Cover page

Exam information
NFYK15800E - Climate Change Thesis 30 ECTS, Niels Bohr Institute - Kontrakt:117255 (Lilja Dahl)

Handed in by
Lilja Dahl
kbc751@alumni.ku.dk

Exam administrators
Studenterservice 35 33 35 33
studenterservice@science.ku.dk

Assessors
Anders Svensson
Examiner
as@nbi.ku.dk
+4535320616

Claus Nordstrøm
Co-examiner
cno@envs.au.dk

Hand-in information

Title, english: Source-Apportionment of Non-Methane Volatile Organic Compounds, Halogenated Species and Non-CO2 Greenhouse Gases at Mt. Cimone (Italy) by applying Positive Matrix Factorization with a Lifetime Correction Method

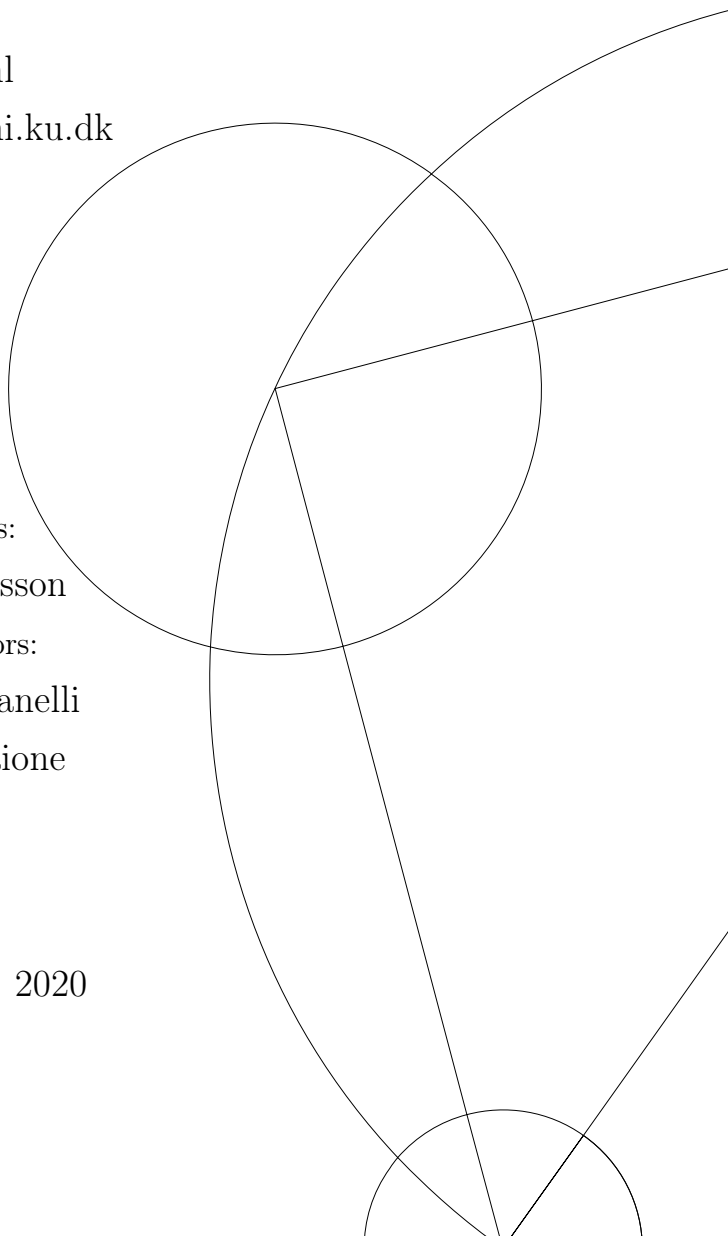
The sworn statement: Yes

UNIVERSITY OF COPENHAGEN

Master of Science in Climate Change

Source-Apportionment of Non-Methane Volatile Organic Compounds, Halogenated Species and Non-CO_2 Greenhouse Gases at Mt. Cimone (Italy) by applying Positive Matrix Factorization with a Lifetime Correction Method

Lilja Dahl
KBC751@alumni.ku.dk



Supervisors:
Anders Svensson
Co-supervisors:
Paolo Cristofanelli
Michela Maione

October 20th, 2020

DEDICATION

I dedicate this thesis work to my beloved father Tórmóur Dahl who passed away when I was a child. He was an environmentalist and his passion and understanding of nature is now very vibrant in me. He taught me how to live and embrace every element of nature. "Now I am walking in your footsteps and I am truly thankful for you showing me the way".

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my main supervisor Anders Svensson for always being available, patient, and helpful.

A special thanks to my thesis supervisor, Paolo Cristofanelli for welcoming me in his research team at the National Research Council of Italy, Institute of Atmospheric Science and Climate. Thanks for supporting and guiding me throughout the process and always being available. A big thanks to the whole research team: Paolo, Cristofanelli, Jgor Arduini, Marco Paglione, Stefania Gilardoni, and Michela Maione, for the exceptional good supervision, collaboration, and continuous support throughout the duration of the thesis project. Their encouragement, patience and invaluable advises helped and motivated me in all the phases of this thesis.

My sincere thanks goes to Professor Alessandro Bigi for introducing me to the project without being a part of it and hosting me at the University of Modena and Reggio Emilia, Departement of Enviromental Engineering. A special thanks to my friend and future collegue Ohad Zivan for proofreading my whole thesis and for giving valuable comments.

Lastly, I owe a special thanks to my beloved partner Alexis and daughter Eira, for supporting me endlessly throughout this thesis work, in particular during the corona lock-down. I am truly grateful to share my journey with you both. Many thanks for for teaching and helping me with latex programming and all your valuable comments. Thank you, my dear mother Wenche, for all the support and guidance you have given me throughout my journey in life. You are my true motivator when I need it the most.

ABSTRACT

Receptor models are applied to atmospheric measurements to discover hidden patterns of variability and identifying relevant sources of pollutants. A source-apportionment investigation was carried out on non-methane volatile organic compounds (NMVOCs), halogenated species and non-CO_2 greenhouse gases (GHGs) at Mt. Cimone in the period 2015-2018, using positive matrix factorization (PMF). NMVOCs are important precursors of tropospheric ozone formation and many halogens contribute to the stratospheric ozone depletion and the greenhouse effect.

In particular, the impact of photochemistry on NMVOCs seasonal cycle is considered in this thesis, and the added value of this work is the application and evaluation of the PMF tool by coupling a lifetime correction method with the PMF model. The lifetime correction proved to be a valid method to apply on data, in order to scale the reactive NMVOCs as a function of their rate constant. Moreover, a consolidated framework of the data pretreatment and analysis process was established, which is essential for performing a sound source apportionment study.

Eight factors were identified during Summer season and seven factors were identified during Winter season. The eight factors characteristics are determined to be: (1) vehicle exhaust; (2) gasoline evaporation; (3) Liquified petroleum gas; (4) solvent evaporation; (5) industrial solvents; (6) chlorinated solvents; (7) octane gasoline; and (8) halogenated species and non-CO_2 GHGs. The last factor (8) is not related to emission sources, but is rather explaining the variability of long-lived halogenated species and non-CO_2 GHGs on a continental scale.

Explorative analysis using cluster analysis and principal component analysis (PCA) gave a better understanding of species variability and correlation among them which was useful for the validation and interpretation of PMF results. Results from PMF and cluster analysis demonstrated similar classification of species. However, PCA was unsuccessful to explain meaningful data variance and was sensitive towards "problematic species".

Overall, the PMF results indicate seven potential emission sources, although further analysis and comparison of PMF results with sophisticated models are needed, in order to evaluate if the obtained factors are robust.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CMN	Mt. Cimone
ARPA	Agenzia Regionale per la Protezione dell'Ambiente
BTEX	Benzene, Toluene, Ethylbenzene, and Xylene
TEX	Toluene, Ethylbenzene and Xylenes
EX	ethylbenzene and xylenes
E/X	Ethylbenzene to Xylene ratio
NMVOCs	Non-Methane Volatile Organic Compounds
VOCs	Volatile Organic Compounds
BVOCs	Biogenic Volatile Organic Compounds
PCA	Principle Component Analysis
PCs	Principal Components
PMF	Positive Matrix Factorization
COS	Carbonyl Sulphide
TCE	Trichloroethylene
PCE	Perchloroethylene
OH	Hydroxyl radical
MP	Montreal Protocol
LPG	Liquified Petroleum Gas

Part I

PROJECT DESCRIPTION

1

INTRODUCTION

This introductory chapter first aim to describe climate change and introduce the reader to the challenges of increasing anthropogenic emissions. It follows an introduction to the study area and its local challenges in curbing anthropogenic emissions and how receptor models can be used to apportion atmospheric species to individual sources. Ultimately a description of the intent of this thesis and finally the structure of the thesis is described.

1.1 Climate change

The climate system is very sensitive to external forcing of greenhouse gases (GHGs) of either natural or anthropogenic origin, pushing climate components out of equilibrium and inducing global warming. The rapid increase of warming since 1970s is higher than any other period over the past 800ka and cannot be explained solely by the natural variability (Milankovitch cycles, solar luminosity, atmospheric composition among others) (Masson-Delmotte et al., 2013). The accelerated warming is caused by the the anthropogenic buildup of GHGs such as CO_2 (417.07 ppm), CH_4 (1873.7 ppb), and N_2O (332.6 ppb)(Tans and Dlugokencky, 2020).

Humans have therefore become the new forcing agent on Earth since industrial revolution, entering a new geological epoch called Anthropocene (Zalasiewicz* et al., 2010). Along with the increase of human population, cities are becoming denser and the public services are also increasing, leading to a major source of GHG emission and air pollution. As a result, the atmospheric composition is changing significantly leading in response to climate change and higher oxidation capacity.

Not long ago, the Paris Agreement initiated a global effort on climate mitigation and adaption in order to keep the global average temperature lower than a 1.5°C temperature increase since preindustrial level. However, the future projections show that in order to stay under the 1.5 °C, a global negative (GHGs) emission pathway has to be achieved (Masson-Delmotte et al., 2018). Hence, there is an urgent need to actively reduce GHG emissions and air pollution nationally and establish mitigation strategies, possibly in connection with air pollution policies. The longer we wait to mitigate, a more rapid decrease of emissions is waiting for us and as time is running short, it implies even more negative emission technology shall be implemented (Masson-Delmotte et al., 2018).

Existing literature, claims that 70 % of anthropogenic GHG emission is coming from urban areas (Hopkins et al., 2016).

1.2 Project background

In Northern Italy, between Apennines in Emilia Romagna and the northern alps, it is located the Po Basin, a highly populated and industrialized area. It is considered to be a 'mega source' in Europe of anthropogenic emissions, facing the challenge of curbing emissions of GHGs and atmospheric pollutants, such as nitrogen oxides (NOx) and volatile organic compounds (VOCs). NOx and VOCs are responsible for the formation of tropospheric ozone, which besides being harmful for human health and ecosystems, it is also a strong GHG (Seinfeld and Pandis, 2016). VOCs emission sources are both natural and anthropogenic. Po Basin is an important agricultural area and a source of natural biogenic VOCs (BVOCs). However, it is also a highly anthropized area and the dominant non-methane VOCs (NMVOCs) sources are fossil fuel burning, vehicle emissions and solvent use (Cristofanelli et al., 2017).

Another important environmental issue is related to halogens and their contribution to stratospheric ozone depletion, meanwhile also being a powerful GHGs in the troposphere. With the successful implementation of the UNEP Montreal Protocol (MP) on ozone depleting substances, a significant reduction of some synthetic halogens is achieved. Although the impacts on stratospheric ozone by some halogens are not regulated by the MP, such as CH_3Cl. Their global natural source is predominant and there is therefore no commitment from Government to report their national inventories of anthropogenic sources (Cristofanelli et al., 2020).

At the regional scale, significant uncertainties still remains for identifying anthropogenic sources of NMVOCs and GHGs, where the effect of photochemical processes of NMVOCs needs to be considered.

This study is in collaboration with the National Research Council of Italy, Institute of Atmospheric Science and Climate (CNR-ISAC) and the University of Urbino, Department of Pure and Applied Sciences (UniUrb, DiSPeA). They carry out continuous VOC and GHG measurements at observatory "O. Vittori", situated at the Mt. Cimone (referred to as CMN), overlooking the Po Basin.

To the attempt of shading light over the underlying chemistry and identify relevant sources of NMVOCs and GHGs at CMN, modern multivariate statistical methods are applied on the long-term/high-frequency time series of atmospheric species obtained from CMN.

Source apportionment of NMVOCs and GHGs is the practice of studying and deriving information of the species to identify, quantify, and characterize their emission source contribution to the ambient air (Belis et al., 2019). There are several approaches and models developed within the discipline of source apportionment among which receptor models (RM) is a class of multivariate statistical analysis. It uses ambient atmospheric concentration measured at a certain site called receptor (CMN) to identify and apportion its sources without taking into account atmospheric transport or dispersion characteristics (e.g meteorology and topography) like in source models. Therefore the receptor model is a relative simple model, although it requires expertise in species and source characteristics as well as a solid knowledge about the applied model.

The top-down approach is necessary for establishing rigorous regional air quality control and lower the environmental risk of human exposure. There are certain species that are particular important for source apportionment studies to prevent human health and climate impacts, such as identifying ozone precursors (Belis et al., 2019). The fundamental principle of receptor to source relationship, is that mass conservation can be assumed. The receptor method is constructed from a mathematical framework based on the equation of continuity, although different RMs include different approaches to solve this equation and quantify source contribution.

$$X_{ij} = \sum_{k}^{p} g_{ik} f_{kj} + e_{ij} \tag{1}$$

The obtained model are estimates from equation (1) where X_{ij} is the input data matrix from measured concentrations at receptor site. Moreover, it is the sum of contributions from different sources (Belis et al., 2019; Comero, Capitani, and Gawlik, 2009). From the variations of concentration patterns in time and space, and with the the right number of factors k, the model gives estimates of the underlying chemistry and sources (Hopke, 2000; Paatero, 1999). Source weight g_{ik} is the sample contribution of source k and source profile f_{kj}, represents a phenomenon present in data, also called chemical profiles and is the fraction of species coming from source k. The residuals e_{ij} is all that is not modelled (Olivieri et al., 2015).

Positive matrix factorization (PMF) is a robust RM used to solve g and f in equation (1). Moreover, the method is extensively used and recognized by EU as one of the reference techniques for source apportionment studies of atmospheric pollutants (Paatero, 1999; Paatero and Tapper, 1994). Traditionally, PMF analysis is used to identify source contribution of aerosols, including secondary aerosols that are not identified by source models (Contini et al., 2016).

Applying PMF on NMVOCs is challenging because the atmospheric lifetime of NMVOCs are generally much shorter than for aerosols and halogens (Sauvage et al., 2009; Yuan et al., 2012). A recent study by He et al., 2019, takes into account the atmospheric lifetime of VOCs prior to PMF, and shows a higher source contribution from the reactive species than without lifetime correction. This is the first attempt to perform PMF on an extensive dataset at a remote site in Italy, taking into account the diverse lifetime of the atmospheric species by applying lifetime correction method.

A study carried out at CMN, characterizes anthropogenic sources of NMVOCs by applying PCA and relates the obtained factors to different air mass ages based on the atmospheric lifetime (Lo Vullo et al., 2016). When classifying NMVOCs source categories, it is of great importance to also consider the photochemical processes. However, this work uses PCA as an exploratory method only, to screen patterns co-variability before applying PMF. Therefore, the aim is not to identify source contributions with PCA, but rather to obtain a preliminary understanding of the data variability. According to Hopke, Jaffe, et al., 2020, PCA is deprecated and should only be used as an exploratory tool.

The application of the receptor models are advanced and occasionally biased, with results strongly depending on the initial set-up of the statistical model (i.e. choice of the species to

be included in the analysis, choosing the number of emission factors, choice of the temporal windows of time series, and choice of the diagnostic metrics). A standard framework is established of data pre-processing, lifetime correction method and the initial-setup of the RM, to be used at CMN for the operative source apportionment studies. Finally, the added value of this thesis work is the application and evaluation of the PMF tool with particular emphasis on NMVOCs lifetime correction.

1.3 Purpose of the study

The overall objective of this study is to perform source-apportionment investigation of NMVOCs, halogenated species, and non-CO_2 GHGs by applying PMF to the long-term/high-frequency time series measured at the remote location CMN. For this purpose, it is followed a multiple approach combining different techniques and performing a sensitivity study on the initial conditions and set-up of the model used. This combined approach will allow an evaluation of the results.

Specific efforts will be devoted to evaluate the effectiveness of applying the different atmospheric lifetime of NMVOCs prior to the PMF analysis.

The outline of this research aims to address the following research question:

- What is the source characterization and contribution of NMVOCs, halogens and non-CO_2 GHGs at CMN?
- How do the emissions from Po basin contribute to the NMVOCs observed at CMN site and what is the effect of using a lifetime correction method prior to PMF?

1.4 Thesis structure

The scope of the study is divided in six parts: (I) Project description, (II) data validation, (III) preliminary analysis, (IV) PMF analysis, (V) discussion and conclusion, and finally (VI) the appendices. The thesis structure follows an introduction to the topic, motivation statement, and a theoretical background of NMVOCs, halogenated species and non-CO_2 GHGs. It follows in part two, an introduction to measurement, data validation, and data pretreatment. In part three, a lifetime correction method is applied on NMVOCs before undergoing a preliminary analysis using cluster Analysis and PCA as the main analysis methods. The methodology structure is illustrated in Figure 1. First step (1) is to apply the lifetime correction method. Second step (2a) is to perform cluster analysis in synergy with (2b) PCA, and (2c) a comparison analysis of the results. Third step (3) is to perform PMF analysis and use (3a, 3b) preliminary analysis results to support the interpretation and identification of PMF factors explained in part four. Lastly, the fifth part is a discussion of the obtained results and interpretation followed by a conclusion and future work. The sixth part is the appendices and includes the R script framework and figures from analysis.

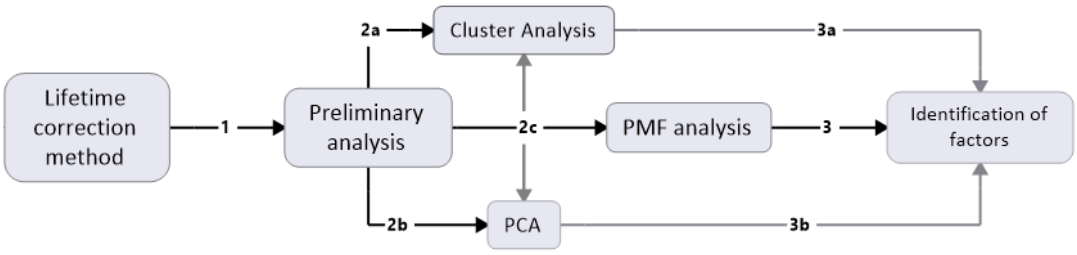


Figure 1: Methodology structure.

2

THEORETICAL BACKGROUND

The present chapter aims to give the reader a general understanding of the concepts, main species characteristics and sources. Moreover, an overview of the species will be useful in the later chapters, when discussing the results and species source contribution.

2.1 Oxidation capacity of the atmosphere

The troposphere is a very reactive medium, extending 10-15 km in altitude and features a high oxidative capacity to self-clean and prevent trace elements accumulation (Seinfeld and Pandis, 2016). Even in the natural atmosphere where pollutants are depleted, the oxidation processes continue to transform, from primary biogenic volatile organic compounds (BVOCs) to secondary products. The atmospheric oxidation capacity mainly depends on four abundant oxidizing species: Hydroxyl radical (OH), ozone (O_3), hydrogen peroxide (H_2O_2), and nitrate radical (NO_3). Out of these, OH radical is the most important during daytime and NO_3 during nighttime (Seinfeld and Pandis, 2016).

The hydroxyl radical (OH) is like a vacuum cleaner in the troposphere, able to clean and reduce the concentration of toxic VOCs species in the atmosphere. Common VOCs from vehicle emissions are benzene, toluene, ethylbenzene, and xylenes (BTEX) and are monitored in most urban areas.

The global average concentration of OH is low in the troposphere, indicating that it is being destroyed rapidly and has an average lifetime of < 1 second (Seinfeld and Pandis, 2016). Despite the short lifetime, OH influence in the troposphere is enormous. Without OH radical, the concentration of toxic species will increase to very unimaginable levels.

Just after sunrise, OH radical concentration starts to peak and will reach a maximum at midday. OH is formed by photodissociation of O_3 that occurs during daytime by the absorption of photons in the troposphere. Photodissociation of O_3 leads to one free oxygen and its primary source is water vapour to form OH radical. OH formation therefore depends on the amount of water vapour, and OH levels tend to decrease with altitude as the temperature of air becomes colder and drier. Moreover, only a small part of the free oxygen is capable to react with water and produce OH radical, while most is being reacted back to O_3 (Seinfeld and Pandis, 2016). This results in a stable concentration that limits OH concentration to increase over a certain threshold. We can therefore rely on one measurement for the whole European domain, where the OH concentration during summer is $1.32 \cdot 10^6$ *molecules* cm^{-3} (Meszaros, Haszpra, and Gelencser, 2004).

The ongoing increase of air pollution and GHGs emission from human activities has led to an increase in atmospheric oxidation capacity since preindustrial time (Monks, 2005). We can therefore expect a more aggressive atmosphere in the near future causing more pollution. In fact, Saiz-Lopez et al., 2017 observed an increased oxidation capacity in the urban city of Madrid in Spain recently. The observation shows a NO_x reduction while the major oxidation species O_3, OH, and NO_3 have increased significantly and so have the overall oxidation capacity in the urban atmosphere (Saiz-Lopez et al., 2017). An increase in O_3 concentration in the troposphere is not only very damaging for human health and vegetation, but it is also a very strong greenhouse gas (GHG). In the lower troposphere, O_3 is a secondary pollutant which means it is not emitted directly from a source, but requires chemical precursor like VOCs and NO_x. Hence, VOCs influence the climate indirectly by increasing O_3 levels through a set of complex reaction mechanisms, which include NO_2 photodissociation followed by O_3 formation (see section 2.2). There exists a nonlinear relationship between VOCs and NO_x, and ozone, meaning that at a given level of VOCs exists a NO_x concentration, at which maximum ozone is produced, as shown in Figure 2. The hyperbolic curve represents a constant concentration of ozone, while the linear (red) curve indicate a linear increase of ozone concentration (Monks, 2005). This makes it challenges to establishing effective air quality control of the toxic pollutant ozone in an urban city (Monks, 2005). However, the favourable mitigation strategy is to decrease both NO_x and VOCs significantly by controlling, e.g., traffic emissions. The ozone isopleth diagram also suggests that if we want to reduce high ozone concentration in a high NO_x polluted environment, it will be more efficient to reduce VOCs concentration (VOCs sensitive region). Likewise, in a high VOCs environment, reducing NO_x will potentially reduce ozone concentration (NO_x sensitive region) (Monks, 2005). NO_x and VOCs sensitive regions for mitigating ozone production are indicated with red circles in the isopleth diagram in Figure 2.

Ultimately, the reactivity in the atmosphere is modified simultaneously with the changes in the chemical composition of the atmosphere, and climate change is the response to the composition change. Since the atmospheric reactivity has the potential to increase in the future together with climate change, there is a need to better understand the atmospheric chemistry and advise policy makers with the right mitigation strategies for the future. It can be necessary to prevent erosion and degradation of materials, damage of agriculture and health. As it may turn out to be expensive and interrupt the overall productive world, food supply and the health care system just to mention a few.

2.2 Non-methane volatile organic compounds

Non methane volatile compounds (NMVOCs) is a group of compounds that are very reactive (with OH) and volatile. Unlike methane, which has intrinsic chemical stability and stable configuration that it barely reacts with the OH radical. This is reflected on methane relative long lifetime of about a decade (9.6 years) until it eventually is oxidized by the OH radical. Consequently, methane has a much higher mixing ratio in the atmosphere than NMVOCs (Seinfeld and Pandis, 2016).

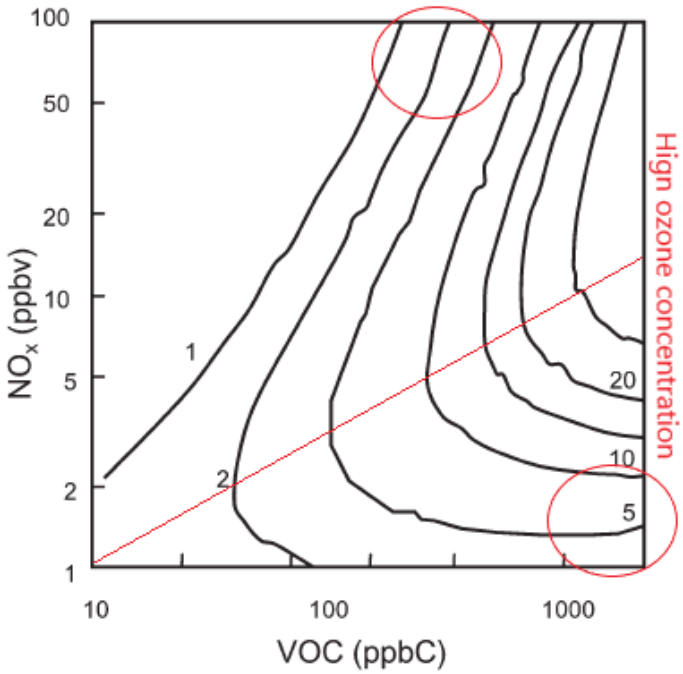


Figure 2: Ozone isopleth diagram: The role of NO_x and VOCs in ozone formation. The red circles indicate VOCs sensitive regions (top) and NO_x sensitive regions (bottom right corner) for mitigating ozone production. Figure is modified from Monks, 2005.

NMVOCs oxidation of OH radical in the troposphere initiates a set of oxidation chain reactions that are highly complex and Figure 3 illustrates the general VOC-OH reaction mechanism. NMVOC reacts with OH radical and breaks down to a peroxyradical (RO_2) to form a water molecule. The abstraction of H atom leaves a radical with an unpaired electron that will react instantly with the abundant oxygen molecules in the atmosphere, and continue to react with oxygen until it reaches a lower molecule weight and produces hydroperoxyl radical (HO_2) that recycle the OH radical. Thus, in a highly anthropized area with high NO_x and VOCs concentration, ozone formation is accelerated.

2.2.1 Atmospheric lifetime

Each individual NMVOC vary a lot from each other by their properties, atmospheric lifetime, and sources. The average life history of each NMVOC is also called the averaged lifetime. It explains how far the compounds travel and stay in the atmosphere before being removed, e.g., by the oxidation of OH radical. The 15 NMVOCs indicated in Table 1, have a short atmospheric lifetime in the atmosphere compared with the lifetime of methane, ranging from a few hours to a few days. Anthropogenic NMVOCs in general have a longer lifetime than biogenic VOCs (BVOCs). In particular, benzene, ethyne, and propane are relatively long-lived in the atmosphere lifetime (>7 days) compared to the other NMVOCs and can be transported longer distances with aged air masses. While toluene, ethylbenzene, and xylenes (TEX) have the shortest lifetime and are therefore indicating fresh air masses coming from nearby emission source (Cristofanelli et al., 2017). The transportation scale of the NMVOCs depends on their

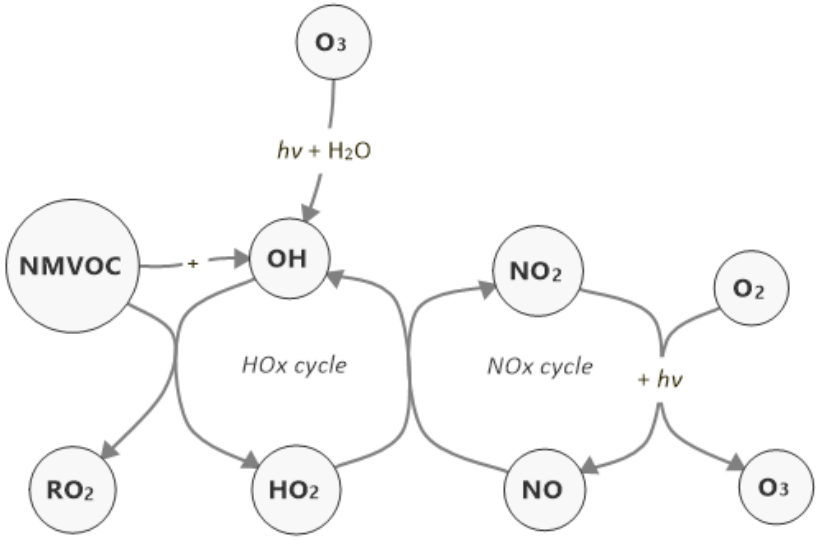


Figure 3: The general reaction mechanism for NMVOCs oxidation of OH radical, modified from Jain et al., 2017.

reactivity and the major atmospheric sink of NMVOCs, which is the reaction with the OH radical.

From a chemical point of view, there is a big difference in NMVOCs lifetime between alkanes and molecules with unsaturated bonds (pi clouds). The latter is more electrophyllic towards OH radical and will easier detach electrons from the unsaturated bond (Seinfeld and Pandis, 2016). In general, the larger the molecule, the more reactive it will become as indicated from the OH rate constants in Table 1 except for isomer compounds e.g. i-butane, i-pentane, i-octane.

2.2.2 Anthropogenic emission sources

The largest sources of NMVOCs globally is biogenic emissions from terrestrial vegetation, where trees are the major contributor to BVOC emissions (Kansal, 2009; Seinfeld and Pandis, 2016). Anthropogenic emissions of NMVOCs sources of biogenic origin include biomass burning and biofuel combustion. Just from biomass burning, substantial amount of hydrocarbons is released to the atmosphere. The main anthropogenic nonbiogenic NMVOCs emission sources are listed in Table 1. Motor vehicles are the main contributor of transportation emissions where alkanes and aromatic compounds are the most abundant species emitted. These aromatic compounds are emitted by incomplete combustion or evaporative emission of fuel such as liquified petroleum gas (LPG) or gasoline. These compounds are therefore prevalent in the urban atmosphere. One important hydrocarbon is benzene, which is a prototype of all aromatic substances and very stable compare to the other NMVOCs. It is highly carcinogenic and has the potential to pose a serious health risk to the surrounded population. In particular, in the Po Basin where the air masses are not so energetic and the mixing of pollutants is not

Table 1: Reaction rate constant, lifetime and sources of 15 NMVOCs measured at CMN.

NMVOCs	kOH ($10^{-12} cm^3 molecules^{-1}$)	lifetime (days)	Main emission sources
Alkanes			
propane	1.09	11	LPG; gasoline evaporation
n-butane	2.36	4.9	LPG; gasoline evaporation
i-butane	2.12	5.5	LPG; gasoline evaporation
n-pentane	3.80	3.0	gasoline evaporation
i-pentane	3.6	3.2	gasoline evaporation
n-hexane	5.20		solvent; gasoline evaporation
n-heptane	6.76		gasoline evaporation
n-octane	8.11		gasoline evaporation
i-octane	3.34		gasoline evaporation
Alkynes			
Ethyne	8.20[1]	14	vehicle exhaust
Aromatic			
benzene	1.22	9.5	vehicle exhaust
toluene	5.63	2.1	solvent-use, vehicle exhaust
ethylbenzene	7.0	1.7	solvent-use, vehicle exhaust
o-xylene	13.6	0.9	solvent-use, vehicle exhaust
(m, p)-xylene	19.0[2]	0.6	solvent-use, vehicle exhaust

[1]From Atkinson et al., 2006.
[2]Calculated average of m-xylene and p-xylene rate constants, from Lo Vullo et al., 2015.
The OH rate constants are measured at 298K and from Atkinson and Arey, 2003, and atmosphere lifetimes and emission sources are from Lo Vullo et al., 2016 and Cristofanelli et al., 2017.

efficient, it has the potential to accumulate benzene concentration at ground level. Other hydrocarbons are also emitted from evaporative sources such as gasoline fuel (Cristofanelli et al., 2017).

2.3 Halogenated species

Chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), and hydrofluorocarbons (HFCs) are man-made synthetic compounds for cooling systems. Table 2 provides a list of the more common halogens and their emission sources, which include refrigerants, airconditioner, foam blowing agents, and repellent fluids (Cristofanelli et al., 2017). CFCs are all very stable species in the troposphere and OH radical cannot break them down. As a result, they do not have a tropospheric sink and will persist in the troposphere for a long time (Table 2). For this reason, they can be transported efficiently with airmasses, reaching a homogenous concentration until they eventually diffuse into the stratosphere, where high ultraviolet radiation with a wavelength of 185-210nm (UV-C) photodissociate them and break off the chlorine atom (Seinfeld and Pandis, 2016). Chlorine and bromine are highly reactive in the stratosphere and are the main cause for destruction of stratospheric ozone (Braesicke

et al., 2019). CFCs could be considered an indirect air pollutant, since ozone depletion causes harmful radiation to reach humans which increases the risk of developing skin cancer.

2.3.1 Ozone depletion

Ozone depletion is a permanent decrease of the ozone layer (20-30km) in the stratosphere at all latitudes (Seinfeld and Pandis, 2016). The simplest halogen cycle is presented below, where CFC-11 is photodissociated with UV-C and releases a free chlorine atom that reacts with ozone and the chlorine atom is recycled. The reaction mechanism happens faster than ozone formation, which causes ozone to deplete in the stratosphere:

$$CFCl_3 + hv \rightarrow CFCl_2 + Cl \tag{2}$$

$$Cl + O_3 \rightarrow ClO + O_2 \tag{3}$$

$$ClO + O \rightarrow Cl + O_2 \tag{4}$$

Montreal protocol banned CFCs in 1987 and it is not commercialized any more. It became a huge scientific success in outphasing CFCs and restore the ozone layer. 20 years ago, ozone concentration stopped to decrease in the ozone layer and this shows the responsiveness of introducing CFCs into the atmosphere (Braesicke et al., 2019). However, this ban resulted in the introduction of HCFCs and HFCs as replacement, that are minor ozone depleting substances, but very strong GHGs. They absorb more infrared radiation than CO_2, CH_4, or N_2O, and thereby having a much higher global warming potential (GWP_{100}) in a 100 years period (IPCC et al., 2014). The GHGs effect depends on the interaction between the infrared radiation and the species covalent bond vibration and GWP_{100} is a metric used to assess the radiative forcing in a 100 years period relative to CO_2 (IPCC et al., 2014).

2.4 Emission sources of halogenated species and non-CO_2 GHGs

Major chlorinated solvents used in industries are chloroform ($CHCl_3$), methyl chloroform (CH_3CCl_3), Trichloroethylene (TCE), and perchloroethylene (PCE) also indicated in Table 2.

Some of the halogens are eventually broken down by the OH radical in the troposphere, preventing some chlorine to reach the stratosphere. While halogens with a long lifetime are diffused into the stratosphere, one of the important sources of chlorine to the stratosphere is methylchloride (CH_3Cl) (Seinfeld and Pandis, 2016). Its major emission sources are from natural origin, such as ocean, biomass burning, tropical plants, and salt marshes. Although the global budget of CH_3Cl is not balanced and uncertainty still remains on the missing source of anthropogenic emissions (Cristofanelli et al., 2020; Seinfeld and Pandis, 2016).

Carbonyl sulphide (COS, also written as OCS) does not belong to the family of halogens, but is a sulfur compound. It is a GHG in the troposphere and a aerosol radiative forcer in the

stratosphere. It has a long average lifetime of 7 years and can therefore reach the stratosphere where it photodissociate and oxidizes into SO_2 and converts into a sulfate aerosol (H_2SO_4) (Seinfeld and Pandis, 2016). Its global natural source is mainly from wetland and ocean, while anthropogenic emission sources originating from industry and biofuel (Cristofanelli et al., 2017). Finally, sulfuryl fluoride (SO_2F_2) is both a sulfur and halogen -containing compound and a strong GHG. The global emission source of SO_2F_2 is from fumigation and is used as a replacement of Montreal Protocol out-phasing fumigant methylbromide (CH_3Br) used in agriculture (Mühle et al., 2009).

Table 2: Lifetime and emission sources of 19 halogens and sulfur compounds measured at CMN.

Compounds	Lifetime (years)	Emission sources
CFCs		
CFC-114	190	refrigerant; propellants in medical aerosol;
CFC-115	1020	refrigerant
HFCs		
HFC-32	5.2	refrigerant
HFC-125	28.2	refrigerant
HFC-134a	13.4	refrigerant
HFC-152a	1.5	foam blowing
HFC-365mfc	8.6	foam blowing
HCFCs		
HCFC-22	11.9	refrigerant; air conditioning; extruded polystyrene foam application
HCFC-142b	17.2	blowing agent in extruded polystyrene board stock;
Halogenated compounds		
CH_3Cl	1.5	tropical vegetation; biomass burning; oceans; salt marshes; coal combustion; chemical feedstock; solvent release;
CH_3Br	0.8	fumigant in agriculture
CH_2Cl_2		industrial solvent usage
$CHCl_3$	0.4	industrial solvent usage
CCl_4	26	solvent usage; raw material for chlorinated chemical production; fire extinguisher
CH_3CCl_3	5	industrial solvent usage
TCE		industrial solvent usage; aluminum degreasing[1]
PCE		industrial solvent usage; dry cleaning[2]; feedstock for HFCs manufacturing
Sulfur compounds		
SO_2F_2		fumigant replacement of CH_3Br[3]
COS		wetlands and oceans; tracer of biomass; biofuel; coal and aluminium production

[1,2]From Mohr, 2020.
[3]From Mühle et al., 2009.
Source information and lifetimes are from Cristofanelli et al., 2017; Cristofanelli et al., 2020; IPCC et al., 2014; Seinfeld and Pandis, 2016.

