공학박사 학위논문

Source apportionment and spatiotemporal analysis of $PM_{2.5}$ using machine learning and receptor models

기계학습과 수용모델을 이용한 초미세먼지 오염원 및 기여도의 시공간 분포 분석

2023년 2월

서울대학교 대학원

건설환경공학부

이 영 수

Source apportionment and spatiotemporal analysis of $PM_{2.5}$ using machine learning and receptor models

지도 교수 김 재 영

이 논문을 공학박사 학위논문으로 제출함
2022년 10월

서울대학교 대학원
건설환경공학부
이 영 수

이영수의 공학박사 학위논문을 인준함
2022년 12월

위 원 장 남 경 필 (인)

부위원장 김 재 영 (인)

위 원 이 승 묵 (인)

위 원 Eun Sug Park (인)

위 원 최 용 주 (인)

Source apportionment and spatiotemporal analysis of $PM_{2.5}$ using machine learning and receptor models

by

Young Su Lee

Advisor: Jae Young Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Department of Civil and Environmental Engineering
The Graduate School
Seoul National University

February 2023

Abstract

Source apportionment and spatiotemporal analysis of $PM_{2.5}$ using machine learning and receptor models

Young Su Lee

Department of Civil and Environment Engineering

The Graduate School

Seoul National University

Particulate matter less than 2.5 micrometers ($PM_{2.5}$) has been a pollutant of interest globally for more than decades, owing to its adverse health effects. For developing effective $PM_{2.5}$ management strategies, it is crucial to identify their sources and quantify how much they contribute to ambient $PM_{2.5}$ concentrations in time and space. Source apportionment is the key to identifying the characteristics of $PM_{2.5}$. Receptor modeling is widely used to identify $PM_{2.5}$ sources as a statistical method of source apportionment. The chemical constituents of $PM_{2.5}$ were used as input data for receptor modeling.

Therefore, this study aimed to investigate the characteristics of $PM_{2.5}$ using

models of source apportionment and spatiotemporal analysis for effective management strategies. Two types of modeling were performed for the source apportionment study. The first is positive matrix factorization modeling, which identifies a specific source type and its contributions to $PM_{2.5}$ from one site. The second is Bayesian spatial multivariate receptor modeling, which derives major sources and their contributions to $PM_{2.5}$ from multiple monitoring sites. In addition, machine learning models were used to predict the concentrations of $PM_{2.5}$, which are important data for receptor modeling. Machine learning models that can be used to increase data integrity and applicability to $PM_{2.5}$ data were assessed.

The sources of $PM_{2.5}$ and their contributions in Siheung, South Korea, were identified using positive matrix factorization modeling. These 10 sources were secondary nitrate (24.3%), secondary sulfate (18.8%), traffic (18.8%), combustion for heating (12.6%), biomass burning (11.8%), coal combustion (3.6%), heavy oil industry (1.8%), smelting industry (4.0%), sea salt (2.7%), and soil (1.7%). Based on the derived sources, the carcinogenic and non-carcinogenic health risks due to $PM_{2.5}$ inhalation were estimated. The contribution to $PM_{2.5}$ mass concentration was low for coal combustion, heavy oil industry, and traffic sources but exceeded the benchmark carcinogenic health risk value (1E-06). Therefore, countermeasures on $PM_{2.5}$ emission sources should be performed based on the $PM_{2.5}$ mass concentration and health risks.

The feature extraction capabilities of the four machine learning models to predict the chemical constituents of $PM_{2.5}$ were assessed by comparing the prediction accuracy depending on input variables, target constituents for prediction, available period, missing ratios of input data, and study sites. The concentrations of $PM_{2.5}$ constituents were predicted at three sites (Seoul, Ulsan, and Baengnyeong) in South

Korea between 2016 and 2018, using four machine learning models: generative adversarial imputation network (GAIN), fully connected deep neural network (FCDNN), random forest (RF), and k-nearest neighbor (kNN). The prediction accuracy identified by the coefficient of determination (R^2) between the prediction and observation was highest in GAIN, followed by FCDNN, RF, and kNN. As the missing ratios (20, 40, 60, and 80%) of the input data increased, the prediction accuracy decreased in the four models and was more noticeable in GAIN and kNN, which are unsupervised models. As the input data period increased, the two deep learning models, GAIN and DNN, had better applicability than the other models, RF and kNN. The study sites with more emission sources exhibited lower prediction accuracy, resulting in the highest R^2 in the BR island and the lowest in Ulsan. Among the target constituent groups, ions and trace elements were predicted to have the highest and lowest R^2, respectively. This study demonstrated that machine learning models can be extended for further air pollution studies depending on model features, required performance, and experimental conditions, such as data availability and time constraints.

The spatial distributions of five $PM_{2.5}$ sources in South Korea were estimated using Bayesian spatial multivariate receptor modeling. Secondary nitrate, secondary sulfate, motor vehicle emissions, industry, and sea salts were determined to be significant contributors to ambient $PM_{2.5}$ concentrations in South Korea. The spatial surface of the daily average contribution for each source in South Korea was derived from measurement data from the eight monitoring sites. The source contributions predicted by the BSMRM were also validated using held-out data from a test site (such as Ansan, Daejeon, and Gwangju). These predicted source contributions can aid in developing effective $PM_{2.5}$ control strategies in cities where no speciated $PM_{2.5}$

monitoring stations are available. They can also be utilized as source-specific exposures in health effect studies, even in cities where no monitoring stations are available.

Keywords: $PM_{2.5}$; Source apportionment; Positive matrix factorization; Machine learning modeling; $PM_{2.5}$ chemical constituents; Bayesian receptor modeling

Student Number: 2019-32839

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Background

The atmospheric environment, along with water quality, waste, and soil, is a major management target for the sustainable prosperity of humans (Arora, 2018; Sauvé et al., 2016). Research on managing the atmospheric environment began in the 1950s and has been actively conducted to expand the research area worldwide (Colvile et al., 2001; Jacobson, 2002; Ramanathan and Feng, 2009). To maintain a sustainable atmospheric environment, it is necessary to identify the situation and efficiently manage air pollution (Melamed et al., 2016). According to the World Health Organization (WHO), “Air pollution is the contamination of air due to the presence of substances in the atmosphere that are harmful to the health of humans and other living beings, or cause damage to the climate or materials” (World Health Organization, 2021).

Air pollutants can be classified into natural and anthropogenic emissions (Jacobson, 1930; Sharma et al., 2018). (Jacobson, 1930; Sharma et al., 2018). Naturally occurring air pollutants, such as yellow dust, emissions from forest fires, and volcanic eruptions, are generated regardless of human activities (Jacobson, 1930). (Jacobson, 1930). Anthropogenic emissions are generated by human activities, such as power plants and automobile exhaust gases (Popescu and Ionel, 2010). A major concern in atmospheric environment management is anthropogenic emissions, which have had an impact on human safety (Jacobson, 2012). The London smog incident is an example in which more than 10,000 people died (Hopke et al., 2020; Jacobson, 2002). Since then, efforts to control anthropogenic air pollutant emissions

have begun, such as investigating the sources of air pollution and enacting air pollution control laws (Hopke, 2016; Jacobson, 2002).

Although there is a reduction in the overall air pollution problem compared with the past, the WHO estimates that 4.2 million people die prematurely every year due to outdoor air pollution (World Health Organization, 2021). Since air pollution may be recognized as a political problem, it may become a cause of conflict between countries. This problem arises because it is difficult to interpret the air pollution phenomenon (Seinfeld and Pandis, 2016). Once generated, air pollutants undergo various reactions and transport processes depending on weather conditions, and their complexity is high (Arya, 1998). For example, reactions to light, long-distance transport, dilution by wind, deposition, and precipitation are affected by many variables in the process (Arya, 1998). This makes the scientific interpretation of the air pollution problem difficult. Therefore, more air pollution studies are required (WHO, 2005). The scientific interpretation of air pollution is an important issue that humans must continue to challenge.

Particulate matter less than 2.5 μm in diameter ($PM_{2.5}$), one of the major air pollutants, is an aerosol composed of various chemical constituents from various emission sources. $PM_{2.5}$ is classified as carcinogenic group 1 by the International Agency for Research on Cancer (IARC) (WHO, 2005; Widziewicz et al., 2016). This group is the same as that for arsenic and benzene. $PM_{2.5}$, known to cause cardiovascular and respiratory diseases, is a crucial air pollutant managed by most countries globally (Choi et al., 2011). However, most countries do not meet the WHO recommendations.

Scientific approaches have been proposed by the United States Environmental Protection Agency (US EPA) to effectively identify and control $PM_{2.5}$ (US EPA, 1997). These can be categorized into four main groups. The first was to measure and analyze the detailed physicochemical characteristics of $PM_{2.5}$. The second is estimating emissions from sources, such as power plants and vehicles. The third is to understand the spatial distribution of $PM_{2.5}$ through spatial modeling. Finally, we aimed to understand the health effects on the human body. Accordingly, various studies have been conducted in each field.

In this thesis, the scientific approaches presented by the US EPA were considered. By researching specific topics, we intended to derive the most scientific results from air pollution research. The following were attempted in this study (1) to derive monitoring data for a specific site in the Republic of Korea by sampling and analyzing $PM_{2.5}$ and its chemical constituents. This is the only result that is no longer available in terms of time and place. (2) To estimate the source types and contributions of $PM_{2.5}$ at a specific site based on the sampled data using receptor models. These results can be used to enhance the understanding of the characteristics of emissions from sources and spatiotemporal characteristics of $PM_{2.5}$. (3) To predict the chemical constituents of $PM_{2.5}$ using machine learning models. This is the application of the latest computer science technology to identify the characteristics of $PM_{2.5}$ in air pollution. (4) To estimate the spatial distribution of $PM_{2.5}$ sources using a multivariate receptor model. This is the first attempt at multivariate spatial distribution modeling in the Republic of Korea. This study draws the latest scientific results from air pollution research.

1.2. Objectives

This thesis aimed to investigate the characteristics of $PM_{2.5}$ for effective management strategies using models of source apportionment and spatiotemporal analysis. The specific objectives for achieving this goal are as follows:

1) To characterize the sources of $PM_{2.5}$ and the inhalation health risks from $PM_{2.5}$-bound heavy metals in a medium-sized industrial city.
2) To assess the applicability of feature extraction using machine learning models to predict the chemical constituents of $PM_{2.5}$ to improve the reliability and availability of the data.
3) To predict latent source-specific $PM_{2.5}$, along with uncertainty estimates at unmonitored sites, using Bayesian multivariate receptor modeling for spatial prediction on a regional scale.

1.3. Dissertation structure

This dissertation comprises six chapters (Fig. 1.1). Chapter 1 describes the background, objectives, and dissertation structure. Chapter 2 reviews previous research related to this study. In Chapter 3, the source apportionment of $PM_{2.5}$ and their health risk by inhalation are demonstrated. Chapter 4 presents the prediction of $PM_{2.5}$ chemical constituents using four machine learning models. The spatial distribution of $PM_{2.5}$ sources in South Korea was estimated using Bayesian spatial multivariate receptor modeling, as described in Chapter 5. Chapter 6 provides a summary and the conclusions of the dissertation.

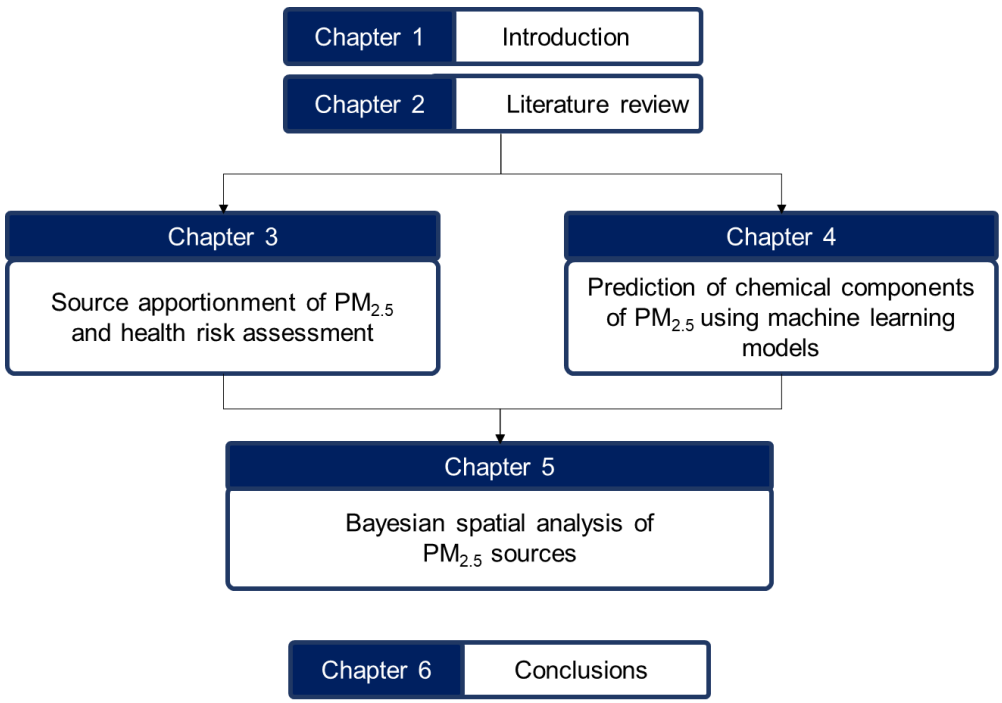


Fig. 1.1. Structure of the dissertation

References

Arora, N.K., 2018. Environmental Sustainability—necessary for survival. Environ. Sustain. 1, 1–2. https://doi.org/10.1007/s42398-018-0013-3

Arya, S.P., 1998. Air Pollution Meteorology and Dispersion - Oxford University Press.

Choi, E., Heo, J.B., Hopke, P.K., Jin, B.B., Yi, S.M., 2011. Identification, apportionment, and photochemical reactivity of non-methane hydrocarbon sources in Busan, Korea. Water. Air. Soil Pollut. 215, 67–82. https://doi.org/10.1007/s11270-010-0459-0

Colvile, R.N., Hutchinson, E.J., Mindell, J.S., Warren, R.F., 2001. The transport sector as a source of air pollution. Atmos. Environ. https://doi.org/10.1016/S1352-2310(00)00551-3

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. https://doi.org/10.1080/10962247.2016.1140693

Hopke, P.K., Dai, Q., Li, L., Feng, Y., 2020. Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.140091

Jacobson, M.Z., 2012. Air Pollution and Global Warming, in: Air Pollution and Global Warming. pp. xix–xx. https://doi.org/10.1017/cbo9781139109444.0020

Jacobson, M.Z., 2002. Atmospheric Pollution: History, Science, and Regulation. https://doi.org/10.1017/CBO9780511802287

Jacobson, M.Z., 1930. Atmospheric Pollution. Lancet.

https://doi.org/10.1016/S0140-6736(01)09050-X

Melamed, M.L., Schmale, J., von Schneidemesser, E., 2016. Sustainable policy—key considerations for air quality and climate change. Curr. Opin. Environ. Sustain. https://doi.org/10.1016/j.cosust.2016.12.003

Popescu, F., Ionel, I., 2010. Anthropogenic Air Pollution Sources, in: Air Quality. https://doi.org/10.5772/9751

Ramanathan, V., Feng, Y., 2009. Air pollution, greenhouse gases and climate change: Global and regional perspectives. Atmos. Environ. 43, 37–50. https://doi.org/10.1016/j.atmosenv.2008.09.063

Sauvé, S., Bernard, S., Sloan, P., 2016. Environmental sciences, sustainable development and circular economy: Alternative concepts for trans-disciplinary research. Environ. Dev. 17, 48–56. https://doi.org/10.1016/j.envdev.2015.09.002

Seinfeld, J.H., Pandis, S.N., 2016. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 3rd Edition | Wiley.

Sharma, N., Agarwal, A.K., Eastwood, P., Gupta, T., Singh, A.P., 2018. Introduction to Air Pollution and Its Control, in: Energy, Environment, and Sustainability. Springer Nature, pp. 3–7. https://doi.org/10.1007/978-981-10-7185-0_1

WHO, 2005. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005 1–21. https://doi.org/10.1016/0004-6981(88)90109-6

Widziewicz, K., Rogula-Kozłowska, W., Loska, K., 2016. Cancer risk from arsenic and chromium species bound to PM2.5 and PM1 – Polish case study. Atmos.

Pollut. Res. 7, 884–894. https://doi.org/10.1016/J.APR.2016.05.002

World Health Organization, 2021. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization.

Chapter 2. Literature review

2.1. Source apportionment and receptor modeling of $PM_{2.5}$

Source apportionment of $PM_{2.5}$ is key to identifying the characteristics of aerosols in the atmosphere (Hopke et al., 2020). As a major tool for source apportionment, receptor modeling based on chemical mass balance and principal component analysis, a statistical method, has been widely used to identify $PM_{2.5}$ sources (Choi et al., 2013; Samara et al., 2003; Yang et al., 2013). Hopke et al. (2020) reviewed the research cases of source apportionment for airborne particulate matter ($PM_{2.5}$ and PM_{10}) from 2014 to 2019 and reported a total of 414 publications conducted in 58 countries worldwide. The number of case studies was 564 and 243 for $PM_{2.5}$ and PM_{10}, respectively. Fig. 2.1 shows source apportionment cases worldwide (Hopke et al., 2020). $PM_{2.5}$ has been studied more recently than PM_{10}. The main pollutant from anthropogenic sources is $PM_{2.5}$ than PM_{10}; $PM_{2.5}$ have many more adverse health effects (Belis et al., 2013; Dai et al., 2015; Park et al., 2004).

The number of identified sources of $PM_{2.5}$ in the literature was primarily five to eight, despite the total range being one to nine (Hopke et al., 2020). However, the characteristics of each source can differ by region and period (E. H. Park et al., 2020; Silva et al., 2020). For example, an industry source is a broad category that can include many relevant sources, such as power plants, incineration, and smelting facilities (Choi et al., 2022). The characteristics of the detailed source, such as the ratio of elements to key elements, differ by region, even though the name of the source is the same (Lv et al., 2021). There are still many limitations, although the names of sources are inferred through key elements and the various characteristics

of each source. It is necessary to accumulate study results that can better reflect the characteristics of the sources in various regions and times (Hopke, 2016). Table. 2.1 shows the results of the classification of source types and their contributions worldwide by Hopke (2020). Such studies are continuously needed to infer the characteristics of a specific region.

