. Albert (1991), "Instance-Based Learning Algorithms", *Machine Learning*, 6 (Jan. 1991), pp. 37-66, DOI: 10.1023/A:1022689900470.

Alex, J. and S. Bernhard (2000), "Sparse greedy matrix approximation for machine learning", in *International Conference on Machine Learning*, 911–918.

Alzona, J., B.L. Cohen, H. Rudolph, H.N. Jow, and J.O. Frohliger (1979), "Indoor-outdoor relationships for airborne particulate matter of outdoor origin", *Atmospheric Environment*, 13, 1, pp. 55-60, ISSN: 0004-6981, DOI: 10.1016/0004-6981(79)90244-0.

Amasyali, K. and N. M. El-Gohary (2018), "A review of data-driven building energy consumption prediction studies", *Renewable and Sustainable Energy Reviews*, 81, pp. 1192-1205, ISSN: 1364-0321, DOI: 10.1016/j.rser.2017.04.095.

Amato, F., I. Rivas, M. Viana, T. Moreno, L. Bouso, C. Reche, M. Àlvarez Pedrerol, A. Alastuey, J. Sunyer, and X. Querol (2014), "Sources of indoor and outdoor PM2.5 concentrations in primary schools", *Science of The Total Environment*, 490, pp. 757-765, ISSN: 0048-9697, DOI: 10.1016/j.scitotenv.2014.05.051.

American Society of Heating Refrigerating and Air-Conditioning Engineers (2017), *Thermal Environmental Conditions for Human Occupancy*, tech. rep., American Society of Heating Refrigerating and Air-Conditioning Engineers.

Andersen, C.M. and R. Bro (2003), "Practical aspects of PARAFAC modeling of fluorescence excitation-emission data", *Journal of Chemometrics*, 17 (Apr. 2003), pp. 200-215, DOI: 10.1002/cem.790.

Andersen, R., V. Fabi, J. Toftum, S. P. Corgnati, and B. W. Olesen (2013), "Window opening behaviour modelled from measurements in Danish dwellings", *Building and Environment*, 69, pp. 101-113.

Anderson, M., E. Daly, S. Miller, and J. Milford (2002), "Source apportionment of exposures to volatile organic compounds: II. Application of receptor models to TEAM study data", *Atmospheric Environment*, 36 (Aug. 2002), pp. 3643-3658, DOI: 10.1016/S1352-2310(02)00280-7.

Astel, A., V. Simeonov, H. Bauer, and H. Puxbaum (2010), "Multidimensional modeling of aerosol monitoring data", *Environmental Pollution*, 158, 10, pp. 3201-

3208, ISSN: 0269-7491, DOI: https://doi.org/10.1016/j.envpol.2010.07.003.

Atul (2022), *What is Machine Learning Machine Learning For Beginners*, https://www.edureka.co/blog/what-is-machine-learning/.

Austin, J., P. Brimblecombe, and W. Sturges (2002), *Air Pollution Science for the 21st Century*, Pergamon.

Bako-Biro, Z., P. Wargocki, C. Weschler, and P.O. Fanger (2004), "Effects of pollution from personal computers on perceived air quality, SBS symptoms and productivity in office", *Indoor air*, 14 (July 2004), pp. 178-187, DOI: 10.1111/j.1600-0668.2004.00218.x.

Bento, C. (2021), *Decision Tree Classifier explained in real-life: picking a vacation destination*, https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575.

Berge, J. (1989), "Convergence of ParaFac preprocessing procedures and the Deming-Stephan method of iterative proportional fitting", in *Multiway data analysis*, pp. 53-63.

Bishop, C. (2006), *Pattern recognition and machine learning*, Springer.

Bishop, C. M. (2012), "Model-based machine learning." *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, DOI: 10.1098/rsta.2012.0222.

Blifford, I. H. and G. O. Meeker (1967), "A factor analysis model of large scale pollution", *Atmospheric Environment (1967)*, 1, 2, pp. 147-157, ISSN: 0004-6981, DOI: 10.1016/0004-6981(67)90042-X.

Box, G.E.P., G. M. Jenkins, and G. Reinsel (1994), *Time Series Analysis: Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Bro, R. (1997), "PARAFAC tutorial and applications", *Chemometrics and Intelligent Laboratory Systems*.

Bro, R. and H. Kiers (2003), "A new Efficient Method for Determining the Number of Components in PARAFAC Models", *Journal of Chemometrics*, 17 (June 2003), pp. 274-286, DOI: 10.1002/cem.801.

Brook, R., B. Franklin, W. Cascio, Y. Hong, G. Howard, M. Lipsett, R. Luepker, M. Mittleman, J. Samet, S. Smith, and I Tager (2004), "Air Pollution and Cardiovascular Disease A Statement for Healthcare Professionals From the Expert Panel on Population and Prevention Science of the American Heart Association", *Circulation*, 109 (July 2004), pp. 2655-2671, DOI: 10.1161/01.CIR.0000128587.30041.C8.

Brown, S. (1997), *National State of the Environment Report - Indoor Air Quality. SoE Technical Report Series*, tech. rep., Dept. Environment, Sports & Territories, Canberra.

Brown, S., A. Frankel, and H. Hafner (2007), "Source apportionment of VOCs in the Los Angeles area using positive matrix factorization", *Atmospheric Environment*, 41, 2, pp. 227-237, ISSN: 1352-2310, DOI: 10.1016/j.atmosenv.2006.08.021.

Cai, C., F. Geng, X. Tie, Q. Yu, and J. An (2010), "Characteristics and source apportionment of VOCs measured in Shanghai, China", *Atmospheric Environment*, 44, 38, pp. 5005-5014, ISSN: 1352-2310, DOI: 10.1016/j.atmosenv.2010.07.059.

Cali, D., M. T. Wesseling, and D. Muller (2018), "WinProGen: A Markov-Chain-based stochastic window status profile generator for the simulation of realistic energy performance in buildings." *Building and Environment*, 136, pp. 240-258.

Carroll, J. D. and J. Chang (1970), "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition", *Psychometrika*, 35, pp. 283-319.

Catelinois, O., A. Rogel, D. Laurier, S. Billon, D. Hemon, P. Verger, and M. Tirmarche (2006), "Lung cancer attributable to indoor radon exposure in france: impact of the risk models and uncertainty analysis", *Environmental health perspectives*, 114, 9, 1361–1366, DOI: 10.1289/ehp.9070.

Cawley, G. C. and N. L. C. Talbot (2010), "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", *Journal of Machine Learning Research*, pp. 2079-2107.