Part II

DATA VALIDATION

3

DATA VALIDATION

This section shows the process done to have a complete dataset before applying any receptor models. This means that a preliminary data analysis was carried out and undergo an extensive data validation. Moreover, a complete framework of data pretreatment is established in the Appendix A (Chapter 12), using the open source programming software R with multiple open source packages. The framework enables anyone to repeat the analysis, thereby reducing any potential mistakes. The analysis has been repeated several times following the scheme in the framework.

In addition, the CMN data measurements brings about multiple challenges, e.g., extracting valuable data and calculating associated instrumental uncertainties. Therefore, the retrieved dataset is undergoing data cleaning, where missing values are replaced, and time series are detrended before performing cluster analysis, PCA, and PMF.

3.1 Measurement site

The high-mountain observatory "O. Vittori" is situated at the peak of CMN (2165m a.s.l.), a remote location in the Northern Apennines, overlooking the Po Basin like a lighthouse in the Mediterranean troposphere (Cristofanelli et al., 2017). The atmospheric composition observations performed at CMN can provide useful hints to investigate the background conditions of the free troposphere as well as the impact of vertical transport of air-masses from the continental planetary boundary layer (PBL). It is therefore a favourable site to study the transport of anthropogenic pollution and climate altering species from the Po Basin and other far emission sources (Lo Vullo et al., 2015). The free troposphere is just above the urban PBL at about 2 km height and occasionally the PBL is observable at CMN, especially during summer when PBL expands as the atmosphere warms up (Seinfeld and Pandis, 2016).

Figure 4, shows the geographical location of CMN and the Po Basin located between the northern Alps and northern Appenines. The observatory at CMN provides continuous high frequency measurements of ozone depleting substances and their substitutes (halogens), VOCs and non-CO_2 GHGs that are analyzed in this work.

Additionally, it is part of the worldwide meteorology organization/global atmosphere watch (WMO/GAW) program and Advanced Global Atmospheric Gases Experiment (AGAGE). It is considered to be one of the most important European climate stations (WMO, 2020).

Figure 4: Geographical location of the Po Basin and observatory O. Vittori at CMN indicated with a circle. The image is created with Google Earth Pro.

3.2 Sampling instrumentation and raw data processing

NMVOCs, halogens, and COS are all measured by a gas chromatography – mass spectrometry (GC-MS) instrument (Agilent 6850-5975), where the MS detector runs a SIM mode (selective ion mode) (Maione et al., 2013). The instrument is collecting continuous measurements every two hours at CMN.

The retrieved data is from the years 2013-2018 and in the time zone UTC+1. Due to technical malfunctions in the GC-MS instrument software, it fails to extract valuable data for all variables during all six years, causing larger gaps of missing values for some species (e.g. i-octane, toluene, and o-xylene contained missing values from 01/01/2013 until 17/07/2014). Therefore, only measurement with associated measurement uncertainty from 01/01/2015 and forward is used in this analysis. This allows for 4 years of continuous high quality measurements for all the considered species.

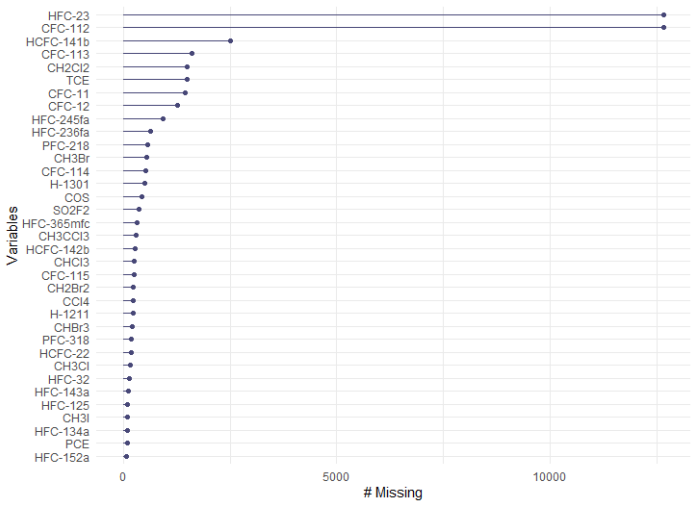
Before initiating data processing and cleaning, data of NMVOCs, halogens, and COS species is divided into two datasets, X_{VOCs} and X_{GHGs}, as explained in Table 3. These datasets (XVOCs and XGHGs,) will be pretreated separately.

Table 3: Data matrices and information

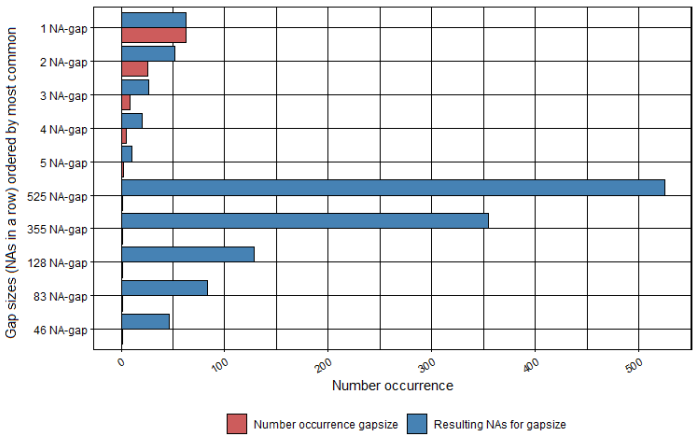
Data matrix	Species	Raw data dim	Final data dim
X_{VOCs}	NMOVCs	18756 x 15 species	12679 x 15 species
X_{GHGs}	Halogens and COS	18756 obs. x 35 species	11672 obs x 19 species

3.3 Assessing species

As illustrated in Figure 5 (a), some species hold a very large fraction of missing values such as HFC-23 and CFC-112 and are removed accordingly. Furthermore, a few species also have large gaps of continuous missing values and large data gaps cannot be handled using artificial data and are therefore cut-off from the time series. For example, in Figure 5 (b), TCE have a large gap that is related to the last period of year 2018 and the period is therefore deleted for all species in X_{GHGs} dataset. This will not affect the PMF results, as PMF analyzes regardless of the time periods given in data.



(a)



(b)

Figure 5: Number of missing values from 2015-2018 of (a) all GHGs and (b) the occurrence of consecutive missing values (NA-gap size) of TCE. The missing values are calculated after removing rows that contain more than 75% missing values (more details on this in section 3.5).

In addition, species that are currently out-phasing as a result of the Montreal Protocol and its amendments, such as first generation of CFCs, are not considered in this study as they are no longer used or emitted, at least in the European domain. Although a persistent increase of CFC-11 is observed globally and is associated with unreported emissions (Montzka et al., 2018). Moreover, species with low emissions in the regional domain or high variability as a

results of poor precision, are excluded from the analysis.
A consolidated list of halogens have been evaluated for this study based on the focus of this thesis (Jgor Arduini and Michela Maione, personal communications). This include species which emission source is already well represented by other species, species that are out-phasing or species where no emission is expected. However, exclusion of species might cause loss of relevant information as they can be varying with other species and help identifying the specific source. Therefore, some species with identical sources are included, as it will benefit the interpretation of the obtained factors from PMF model.
Some species that are strongly affected by natural sources such as COS, CH_3Br, and CH_3Cl, remained in this analysis to emphasize species source contribution from specific natural sources (i.e. oceanic emissions). The various reasons for excluding certain halogens from X_{GHGs} is given in Table 4. No NMVOCs species were deleted from the X_{VOCs}.

Table 4: Halogens that were removed from dataset and their reasons.

Halogens and GHGs	Reasons of exclusion
HCFC-141b	poor precision
CFC-112	Montreal Protoca - no emission expected
CFC-11	Montreal Protocal - no emission expected
CFC-12	Montreal Protocal - no emission expected
HFC-143a (redundant)	represented by HFC-125 and HFC-32
HFC-227 (redundant)	represented by HFC-365
HFC-236fa (redundant)	represented by HFC-365
HFC-245fa (redundant)	represented by HFC-365
H-1211	no emissions expected
H-1301	no emissions expected
CH_3I	mainly biogenic -marine- origin
CH_2Br_2	mainly biogenic -marine- origin
$CHBr_3$	mainly biogenic -marine- origin
PFC-218	Montreal Protocol - no emissions expected
PFC-318	Montreal Protocolt - no emissions expected

3.4 Time series plots

The diurnal time series plots of raw X data gives a 'zigzag' pattern as seen in Figure 6 (a) and (b) of ethylbenzene and (m,p)-xylene, respectively. The sampling time is not fixed over the whole period, but is anyhow regularly acquired every two hours and both X_{VOCs} and X_{GHGs} have the same timestamp. In order to merge and synchronize X_{VOCs} and X_{GHGs} datasets due to the different data coverage for different hours, each data set is averaged on a 2-hour moving interval, by applying at time average algorithm from openair package. The diurnal variability of ethylbenzene and (m,p)-xylene after applying the time average algorithm is given in Figure 6 (c) and (d).

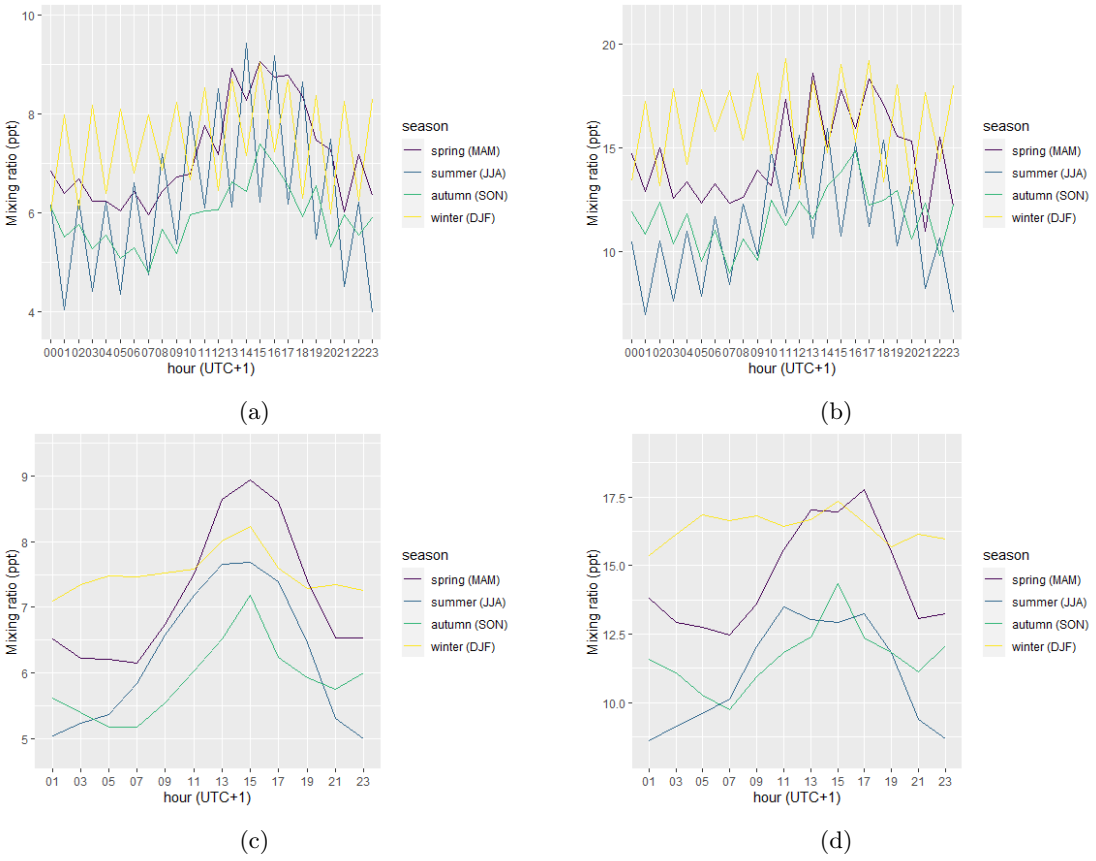


Figure 6: Diurnal variability of ethylbenzene and (m,p)-xylene at CMN. Diurnal plots of (a) ethylbenzene and (b) (m,p)-xylene are based on raw data, while (c) and (d) have applied a 2 hour time average algorithm.

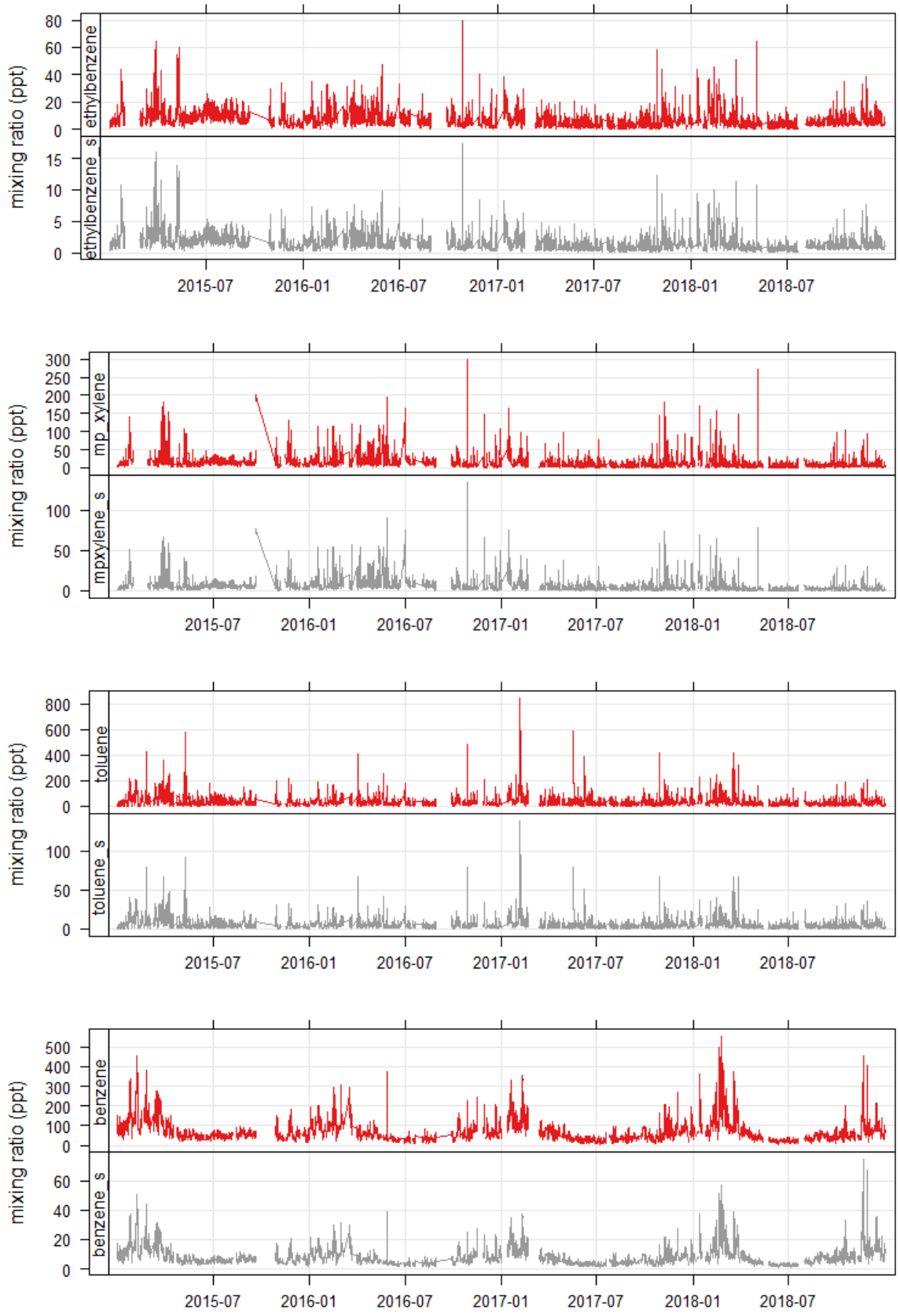


Figure 7: Time series plots of ethylbenzene, (m,p)-xylene, toluene and benzene in red, with their associated uncertainties indicated in grey. Time series are plotted based on raw data.

Time series of NMVOCs species are plotted to explore their variability and evaluate potential problematic peaks that might be challenging in the preliminary analysis. Time series of ethylbenzene, (m,p)-xylenes, toluene, and benzene with their associated uncertainties are illustrated in Figure 7. Winter season is more representative of the difference in emission source (both spatial and temporal), due to having more conservative conditions during the transport of air masses to receptor. While Summer season shows a reduced variability compared to winter. This is consistent with the higher impact of OH removal on NMVOCs and thereby a reduced variability related to different air/mass transport regimes. The comparison of both seasons gives a greater level of understanding of the role OH radical has on the budget of NMVOCs in the troposphere as demonstrated both in the time series plots and Table 5,6.

The more reactive NMVOCs species such as (m,p)-xylene does not have the same seasonal variability as seen in benzene. This is because of the different OH reactivity, where (m,p)-xylene OH rate constant is ≈ 15 times higher than that of benzene.

3.5 Missing data filling

Missing data values cannot be included in the PCA and PMF model and therefore the missing values (NA's) in X_{VOCs} and X_{GHGs} datasets must be evaluated. To begin with, every row in the data matrix that holds more than 75 % missing values across species is removed. Thereafter, X_{GHGs} contain 5292 missing values and X_{VOCs} contain 4532 missing values to be replaced by estimated values. The process of filling in the missing values includes two steps for both datasets. Firstly, the consecutive missing values that last less than 12 hours are filled in by linear interpolation from "imputeTS" package in R. Secondly, for large missing gaps that last longer than 12 hours, seasonality is also considered. The function "na.interp" from the Forecast package in R, is used to fill in missing values by using Seasonal Decomposition of Time Series by Loess (STL). The algorithm uses linear interpolation for time series that do not have a significant seasonal trend (this is true for many halogens) and a periodic STL decomposition for time series with strong seasonality. It is an easy way to cope with the data filling in one single step although some higher frequency variability for the filled period will be lost. An example of data filling is shown in Figure 8 of COS, indicated in red.

Some species have a huge gapsize with many missing values that cannot be replaced by estimates based on STL. Therefore, as mentioned earlier, certain time periods with large gaps of missing values are removed from all time series (e.g. for TCE).

3.6 Detrending time series

Long-term trends are removed from every time series in X_{VOCs} and X_{GHGs} in order to detect patterns that are otherwise masked by the trends. A preliminary cluster and PCA analysis was performed on data matrix X, showing that species grouped close together were mainly related due to an increasing or decreasing trend. The results explained little data variance and in particular halogens were dominated by their trends, as many of them are currently

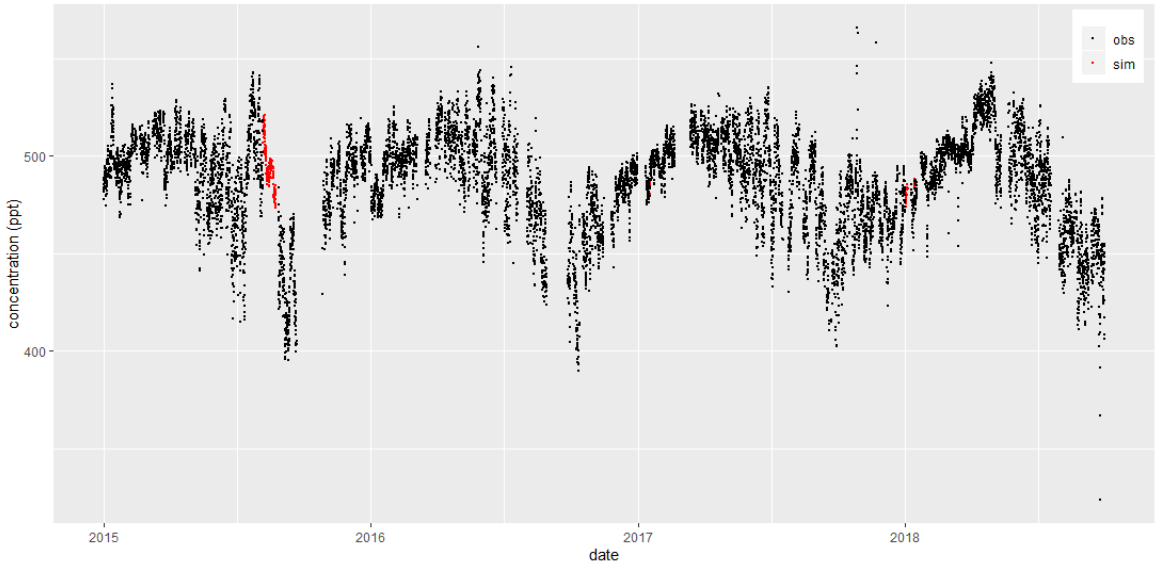


Figure 8: Data filling of COS, where red points indicated simulated values and black points refers to the original data measurements.

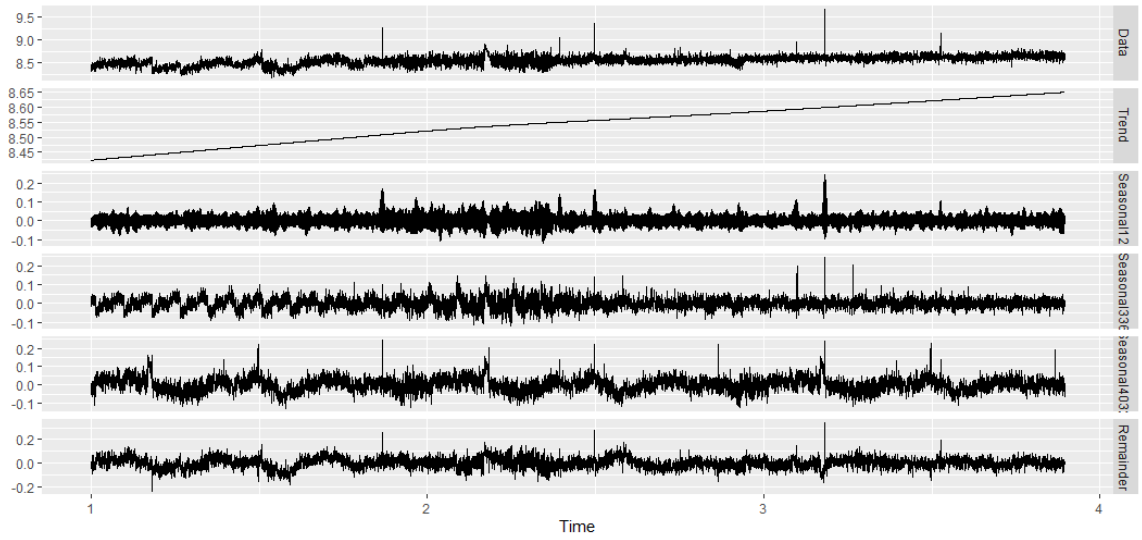
out-phasing. Detrending all timeseries might therefore improve model results and explain more data variance in PCA, as well as identifying real sources when applying PMF.

$$Timeseries = seasonal + trend + remainder \tag{5}$$

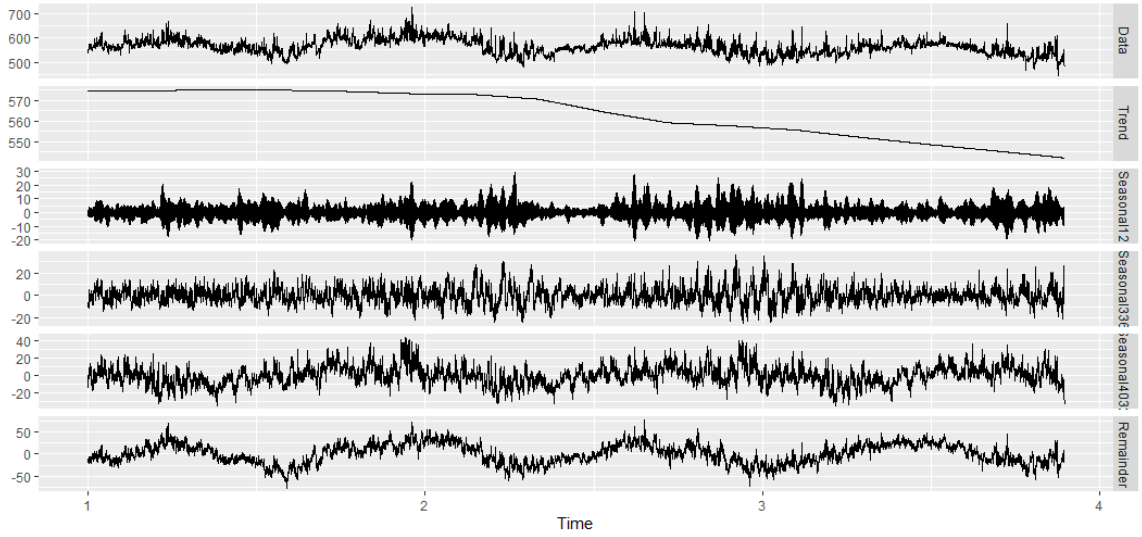
STL function is a robust method that can be used to detrend the time series in data matrix X, by decomposing each timeseries into seasonal, trend and remainder as demonstrated in equation 5. In STL, trend is calculated based on the seasonality given and in this case weekly, monthly, and yearly seasonality is calculated. The trend is illustrated in the second time series in Figure 9 of CFC-115 and CH_3Cl. When detrending the time series, the trend is subtracted from the original time series. Interestingly enough, CFC-115 shows an increasing trend in the period 2015-2018, even though it is expected to be out-phasing due to Montreal Protocol. This is expected because CFC-115 has a very long lifetime and the effect of reduced emissions takes time to be visible. Methylchloride (CH_3Cl) on the other hand shows a decreasing trend in the middle of the time series corresponding to approx. year 2017 and 2018. Furthermore, CH_3Cl have a similar seasonal variability like NMVOCs.

3.7 Input data file

The two datasets, X_{VOCs} and X_{GHGs}, are merged into one data matrix X used as an input file for further analysis with cluster analysis, PCA, and PMF. A few datapoints are lost when merging the two datasets (359 datapoints), due to the different time coverage. Summer and Winter seasons are separated in two data matrices X_{Summer} and X_{Winter}. This will help interpreting the results of cluster analysis, PCA, and PMF implicitly by taking into account the role of OH radical and photochemistry in affecting NMVOCs variability.



(a)



(b)

Figure 9: STL of (a) CFC-115 and and (b) CH_3Cl. 1. original data, 2. trend, 3. weekly seasonality, 4. monthly seasonality, 5. yearly seasonality, and 6. remainder.

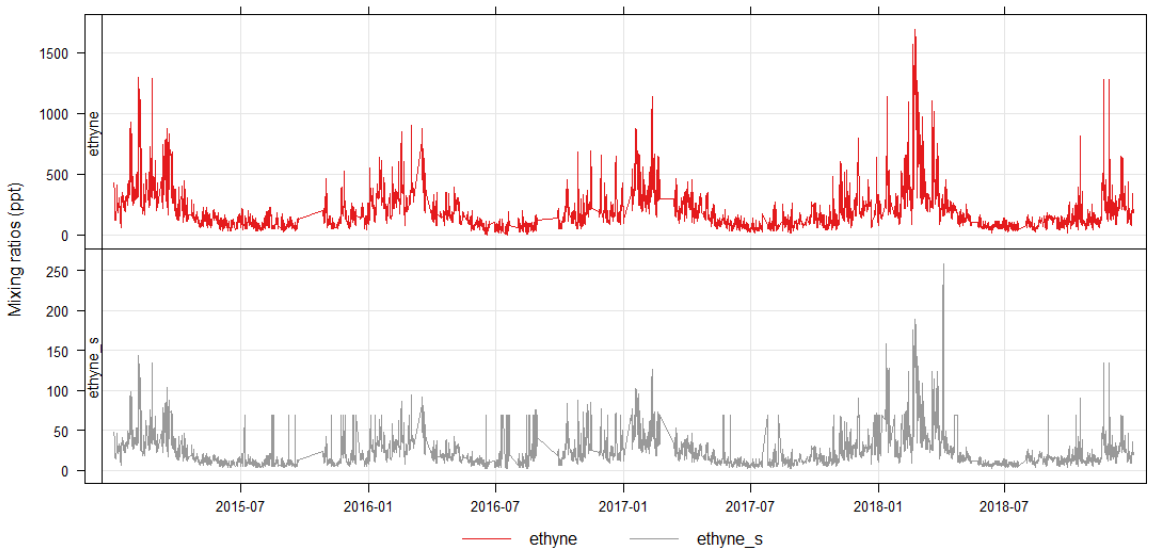


Figure 10: Time series of ethyne (red) and its associated uncertainty time series S (grey), where missing values in the S are substituted with the geometric mean multiplied with 4.

3.8 Uncertainty data matrix

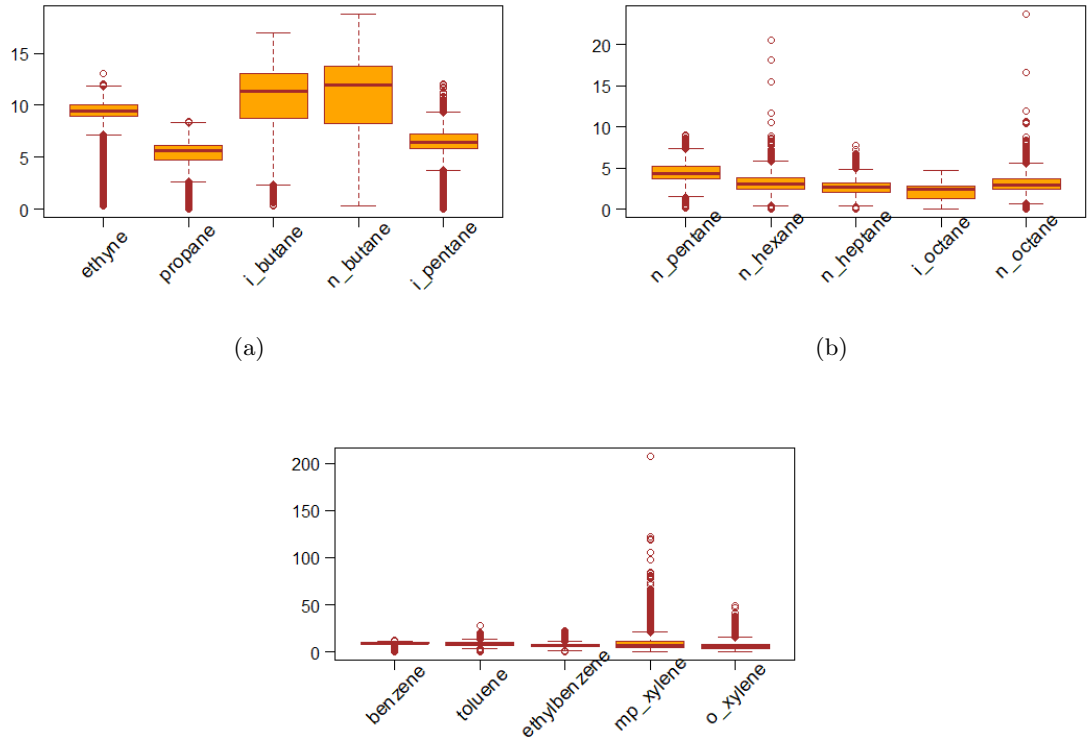
PMF requires both a data matrix (X) and associated uncertainty matrix (S) and in this work analytical uncertainty (S) is used from every sample of the data matrix (X).

There is not a common recognized uncertainty methodology for estimating uncertainties of halogens and GHGs, that on the contrary exist for NMVOCs. The analytical uncertainty of NMVOCs are estimated from measurement repeatability and scale propagation error. In general, the measurement total uncertainty for NMVOC is always larger than the uncertainty estimated for halogens.

The uncertainty data matrix (S) in PMF is almost as important as the concentration matrix (X) because the uncertainty values are the weight of the variables and therefore determine the final result (Belis et al., 2014). Furthermore, PMF can be sensitive to the uncertainty matrix.

An important point to address, is related to how to estimate the missing values in the uncertainty matrix (S). Where the missing values of data matrix (X) was filled using interpolation, its associated uncertainty matrix (S) contain missing values as well. The estimated missing values of matrix S has to be large in order to minimize the influence of the interpolated values in data matrix X when applying the PMF model.

There are no guidelines on which method is most appropriate to use. Therefore, the approach is to try different options and then demonstrate that the choice was meaningful or not with the PMF results. In this thesis, a common approach introduced first by Polissar et al., 1998 is used, where the missing values of uncertainty matrix (S) is substituted by its geometric mean multiplied with 4 (Belis et al., 2014; Reff, Eberly, and Bhave, 2007). This is illustrated in Figure 10.



(a) (b)

(c)

Figure 11: S/N ratios for the whole dataset of NMVOCs species, where data contain no missing values. (a): Ethyne, propane, i-butane, n-butane, and n-pentane; (b): n-pentane, n-hexane, n-heptane, i-octane, and n-octane; (c): benzene, toluene, ethylbenzene, mp-xylene, and o-xylene.

3.9 Signal to noise ratio

In order to evaluate the robustness of each species contribution to the PMF model output, the relationship between the concentration (X) and uncertainty (S) has to be defined, known as the signal to noise (S/N) ratio. S/N distribution are carefully investigated for all species. In order to achieve better results, S/N should be larger than 2 (see section 10.2). Otherwise variables are classified as weak or bad variables, and weights are introduced in the PMF model to constrain the algorithm.

Results are plotted in Figure 11. Species with high variability, many outliers, and a low S/N ratio are the following: n-hexane, n-heptane, i-octane, n-octane, (m,p)-xylene, and o-xylene. Especially n-hexane, n-heptane, and (i,n)-octane are having long periods where uncertainties are very large. Before applying PMF analysis, species are grouped as strong, weak or bad variables depending on their S/N ratio and the final S/N ratio are summarized in Table 5 and Table 6 for Summer and Winter, respectively. In addition, the minimum, median and maximum mixing ratios of all species are summarized in the same tables and these are the data used for the PMF analysis. An additional correction for the lifetime has been tested and applied on the datasets, based on the OH exposure method, as explained in the following chapter.

Table 5: Summer dataset statistics.

Species	S/N	Min (ppt)	Median (ppt)	Max (ppt)
SO_2F_2	7.6	1.91663	2.22547	59.88395
HFC-32	10	10.64663	14.98042	43.97495
HFC-125	10	17.71908	21.7168	67.05466
HFC-134a	10	83.89794	95.4909	155.22598
HFC-152a	10	7.67518	10.86015	27.50471
HFC-365mfc	4.3	0.81249	1.1801	5.69075
HCFC-22	10	231.32428	240.31063	265.66247
HCFC-142b	10	22.47877	23.85912	28.38858
CFC-114	10	15.86305	16.30386	16.95074
CFC-115	9.9	8.12144	8.42898	8.79124
CH_3Cl	10	494.12828	565.62247	688.48333
CH_3Br	10	6.44889	7.67966	11.83729
CH_2Cl_2	10	32.74607	58.3796	227.61684
$CHCl_3$	10	5.83247	12.30982	25.14529
CCl_4	10	76.07161	78.90963	85.767
CH_3CCl_3	9.9	3.07329	3.55165	17.42013
TCE	3	0.36236	0.78803	4.59543
PCE	10	2.88765	6.27629	51.27296
COS	10	419.33281	496.28785	554.07392
ethyne	9.3	42.45691	112.47768	315.90999
propane	5.2	50.81561	232.96507	975.81715
i-butane	9.2	12.42267	48.02323	212.00289
n-butane	9.5	21.19955	83.51012	415.10216
i-pentane	9	14.16457	54.5075	405.13027
n-pentane	5.8	8.7669	27.09622	205.53149
n-hexane	3.7	0.36433	5.74964	44.78401
n-heptane	3.4	0.92235	3.9567	31.47713
i-octane	2.1	0.3536	2.83187	15.32556
n-octane	3.3	0.67745	2.32275	11.25879
benzene	9.7	27.5553	55.59146	138.67187
toluene	9.8	11.21787	43.09407	538.90388
ethylbenzene	9.3	5.78819	13.86782	68.3724
mp-xylene	9.1	20.67209	49.52794	244.18716
o-xylene	9	5.48634	11.35577	44.60328

Table 6: Winter dataset statistics.