In Korea, research results are insufficient. According to Hopke et al. (2020), there are only five source apportionment studies on PM in South Korea. Since then, only a few studies have been published on this topic. Table 2.2 shows the research cases of source apportionment using the PMF model, including domestic and international journal papers. There are fewer than 10 studies. Due to these challenges, there are many difficulties in estimating the source of PM in Korea in detail. Through the accumulation of research results, a consensus can be created on the interpretation of air pollution phenomena. Therefore, it is necessary to gather data on source apportionment through various studies.

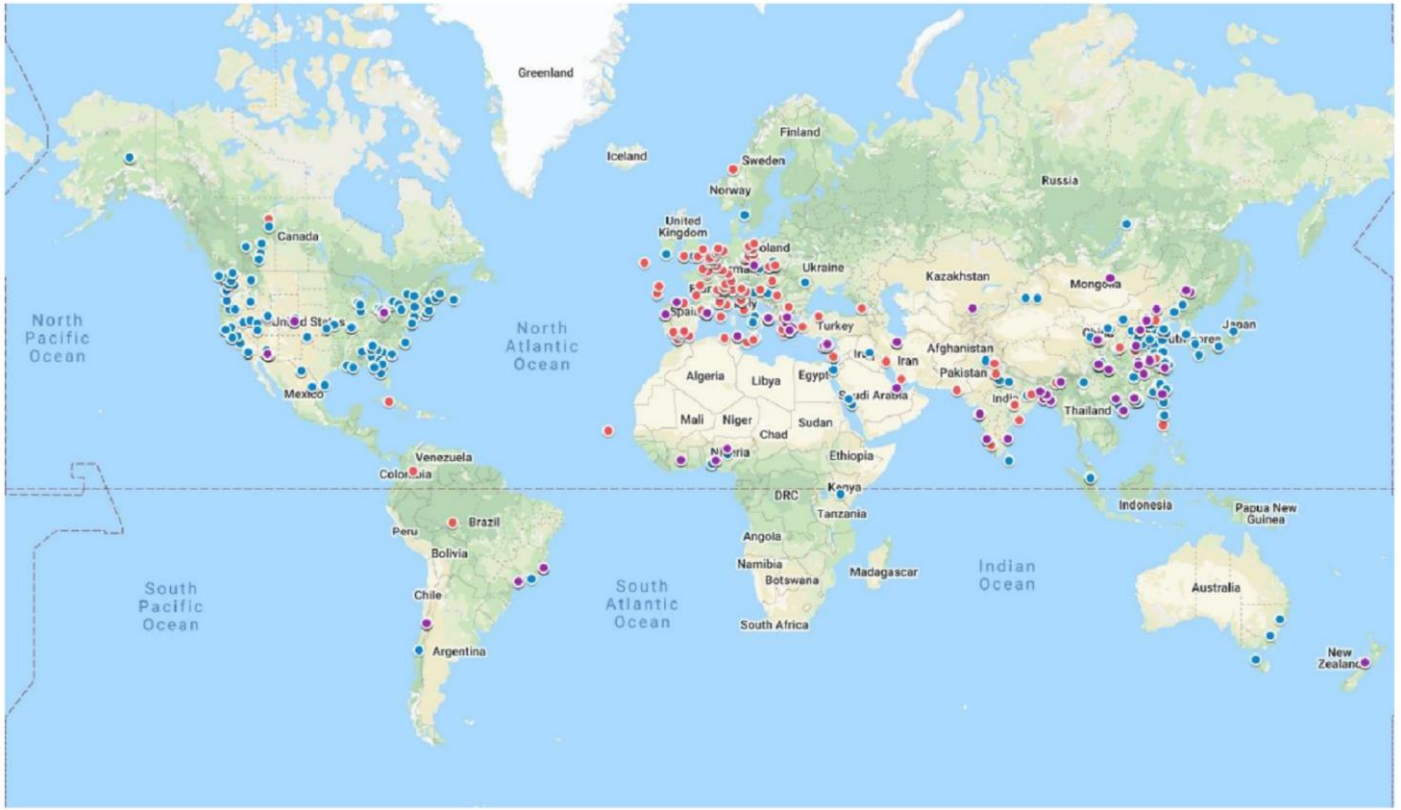


Fig. 2.1. Map showing the locations of the sampling sites for $PM_{2.5}$ (blue points), PM_{10} (red points), and combined $PM_{2.5}/PM_{10}$ (purple points) reported in the identified apportionment publications. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of Hopke et al. (2020)) (Hopke et al., 2020)

Table 2.1. Tabulation of the fractional apportionments by global region or country (Hopke et al., 2020)

Source	Australia and New Zealand	Central America & Caribbean	Eastern Asia -Not China	Eastern Europe	Northern Africa	Northern America -Canada	Northern America -USA	Northern Europe	South America - Brazil	South America - other	South-eastern Asia	Southern Asia	Southern Europe	Western Africa	Western Asia (M49)	Western Europe (M49)	Northern China	Southern China
No. of reports	4	4	27	20	4	23	148	8	7	4	14	46	86	10	12	60	205	41
$PM_{2.5}$ ($\mu g/m^3$)	15.8		27.5	24.7	25.0	8.5	11.0	14.9	22.4	35.1	24.8	102.5	20.4	50.7	46.7	15.6	100.9	50.3
Mixed SIA (%)	11.5		42.3	31.5	0.0	26.8	46.2	25.0	17.1	20.1	23.9	34.3	19.1	21.3	51.8	36.0	31.0	31.0
Sea salt (%)	7.0		6.8	27.1	0.2	7.5	7.2	10.2		16.0	7.8	4.8	19.5	1.4	5.6	4.5	4.1	4.1
Dust (%)	5.1		7.2	18.5	20.0	5.4	2.4	18.8	3.8	13.5	19.1	9.8	19.8	18.4	6.0	12.3	8.9	8.9
Traffic (%)	10.4		17.6	23.8		18.6	8.0	50.4	23.5	35.5	23.0	25.2	26.6	17.9	14.9	19.2	19.3	19.3
Industry (%)	21.5		8.9	15.9		6.7	17.8	8.5	32.5	38.6	15.4	6.4	5.9	8.7	1.0	17.7	16.8	16.8
Biomass. burning (%)	75.1		10.1	17.8		13.8	4.4	15.4	22.4	12.2	16.3	14.9	10.7	2.8	13.5	12.0	10.3	10.3
Coal or no. 6 oil combustion (%)			14.5	32.4		7.9	1.4	11.1	4.4	29.2	13.8	7.1	56.4	13.0	15.9	16.1	10.9	10.9
Other (%)	12.0		11.9	18.5	24.0	10.1	26.0	4.1	17.9	15.7	23.9	10.8	14.5	15.7	1.5	15.2	10.5	10.5
PM_{10} ($\mu g/m^3$)	20.5	35.4	40.5	13.8	56.0	16.2	18.4	28.9	42.5	51.9	92.8	190.0	32.3	220.2	85.3	30.6	164.6	110.0
Mixed SIA (%)	6.0		0.0	48.5			18.6	32.7	17.9	21.0		17.5	26.2	22.2	16.7	35.3	22.3	28.9
Sea salt (%)	13.0	1.0		8.0	4.5		3.3	15.6	14.0			16.5	8.9	2.2	6.2	10.1	6.9	6.0
Dust (%)	5.0	1.0	44.0	16.7	44.1	58.0	24.3	24.6	25.3	28.0		33.5	22.2	26.9	37.3	13.0	30.5	15.2
Traffic (%)		1.0	8.0	17.2	6.0	3.0	15.8	4.9	38.4	23.0		21.4	20.9	20.0	16.6	19.8	14.6	25.6
Industry (%)	48.0			5.9			11.2	6.6	11.1	4.9		19.8	7.4	17.9	9.3	4.5	12.1	19.0
Biomass. burning (%)				14.0		17.0	12.3	3.3	7.6			17.0	11.8		9.4	14.4	14.3	5.9
Coal or no. 6 oil combustion (%)		1.0	25.0	41.7		14.0	4.9	5.6				12.9	5.7	22.1	18.0	7.8	20.2	17.3
Other (%)			23.0	8.1	8.7	7.0	31.4	22.8		23.0		23.2	8.7	35.0	13.8	9.5	11.0	9.0

Table 2.2. Research on source apportionment using PMF model in Korea

Location	Period	No. of source	Contribution	Reference
Seoul, Daejeon, Gwangju, Ulsan	2014-2018	9-10	Secondary sulfate, secondary nitrate, mobile, biomass burning, incinerator, soil, industry, coal combustion, oil combustion, aged sea salt	(Kim et al., 2022)
Seoul	2014−2015	9	Secondary sulfate (20.1%), secondary nitrate (19.0%), vehicles (23.3%), oil combustion (9.07%), soil (8.20%), roadway (3.03%), coal combustion (4.20%), biomass burning (12.2%)	(Park et al., 2020)
Seoul	2014	10	Secondary sulfate (20.8%), secondary nitrate (24.3%), vehicles (15.7%), industry (4.2%), oil combustion (3.4%), soil (2.5%), road dust (1.8%), incinerator (6.8%), coal combustion (9.3%), wood/field burning (13.8%)	(Hwang et al., 2020)
Daejeon	2014	9	Secondary sulfate (20.7%), secondary nitrate (25.3%), vehicles (14.1%), industry (1.6%), oil combustion (4.4%), soil (8.1%), road dust (4.0%), coal combustion (13.4%), wood/field burning (8.4%)	(Hwang et al., 2020)

Busan	2013	8	Secondary sulfate (31%), secondary nitrate (19%), diesel vehicle (6%), gasoline vehicle (12%), industry (3%), road dust (4%), ship (7%), soil (18%)	(Jeong et al., 2017)
Seoul	2013−2014	10	Secondary aerosol (31.2%), motor vehicle (19.2%), break and tire wear (3.5%), coal burning (17.3%), oil combustion (2.0%), waste incineration (9.8%), biomass burning (6.7%), industry (3.7%), sea salt (4.6%), road dust (1.9%)	(Park et al., 2019)
Daebu	2016.05 − 2016.11	9	Secondary sulfate (29%), secondary nitrate (13%), mobile (22%), oil combustion (10%), soil (6%), coal combustion (9%), aged sea salt (8%), industrial activities (1%), non-ferrous smelter (2%)	(Kim et al., 2018)
Gyeongsan	2010.09 − 2012.12	8	secondary sulfate (16.0%), secondary nitrate (20.6%), biomass burning (15.5%), industry (10.4%), soil (7.0%), gasoline (9.1%), incinerator (10.4%), diesel emission (11.0%)	(Jeong and Hwang, 2015)

The following is a detailed summary of the world's source apportionment research cases: Han et al. (2017) identified seven sources and their contributions to $PM_{2.5}$ based on six-year data in Baton Rouge, Louisiana, United States, using PMF modeling. The sources identified were secondary sulfate, secondary nitrate, industrial emissions, traffic, crustal dust, road dust, and sea salt, with contributions of 38.4, 17.6, 18.7, 11.5, 6.1, 4.2, and 3.6%, respectively (Han et al., 2017).

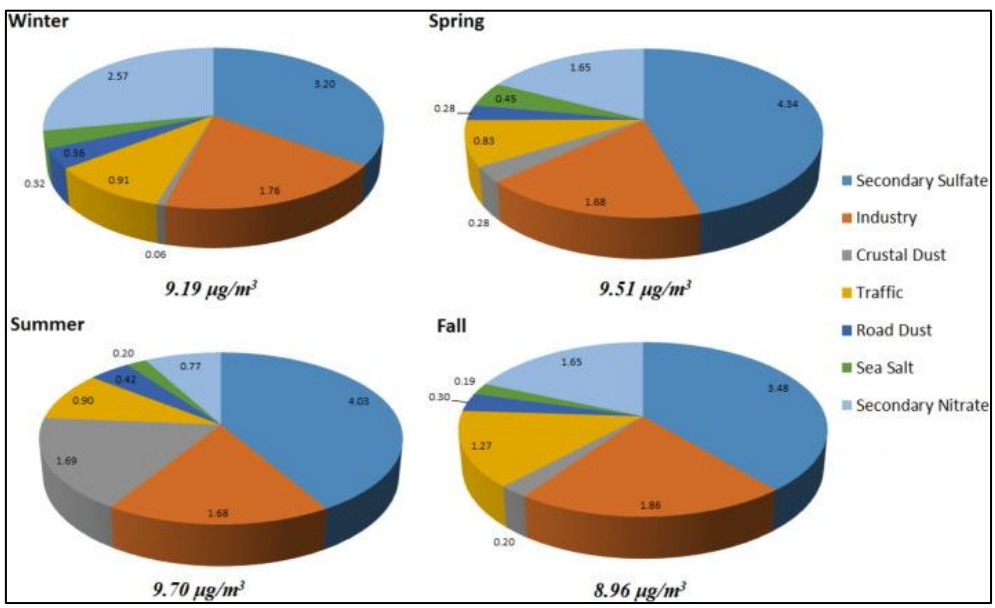


Fig. 2.2. Seasonal variation of source contributions to $PM_{2.5}$ from 2009 to 2014 in Baton Rouge, Louisiana, United States (Han et al., 2017)

Dai et al. (2020) investigated the changes in source contributions of $PM_{2.5}$ after the COVID-19 lockdown. Dispersion-normalized PMF was used for the hourly $PM_{2.5}$ chemical constituents data measured from January 1, 2020, to February 15, 2020, at Nankai University in the Jinan district of Tianjin, China. Fig. 2.2 shows the time series contribution of $PM_{2.5}$, from the study by Dai et al. (2020). Six sources

were identified. The differences between the PMF and dispersion-normalized PMF were analyzed. Additionally, the effects of COVID-19 were studied.

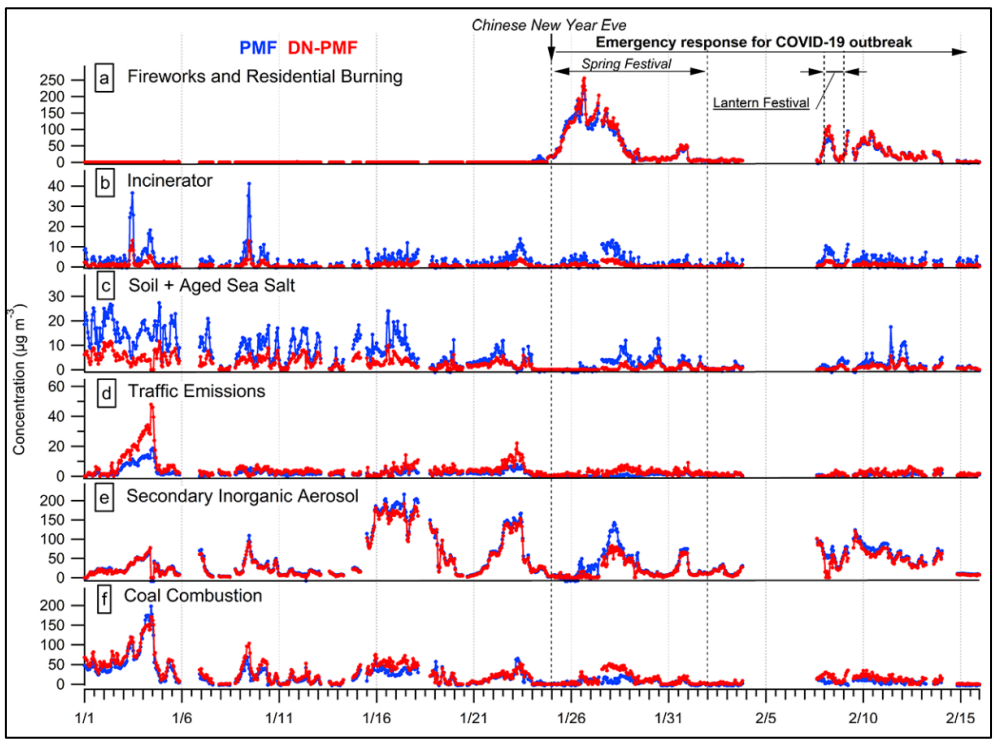


Fig. 2.3. Time series of source contributions estimated using constrained PMF and dispersion-normalized PMF

Park et al. (2020) investigated the long-term trends of source contributions of $PM_{2.5}$ in Seoul, Republic of Korea. PMF modeling was conducted using data from 2014 to 2015. The results were compared with the study that investigated the sources of $PM_{2.5}$ from 2003 to 2007 (Heo et al., 2009). The results reveal that the contribution of mobile sources decreased from 2003 to 2015 (E. H. Park et al., 2020)

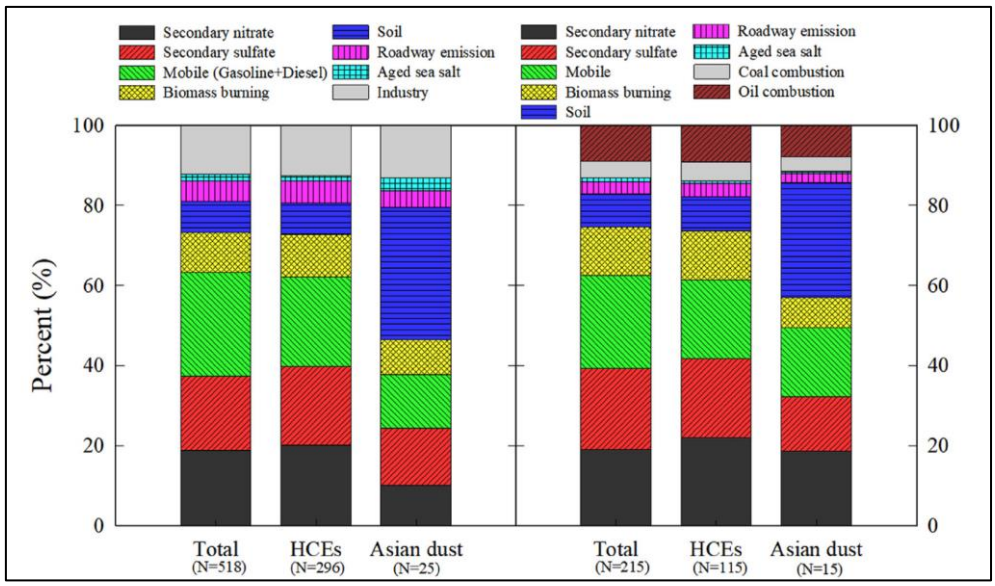


Fig. 2.4. Comparison of the source contributions to $PM_{2.5}$ of Seoul, the Republic of Korea in Heo et al. (left, from 2003 to 2007) and Park et al. (right, from 2014 to 2015)

Positive matrix factorization (PMF) is a widely used model globally as a tool for source appointment of $PM_{2.5}$. The PMF model was developed and distributed by the US EPA. Fig. 2.5 shows the execution image of the PMF modeling program.