Centers for Disease Control and Prevention (2004), *About SARS*, https://www.cdc.gov/sars/about/faq.html (visited on 04/2004).

Chen, C. and B. Zhao (2011), "Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor", *Atmospheric Environment*, 45 (Jan. 2011), pp. 275-288, DOI: 10.1016/j.atmosenv.2010.09.048.

Chen, S., K. Mihara, and J. Wen (2018), "Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points", *Building and Environment*, 146, pp. 238-246, ISSN: 0360-1323, DOI: 10.1016/j.buildenv.2018.09.054.

Cheng, Y. and Y. Lin (2010), "Measurement of Particle Mass Concentrations and Size Distributions in an Underground Station", *Aerosol and Air Quality Research*, 10 (Feb. 2010), pp. 22-29, DOI: 10.4209/aaqr.2009.05.0037.

Cherry, E. C. (1953), "Some Experiments on the Recognition of Speech, with One and with Two Ears", *The Journal of the Acoustical Society of America*, 25, 5, pp. 975-979, DOI: 10.1121/1.1907229.

Choi, S., H. Hong, H. Glotin, and F. Berthommier (2002), "Multichannel signal separation for cocktail party speech recognition: a dynamic recurrent network", *Neurocomputing*, 49, 1, pp. 299-314, ISSN: 0925-2312, DOI: 10.1016/S0925-2312(02)00522-2.

Chow, J. and J. Watson (1998), *Guideline On Speciated Particulate Monitoring*, pp. -.

Christensen, W. and R. Gunst (2004), "Measurement Error Models in Chemical Mass Balance Analysis of Air Quality Data", *Atmospheric Environment*, 38 (Feb. 2004), pp. 733-744, DOI: 10.1016/j.atmosenv.2003.10.018.

Cichocki, A., R. Zdunek, and S. Amari (2006), "Csiszár's Divergences for Non-negative Matrix Factorization: Family of New Algorithms", in, vol. 3889, pp. 32-39, ISBN: 978-3-540-32630-4, DOI: 10.1007/11679363_5.

Cogliano, V., Y. Grosse, R. Baan, K. Straif, B. Lauby-Secretan, and Fatiha Ghissassi (2005), "Meeting Report: Summary of IARC Monographs on Formaldehyde, 2-Butoxyethanol, and 1-tert-Butoxy-2-Propanol", *Environmental health perspectives*, 113 (Oct. 2005), pp. 1205-1208, DOI: 10.1289/ehp.7542.

Colleen, M. (2015), "Chapter 7 - Identification, Characterization, and Modeling", in *Data Mining and Predictive Analysis (Second Edition)*, ed. by McCue Colleen, Second Edition, Butterworth-Heinemann, Boston, pp. 137-155, ISBN: 978-0-12-800229-2, DOI: 10.1016/B978-0-12-800229-2.00007-9.

Colucci, J. M. and C. R. Begeman (1965), "The Automotive Contribution to Air-Borne Polynuclear Aromatic Hydrocarbons in Detroit", *Journal of the Air Pollution Control Association*, 15, 3, pp. 113-122, DOI: 10.1080/00022470.1965.10468342.

Comon, P. and C. Jutten (2007), *Séparation de sources 1 concepts de base et analyse en composantes indpendantes*, Lavoisier.

Cortes, C., M. Mohri, and A. Talwalkar (2010), "On the Impact of Kernel Approximation on Learning Accuracy." *Journal of Machine Learning Research - Proceedings Track*, 9 (Jan. 2010), pp. 113-120.

Crook, B. and N. C. Burton (2010), "Indoor moulds, Sick Building Syndrome and building related illness", *Fungal Biology Reviews*, 24, 3, pp. 106-113, ISSN: 1749-4613, DOI: 10.1016/j.fbr.2010.05.001.

Dai, X., J. Liu, and X Zhang (2020), "A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings." *Energy and Buildings*, 223, pp. 110-159.

Delmaire, G., G. Roussel, D. Hleis, and F. Ledoux (2010), "Une version pondérée de la Factorisation Matricielle Non négative pour l'identification de sources de particules atmosphériques. Application au littoral de la Mer du Nord", *Journal Européen des Systèmes Automatisés*, 44 (May 2010), DOI: 10.3166/jesa.44.547-566.

Derbez, M., A. Gregoire, O. Ramalho, J. Garrigue, and S. Kirchner (2006), "French permanent survey on indoor air quality-Part 2.: Questionnaires and validation procedure of collected data", *Proceedings Healthy Building 2006*, 3 (June 2006), pp. 327-331.

Després, V. R., J. Huffman, S. Burrows Burrows, Corinna H., AleksandrS. Safatov, Galina Buryak, Janine Fröhlich-Nowoisky, Wolfgang Elbert, MeinratO. Andreae, Ulrich Pöschl, and Ruprecht Jaenicke (2012), "Primary biological aerosol particles in the atmosphere: a review", *Tellus B: Chemical and Physical Meteorology*, 64, 1, p. 15598, DOI: 10.3402/tellusb.v64i0.15598.

Destaillats, H., M. Lunden, B. Singer, B. Coleman, A. Hodgson, C. Weschler, and W. Nazaroff (2006), "Indoor Secondary Pollutants from Household Product Emissions in the Presence of Ozone: A Bench-Scale Chamber Study", *Environmental Science and Technology*, 40 (Aug. 2006), pp. 4421-4428, DOI: 10.1021/es052198z.

Destaillats, H., R. Maddalena, B. Singer, A. Hodgson, and T. McKone (2008), "Indoor pollutants emitted by office equipment: A review of reported data and information needs", *Atmospheric Environment*, 42 (Jan. 2008), pp. 1371-1388.

D'Oca, S. and T. Hong (2014), "A data-mining approach to discover patterns of window opening and closing behavior in offices", *Building and Environment*, 82, pp. 726-739.

Dorothée, M., F. Chaventré, O. Ramalho, J. Laffitte, B. Collignan, and K. Weiss (2013), "De l'évaluation du risque à la gestion de la crise : le cas du syndrome des bâtiments malsains", *Environnement, Risques & Sant* (July 2013), 325–329.

Dreiseitl, S. and L. Ohno-Machado (2002), "Logistic regression and artificial neural network classification models: a methodology review", *Journal of Biomedical Informatics*, 35, 5, pp. 352-359, ISSN: 1532-0464, DOI: 10.1016/S1532-0464(03)00034-0.

Education, IBM Cloud (2021), *Underfitting*, https://www.ibm.com/cloud/learn/underfitting.