Species	S/N	Min (ppt)	Median (ppt)	Max (ppt)
SO_2F_2	7.2	1.8917	2.14638	2.61017
HFC-32	9.9	9.76125	13.04012	27.50988
HFC-125	10	17.50478	19.86663	35.04822
HFC-134a	10	81.72684	88.71379	122.83512
HFC-152a	10	7.27641	10.84001	37.34464
HFC-365mfc	3.8	0.77642	1.07159	3.7035
HCFC-22	10	230.29709	240.72917	283.23517
HCFC-142b	10	22.66013	23.50511	29.9665
CFC-114	10	16.01933	16.34838	16.68162
CFC-115	9.9	8.2214	8.43813	9.23165
CH_3Cl	10	520.59746	578.42217	671.14399
CH_3Br	9.8	6.67339	7.25854	7.91101
CH_2Cl_2	10	39.66116	63.92442	167.53329
$CHCl_3$	10	8.68793	13.84087	21.75432
CCl_4	10	77.11333	79.01024	81.91518
CH_3CCl_3	9.6	3.25113	3.50144	5.3753
TCE	4.2	0.29566	1.13611	9.29776
PCE	9.9	1.73887	5.83089	43.62095
COS	10	434.81013	501.09719	536.845
ethyne	9.2	60.75802	322.59623	1780.76073
propane	5.5	123.80042	755.87038	2235.27602
i-butane	9.2	14.29832	128.40679	484.14547
n-butane	9.5	32.7066	245.59569	719.52078
i-pentane	7.3	6.61945	81.07247	346.85999
n-pentane	4	3.60214	60.81414	320.90026
n-hexane	2.6	0.51286	15.28838	174.13807
n-heptane	2.5	0.21668	6.52221	58.6753
i-octane	1.2	0.22043	3.19582	21.37665
n-octane	2.7	0.20595	2.97547	19.04787
benzene	9.6	29.26983	120.48134	603.11542
toluene	8.5	1.26623	55.27111	957.41799
ethylbenzene	7.7	0.86924	13.77	56.71226
mp-xylene	7	3.10443	49.17857	202.5438
o-xylene	7.1	1.01401	11.621	91.53672

Part III

PRELIMINARY ANALYSIS

4

LIFETIME CORRECTION METHOD

The aim of this chapter is to investigate the role of the diverse atmospheric lifetime of NMVOCs. Each individual NMVOC have different lifetime, but they have generally a much shorter lifetime than halogenated species. Specifically, when classifying NMVOCs source categories, it is of great importance to also consider the photochemical processes occurring during the transport to the receptor. The atmospheric concentration of NMVOCs strongly depends on the tropospheric oxidation by OH radical and correspondingly, affect the presumable range of transport from source to receptor according to their atmospheric lifetime.

With the aim of evaluating the impact of lifetime correction to the PCA and PMF results, the age of air masses at measurement site can be estimated by calculating the OH exposure, using atmospheric mixing ratios of NMVOCs (De Gouw et al., 2005).

The OH exposure equation (6) also called "photochemical age", was first introduced by Roberts et al., 1985, using toluene/benzene ratio. Here, E/X ratio is used in this research based on the criteria that they are both emitted from a common traffic source, are strongly correlated and have different atmospheric lifetime (He et al., 2019). The lifetime of (m,p)-xylene is considerably shorter than ethylbenzene (Table (1)), and can therefore demonstrate the age of an air mass as X/E mixing ratio decreases significantly when moving away from the emission source (Monod et al., 2001). The initial NMVOCs concentration is calculated using equation (8), whereby the obtained concentration is corrected for the atmospheric lifetime of the species.

- OH exposure:

$$[OH] \cdot \Delta t = \frac{1}{k_{\mathrm{E}} - k_{\mathrm{X}}} \times \left(\ln \frac{[E]}{[X]}\bigg|_{\text{source}} - \ln \frac{[E]}{[X]}\bigg|_{\text{CMN}} \right) \tag{6}$$

$$\Updownarrow$$

$$\frac{[E]}{[X]}\bigg|_{\text{source}} = \frac{[E]}{[X]}\bigg|_{\text{CMN}} \times \exp\left((k_{\mathrm{E}} - k_{\mathrm{X}}) \cdot [OH] \cdot \Delta t\right) \tag{7}$$

- Lifetime correction method:

$$[NMVOC]_{initial} = [NMVOC]_{measured} \times \exp(k_{\mathrm{NMVOC}} \cdot [OH] \cdot \Delta t) \tag{8}$$

Where $\frac{[E]}{[X]}|_{\text{source}}$ represents ethylbenzene/(m,p)-xylene source ratio measured at Po Basin and $\frac{[E]}{[X]}|_{\text{CMN}}$ is the observations at CMN; Δt indicate the age of the air mass and k_{E} and k_{X} are the OH rate constants of ethylbenzene and (m,p)-xylene; $[NMVOC]_{measured}$ is the measured concentration of NMVOC at CMN and k_{NMVOC} is the corresponding OH rate constants.

This method is based on the assumption that the transport of NMVOCs to CMN is mainly coming from Po Basin domain without mixing with fresh emissions nearby receptor or very far emission sources. However, the latter is valid only for the less reactive species (like benzene) where the impact of OH chemistry is smaller.

Finally, a sensitivity study is performed of NMVOCs with and without lifetime correction method obtained by the application of equation (8). The initial NMVOCs dataset is merged together with the halogens and GHGs and used as an input data file for further analysis with PCA and PMF.

4.1 Po Basin source ratio

The lifetime correction of NMVOCs involve two steps, first OH exposure is calculated from equation (6) based on ethylbenzene/(m,p)-xylene (E/X) source ratio and secondly, the initial value of NMVOC can be calculated by equation (8).

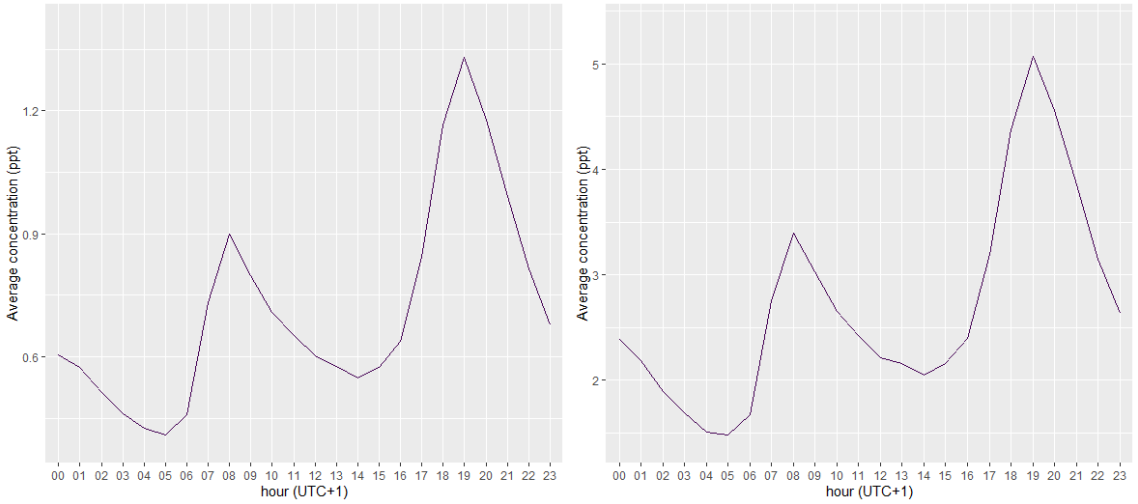
The E/X source ratio is needed as a reference ratio at the emission source at $t = 0$, which is in this case the Po Basin. It is estimated from urban traffic environment in Po Basin at dark hours from 18:00-06:00 ($UTC + 1$), to ensure a low degree of OH photochemistry. The environmental and protection agency ARPAE-Emilia Romagna, provided BTEX emissions collected from 9 ambient monitoring stations in Emilia Romagna (one ambient monitoring station from each city), that are located in a traffic environment. ARPAE-Emila Romagna only provide a NMVOC dataset spanning from date - date and a total of 1200 observations and 45 variables from the period 1st October 2019 to 19th November 2019 is analyzed and the locations and names of the ambient monitoring stations are given in Table 7. To minimise the possible impact of OH radical near the emission source, only autumn data are considered.

Table 7: Ambient monitoring stations in Emilia Romagna.

Geographical locations	Name of monitoring station
Piacenza (PC)	Giordani-Farnese
Parma (PR)	Montebello
Reggio Emilia (RE)	Timavo
Modena (MO)	Giardini
Bologna (BO)	Porta San Felice
Ferrara (FE)	Isonzo
Ravenna (RA)	Zalamella
Forlì (FO)	Roma
Rimini (RN)	Flaminia

Before deciding on using E/X ratio as a reference ratio for calculating OH exposure and estimating the photochemical age, also benzene/toluene (B/T) ratio was calculated and evaluated, as performed by Roberts et al., 1985. The calculated initial value from equation (8) using B/T ratio gave very high values for the more reactive species, such as (m,p)-xylene and o-xylene. For this reason, B/T ratio was not considered suitable to use in this context as a reference ratio, because some species are more reactive than toluene. Besides the need to include species with different atmospheric lifetimes, they also need to be correlated, assuming

that they only come from the same traffic source (He et al., 2019). Pearson correlation is used for comparative evaluation of correlation among species. The Pearson correlation between benzene and toluene during summer is $r = 0.50$ at CMN. In comparison, ethylbenzene and (m,p)-xylene correlation is $r = 0.76$ and significantly higher. From Figure 12, the diurnal variability of ethylbenzene and (m,p)-xylene illustrates a strong correlation between the two species in the PBL of Emilia-Romagna and is calculated to be $r = 0.9926$. Furthermore, two significant traffic peaks are related to the daily behaviour pattern of car commuters going to and from work during rush hour as well as to the PBL height evolution during the day.



(a) ethylbenzene (b) (m,p)-xylene

Figure 12: Averaged diurnal time series in Emilia Romagna in the period 01/10/19 - 19/11/19 (1200 observations), (a) for ethylbenzene and (b) for (m,p)-xylene.

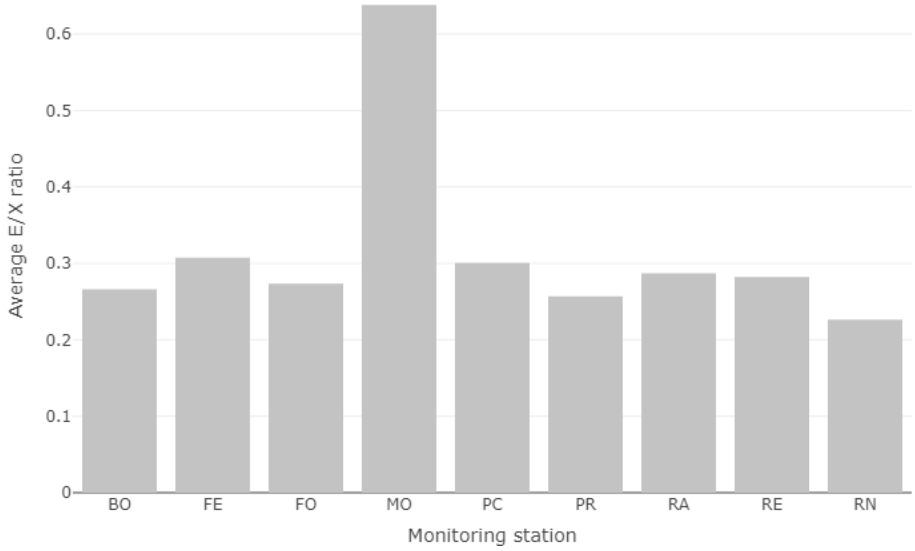
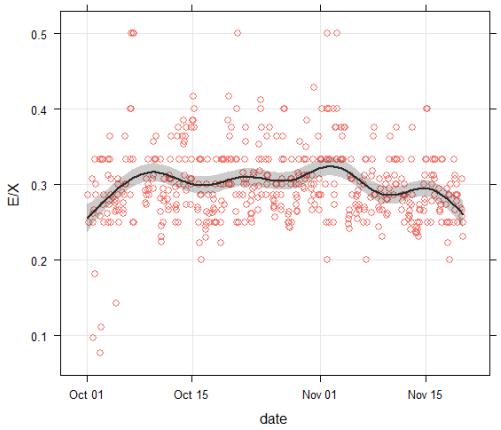


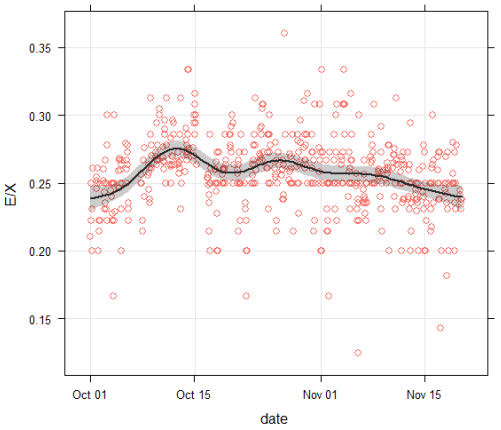
Figure 13: Average E/X source ratio of 9 ambient monitoring stations located in Emilia Romagna (the full names are available in Table 7).

Based on these evidences, E/X ratio is chosen as a reference source ratio that represents the emissions in the Po Basin and provides an estimate of the age of NMVOCs emissions in

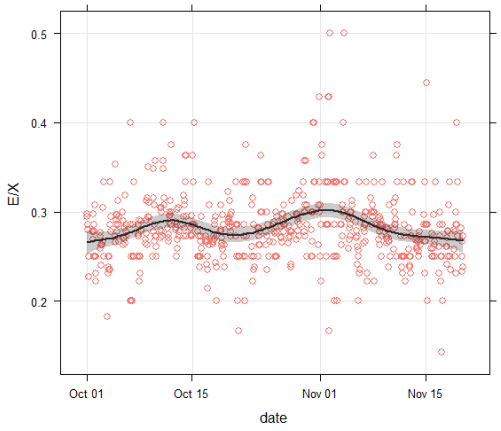
the sampled air masses. However one ambient monitoring station located in Modena has a much higher E/X ratio on average, attributable to a low (m,p)-xylene concentration compared to the other stations as seen in Figure 13. The E/X ratio should be similar for all cities as their sources are identical in an urban environment (Monod et al., 2001). Therefore, the monitoring station in Modena is removed from this analysis and the averaged E/X source ratio is calculated to be 0.28 (X/E = 3.76). A comparative study of X/E mixing ratio in urban environment at several different locations and sources is conducted by Monod et al., 2001, and indicate a mixing ratio ranging from 2.8 to 4.6. The obtained X/E source ratio from Po Basin is 3.76, which is within this range. The X/E value obtained from the autumn ARPAE dataset is particularly close to the measured X/E ratio from traffic tunnel and roadside studies carried out at 10 different cities worldwide, where X/E ratio is 3.64 and 3.18, respectively (Monod et al., 2001). A recent study calculates a E/X source ratio to be 0.62 in China, however that study is conducted much closer to the source with an air mass age of only 3 hours (He et al., 2019).



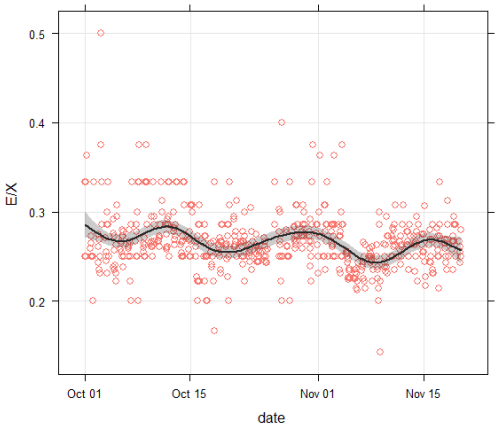
(a) Piacenza



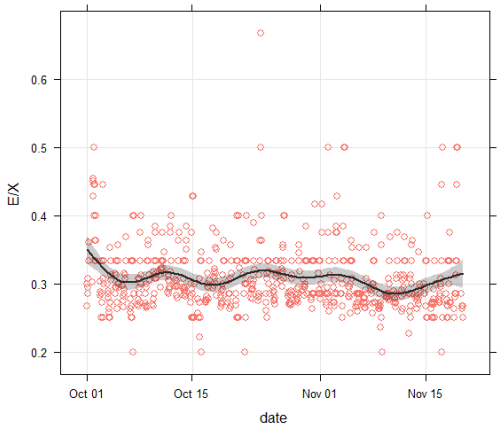
(b) Parma



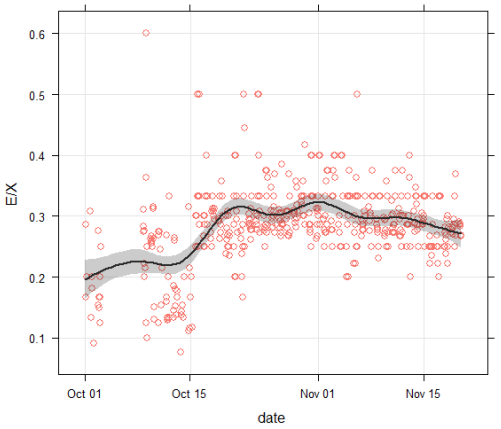
(c) Reggio Emilia



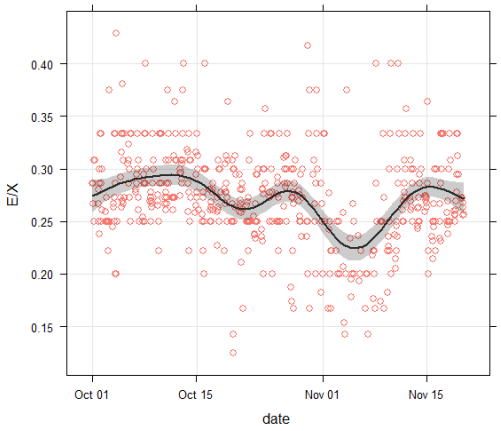
(d) Bologna



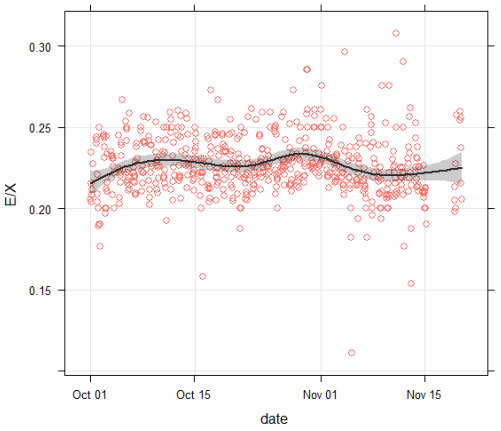
(e) Ferrara



(f) Ravenna



(g) Forlì



(h) Rimini

Figure 14: Time series of E/X source ratios from 8 ambient monitoring stations located in Emilia Romagna, Po Basin. Red circles indicate samples, black line represents the mean.

4.2 OH exposure

The average photochemical age of measured air masses at CMN is calculated according to equations (9),(10),(11). The air mass age is estimated to be ≈ 12 hours during Summer season, by using an OH concentration representing the whole European domain as described in Chapter 2. This value is roughly consistent with the typical wind speed value (average value: 6.4 m/s) observed at CMN during Summer season (Cristofanelli et al., 2017). At CMN, 12 hours roughly corresponds to an average catchment area of about 40.8 km equivalent radius as defined by the modelling work carried out by Henne et al., 2010.

The OH rate constants k_E and k_X belong to ethylbenzene and (m,p)-xylene respectively, and $\frac{E}{X}|_{CMN}$ is here calculated as the average mixing ratio during Summer at CMN.

$$\begin{aligned}[OH]\Delta t &= \frac{1}{(7.0 \cdot 10^{-12} - 19.0 \cdot 10^{-12})} \times (\ln(0.28)|_{\text{source}} - \ln(0.56)|_{CMN}) \\ &= 5.78 \cdot 10^{10} cm^{-3} molecule\ s\end{aligned} \tag{9}$$

$$\Delta t = \frac{5.78 \cdot \{10^{10} cm^{-3}\ molecules}{1.32 \cdot 10^{6} molecules\ cm^{-3}} = 43787.9s \tag{10}$$

$$\Delta t_{hour} = \frac{43787.9s}{60s\ min^{-1} \cdot 60min\ h^{-1}} = 12.2h \tag{11}$$

When applying OH exposure to the observational dataset, the method is applied for every single measurement as shown in Figure 15. It should be noted, that the calculated OH exposure based on ethylbenzene/(m,p)-xylene source gives a few negative values. The fraction of negative OH exposure events was calculated to be 3.8 % of total NMVOCs dataset (a total of 481 negative events). According to equation (6), the negative events can be explained by the fact that the true E/X emission ratio at receptor site (CMN) deviates from the E/X source ratio estimated from the Po Basin. This can suggest that some NMVOCs events originating from local sources, that specifically enhance only the (m,p)-xylene concentration. Negative OH exposure values has also been evident in a study using similar approach, showing that initial VOCs values are lower than the original ones, thus implying negative OH exposure for some periods (Yuan et al., 2012).

To validate if the negative events should be excluded or not, a few methods and sensitivity tests have been carried out. One method was to combine OH exposure data with model calculations to find the best estimate e.g. by using a simple "smoother" which takes the mean of a moving window every week. Another approach was to exclude all negative events and ultimately every method contain a certain degree of arbitrary. In the end, only specific coherent negative events that last less than 6 hours have been removed as it is not robust to apply the same lifetime correction method for these periods. A lifetime of 6 hours and below corresponds to air masses that are likely related to local sources and not from the Po Basin. Moreover, four evident events with very high peaks related to particular behaviour of atmospheric species have been excluded.

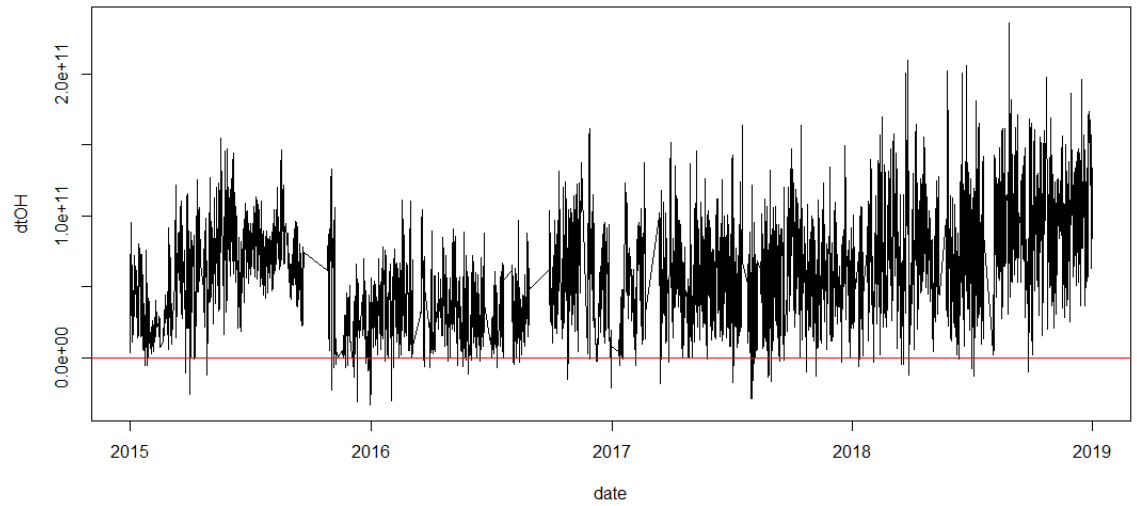


Figure 15: Calculated OH exposure, for every NMVOCs sample where negative event that last $>$ 6 hours are removed. The red line is a zero threshold, to denote negative samples.

4.3 Lifetime correction

The initial NMVOCs concentration are calculated by equation (8) and the effect of applying atmospheric lifetime correction on data is plotted in Figure 16. The NMVOCs initial/NMVOCs observed represents with and without lifetime correction and is calculated based on the average concentration of NMVOC. The result shows, that the more reactive species concentrations increases significantly after applying the OH exposure method. For example, (m,p)-xylene increases 4 times in concentration. The high correction for reactive species, in particular xylenes, is consistent with the relatively remoteness of CMN.

Finally, the next step involves cluster analysis using the calculated initial values for NMVOCs together with halogenated species and non-CO_2 GHGs, to explore correlation and groups among the species.

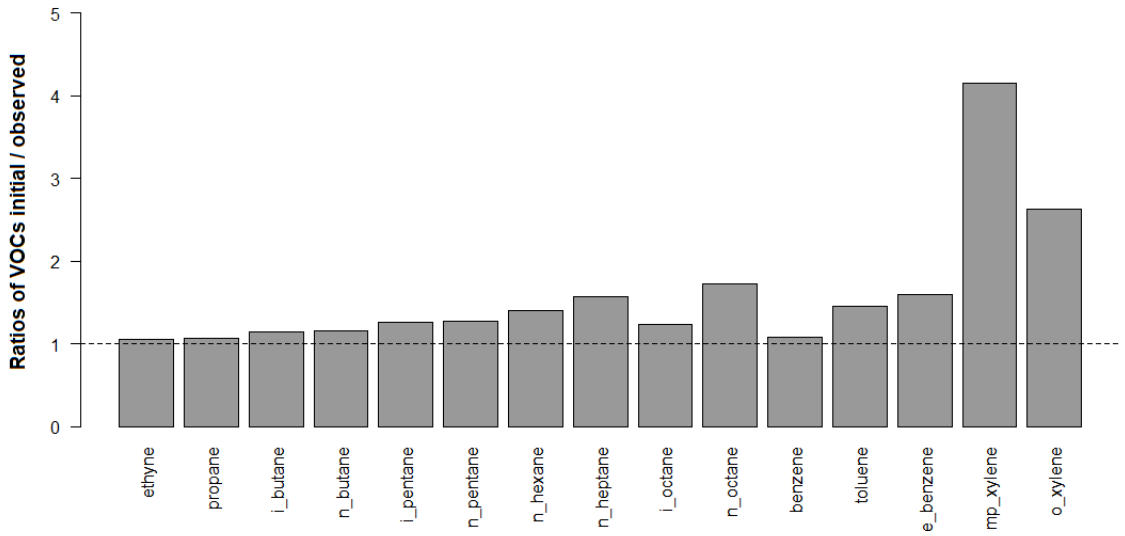


Figure 16: NMVOCs initial values with liftime correction method / NMVOCs observed values at CMN.

5

METHODOLOGY

This chapter introduces the preliminary analysis methods, cluster analysis and principal component analysis (PCA) with the aim to explore the correlation among the species and classifying groups within the data. Cluster analysis and PCA have been performed for defining a preliminary strategy for setting the PMF analysis. The chapter starts with an introduction to cluster analysis methods and the calculation procedure for the analysis. Then follows an introduction to PCA, explanation of the model equation, model constraints and finally the standardization procedure for the input data.

Both cluster analysis and PCA are performed using R software (For further information, see Chapter 12).

5.1 Hierarchical cluster analysis

Cluster analysis is used to find groups in data and classify similarities/dissimilarities of variables (Kaufman and Rousseeuw, 2009). In this research, hierarchical cluster analysis is performed based on the Pearson linear correlation distance, using "hclust" algorithm from "stat" package in R software (more details can be found in Chapter 12).
Hierarchical clustering can be performed by different algorithms and distance metrics of dissimilarity (e.g. Euclidean or correlation distance). Furthermore, there are two approaches when constructing a hierarchical cluster: Agglomerative bottom-up approach and Divisive top-down approach. The agglomerative bottom-up approach starts by clustering the species when they are all apart and then grouping species pairwise step by step into clusters, until only one species is left. The Divisive approach, groups species in the opposite direction by starting with one big cluster and then separates them in two or more clusters.

As mentioned before, there are several techniques on how to group the species together using agglomerative clustering and four the common methods are the following: Ward's method, Single linkage, Complete linkage, and Group average.

After using cluster analysis as an exploratory tool by testing different distance measures and algorithm, the results gave different grouping of the species which is also one of the most challenging part when deciding which method is more robust. We decided to use the "complete linkage", method (also called "furthest neighbor") based on a series of preliminary tests performed on the dataset. In the "complete linkage" method, the link between two clusters contains all element pairs, and the distance between clusters equals the distance

between those two elements (one in each cluster) that are farthest away from each other (Vigni, Durante, and Cocchi, 2013).

The agglomerative hierarcical clustering is based on Pearson correlation as metric for dissimilarity and is calculated by equation (12) to find a linear relationship between two variables f and g:

$$r_{f,g} = (f, g) = \frac{\sum_{i=1}^{n}(x_{if} - m_f)(x_{ig} - m_g)}{\sqrt{\sum_{i=1}^{n}\left(x_{if} - m_f\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ig} - m_g\right)^2}} \tag{12}$$

It explains the covariance between variable f and g in data measurements x and its associated sample mean m, divided by its standard deviation σ_f and σ_g. Furthermore, if there is no correlation between f and g $r = 0$ and perfect correlation when $r = 1$ (Kaufman and Rousseeuw, 2009).

Most clustering algorithm are designed for dissimilarities, and therefore the distance based on Pearson correlation is calculated for dissimilarities before applying hierarchical clustering algorithm (Kaufman and Rousseeuw, 2009). The correlation distance is given by:

$$dist = \frac{1 - r_{f,g}}{2} \tag{13}$$

The hierarchical agglomerative clustering results are graphical visualized in a dendrogram, which represents different groups in data.

5.2 Principal component analysis

PCA is a widely used exploratory tool and based on the original data variance, PCA present graphically sample similarities/dissimilarities and correlation between the measured species. PCA is performed using "prcomp" from "stat" package and "factoextra" for data visualization, using R software (further details in Appendix B, Chapter 12).
According to Hopke, Jaffe, et al., 2020, PCA should not be used as a source apportionment method, even though it is classified as a receptor model. Firstly, it does not consider the uncertainties of measurements and is based on a unweighted least square fit, unlike in PMF where each individual data point is weighted. Secondly, the data input file are standardized meaning that the original scale of variables and the absolute concentration is lost and cannot apportion species to sources (Hopke, Jaffe, et al., 2020). Thus, PCA is not used to identify source contribution of each principal components, but rater as a screening tool as already emphasized in Chapter 1.

In this analysis PCA is carried out on the correlation matrix by an eigenvector analysis (Seinfeld and Pandis, 2016). PCA decomposes the large dataset X into a few factors (p) called principal components (PCs). The dataset undergo a linear transformation from multi dimension to a projection on a hyperplane. This hyperplane is spanned by orthogonal components (PCs).

PCA is explained by equation 1 or from a matrix approach:

$$X_{ij} = T_p \cdot V_p^T + E_{ij} \tag{14}$$

Scores T are original samples in PC space explaining the similarity/dissimilarity between each sample based on euclidean distance and represent the high and low intensity. Loadings V, are original variables in PC space, explaining the correlation between each variable based on the angle. Close angles between two variables means that they are closely correlated while variables orthogonal to each other are not correlated. Same applies for variables that are close to PC and can be interpreted as a dominant variable, contributing to highest variance to the PC. The number of PCs is denoted as p and residuals matrix E, is the everything not explained by the model(Vigni, Durante, and Cocchi, 2013).
The obtained PCs can be graphically represented as scores (T) and loadings (V), showing scatter plots of samples and variables in PC space.

5.2.1 PCA onstraints

PCs are a linear combination of original variables which are explained by:

$$\begin{aligned} t_1 &= X_{ij} \cdot v_1 \\ t_2 &= X_{ij} \cdot v_2 \end{aligned} \tag{15}$$

Where t_1 and t_2 are score vectors and v_1 and v_2 are the loading vectors in PC1 and PC2 respectively. X is the original data matrix (Vigni, Durante, and Cocchi, 2013).
The PCs are eigenvectors of the covariance matrix with their corresponding eigenvalues. The covariance matrix (cov(X)) multiplied with a vector v_1, rotates the vector towards the direction of maximum variance, with the largest eigenvalue λ_1, while vector v_2 of cov(X) gives the v_2 with corresponding second largest eigenvalue λ_2, this is valid for the increasing number of PCs for a=1...j:

$$cov(X)v_a = \lambda_a v_a \tag{16}$$

In addition, the PCs have to be normalized and orthogonal to each other (uncorrelated):

$$\begin{aligned} v_1 \cdot v_1 &= 1 \\ v_1 \cdot v_2 &= 0 \end{aligned} \tag{17}$$

In PCA, each PC needs to explain most of data variance which means that the first PC (PC1) describes the maximum spread of the data points projected on PC1. The second PC (PC2) is orthogonal to PC1 and describes the remaining variance (with second largest variance) and this is true for the increasing number of PCs (Vigni, Durante, and Cocchi, 2013).

5.2.2 Standardized data

The input data matrix is standardized which means that all variables have the same metric and all the species are weighted equally. There is zero mean (columns are centered) and unit standard deviation (columns are scaled) and PCA is therefore carried out on the correlation matrix (Seinfeld and Pandis, 2016).

6

CLUSTER ANALYSIS RESULTS

Cluster analysis (CA) is performed before PCA to understand species correlation and group them accordingly. This chapter presents the hierarchical clustering results using "complete linkage" agglomeration method and distance correlation.

The results are illustrated in a dendrogram and are grouped together according to species correlation. Values grouped together close to 0 are similar and highly correlated. Furthermore, the groups are classified according to species main emission source from Table 1 and Table 2.

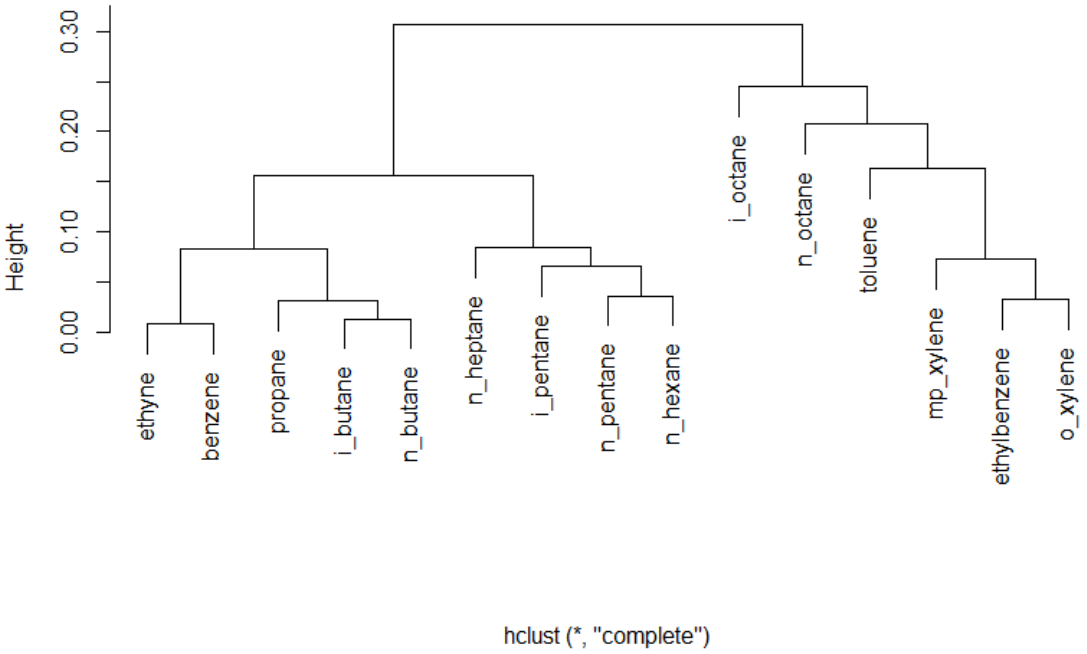
First, cluster analysis is performed on raw dataset X_{VOCs} and X_{GHGs}. Thereafter, to assess the impact of the NMVOCs lifetime correction to the cluster analysis results, clustering is performed on the whole dataset X with NMVOCs lifetime correction.

When performing cluster analysis on raw data, the missing values have to be treated and are removed casewise in the cluster algorithm. A dendrogram of NMVOCs is illustrated in Figure 17 (a), while halogenated species and GHGs are illustrated 17 (b).

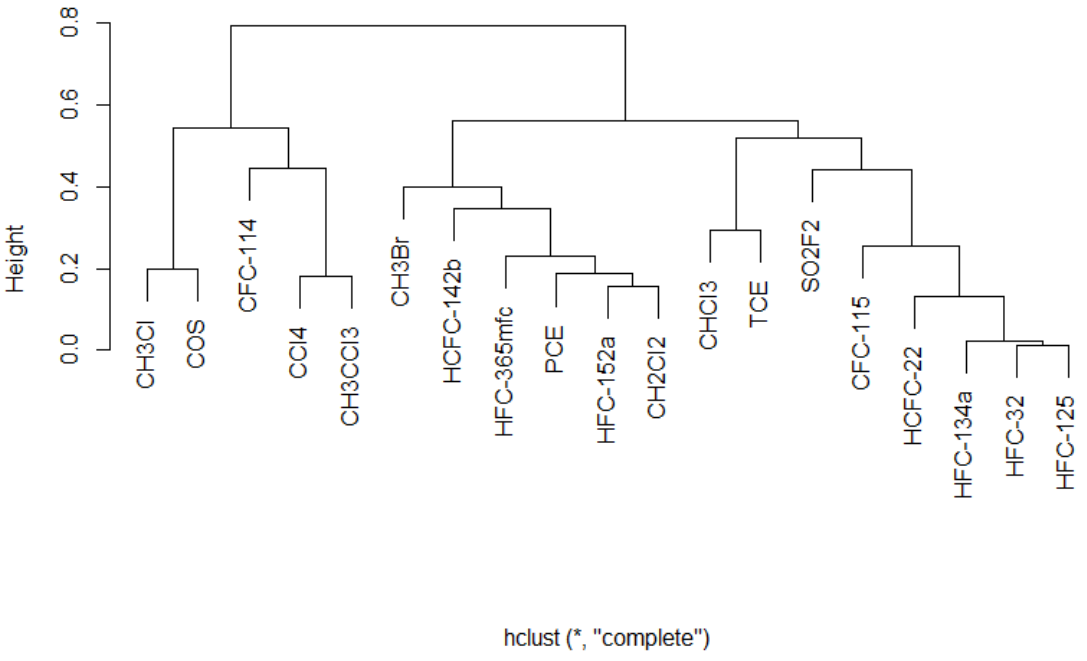
6.0.1 NMVOCs

Two main cluster are distinguished by their variability and lifetime: Species with relative long lifetime (left) and species with relative short lifetime (right). The left cluster contain two subgroups, where ethyne and benzene have the highest degree of similarities and are strongly correlated and both emitted from vehicle exhaust. They are linked with a group containing propane and (i,n)-butane emitted from activities related to liquified petroleum gases (LPG) industry (distilling, storing, distributing, consumption etc.). Furthermore, the subgroup is connected with the second subgroup containing (i,n)-pentane, n-hexane and i-heptane related to gasoline evaporation.