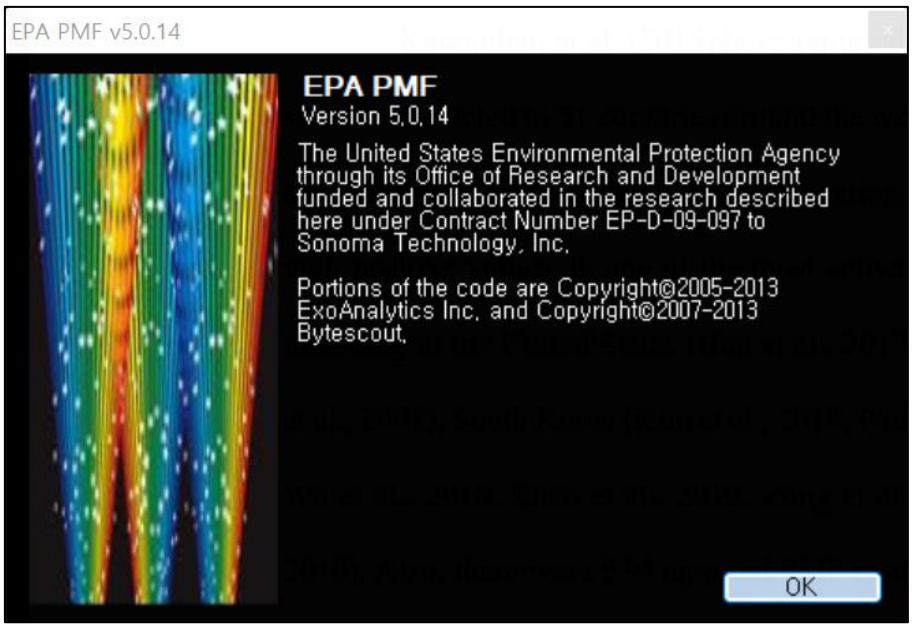


Fig. 2.5. The execution image of the PMF modeling program

Karagulian et al. (2015) reported 419 source apportionment studies conducted in 51 countries worldwide. Among the principal component analysis methods, positive matrix factorization (PMF), which limits factors to those with positive values, is one of the most actively used receptor models worldwide, including in the United States (Han et al., 2017; Paatero and Tapper, 1994; Polissar et al., 2001), South Korea (Kim et al., 2018; E. H. Park et al., 2020), China (Lv et al., 2021; Wu et al., 2018; Zhao et al., 2019; Zong et al., 2016), and Vietnam (Cohen et al., 2010). In addition, there were 539 reported PMF results by Hopke et al. (2020). The PMF model is the most utilized and studied model among existing receptor models (Belis et al., 2013; Hopke et al., 2020; Kumar et al., 2022; Pant and Harrison, 2012). The basic calculation formulae and applications of the PMF model are discussed in Chapter 3.

Notably, PMF modeling has error review capabilities, such as bootstrapping (BS) and displacement (DISP), which lead to relatively accurate source apportionment and are useful for interpreting source profiles based on domain knowledge (Hopke, 2016; Paatero, 1997). Due to these advantages, the PMF model is used the most and is emphasized as an important application point (Hopke et al., 2020). In addition, new approaches have been proposed to improve usability (Brown et al., 2015; Du et al., 2021; Wang et al., 2018). More recently, advanced methods, such as dispersion-normalized (DN) PMF have emerged (Dai et al., 2021, 2020). Matrix factorization with Bayesian methodology has also been used in receptor models (Park et al., 2021, 2018; Park and Oh, 2015). It is necessary to increase the number of research cases in Korea to apply these methods.

2.2. Toxicity and health risk of assessment $PM_{2.5}$

In particular, $PM_{2.5}$ is harmful to human health; accordingly, $PM_{2.5}$ is classified as carcinogenic group 1 by the International Agency for Research on Cancer (IARC) (WHO, 2005; Widziewicz et al., 2016). This group is the same as that of arsenic and benzene, as described in Chapter 1. $PM_{2.5}$ enters the lungs during respiration, adversely affecting human health (WHO, 2005). Table 2.3 shows the penetrability according to the aerosol particle size (Manisalidis et al., 2020).

Table 2.3. Penetrability according to particle size (Manisalidis et al., 2020)

Particle size (μm)	Penetration degree in the human respiratory system
> 11	Passage into nostrils and upper respiratory tract
7–11	Passage into the nasal cavity
4.7–7	Passage into larynx
3.3–4.7	Passage into the trachea-bronchial area
2.1–3.3	Secondary bronchial area passage
1.1–2.1	Terminal bronchial area passage
0.65–1.1	Bronchioles penetrability
0.43–0.65	Alveolar penetrability

Many epidemiological studies have revealed that $PM_{2.5}$ causes respiratory diseases as well as cardiovascular diseases (Atkinson et al., 2014; Hamanaka and Mutlu, 2018; Hopke et al., 2020; Kim et al., 2015, 2022; Li et al., 2013; Manisalidis et al., 2020; Thangavel et al., 2022). Diseases caused by $PM_{2.5}$ are found to be cardiopulmonary disease, cerebrovascular diseases, neurodegenerative diseases,

bronchitis, emphysema, irritation of the eye, asthma, and respiratory infections (Thangavel et al., 2022). However, the mechanisms by which $PM_{2.5}$ affects the human body are still unclear (Thangavel et al., 2022). At the current level of understanding, it is hypothesized that PM inhaled into the lungs causes cellular inflammation, produces free radicals, or causes an imbalance in the nervous system (Manisalidis et al., 2020; Thangavel et al., 2022).

Table 2.4 shows health complications caused by $PM_{2.5}$ (Thangavel et al., 2022). As shown in Table 2.4, $PM_{2.5}$ affects health on short-term as well as long-term exposure. The four effects of $PM_{2.5}$ toxicity (1) pulmonary diseases, (2) cardiovascular diseases, (3) cancers, and (4) neurodegenerative diseases are to be examined in detail. This primarily refers to the literature review of the health effects of $PM_{2.5}$ exposure (Thangavel et al., 2022). In addition, the figures for each health effect were referred to because well represented in the same literature (Thangavel et al., 2022).

Table 2.4. Health complications caused by $PM_{2.5}$ (Thangavel et al., 2022)

Exposure	System Affected	Health Effects
Short term	Cardiovascular	Increased rates of myocardial infarction and ischemia in those at risk Exacerbation of cardiac failure
	Respiratory	Increased incidence of arrhythmia Increased incidence of deep vein thrombosis Increased incidence of stroke Increased wheeze Exacerbation of asthma Exacerbation of chronic obstructive pulmonary disease Bronchiolitis and other respiratory infections Increased incidence of emergency department visits
Long term	Cardiovascular	Increased rates of myocardial infarction Accelerated development of atherosclerosis Increased blood coagulability
	Respiratory	Increase in systemic inflammatory markers Increased incidence of pneumonia Increased incidence of lung cancer Impaired lung development in children Development of new asthma
	Reproductive	Increased incidence of preterm birth Increased incidence of low birth weight
	Brain	Increased risk of Alzheimer's Increased risk of Parkinson's Increased risk of neurodegenerative diseases

Figure 2.6 indicates the underlying mechanisms of chronic obstructive pulmonary disease and asthma from $PM_{2.5}$ (Thangavel et al., 2022). The chemical constituents of $PM_{2.5}$ and $PM_{2.5}$-induced reactive oxygen species (ROS) pose a risk to the respiratory health (Thangavel et al., 2022; Wu et al., 2016). For example, increasing levels of PM increase sore throat, cough, sputum production, wheezing, and dyspnea (Wu et al., 2016).

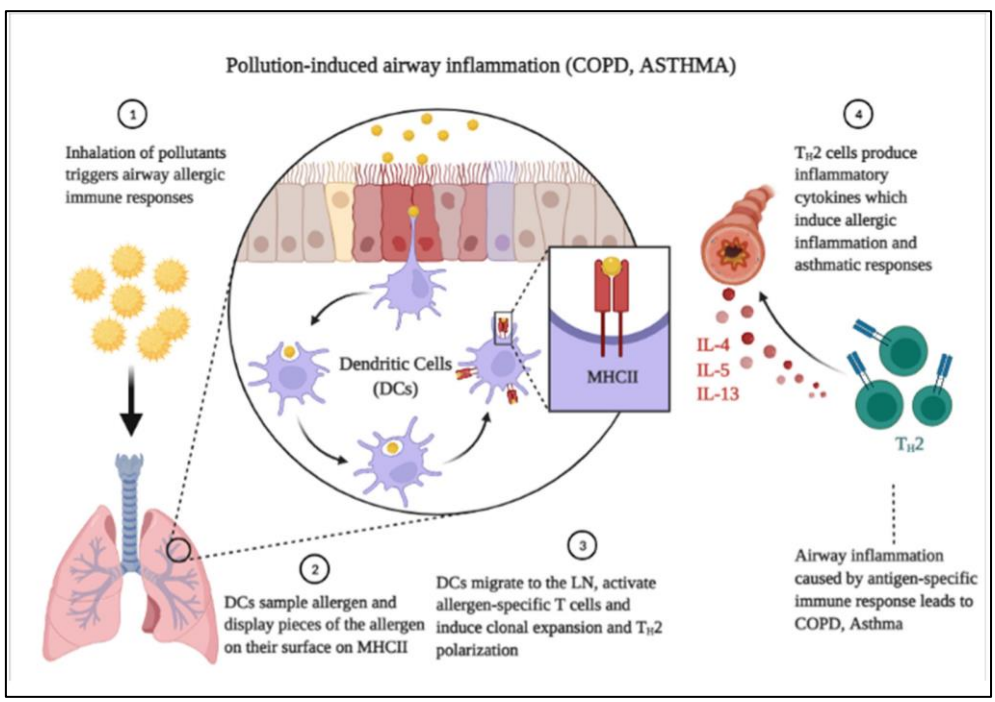


Fig. 2.6. An illustration of underlying mechanisms of $PM_{2.5}$-induced chronic obstructive pulmonary disease and asthma (Thangavel et al., 2022)

Figure 2.7 shows the pathways by which PM promotes cardiovascular impairments (Thangavel et al., 2022). Oxidative stress is the primary response to PM exposure. A recent study suggested that $PM_{2.5}$ causes both cardiac and vascular dysfunctions (Thangavel et al., 2022).

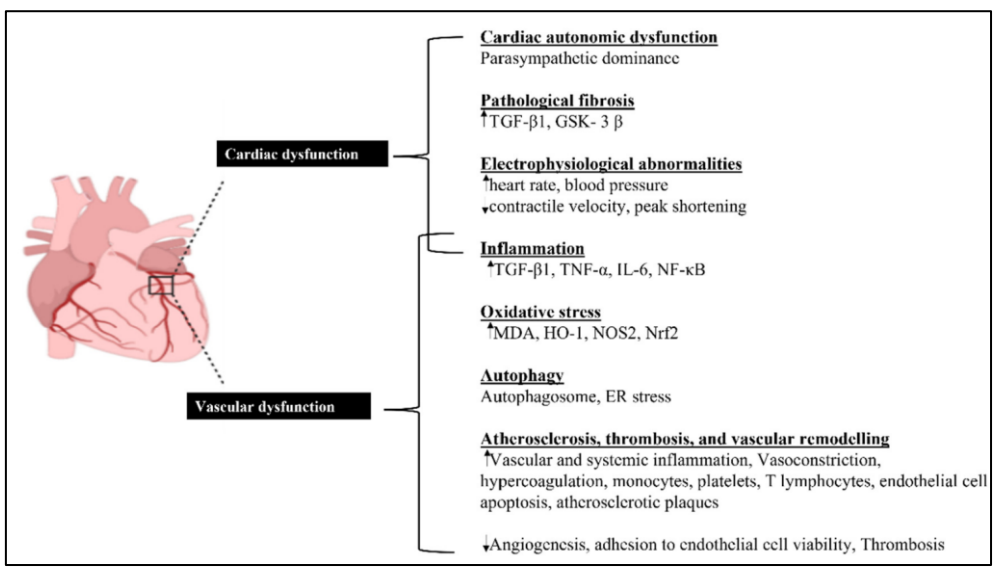


Fig. 2.7. Biological pathways whereby PM particles promote cardiovascular impairments (Thangavel et al., 2022)

A positive correlation between the risk of lung cancer and PM exposure has been previously reported (Hamra et al., 2014). In addition, the American Cancer Society's prospective Cancer Prevention Study II found that $PM_{2.5}$ was significantly positively associated with the death of kidney and bladder cancers from the monitoring data of 623,048 individuals for 22 years (1982–2004) (Thangavel et al., 2022; Turner et al., 2017). Figure 2.8 indicates the potential molecular pathways involved in lung cancer (Thangavel et al., 2022).

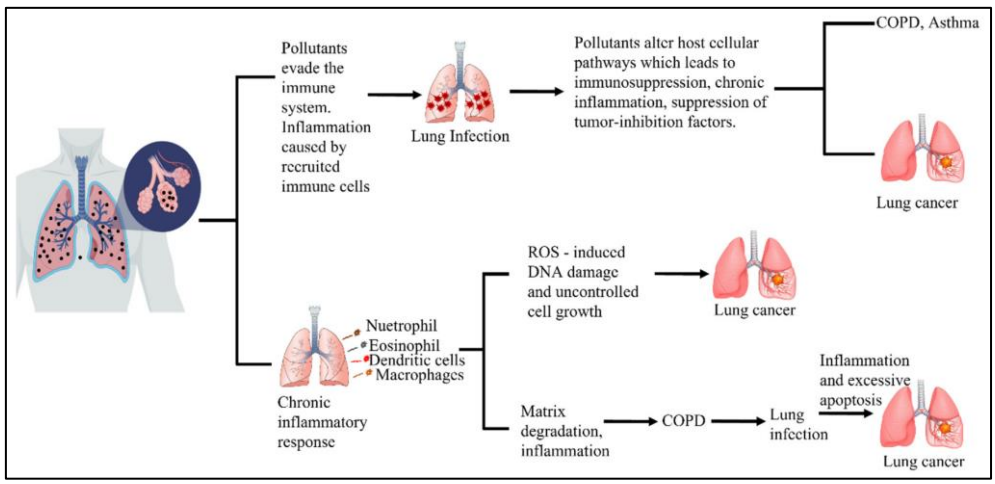


Fig. 2.8. Potential molecular pathways in air pollution-related lung cancer (Thangavel et al., 2022)

It has also been suggested that PM affects the central nervous system and causes neurodegenerative diseases (Costa et al., 2017; Thangavel et al., 2022). PM from diesel exhaust causes electroencephalogram alterations and a general cortical stress response (Crüts et al., 2008). Fig 2.9 shows the effects of air pollution on the nervous system (Thangavel et al., 2022).

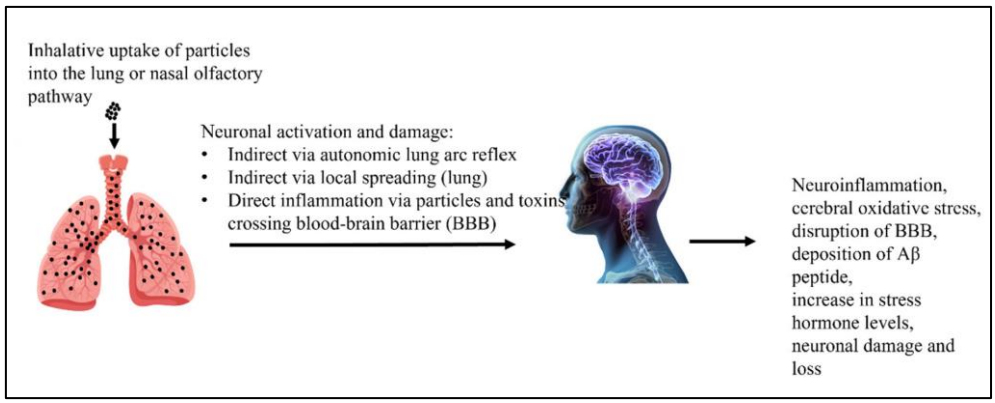


Fig. 2.9. Effects of air pollution on the nervous system and its possible role in neurodegenerative disorders (Thangavel et al., 2022)

Research on the health effects of $PM_{2.5}$ has revealed its toxicity. Epidemiological studies, including experiments and molecular analyses, have also been conducted (Thangavel et al., 2022). However, the toxicity value of PM for health risk assessment has not yet been determined. To assess possible health risks, researchers have performed a health risk assessment for each component based on the concentration of detailed chemical constituents in the PM (Briffa et al., 2020; Choi et al., 2022; Hu et al., 2012; Khillare and Sarkar, 2012; Kim et al., 2022; Lee et al., 2022; Sakunkoo et al., 2022; Yang et al., 2013; Zhao et al., 2021; Zheng et al., 2019). Health risk assessment was conducted using the method described by the US EPA (US EPA, 2009).

The human health risks caused by $PM_{2.5}$-bound heavy metals were calculated using this method. The principal pathway considered is the inhalation of ambient air (Sakunkoo et al., 2022). The health risks posed by heavy metals are divided into non-carcinogenic and carcinogenic (Fan et al., 2021). The International Agency for

Research on Cancer (IARC) classifies As, Ni, Cd, and Cr as Group 1 (carcinogenic to humans), Pb as Group 2A (probably carcinogenic to humans), and Group 2B (possibly carcinogenic to humans) (Zheng et al., 2019). The non-cancer risk was calculated using the hazard quotient (HQ) (Lee et al., 2022; Zhao et al., 2021; Zheng et al., 2019). The detailed calculation procedure is described in Section 3.