Edwards, R.E., J. New, and L. Parker (2012), "Predicting future hourly residential electrical consumption: A machine learning case study", *Energy and Buildings*, 49, pp. 591-603, ISSN: 0378-7788, DOI: 10.1016/j.enbuild.2012.03.010.

El Naqa, I. and M. J. Murphy (2015), "What Is Machine Learning", in, *Machine Learning in Radiation Oncology: Theory and Applications*, ed. by Issam El Naqa, Ruijiang Li, and Martin J. Murphy, Springer International Publishing, Cham, pp. 3-11, ISBN: 978-3-319-18305-3, DOI: 10.1007/978-3-319-18305-3_1.

Environmental Protection Agency (2014), *ENERGY STAR Score for Offices in the United States*, tech. rep., EPA.

European Environment Agency (2014), *Air quality in Europe — 2014 report*, tech. rep., EEA, https://www.eea.europa.eu/publications/air-quality-in-europe-2014.

Everitt, B.S. and A. Skrondal (2010), *Cambridge Dictionary of Statistics*, tech. rep.

Ezzati, M. (2005), "Indoor air pollution and health in developing countries", *The Lancet*, 366 (July 2005), pp. 104-106, DOI: 10.1016/S0140-6736(05)66845-6.

Fabi, V., R. Andersen, S. Corgnati, and B Olesen (2012), "Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models", *Building and Environment*, 58, pp. 188-198.

Fenech, A, M. Strlič, K. C. Irena, A. Levart, T. Lorraine, B. Gerrit, N. Konstantinos, K. Jana, and C. May (2010), "Volatile aldehydes in libraries and archives", *Atmospheric Environment*, 44, 17, pp. 2067-2073, ISSN: 1352-2310, DOI: 10.1016/j.atmosenv.2010.03.021.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, pp. 179-188.

Fisk, W., E. Ekaterina, and M. Mendell (2010), "Association of residential dampness and mold with respiratory tract infections and bronchitis: A meta-analysis", *Environmental health : a global access science source*, 9 (Nov. 2010), p. 72, DOI: 10.1186/1476-069X-9-72.

Fisk, W. and A. Rosenfeld (2004), "Improved Productivity and Health from Better Indoor Environments", *Indoor Air*, 7 (Apr. 2004), pp. 158-172, DOI: 10.1111/j.1600-0668.1997.t01-1-00002.x.

Fix, E. and J. L Hodges (1951), *Discriminatory analysis : nonparametric discrimination, consistency properties*, USAF School of Aviation Medicine.

Gilbert, K. (2019), *A framework for multiple algorithm source separation*, PhD thesis.

Godish, T. and J. D. Spengler (1996), "Relationships Between Ventilation and Indoor Air Quality: A Review", *Indoor Air*, 6, 2, pp. 135-145, DOI: 10.1111/j.1600-0668.1996.00010.x.

Green, B. F. (1966), "The computer revolution in psychometrics", *Psychometrika*, 31, 4, pp. 437-445, DOI: 10.1007/BF02289515.

Gundel, L. A., G. S. Richard, S. R. Lev, and H. Naomi (2005), "Aerosol physics and chemistry: indoor perspective", *Aerosol Handbook: Measurement, Dosimetry and Health Effects*.

Guo, H. (2011), "Source apportionment of volatile organic compounds in Hong Kong homes", *Building and Environment*, 46, 11, pp. 2280-2286, ISSN: 0360-1323, DOI: 10.1016/j.buildenv.2011.05.008.

Géhin, E., O. Ramalho, and S. Kirchner (2008), "Size distribution and emission rate measurement of fine and ultrafine particle from indoor human activities", *Atmospheric Environment*, 42 (Nov. 2008), pp. 8341-8352, DOI: 10.1016/j.atmosenv.2008.07.021.

Hamilton, W.D. and T.M. Lenton (1998), "Spora and Gaia: how microbes fly with their clouds", *Ethology Ecology & Evolution*, 10, 1, pp. 1-16, DOI: 10.1080/08927014.1998.9522867.

Han, Y., M. Qi, Y; Chen, H. Shen, J. Liu, Ye Huang, Han Chen, Wenxin Liu, Xilong Wang, Junfeng Liu, Baoshan Xing, and Shu Tao (2015), "Influences of ambient air PM2.5 concentration and meteorological condition on the indoor PM2.5 concentrations in a residential apartment in Beijing using a new approach", *Environmental pollution*, 205 (June 2015), pp. 307-314, DOI: 10.1016/j.envpol.2015.04.026.

Harley, Naomi (2020), *RADON AND LUNG CANCER*, pp. 877-909, ISBN: 9781119438809, DOI: 10.1002/9781119438922.ch23.

Harshman, R. A. (1970), "Foundations of the PARAFAC procedure: Models and conditions for an ëxplanatorym̈ulti-model factor analysis", in *UCLA Working Papers in Phonetics*.

Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Hinds, W. C. (1999), *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*, Wiley.

Hitchcock, F. L. (1927), "The expression of a tensor or a polyadic as a sum of products", *J. Math. Phys 6*, pp. 164-189.

Ho, T. (2016), "Random Decision Forests", in *3rd International Conference on Document Analysis and Recognition*, 278–282.

Hopke, P. (2010), "The Application of Receptor Modeling to Air Quality Data", *Pollution Atmospherique* (Jan. 2010), pp. 91-109.

Hopke, P., Z. Ramadan, P. Paatero, G. Norris, M. Landis, R. Williams, and C. Lewis (2003), "Receptor modeling of ambient and personal exposure samples: 1998 Baltimore Particulate Matter Epidemiology-Exposure Study", *Atmospheric Environment*, 37 (July 2003), pp. 3289-3302, DOI: 10.1016/S1352-2310(03)00331-5.

Hopke, P. K. (1991), "Receptor modeling for air quality management", in *Data Handling in Science and Technology*, Elsevier Science.

– (2016), "Review of receptor modeling methods for source apportionment", *Journal of the Air & Waste Management Association*, 66, 3, PMID: 26756961, pp. 237-259, DOI: 10.1080/10962247.2016.1140693.

Hoskins, J. A (2003), "Health Effects due to Indoor Air Pollution", *Indoor and Built Environment*, 12, 6, pp. 427-433, DOI: 10.1177/1420326X03037109.

Hosmer, D. and S. Lemeshow (2000), *Applied Logistic Regression. Hoboken*, vol. 354, DOI: 10.1002/0471722146.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, 24, 417–441.