The right cluster contain TEX (Toluene, Ethylbenzene and Xylenes) and (i,n)-octane and could be classified as evaporative emission sources from multiple sources, mainly gasoline, but also solvent industry and solvent use. Especially EX show high similarities suggesting a common source (solvent use and vehicle exhaust).



(a)



(b)

Figure 17: Dendrograms of (a) NMVOCs and (b) halogenated species and non-CO_2 GHGs.

6.0.2 Halogenated species and non-CO_2 GHGs

The dendrogram of halogenated species and GHGs is indicating two clusters. The smaller cluster (left) show high similarities of COS and CH_3Cl variability and both compounds have indeed emission strongly affected by natural sources compared to other halogenated species. Furthermore, they are connected to a group containing CFC-114, CCl_4, and CH_3CCl_3, which sources represents industrial solvents and refrigerant. The same cluster is dissimilar to species variability belonging to the other (right) cluster. Right cluster is divided in two smaller subgroups, where HFC-134a, HFC-32, HFC-125 show a high degree of similarities and are all refrigerants. Moreover, they are grouped with HCFC-22 (refrigerant), CFC-115 (refrigerant) and SO_2F_2 (fumigant) and are connected with TCE and $CHCl_3$ (industrial solvents). This subgroup represents mainly refrigerants connected with a group of industrial solvents. While the left subgroup, show highest similarities between HFC-152a (foam blowing agent) and CH_2Cl_2 (industrial solvent). However they are connected with PCE (solvent), HFC-365mfc (foam blowing agent), HCFC-142b (blowing agent), and CH_3Br (fumigant). The splitting of this subgroup in two branches well capture the two main different activities in which these compounds are involved (commercialized and used): foam blowing industries and refrigeration activities.

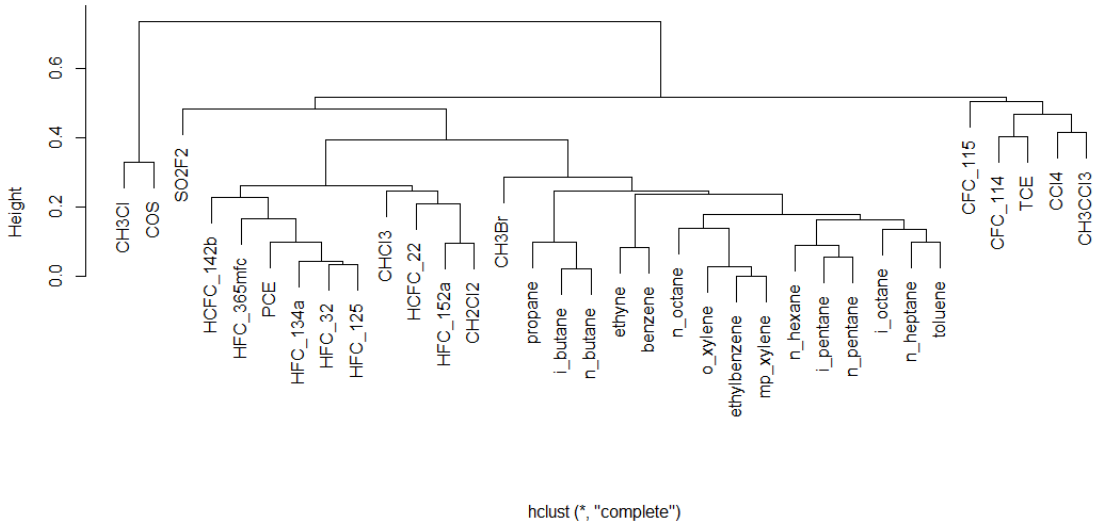
6.1 With lifetime correction

Cluster analysis is performed on dataset X_{Summer} and X_{Winter} with lifetime correction as shown in Figure 18, merging NMVOCs and halogens together. It is evident in both dendrograms that a few species, are dissimilar and distinguished from the rest. These compounds are CFC-115, CFC-114, TCE, CCl_4, CH_3CCl_3, SO_2F_2. Comparing Summer and Winter dendrograms, HCFC-22 and HCFC-142b are also distinguished from the larger cluster during Winter are related to extruded polystyrene foam application. Moreover, CH_3Br is also a "outlier" species grouped with CH_3Cl and COS during Winter, and are all mainly from natural source emissions. The same group is linked with a group containing SO_2F_2 and CFC-115 which cluster is dissimilar to the all other species. Finally, the comparison of Summer and Winter dendrogram show the same (8 to 10) "outlier" species separated and dissimilar to species belonging to the larger cluster.

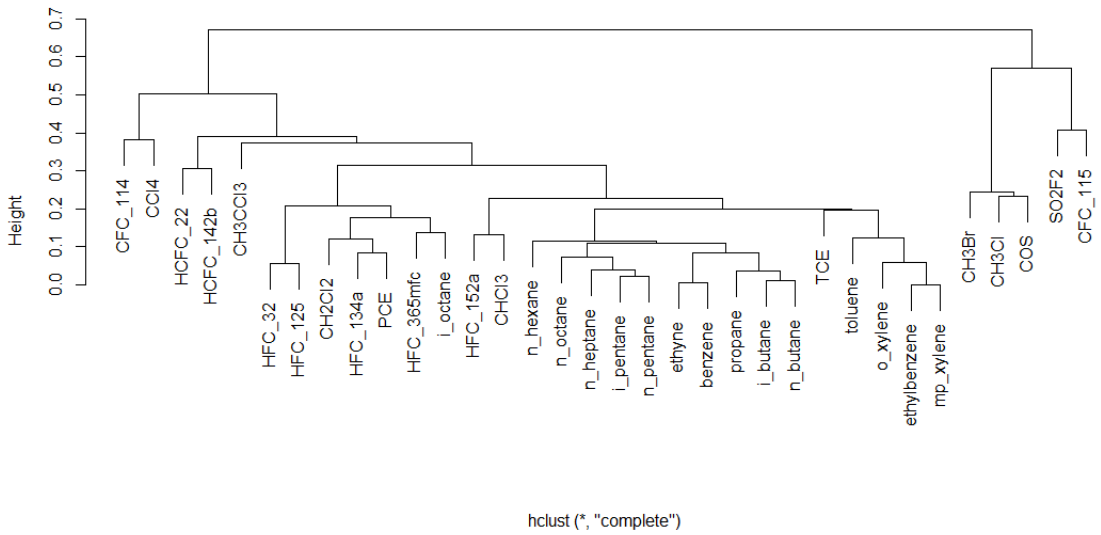
It is to a certain degree subjective when determining the number of clusters obtained, as it depends on the cut-off height on the dendrogram. The following analysis is considering a cut-off height at 0.3, where two main clusters are obtained in both dengrograms; grouping halogenated species and GHGs (left) and NMVOCs (right).

6.1.1 Suummer (JJA)

In Summer, the halogenated species and GHGs (left) are divided in two subgroups, while NMVOCs (right) are divided in several subgroups linked with CH_3Br. Comparing the dendrograms with and without lifetime correction (Figure 17), the NMVOCs groupings are



(a)



(b)

Figure 18: Dendrogram of X_{Summer} (a) and X_{Winter} (b) with lifetime correction.

slightly different. The more reactive species EX and n-octane are now grouped together and less reactive species (benzene, ethyne, propane, and (i,n)-butane) remain in the same cluster probably tracing high/low spatial proximity of emission sources. Furthermore, toluene and n-heptane is closely correlated and grouped together with i-octane with n-heptane, i-octane, and toluene.

6.1.2 Winter (DJF)

During winter, NMVOCs cluster (right) is divided in two main subgroups, where TEX is grouped with TCE and have a dissimilar variability than the rest of NMVOCs. In comparison

with Figure 17 a) without lifetime correction, TEX are now grouped with TCE and are all used as solvents in industry, therefore the group source category can be classified as industrial solvent usage. The other subgroup show almost same cluster as illustrated in 17 a), but with a higher similarity between n-pentane, n-heptane, and n-octane all related to gasoline evaporation. NMVOCs cluster is linked with a single group consisting of HFC-152a and $CHCl_3$ and have different variability than NMVOCs cluster and dissimilar to halogenated and GHG species belonging to the other cluster.

In the (left) cluster containing halogenated species and GHGs, i-octane is highly correlated with the foam blowing agent HFC-365mfc. Moreover, HFC-134a, PCE, and CH_2Cl_2 are correlated and mainly used as chlorinated solvents and feedstock for manufacturing refrigerants (HFCs). A single group of refrigerants contain HFC-32 and HFC-125 and the cluster analysis is keen to capture the variability of these two HFCs that are often used combined in mixtures for the small to medium size domestic and industrial AC conditioners.

7

PCA RESULTS

PCA is performed as a preliminary step prior to PMF analysis. The main objective, is to identify the optimum number of principal components (PCs) that summarizes the data variance. Thus the optimal solution is investigated using four analysis approaches. Furthermore, PCA results are compared with the dendrograms obtained from cluster analysis to investigate if the grouping of species are similar and to validate if the PCA method is robust.

7.1 Input data matrix

PCA is applied on both Summer and Winter data matrix, X_{Summer} and X_{Winter}. Moreover, since the transport of air masses can be roughly distinguished in 1) thermal transport (advection) from the Po valley (PBL) up to the mountain and 2) long range transport, we decided to split the database in "daytime" and "nighttime", in order to exclude a bias induced on the total variability due to the two different patterns. Daytime (10a.m. to 6p.m.) and nighttime (12a.m. to 4a.m.) of each data matrix is also analyzed separately. Every dataset is standardized otherwise most variance (PC1) is dominated by e.g. propane (with high variability and high mixing ratios) and thereby failing to explain other species.

7.1.1 Determining the number of principal components

To determine the optimum number of PCs, eigenvalues and the cumulative variance explained by the PCs are examined. Four approaches are used to evaluate the optimal number of PCs:

1. Using a cut of level = 90 % cumulative variance.
2. Only PC with eigenvalues larger than 1 are retained ($\lambda > 1$).
3. Each PC must explain at least 5% variance.
4. Evaluating scree plots by determining “elbow” points and evaluate loading plots.

The first three approaches are used to statistically evaluate the PCs, while the fourth is following a more intuitive approach. The intuitive approach is to visualize graphically the eigenvalues and variance as function of PCs, known as scree plot in PCA and is illustrated in Figure 19 and 22. These are common strategies for interpreting PCA results (Vigni, Durante, and Cocchi, 2013). The retained PCs should all explain real variation in data compared to

PCs explaining only unsystematic variance. This means, that the retained PC should explain a higher and larger proportion of variance than PC explaining mainly "noise". This is can be detected as an "elbow" or inflection point on the scree plot, where the following PCs do not show a large change in explained variance (Vigni, Durante, and Cocchi, 2013). Furthermore, also loadings are evaluated to understand if collinear variables exists, explaining the same variance in data.

The first statistical approach is to retain PCs explaining a total data variance between 70 % and 99 %. Therefore a cut-off level at 90 % is set to understand how many components are needed in order to explain a large fraction of total data variance. Second approach is to only retain PCs with eigenvalues (variance) greater than 1. According to the eigenvalue criteria, only $\lambda > 1$ explain meaningful data variance (Vigni, Durante, and Cocchi, 2013). The third approach ensures that the PCs do not explain unsystematic variance and include only PCs explaining at least 5 % of data variance. Finally the results from following the three statistical approaches are summarized in Table 8, for Summer and Winter season.

Table 8: Determining the number of PCs of Summer and Winter season

Data set	$\lambda > 1$	5 % variance	90 % variance
Summer season (JJA)	7 PCs (74.8 %)	3 PCs (60.4 %)	15 PCs
Summer daytime	7 PCs (78.0 %)	4 PCs (66.9 %)	13 PCs
Summer nighttime	8 PCs (81.53 %)	4 PCs (67.51 %)	12 PCs
Winter season (DJF)	6 PCs (77.36 %)	3 PCs (66.03 %)	13 PCs
Winter daytime	6 PCs (75.35 %)	4 PCs (68.44 %)	14 PCs
Winter nighttime	7 PCs (78.27 %)	4 PCs (68.32 %)	13 PCs

7.2 Summer (JJA)

The explained variance is plotted as a function of PCs and illustrated in Figure 19 a), where PC1 explain 47.1 % data variance. From Figure 19 b), the number of PCs are suggested to be 3, each of them explaining ≥ 5 % data variance and results in a cumulative variance of 60.4 %. Also, one "elbow" point is noted at PC3. Furthermore, eigenvalues are plotted against the number of PCs in Figure 19 c), where two "elbow" points are noted at PC3 and PC6, respectively. According to the eigenvalue criteria ($\lambda > 1$), 7 PCs are obtained explaining 74.8 % of total data variance, while 15 PCs accounts for 90 % of total data variance shown in Figure 19 d).

To summarize, the 15 PCs explain individually too low variance and 3 PCs explain insufficient amount of cumulative data variance. Therefore, 7 PCs are retained following the eigenvalue criteria and explain an acceptable amount of data variance. Moreover, PCA is also performed on daytime and nighttime dataset and the results are compared. Daytime and nighttime results for $\lambda > 1$ (Table 8), suggest 7 and 8 PCs, explaining up to 81.53 % of total data variance.

The variables contributing to the 7 PCs identified for the Summer season are explored and 4 PCs are depicted in Figure 20 and 21 (see Appendix B, Chapter 12 for the remaining PCs).

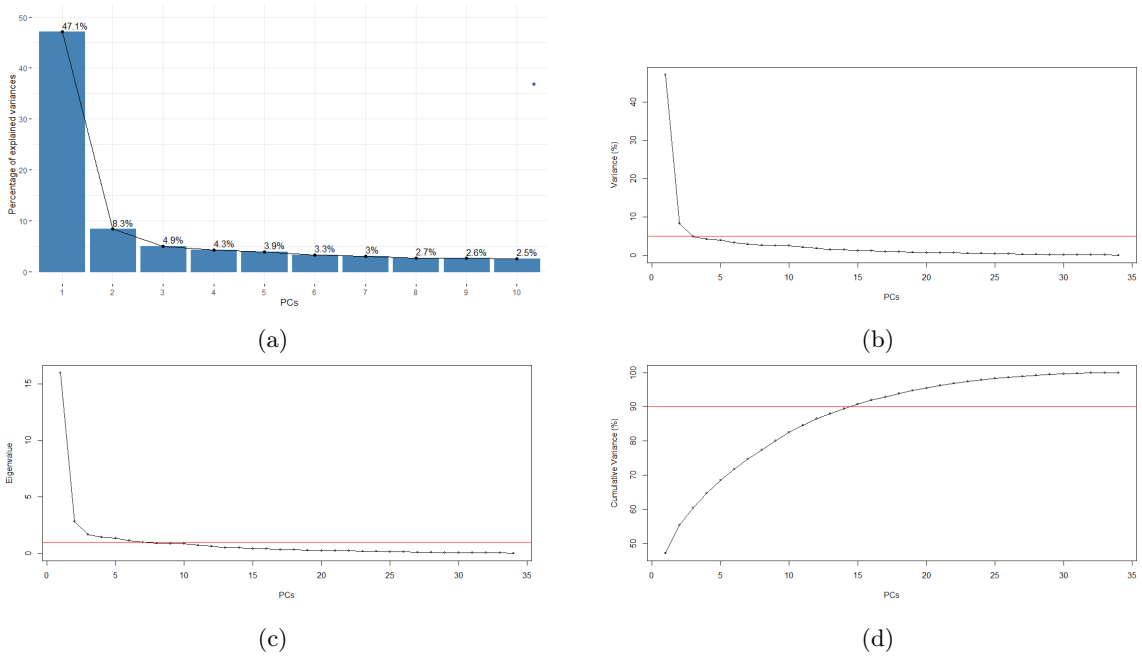
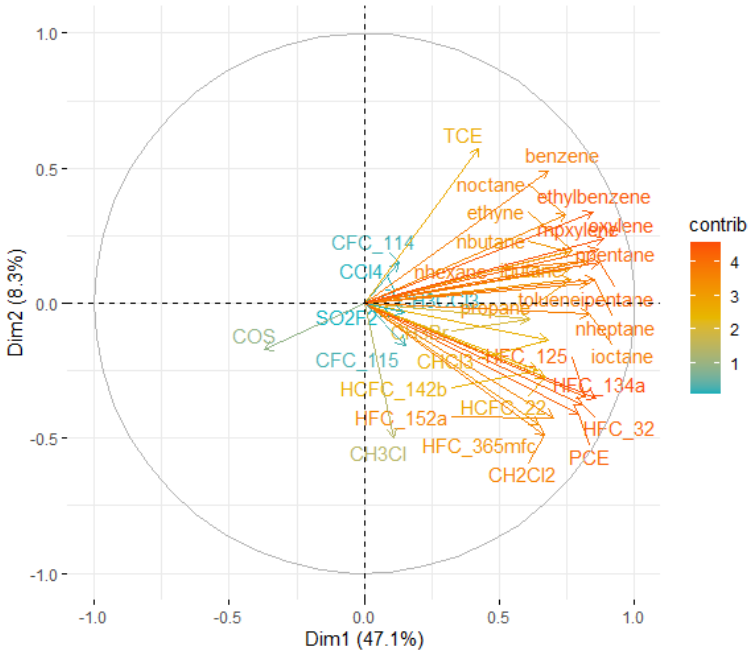
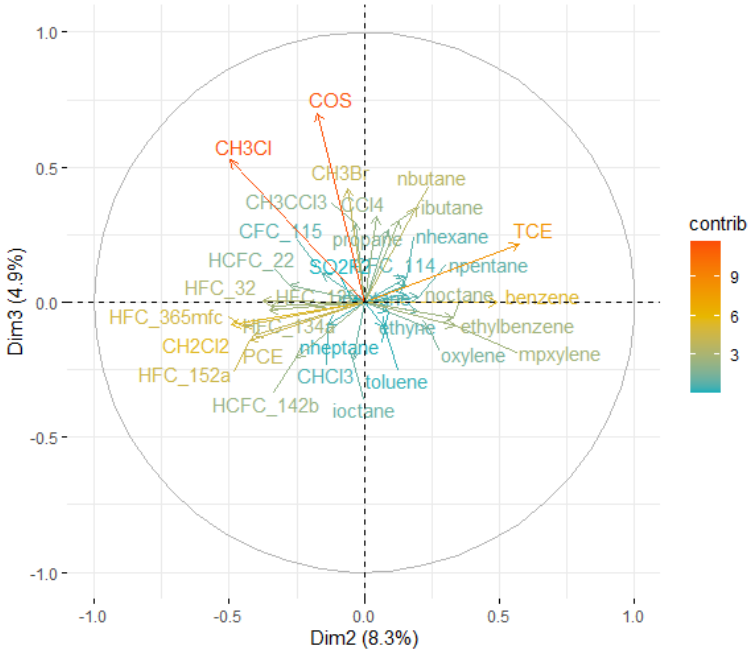


Figure 19: Scree plots of Summer season. (a) Bar chart of explained variance as a function of PCs; (b) graph with a threshold at 5 % variance; (c) eigenvalues as a function of factors with a threshold a $\lambda = 1$; (d) cumulative variance with a cut-off level at 90 %.

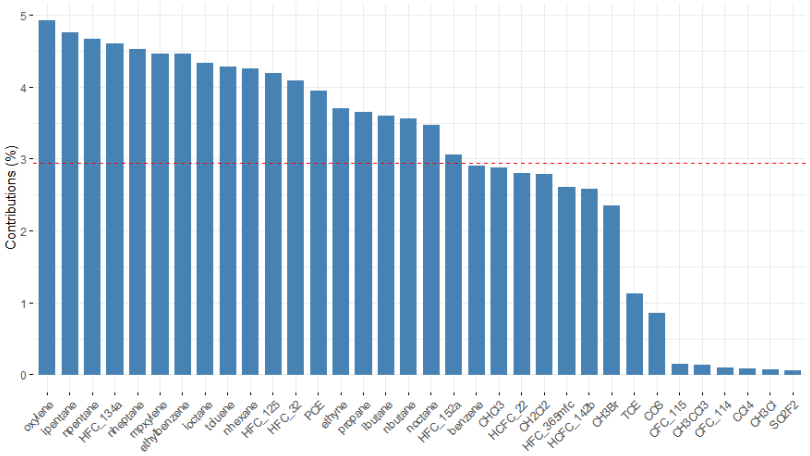
Loadings illustrating PC1 vs. PC2, distinguishes NMVOCs from halogenated species in positive and negative PC space. Moreover, the loading plots did not reveal any interesting correlation among variables, but identified a few "problematic species" to be: SO_2F_2, CFC-115, CFC-114, CH_3Cl, CH_3Br, CH_3CCl_3, CCl_4, TCE, and COS. These species are dominating most of remaining PCs, apart from PC1 and PC2. This means, that the variability related to these species are providing little information to explain other species. Same species are identified as "outliers" in the dendrograms obtained from cluster analysis (Chapter 6.1.2) and are distinguished from the major clusters. A further analysis was carried out (not included in this work), by removing all the "problematic species" from the input matrix to understand how they are effecting the model output. They were introduced back in the model one by one, in order to understand how they affect the explained model variance (PCs). When introducing more than one "problematic species" back in the model such as CH_3Cl and CH_3Br, the obtained PCs distinguished the "problematic species" from the main clusters of species as demonstrated in cluster analysis. Hence, the "problematic species" needs special attention and to achieve better PCA results, further analysis of potential outliers or events should be examined or removing the species from the input matrix.



(a)

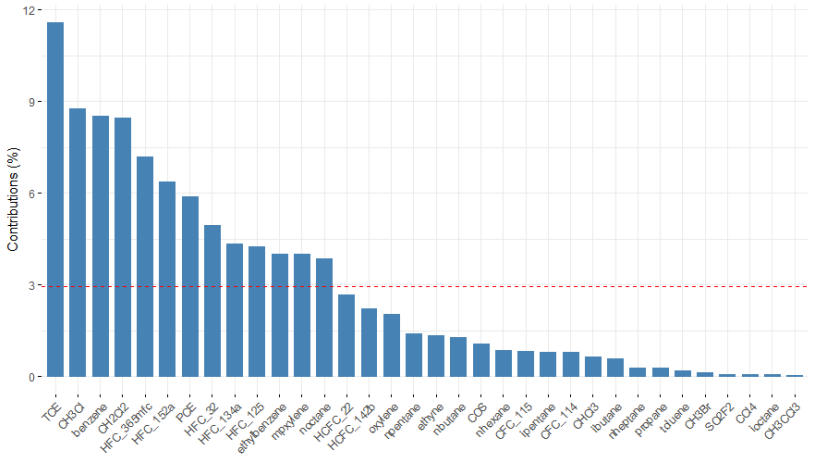


(b)

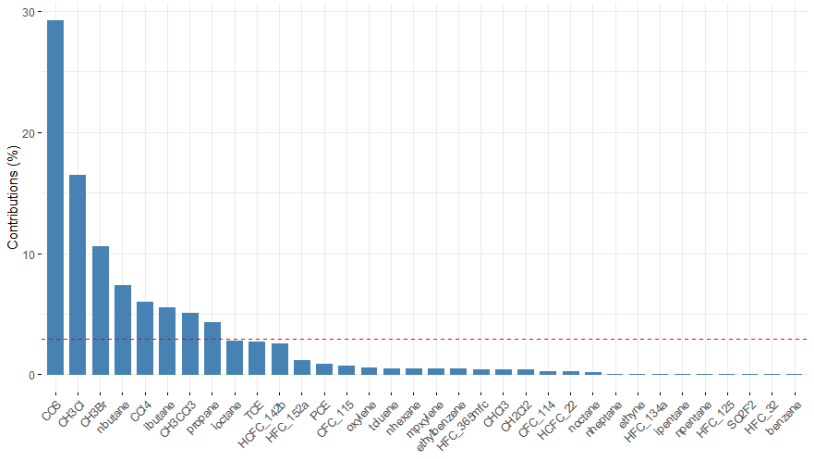


(c)

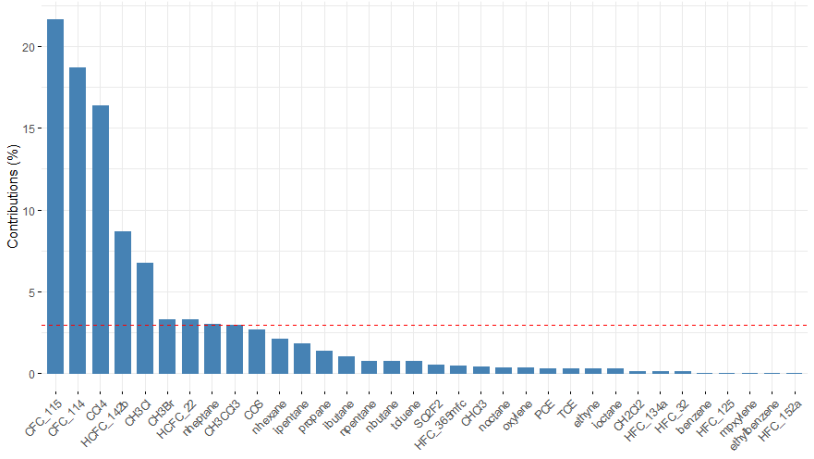
Figure 20: PCA on Summer season. Loading plots of PC2 as a function of PC1 (a); PC4 as a function of PC2 (b); Species contribution of PC1 (c). "Dim" (dimension) is the PC.



(a)



(b)



(c)

Figure 21: Summer season results and species contribution to a) PC2; b) PC3; c) PC4. Red line indicate the average contribution for all 34 variables.

7.3 Winter season (DJF)

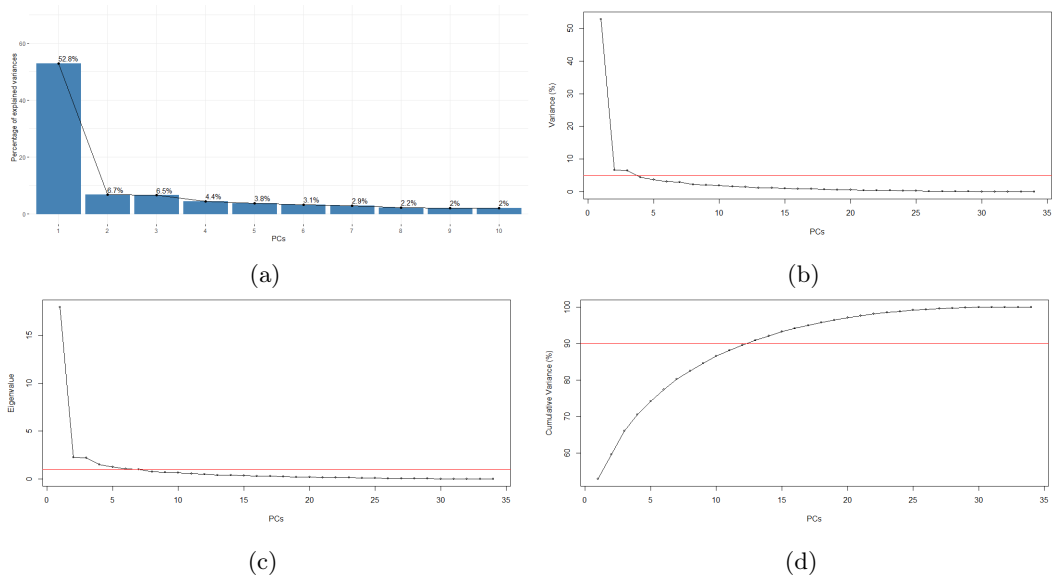


Figure 22: Scree plots of Winter season. (a) Bar chart of explained variance as a function of PCs; (b) graph with a threshold at 5 % variance; (c) eigenvalues as a function of factors with a threshold a $\lambda = 1$; (d) cumulative variance with a cut-off level at 90 %.

PCA results performed on the winter dataset are illustrated on the scree plot Figure 22. In general, higher variance is explained during Winter by fewer PCs compared to Summer season and PC1 explain 52.8 % of data variance (Figure 22 a). Furthermore, two "elbow" points are apparent at PC2 and PC4 in Figure 22 b) and c). Also, when moving from PC7 to PC8, a discrete inflection point is observed. Only 3 PCs are retained explaining ≥ 5 % data variance each, while $\lambda > 1$ suggests 6-7 PCs. A total data variance of 78.3 % is explained by 7 PCs from nighttime data matrix (Table 8), while 13 PCs are necessary to explain 90 % of the total data variance. According to the four approaches, 6-7 PCs seem to be explaining meaningful data variance, although it can be argued that the choice of retained PCs is largely subjective. Recalling the intention of using PCA prior to PMF was solely to explore if there are any underlying phenomena presented in data that can be explained by an optimum number of PCs. In this analysis, eigenvalues above 1 and scree plots turned out to be the most suitable approaches for determining the appropriate number of PCs to retain.

Loadings of Winter season is given in Figure 23 and 24. PC1 vs. PC2 clearly distinguish CH_3Br, COS and CH_3Cl from PC1, which explains variance of a large cluster of species and can be related to natural (oceanic) emissions. This is also evident from the dendrogram of Winter season in cluster analysis. Furthermore PC3(6.5 %) has largest contribution from HFC-32 and HFC-125 that are given a high similarity and correlation in the dendrogram. PC4(4.4 %) explains mainly SO_2F_2, CFC-115, and to a lesser extent also CCl_4 and CH_3Cl (see Appendix B, Chapter 12 for the remaining PCs). To conclude, same "problematic species" observed during Summer are also dominating the PCs during Winter. Results explain mainly the variance of the "problematic species", except for PC1. This clearly points out that PCA

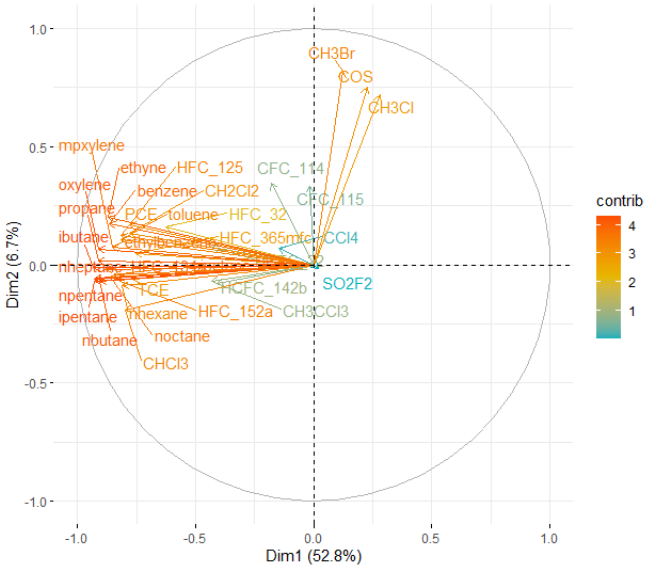
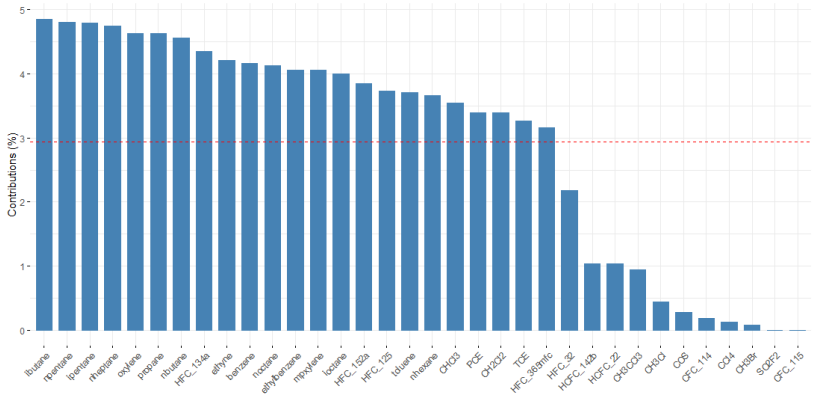
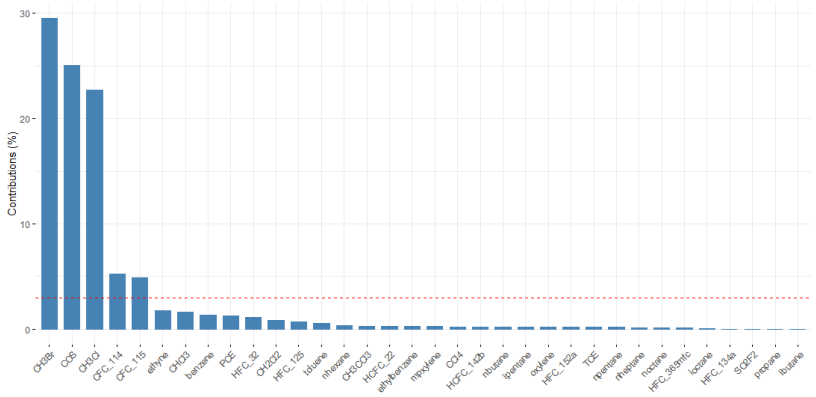


Figure 23: PCA of Winter season. Loading plot of PC2 as a function of PC1, where "Dim" (dimension) is the PCs.

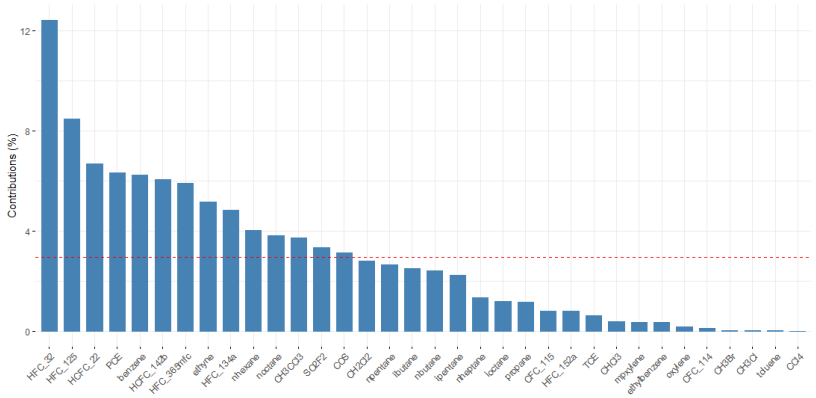
is sensitive to the "problematic species" and is therefore not effective in attributing potential emission sources of NMVOCs, halogenated species and non-CO_2 GHGs.



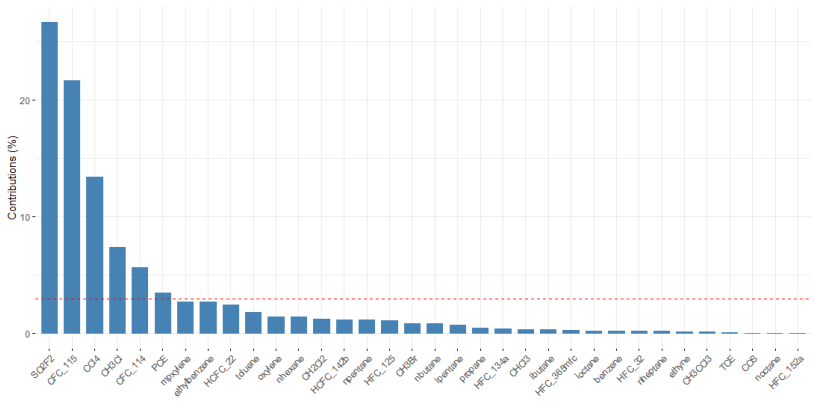
(a)



(b)



(c)



(d)

Figure 24: Winter season results and species contribution to a) PC1; b) PC2; c) PC3; d) PC4. Red line indicate the average contribution for all 34 variables.

Part IV

PMF ANALYSIS

8

METHODOLOGY

Prior to the PMF analysis, validation of data and explorative analysis have been accomplished. This include a realistic evaluation of both molar ratios and uncertainty dataset as well as an explorative analysis of trends and seasonality, basic statistics, S/N ratios, cluster analysis, and PCA.