Zheng et al. (2019) reported health risk assessment results using $PM_{2.5}$ collected from 2014 to 2016 in Nanjing, China. The results showed that the carcinogenic risks were within the tolerance or acceptable level (1×10^{-6}–1×10^{-4}). The HQ values were less than 1, which implies that there was no significant risk of non-carcinogenic effects and was set by the US EPA. Fig. 2.10 showed the health risk assessment of Zheng et al. (2019).

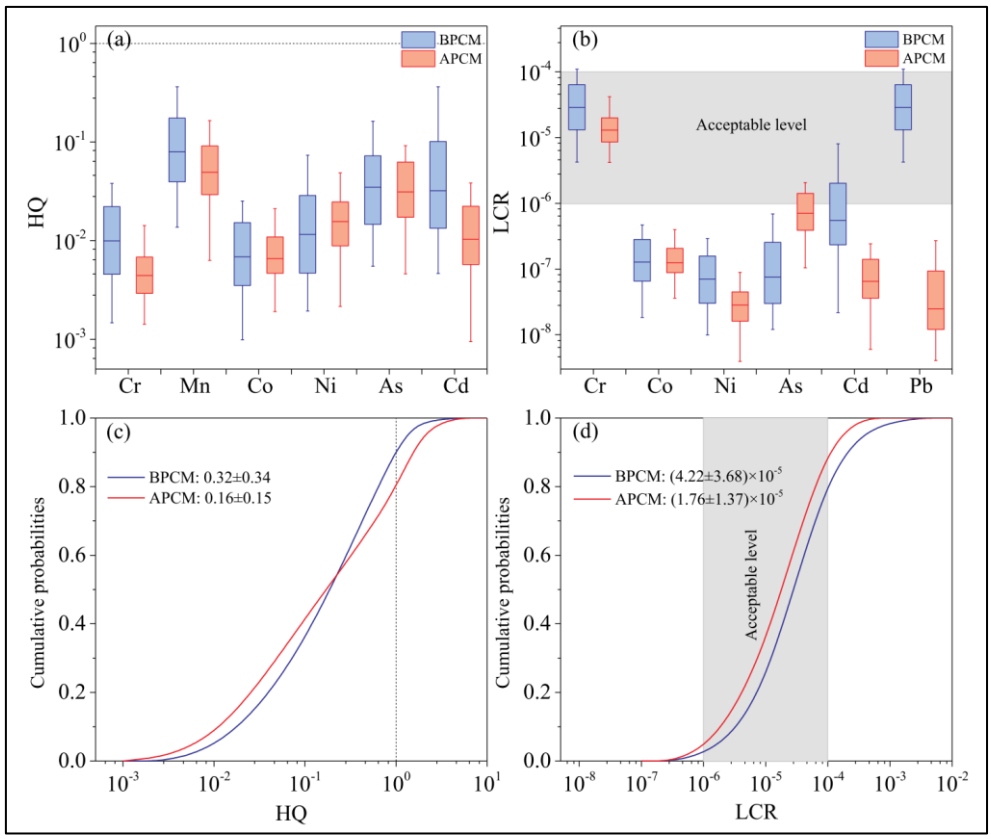


Fig. 2.10. Non-cancer (a) and cancer risks (b) of selected trace elements and their total risk cumulative probabilities (c, d) in $PM_{2.5}$, before and after the release of pollution control measures (BPCM: Jan.–Nov. 2014 and APCM: Nov. 2015–Jul. 2016). Box and whisker plots are constructed by 25–75th and 5–95th percentiles, respectively. (Zheng et al., 2019)

Zhao et al. (2021) performed a health risk assessment using $PM_{2.5}$ from coal-fired power plants in Fuxin, China. The non-carcinogenic risk values of As for children and adults were 45.7 and 4.90, respectively. The carcinogenic risk values of Cr for adults and children were the highest, with values of 3.66×10^{-5} and 2.06×10^{-5}, respectively. These results indicate the need for a response to the high health impact of $PM_{2.5}$.

Khillare and Sarkar (2012) evaluated the health effects of Cr and Ni in PM in Delhi, India. The ILCR values were 1.51×10^{-4} and 1.5×10^{-5} for Cr(VI) and Ni, respectively. It can impact health risks from PM in Delhi, considering lifetime inhalation exposure.

Sakunkoo et al. (2022) reported the human health risks of $PM_{2.5}$-bound heavy metals from anthropogenic sources in Khon Kaen Province, Thailand, between December 2020 and February 2021. According to the results, adults were exposed to risks that were beyond the safe level, showing a high carcinogenic risk in urban areas (residential), industrial zones, and agricultural zones.

As shown thus far, there are many cases where non-carcinogenic and carcinogenic risks are higher than the safety level in health risk assessment studies conducted in East Asia. However, these studies have a limitation such that it was possible to evaluate only the components whose toxicity values were provided by the US EPA. This means that the health impact may be underestimated compared with the actual health impact. Therefore, the health risk assessment of $PM_{2.5}$ needs to be studied constantly.

2.3. Machine learning approaches in the prediction of $PM_{2.5}$

Machine learning models, which have recently been in the spotlight, can be used to interpret complex phenomena (Jordan and Mitchell, 2015). Accordingly, concepts introduced in computer science are used in the analysis of earth sciences (Kelp et al., 2020; Zhong et al., 2021). They have been successfully used in flow pattern analysis, weather analysis, and air quality prediction (Hadeed et al., 2020; Hu et al., 2017; Lyu et al., 2019; Yao and Ruzzo, 2006).

Recently, attempts have been made to develop models to predict air pollution using machine learning (Chang et al., 2020; Kelp et al., 2020; Reichstein et al., 2019; Zhong et al., 2021). Machine learning models work by analyzing data, looking for specific patterns and rules, and making predictions when given a sufficient amount of data (Alpaydin, 2020). Previous studies have successfully predicted the concentrations of $PM_{2.5}$, PM_{10}, and gaseous air pollutants (such as sulfur dioxide [SO_2], nitrogen dioxide [NO_2], and ozone [O_3]) using machine learning (Castelli et al., 2020; Chang et al., 2020; Zhong et al., 2021).

Lyu et al. (2019) improved the accuracy of $PM_{2.5}$ predictions in China using an ensemble of a deep neural network and a community multiscale air quality (CMAQ) model. The results indicated that the prediction of accuracy concentration of $PM_{2.5}$ increased from 0.39 to 0.64 in R^2, and the root mean squared error (RMSE) decreased from 33.7 to 24.8 $\mu g/m^3$ (Lyu et al., 2019). Fig 2.11. shows the prediction accuracy results using (a) the CMAQ model only and (b) the fusion model.

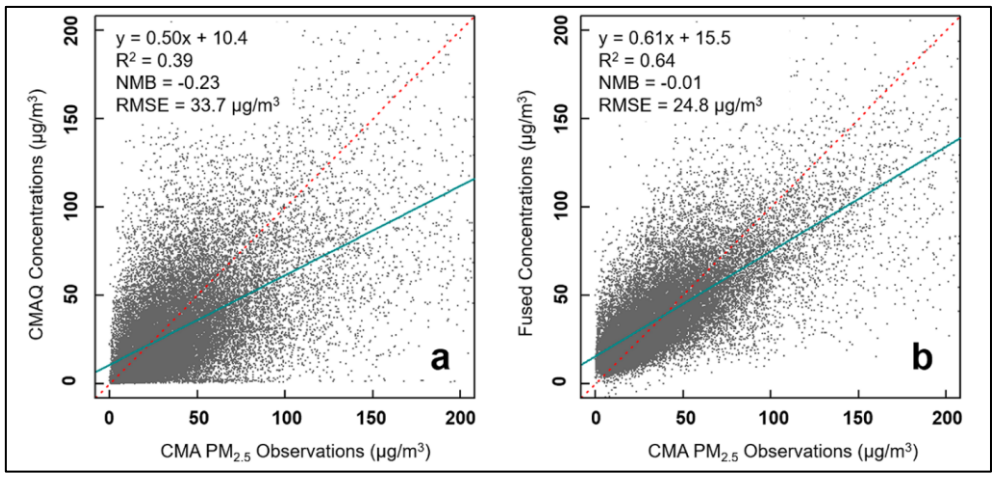


Fig. 2.11. Scatter plots of (a) raw CMAQ simulations and (b) final fusion product, evaluated with independent China Meteorology Agency (CMA) observations in 2016. The green line reflects the linear regression of predictions against observations; the dashed red line represents the one-to-one line indicating perfect agreement (Lyu et al., 2019)

Hu et al. (2017) estimated daily average $PM_{2.5}$ concentrations in the United States with an accuracy of $R^2 = 0.80$ using the Random Forest (RF) algorithm, a machine learning technique. Fig. 2.12 shows the prediction results and the differences between the predicted and observed $PM_{2.5}$ concentrations of Hu et al. (2017).

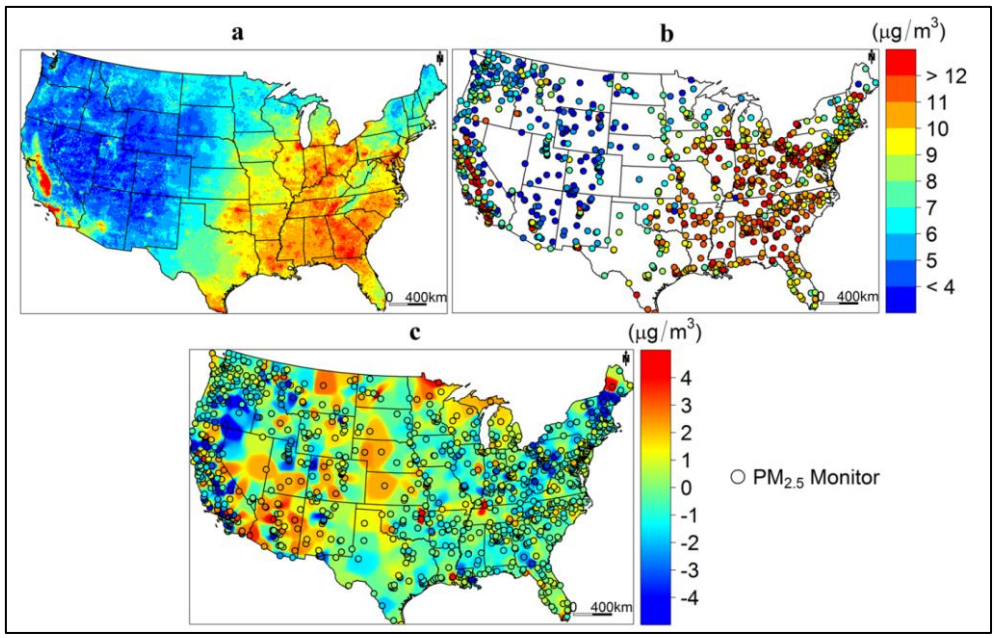


Fig. 2.12. Annual mean predictions. (a) Annual mean $PM_{2.5}$ predictions over the continental United States for 2011; (b) annual mean $PM_{2.5}$ measurements at ground monitors; (c) difference between annual mean predictions and observations at ground monitors and difference interpolations over the continental United States (Hu et al., 2017)

Table 2.5 presents recent studies to predict air pollution using machine learning models. Most of the studies predicted the concentrations of PM_{10}, $PM_{2.5}$, O_3, NO_2, CO, and SO_2. However, most of them predicted the mass concentration of PM or the concentration of gaseous air pollutants. The use of machine learning in predicting $PM_{2.5}$ constituents has not been reported, even though $PM_{2.5}$ chemical constituents provide information about the origin and hazard of $PM_{2.5}$ (Zheng et al., 2019).

Table 2.5. Research on predicting air pollution using machine learning models

Location	Model used	Prediction target	Additional data used	Prediction accuracy	Reference
USA	RF	$PM_{2.5}$ (Multiple sites)	Aerosol optical depth, land use variables, meteorological data	0.80 in R^2	(Hu et al., 2017)
USA	Various models (8 models)	Missing value of $PM_{2.5}$ (within 24 hours)	$PM_{2.5}$	0.32−0.65 in R^2	(Hadeed et al., 2020)
USA (California)	SVR*	Air quality index (category)	SO_2, O_3, NO_2, CO, $PM_{2.5}$, wind speed, temperature, humidity	94.1% accuracy	(Castelli et al., 2020)

China	DNN, RF (ensemble)	$PM_{2.5}$ (Multiple sites)	Numerically modeled data	0.39−0.64 in R^2	(Lyu et al., 2019)
Chile	kNN, linear regression, etc.	Missing value of $PM_{2.5}$ (Daily average)	$PM_{2.5}$, PM_{10}, NO_x, O_3, CO, temperature, humidity, wind speed, rainfall	0.37−0.91 in R^2	(Quinteros et al., 2019)
Taiwan	RNN** (LSTM***)	$PM_{2.5}$, PM_{10} (hourly future)	SO_2, O_3, NO, NO_2, NO_x, CO, rainfall, data time, month, weekday, and hour	30−40% error in 8-hour prediction	(Chang et al., 2020)

* SVR: Support vector regression

** RNN: Recurrent neural network

***LSTM: Long short-term memory

The commonly used machine learning models include (1) fully connected deep neural networks, (2) Random Forest, (3) k-nearest neighbor, and (4) generative adversarial imputation networks.

The fully connected deep neural network is specialized in feature extraction and is one of the most widely used neural network models for nonlinear regression (Hinton and Salakhutdinov, 2006; Hwangbo et al., 2021). The DNN model was trained by adjusting the weights and biases of the hidden layer neurons to correspond to the input and output data, respectively. For the models, avoiding overfitting and optimizing hyperparameters is crucial to develop a model with high prediction accuracy, with training as well as actual field data (Montavon et al., 2018). Fig. 2.13 shows the structure of the deep neural network model. The hyperparameters of the model are indicated by blue boxes. The input and output data can be adjusted according to the convenience of the analyst.

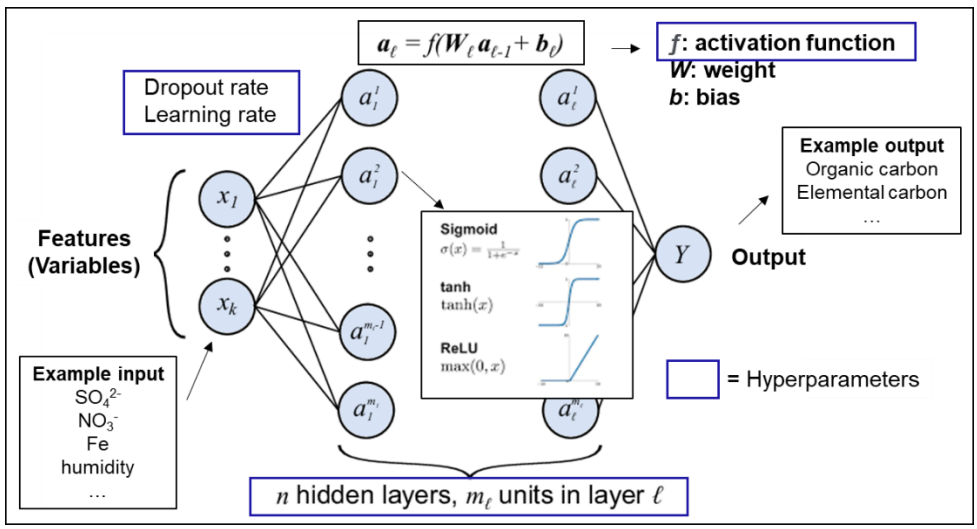


Fig. 2.13. Structure of a deep neural networks model

A generative adversarial imputation network is a missing-value processing model that competes with learning and improves accuracy using a generator and discriminator (Li et al., 2019; Nazábal et al., 2020; Yoon et al., 2018). Fig. 2.14 shows the architecture and learning process of the Generative adversarial imputation network. This model was presented first by Yoon et al. (Yoon et al., 2018). A generative adversarial imputation network has the characteristic of being able to use data with missing values without modification (Ivanov et al., 2018) and has been recently used in various fields for processing missing values (Andrews and Gorell, 2020; Popolizio et al., 2021; Viñas et al., 2020). This is based on the basic assumption that missing values in the data occur randomly (Yoon et al., 2018).

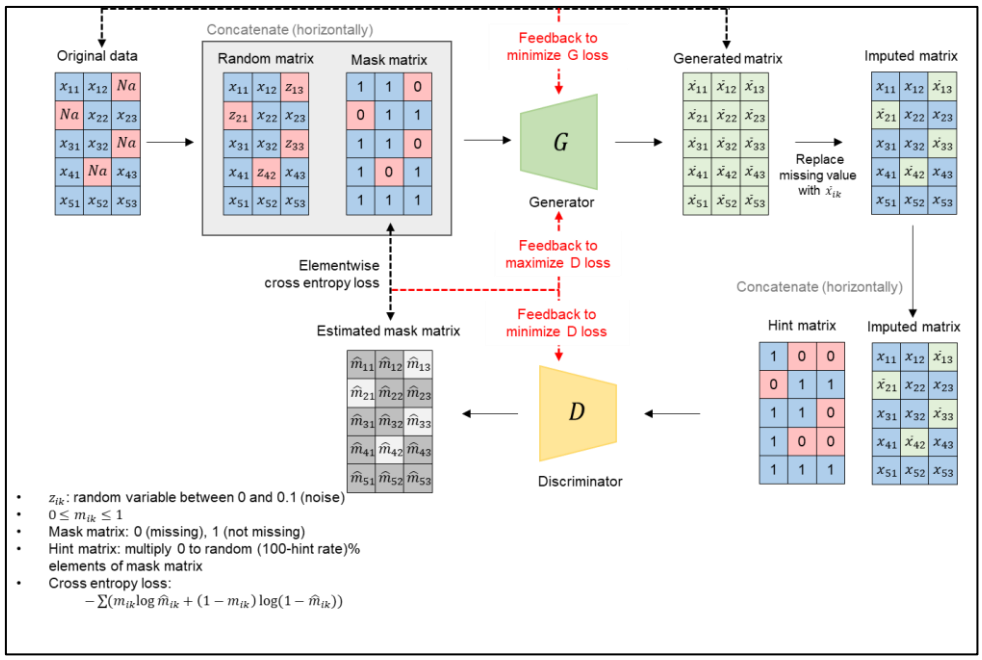


Fig. 2.14. Architecture and training process of the generative adversarial imputation network model

The RandomForest algorithm, proposed by Breiman (2001), is an ensemble model widely employed for multi-dimensional classification and regression problems (Breiman, 2001). Various decision trees in RandomForest models are trained to enhance the model performance (Tella et al., 2021). Fig. 2.15 shows a brief description of the branch division of a tree and the calculation of the predicted value. RandomForest has shown outstanding prediction results in situations where the number of variables is larger than the number of monitored data (Biau and Scornet, 2016).