Hsieh, C., S. Si, and D. Inderjit (2014), "A divide-and-conquer solver for kernel support vector machines", in *International Conference on Machine Learning*, 566–574.

Hu, M., J. Peng, K. Sun, D. Yue, S. Guo, A. Wiedensohler, and Z. Wu (2012), "Estimation of Size-Resolved Ambient Particle Density Based on the Measurement of Aerosol Number, Mass, and Chemical Size Distributions in the Winter in Beijing", *Environmental science & technology*, 46 (Mar. 2012), pp. 9941-9947, DOI: 10.1021/es204073t.

Ilacqua, V., J. Dawson, M. Breen, S. Singer, and A. Berg (2017), "Effects of climate change on residential infiltration and air pollution exposure", *Journal of exposure science & environmental epidemiology*, 27 (May 2017), pp. 16-23, DOI: 10.1038/jes.2015.38.

Institute of Medicine (2011), *Climate change, the indoor environment, and health*, pp. 1-272, DOI: 10.17226/13115.

Ionescu, A. (2010), *Retour aux sources de pollution atmosphérique : point de vue des scientifiques français.* Tech. rep., Association pour la Prvention de la Pollution Atmosphrique.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning: With Applications in R*, ISBN: 9781461471370.

Jirina, M. (2011), "Classifiers Based on Inverted Distances", in, pp. 369-387, ISBN: 978-953-307-547-1, DOI: 10.5772/13971.

Jones, A.P. (1999), "Indoor air quality and health", *Atmospheric Environment*, 33, 28, pp. 4535-4564, ISSN: 1352-2310, DOI: 10.1016/S1352-2310(99)00272-1.

Kaelbling, L., M. Littman, and A. Moore (1996), "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research*, pp. 237-285.

Kagi, N., S. Fujii, Y. Horiba, N. Namiki, Y. Ohtani, H. Emi, H. Tamura, and Y. Kim (2007), "Indoor air quality for chemical and ultrafine particle contaminants from printers", *Building and Environment*, 42 (May 2007), pp. 1949-1954, DOI: 10.1016/j.buildenv.2006.04.008.

Kara, M., P. Hopke, Y. Dumanoglu, H. Altiok, T. Elbir, M. Odabasi, and A. Bayram (2015), "Characterization of PM Using Multiple Site Data in a Heavily Industrialized Region of Turkey", *Aerosol and Air Quality Research*, 15 (Feb. 2015), pp. 11-27, DOI: 10.4209/aaqr.2014.02.0039.

Keton, W. (2021), *Sample*, https://www.investopedia.com/terms/s/sample.asp.

Kfoury, A., F. Ledoux, A. Limem, G. Delmaire, G. Roussel, and D." Courcot (2014), "The Use of a Non Negative Matrix Factorization Method Combined to PM2.5 Chemical Data for a Source Apportionment Study in Different Environments", in *Air Pollution Modeling and its Application XXIII*, ed. by Douw Steyn and Rohit Mathur, Springer International Publishing, Cham, pp. 79-84, ISBN: 978-3-319-04379-1.

Kfoury, A., F. Ledoux, C. Roche, G. Delmaire, G. Roussel, and D. Courcot (2016), "PM2.5 source apportionment in a French urban coastal site under steelworks emission influences using constrained non-negative matrix factorization receptor model", *Journal of Environmental Sciences*, 40, Changing Complexity of Air Pollution, pp. 114-128, ISSN: 1001-0742, DOI: https://doi.org/10.1016/j.jes.2015.10.025.

Kirchner, S., J. Arenes, C. Cochet, M. Derbez, C. Duboudin, P. Elias, A. Gregoire, B. Jedor, J. Lucas, and et al. Pasquier N. (2007), *Campagne Nationale Logements : état de la qualité de l'air dans les logements français.* Tech. rep., Institut de veille sanitaire, Observatoire de la qualit de l'air intrieur.

Kirchner, S., A. Buchmann, C. Cochet, C. Dassonville, M. Derbez, Yves Leers, Jean-Paul Lucas, Corinne Mandin, Mory Ouattara, and Olivier Ramalho (2011), *Qualité dáir intrieur, qualit de vie. 10 ans de recherche pour mieux respirer*, tech. rep.

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

Koivisto, A., T. Hussein, R. Niemelä, T. Tuomi, and K. Hämeri (2010), "Impact of particle emissions of new laser printers on modeled office room", *Atmospheric Environment*, 44 (June 2010), pp. 2140-2146, DOI: 10.1016/j.atmosenv.2010.02.023.

Kopperud, R., A. Ferro, and L. Hildemann (2004), "Outdoor Versus Indoor Contributions to Indoor Particulate Matter (PM) Determined by Mass Balance Methods", *Journal of the Air & Waste Management Association*, 54 (Oct. 2004), pp. 1188-1196, DOI: 10.1080/10473289.2004.10470983.

Kubba, S. (2017), "Chapter Seven - Indoor Environmental Quality", in *Handbook of Green Building Design and Construction (Second Edition)*, ed. by Sam Kubba, Second Edition, Butterworth-Heinemann, pp. 353-412, ISBN: 978-0-12-810433-0, DOI: 10.1016/B978-0-12-810433-0.00007-1.

Larson, T., T. Gould, C. Simpson, L Liu, C. Claiborn, and J. Lewtas (2004), "Source Apportionment of Indoor, Outdoor, and Personal PM2.5 in Seattle, Washington, Using Positive Matrix Factorization", *Journal of the Air & Waste Management Association*, 54 (Sept. 2004), pp. 1175-1187, DOI: 10.1080/10473289.2004.10470976.

Le, V., T. Sarlos, and A. Smola (2013), "Fastfood: Approximate Kernel Expansions in Loglinear Time", *30th International Conference on Machine Learning, ICML 2013*, 28 (Aug. 2013), 244–252.

Scipy lectures.org (2022), *Nearest-neighbor prediction on iris*, http://scipy-lectures.org/index.html.

Lee, C. and D. Hsu (2007), "Measurements of fine and ultrafine particles formation in photocopy centers in Taiwan", *Atmospheric Environment*, 41 (Oct. 2007), pp. 6598-6609, DOI: 10.1016/j.atmosenv.2007.04.016.

Lee, D. and H. Seung (1999), "Learning the Parts of Objects by Non-Negative Matrix Factorization", *Nature*, 401 (Nov. 1999), pp. 788-791, DOI: 10.1038/44565.

Lee, E., C. Chan, and P. Paatero (1999), "Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants in Hong Kong", *Atmospheric Environment*, 33 (Aug. 1999), pp. 3201-3212, DOI: 10.1016/S1352-2310(99)00113-2.