Factor analysis is described by equation (1) or by the matrix algebra formula (Reff, Eberly, and Bhave, 2007):

$$X = GF + E \tag{18}$$

X is the NxM data matrix, where each row (N) contain one sample for each chemical species and each column represent the time record of certain species (M). X is decomposed into two smaller matrices, namely G and F. They have smaller dimensions and are explained by a few factors P, that are linear combination of old variables. The aim is to reduce the data matrix X into G and F without loosing too much information and where the number of factor (P) solution is less than the number of species (P $<$ M) and may be related to real sources or phenomena present in data (Comero, Capitani, and Gawlik, 2009). It is therefore important to carefully choose the optimum number of factors because the model solution will change with respect to the factor number.

G is a NxP matrix explaining the source contribution of factor/source P, also called factor Scores, and correspond to original samples from data matrix X. F is a PxM matrix and is the factor profiles, also called factor Loadings, explaining the concentration/percentage of species m contributing to factor/source P. Finally, E is the residuals and is the difference between input data X and modeled data (X - GF).

8.1 Comparison of PMF and PCA models

Both PMF and PCA are multivariate statistical models originated from the same fundamental equation of continuity and classified as receptor models (as stated in Chapter 1). Despite these analogies, the models uses different algorithm to solve the equation of continuity. PMF uses least square fit weighted with species associated uncertainties (S), while PCA uses no weighted uncertainties matrix. PMF input files include both a concentration matrix (X) and uncertainty matrix (S) that is based on analytical uncertainties, thereby applying a weight on each data point.

There are no applied orthogonal constraint in PMF, while PMF constraint is that all elements of G and F are positive. The positive constraint ensures that the predicted source contribution is always positive (positive emission from source) which is not always the case in PCA and can result in negative apportionment. Another important difference is that PCA is based on a correlation matrix (variables are usually standardized) and therefore the result are in arbitrary units and not based on the absolute concentration. Source apportionment with PMF is based on the absolute concentration and the obtained information explains the influence of species concentration in the ambient air measured at the receptor site. Finally, both models requires expert knowledge of the study area and chemical species in order to interpret the output factors/principal components (Comero, Capitani, and Gawlik, 2009).

8.2 Signal-to-noise ratio

In receptor models, Signal-to-Noise ratio (S/N) is calculated to better understand the relationship between the molar ratios (X) and associated uncertainties (S) and are plotted in Chapter 3.9. According to Table 9, the variables are classified into good, weak and bad variables. Bad variables are excluded from the analysis, while weak variables are downweighted by 10 % (Belis et al., 2014).

Table 9: S/N ratio

$S/N > 2$	Strong variable
$0.2 < S/N < 2.0$	Weak variable
$S/N < 0.2$	Bad variable

(Belis et al., 2014) [1]

8.3 Q function

Q is the weighted least square function which PMF tries to minimize. The Q value is given in equation (19):

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{e_{ij}}{S_{ij}}\right)^2 \tag{19}$$

Where S_{ij} is the uncertainty matrix, and E_{ij} is everything that was not modeled. Q is therefore the sum of square scaled residuals. By assuming that the uncertainty matrix (S) is correct, then Q should follow χ^2 distribution. A good fit of the data will result in the correct minimum value of Q. The minimum Q value is calculated from the PMF model itself either in robust mode (without outliers) or true mode (with outliers) (Reff, Eberly, and Bhave, 2007).

For every PMF solution exist multiple Q values, which means that there can be different relative minimum in a solution. The algorithm starts with a random starting point and should converge to the absolute minimum. If the model does not converge, it means that there is a local minimum and the model best solution has to be evaluated. Therefore the base model needs to run with minimum 10 trials to make sure that the algorithm converges to the absolute

minimum Q. The relationship of $Q/Q_{expected}$ is considered when evaluating the number of factors to retain and is plotted against the rank of factors. The theoretical Q value ($Q_{expected}$) is given by the number of degrees of freedom in data and $Q/Q_{expected}$ is approximately the sum of square residuals divided the number of data points (Norris et al., 2014).

$$Q_{theoretical} \approx (N \cdot M) - P \cdot (N + M) \tag{20}$$

If plotting too many factors $Q/Q_{expected}$ function can have notable inflection point at a given factor (Seinfeld and Pandis, 2016). Moreover, $Q/Q_{expected}$ should be approximately 1 and if $Q/Q_{expected}$ value exceeds »1 it could indicate an underestimation of the input data uncertainties and PMF solution is therefore not optimal (Contini et al., 2016). However, evaluating the optimum number of factor by this approach might be misleading and Q/Q_{exp} should be carefully examined, as Q value also relates to the uncertainties data matrix (S) and tells if it is correctly estimated or not. Furthermore, changing the uncertainty also affects the Q value and it is therefore not recommended to change uncertainties just to force Q to be close to 1 (Belis et al., 2019).

8.4 Residuals

There are different ways to see whether there are problems with the uncertainty (S) matrix based on the result obtained, and in particular based on the distribution of the residuals. In general, the scaled residuals should be normally distributed between -3 to $+3$. A very large distribution of scaled residuals may be due to an underestimation of uncertainties. Likewise, if the distribution is too sharp or too narrow there are generally problems with the uncertainties (Comero, Capitani, and Gawlik, 2009). Furthermore, there are also different ways to evaluate the uncertainties besides the shape of the residual. To verify that variables are decomposed effectively, the residuals is plotted as a function of the variable. The residuals should be normally and randomly distributed around zero and abnormal spikes are related to outliers.

9

PMF RESULTS AND DIAGNOSTICS

This research is the first attempt to perform EPA-PMF on NMVOCs, halogens and non-CO_2 all together to a long-term high frequency data recorded at CMN.

The detrended dataset is divided into Summer and Winter seasons and run separately in PMF. A 8 factor solution is identified for Summer season (JJA) and a 7 factor solution for the Winter season (DJF).

At first, the best PMF solution is determined by examining different solutions composed by 2 to 9 factor solutions without extra modeling uncertainties or rotations. Each model was calculated based on 20 base runs, to make sure that the algorithm did not converge to a local minimum, but rather an absolute minimum Q value. The results showed, that not all base model runs converged to an absolute minimum. This was especially evident when the number of factors exceeded 7 factors. When the model does not converged in all 20 base runs, the solution needs to be evaluated. However, the issue was resolved when adding an extra modeling uncertainty of 5 % to all variables and sample. As a result, all 20 base runs converged.

There are no statistical tests in PMF that can determine the true factor solution. In this work, the choice of factors is determined by:

1. Interpretation of PMF solutions with the help of expert knowledge to associate obtained factors with related sources.
2. Plotting $Q/Q_{expected}$ relationship to identify any noteable inflection points.
3. Comparing results with literature profiles.
4. Validating and comparing PMF solutions with PCA results and cluster analysis.

9.1 Relationship between $Q/Q_{expected}$

The $Q/Q_{expected}$ relationship is plotted against the number of factors and is explored on Summer and Winter PMF results. In Figure 25, a) shows no apparent inflection point, but only a greater slope between 2 and 3 factors. While b), gives several "elbow" points where the slope is greater moving from one factor to another at factor 3,4, and 7. In c), a small inflection point is notable at factor 5 and 7, and moving from 7 to 9 factor the slope is not changing significantly. In d), 3 elbow points are remarkable at factor 3, 5, and 7 respectively.

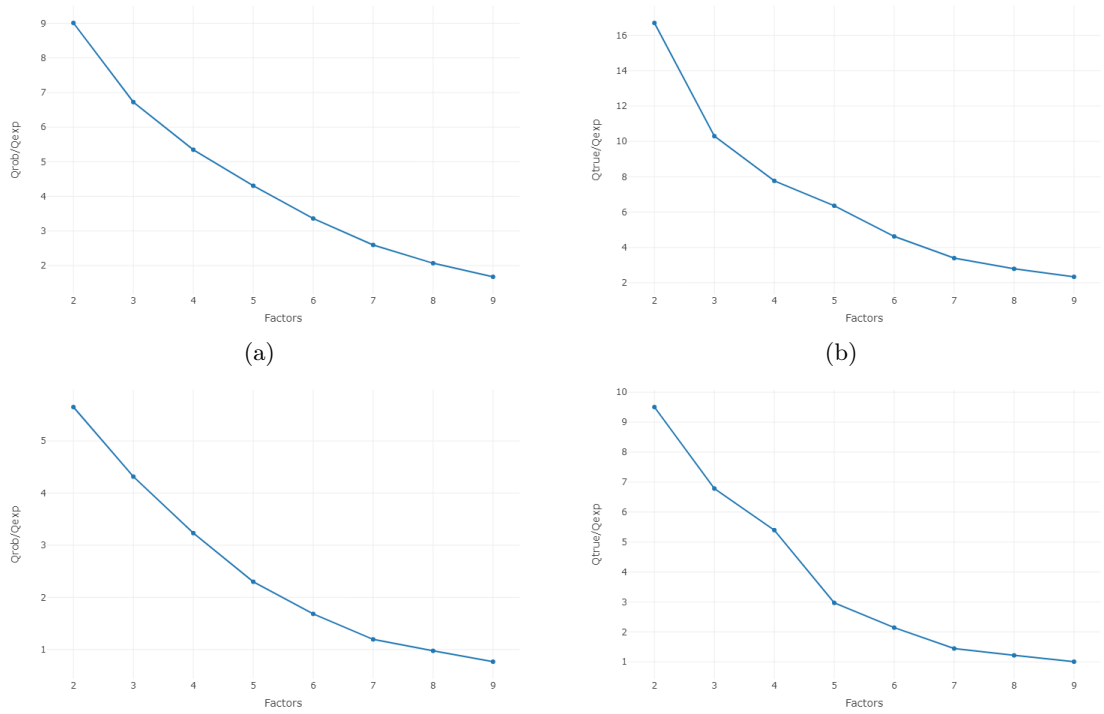


Figure 25: $Q/Q_{expected}$ as a function of factor ranking, where (a) and (c) is based on Q_{robust} mode and (b) and (d) is based on Q_{true} mode. Top row for Summer and bottom row for Winter season.

9.2 Summer (JJA)

The factor solutions with either 7 factors or 8 factors are compared where the 8 factor solutions explains some profiles that are also identified during the Winter season. Meanwhile, the 9 factors solution can be excluded by the reason that factor 9 explains mainly PCE and CH_2Cl_2, which are the same species explained by factor 1. Factor 9 is therefore duplicating factor 1 and furthermore the base model runs did not converge to the minimum Q value in all 20 base runs.

When comparing observed/predicted scatter plots from EPA-PMF, there are some species that are not explained and fitted by the model in the 8 factor solution. 10 species have poor model fit and a low correlation $R^2 < 0.6$ as indicated in Table 10 and are listed here: HFC-365mfc, CH_3Br, CH_3Cl, HCFC-22, HCFC-142b, CH_3CCl_3, CCl_4, SO_2F_2, CFC-114, and CFC-115. Furthermore, scaled residuals of three species are illustrated in Figure 26, and show a normal distribution within the interval -3 and 3. Scaled residuals of SO_2F_2 (Figure 26b) shows a more narrow distribution than toluene (Figure 26c), which can indicate that the uncertainties of SO_2F_2 is either underestimated or overestimated.

$Q/Q_{expected}$ is plotted in Figure 27 and used to effectively evaluate residuals of the PMF solution and understand which species are having high residuals ($Q/Q_{exp}>2$) which is not well explained in the model solution (Norris et al., 2014). As for instance, SO_2F_2 residual is 2.3 ($>$ 2) and also scaled residuals gives a narrow distribution, which means that further evaluation

in future analysis is needed. In addition, 3 high peaks in Q/Q_{exp} sample contribution were evaluated and removed. Finally, After thorough evaluation, both statistically and supported by expert knowledge, 8 factors are retained where factor profiles and contribution are plotted in Figure 28 to 35. To ease the reader it is decided to add a brief description of each figure and its mean within the figure's caption.

The factor profiles (also called source profiles) show a histogram indicating the concentration (log scale) of the species (on the left y-axis) apportioned to the factor/source. Additionally and more importantly, is the red squares explaining the percentage of species related to the factor/source (on the right y-axis). Finally, the factor contribution (also called source contribution) is illustrating a time series trend of factor/source. It is explaining the contribution of each sample to the factor p and the factor contribution is normalized (Norris et al., 2014).

Table 10: PMF model input data and diagnostics of Summer (JJA).

Base model run with 8 factors	
Model input data	
Samples	3306
Species	34
Factors	8
Base run	20
N of weak species	0
N of outliers	3
Fpeak	0
Model diagnostics	
N of species with $R^2 < 0.6$	10
Extra modeling uncertainty	0 % and 5 %
Q robust	177079 (0 %), 36819.2 (5 %)
Q true	238559 (0 %), 39335.7 (5 %)
Qrob / Qexp	2.0685 (0 %), 0.43010 (5 %)

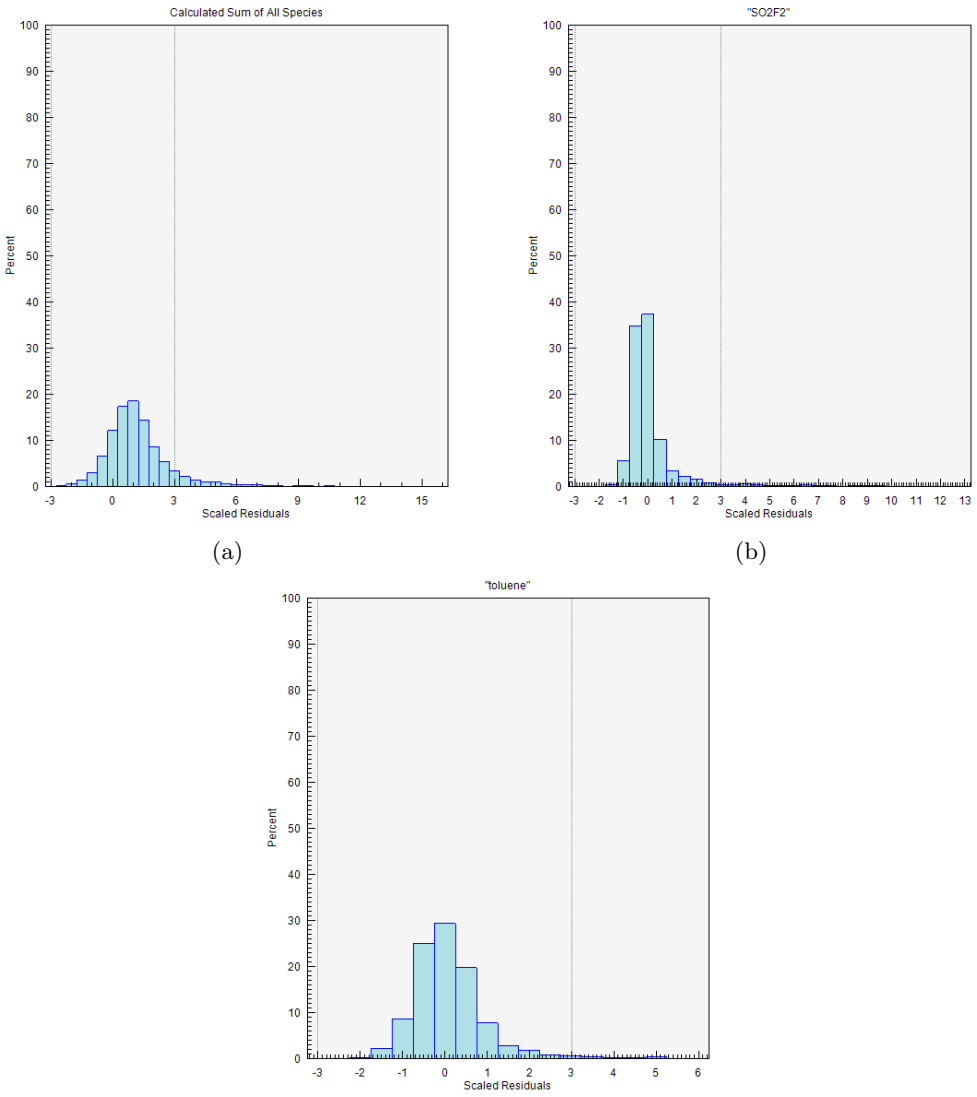


Figure 26: Scaled residuals of (a) all species, (b) SO_2F_2, and (c) toluene. Scaled residuals should be normally distributed and within $\pm 3\sigma$.

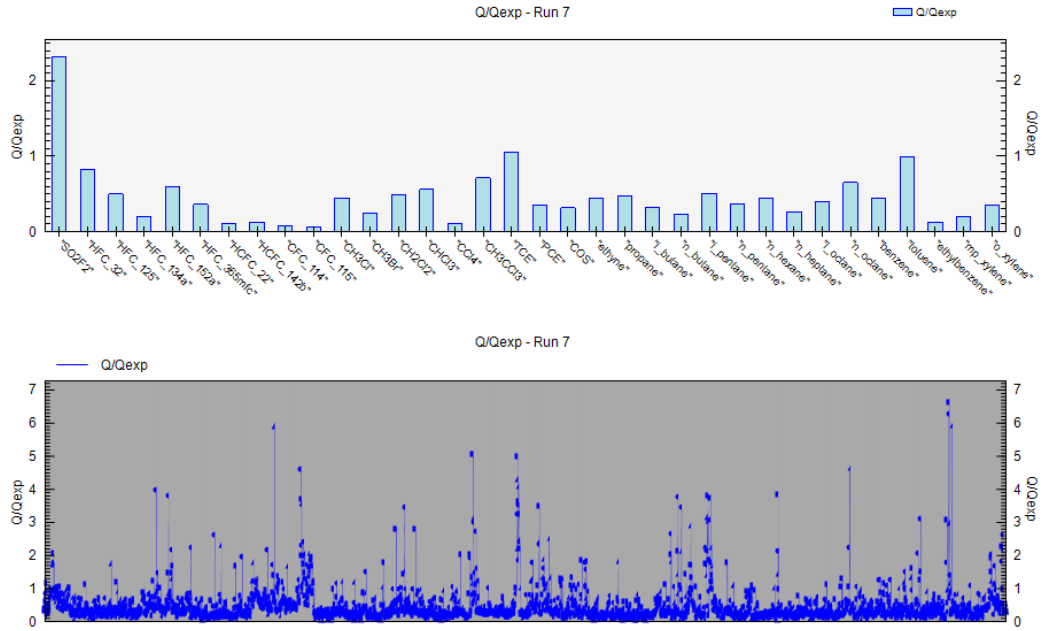


Figure 27: $Q/Q_{expected}$ of (a) factor profile and (b) factor contribution for a 8 factor solution. SO_2F_2 gives the highest residuals, where Q/Q_{exp} is 2.3.

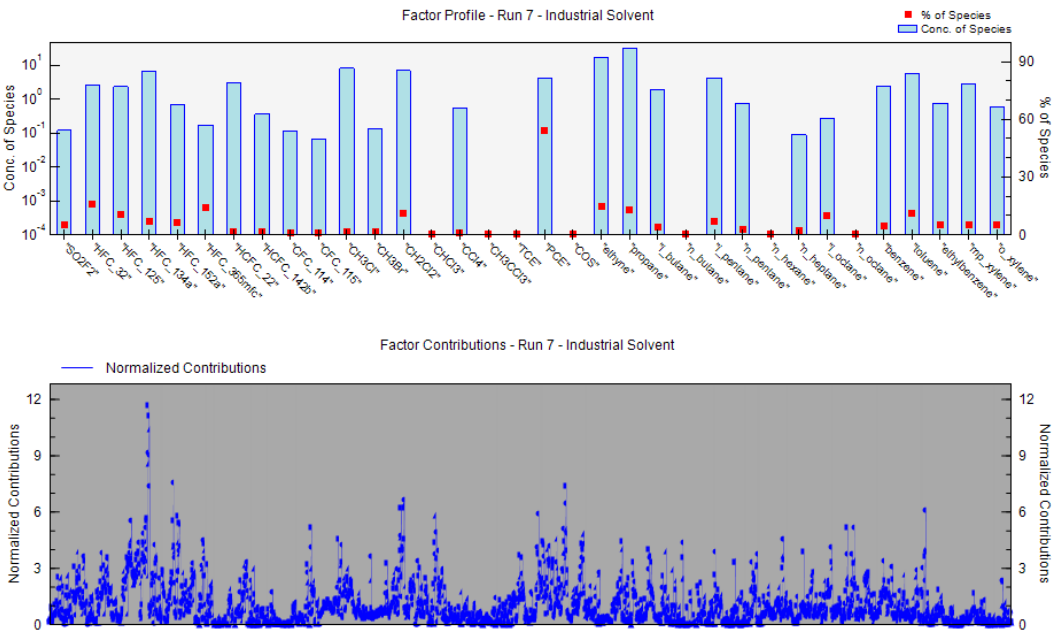


Figure 28: 54.1 % of PCE variability is related to factor 1 and is the main species explaining this factor, with a small attribution from HFC-32 (16.0 %), HFC-365mfc (13.8 %), CH_2Cl_2 (11.3 %), and ethyne (14.9 %) to mention a few. A high peak and higher annual variability is observed in factor contribution at the beginning of the time series.

Factor 2: Liquified Petroleum Gas

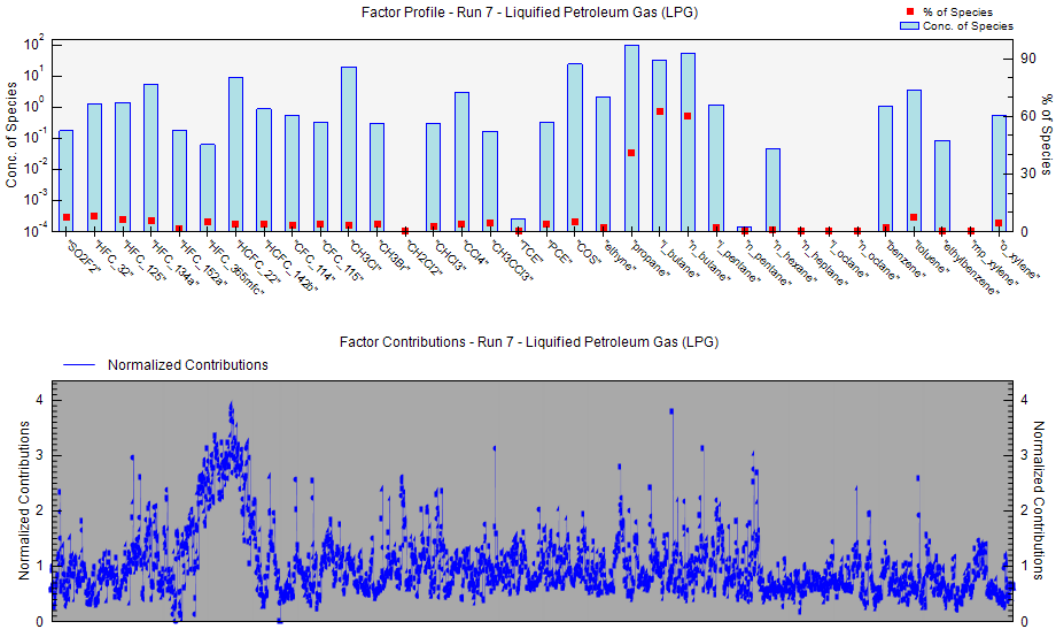


Figure 29: Factor profile illustrates mainly (i,n)-butane variability (62.7 % and 60.1 %). In addition, 41.1 % of propane variability is apportioned to factor this factor. In factor contribution, A high peak is observed early in the time series.

Factor 3: Halogens and GHGs

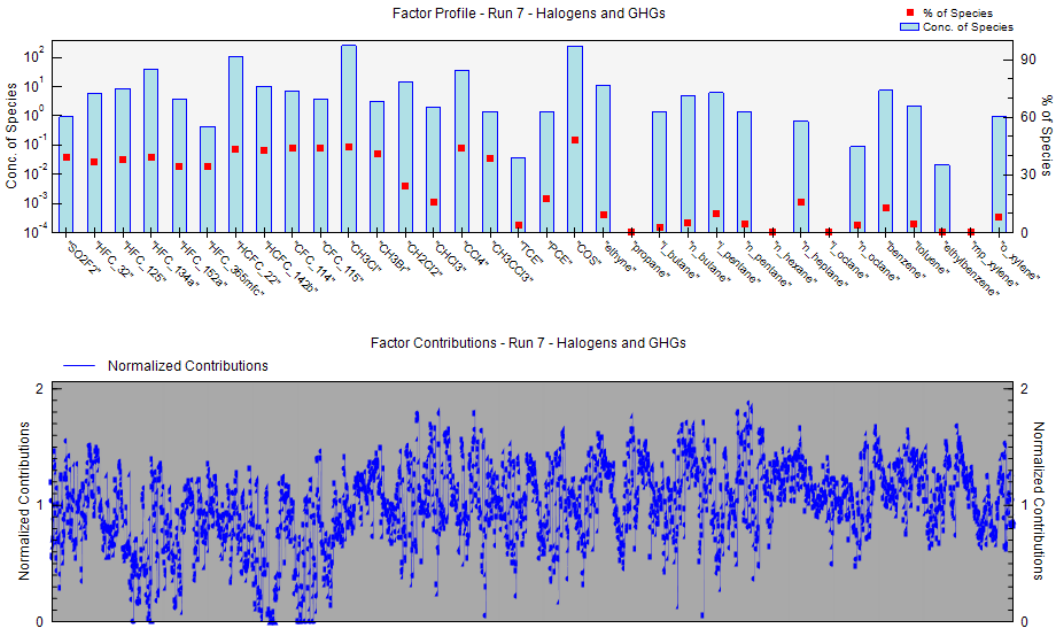


Figure 30: Factor 3 represents the halogens and COS explaining ≈ 45% of their variability. COS (48.3 %) and CH_3Cl (44.3 %) explains largest variability in factor 3. Meanwhile, $CHCl_3$ (16.0 %), CH_2Cl_2 (24.0 %), and PCE (17.8 %) contribute less as they are also apportioned to another factor. Factor contribution shows a very high variability.

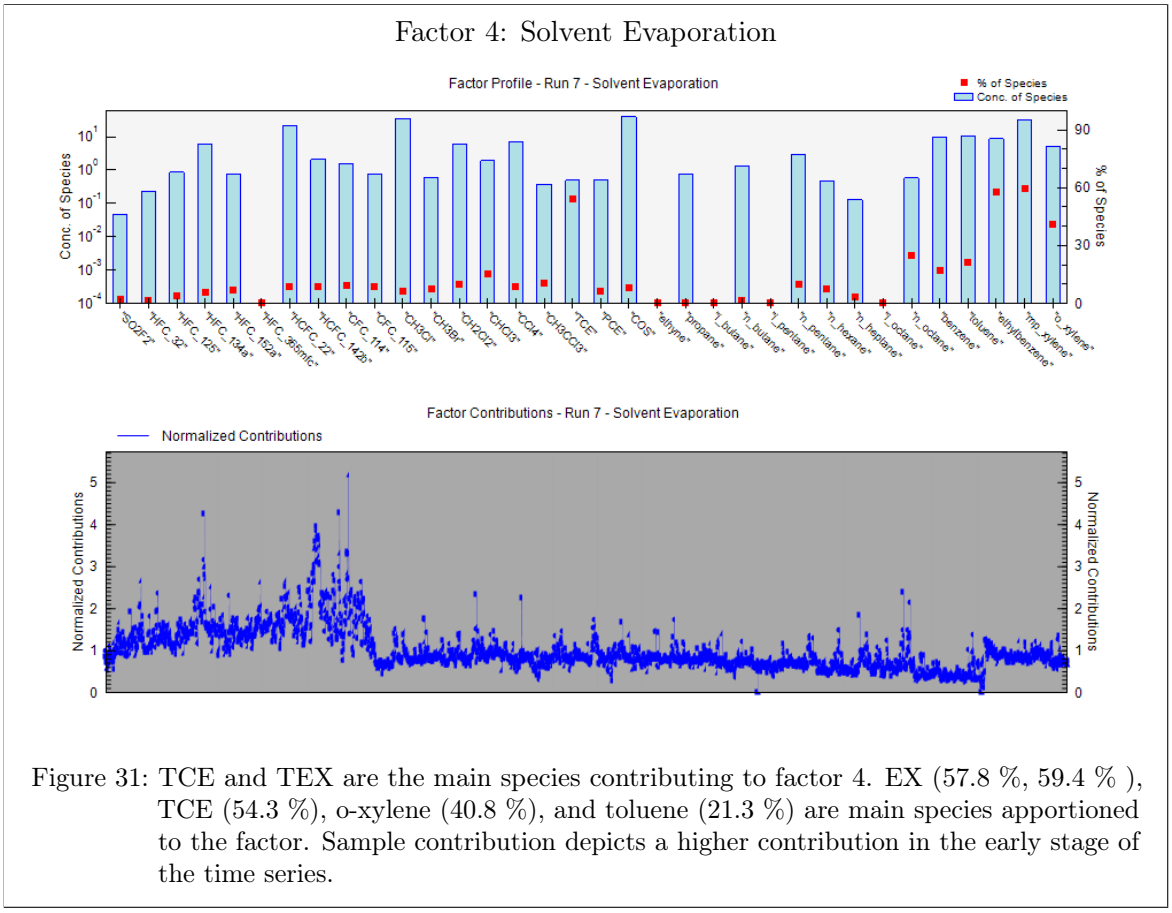


Figure 31: TCE and TEX are the main species contributing to factor 4. EX (57.8 %, 59.4 %), TCE (54.3 %), o-xylene (40.8 %), and toluene (21.3 %) are main species apportioned to the factor. Sample contribution depicts a higher contribution in the early stage of the time series.

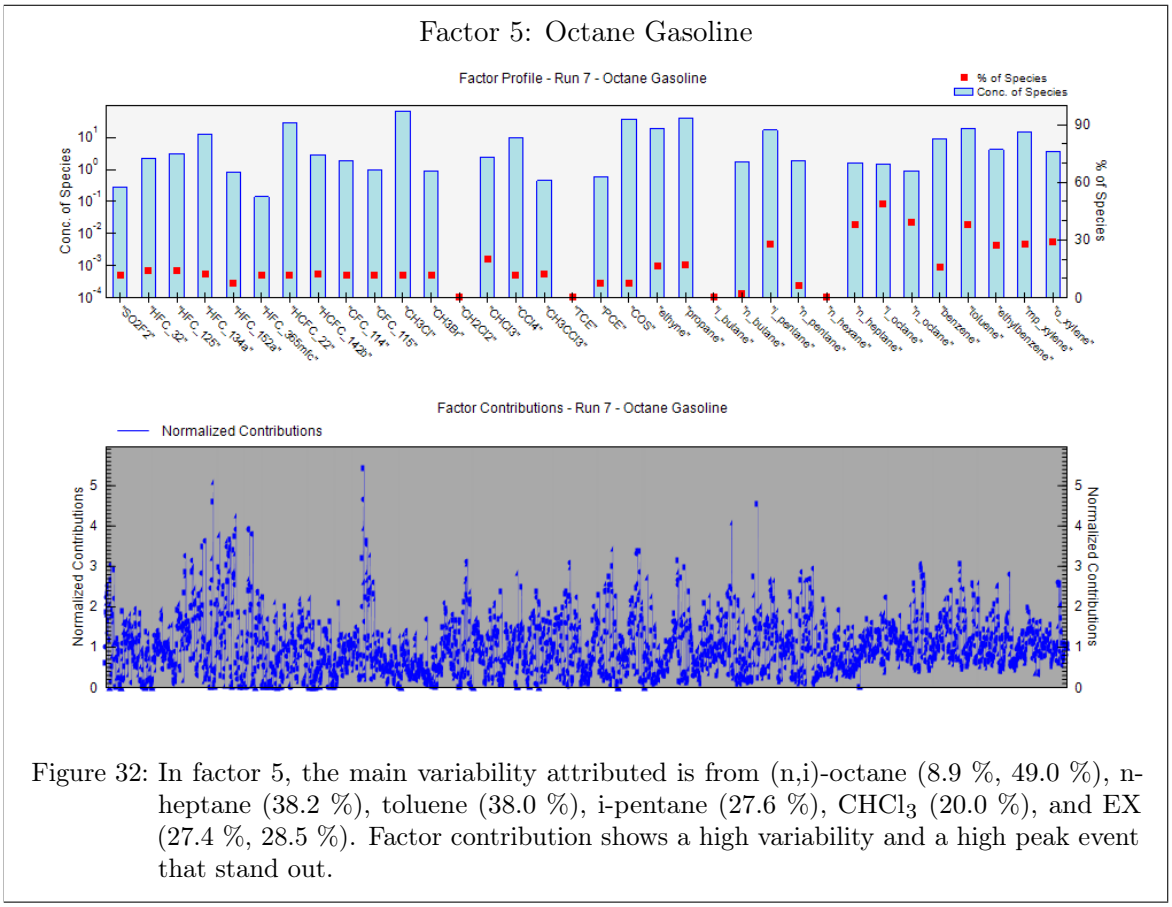


Figure 32: In factor 5, the main variability attributed is from (n,i)-octane (8.9 %, 49.0 %), n-heptane (38.2 %), toluene (38.0 %), i-pentane (27.6 %), $CHCl_3$ (20.0 %), and EX (27.4 %, 28.5 %). Factor contribution shows a high variability and a high peak event that stand out.

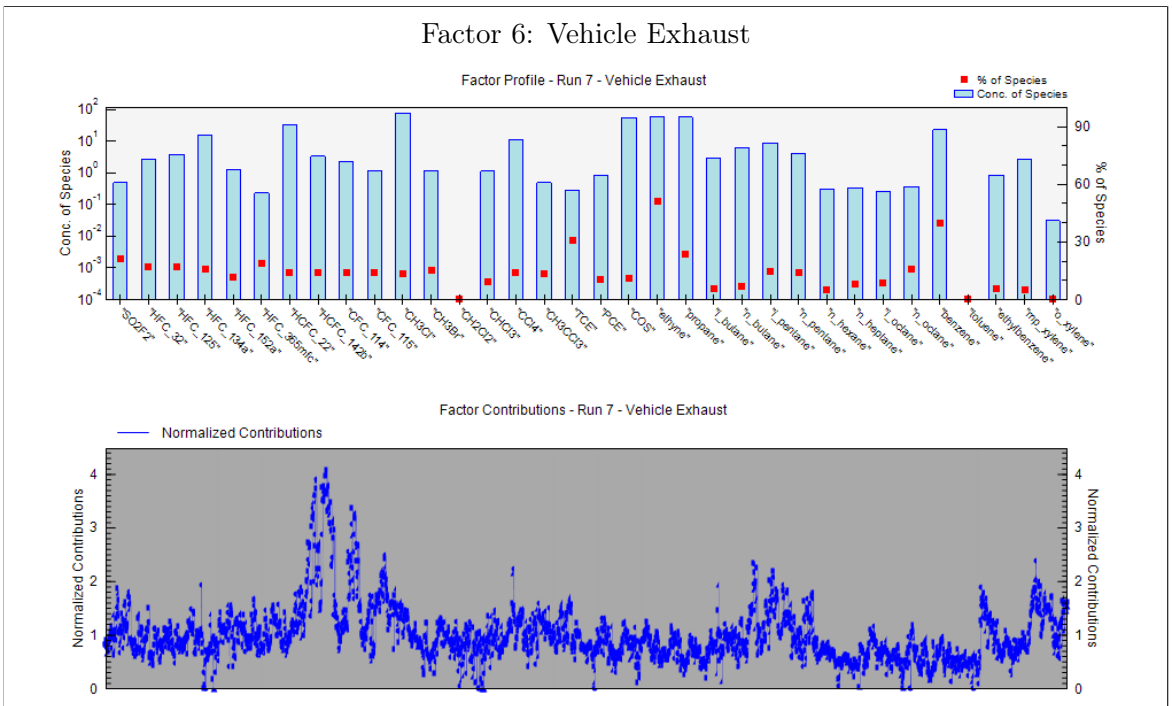


Figure 33: Ethyne (51.0 %) and benzene (39.9 %) attributes most to this factor Also, 30.9 % of TCE variability is apportioned factor 6.