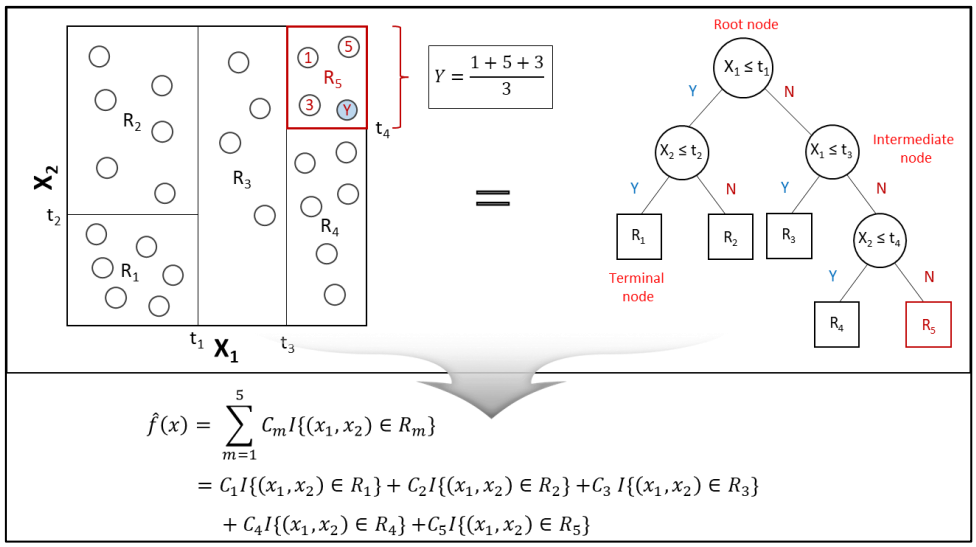


Fig. 2.15. Schematic diagram of a tree, branch division, and predicted value calculation

The k-nearest neighbor algorithm is a non-parametric model for classification and regression, wherein the prediction object is calculated as the average of the k values closest to the prediction point (Tella et al., 2021; Yao and Ruzzo, 2006). The Euclidean distance for the judgment of the nearest neighbor is used to calculate the distance in the k-nearest neighbor algorithm. Fig. 2.16 shows the calculation procedure for the unknown value.

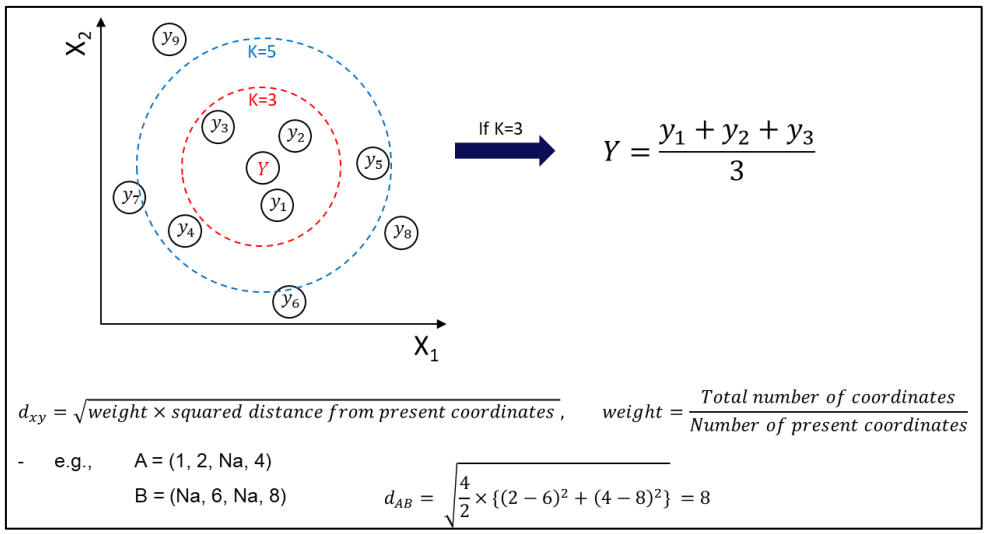


Fig. 2.16. Schematic diagram of the calculation process of k-nearest neighboring algorithm

Thus far, we have investigated commonly used machine learning models. However, as mentioned in Table 2.5, there are a few examples of such machine learning research applications. There are especially few applications in Korea. Therefore, there is a need to diversify studies on machine learning applications in air pollution.

2.4. Bayesian approach in source apportionment

The Bayesian method has been of interest in source apportionment studies in recent decades (Hopke, 2016). Bayesian factor analysis has advantages that can overcome challenging problems in factor analysis, such as uncertainty estimation and rotational ambiguity (Park and Tauler, 2020). Domain knowledge can be incorporated into parameter estimations in Bayesian source apportionment models (Park and Tauler, 2020).

Despite these strengths, there have been limited studies on source apportionment using the Bayesian method. Hopke (2016) pointed out that the conceptual framework and statistical computations of Bayesian source apportionment are complex, which makes it difficult to use the model. Bayesian source apportionment has not been widely applied (Park and Tauler, 2020). Continuous research is needed to increase usability and exploit its advantages for advancing source apportionment techniques. This chapter thoroughly investigates the literature using the Bayesian approach in source apportionment. Table 2.6 shows the applications of Bayesian factor analysis to source apportionment.

Table 2.6. Research using Bayesian approach in source apportionment

Research summary	Location	Reference
- Source apportionment of particle number size distribution using Bayesian Dirichlet process model - Identification of a sources in London Gatwick Airport	UK	(Baerenbold et al., 2022)
- Presenting user-friendly software tools to implement Bayesian multivariate receptor modeling - Example analysis of $PM_{2.5}$ dataset from El Paso, USA (4 sources identified)	USA	(Park et al., 2021)
- Incorporating latent source profiles and meteorological conditions using Bayesian hierarchical source apportionment model - Identification of major sources in two study areas of northern Taiwan	Taiwan	(Tang et al., 2020)
- Development of Bayesian spatial multivariate receptor model to enable predictions of source contributions at any unmonitored site - Identification of 5 sources from 9 monitoring sites in Harris County, Texas	USA	(Park et al., 2018)
- present a source-specific health effects evaluation approach within a Bayesian framework that can handle both parameter uncertainty and model uncertainty in source apportionment under Poisson health outcome models	USA	(Park and Oh, 2018)

- Presenting a new flexible source apportionment approach, Bayesian quantile multivariate receptor modeling - Dealing with the non-normality of air pollution data and outliers	USA	(Park and Oh, 2016)
- Extending the previous Bayesian multivariate receptor modeling to account for (1) nonnegativity constraints and (2) outliers by considering a heavy-tailed error distribution - Identification of 6 sources in Phoenix, Arizona	USA	(Park and Oh, 2015)
- Development of a multipollutant approach that incorporates both sources of uncertainty into the assessment of source-specific health effects - Development of enhanced multivariate receptor models that can account for spatial correlations in the multipollutant data collected from multiple sites	USA	(Park et al., 2015)
- Bayesian receptor modeling incorporating a priori information about the source emissions from national database - Application of the model in 2 locations in USA (Boston, Massachusetts and Phoenix, Arizona)	USA	(Hackstadt and Peng, 2014)
- Evaluating the source-specific health effects associated with an unknown number of major sources of multiple air pollutants - Estimating source contributions along with their uncertainties and model uncertainty	USA	(Peak et al., 2014)

- Using Dirichlet distribution to extend the receptor model for time-varying source profiles - Evaluation of the extended model using the dataset of St. Louis, Illinois	USA	(Heaton et al., 2010)
- Dirichlet based Bayesian receptor modeling to incorporate a prior information on source profiles - Comparison the simulation results of the Bayesian receptor modeling to PMF modeling	USA	(Lingwall et al., 2008)
- Estimating the source spatial profiles using Bayesian approach - Identification of 2-3 sources of PM_{10} in Seoul using the data of 17 monitoring sites	Korea	(Park et al., 2004)
- Proposing Bayesian approach that can handle the unknown number of pollution sources and identifiability conditions - Dealing with model uncertainties in receptor models by using Makov chain Monte Carlo (MCMC) schemes	USA	(Park et al., 2002)
- Development of time-series extension of multivariate receptor models to account for temporal correlation in parameter estimation - Application of the model in Atlanta	USA	(Park et al., 2001)

Baerenbold et al. (2022) applied the Bayesian Dirichlet process model for source apportionment of the particle number size distribution measured near London Gatwick Airport, UK, in 2019. Nine sources were identified, and the results were compared with those of Tremper et al. (2022). The estimated particle-size distributions for each source are shown in Fig. 2.17.

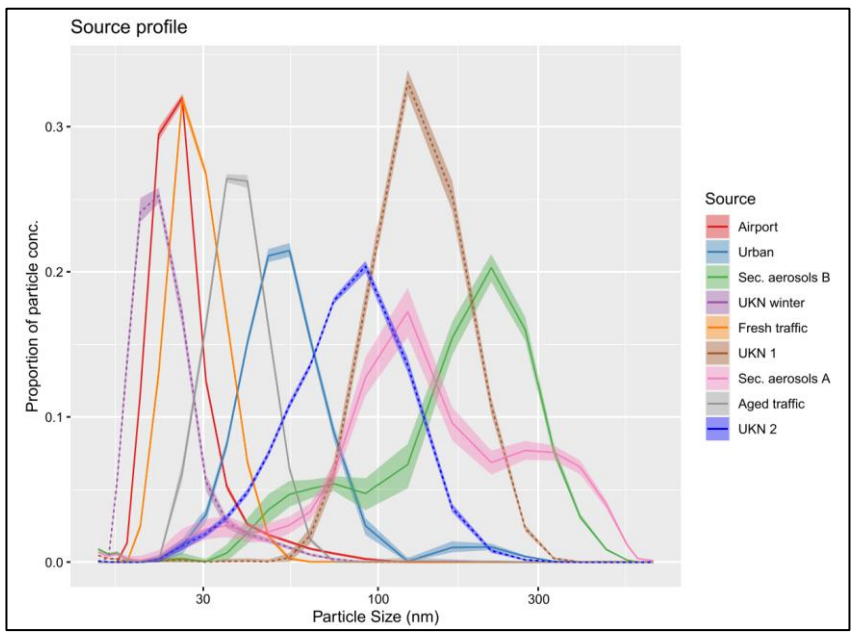


Fig. 2.17. Particle size distribution for the 9 sources identified by the model. Solid lines represent the sources which were also found using PMF in Tremper et al. (2022), while dashed lines are from Baerenbold et al. (2022)

Park et al. (2021) presented user-friendly software tools to implement Bayesian receptor modeling for the convenience of the investigators. The tools were developed for use in MATLAB and R software. This is expected to solve the problem of the low accessibility of Bayesian source apportionment modeling, which has been mentioned previously (Hopke, 2016; Park and Tauler, 2020).

Tang et al. (2020) identified major sources in two study areas of northern Taiwan (Shimen and Taipei) using a Bayesian hierarchical model. The results of Bayesian receptor modeling were compared with the results of PMF modeling using simulated data to estimate the performance of the models. Fig. 2.18 shows a comparison of the source profiles obtained from the Bayesian model and the PMF model (Tang et al., 2020). The Bayesian model showed a better performance. Based on these results, Tang et al. (2020) proposed a multivariate source apportionment model using a Bayesian framework for latent source profiles to incorporate domain knowledge, such as emissions and meteorological data. This method can be used to avoid restrictive assumptions (Tang et al., 2020).

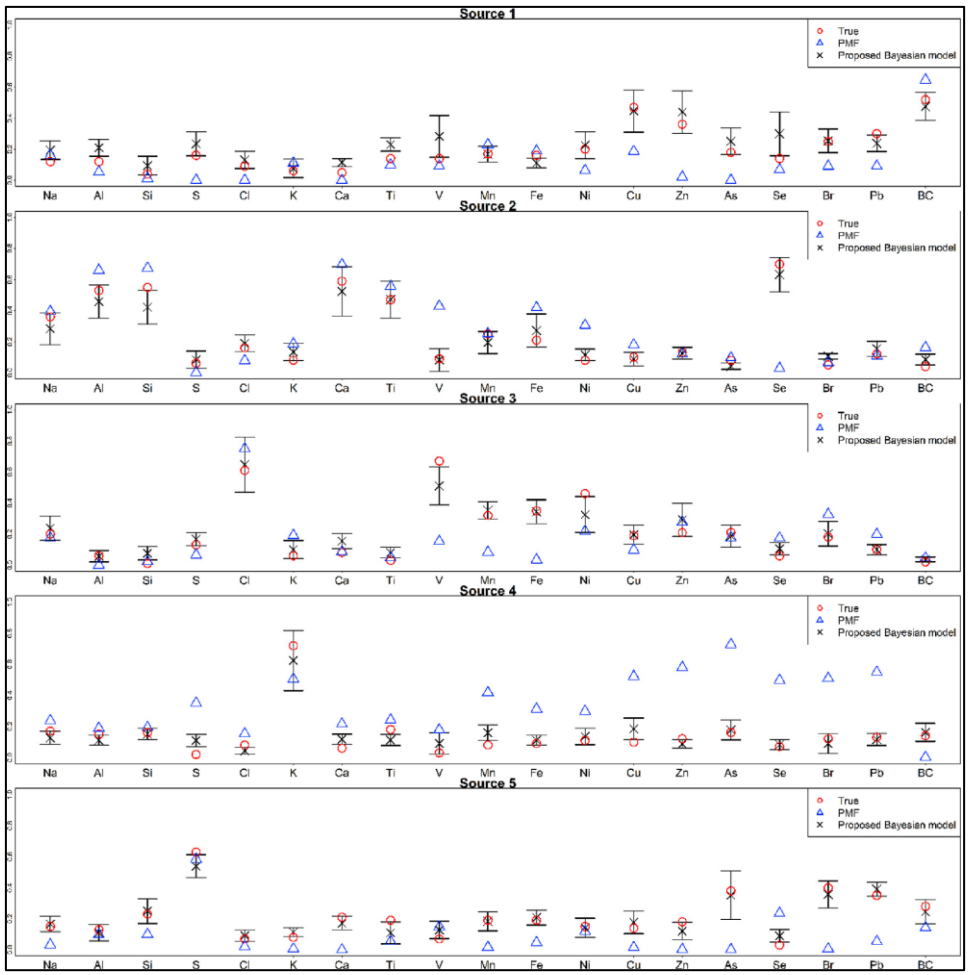


Fig. 2.18. Comparison of the estimated source profiles obtained from the proposed Bayesian and PMF models with the true values (Tang et al., 2020)

Park et al. (2018) proposed a Bayesian spatial multivariate receptor model that can incorporate multisite multipollutant data and predict the source apportionment results at any unmonitored location. The model used 17 volatile organic compound data collected from nine monitoring sites in Harris County, Texas, United States, and predicted the source contributions of five major sources (Park et al., 2018). Fig. 2.19 shows the predicted surface of the source contribution from

Bayesian spatial multivariate receptor modeling. This is the first study to predict the surface map of the source contributions using Bayesian receptor modeling. The method and outcome of this research can considerably aid in developing effective pollution control strategies in cities with no multi-pollutant data. They are expected to be used in various applications.

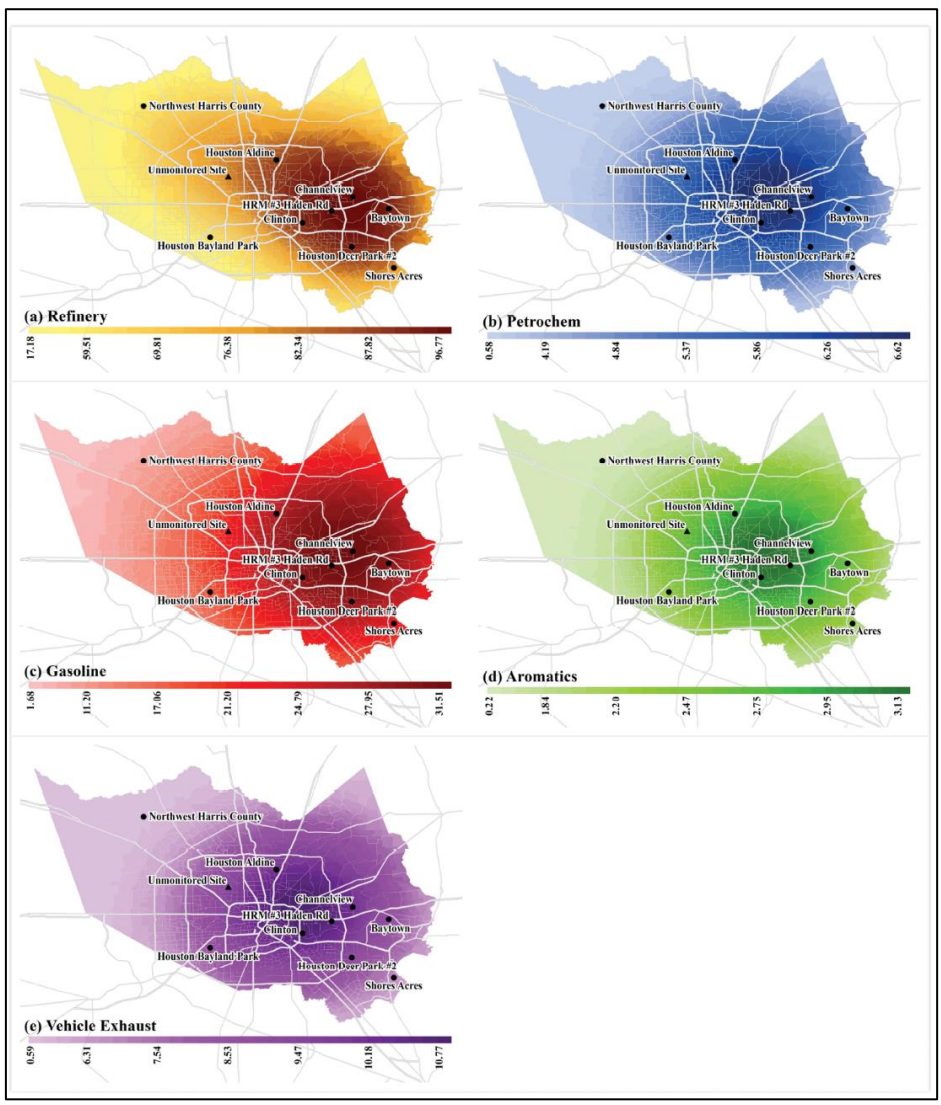


Fig. 2.19. Predicted source contribution surface for Harris County on December 12, 2005 (Park et al., 2018)

Park and Oh (2015) proposed a robust Bayesian receptor model to estimate the uncertainty of source contributions and source profiles by extending previous Bayesian multivariate receptor modeling to account for (1) non-negativity constraints and (2) outliers by considering a heavy-tailed error distribution (Park and Oh, 2015). The proposed robust Bayesian receptor modeling was investigated using simulated data and monitored $PM_{2.5}$ speciation data from Pheonix, Arizona, USA. Fig. 2.20 and Fig. 2.21 show the results of robust Bayesian receptor modeling of the simulated data and the monitored data, respectively (Park and Oh, 2015). In the simulation results, the modeling results tended to agree well when the data contained outliers (Fig. 2.20). In practical applications, six sources were identified with uncertainty estimates of 95% posterior intervals (Fig. 2. 21). This approach can provide uncertainty estimates for both source contributions and profiles, coping with unknown identifiability conditions.