Lee, S., H. Liu, M. Kim, J. Kim, and C. Yoo (2014), "Online monitoring and interpretation of periodic diurnal and seasonal variations of indoor air pollutants in a subway station using parallel factor analysis (PARAFAC)", *Energy and Buildings*, 68, pp. 87-98, ISSN: 0378-7788, DOI: https://doi.org/10.1016/j.enbuild.2013.09.022.

Leuchner, M. and B. Rappenglück (2010), "VOC source–receptor relationships in Houston during TexAQS-II", *Atmospheric Environment*, 44, 33, pp. 4056-4067, ISSN: 1352-2310, DOI: 10.1016/j.atmosenv.2009.02.029.

Li, S., X. Hou, H. Zhang, and Q. Cheng (2001), "Learning Spatially Localized, Parts-Based Representation.", in, vol. 1, pp. 207-212, DOI: 10.1109/CVPR.2001.990477.

Limem, A., G. Delmaire, M. Puigt, G. Roussel, and D. Courcot (2014), "Non-negative Matrix Factorization under equality constraints—a study of industrial source identification", *Applied Numerical Mathematics*, 85, pp. 1-15, ISSN: 0168-9274, DOI: https://doi.org/10.1016/j.apnum.2014.05.009.

Liu, F., X. Huang, Y. Chen, and J. Suykens (2021), "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, DOI: 10.1109/TPAMI.2021.3097011.

Lévesque, B., P. L. Auger, J. Bourbeau, J.-F. Duchesne, P. Lajoie, and D. Menzies (2003), "Qualité de l'air", in.

Löndahl, J. (2014), "Physical and Biological Properties of Bioaerosols", in, pp. 33-48, ISBN: 978-1-4419-5581-4, DOI: 10.1007/978-1-4419-5582-1_3.

Mark, R. M. T., M. Robert, I. Miyamoto, L. M. Jason, and D. Kaye (2015), "Chapter 1 - Analytics Defined", in *Information Security Analytics*, ed. by Ryan M. Talabis Mark, McPherson Robert, I. Miyamoto, L. Martin Jason, and D. Kaye, Syngress, Boston, pp. 1-12, ISBN: 978-0-12-800207-0, DOI: 10.1016/B978-0-12-800207-0.00001-0.

Markovic, R., E. Grintal, D. Wölki, J. Frisch, and C. van Treeck (2018), "Window opening model using deep learning methods", *Building and Environment*, 145, ISSN: 03601323, DOI: 10.1016/j.buildenv.2018.09.024.

Maroni, M., B. Seifert, and T. Lindvall (1995), "Indoor air quality : a comprehensive reference book", in.

Martuzevicius, D., S. Grinshpun, T. Lee, S. Hu, P. Biswas, T. Reponen, and G. Lemasters (2008), "Traffic-related PM2.5 aerosol in residential houses located near major highways: Indoor versus outdoor concentrations", *Atmospheric Environment*, 42 (Sept. 2008), pp. 6575-6585, DOI: 10.1016/j.atmosenv.2008.05.009.

Martínez-Comesaña, M., P. Eguía-Oller, J. Martínez-Torres, L. Febrero-Garrido, and E. Granada-Álvarez (2022), "Optimisation of thermal comfort and indoor air quality estimations applied to in-use buildings combining NSGA-III and XGBoost", *Sustainable Cities and Society*, 80, p. 103723, ISSN: 2210-6707, DOI: 10.1016/j.scs.2022.103723.

McDermott, J. H. (2009), "The cocktail party problem", *Current Biology*, 19, 22, R1024-R1027, ISSN: 0960-9822, DOI: 10.1016/j.cub.2009.09.005.

Mendell, M., W. Fisk, K. Kreiss, H. Levin, Darryl Alexander, William Cain, John Girman, Cynthia Hines, Paul Jensen, Donald Milton, Larry Rexroat, and Kenneth Wallingford (2002), "Improving the Health of Workers in Indoor Environments: Priority Research Needs for a National Occupational Research Agenda", *American journal of public health*, 92 (Oct. 2002), pp. 1430-1440, DOI: 10.2105/AJPH.92.9.1430.

Miller, S., M. Anderson, E. Daly Daly, and J. Milford (2002), "Source apportionment of exposures to volatile organic compounds. I. Evaluation of receptor models using simulated exposure data", *Atmospheric Environment*, 36, 22, pp. 3629-3641, ISSN: 1352-2310, DOI: 10.1016/S1352-2310(02)00279-0.

Molnar, P., S. Johannesson, and U. Quass (2014), "Source Apportionment of PM2.5 Using Positive Matrix Factorization (PMF) and PMF with Factor Selection", *Aerosol and Air Quality Research*, 14, 3, pp. 725-733, DOI: 10.4209/aaqr.2013.11.0335.

Mooibroek, D., M. Schaap, E.P. Weijers, and R. Hoogerbrugge (2011), "Source apportionment and spatial variability of PM2.5 using measurements at five sites in the Netherlands", *Atmospheric Environment*, 45, 25, pp. 4180-4191, ISSN: 1352-2310, DOI: 10.1016/j.atmosenv.2011.05.017.

Mosqueron, L., I Momas, and Y Moullec (2002), "Personal exposure of Paris office workers to nitrogen dioxide and fine particles", *Occupational and environmental medicine*, 59 (Sept. 2002), pp. 550-555, DOI: 10.1136/oem.59.8.550.

Nadadur, S. and J. Hollingsworth (2015), *Air Pollution and Health Effects*, ISBN: 978-1-4471-6668-9, DOI: 10.1007/978-1-4471-6669-6.

Nakaoka, H., E. Todaka, H. Seto, I. Saito, M. Hanazato, Masahiro Watanabe, and Chisato Mori (2014), "Correlating the symptoms of sick-building syndrome to indoor VOCs concentration levels and odour", *Indoor and Built Environment*, 23 (Oct. 2014), pp. 804-813, DOI: 10.1177/1420326X13500975.

Nam, H. and J. Ryu (2018), "Indoor Radon and Lung Cancer: National Radon Action Plans Are Urgently Required", *Yonsei Medical Journal*, 59 (Nov. 2018), pp. 10-13, DOI: 10.3349/ymj.2018.59.9.1013.

Natekin, A. and A. Knoll (2013), "Gradient Boosting Machines, A Tutorial", *Frontiers in neurorobotics*, 7 (Dec. 2013), p. 21, DOI: 10.3389/fnbot.2013.00021.