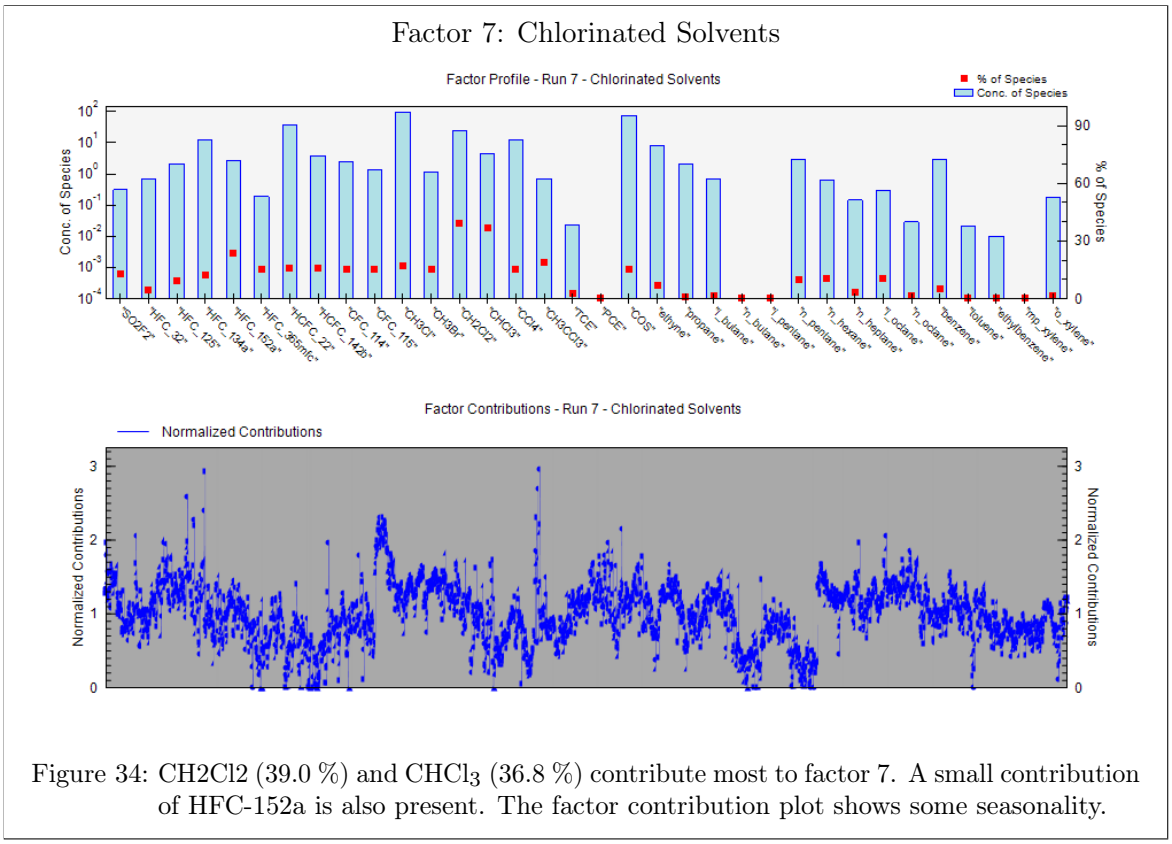


Figure 34: CH2Cl2 (39.0 %) and $CHCl_3$ (36.8 %) contribute most to factor 7. A small contribution of HFC-152a is also present. The factor contribution plot shows some seasonality.

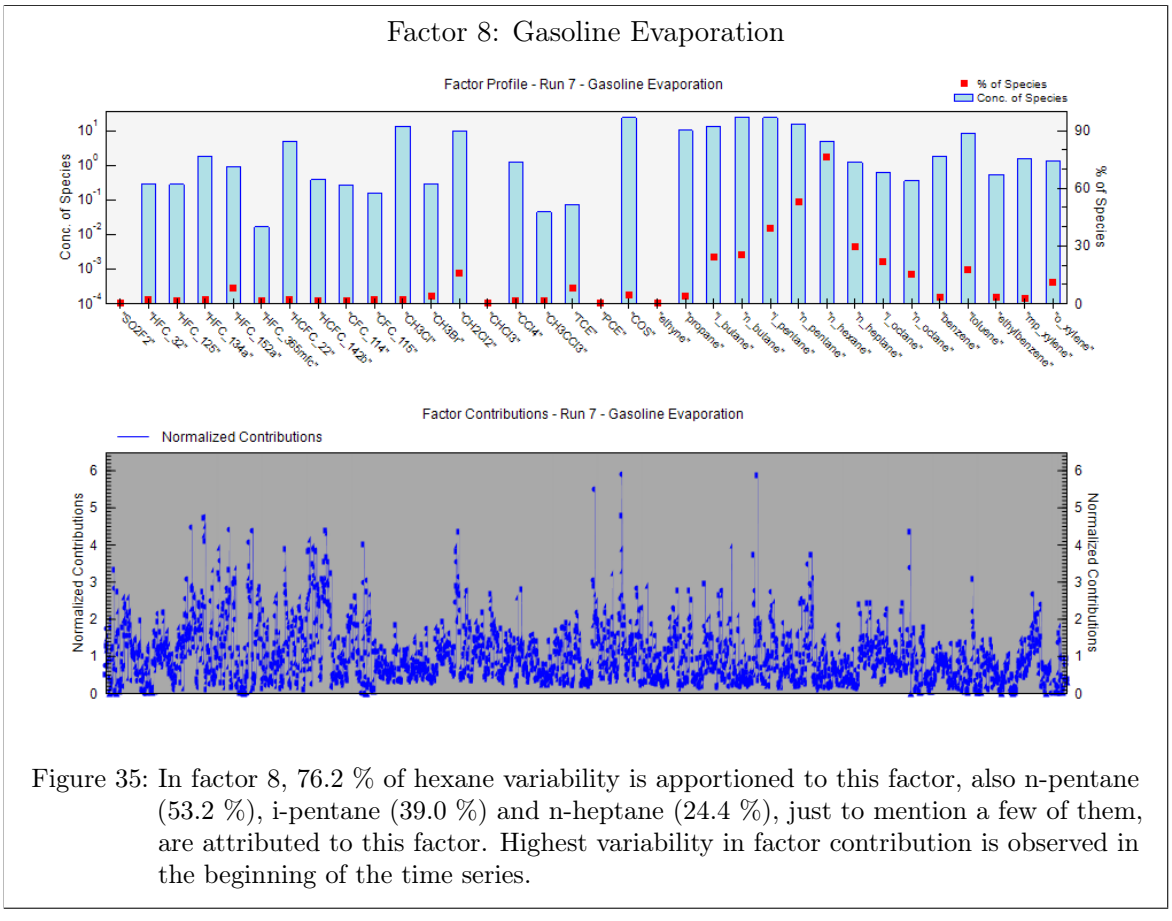


Figure 35: In factor 8, 76.2 % of hexane variability is apportioned to this factor, also n-pentane (53.2 %), i-pentane (39.0 %) and n-heptane (24.4 %), just to mention a few of them, are attributed to this factor. Highest variability in factor contribution is observed in the beginning of the time series.

Factor 9: PCE

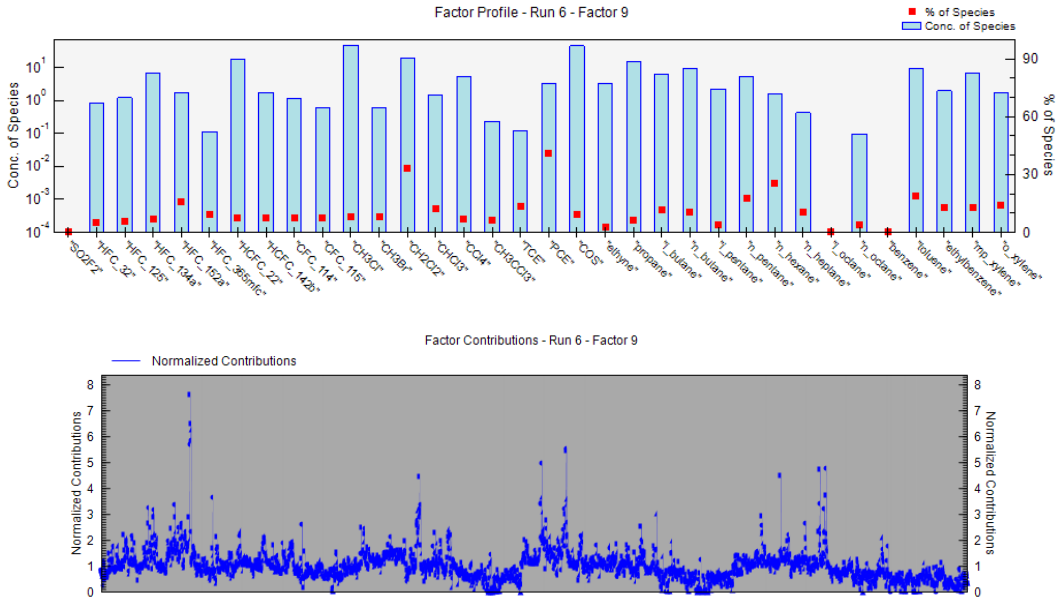


Figure 36: Factor 9 is only included in the results to argument to argument why 8 factor solution is better. PCE (40.8 %) and CH_2Cl_2 (33.1 %) are the main species contributing to the factor.

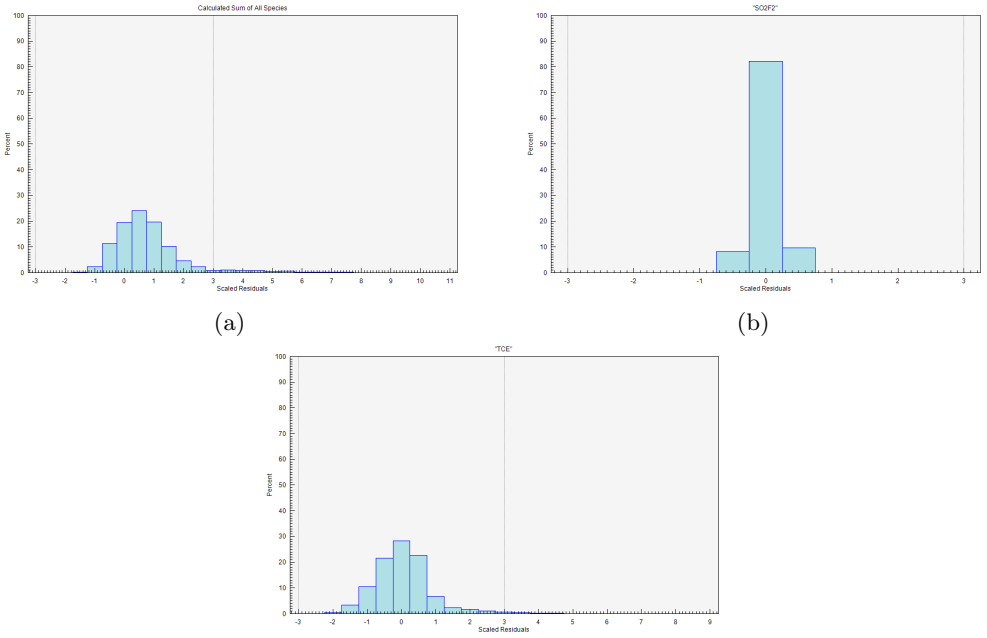
9.3 Winter (DJF)

For Winter season dataset, 7 factors model solution is obtained based on careful evaluation and interpretation of each factor. Although, the added factor explains mainly one single species (CH_2Cl_2) as depicted in Figure 43 it is retained in this analysis as it is also present during Summer season (Figure 34). Likewise, the 8 factor solution is not retained in this analysis as it mainly explains a relative small percentage of TCE and PCE species (Figure 46). Another argument to exclude 8 factor solution, is that TCE and PCE fingerprint is not observed during Summer season.

The model diagnostics are summarized in Table 11, and 10 species are not effectively modelled by PMF. This is evident when assessing the correlation of the observed/predicted values and poorly correlated species ($R^2 < 0.6$) are listed in the following order: COS, CH_3Br, CH_3Cl, HCFC-22, CH_3CCl_3, HCFC-142b, SO_2F_2, CFC-115, CCl_4, and CFC-114 (TCE correlation $R^2 = 0.60$). Furthermore, i-octane is set as a weak variable, for the reason that S/N ratio of i-octane is 1.2, thereby less than 2 (Table 9). Also, one high peak observed in Q/Q_{exp} factor contributions (Figure ??), has been deleted (2/5/2018 15:00). From the three scaled residuals plots (Figure 37) and Q/Q_{exp} residuals (Figure 38), the overall residuals of the model is acceptable. Scaled residuals of SO_2F_2 are very narrow and the correlation of observed/predicted model values are small, like the Q/Q_{exp} residuals. This could indicate that the scaled residuals of SO_2F_2 are underestimated and needs special attention. Highest Q/Q_{exp} residuals is observed for TCE (<2), but scaled residuals gives a normal distribution within $\pm 3\sigma$.

Table 11: Model input data and diagnostics of Winter season.

Base model run with 7 factors	
Model input data	
Samples	2856
Species	34
Factors	7
Base run	20
N of weak species	1
Q theoretical	76914
Fpeak	0
Model diagnostics	
N of species with $R^2 < 0.6$	10
Extra modeling uncertainty	0 % and 10 %
Q_{robust}	91900.3 (0 %), 16887.5 (5 %)
Q_{true}	107078 (0 %), 17211.7 (5 %)
Qtrue / Qexp	1.19 (0 %), 0.22 (5 %)



(a) (b)

(c)

Figure 37: Scaled residuals of (a) all species, (b) SO_2F_2, and (c) TCE. Scaled residuals should be normally distributed and within $\pm 3\sigma$.

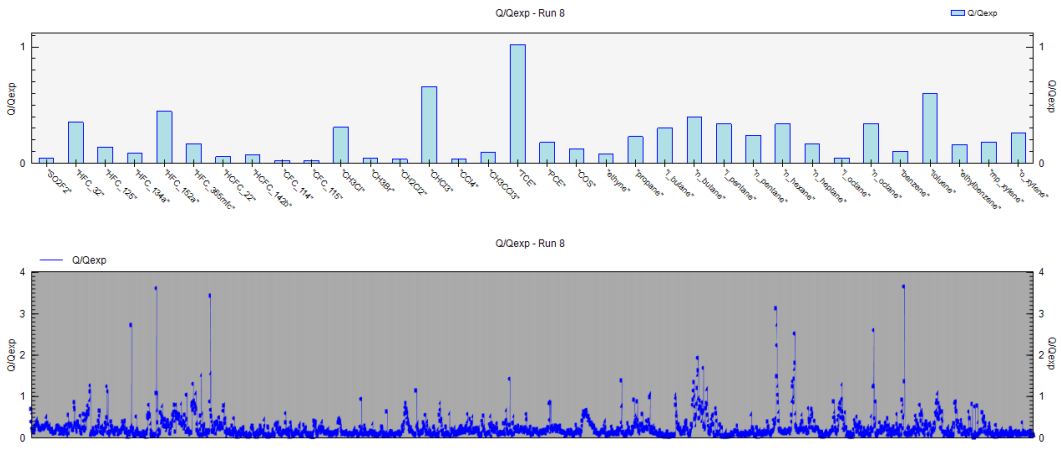


Figure 38: $Q/Q_{expected}$ of (a) factor profile and (b) factor contribution for a 7 factor solution

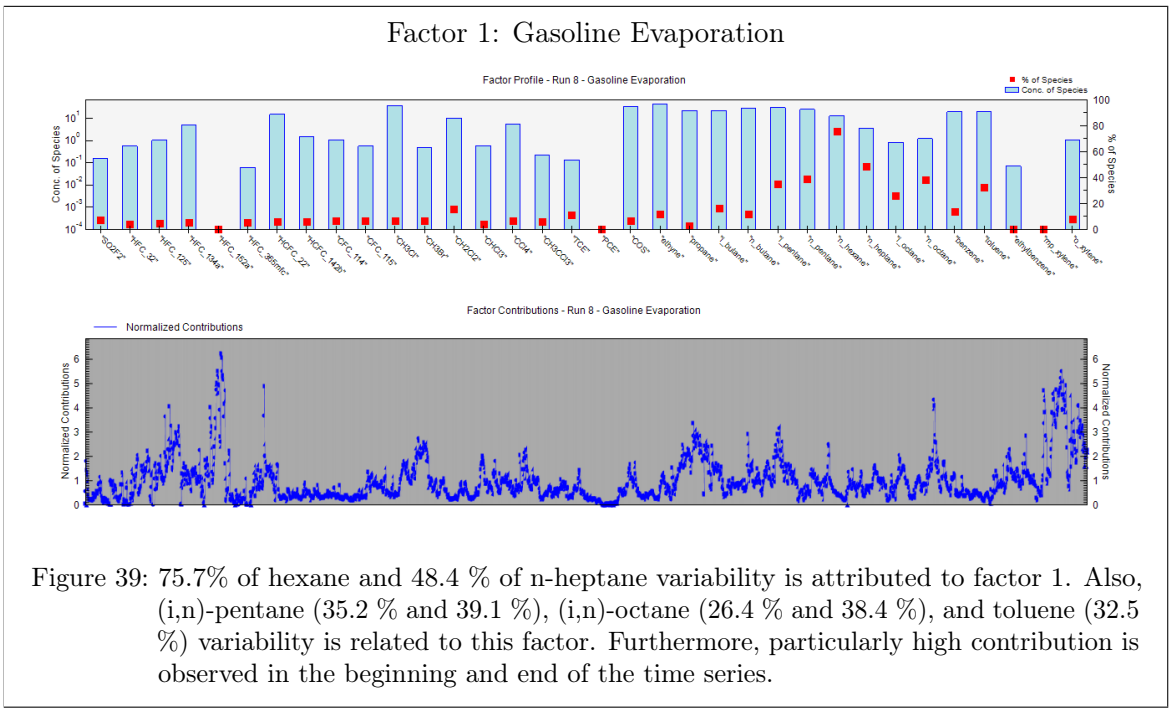


Figure 39: 75.7% of hexane and 48.4 % of n-heptane variability is attributed to factor 1. Also, (i,n)-pentane (35.2 % and 39.1 %), (i,n)-octane (26.4 % and 38.4 %), and toluene (32.5 %) variability is related to this factor. Furthermore, particularly high contribution is observed in the beginning and end of the time series.

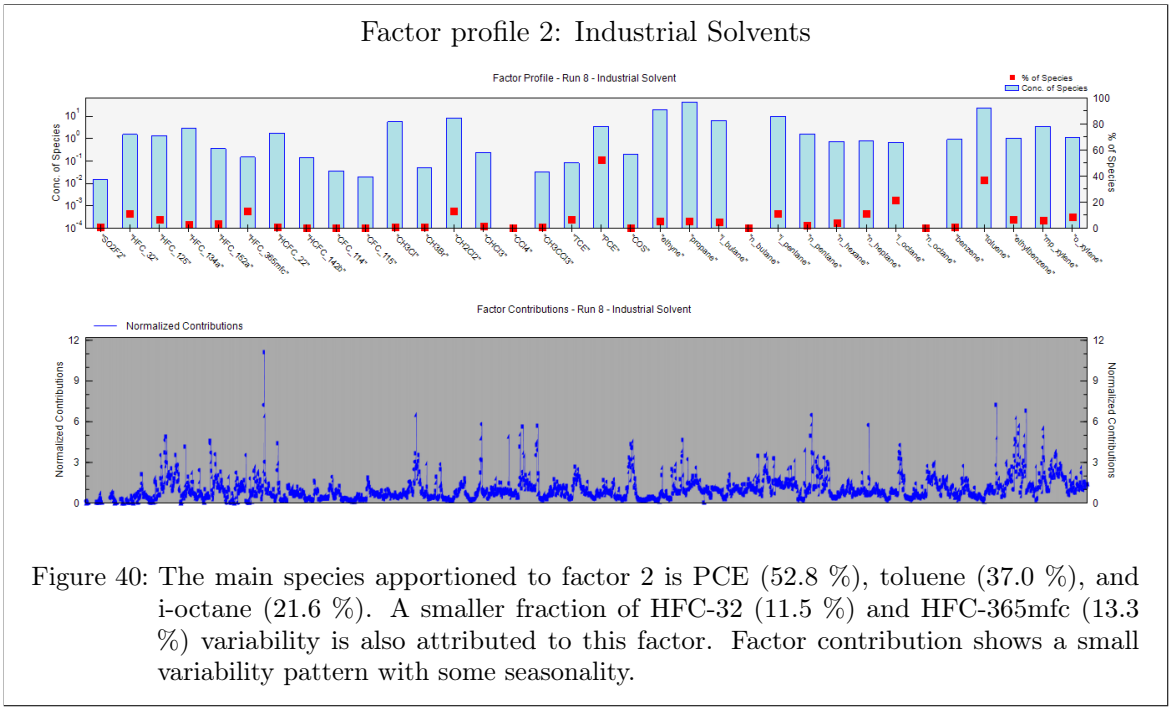


Figure 40: The main species apportioned to factor 2 is PCE (52.8 %), toluene (37.0 %), and i-octane (21.6 %). A smaller fraction of HFC-32 (11.5 %) and HFC-365mfc (13.3 %) variability is also attributed to this factor. Factor contribution shows a small variability pattern with some seasonality.

Factor 3: Vehicle Exhaust

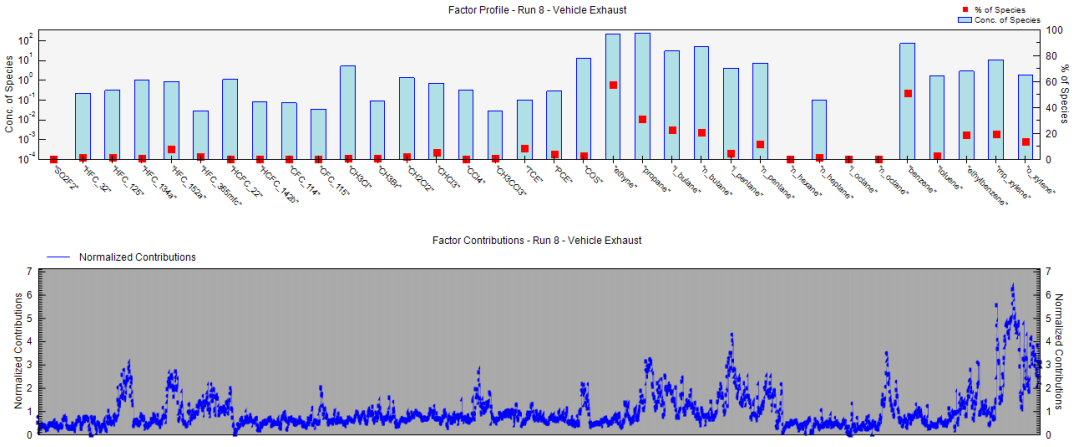


Figure 41: ethyne (57.9 %) and benzene (51.6 %) are the main species apportioned to factor 3. While a smaller fraction of variability is from (i,n)- butane (23.1 % and 21.1 %), propane (31.2 %), and EX (≈ 17.4 %) and related to the factor. In addition, factor contribution shows certain events of higher source contribution, mainly from ethyne and benzene.

Factor 4: Solvent Evaporation

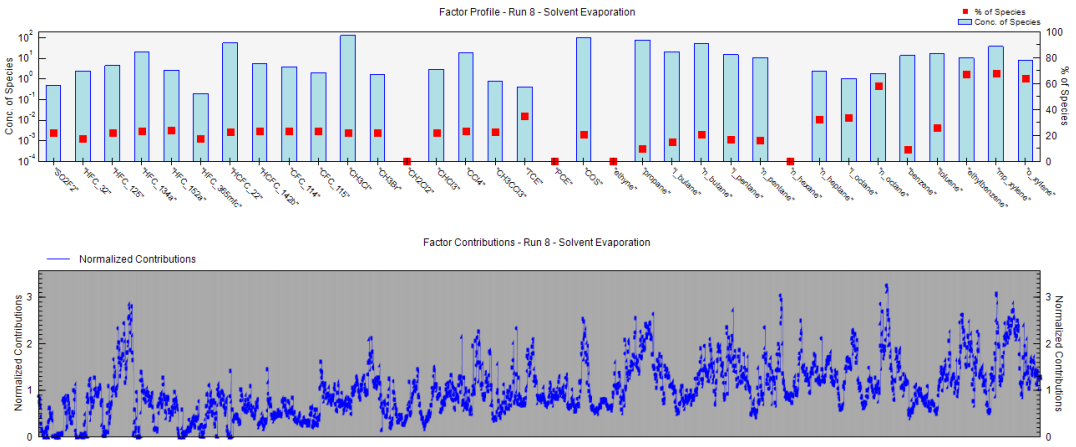


Figure 42: EX is apportioned to factor 4 by ≈ 66.5 % and other species related to this factors are (i,n)-octane (34.0 % and 58.6 %), TCE (35.0 %), n-heptane (32.7 %), and toluene (25.8 %). Factor contribution shows a high variability.

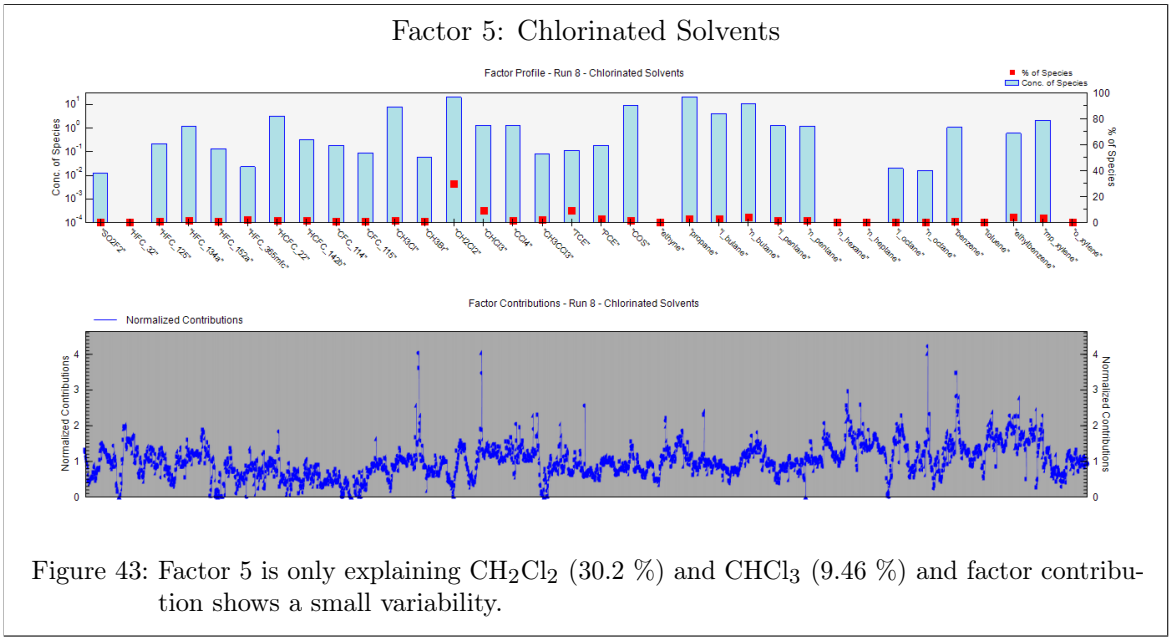


Figure 43: Factor 5 is only explaining CH_2Cl_2 (30.2 %) and $CHCl_3$ (9.46 %) and factor contribution shows a small variability.

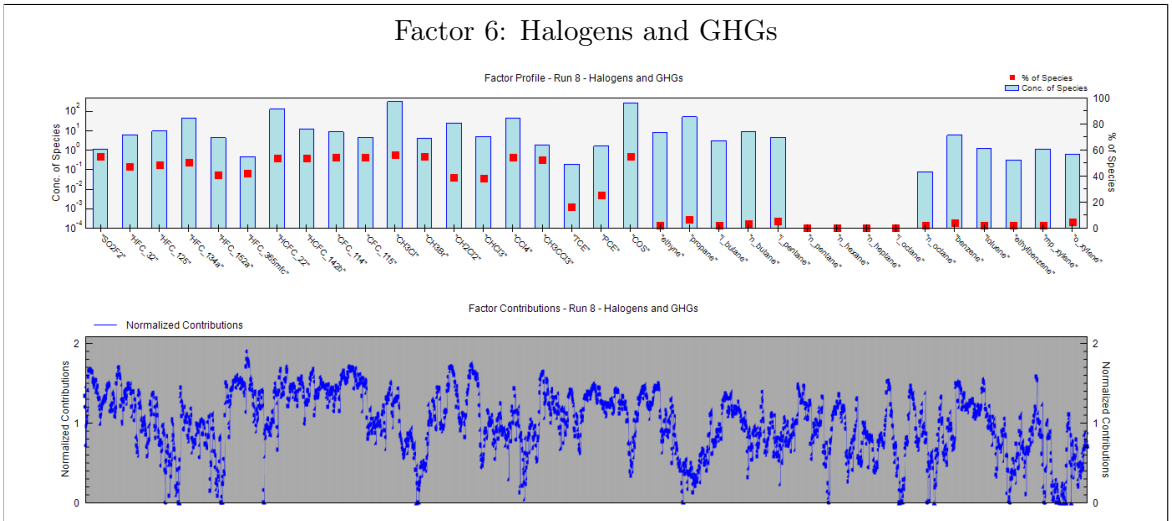


Figure 44: Factor 6 is mainly explaining halogens and COS (55.3 %). Among the halogens, CH_3Cl (56.2 %), CH_3Br (55.4 %), and SO_2F_2 (55.0 %) have the highest fraction of variability apportioned to this factor. Factor contribution shows a very high variability.

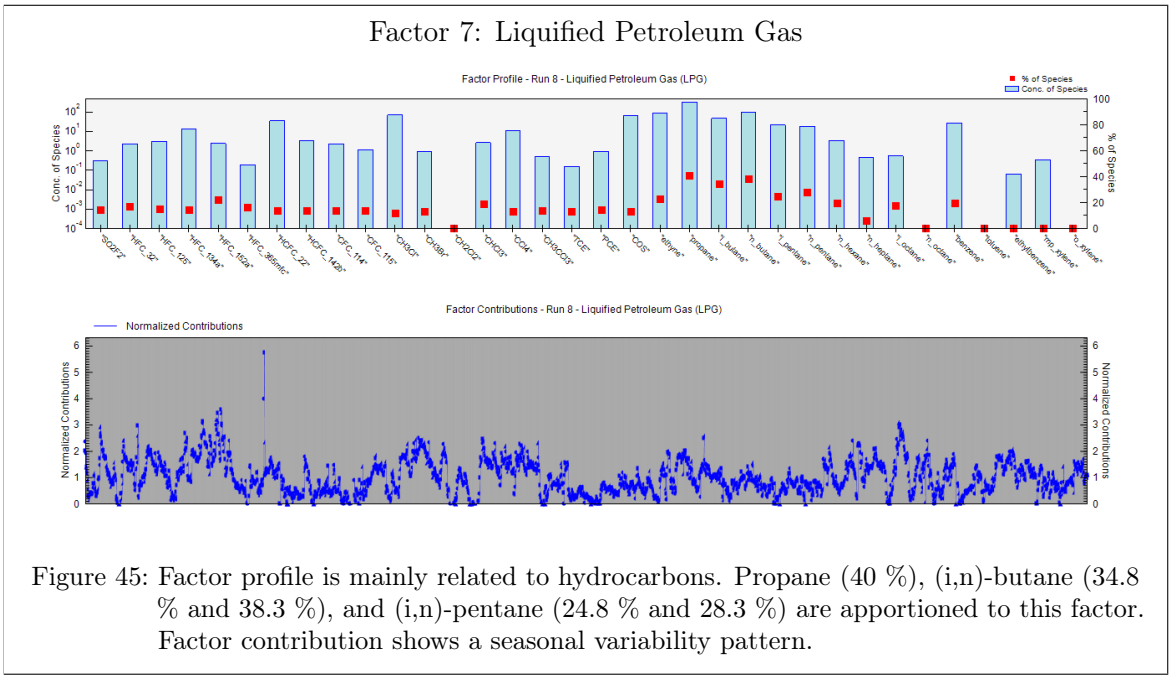


Figure 45: Factor profile is mainly related to hydrocarbons. Propane (40 %), (i,n)-butane (34.8 % and 38.3 %), and (i,n)-pentane (24.8 % and 28.3 %) are apportioned to this factor. Factor contribution shows a seasonal variability pattern.

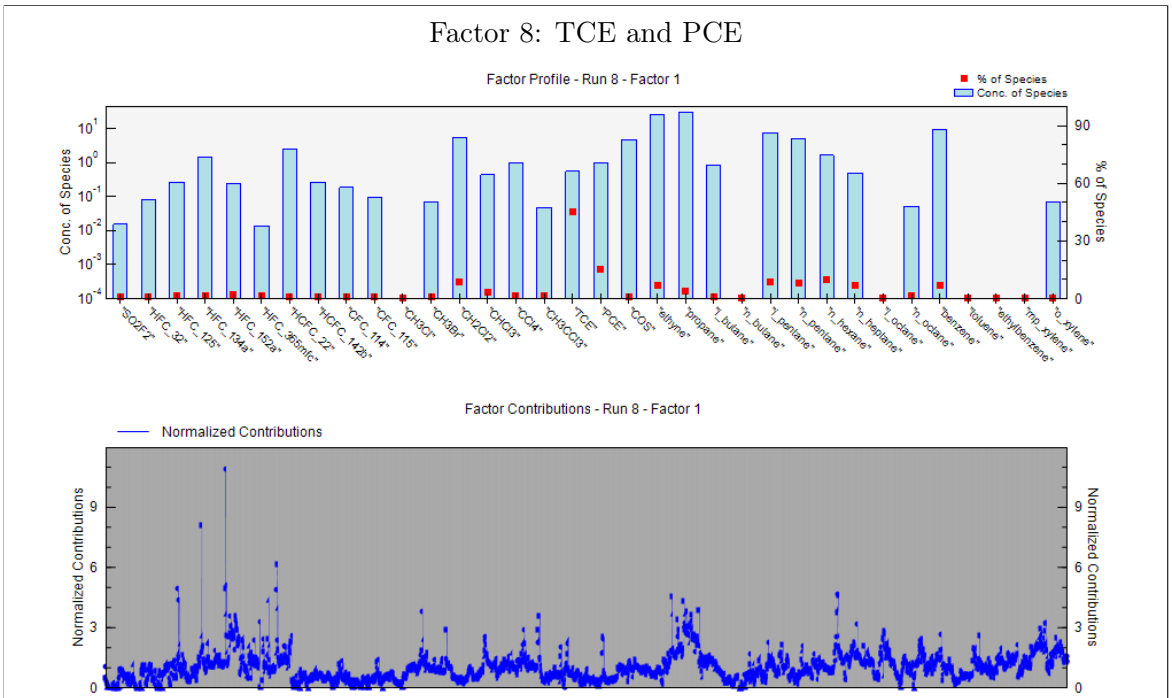
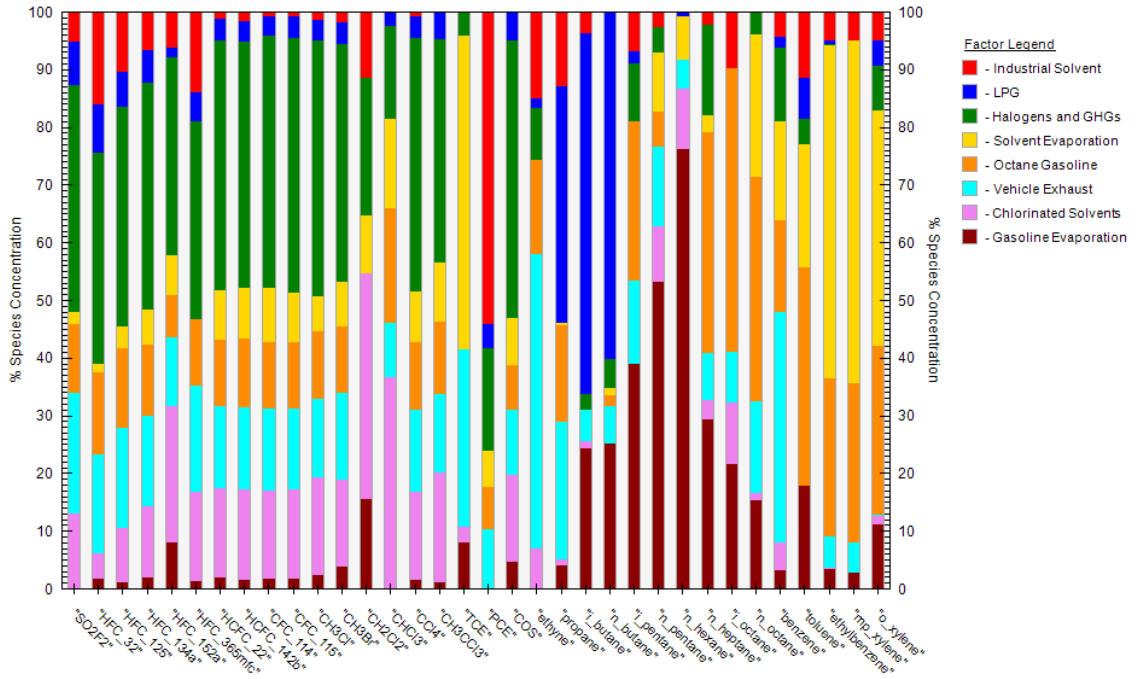


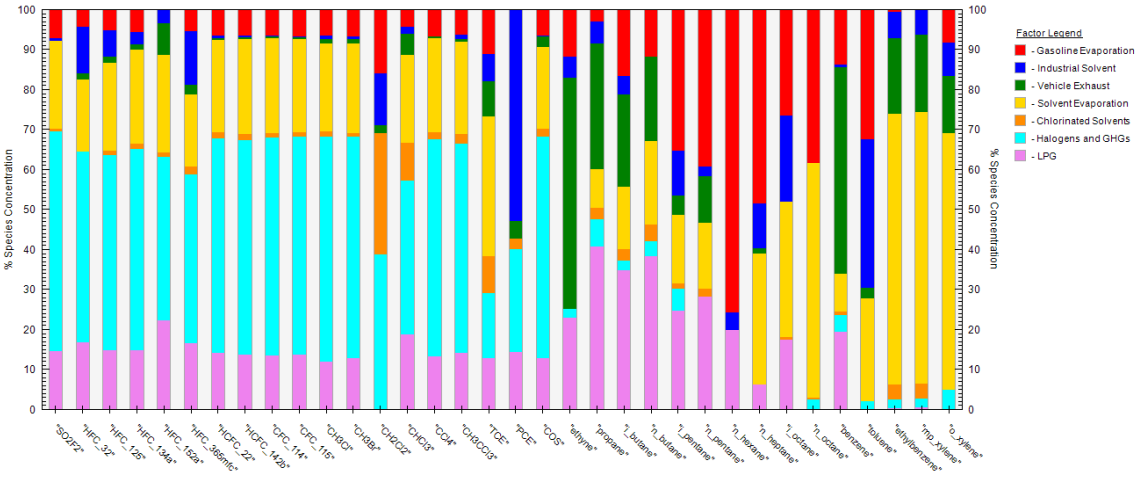
Figure 46: This factor is not selected as a factor solution and is only explained by TCE (45.1 %) and a small contribution from PCE (15.2 %). Factor contribution shows small variability only related to TCE.