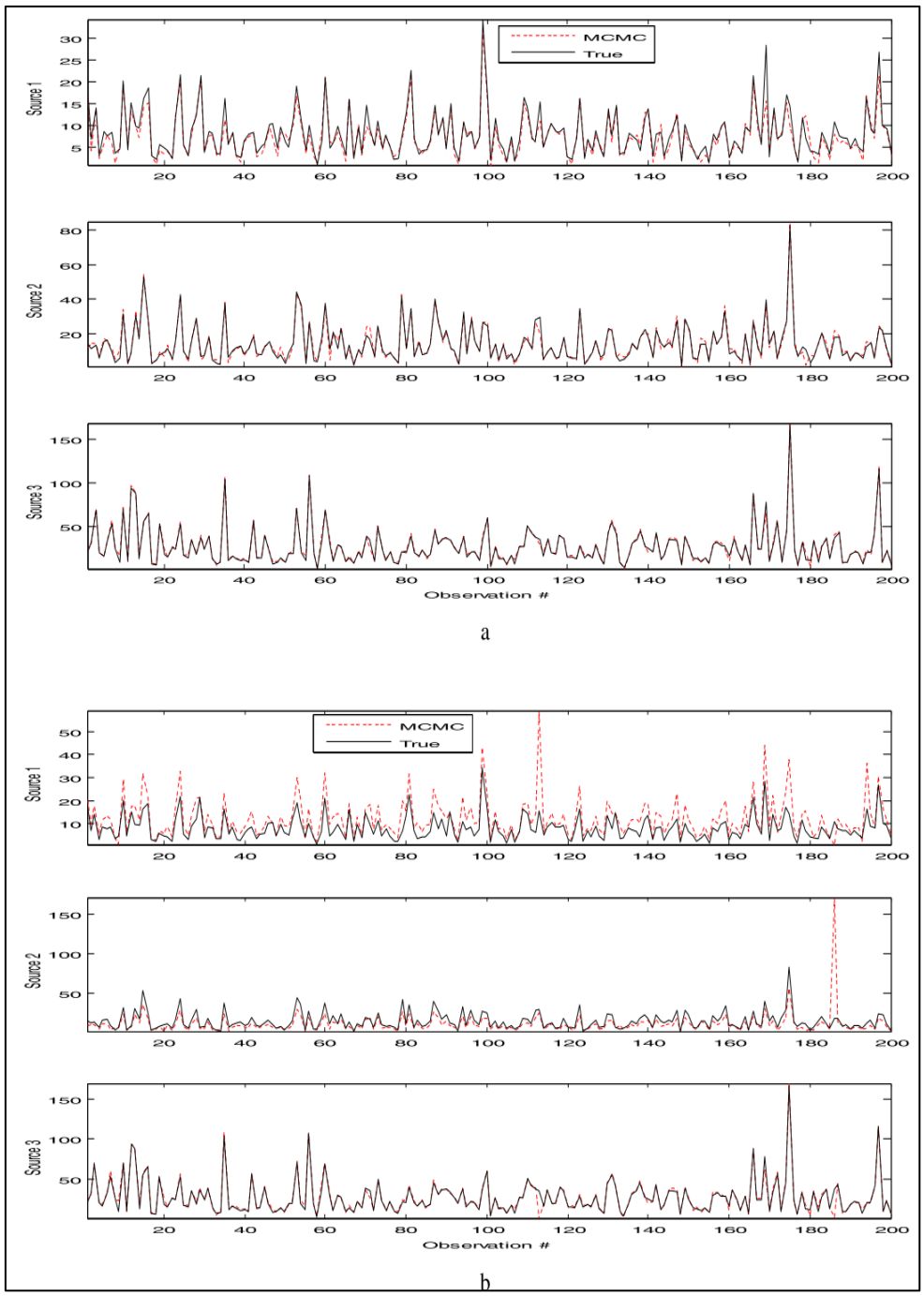


Fig. 2.20. Time series plots of the true source contributions and estimated source contributions using (a) Method T and (b) Method G when the data contain outliers (Park and Oh, 2015)

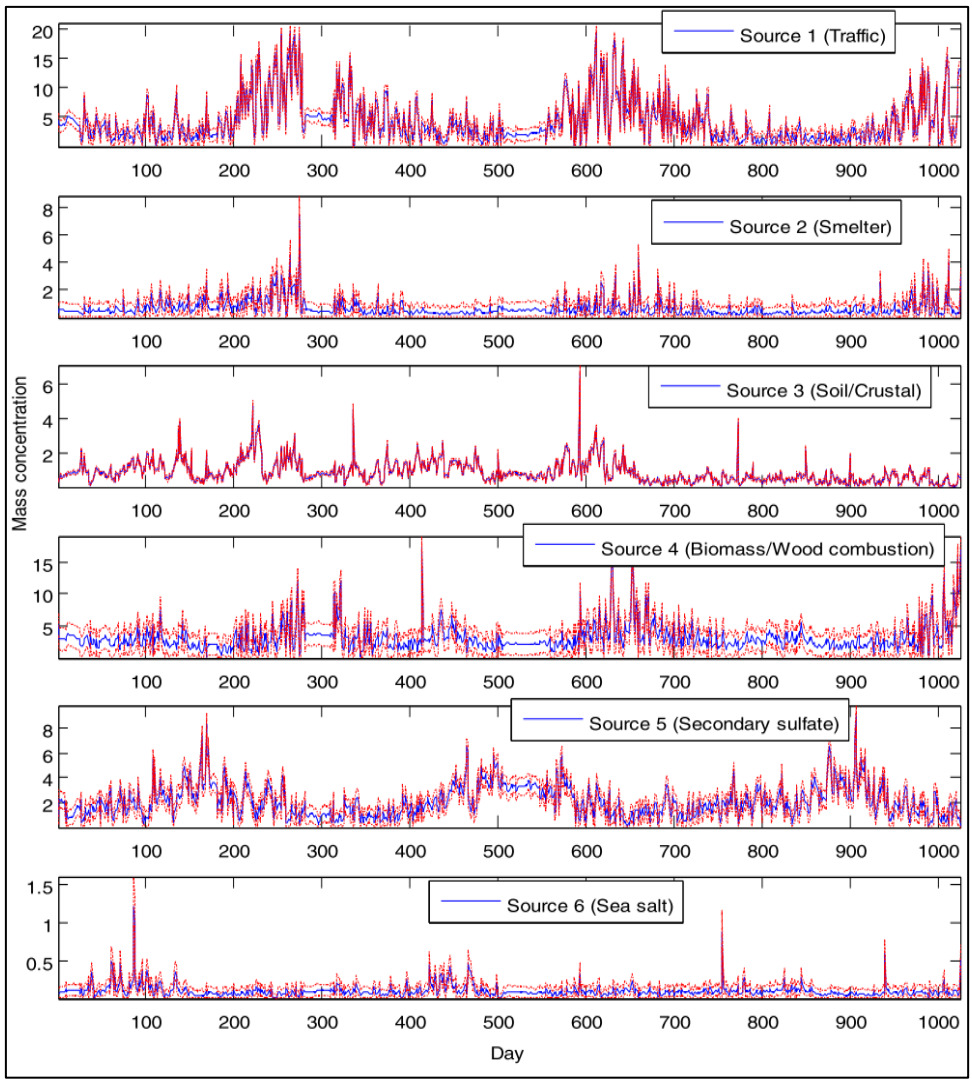


Fig. 2.21. Time series plots of the estimated source contributions (in μg/m3) using Method G for 1027 days with their uncertainty estimates (95% posterior intervals) represented by dashed lines (Park and Oh, 2015)

Hackstadt and Peng (2014) proposed a Bayesian source apportionment model that incorporates a priori information about source emissions from a national database. The proposed model was also applied to two locations in the USA (Boston, Massachusetts, and Phoenix, Arizona). The authors concluded that uncertainties in

the source contributions should not be ignored.

Heaton et al. (2010) used Dirichlet distribution to extend the receptor model for time-varying source profiles. Fig. 2.22 shows the source profile of the zinc smelter source from the results of time-varying receptor modeling (model proposed by Heaton et al. (2010)) and time-constant receptor modeling (PMF model). The authors pointed out that time-varying source profiles were empirically and physically justifiable and could reduce the estimation error (Heaton et al., 2010).

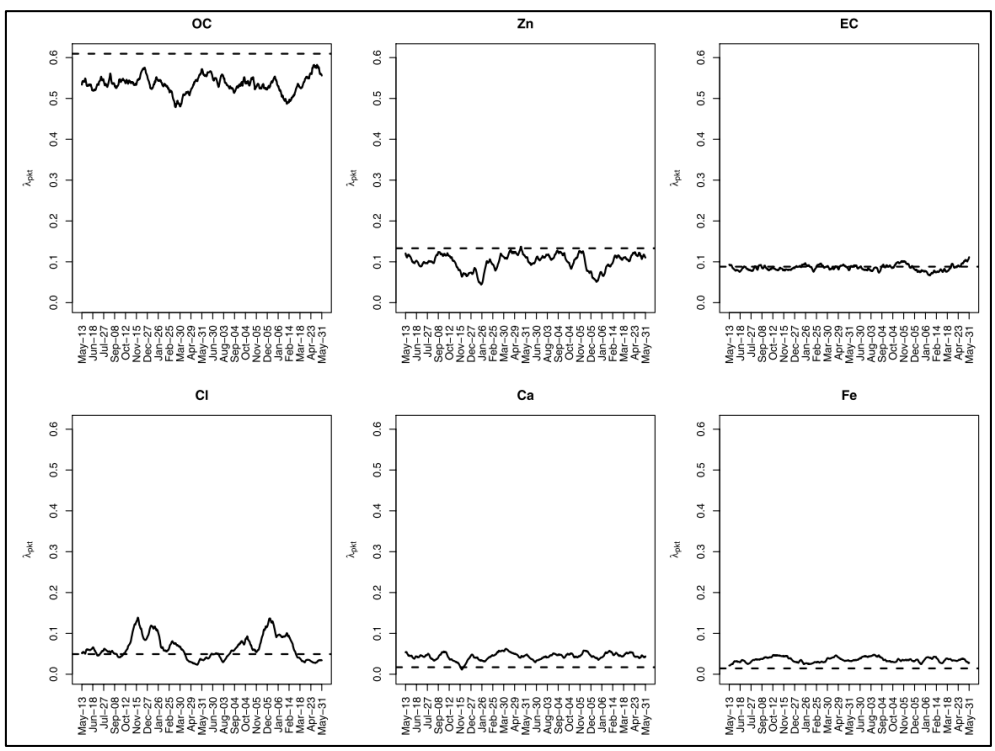


Fig. 2.22. Time plot of the six largest elements for the zinc smelter profile as identified by the Dirichlet process model (solid line). The dashed lines correspond to the time-constant PMF estimate.

Park et al. (2004) estimated the major source regions of PM_{10} using the data from 17 monitoring sites in Seoul using the Bayesian spatial receptor modeling.

Sixteen candidate models were considered and two models were selected as the best model based on the value of the estimated marginal likelihood (Park et al., 2004). Fig. 2.23 shows the result of the Bayesian receptor modeling in winter in Seoul (Park et al., 2004).

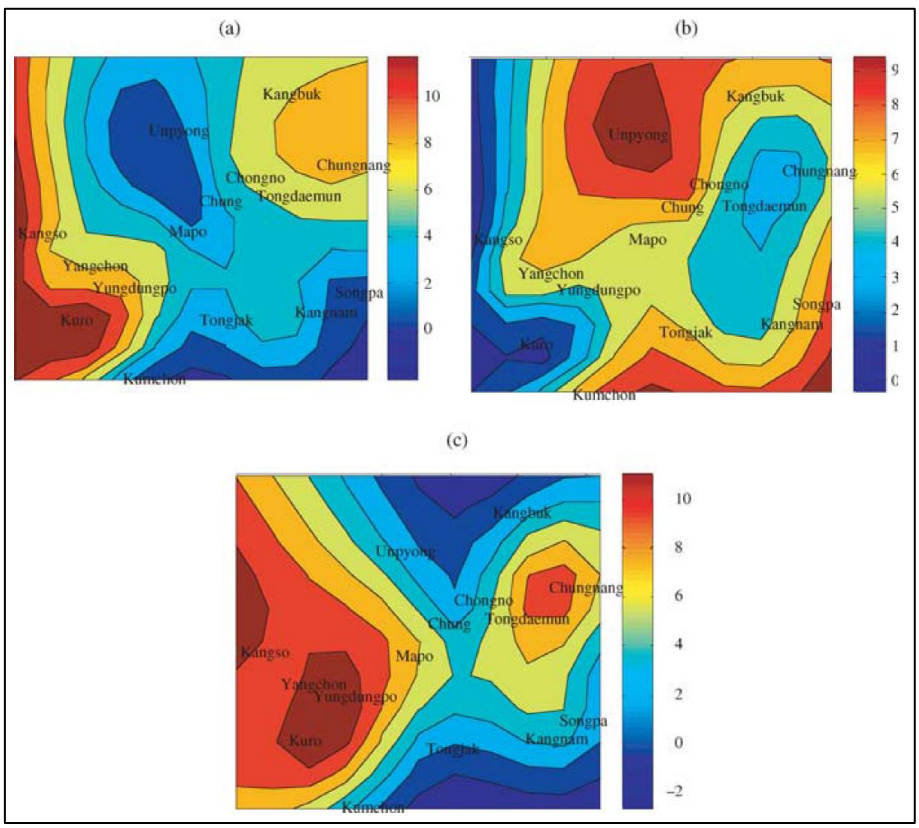


Fig. 2.23. Spatial profiles for (a) Source 1, (b) Source 2, and (c) Source 3 in Winter. The first letter of each site name corresponds to the actual location of the monitoring station, and "-gu" is omitted from the site name for the space (Park et al., 2004)

Although the Bayesian approach to air pollution is an emerging field of research with many advantages, there are not many applications because of the difficulty for investigators to start (Hopke, 2016). Therefore, additional studies are required to understand Bayesian methods in receptor modeling and air pollution phenomena.

References

Alpaydin, E., 2020. Introduction to machine learning. MIT press.