National Research Council (US) Committee on Indoor Pollutants (1981), *Indoor Pollutants*, National Academies Press (US).

Nazaroff, W., C.s Weschler, and R. Corsi (2003), "Indoor air chemistry and physics", *Atmospheric Environment*, 37 (Dec. 2003), 5451–5453, DOI: 10.1016/j.atmosenv.2003.09.021.

Nezis, I., G. Biskos, K. Eleftheriadis, P. Fetfatzis, O. Popovicheva, Nikolay Sitnikov, and Olga-Ioanna Kalantzi (2022), "Linking indoor particulate matter and black carbon with sick building syndrome symptoms in a public office building", *Atmospheric Pollution Research*, 13, 1, p. 101292, ISSN: 1309-1042, DOI: 10.1016/j.apr.2021.101292.

Norvig, S. and Peter (1995), *Artificial Intelligence: A Modern Approach (AIMA)*, Prentice Hall.

Ouaret, R., A. Ionescu, and O. Ramalho (2021), "Non-negative matrix factorization for the analysis of particle number concentrations: Characterization of the temporal variability of sources in indoor workplace", *Building and Environment*, 203, p. 108055, ISSN: 0360-1323, DOI: 10.1016/j.buildenv.2021.108055.

Paatero, P. (1997), "Least squares formulation of robust non-negative factor analysis", *Chemometrics and Intelligent Laboratory Systems*, 37, 1, pp. 23-35, ISSN: 0169-7439, DOI: 10.1016/S0169-7439(96)00044-5.

– (1999), "The Multilinear Engine–A Table-Driven, Least Squares Program for Solving Multilinear Problems, Including the n-Way Parallel Factor Analysis Model", *Journal of Computational and Graphical Statistics*, 8, 4, Cited by: 616, 854 – 888, DOI: 10.1080/10618600.1999.10474853.

Paatero, P. and U. Tapper (1994), "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", *Environmetrics*, 5, 2, pp. 111-126, DOI: 10.1002/env.3170050203.

Pan, S., Y. Xiong, Y. Han, X. Zhang, L. Xia, S. Wei, J. Wu, and M. Han (2018), "A study on influential factors of occupant window-opening behavior in an office building in China", *Building and Environment*, 133, pp. 41-50.

Park, J. and C. Choi (2018), "Modeling Occupant Behavior of the Manual Control of Windows in Residential Buildings", *Indoor Air*, 29 (Nov. 2018), DOI: 10.1111/ina.12522.

Park, J., B. Jeong, Y. Chae, and J. Jeong (2020), "Machine learning algorithms for predicting occupantsbehaviour in the manual control of windows for cross-ventilation in homes", *Indoor and Built Environment*, 30, 8, pp. 1106-1123, DOI: 10.1177/1420326X20927070.

Park, J.S. (2013), "Long-term field measurement on effects of wind speed and directional fluctuation on wind-driven cross ventilation in a mock-up building", *Building and Environment*, 62, pp. 1-8, ISSN: 0360-1323, DOI: 10.1016/j.buildenv.2012.12.013.

Pearson, K. F.R.S. (1901), "LIII. On lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 11, pp. 559-572, DOI: 10.1080/14786440109462720.

Pedregosa, F., G. Varoquaux, A. Gramfort, and V. Michel (2022), *6.7. Kernel Approximation*, https://scikit-learn.org/stable/modules/kernel\%5Fapproximation.html.

Pitz, M., J. Cyrys, E. Karg, A. Wiedensohler, H. Wichmann, and J. Heinrich (2003), "Variability of Apparent Particle Density of an Urban Aerosol", *Environmental Science & Technology*, 37, 19, PMID: 14572082, pp. 4336-4342, DOI: 10.1021/es034322p.

Polissar, A., P. Hopke, W. Malm, and J. Sisler (1996), "The ratio of aerosol optical absorption coefficients to sulfur concentrations, as an indicator of smoke from forest fires when sampling in polar regions", *Atmospheric Environment*, 30, 7, pp. 1147-1157, ISSN: 1352-2310, DOI: https://doi.org/10.1016/1352-2310(95)00334-7.

– (1998), "Atmospheric aerosol over Alaska - 1. Spatial and seasonal variability", *JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES*, 103 (Aug. 1998), pp. 19035-19044, DOI: 10.1029/98JD01365.

Quinlan, J. R. (1986), "Induction of decision trees", *Machine Learning*, 1, pp. 81-106, DOI: 10.1007/BF00116251.

Radford, E. P (1976), "Health Aspects of Housing", *Journal of Occupational Medicine*, 18, 105–108.

Rahimi, A. and B. Recht (2008), "Random features for large scale kernel machines", *Advances in Neural Information Processing Systems*, 20 (Jan. 2008), pp. 1177-1184.

Raja, I. A, J. Nicol, J McCartney, and A Humphreys (2001), "Thermal comfort: use of controls in naturally ventilated buildings", *Energy and Buildings*, 33, 3, pp. 235-244, ISSN: 0378-7788, DOI: 10.1016/S0378-7788(00)00087-6.

Ramalho, O., J. Lucas, C. Mandin, and M. Derbez (2012), "Niveaux de particules dans les environnements intrieurs en France", *Pollution Atmospherique* (Nov. 2012), 37–42.

Ramalho, O., R. Ouaret, A. Ionescu, E. Le Ponner, and Y. Candau (2016), *TRIBU - Suivi dynamique en Temps Rèel de la qualité de l'air Intérieur dans un environnement de BUreaux. Contributions des sources et Modèle prévisionnel rapport, PRIMEQUAL*

APR EIAI / projet TRIBU, tech. rep., CSTB, https://www.primequal.fr/sites/default/files/tribu\%5Frf.pdf.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, DOI: 10.1017/CBO9780511812651.

Robinson, J and Wc Nelson (1995), "National Human Activity Pattern Survey Data Base" (Jan. 1995).

Roll, J. (1981), *Contribution à la proprioception musculaire, à la perception et au contrôle du mouvement chez l'homme.* Tech. rep., Th. Sci. nat. Aix-Marseille 1.

Sahu, S. (2021), *Decision boundary for classifiers: An introduction*, https://medium.com/analytics-vidhya/decision-boundary-for-classifiers-an-introduction-cc67c6d3da0e.

Salthammer, T., S. Mentese, and R. Marutzky (2010), "Formaldehyde in the Indoor Environment", *Chemical reviews*, 110 (Apr. 2010), pp. 2536-72, DOI: 10.1021/cr800399g.