9.4 Factor fingerprints

All Factor profiles results are now gathered in one figure. Figure 47(a) for Summer season with 8 factors and Figure 47(b) for Winter with 7 factors. It is plotted to better understand the distribution of the factor profiles for each species (Norris et al., 2014).



(a)



(b)

Figure 47: Factor fingerprints are depicted in (a) for Summer (JJA) and (b) for Winter (DJF).

Part V

DISCUSSION AND CONCLUSION

10

DISCUSSION

In general, during Summer the measured concentrations will be less influenced by aged air masses and more affected by regional and Italian emissions. The age of sampled air masses measured at CMN during summer is calculated to be on average ≈ 12 hours. On the contrary, the lower OH concentration during Winter caused by a decrease in UV-light, results in a longer lifetime of NMVOCs. For this same reason, in order to reduce the variability on the whole dataset due to the changing lifetime (seasonal cycle), it is easier to interpret the PMF results on Summer and Winter season, respectively.

Several challenges are met during the initial phase of this work that affect the PMF model output and interpretation of the factors (i.e. missing values, uncertainties, seasonality, and trends). Among others, one is related to having a multi year dataset. The expected outcome can be a factor explaining mainly the multi-year trend of the observed species as well as the seasonal variability of the sources. Species time series trends, underlying the whole time period (2015-2018) considered, are removed to make sure that resulting factors will not be biased by the trends themselves, as it had been found in the first attempt to perform explorative analysis by PCA.

From PMF diagnostics, a total of ten "problematic species" are identified and for Summer season they are: HFC-365mfc, CH_3Br, CH_3Cl, HCFC-22, HCFC-142b, CH_3CCl_3, CCl_4, SO_2F_2, CFC-114, and CFC-115. For Winter season: COS, CH_3Br, CH_3Cl, HCFC-22, CH_3CCl_3, HCFC-142b, SO_2F_2, CFC-115, CCl_4, and CFC-114. Same species are identified as "problematic species" in cluster analysis and PCA. In PMF they are not effectively modelled, which means that sources attributed to these species needs to be carefully interpreted.

To determine the appropriate number of factors, model statistics from PCA and PMF was first thoroughly analyzed. The optimum number of principal components (PCs) during Summer is 7 and 8 PCs, explaining up to 81.53 % of total variance. Whereas, 6 and 7 PCs are retained for Winter season, explaining up to 78.27 % total data variance. Same number of emission factors were identified in PMF analysis suggesting that although PCA is sensitive to certain "problematic species", potential sample outliers, and long-term trends in data, it can support PMF analysis in determining the optimum number of factors to retain.

10.1 Source apportionment

From the PMF results, 8 emission factors are identified in Summer season and 7 emission factors are identified in Winter season. There is a large subjectivity in the interpretation of

the emission factors and will require a deeper argumentation and comparison with source models to be more sustainable. The following discussion is therefore a first attempt to classify factors' sources on CMN, mainly by comparing the obtained results with literature profiles and cluster analysis.

10.1.1 Vehicle exhaust

"Vehicle exhaust" factor is both present during Summer and Winter as shown in Figure 33 and Figure 41, respectively. Ethyne and benzene are the main species apportioned to this factor explaining $\approx 40-58\%$ of species variability. From cluster analysis (Chapter 6.1.2), benzene and ethyne are highly correlated.

Summer factor of "vehicle exhaust", include more species variability especially from the background species (halogens and COS) compared to Winter factor. Comparing factor contribution, they have a slightly different seasonal variability.

Research shows, that propane, ethyne, and benzene are linked with fossil fuel consumption (Lo Vullo et al., 2015). Moreover, Lo Vullo et al., 2016 perform PCA at CMN site, and identifies ethyne and benzene main source to be vehicle exhaust. Another recent source apportionment study using PMF is by Debevec et al., 2020. The receptor site in this study is located at Corsica Island, near the western coast of Italy where a combustion factor is identified explained mainly by ethyne and benzene. Finally, Debevec et al., 2020 classifies the main annual source to be coming from residential heating, due to the low contribution from toluene to factor.

To conclude, this factor is source categorized to be from vehicle exhaust, but could also include other combustion sources such as residential heating. Furthermore, benzene and ethyne have the longest lifetime of all NMVOCs, 9.5 to 14 days respectively (Table 1), which are referred to as aged air masses (Lo Vullo et al., 2016). Other medium long-lived NMVOCs species are propane, (i,n)-butanes, and (i,n)-pentanes (Table 1). Long-lived NMVOCs and aged air masses are more related to long-range transport than from regional emissions (Lo Vullo et al., 2015).

10.1.2 Halogens and non-CO_2 GHGs

"Halogens and GHGs" factor is separating the halogens and COS from most NMVOCs by their differences in variability. This factor is observed both during Summer and Winter as illustrated Figure 30 and Figure 44, respectively. PCA also identifies this factor, which explain most variance in the data. In addition, this factor was present in every PMF solution, when running the explorative PMF analysis using 2 to 9 factor solutions.

This factor is not representing a particular source, but is interpreted as the result of the variability on a continental scale. This is explained by the fact that halogens and COS have a longer lifetime and can be transported over long scale distances. Nevertheless, it is interesting to note that a not negligible fraction of the halogenated species variability is attributed to factors "Solvent evaporation" (Section 9.3, Figure 42) and "LPG" (Section 9.3, Figure 45) in

Winter, pointing towards a spatial co-emission with the typical NMVOCs related to these factors.

Furthermore, it demonstrates the challenges associated with applying PMF on NMVOCs, halogens, and COS all together and the impact of a multi year dataset which may only explain seasonal variability.

10.1.3 Gasoline evaporation

"Gasoline evaporation" factor is both present during Summer (Figure 35) and Winter (Figure 39). The main species apportioned to this factor are n-hexane, n-heptane, (i,n)-pentane, toluene, and (i,n)-octane. In general, a slightly higher source contribution is observed during Summer season, confirming the fact that evaporative processes are more efficient with higher ambient temperature.

Hexane is the main species attributed to this factor, contributing with $\approx$ 60 % of its variability. According to Table 1, they are all evaporative sources of gasoline fuel. Debevec et al., 2020 identifies an evaporative source factor which fingerprints are comparable to this "gasoline evaporation" factor.

By comparing obtained factor with cluster analysis, it is evident that n-hexane, n-heptane, (i,n)-pentane, and n-octane are correlated in Winter season. While during Summer, n-hexane is correlated with (n,i)-pentane. To conclude, cluster analysis is consistent with obtained factor fingerprints.

10.2 Liquified petroleum gas

"LPG" (liquefied petroleum gas) factor is mainly related to hydrocarbons and are both present during Summer (29) and Winter (45). Propane and (i,n)-butane are the main species apportioned to this factor, $\approx$ 60 % during Summer and $\approx$ 40 % during Winter. Also, a smaller fraction of explained variability from (i,n)-pentane is especially present during summer (40.1 %). Overall, there is a higher species contribution to "LPG" source in Summer compared to Winter. This factor is identified as "LPG" factor because butanes and propane are the main constituents in LPG and these alkanes are all identified in LPG composition. In general, propane is an important constituent in natural gas use (Debevec et al., 2020). Cluster analysis show a high correlation of propane and (i,n)-butanes and the cluster is present in both seasons which demonstrates the obtained fingerprints.

10.2.1 Solvent evaporation

"Solvent evaporation" factor is present in both seasons as represented in Figure 31 and Figure 42, respectively. TCE and TEX are the main species contributing to this factor. Ethylbenzene and xylenes (EX) variability explained in this factor, is higher during Winter ($\approx$ 65.5 %) than during Summer season, whereby TCE is contributing more during Summer season (54.3 %). Also, n-octane is apportioned to this factor with more variability explained during Winter

(58.6 %). This is also evident in cluster analysis, where in Winter TCE and TEX are clustered together and in Summer n-octane and EX are clustered together.

The time series contributions illustrates different trends during Summer and Winter, with a higher variability observed during Winter. While for Summer season, a higher contribution is observed in the early stage of the time series followed by an abrupt decrease. Since the times series contributions show Winter and Summer separately, the abrupt changes in variability are more related to the difference in annual variability than actual events.

TEX and n-octane have the shortest lifetime of all NMVOCs, due to a higher OH rate constant (Table 1), and are therefore expected to be found closer to the emission source. This factor can therefore be representing short-lived species. Furthermore, the main source of TEX and TCE emissions is from solvent usage (Table 1), and this factor can therefore be indicating the strength of "solvent use" sector (Lo Vullo et al., 2015).

10.2.2 Industrial solvents

PCE is the main species apportioned to "industrial solvents" factor $\approx$ 53 % and a slightly greater amount of TCE variability is explained in Summer. Although the obtained fingerprints for Summer and Winter are very similar as depicted in Figure 28 and (Figure 40).

PCE is mainly used in industry as a solvent (Table 2), and this is also the case for CH_2Cl_2, toluene, and i-octane that are also apportioned to this factor, but in a smaller amount. During Winter, toluene is explaining a higher variability (37.0 %) than in Summer. Furthermore, a small variability explained by this factor is from two HFCs, namely HFC-32 and HFC-365mfc. PCE is also used as a solvent for manufacturing refrigerants. Therefore, this factor could potentially explain industrial solvent usage, although there is not enough information to identify the HFCs.

From cluster analysis, the dendrogram of Winter season show that PCE, HFC-365mfc, and HFC-32 belong to the same cluster family.

10.2.3 Chlorinated solvents

The main variability attributed to "chlorinated solvents" factor is from CH_2Cl_2 ($\approx$ 30 to 40 %) and $CHCl_3$. Higher percentage of species variability is explained in Summer (Figure 34), in particular $CHCl_3$ explaining 36.8 % Summer and 9.46 % during Winter, compared to Winter (Figure 43). They are both industrial solvents, also used to manufacture HFCs (Table 2) and the results indicate a higher solvent evaporation in Summer compare to Winter.

A small contribution from HFC-152a is observed in Summer and from the dendrogram of Summer season, CH_2Cl_2 is clustered together with HFC-152a and correlated.

It can be argued that this factor is similar to the above "industrial solvent" factor. Therefore, the factor is named "chlorinated solvents" to distinguish from the other identified solvent factors.

10.2.4 Octane gasoline

"Octane gasoline" factor is only obtained from PMF on Summer season (Figure 32) and the time series contribution to this factor shows a very high variability.

I-octane (49.0 %) explains the main variability of this factor. Also, toluene, n-heptane, i-pentane and EX are contributing to this factor. From cluster analysis of Summer season, these species are all clustered together and especially i-octane, n-heptane and toluene are highly correlated. These compounds source can be identified as gasoline surrogates and for this reason the factor is named "octane gasoline" (Knop et al., 2014).

11

CONCLUSION

The main objectives of the study was to identify source contribution of NMVOCs, halogens and non-CO_2 GHGs at a remote mountain site using positive matrix factorization (PMF) as a source apportionment method. One of the main topic discussed in this work was how to consider the impact of hydroxyl (OH) radical on NMVOCs seasonal cycle from the emission source to the receptor. However, how to integrate a NMVOCs lifetime correction method in a source apportionment study using PMF, remains a gap in literature. Therefore, this work contributes to the development of integrating a lifetime correction method with PMF, in order to improve source apportionment of the most reactive NMVOCs measured at a remote site.

This source apportionment study consist of several phases. The initial phase included validation of data, quality checks, basic statistics, and observations of time series plots. Along the process, several caveats have been found, where decisions had to be taken and thoroughly evaluated, to make sure not to affect the final model output. This implied removal or data filling of missing values with best estimates also for associated data uncertainties, and removal of long-term trends from time series to aid interpreting the final results.

As already stated, one of the crucial questions discussed in this work, is the effect of applying a lifetime correction method on processed NMVOCs data. The average photochemical age of air mass reaching CMN was estimated to be $\approx$ 12 hours during Summer. A comparison analysis of NMVOCs with and without the lifetime correction method, revealed that reactive NMVOCs mixing ratios increased significantly after applied lifetime correction method. This implies, that a higher source contribution can be attributed to the reactive species. Moreover, the lifetime correction method of NMVOCs was also evaluated in cluster analysis, which was the second phase of the source apportionment study.

To discover correlation among species and groups within the data, hierarchical agglomerative cluster analysis based on Pearson correlation was performed on Summer and Winter season. The main differences observed by comparing dendrograms with and without lifetime correction, was a rearrangement of the reactive NMVOCs species and groups, according to the different seasons. With applied lifetime correction, the three most reactive species (EX and n-octane) are clustered together during Summer and dissimilar to long-lived NMVOCs such as ethyne and benzene.

Thereafter, an exploratory analysis was carried out using principal component analysis (PCA) and a comparison with cluster analysis confirmed that PCA is data-sensitive towards ten "problematic" species and failed to explain the variability of the other species. Same species are determined as "outliers" in cluster analysis and not effectively modelled in PMF. Nevertheless, the preliminary and exploratory evaluation helped in understanding species

behaviour and correlations, which was essential when identifying the source factors obtained by the PMF.

Final phase of the source apportionment study was the PMF analysis, that proved to be a valuable multivariate analysis tool for source categorization of the 34 measured species. The number of identified factors were 8 during Summer and 7 during Winter, all describing emission sectors except for one factor. This factor represented halogenated species and non-CO_2 GHGs and was present in both seasons explaining their variability on a continental scale. The other source factors are: (1) vehicle exhaust; (2) gasoline evaporation; (3) Liquified petroleum gas; (4) solvent evaporation; (5) industrial solvents; (6) chlorinated solvents; and (7) octane gasoline.

The obtained factors were critically selected, interpreted and in agreement with literature profiles. Although, there is a degree of subjectivity involved when interpreting the results and determining the number of emission factors to retain.

In general, source apportionment of atmospheric species using PMF, is an iterative process that requires many runs of analysis to cope with possible interferences. This study was the first attempt to use PMF on a long-term/high frequency dataset on the remote mountain site CMN.

11.1 Future work

A future research perspective includes a PMF analysis on the whole detrended dataset and also by analyzing NMVOCs and halogenated species separately and compare obtained factors. In addition, identifying source apportionment in different time windows could benefit the interpretation of emission sources e.g. by distinguishing daytime from nighttime of Summer and Winter season, as well as analyzing individual years. Furthermore, to evaluate the lifetime correction method effect on source apportionment, PMF should be analyzed with lifetime correction and without (He et al., 2019). Ultimately, the PMF model uncertainty should include all possible contributions from sampling, data pretreatment, and the PMF algorithm should therefore be adapted in some way.

One prominent limitation of this work, is the fact that the study is conducted at one single measurement site, thus the results will be limited to CMN only. Including several site studies might benefit the interpretation of specific sources that influence CMN, meanwhile other sources might influence different study sites. However, CMN is representing the Southern European atmospheric region.

Reasonably, only one factor model is used for the analysis (the EPA-PMF model). Although a comparison with other receptor models or source models should provide further hints about the uncertainty of obtained results. Additionally, comparing results with more sophisticated source models, involving back trajectories/air mass transport is needed in order to know more about their source location and to better describe the variability on the concentrations for the different classes of halogenated compounds that the PMF/PCA models are not able to capture properly.

Finally, when identifying obtained factors to specific sources, all knowledge considering CMN site needs to be consider, including meteorology. Therefore, polar plots should also be considered in this research.

In future, source apportionment of NMVOCs using PMF with a lifetime correction method has the potential to aid policy makers to develop more effective pollution control concerning anthropogenic NMVOCs.

Part VI

APPENDICES

12

APPENDIX A

```
### GHGs data pretreatment ###

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# GHGs_raw data dimension 18756 x 36

#Deleting species not used in this analysis
library(dplyr)
GHGs<- select(GHGs_raw, -c(2,3,5,9,17,18,19,20,23,24,27,31,32,15, 11, 12))
GHGs_s<- select(GHGs_sraw, -c(2,3,5,9,17,18,19,20,23,24,27,31,32,15, 11, 12))

# Setting the time zone to be UTC +1
library(lubridate)
names(GHGs)[1]<- "date"
GHGs$date <- ymd_hms(GHGs$date, tz = "Etc/GMT-1")
GHGs_unc$date <- ymd_hms(GHGs_unc$date, tz = "Etc/GMT-1")

# Applying a time average function
library(openair)
GHGs<- timeAverage(GHGs, avg.time = "2 hour")
GHGs_unc<- timeAverage(GHGs_unc, avg.time = "2 hour")

library(dplyr)
# Using only years 2015 to 2018
GHGs<- slice(GHGs, 8761:26291)
GHGs_unc<- slice(GHGs_unc, 8761:26291)

# Missing values: converting  NaN to NA in a dataframe
is.nan.data.frame <- function(x)
do.call(cbind, lapply(x, is.nan))

GHGs[is.nan(GHGs)] <- NA
GHGs_unc[is.nan(GHGs_unc)] <- NA

# all zero values must be omitted or replaced by other values
library(dplyr)
sum(is.na(GHGs))
GHGs<- na_if(GHGs,0)
sum(is.na(GHGs))# 33 zero values a replaced by NA
sum(is.na(GHGs_unc))
GHGs_unc<- na_if(GHGs_unc,0)
```

```
sum(is.na(GHGs_unc))


#removing rows with more than 75% NA's
GHGs<- GHGs[rowSums(is.na(GHGs))/ncol(GHGs) <0.75, ] #removing all NA's from rows using a treshold of 75%

# remove large TCE NA's gap from all species
GHGs<- slice(GHGs, 1:11672)

#diurnal timeseries
library(openair)
GHGs<-cutData(GHGs, type = "hour")
GHGs<-cutData(GHGs, type = "season")


ggp<- ggplot(data=GHGs, aes(x = hour, y = CH3Cl, group=season, color= season))
ggp + ggtitle("Diurnal variability 2015-2018")+ geom_line(stat="summary", fun.y="mean") + xlab("hour (UTC+1)") + ylab("CH3Cl (ppt)")

#plot timeseries and uncertainty
library(openair)
timePlot(GHGs_all, pollutant = c("COS", "COS_s"), y.relation = "free", ylab = "concentration (ppt)", main = "Timeseries: COS with uncertainty")

#many timeseries in one graph
library(reshape2)
df <- melt(GHGs_all[, c("date", "SO2F2", "SO2F2_s")], id="date")
ggplot(df) + geom_line(aes(x=date, y=value, color=variable)) + labs(title="Timeseries: SO2F2 and uncertainty")

#Clustering
GHGs_cor<-cor(GHGs[,2:20], method = "pearson", use = "complete.obs")
GHGs_dist<- as.dist((1 - GHGs_cor)/2)
plot(hclust(GHGs_dist, method = "complete"), main = "", xlab= "") #Clustering methods: Agglomerative, CorDistance, Complete linkage")

#filter season
library(dplyr)
GHGs_summer<- filter(GHGs, season == "summer (JJA)")
GHGs_winter<- filter(GHGs, season == "winter (DJF)")

#NA's values
library(imputeTS)
ggplot_na_gapsize(GHGs$TCE)
library(naniar)
gg_miss_var(GHGs[2:21])
gg_miss_var(GHGs[2:36])

#linear interpolation
GHGs_int<- GHGs[,1:20]
```

```
colnames(GHGs_int) <- c("date","SO2F2", "HFC_32", "HFC_125", "HFC_134a", "HFC_
    ↪ 152a", "HFC_365mfc", "HCFC_22", "HCFC_142b", "CFC_114", "CFC_115", "
    ↪ CH3Cl","CH3Br", "CH2Cl2", "CHCl3", "CCl4", "CH3CCl3", "TCE", "PCE", "COS
    ↪ ")

library("imputeTS")
GHGs_int$COS<- na_interpolation(GHGs_int$COS, option = "linear",  maxgap = 6)
plotNA.gapsize(GHGs_int$SO2F2)

#creating timeseries using mstl function
library("forecast")
GHGs_imp<-msts(GHGs_int[,2:20], seasonal.periods=c(12*7*4)) # monthly
    ↪ seasonality

#using na.interp algorithm to interpolate missing values (uses linear
    ↪ interpolation for non-seasonal series and a periodic STL-decomposition
    ↪ with seasonal series to replace missing values)
library("forecast")
COS<- na.interp(GHGs_imp[,19],lambda = NULL,linear = (frequency(GHGs_imp[,19])
    ↪  <= 1 | sum(!is.na(GHGs_imp[,19])) <= 2 * frequency(GHGs_imp[,19])))

GHGs_sim<- data.frame(GHGs_int$date, SO2F2)

#plotting simulated NA's vs observed values for a quick view
plot.ts(GHGs_COS, ylab = expression("concentration␣(ppt)"), cex.main = 0.85,
    ↪ type = 'o', cex = 0.3, pch = 16, main = "Replacing␣COS␣missing␣values␣
    ↪ with␣estimates", col = 'red')
points(GHGs_imp[,17], cex = 0.3, pch = 16)
legend("topright",legend = 'Imputed␣values', lty = 1, col = 'red', cex = 0.5)

#mSTL decomposition
library("forecast")
library(ggplot2)

COS<- mstl(GHGs_imp[,20], s.window = 7)  # mSTL: multiple seasonal
    ↪ decomposition
autoplot(COS) + ggtitle("COS")


#plotting simulated NA's vs observed values with ggplot
COS_df<- data.frame(GHGs_int$date, COS)
COS_df<- data.frame(COS_df, GHGs_imp[,19])
names(COS_df)[1]<- "date"
names(COS_df)[2]<- "sim"
names(COS_df)[3]<- "obs"

ggplot(COS_df, aes(x = date, y = sim))+
  geom_point(aes(color = "sim"), size = 0.5)+
  geom_point(aes(y = obs, color= "obs"), na.rm=TRUE, size = 0.5)+
  labs(colour= NULL, x= "date", y = "concentration␣(ppt)", title = "Replacing␣
      ↪ COS␣missing␣values␣with␣estimates")+
  theme(plot.title = element_text(hjust = 0.5))+
```

```
  theme(legend.position = c(0.99, 0.99),legend.justification = c("right", "top
      ↪ "), legend.box.just = "right", legend.margin = margin(6, 6, 6, 6))+
  scale_color_manual(values = c("black","red"))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

## GHGs Uncertainties ##
sum(is.na(GHGs_s))

# Geometric mean
library(psych)

GHGs_unc_geom<- GHGs_unc
#Repeat for all GHGs species
GHGs_unc_geom$`SO2F2-1s`[is.na(GHGs_unc_geom$`SO2F2-1s`)] <- geometric.mean(
    ↪ GHGs_unc$`SO2F2-1s`, na.rm = TRUE)*4

sum(is.na(GHGs_unc_geom))
#plot timeseries and uncertainty
library(openair)
timePlot(GHG_s_all, pollutant = c("COS_s", "COS_g"), y.relation = "free", ylab
    ↪  = "concentration (ppt)", main = "Timeseries: COS uncertainty")

# S/N ratio
boxplot(x=GHGs_s_ratio,
        main="S/N ratio",
        xlab="Species",
        ylab="",
        col="orange",
        border="brown")

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# detrend timeseries using STL - remove long-term trends


library(forecast)
library(ggplot2)

#Using STL.

# timeseries
GHGs_STL<-msts(GHGs_sim[,2:20], seasonal.periods=c(12*7,12*7*4, 12*7*4*12)) #
    ↪ weekly, monthly and yearly seasonality

COS<- mstl(GHGs_STL[,19], s.window = 7)  # mSTL: multiple seasonal
    ↪ decomposition
autoplot(COS) + ggtitle("COS STL")+ theme(plot.title = element_text(hjust =
    ↪ 0.5))

COS_ts<- COS[,1] # Transform into a time series
plot(COS_ts)
```

```
COS_trend<- COS[,2] # Transform into a time series
plot(COS_trend)

COS_detrend <- COS_ts - COS_trend
plot(COS_detrend)
abline(h=0)

COS_detrend<- COS_detrend + COS_trend[1] # detrended ts normalised for the
    ↪ first data in time series
plot(COS_detrend)
abline(h=0)

COS_detrend <- data.frame(COS_detrend)
GHGs_sim <- cbind(GHGs_sim, COS_detrend)

x<- COS_detrend[COS_detrend < 0]
plot(x, cex = 1)

test<- GHGs_sim$COS - GHGs_sim$COS_detrend
plot(test)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

### NMVOCs data pretreatment ###

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

#Open VOCs excel/csv file and set NA to be 999999.99

library(openair)
library(dplyr)

VOCs<- VOCs_raw
VOCs_unc <- VOCs_raw_unc
library(lubridate)
VOCs$date <- ymd_hms(VOCs$date, tz = "Etc/GMT-1")
VOCs_unc$date <- ymd_hms(VOCs_unc$date, tz = "Etc/GMT-1")

#calculating the time average before slicing the years and deleting missing
    ↪ values, makes us loose less datapoints
VOCs<- timeAverage(VOCs, avg.time = "2␣hour")
VOCs_unc<- timeAverage(VOCs_unc, avg.time = "2␣hour")

#Only considering year 2015 to 2018
VOCs<- slice(VOCs, 8761:26291)
VOCs_unc<- slice(VOCs_unc, 8761:26291)

# Missing values: converting  NaN to NA in a dataframe
is.nan.data.frame <- function(x)
  do.call(cbind, lapply(x, is.nan))

VOCs[is.nan(VOCs)] <- NA
```

```
VOCs_unc[is.nan(VOCs_unc)] <- NA

#removing all NA's from the rows using a treshold of 75%

VOCs<- VOCs[rowSums(is.na(VOCs))/ncol(VOCs) <0.75, ]
VOCs_unc<- VOCs_unc[rowSums(is.na(VOCs_unc))/ncol(VOCs_unc) <0.75, ]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#Plot timeseries

#Diurnal timeseries
library(openair)
VOCs<-cutData(VOCs, type = "hour")
VOCs<-cutData(VOCs, type = "season")

#Filter season
library(dplyr)
VOCs_S<- filter(VOCs, season == "summer (JJA)")
VOCs_W<- filter(VOCs, season == "winter (DJF)")

library(ggplot2)
ggp<- ggplot(data=VOCs, aes(x = hour, y = mp_xylene, group=season, color=
    ↪ season))
ggp + ggtitle("(m,p)-xylene")+ geom_line(stat="summary", fun.y="mean") + xlab
    ↪ ("hour (UTC+1)") + ylab("Mixing ratio (ppt)") +theme(plot.title =
    ↪ element_text(hjust = 0.5))

#plot timeseries and uncertainty
library(openair)
timePlot(VOCs_all, pollutant = c("toluene", "toluene_s"), y.relation = "free",
    ↪  ylab = "mixing ratio (ppt)", main = "")

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#Clustering: Hierarchical agglomerative cluster analysis using pearson
    ↪ correlation

VOCs_cor<-cor(VOCs[,2:16], method = "pearson", use = "complete.obs")
VOCs_dist<- as.dist((1 - VOCs_cor)/2)
plot(hclust(VOCs_dist, method = "complete"), main = "", xlab= "") #Clustering
    ↪ methods: Agglomerative, CorDistance, Complete linkage"



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# Imputing missing values

#plotting NA's values
sum(is.na(VOCs)) # 4532 missing values needs to be imputed
library("imputeTS")
ggplot_na_gapsize(VOCs$propane)
library(naniar)
gg_miss_var(VOCs[2:16])
library(VIM)
```

```
matrixplot(VOCs[2:16], sortby = NULL)


#Linear interpolation
VOCs_int<- VOCs
library("imputeTS")
#Repeat for every species
VOCs_int$o_xylene<- na_interpolation(VOCs_int$o_xylene, option = "linear",
    ↪ maxgap = 6)
ggplot_na_gapsize(VOCs$propane)


#creating timeseries using mstl function
library("forecast")
VOCs_imp<-msts(VOCs_int[,2:16], seasonal.periods=c(12*7*4)) # monthly
    ↪ seasonality

#using na.interp algorithm to interpolate missing values (uses linear
    ↪ interpolation for non-seasonal series and a periodic STL-decomposition
    ↪ with seasonal series to replace missing values)
#repeat for every species
VOCs_oxylene<- na.interp(VOCs_imp[,15],lambda = NULL,linear = (frequency(VOCs_
    ↪ imp[,15]) <= 1 | sum(!is.na(VOCs_imp[,15])) <= 2 * frequency(VOCs_imp
    ↪ [,15])))
VOCs_oxylene[VOCs_oxylene < 0]
VOCs_propane[VOCs_propane < 0] <- NA #propane have two negative values
VOCs_imp$propane<- na_interpolation(VOCs_imp$propane, option = "linear",
    ↪ maxgap = 6)


VOCs_oxylene <- data.frame(VOCs_oxylene)
VOCs_imp <- cbind(VOCs_imp, VOCs_oxylene)

colnames(VOCs_imp) <- c("date","ethyne", "propane", "i_butane", "n_butane", "i
    ↪ _pentane", "n_pentane", "n_hexane", "n_heptane", "i_octane", "n_octane",
    ↪  "benzene", "toluene", "ethylbenzene", "mp_xylene", "o_xylene")
sum(is.na(VOCs_imp[,2:16]))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#plotting simulated NA's timeseries vs. observed

#For a quick view
plot.ts(VOCs_propane, ylab = expression("Mixing␣ratios␣(ppt)"), cex.main =
    ↪ 0.85, type = 'o', cex = 0.3, pch = 16, main = "Replacing␣ethyne␣missing␣
    ↪ values␣with␣estimates", col = 'red')
points(VOCs_imp[,2], cex = 0.3, pch = 16)
legend("topright",legend = 'Imputed␣values', lty = 1, col = 'red', cex = 0.5)
30, 65,

#plotting simulated NA's vs observed values with ggplot
oxylene_df<- data.frame(VOCs_int$date, VOCs_oxylene)
oxylene_df<- data.frame(oxylene_df, VOCs_imp[,15])
```

```
names(oxylene_df)[1]<- "date"
names(oxylene_df)[2]<- "sim"
names(oxylene_df)[3]<- "obs"

ggplot(oxylene_df, aes(x = date, y = sim))+
  geom_point(aes(color = "sim"), size = 0.5)+
  geom_point(aes(y = obs, color= "obs"), na.rm=TRUE, size = 0.5)+
  labs(colour= NULL, x= "date", y = "concentration (ppt)", title = "Replacing
      ↪ o-xylene missing values with estimates")+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.position = c(0.99, 0.99),legend.justification = c("right", "top
      ↪ "), legend.box.just = "right", legend.margin = margin(6, 6, 6, 6))+
  scale_color_manual(values = c("black","red"))


#mSTL decomposition
library("forecast")
oxylene<- mstl(VOCs_imp[,15], s.window = 7)  # mSTL: multiple seasonal
    ↪ decomposition
autoplot(oxylene) + ggtitle("o-xylene STL")+ theme(plot.title = element_text(
    ↪ hjust = 0.5))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

#*** VOCs uncertainties***
# Geometric mean
library(psych)
VOCs_unc_g<- VOCs_unc
#repeat for all species
VOCs_unc_g$ethyne_s[is.na(VOCs_unc_g$ethyne_s)] <- geometric.mean(VOCs_unc$
    ↪ ethyne_s, na.rm = TRUE)*4

sum(is.na(VOCs_unc_g))

#plot timeseries and uncertainty
library(openair)

VOCs_all <- merge.data.frame(VOCs_imp, VOCs_unc_g, by = "date")
timePlot(VOCs_all, pollutant = c("ethyne", "ethyne_s"), y.relation = "free",
    ↪ ylab = "Mixing ratios (ppt)", main = "")

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# Signal-to-noise ratio

# boxplot
VOCs_unc_ratio<- select(VOCs_all, c(2:16))/select(VOCs_all, c(17:31))

qplot(PMF_all$ethyne, PMF_all$ethyne_unc, main = "SO2F2 vs uncertainty ", xlab
    ↪  = "concentration", ylab = "uncertainty (ppt)")

boxplot(x=VOCs_unc_ratio,
        main="Different boxplots for each month",
```

```
        xlab="Month Number",
        ylab="Degree Fahrenheit",
        col="orange",
        border="brown")

boxplot(x=VOCs_SNratio[11:15], xaxt = "n", yaxt = "n", col="orange", border="
    ↪ brown", ylab="concentration (ppt)")
axis(side = 1, labels = FALSE)
axis(side = 2, las = 2, mgp = c(3, 0.75, 0))
text(x = 1:length(VOCs_SNratio[,11:15]),
     labels = names(VOCs_SNratio[,11:15]),
     xpd = NA,
     ## Rotate the labels by 90 degrees.
     srt = 90,
     cex = 1.2,
     y = par("usr")[2] - 22) # try 10, 60



p <- ggplot(VOCs_unc_ratio, aes(x = "ethyne", y = "ethyne"))
p + geom_boxplot(outlier.colour = "red", outlier.shape = 1)
geom_quantile()

ggplot(VOCs_s_ratio, aes(x = "ethyne", y = values))+ geom_boxplot()

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# calculating OH exposure (dtoh)

# OH rate constants
kOH<- c(8.2e-13, 1.09e-12, 2.12e-12, 2.36e-12, 3.6e-12, 3.8e-12, 5.2e-12, 6.76
    ↪ e-12, 3.34e-12, 8.11e-12, 1.22e-12, 5.63e-12, 7.0e-12, 1.90e-11, 1.36e
    ↪ -11)
ethyne 8.20*10^(-13)
propane 1.09*10^(-12)
i_butane 2.12*10^(-12)
n_butane 2.36*10^(-12)
i_pentane 3.6*10^(-12)
n_pentane 3.80*10^(-12)
n_hexane 5.20*10^(-12)
n_heptane 6.76*10^(-12)
i_octane 3.34*10^(-12)
n_octane 8.11*10^(-12)
benzene 1.22*10^(-12)
toluene 5.63*10^(-12)
ethylbenzene 7.0*10^(-12)
mp_xylene 19.0*10^(-12)
o_xylene 13.6*10^(-12)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
#OH exposure :BTEX data provided by ARPA in the period 1 Oct. 2019 - 19 Nov.
    ↪ 2019.
#Used for calculated source ratio in OH exposure
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
X1<- select(BTEX, c(1,11:28)) #xyleni and etilbenzene
X1<- X1[2:1201,1:19] #X = (m,p)-xylene + ethylbenzene

library(lubridate)
names(X1)[1]<- "date"
X1$date <- ymd_hms(X1$date, tz = "Etc/GMT-1")
library(openair)
X1<-cutData(X1, type = "hour") # X = (m,p)-xylene + ethylbenzene

E1<- select(X1, c(1,11:20))
colnames(E1)<- c("date", "RN", "MO", "PC", "FE", "PR", "BO", "RE", "RA", "FO",
    ↪ "hour")
E1<- select(E1, -c(3))
library(openair)
E1<- cutData(E1, type = "season")
E2<-rowMeans(E1[,2:9], na.rm = TRUE)
Edf<- data.frame(E1$date, E2)
Edf<- data.frame(Edf, E1$hour)
Edf<- data.frame(Edf, E1$season)

names(Edf)[1]<-"date"
names(Edf)[2]<-"ethylbenzene"
names(Edf)[3]<-"hour"
names(Edf)[4]<-"season"

library(ggplot2)
ggp<- ggplot(data= Edf, aes(x = hour, y = ethylbenzene, group= season, color=
    ↪ season))
ggp +ggtitle("Diurnal␣variability␣of␣ethylbenzene␣(source)")+ theme(plot.title
    ↪ = element_text(hjust = 0.5)) + geom_line(stat="summary", fun.y="mean")
    ↪ + xlab("hour␣(UTC+1)") + ylab("Average␣concentration␣(ppt)")

X1<- select(X1, c(1:10,20))
colnames(X1)<- c("date", "RN", "MO", "PC", "FE", "PR", "BO", "RE", "RA", "FO",
    ↪ "hour")
X1<- select(X1, -c(3))
X1<- cutData(X1, type = "season")
X2<-rowMeans(X1[,2:9], na.rm = TRUE)
Xdf<- data.frame(X1$date, X2)
Xdf<- data.frame(Xdf, X1$hour)
Xdf<- data.frame(Xdf, X1$season)

names(Xdf)[1]<-"date"
names(Xdf)[2]<-"xylene"
names(Xdf)[3]<-"hour"
names(Xdf)[4]<-"season"
```

```
ggp<- ggplot(data= Edf, aes(x = hour, y = ethylbenzene, group= season, color=
    ↪ season))
ggp + geom_line(stat="summary", fun.y="mean") + xlab("hour (UTC+1)") + ylab("
    ↪ Average concentration (ppt)") + theme(legend.position = "none")
ggplot(test_data, aes(date)) +



# Consider only night time  18:00-06:00 UTC +1
X<- filter(X, hour== "00"| hour== "01"| hour== "02"| hour== "03"| hour== "04"|
    ↪  hour== "05"| hour== "06"| hour=="18"| hour=="19"| hour == "20"| hour ==
    ↪ "21"| hour== "22"| hour== "23") # X = (m,p)-xylene
E<- select(X, c(1,11:19)) # ethylbenzene
sum(is.na(E))
X<- select(X, c(1:10)) # (m,p)-xylene
sum(is.na(X))

E_X<- E[,2:10]/X[,2:10]
sum(is.na(E_X))
E_X<- E[,2:9]/X[,2:9]
sum(is.na(E_X))

#E_X_mean = 0.2263  + (0.6381) + 0.3007 + 0.3075 + 0.2569 + 0.2659 + 0.2822 +
    ↪ 0.2868 + 0.2734 = 2.1997/8 = 0.275
#X_E_mean = 4.451 + 3.433 + 3.322 + 3.942 + 3.815 + 3.611 + 3.735 + 3.768 =
    ↪ 30.077/8 = 3.76
#E_mpxylene 0.3620 + 0.4803) + 0.5028) + 0.3947) + 0.4238) + 0.4604) + 0.4475)
    ↪  + 0.4760/8 = 0.4434


E_X<- data.frame(E$date, E_X)
names(E_X)[1]<- "date"
colnames(E_X)<- c("date", "RN", "MO", "PC", "FE", "PR", "BO", "RE", "RA", "FO"
    ↪ )

library(plotly)
x <- list(
  title = "Monitoring station")
y <- list(
  title = "Average E/X ratio ")
fig<- plot_ly(E_Xdf, x=~x, y=~y, linetype = I("solid"),
              marker = list(color = "rgb(195, 195, 195)"))
fig <- fig %>% layout(xaxis = x, yaxis = y)
fig

library(openair)
scatterPlot(E_X, x = "date", y = "RA", smooth = TRUE, statistic = "mean", ylab
    ↪  = "E/X", titel = "FO" )

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# Calculating OH exposure of VOCs

k_E<- 7e-12
k_X<- 1.9e-11
```

```
dtoh<-(1 / (k_E - k_X)) * (log(0.28) - log(VOCs_sim$ethylbenzene/VOCs_sim$mp_
    ↪ xylene))
dtoh_df<- data.frame(VOCs_sim$date, dtoh)

dtoh<-(1 / (k_E - k_X)) * (log(0.28) - log(VOCs_detr$ethylbenzene_detrend/VOCs
    ↪ _detr$mpxylene_detrend))
dtoh_df<- data.frame(VOCs_detr$date, dtoh)


plot(dtoh_df, type = "o", cex = .1, main = "OH exposure", xlab = "date", ylab
    ↪ = "dtOH")
abline(h=0, col = "red")

#(negative OH exposure events)
neg_dtoh<- dtoh[dtoh < 0]
plot(neg_dtoh, type = "o", cex = .1, main = "Negative OH exposure events",
    ↪ ylab = "dtOH") #3600 values are negative

#deleting events that last more than 6 h
library(dplyr)
names(dtoh_df)[1]<-"date"
dtoh_1<- dtoh_df %>%
  mutate(date_end = lead(date, default = last(date))) %>%
  group_by(group = data.table::rleid(dtoh < 0)) %>%
  filter(is.na(dtoh) | dtoh > 0 | difftime(max(date_end), min(date), units = "
      ↪ hours") < 6) %>%
  ungroup()
dtoh_1<- select(dtoh_1, c(1,2))

plot(dtoh_1, type = "l", cex = .2, main = "OH exposure without negative events
    ↪ that last > 6h", xlab = "date", ylab = "dtOH")
abline(h=0, col = "red")

#removing outliers
dtoh_2<- dtoh_1[-c(4795, 4796, 4797, 7868),]
plot(dtoh_2, type = "l", cex = .2, main = "OH exposure without negative events
    ↪ that last > 6h", xlab = "date", ylab = "dtOH")
abline(h=0, col = "red")

#(deleting all negative events)
out<- which(dtoh_oX$dtoh_oX<0)
dtoh_oX2<-dtoh_oX2[-c(out),]
plot(dtoh_oX2, type = "l", cex = .2, main = "OH exposure without negative
    ↪ events", xlab = "date", ylab = "dtOH")
abline(h=0, col = "red")

dtoh<- dtoh_oX1[dtoh_oX1< 0]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
# Calculating VOCs initial