Andrews, J., Gorell, S., 2020. Generating Missing Unconventional Oilfield Data using a Generative Adversarial Imputation Network (GAIN). OnePetro. https://doi.org/10.15530/urtec-2020-3014

Atkinson, R.W., Kang, S., Anderson, H.R., Mills, I.C., Walton, H.A., 2014. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax 69, 660–665. https://doi.org/10.1136/thoraxjnl-2013-204492

Baerenbold, O., Meis, M., Martínez-Hernández, I., Euán, C., Burr, W.S., Tremper, A., Fuller, G., Pirani, M., Blangiardo, M., 2022. A dependent Bayesian Dirichlet process model for source apportionment of particle number size distribution. Environmetrics 1–18. https://doi.org/10.1002/env.2763

Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2012.11.009

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/S11749-016-0481-7/FIGURES/4

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Briffa, J., Sinagra, E., Blundell, R., 2020. Heavy metal pollution in the environment and their toxicological effects on humans. Heliyon 6, e04691. https://doi.org/10.1016/j.heliyon.2020.e04691

Brown, S.G., Eberly, S., Paatero, P., Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Sci. Total Environ. 518–519, 626–635. https://doi.org/10.1016/j.scitotenv.2015.01.022

Castelli, M., Clemente, F.M., Popovič, A., Silva, S., Vanneschi, L., 2020. A Machine Learning Approach to Predict Air Quality in California. Complexity 2020. https://doi.org/10.1155/2020/8049504

Chang, Y.S., Chiao, H.T., Abimannan, S., Huang, Y.P., Tsai, Y.T., Lin, K.M., 2020. An LSTM-based aggregated model for air pollution forecasting. Atmos. Pollut. Res. 11, 1451–1463. https://doi.org/10.1016/J.APR.2020.05.015

Choi, E., Yi, S.M., Lee, Y.S., Jo, H., Baek, S.O., Heo, J.B., 2022. Sources of airborne particulate matter-bound metals and spatial-seasonal variability of health risk potentials in four large cities, South Korea. Environ. Sci. Pollut. Res. 29, 28359–28374. https://doi.org/10.1007/s11356-021-18445-8

Choi, J. kyu, Heo, J.B., Ban, S.J., Yi, S.M., Zoh, K.D., 2013. Source apportionment of PM2.5 at the coastal area in Korea. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2012.12.047

Cohen, D.D., Crawford, J., Stelcer, E., Bac, V.T., 2010. Characterisation and source apportionment of fine particulate sources at Hanoi from 2001 to 2008. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2009.10.037

Costa, L.G., Cole, T.B., Coburn, J., Chang, Y.C., Dao, K., Roqué, P.J., 2017. Neurotoxicity of traffic-related air pollution. Neurotoxicology 59, 133–139. https://doi.org/10.1016/J.NEURO.2015.11.008

Crüts, B., van Etten, L., Törnqvist, H., Blomberg, A., Sandström, T., Mills, N.L.,

Borm, P.J., 2008. Exposure to diesel exhaust induces changes in EEG in human volunteers. Part. Fibre Toxicol. 5, 1–6. https://doi.org/10.1186/1743-8977-5-4/FIGURES/3

Dai, Q., Ding, J., Song, C., Liu, B., Bi, X., Wu, J., Zhang, Y., Feng, Y., Hopke, P.K., 2021. Changes in source contributions to particle number concentrations after the COVID-19 outbreak: Insights from a dispersion normalized PMF. Sci. Total Environ. 759. https://doi.org/10.1016/j.scitotenv.2020.143548

Dai, Q., Liu, B., Bi, X., Wu, J., Liang, D., Zhang, Y., Feng, Y., Hopke, P.K., 2020. Dispersion normalized PMF provides insights into the significant changes in source contributions to PM2.5 after the CoviD-19 outbreak. Environ. Sci. Technol. 54, 9917–9927. https://doi.org/10.1021/acs.est.0c02776

Dai, Q.L., Bi, X.H., Wu, J.H., Zhang, Y.F., Wang, J., Xu, H., Yao, L., Jiao, L., Feng, Y.C., 2015. Characterization and source identification of heavy metals in ambient PM10 and PM2.5 in an integrated Iron and Steel industry zone compared with a background site. Aerosol Air Qual. Res. 15, 875–887. https://doi.org/10.4209/aaqr.2014.09.0226

Du, X., Yang, J., Xiao, Z., Tian, Y., Chen, K., Feng, Y., 2021. Source apportionment of PM2.5 during different haze episodes by PMF and random forest method based on hourly measured atmospheric pollutant. Environ. Sci. Pollut. Res. 2021 1–12. https://doi.org/10.1007/S11356-021-14487-0

Fan, M.Y., Zhang, Y.L., Lin, Y.C., Cao, F., Sun, Y., Qiu, Y., Xing, G., Dao, X., Fu, P., 2021. Specific sources of health risks induced by metallic elements in PM2.5 during the wintertime in Beijing, China. Atmos. Environ. 246, 118112. https://doi.org/10.1016/j.atmosenv.2020.118112

Hackstadt, A.J., Peng, R.D., 2014. A Bayesian multivariate receptor model for estimating source contributions to particulate matter pollution using national databases. Environmetrics 25, 513–527. https://doi.org/10.1002/ENV.2296

Hadeed, S.J., O'Rourke, M.K., Burgess, J.L., Harris, R.B., Canales, R.A., 2020. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Sci. Total Environ. 730, 139140. https://doi.org/10.1016/J.SCITOTENV.2020.139140

Hamanaka, R.B., Mutlu, G.M., 2018. Particulate Matter Air Pollution: Effects on the Cardiovascular System. Front. Endocrinol. (Lausanne). 9. https://doi.org/10.3389/fendo.2018.00680

Hamra, G.B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J.M., Vineis, P., Forastiere, F., Saldiva, P., Yorifuji, T., Loomis, D., 2014. Outdoor Particulate Matter Exposure and Lung Cancer: A Systematic Review and Meta-Analysis. Environ. Health Perspect. 122, 906–911. https://doi.org/10.1289/EHP/1408092

Han, F., Kota, S.H., Wang, Y., Zhang, H., 2017. Source apportionment of PM2.5 in Baton Rouge, Louisiana during 2009–2014. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2017.01.189

Heaton, M.J., Reese, C.S., Christensen, W.F., 2010. Incorporating Time-Dependent Source Profiles Using the Dirichlet Distribution in Multivariate Receptor Models. Technometrics 52, 67–79. https://doi.org/10.1198/TECH.2009.08134

Heo, J.-B., Hopke, P.K., Yi, S.-M., 2009. Source apportionment of $PM_{2.5}$ in Seoul, Korea. Atmos. Chem. Phys. 9, 4957–4971. https://doi.org/10.5194/acp-9-4957-2009

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science (80-.). 313, 504–507. https://doi.org/10.1126/SCIENCE.1127647

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. https://doi.org/10.1080/10962247.2016.1140693

Hopke, P.K., Dai, Q., Li, L., Feng, Y., 2020. Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.140091

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs.est.7b01210

Hu, X., Zhang, Y., Ding, Z., Wang, T., Lian, H., Sun, Y., Wu, J., 2012. Bioaccessibility and health risk of arsenic and heavy metals (Cd, Co, Cr, Cu, Ni, Pb, Zn and Mn) in TSP and PM2.5 in Nanjing, China. Atmos. Environ. 57, 146–152. https://doi.org/10.1016/j.atmosenv.2012.04.056

Hwang, I.J., Yi, S.M., Park, J., 2020. Estimation of Source Apportionment for Filter-based PM2.5 Data using the EPA-PMF Model at Air Pollution Monitoring Supersites. J. Korean Soc. Atmos. Environ. 36, 620–632. https://doi.org/10.5572/KOSAE.2020.36.5.620

Hwangbo, S., Al, R., Chen, X., Sin, G., 2021. Integrated Model for Understanding N2O Emissions from Wastewater Treatment Plants: A Deep Learning Approach. Environ. Sci. Technol. 55, 2143–2151. https://doi.org/10.1021/acs.est.0c05231

Ivanov, O., Figurnov, M., Vetrov, D., 2018. Variational Autoencoder with Arbitrary Conditioning. 7th Int. Conf. Learn. Represent. ICLR 2019.

Jeong, J.H., Shon, Z.H., Kang, M., Song, S.K., Kim, Y.K., Park, J., Kim, H., 2017. Comparison of source apportionment of PM2.5 using receptor models in the main hub port city of East Asia: Busan. Atmos. Environ. 148, 115–127. https://doi.org/10.1016/j.atmosenv.2016.10.055

Jeong, Y., Hwang, I., 2015. Source Apportionment of PM2.5 in Gyeongsan Using the PMF Model. J. Korean Soc. Atmos. Environ. 31, 508–519. https://doi.org/10.5572/kosae.2015.31.6.508

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science (80-.). https://doi.org/10.1126/science.aaa8415

Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., Amann, M., 2015. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087

Kelp, M.M., Jacob, D.J., Kutz, J.N., Marshall, J.D., Tessum, C.W., 2020. Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System. J. Geophys. Res. Atmos. 125, 1–13. https://doi.org/10.1029/2020JD032759

Khillare, P.S., Sarkar, S., 2012. Airborne inhalable metals in residential areas of Delhi, India: Distribution, source apportionment and health risks. Atmos. Pollut. Res. 3, 46–54. https://doi.org/10.5094/APR.2012.004

Kim, K.H., Kabir, E., Kabir, S., 2015. A review on the human health impact of

airborne particulate matter. Environ. Int. 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005

Kim, S., Kim, T.Y., Yi, S.M., Heo, J., 2018. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. J. Environ. Manage. 214, 325–334. https://doi.org/10.1016/j.jenvman.2018.03.027

Kim, S., Yang, J., Park, J., Song, I., Kim, D.G., Jeon, K., Kim, H., Yi, S.M., 2022. Health effects of PM2.5 constituents and source contributions in major metropolitan cities, South Korea. Environ. Sci. Pollut. Res. 1, 1–15. https://doi.org/10.1007/S11356-022-21592-1

Kumar, V., Sahu, M., Biswas, P., 2022. Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms. Aerosol Air Qual. Res. 22, 210240. https://doi.org/10.4209/AAQR.210240

Lee, Y.S., Kim, Y.K., Choi, E., Jo, H., Hyun, H., Yi, S.-M., Kim, J.Y., 2022. Health risk assessment and source apportionment of PM2.5-bound toxic elements in the industrial city of Siheung, Korea. Environ. Sci. Pollut. Res. 1, 1–14. https://doi.org/10.1007/s11356-022-20462-0

Li, H., Qian, X., Wang, Q., 2013. Heavy metals in atmospheric particulate matter: A comprehensive understanding is needed for monitoring and risk mitigation. Environ. Sci. Technol. 47, 13210–13211. https://doi.org/10.1021/es404751a

Li, S.C.X., Marlin, B.M., Jiang, B., 2019. Misgan: Learning from incomplete data with generative adversarial networks, in: 7th International Conference on Learning Representations, ICLR 2019. International Conference on Learning Representations, ICLR.

Lingwall, J.W., Christensen, W.F., Reese, C.S., 2008. Dirichlet based Bayesian

multivariate receptor modeling. Environmetrics 19, 618–629. https://doi.org/10.1002/env.902

Lv, L., Chen, Y., Han, Y., Cui, M., Wei, P., Zheng, M., Hu, J., 2021. High-time-resolution PM2.5 source apportionment based on multi-model with organic tracers in Beijing during haze episodes. Sci. Total Environ. 772, 144766. https://doi.org/10.1016/j.scitotenv.2020.144766

Lyu, B., Hu, Y., Zhang, W., Du, Y., Luo, B., Sun, X., Sun, Z., Deng, Z., Wang, Xiaojiang, Liu, J., Wang, Xuesong, Russell, A.G., 2019. Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved PM2.5 Exposure Fields in 2014-2017. Environ. Sci. Technol. 53, 7306–7315. https://doi.org/10.1021/acs.est.9b01117

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and Health Impacts of Air Pollution: A Review. Front. public Heal. 8, 14. https://doi.org/10.3389/FPUBH.2020.00014

Montavon, G., Samek, W., Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. A Rev. J. 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nazábal, A., Olmos, P.M., Ghahramani, Z., Valera, I., 2020. Handling incomplete heterogeneous data using VAEs. Pattern Recognit. 107. https://doi.org/10.1016/j.patcog.2020.107501

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemom. Intell. Lab. Syst. 37, 23–35. https://doi.org/10.1016/S0169-

7439(96)00044-5

Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126. https://doi.org/10.1002/env.3170050203

Pant, P., Harrison, R.M., 2012. Critical review of receptor modelling for particulate matter: A case study of India. Atmos. Environ. 49, 1–12. https://doi.org/10.1016/J.ATMOSENV.2011.11.060

Park, M. Bin, Lee, T.J., Lee, E.S., Kim, D.S., 2019. Enhancing source identification of hourly PM2.5 data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). Atmos. Pollut. Res. 10, 1042–1059. https://doi.org/10.1016/j.apr.2019.01.013

Park, E.H., Heo, J., Kim, H., Yi, S.-M., 2020. Long term trends of chemical constituents and source contributions of PM2.5 in Seoul. Chemosphere 251, 126371. https://doi.org/10.1016/j.chemosphere.2020.126371

Park, E.S., Guttorp, P., Henry, R.C., 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. J. Am. Stat. Assoc. 96, 1171–1183. https://doi.org/10.1198/016214501753381823

Park, E.S., Guttorp, P., Kim, H., 2004. Locating major PM 10 source areas in Seoul using multivariate receptor modeling. Environ. Ecol. Stat. 11, 9–19. https://doi.org/10.1023/B:EEST.0000011361.33942.be

Park, E.S., Hopke, P.K., Kim, I., Tan, S., Spiegelman, C.H., 2018. Bayesian Spatial Multivariate Receptor Modeling for Multisite Multipollutant Data. Technometrics 60, 306–318. https://doi.org/10.1080/00401706.2017.1366948

Park, E.S., Lee, E.K., Oh, M.S., 2021. Bayesian multivariate receptor modeling

software: BNFA and bayesMRM. Chemom. Intell. Lab. Syst. 211, 104280. https://doi.org/10.1016/j.chemolab.2021.104280

Park, E. S.; Oh, M. S. Accounting for Uncertainty in Source-Specific Exposures in the Evaluation of Health Effects of Pollution Sources on Daily Cause-Specific Mortality. Environmetrics 2018, 29 (1), 2484

Park, E. S.; Oh, M. S. Bayesian Quantile Multivariate Receptor Modeling. Chemom. Intell. Lab. Syst. 2016, 159, 174–180

Park, E.S., Oh, M.S., 2015. Robust Bayesian multivariate receptor modeling. Chemom. Intell. Lab. Syst. 149, 215–226. https://doi.org/10.1016/j.chemolab.2015.08.021

Park, E. S.; Symanski, E.; Han, D.; Spiegelman, C. H. Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated With an Unknown Number of Major Sources of Multiple Air Pollutants. In Development of Statistical Methods for Multipollutant Research, Health Effects Institute: Boston, MA, 2015. Research Report 183

Park, E. S.; Hopke, P. K.; Oh, M. S.; Han, D.; Symanski, E.; Spiegelman, C. H. Assessment of Source Specific Health Effects Associated With an Unknown Number of Major Sources of Multiple Air Pollutants: A Unified Bayesian Approach. Biostatistics 2014, 15, 484–497

Park, E.S., Oh, M.S., Guttorp, P., 2002. Multivariate receptor models and model uncertainty. Chemom. Intell. Lab. Syst. 60, 49–67. https://doi.org/10.1016/S0169-7439(01)00185-X

Park, E.S., Tauler, R., 2020. Bayesian Methods for Factor Analysis in Chemometrics. Compr. Chemom. 355–369. https://doi.org/10.1016/B978-0-12-409547-

2.14876-0

Park, M.H., Ju, M., Kim, J.Y., 2020. Bayesian approach in estimating flood waste generation: A case study in South Korea. J. Environ. Manage. 265, 110552. https://doi.org/10.1016/J.JENVMAN.2020.110552

Polissar, A. V., Hopke, P.K., Harris, J.M., 2001. Source regions for atmospheric aerosol measured at Barrow, Alaska. Environ. Sci. Technol. 35, 4214–4226. https://doi.org/10.1021/es0107529

Popolizio, M., Amato, A., Liquori, F., Politi, T., Quarto, A., Lecce, V. Di, 2021. The GAIN Method for the Completion of Multidimensional Numerical Series of Meteorological Data. IAENG Int. J. Comput. Sci. 48, 1–11.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1

Sakunkoo, P., Thonglua, T., Sangkham, S., Jirapornkul, C., Limmongkon, Y., Daduang, S., Tessiri, T., Rayubkul, J., Thongtip, S., Maneenin, N., Pimonsree, S., 2022. Human health risk assessment of PM2.5-bound heavy metal of anthropogenic sources in the Khon Kaen Province of Northeast Thailand. Heliyon 8, e09572. https://doi.org/10.1016/J.HELIYON.2022.E09572

Samara, C., Kouimtzis, T., Tsitouridou, R., Kanias, G., Simeonov, V., 2003. Chemical mass balance source apportionment of PM10 in an industrialized urban area of Northern Greece. Atmos. Environ. 37, 41–54. https://doi.org/10.1016/S1352-2310(02)00772-0

Silva, A.V., Oliveira, C.M., Canha, N., Miranda, A.I., Almeida, S.M., 2020. Long-

Term Assessment of Air Quality and Identification of Aerosol Sources at Setúbal, Portugal. Int. J. Environ. Res. Public Health 17, 1–23. https://doi.org/10.3390/IJERPH17155447

Tang, J.H., Candice Lung, S.C., Hwang, J.S., 2020. Source apportionment of PM2.5 concentrations with a Bayesian hierarchical model on latent source profiles. Atmos. Pollut. Res. 11, 1715–1727. https://doi.org/10.1016/J.APR.2020.06.013

Tella, A., Balogun, A.L., Adebisi, N., Abdullah, S., 2021. Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. Atmos. Pollut. Res. 12, 101202. https://doi.org/10.1016/J.APR.2021.101202

Thangavel, P., Park, D., Lee, Y.C., 2022. Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview. Int. J. Environ. Res. Public Heal. 2022, Vol. 19, Page 7511 19, 7511. https://doi.org/10.3390/IJERPH19127511

Tremper, A.H., Jephcote, C., Gulliver, J., Hibbs, L., Green, D.C., Font, A., Priestman, M., Hansell, A.L., Fuller, G.W., 2022. Sources of particle number concentration and noise near London Gatwick Airport. Environ. Int. 161, 107092. https://doi.org/10.1016/J.ENVINT.2022.107092

Turner, M.C., Krewski, D., Ryan Diver, W., Arden Pope, C., Burnett, R.T., Jerrett, M., Marshall, J.D., Gapstur, S.M., 2017. Ambient air pollution and cancer mortality in the cancer prevention study II. Environ. Health Perspect. 125. https://doi.org/10.1289/EHP1249

US EPA, 2009. Risk Assessment Guidance for Superfund Volume I: Human Health Evaluation Manual (Part F, Supplemental Guidance for Inhalation Risk

Assessment). Off. Superfund Remediat. Technol. Innov. Environ. Prot. Agency I, 1–68.

Viñas, R., Azevedo, T., Gamazon, E.R., Liò, P., 2020. Gene expression imputation with Generative Adversarial Imputation Nets. bioRxiv. https://doi.org/10.1101/2020.06.09.141689

Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., Yu, J.Z., 2018. Source Apportionment of PM2.5 Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-Resolution Influence. J. Geophys. Res. Atmos. 123, 5284–5300. https://doi.org/10.1029/2017JD027877

WHO, 2005. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005 1–21. https://doi.org/10.1016/0004-6981(88)90109-6

Widziewicz, K., Rogula-Kozłowska, W., Loska, K., 2016. Cancer risk from arsenic and chromium species bound to PM2.5 and PM1 – Polish case study. Atmos. Pollut. Res. 7, 884–894. https://doi.org/10.1016/J.APR.2016.05.002

Wu, S., Ni, Y., Li, H., Pan, L., Yang, D., Baccarelli, A.A., Deng, F., Chen, Y., Shima, M., Guo, X., 2016. Short-term exposure to high ambient air pollution increases airway inflammation and respiratory symptoms in chronic obstructive pulmonary disease patients in Beijing, China. Environ. Int. 94, 76–82. https://doi.org/10.1016/J.ENVINT.2016.05.004

Wu, X., Vu, T. V., Shi, Z., Harrison, R.M., Liu, D., Cen, K., 2018. Characterization and source apportionment of carbonaceous PM2.5 particles in China - A review. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2018.06.025

Yang, L., Cheng, S., Wang, X., Nie, W., Xu, P., Gao, X., Yuan, C., Wang, W., 2013. Source identification and health impact of PM2.5 in a heavily polluted urban atmosphere in China. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2013.04.058

Yao, Z., Ruzzo, W.L., 2006. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7, S11. https://doi.org/10.1186/1471-2105-7-S1-S11

Yoon, J., Jordon, J., Van Der Schaar, M., 2018. Supplementary materials GAIN: Missing data imputation using generative adversarial nets, in: 35th International Conference on Machine Learning, ICML 2018. PMLR, pp. 9052–9059.