Saraga, D., S. Pateraki, A. Papadopoulos, C. Vasilakos, and T. Maggos (2011), "Studying the indoor air quality in three non-residential environments of different use: A museum, a printery industry and an office", *Building and Environment*, 46 (Nov. 2011), pp. 2333-2341, DOI: 10.1016/j.buildenv.2011.05.013.

Sarangi, B., S. G. Aggarwal, D. Sinha, and P. K. Gupta (2016), "Aerosol effective density measurement using scanning mobility particle sizer and quartz crystal microbalance with the estimation of involved uncertainty", *Atmospheric Measurement Techniques*, 9, 3, pp. 859-875, DOI: 10.5194/amt-9-859-2016.

Sarkhosh, M., A. Najafpoor, H. Alidadi, J. Shamsara, H. Amiri, T. Andrea, and F. Kariminejad (2021), "Indoor Air Quality associations with sick building syndrome: An application of decision tree technology", *Building and Environment*, 188.

Schauer, J., G. Lough, W. Christensen, M. Arndt, J. DeMinter, and J. Park (2006), "Characterization of Metals Emitted from Motor Vehicles", *Research report (Health Effects Institute)*, 133 (Apr. 2006), 1-76; discussion 77.

Schripp, T., S.J. Mulakampilly, W. Delius, E Uhde, M. Wensing, T. Salthammer, R Kreuzig, M. Bahadir, L. Wang, and L. Morawska (2009), "Comparison of ultrafine particle release from hardcopy devices in emission test chambers and office rooms", *Gefahrstoffe Reinhalt. Luft*, 69 (Mar. 2009).

Schölkopf, B. and A. Smola (2002), *Learning with Kernels*.

Seifert, B. and D. Ullrich (1987), "Methodologies for evaluating sources of volatile organic chemicals (VOC) in homes", *Atmospheric Environment*, 21, 2 (Jan. 1987), pp. 395-404, DOI: 10.1016/0004-6981(87)90018-7.

Seinfeld, J. H. and S. N. Pandis (2012), *Atmospheric chemistry and physics : from air pollution to climate change*, John Wiley & Sons.

Seltzer, J. M. (1994), "Building-related illnesses", *Journal of Allergy and Clinical Immunology*, 94, 2, Part 2, Building- and Home-Related Complaints and Illnesses: Sick Building Syndrome, pp. 351-361, ISSN: 0091-6749, DOI: 10.1053/ai.1994.v94.a57114.

Serfozo, R. (2009), *Basics of Applied Stochastic Processes*, ISBN: 978-3-540-89331-8, DOI: 10.1007/978-3-540-89332-5.

Singer, B., B. Coleman, H. Destaillats, A. Hodgson, M. Lunden, C. Weschler, and W. Nazaroff (2006), "Indoor secondary pollutants from cleaning product and air freshener use in the presence of ozone", *Atmospheric Environment*, 40 (Nov. 2006), pp. 6696-6710, DOI: 10.1016/j.atmosenv.2006.06.005.

Spengler, J., J. Samet, and J. McCarthy (2001), *Indoor Air Quality Handbook*, McGraw-Hill Education.

Spengler, J. D. and K. Sexton (1983), "Indoor Air Pollution: A Public Health Perspective", *Science*, 221, 4605, pp. 9-17, DOI: 10.1126/science.6857273.

Stanimirova, I. and V. Simeonov (2005), "Modeling of environmental four-way data from air quality control", *Chemometrics and Intelligent Laboratory Systems*, 77, 1, pp. 115-121, ISSN: 0169-7439, DOI: 10.1016/j.chemolab.2004.11.005.

Stehman, S. (1997), "Selecting and interpreting measures of thematic classification accuracy", *Remote Sensing of Environment*, 62 (Oct. 1997), pp. 77-89, DOI: 10.1016/S0034-4257(97)00083-7.

Stolwik, J. A. J. (1992), "Risk Assessment of Acute Health and Comfort Effects of Indoor Air Pollution", *Annals of the New York Academy of Sciences*, 641, 1, pp. 56-62, DOI: 10.1111/j.1749-6632.1992.tb16532.x.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 2, pp. 111-133, DOI: 10.1111/j.2517-6161.1974.tb00994.x.

Sundell, J. (2004), "On the history of indoor air quality and health", *Indoor air*, 14 Suppl 7 (Feb. 2004), pp. 51-58, DOI: 10.1111/j.1600-0668.2004.00273.x.

Suryawanshi, S., A. Chauhan, R. Verma, and T. Gupta (2016), "Identification and quantification of indoor air pollutant sources within a residential academic campus", *Science of The Total Environment*, 569-570, pp. 46-52, ISSN: 0048-9697, DOI: 10.1016/j.scitotenv.2016.06.061.

Suzuki, N., H. Nakaoka, Y. Nakayama, K. Tsumura, K. Takaguchi, Kazunari Takaya, Akifumi Eguchi, Masamichi Hanazato, Emiko Todaka, and Chisato Mori (2020), "Association between sum of volatile organic compounds and occurrence of building-related symptoms in humans: A study in real full-scale laboratory houses", *Science of The Total Environment*, 750 (Aug. 2020), p. 141635, DOI: 10.1016/j.scitotenv.2020.141635.

Tan, P., M. Steinbach, M. Adeyeye Oshin, and V. Kumar (2005), *Introduction to Data Mining*.

Tien, P., S. Wei, T. Liu, J. Calautit, J. Darkwa, and C. Wood (2021), "A deep learning approach towards the detection and recognition of opening of windows for effective management of building ventilation heat losses and reducing space heating demand", *Renewable Energy*, 177, pp. 603-625, ISSN: 0960-1481, DOI: 10.1016/j.renene.2021.05.155.

Uhde, E., C. He, and M. W. Wensing (2006), "Characterization of Ultra-fine Particle Emissions from a Laser Printer", in, pp. 479-482.

United States Environmental Protection Agency (2022), *Air Pollutant Receptor Modeling*, {https://www.epa.gov/scram/air-pollutant-receptor-modeling}.

U.S. Environmental Protection Agency (1989), *Report to Congress on indoor air quality: Assessment and Control of Indoor Air Pollution*, tech. rep.

Vianaa, M., T.A. Kuhlbuschb, X. Querola, A. Alastueya, R. Harrisonc, P.K. Hopked, W. Winiwartere, M. Valliusg, S. Szidath, A.S. Prvôti, C. Hueglinj, H. Bloemenk, P. Wåhlinl, R. Vecchim, A.I. Mirandan, A. Kasper-Gieblo, W. Maenhautp, and R Hitzenbergerq (2008), "Source apportionment of particulate matter in Europe : A review of methods and results", *Journal of Aerosol Science*, 39, 827–849.