```

```
VOCs_0<- merge.data.frame(VOCs_sim, dtoh_2, by = "date")
kOH<- c(8.2e-13, 1.09e-12, 2.12e-12, 2.36e-12, 3.6e-12, 3.8e-12, 5.2e-12, 6.76
    ↪ e-12, 3.34e-12, 8.11e-12, 1.22e-12, 5.63e-12, 7.0e-12, 1.90e-11, 1.36e
    ↪ -11)
VOCs_initial <- as.matrix(VOCs_0[,2:16])*exp(outer(VOCs_0$dtoh,kOH,"*"))
VOCs_ratio<- VOCs_initial/VOCs_0[,2:16]
names(VOCs_ratio)[13]<- "e_benzene"

VOCs_initial<- data.frame(VOCs_0$date, VOCs_initial)
names(VOCs_initial)[1]<- "date"

#plotting a barplot of the ratio

V1<- colnames(VOCs_ratio)
V2<- colMeans(VOCs_ratio)
ratio_box<- data.frame(V1, V2)
barplot(ratio_box[[2]], main = "Ratios␣of␣VOCs␣initial␣/␣observed", col = "
    ↪ gray60", names.arg = ratio_box$V1,las = 2, ylim=c(0,5))
abline(h=1, v = NULL, col = "black", lwd=1, lty = 2)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#Detrending timeseries

library(forecast)
library(ggplot2)

#Using STL.

# timeseries
VOCs_STL<-msts(VOCs_initial[,2:16], seasonal.periods=c(12*7,12*7*4, 12*7*4*12)
    ↪ ) # weekly, monthly and yearly seasonality

oxylene<- mstl(VOCs_STL[,15], s.window = 7)  # mSTL: multiple seasonal
    ↪ decomposition
autoplot(oxylene) + ggtitle("o-xylene␣STL")+ theme(plot.title = element_text(
    ↪ hjust = 0.5))

oxylene_ts<- oxylene[,1] # Transform into a time series
plot(oxylene_ts)
oxylene_trend<- oxylene[,2] # Transform into a time series
plot(oxylene_trend)

oxylene_detrend <- oxylene_ts - oxylene_trend
plot(oxylene_detrend)
abline(h=0)

oxylene_detrend<- oxylene_detrend + oxylene_trend[1] # detrended ts normalised
    ↪  for the first data in time series
plot(oxylene_detrend)
abline(h=0)

oxylene_detrend <- data.frame(oxylene_detrend)
VOCs_initial <- cbind(VOCs_initial, oxylene_detrend)
```

```
test<- VOCs_sim$o_xylene - VOCs_sim$oxylene_detrend
plot(test)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#Principal Component Analysis

#PCA The script below is used in the analysis
library(backports)
library(devtools)
library(ggplot2)
library(dplyr)
library(ggplot2)
library(factoextra)

detr_S<- select(PMFdetr_S, -c(2,10,11,13,17,18,19))
p<- select(PMFS_night, c(2:35))

colnames(p) <- c("SO2F2", "HFC_32", "HFC_125", "HFC_134a", "HFC_152a", "HFC_
    ↪ 365mfc", "HCFC_22", "HCFC_142b", "CFC_114", "CFC_115", "CH3Cl","CH3Br",
    ↪ "CH2Cl2", "CHCl3", "CCl4", "CH3CCl3", "TCE", "PCE", "COS", "ethyne", "
    ↪ propane", "ibutane", "nbutane", "ipentane", "npentane", "nhexane", "
    ↪ nheptane", "ioctane", "noctane", "benzene", "toluene", "ethylbenzene", "
    ↪ mp_xylene", "o_xylene")


res_p<-prcomp(p, center = TRUE, scale. = TRUE)
summary(res_p)

#Scree plot - variance explained

library(factoextra)
fviz_screeplot(res_p, addlabels = TRUE, ylim = c(0, 75), xlab = "PCs")
get_eig(res_p)
x<-get_eig(res_p)
plot(x$eigenvalue, type = "o", cex = .5, main = "Scree␣plot", xlab = "PCs",
    ↪ ylab = "Eigenvalue")
abline(h=1, col = "red")

plot(x$cumulative.variance.percent, type = "o", cex = .5, main = "Scree␣plot",
    ↪  xlab = "PCs", ylab = "Cumulative␣Variance␣(%)")
abline(h=90, col = "red")

plot(x$variance.percent, type = "o", cex = .5, main = "Scree␣plot", xlab = "
    ↪ PCs", ylab = "Variance␣(%)")
abline(h=5, col = "red")

# Loadings, Scores
library(ggplot2)
loadings<- res_p$rotation
print(loadings, cutoff = 0.0)
```

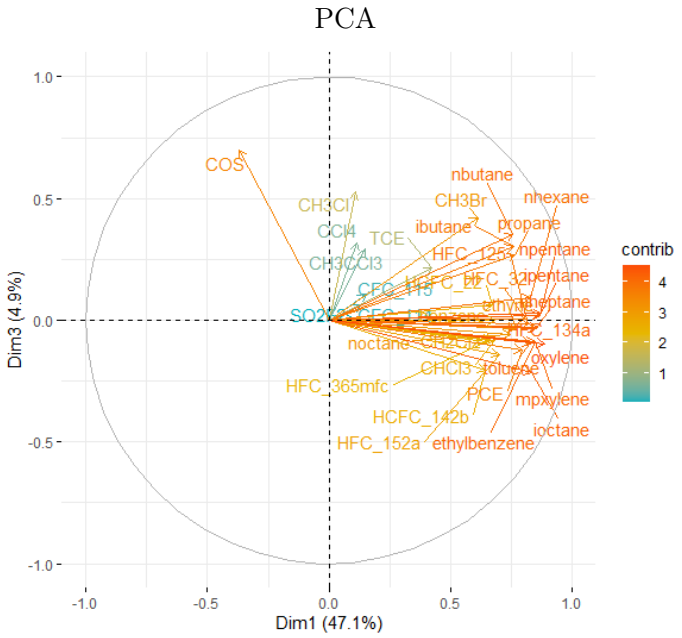
```
scores<- res_p$x

#plot scores
qplot(scores[,1], scores[,2], data = PMF_res, colour = year, xlab = "PC1",
    ↪ ylab = "PC2", main = "PCA: Scores")

#plot loadings
loadings<- as.data.frame(loadings)
rownames(loadings) <- c("SO2F2", "HFC_32", "HFC_125", "HFC_134a", "HFC_152a",
    ↪ "HFC_365mfc", "HCFC_22", "HCFC_142b", "CFC_114", "CFC_115", "CH3Cl","
    ↪ CH3Br", "CH2Cl2", "CHCl3", "CCl4", "CH3CCl3", "TCE", "PCE", "COS", "
    ↪ ethyne", "propane", "ibutane", "nbutane", "ipentane", "npentane", "
    ↪ nhexane", "nheptane", "ioctane", "noctane", "benzene", "toluene", "
    ↪ ethylbenzene", "mpxylene", "oxylene")
plot <- ggplot(loadings, aes(PC1, PC5)) + geom_point(color = "red")
plot + geom_text_repel(label= rownames(loadings)) + labs(title = "PCA:
    ↪ Loadings")+
  geom_vline(xintercept=c(0), linetype="dotted")+
  geom_hline(yintercept = c(0), linetype="dotted")


fviz_pca_var(res_p, col.var = "black", repel = TRUE)

fviz_pca_var(res_p, col.var="contrib", axes = c(2,4),
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)

#loadings contributions
fviz_contrib(res_p, choice = "var", axes = 1, top = 34)

#biplot
fviz_pca_biplot(res_p, repel = TRUE)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#PMF data preparation

library(openair)
library(dplyr)
#summer
PMF_S<-cutData(PMF_detr, type = "season")
PMF_S<- filter(PMF_S, season == "summer (JJA)")

#daytime and nighttime
PMFS_h<-cutData(PMF_S, type = "hour")
PMFS_day <-filter(PMFS_h, hour == c("10","11","12","13","14","15","16","17","
    ↪ 18"))
PMFS_night <- filter(PMFS_h, hour== c("00","01","02","03","04"))

#winter
PMF_W<-cutData(PMF_detr, type = "season")
PMF_W<- filter(PMF_W, season == "winter (DJF)")
#
```

```
PMFW_h<-cutData(PMF_W, type = "hour")
PMFW_day <-filter(PMFW_h, hour == c("10","11","12","13","14","15","16","17","
    ↪ 18"))
PMFW_night <- filter(PMFW_h, hour== c("00","01","02","03","04"))

# Plotting residuals
library(ggplot2)
library(plotly)

plot_ly(X, x=~date, y=~toluene)

#Qtrue/Qexp
library(plotly)
fig<- plot_ly(Q, x=~factors, y=~Q, linetype = I("solid"))

x <- list(
  title = "Factors")
y <- list(
  title = "Qrob/Qexp")

fig <- fig %>% layout(xaxis = x, yaxis = y)
fig
```

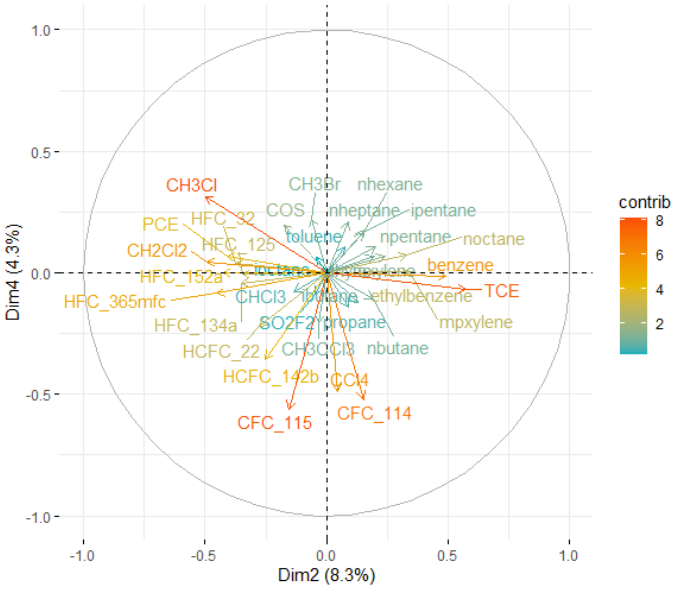
Listing 12.1: caption

13

APPENDIX B



(a)



(b)

Figure 48: Loadings of Summer season: (a) PC1 vs. PC3 and (b) PC2 vs. PC4.

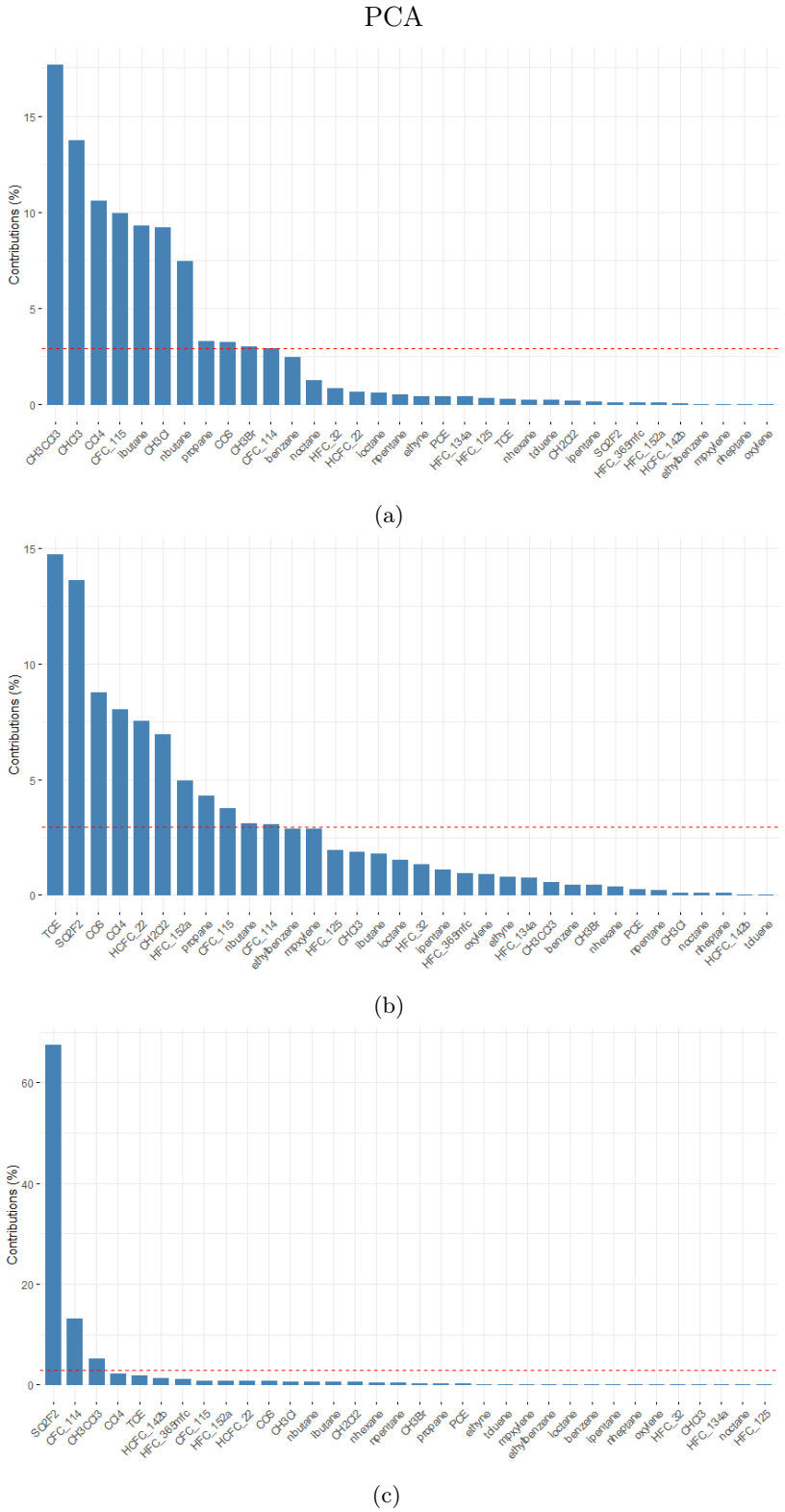
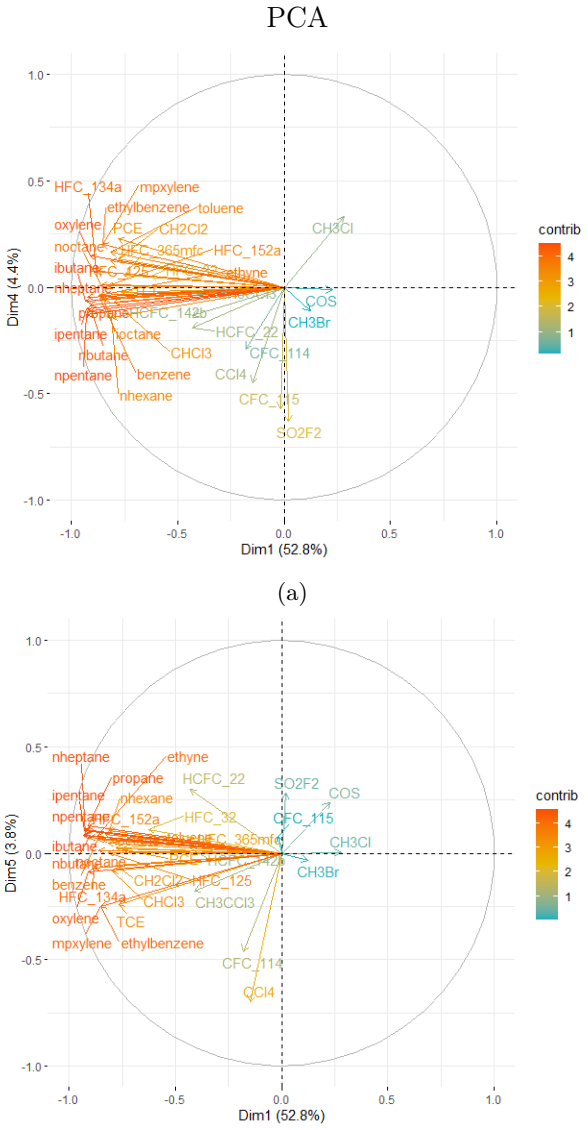


Figure 49: Loadings of Summer season: (a) PC5, (b) PC6 and (c) PC7.



(a)

(b)

Figure 50: Winter season: (a) PC1 vs. PC4 and (b) PC1 vs. PC5.

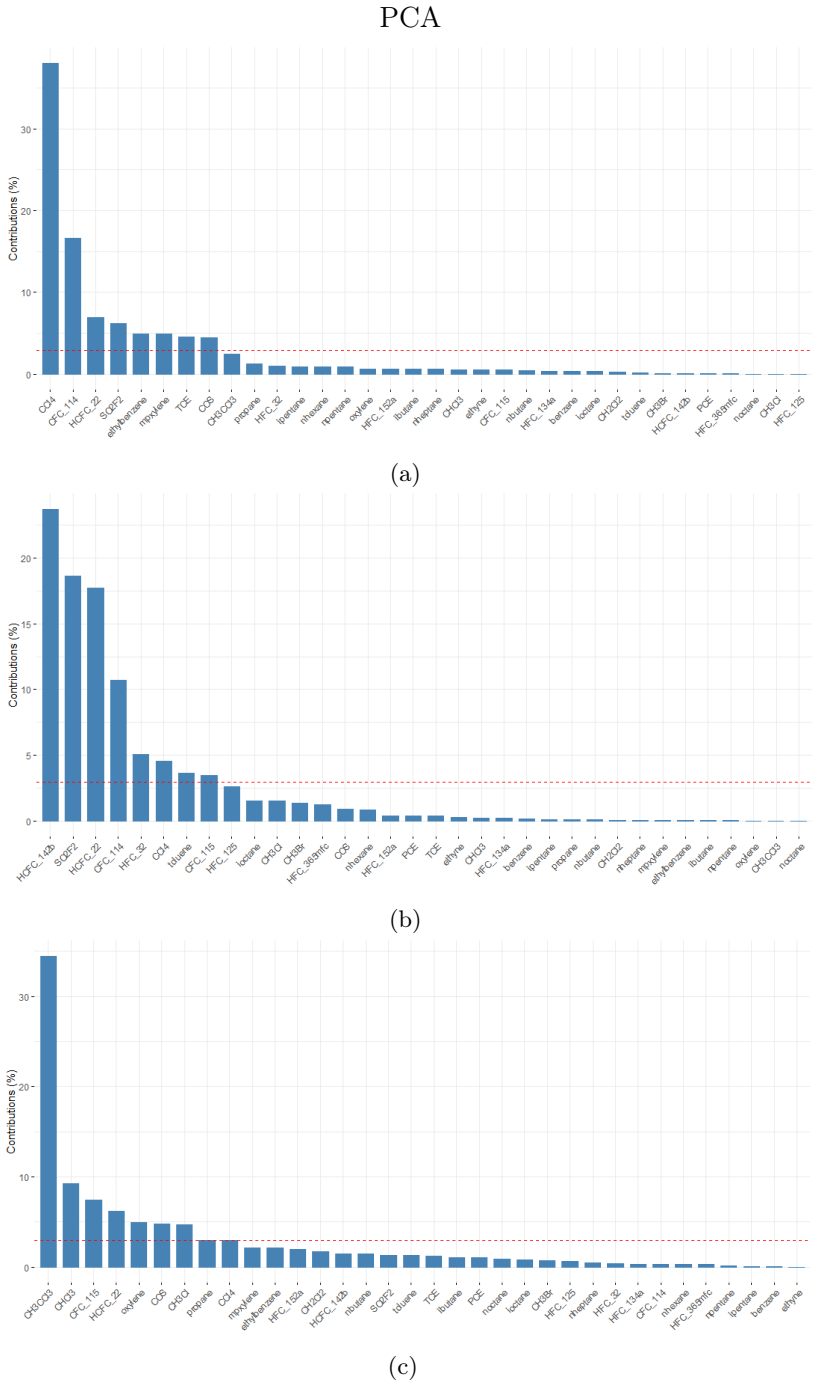


Figure 51: Loadings of Winter season: (a) PC5, (b) PC6, and (c) PC7.

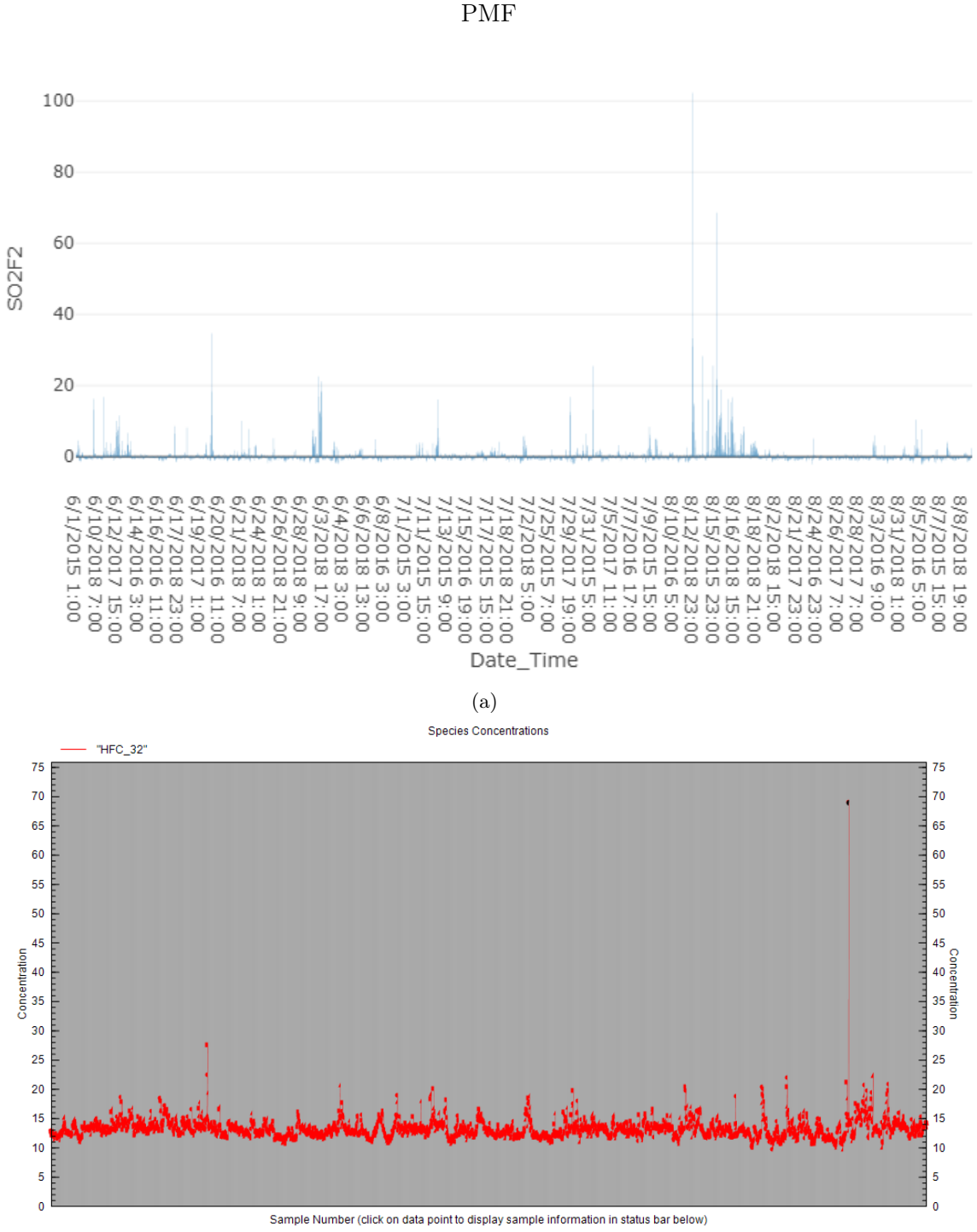


Figure 52: Evaluating and removing outliers from PMF analysis: (a) Residuals of SO_2F_2 and three peaks are removed; (b) time series HFC-32 during Winter with high Q/Qexp due to one outlier (highlighted in black) and is removed prior to PMF analysis.

BIBLIOGRAPHY

Atkinson, R et al. (2006). "Evaluated kinetic and photochemical data for atmospheric chemistry: Volume II–gas phase reactions of organic species". In: Atmospheric chemistry and physics 6.11, pp. 3625–4055.

Atkinson, Roger and Janet Arey (2003). "Atmospheric degradation of volatile organic compounds". In: Chemical reviews 103.12, pp. 4605–4638.

Belis, CA et al. (2019). "European Guide on Air Pollution Source Apportionment With Receptor Models–Revised Version 2019. JRC117306". In: EUR. Vol. 29816. EN, Publications Office of the European Union Luxembourg.

Belis, Claudio A et al. (2014). European guide on air pollution source apportionment with receptor models.

Braesicke, Peter et al. (2019). SCIENTIFIC ASSESSMENT OF OZONE DEPLETION: 2018 World Meteorological Organization Global Ozone Research and Monitoring Project-Report No. 58 World Meteorological Organization United Nations Environment Programme National Oceanic and Atmospheric Administration National Aeronautics and Space Administration European Commission.

Comero, S, L Capitani, and BM Gawlik (2009). "Positive Matrix Factorisation (PMF)–An introduction to the chemometric evaluation of environmental monitoring data using PMF". In: Office for Official Publications of the European Communities, Luxembourg, p. 59.

Contini, Daniele et al. (2016). "Application of PMF and CMB receptor models for the evaluation of the contribution of a large coal-fired power plant to PM10 concentrations". In: Science of the Total Environment 560, pp. 131–140.

Cristofanelli, Paolo et al. (2017). High-mountain Atmospheric Research: The Italian Mt. Cimone WMO/GAW Global Station (2165 M Asl). Springer.

Cristofanelli, Paolo et al. (2020). "First Evidences of Methyl Chloride (CH3Cl) Transport from the Northern Italy Boundary Layer during Summer 2017". In: Atmosphere 11.3, p. 238.

De Gouw, JA et al. (2005). "Budget of organic carbon in a polluted atmosphere: Results from the New England Air Quality Study in 2002". In: Journal of Geophysical Research: Atmospheres 110.D16.

Debevec, Cécile et al. (2020). "Seasonal variation and origins of volatile organic compounds observed during two years at a western Mediterranean remote background site (Ersa, Cape Corsica)". In: Atmospheric Chemistry and Physics Discussions, pp. 1–63.

He, ZR et al. (2019). "Contributions of different anthropogenic volatile organic compound sources to ozone formation at a receptor site in the Pearl River Delta region and its policy implications". In: Atmospheric chemistry and physics.

Henne, Stephan et al. (2010). "Assessment of parameters describing representativeness of air quality in-situ measurement sites". In:

Hopke, Philip K (2000). "A guide to positive matrix factorization". In: Workshop on UNMIX and PMF as Applied to PM2. Vol. 5, p. 600.

Hopke, Philip K, Daniel A Jaffe, et al. (2020). "Letter to the Editor: Ending the Use of Obsolete Data Analysis Methods". In: Aerosol and Air Quality Research 20.4, pp. 688–689.

Hopkins, Francesca M et al. (2016). "Mitigation of methane emissions in cities: How new measurements and partnerships can contribute to emissions reduction strategies". In: Earth's Future 4.9, pp. 408–425.

IPCC Stocker, Thomas F et al. (2014). Climate Change 2013: The physical science basis. contribution of working group I to the fifth assessment report of IPCC the intergovernmental panel on climate change. Cambridge University Press.

Jain, C. D. et al. (2017). "Volatile organic compounds (VOCs) in the air, their importance and measurements". In:

Kansal, Ankur (2009). "Sources and reactivity of NMHCs and VOCs in the atmosphere: A review". In: Journal of hazardous materials 166.1, pp. 17–26.

Kaufman, Leonard and Peter J Rousseeuw (2009). Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons.

Knop, Vincent et al. (2014). "A linear-by-mole blending rule for octane numbers of n-heptane/iso-octane/toluene mixtures". In: Fuel 115, pp. 666–673.

Lo Vullo, Eleonora Lo et al. (2015). "Non-methane volatile organic compounds in the background atmospheres of a Southern European mountain site (Mt. Cimone, Italy): Annual and seasonal variability". In: Aerosol and Air Quality Research 16.3, pp. 581–592.

Lo Vullo, Eleonora Lo et al. (2016). "Anthropogenic non-methane volatile hydrocarbons at Mt. Cimone (2165 m asl, Italy): Impact of sources and transport on atmospheric composition". In: Atmospheric Environment 140, pp. 395–403.

Maione, Michela et al. (2013). "Ten years of continuous observations of stratospheric ozone depleting gases at Monte Cimone (Italy)—Comments on the effectiveness of the Montreal Protocol from a regional perspective". In: Science of the total environment 445, pp. 155–164.

WMO (2020). WMO/GAW program. url: https://public.wmo.int/en/programmes/global-atmosphere-watch-programme.

Masson-Delmotte, TWV et al. (2018). "IPCC, 2018: Summary for Policymakers. In: Global warming of 1.5 C. An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global". In: World Meteorological Organization, Geneva, Tech. Rep.

Masson-Delmotte, Valérie et al. (2013). "Information from paleoclimate archives". In:

Meszaros, T, L Haszpra, and A Gelencser (2004). "The assessment of the seasonal contribution of the anthropogenic sources to the carbon monoxide budget in Europe". In: Atmospheric Environment 38.25, pp. 4147–4154.

Mohr, Thomas KG (2020). "1 Chlorinated Solvents and Solvent Stabilizers". In: Environmental Investigation and Remediation: 1, 4-Dioxane and other Solvent Stabilizers, p. 1.

Monks, Paul S (2005). "Gas-phase radical chemistry in the troposphere". In: Chemical Society Reviews 34.5, pp. 376–395.

Monod, Anne et al. (2001). "Monoaromatic compounds in ambient air of various cities: a focus on correlations between the xylenes and ethylbenzene". In: Atmospheric Environment 35.1, pp. 135–149.

Montzka, Stephen A et al. (2018). "An unexpected and persistent increase in global emissions of ozone-depleting CFC-11". In: Nature 557.7705, pp. 413–417.

Mühle, J et al. (2009). "Sulfuryl fluoride in the global atmosphere". In: Journal of Geophysical Research: Atmospheres 114.D5.

Norris, G et al. (2014). "Epa positive matrix factorization (PMF) 5.0 fundamentals and user guide prepared for the US environmental protection agency office of research and development, Washington, DC". In: Inc., Petaluma.

Olivieri, Alejandro C et al. (2015). Fundamentals and analytical applications of multiway calibration. Elsevier.

Paatero, Pentti (1999). "The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model". In: Journal of Computational and Graphical Statistics 8.4, pp. 854–888.

Paatero, Pentti and Unto Tapper (1994). "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values". In: Environmetrics 5.2, pp. 111–126.

Polissar, Alexandr V et al. (1998). "Atmospheric aerosol over Alaska: 2. Elemental composition and sources". In: Journal of Geophysical Research: Atmospheres 103.D15, pp. 19045–19057.

Reff, Adam, Shelly I Eberly, and Prakash V Bhave (2007). "Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods". In: Journal of the Air & Waste Management Association 57.2, pp. 146–154.

Roberts, James M et al. (1985). "Measurements of anthropogenic hydrocarbon concentration ratios in the rural troposphere: Discrimination between background and urban sources". In: Atmospheric Environment (1967) 19.11, pp. 1945–1950.

Saiz-Lopez, Alfonso et al. (2017). "Unexpected increase in the oxidation capacity of the urban atmosphere of Madrid, Spain". In: Scientific reports 7, p. 45956.

Sauvage, S et al. (2009). "Long term measurement and source apportionment of non-methane hydrocarbons in three French rural areas". In: Atmospheric Environment 43.15, pp. 2430–2441.

Seinfeld, John H and Spyros N Pandis (2016). Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons.

Tans, Pieter and E Dlugokencky (2020). "NOAA/ESRL (www. esrl. noaa. gov/gmd/ccgg/trends/)". In: Retrieved July 7.

Vigni, Mario Li, Caterina Durante, and Marina Cocchi (2013). "Exploratory data analysis". In: Data Handling in Science and Technology. Vol. 28. Elsevier, pp. 55–126.

Yuan, Bin et al. (2012). "Volatile organic compounds (VOCs) in urban air: How chemistry affects the interpretation of positive matrix factorization (PMF) analysis". In: Journal of Geophysical Research: Atmospheres 117.D24.

Zalasiewicz*, Jan et al. (2010). The new world of the Anthropocene.