Vincent, D., I. Annesi, B. Festy, and J. Lambrozo (1997), "Ventilation System, Indoor Air Quality, and Health Outcomes in Parisian Modern Office Workers", *Environmental Research*, 75, 2, pp. 100-112, ISSN: 0013-9351, DOI: 10.1006/enrs.1997.3764.

Vincent, J. (2007), "Aerosol Sampling. Science, Standards, Instrumentation and Applications", *Direct-reading Aerosol Sampling Instruments* (Mar. 2007), pp. 489-513, DOI: 10.1002/9780470060230.ch10.

Wallace, L., H. Mitchell, G. OĆonnor, L. Neas, M. Lippmann, Meyer Kattan, Jane Koenig, James Stout, Ben Vaughn, Dennis Wallace, Michelle Walter, Ken Adams, and Lee-Jane Liu (2003), "Particle Concentrations in Inner-City Homes of Children with Asthma: The Effect of Smoking, Cooking, and Outdoor Pollution", *Environmental health perspectives*, 111 (Aug. 2003), pp. 1265-1272, DOI: 10.1289/ehp.6135.

Wallace, L. A., S. J. Emmerich, and C. Howard-Reed (2004), "Source Strengths of Ultrafine and Fine Particles Due to Cooking with a Gas Stove", *Environmental Science & Technology*, 38, 8, PMID: 15116834, pp. 2304-2311, DOI: 10.1021/es0306260.

Wei, W., O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little, and C. Mandin (2019), "Machine learning and statistical models for predicting indoor air quality", *Indoor Air*, 29, 5, pp. 704-726, DOI: 10.1111/ina.12580.

Welsh, M. and S. Begg (2016), "What have we learned Insights from a decade of bias research", *The APPEA Journal*, pp. 435-450.

Wensing, M, G. Pinz, M Bednarek, T. Schripp, E Uhde, and T. Salthammer (2006), "Particle Measurement of Hardcopy Devices", *Proc Int Conf Healthy Build*, 2 (Jan. 2006), pp. 4-8.

WHO-Europe (2000), *Air quality guidelines for Europe*, tech. rep.

Williams, C. and M. Seeger (2001), "Using the Nystr¨om method to speed up kernel machines", *Advances in Neural Information Processing Systems*, 682–688.

Willis, R.D. (2000), *Final Report, vol. 5. Workshop on UNMIX and PMF as Applied to PM2.5.* Tech. rep., U.S EPA.

Wolff, M. (2003), "Apports de l'analyse géométrique des données pour la modélisation de l'activité", in *Formalismes de modélisation pour l'analyse du travail et l'ergonomie*, Presses Universitaires de France.

Wolkoff, P. (2013), "Indoor air pollutants in office environments: Assessment of comfort, health, and performance", *International Journal of Hygiene and Environmental Health*, 216 (July 2013), pp. 371-394, DOI: 10.1016/j.ijheh.2012.08.001.

Wolkoff, P. and G. Nielsen (2010), "Non-cancer effects of formaldehyde and relevance for setting an indoor air guideline", *Environment International*, 36 (Oct. 2010), pp. 788-799, DOI: 10.1016/j.envint.2010.05.012.

Yakovleva, E., P. K. Hopke, and L. Wallace (1999), "Receptor Modeling Assessment of Particle Total Exposure Assessment Methodology Data", *Environmental Science & Technology*, 33, 20, pp. 3645-3652, DOI: 10.1021/es981122i.

Yao, M. and B. Zhao (2017), "Window opening behavior of occupants in residential buildings in Beijing", *Building and Environment*, 124, pp. 441-449.

Yi, T., B. Pratim, E. Sotiris, and J. Jin (1990), "Receptor Modeling for Contaminant Particle Source Apportionment in Clean Rooms", *Aerosol Science and Technology*, 12, 4, pp. 805-812, DOI: 10.1080/02786829008959394.

Yu, Y., K. Diamantaras, T. McKelvey, and S.Y. Kung (2018), "Chapter 6 - Kernel Subspace Learning for Pattern Classification", in *Adaptive Learning Methods for Nonlinear System Modeling*, ed. by Danilo Comminiello Príncipe and Jos C., Butterworth-Heinemann, pp. 127-147.

Zdunek, R. and A. Cichocki (2007), "Nonnegative matrix factorization with constrained second-order optimization", *Signal Processing*, 87 (Aug. 2007), pp. 1904-1916, DOI: 10.1016/j.sigpro.2007.01.024.

Zeghnoun, A., F. Dor, and A. Grégoire (2010), *Description du budget espace temps et estimation de l'exposition de la population française dans son logement.* Tech. rep., Institut de veille sanitaire, Observatoire de la qualit de l'air intrieur.

Zhang, L., R. Lunn, G. Jahnke, D. Spencer, G. S, S. Atwood, Carter G, Ewens A, Greenwood D, Ratcliffe J, Desrosiers T, Haseman J, Jameson CW, Darden E, Saunders T, Riojas JC, Susan Dakin, McMartin KE, Akbar-Khanzadeh F, and Elmore SA (2010), *Final Report on Carcinogens Background Document for Formaldehyde*, tech. rep.

Zhang, M., S. Zhang, G. Feng, H. Su, F. Zhu, M.g Ren, and Z. Cai (2017), "Indoor airborne particle sources and outdoor haze days effect in urban office areas in Guangzhou", *Environmental Research*, 154, pp. 60-65, ISSN: 0013-9351, DOI: 10.1016/j.envres.2016.12.021.

Zhao, S., Y. Yu, D. Yin, and J. He (2017), "Effective Density of Submicron Aerosol Particles in a Typical Valley City, Western China", *Aerosol and Air Quality Research*, 17 (Jan. 2017), 1–13, DOI: 10.4209/aaqr.2015.11.0641.

Zhao, W., P. K. Hopke, E. W. Gelfand, and N. Rabinovitch (2007), "Use of an expanded receptor model for personal exposure analysis in schoolchildren with asthma", English, *Atmospheric Environment*, 41, 19, pp. 4084-4096, DOI: 10.1016/j.atmosenv.2007.01.037.

Zhao, Y., R. Qiu, X. Zhao, and B. Wang (2016), "Speech enhancement method based on low-rank approximation in a reproducing kernel Hilbert space", *Applied Acoustics*, 112, pp. 79-83, ISSN: 0003-682X, DOI: 10.1016/j.apacoust.2016.05.008.