Zhao, X., Liu, Y., Han, F., Touseef, B., Yue, Y., Guo, J., 2021. Source profile and health risk assessment of PM2.5 from coal-fired power plants in Fuxin, China. Environ. Sci. Pollut. Res. 28, 40151–40159. https://doi.org/10.1007/s11356-020-11378-8

Zhao, Z., Lv, S., Zhang, Y., Zhao, Q., Shen, L., Xu, S., Yu, J., Hou, J., Jin, C., 2019. Characteristics and source apportionment of PM2.5 in Jiaxing, China. Environ. Sci. Pollut. Res. 2019 268 26, 7497–7511. https://doi.org/10.1007/S11356-019-04205-2

Zheng, H., Kong, S., Yan, Q., Wu, F., Cheng, Y., Zheng, S., Wu, J., Yang, G., Zheng, M., Tang, L., Yin, Y., Chen, K., Zhao, T., Liu, D., Li, S., Qi, S., Zhao, D., Zhang, T., Ruan, J., Huang, M., 2019. The impacts of pollution control measures on PM2.5 reduction: Insights of chemical composition, source variation and health risk. Atmos. Environ. 197, 103–117.

https://doi.org/10.1016/j.atmosenv.2018.10.023

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.J., Zhang, H., 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environ. Sci. Technol. 55, 12741–12754. https://doi.org/10.1021/acs.est.1c01339

Zong, Z., Wang, X., Tian, C., Chen, Y., Qu, L., Ji, L., Zhi, G., Li, J., Zhang, G., 2016. Source apportionment of PM 2.5 at a regional background site in North China using PMF linked with radiocarbon analysis: insight into the contribution of biomass burning. Atmos. Chem. Phys 16, 11249–11265. https://doi.org/10.5194/acp-16-11249-2016

Chapter 3. Source apportionment of $PM_{2.5}$ using PMF model and health risk assessment by inhalation[1]

3.1. Introduction

Fine particulate matter ($PM_{2.5}$) in the atmosphere is classified as a Group 1 carcinogen by the World Health Organization (WHO) owing to its carcinogenicity to humans (Anderson, 2009; WHO, 2005). In many countries, $PM_{2.5}$ concentration is used as a major indicator of air quality, and significant efforts have been made to reduce $PM_{2.5}$ pollution (Nazarenko et al., 2021; Riojas-Rodríguez et al., 2016). For a proper $PM_{2.5}$ management, pollution sources should be accurately managed by determining the relationship between the source characteristics and atmospheric concentrations (Fang et al., 2020; Kim et al., 2019; Long et al., 2021). However, when $PM_{2.5}$ is released into the atmosphere, it immediately goes through complex mechanisms such as advection, diffusion, reaction, and deposition; therefore, it is difficult to identify its source (Anderson, 2009; Riojas-Rodríguez et al., 2016). Thus, to effectively clarify the mechanisms and characteristics of $PM_{2.5}$ pollution and improve air quality, scientific methods should be applied to identify and quantify $PM_{2.5}$ sources (Belis et al., 2013; Hopke, 2016; Wang et al., 2012). In addition, as the impacts on human health vary according to $PM_{2.5}$ source, management priorities should be defined based on the evaluation of health impacts and source

[1] A significant portion of this chapter was published in the following article: Lee, Y.S., Kim, Y.K., Choi, E., Jo, H., Hyun, H., Yi, S.-M., Kim, J.Y., 2022. Health risk assessment and source apportionment of PM2.5-bound toxic elements in the industrial city of Siheung, Korea. Environ. Sci. Pollut. Res. 1, 1–14. https://doi.org/10.1007/s11356-022-20462-0.

apportionment (Kim et al., 2015; Yang et al., 2013).

The health risk assessment coupled with source apportionment can be used to develop more specific environmental health policies because the health risks due to exposure to $PM_{2.5}$ may vary depending on the emission source. (Kim et al., 2019; Leogrande et al., 2019; Wang et al., 2020; Yang et al., 2013; Zhang et al., 2020). It is shown that oxidative potentials per PM mass differs greatly depending on the emission sources such as vehicle exhaust and secondary aerosols (Shiraiwa et al., 2017). Accordingly, health risk assessments by sources were considered essential for comprehensive understanding behavior of particulate matter (PM) (Choi et al., 2022; Fan et al., 2021; Li et al., 2013). Also, although the importance of evaluation of ambient PM that takes into consideration size, chemical composition, and source of particles has been pointed out (Cassee et al., 2013), those factors have rarely been involved in the health or toxicity assessment (Fushimi et al., 2021; Hannigan et al., 2005; Kim et al., 2020). Recent relevant studies investigate specific sources and chemical components of air pollution that affect human health and compared the assessment results to those of other regions, but these studies are still lacking (Fan et al., 2021). Furthermore, some studies show that health effects are still indicated in developed countries with low $PM_{2.5}$ concentrations, it is still necessary to study on which pollutants and how they affect human health (Ma et al., 2022; Thurston et al., 2021; Christidis et al., 2019).

To date, far too little attention has been paid to conduct both source apportionment and health risk assessment simultaneously in middle-sized industrial cities that could exist in any country in the world, and rather, only some large cities are being studied (Fu et al., 2021; Hu et al., 2012; Yang et al., 2013). Air pollution is

generally more severe in industrial areas, owing to local industrial emissions (Fu et al., 2021; Shende and Qureshi, 2022). The negative impact to human health in these areas are expected to be greater than those to humans in areas with less pollution because of the presence of pollutants such as heavy metals, organic carbon (OC), or elemental carbon (EC) (Kumar et al., 2020; Samara et al., 2003). Therefore, the method source apportionment integrated with health risk assessment needs to be applied as a basis for the development of air pollution management policies, especially in industrial areas.

The main purpose of this study was to identify the sources of $PM_{2.5}$ and to evaluate the health risk of each source type in Siheung, which is a city with national industrial complexes located in the Republic of Korea. The specific aims of this study were to (1) identify and apportion $PM_{2.5}$ sources with error estimation, (2) assess health risks of $PM_{2.5}$ inhalation and the contribution of each source to these health risks from heavy metals in $PM_{2.5}$, and (3) identify the characteristics of the sources that represent higher health risks and explore appropriate $PM_{2.5}$ reduction measures based on a source-based health risk assessment. The target area of this study is a medium-sized industrial city, which is similar to many other industrial cities worldwide.

3.2. Materials and methods

3.2.1 Study site, sampling, and analysis

Siheung City is located at approximately 20 km southwest of Seoul, Republic of Korea, and it has a population of approximately 0.56 million (as of 2021). In the southwest of Siheung City, 10,000 factories are located in a national industrial complex, with an area of approximately 165 million m^2 (Siheung City's official website, https://www.siheung.go.kr/english/, last access: 10 August 2021). The main industrial fields include textiles, chemicals, metal smelting, printing, and paper, Siheung City has high accessibility to Seoul owing to the highways and nearby ports; therefore, industrial activities are prominent in that area. It shares city-regional characteristics with medium-sized industrial cities in other major countries worldwide. Fig. 3.1 illustrates the location of Siheung City and its industrial complexes. The daily average $PM_{2.5}$ concentrations in Siheung City were compared with those of other industrial cities in Korea, China, and Germany.

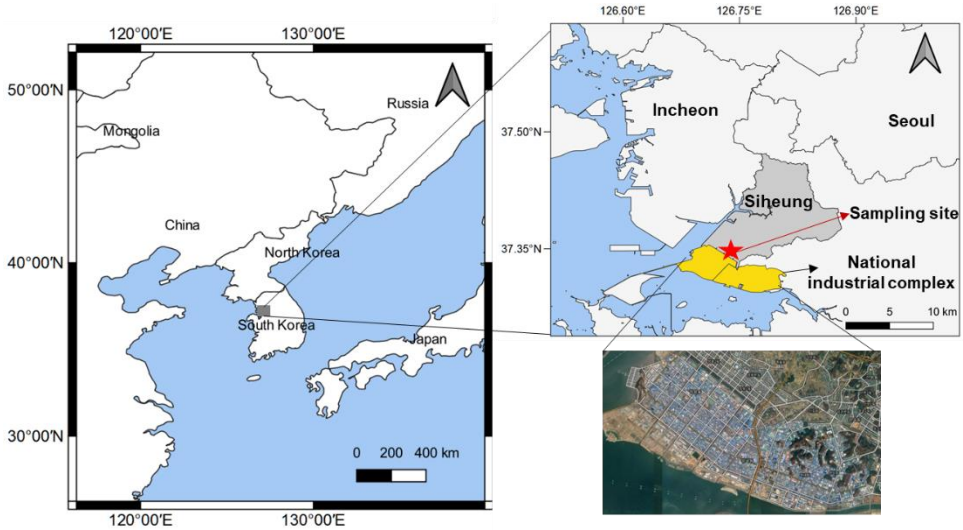


Fig. 3.1. Locations of this study site (Siheung city and sampling site)

Fig. 3.2 shows the $PM_{2.5}$ concentration levels of industrial cities in China and Germany (Beijing, Shanghai, Hamburg, Kassel), in Korea (Ulsan, Yeosu, Incheon, and Daebudo), and Seoul, the capital city of Korea. For the data, the air quality index value obtained from the Air Quality Historical Data Platform (https://aqicn.org/, last access: 10 August 2021) was converted into mass concentration.

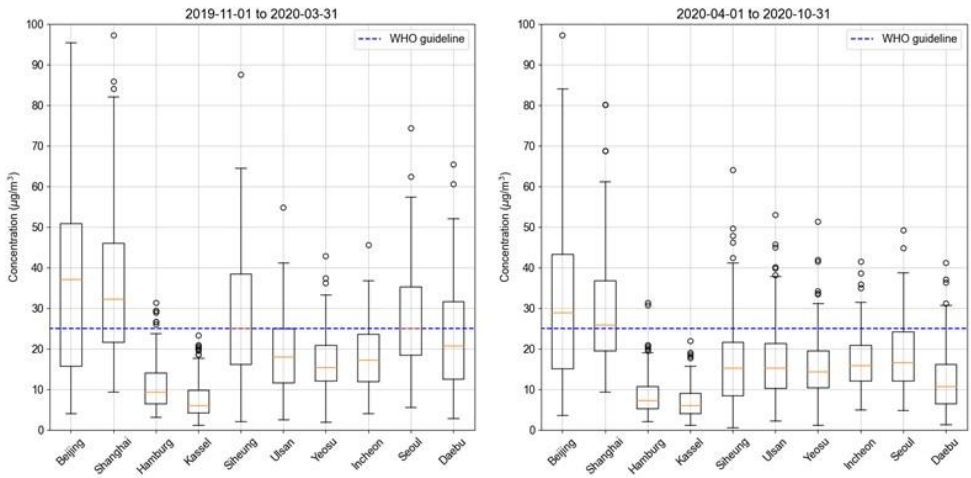


Fig. 3.2. Average daily $PM_{2.5}$ concentration comparisons between the sampling site and other sites

To quantify the chemical composition of $PM_{2.5}$ samples were collected every three or four times a week over 24 h from November 2019 to December 2020 at the rooftop of Jeongwnag-dong National Air Quality Measuring Station (37.3472°N, 126.7399°E, shown as a red star in Fig. 3.1), which is approximately 10 m above the ground level. A $PM_{2.5}$ sampler (PMS-204, APM Engineering, South Korea) with three parallel channels was used to collect $PM_{2.5}$ samples. Two channels were installed with Teflon filters (2 μm pore size and 47 mm diameter, Measurement Technology Laboratories, USA) and one channel with a quartz filter (47mm diameter, Pall Corporation, USA). Each sampler was operated for 24 h at a 16.67 L/min flow rate. The mass concentration, ionic component, OC, EC, and elemental components of $PM_{2.5}$ were analyzed as follows. The mass concentration was calculated by measuring the weight of a 24 h dried Teflon filter (PT47P, MTL, US) before and after sample collection, and then dividing the obtained value by the collected air volume. The weight of the filters was measured after removing static electricity at a constant

temperature (21±1.5°C) and humidity (35±5%). Moreover, the weight of the blank filter was measured and used for correction. Ion component analysis was performed by ion chromatography (930 Compact IC Flex, Metrohm, Switzerland) using a Teflon filter (TF-10000, PALL, USA). In the analysis, each of the entire sampled filter was extracted for 120 min in a bath-type sonicator using 40 ml of distilled water, and then filtered using a 0.45 μm membrane. For OC and EC, a quartz fiber filter paper (7407, PALL, USA) cut to a diameter of 4 mm in the sampled portion was used, and the analysis was performed using the thermal optical transmittance (TOT) method in a carbon analyzer (laboratory OC-EC aerosol analyzer, Sunset Lab, USA), and the analysis conditions followed the NIOSH 5040 protocol. The trace elements were analyzed by energy dispersive X-ray fluorescence (ED-XRF) spectroscopy (ARL QUANT'X ED XRF Spectrometer, Thermo Fisher Scientific, USA) using Teflon filters (PT47P, MTL, US) without additional pretreatment. Namely, each of the entire sampled filter was used in the measurement. A total of 29 components were analyzed. Including the mass concentration analysis, 6 ionic species (NO_3^-, SO_4^{2-}, NH_4^+, K^+, Na^+, and Cl^-), carbons (OC and EC), and 21 species of elemental components (Na, Mg, Al, Si, S, Cl, K, Ca, Ti, V, Cr, Mn, Ba, Fe, Ni, Cu, Zn, As, Se, Br, and Pb) were quantified.

3.2.2 Positive matrix factorization (PMF) modeling and combined analysis with meteorological data

The PMF model has been widely used as a method of factor analysis to derive air pollution sources from speciated sample data (Hopke, 2016; Paatero, 1997; Paatero and Tapper, 1994). The data matrix can be separated into factor contributions (G) and factor profiles (F) (United States Environmental Protection Agency (US EPA) 2014). The equation for the PMF model is given by (Paatero and Tapper 1994).

$$X = G \times F + E \qquad \text{Eq. 3.1}$$

where X is a matrix of the sample dataset (e.g., n × j matrix, where n is the sampled date and j is the chemical species of the data), G is the source contribution matrix (e.g., n × q matrix, where q is the source contribution), F is the source profile matrix (e.g., q × j matrix), and E is a residual matrix (e.g., n × j matrix).

In Eq. 3.1, all elements of matrices G and F are constrained to positive values. To derive the appropriate G and F matrices, the objective function Q in Eq. 3.2 was minimized (Paatero, 1997).

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 \qquad \text{Eq. 3.2}$$

where n is the number of samples, m is the number of species, e_{ij} is the residual (e.g., element of matrix E), and σ_{ij} is the data uncertainty (e.g., uncertainty of chemical species j at date i).

The US EPA PMF version 5.0.14 was used to estimate the source contribution and profile in the target area. The concentration data for the modeling included the pre-processed chemical composition analysis of 22 substances (NO_3^-, SO_4^{2-}, NH_4^+, K^+, Na^+, Cl^-, OC, EC, Mg, Al, Si, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, and Pb) and $PM_{2.5}$ mass concentration. The pretreatment process considered the ratio of cations and anions in $PM_{2.5}$, and data were excluded if concentrations were below the detection limit or when an outlier was detected. If there were duplicate measurements, one was selected for use. Data with an S/N ratio of 0.2 or less were also removed. This method is an established procedure reported in previous studies (Choi et al., 2013; Kim et al., 2018; E. H. Park et al., 2020). The data uncertainty was calculated using Eq. 3.3, according to the US EPA guidelines (US -EPA 2014).

$$\sigma_{ij} = \begin{cases} (5/6) \times \mathrm{MDL} & (\text{if Conc.} \leq \mathrm{MDL}) \\ \sqrt{(Conc. \times 0.1)^2 + (0.5 \times \mathrm{MDL})^2} & (\text{if Conc.} > \mathrm{MDL}) \end{cases} \quad \text{Eq. 3.3}$$

where MDL is the method detection limit and Conc. is the concentration ($\mu g/m^3$) of the species, (e.g., X_{ij}). MDL values of the elemental components are listed in Table 3.1.

Table 3.1. Method detection limit (MDL) values of the elemental components (unit: ng m^{-3})

Al	Si	Ca	Ti	V	Cr	Mn	Fe	Ni	Cu	Zn	As	Pb
6.69	5.54	4.39	3.72	0.201	0.726	0.969	7.04	0.609	0.242	1.22	1.42	3.19

The data used for the modeling included 95 daily average values. The number of sources (e.g., q) in the model was selected by repeated modeling. Moreover, BS and DISP analyses in the US EPA PMF 5.0 were conducted to confirm the appropriate range of major chemical species by source. These functions are widely used to investigate errors and rotational ambiguity (Dai et al., 2020b). PMF results of 8 to 10 factors were considered for the best solution.

The CPF analysis was applied to investigate source directionality and the PSCF analysis was applied to locate possible source areas. The hybrid single-particle Lagrangian integrated trajectory (HYSPLIT 5) model and gridded meteorological data from the US National Oceanic and Atmospheric Administration were used to calculate air parcel backward trajectories.

The conditional probability function (CPF) enable to analyze the changes in $PM_{2.5}$ concentrations for each source according to wind direction and speed (Carslaw, 2015).. The CPF is defined as CPF = m_θ/n_θ, where m_θ represents the samples above a certain concentration in the wind direction θ, and n_θ is the total numbers of samples in the same wind direction. CPF values were visualized using hourly wind direction and speed data combined with PMF source contributions using the OpenAir package in R (version 4.0.3, Vienna, Austria). Meteorological data were obtained from the weather station located at the same position as the sampling site (37°20'48"N 126°44'24"E) and operated by the Korea Meteorological Administration (data are available at https://data.kma.go.kr/, last access: 10 August 2021). The upper 25% of PMF source contributions was used as the threshold criteria.

Subsequently, backward trajectory analysis was conducted using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model. The

transboundary airmass transport pathways from the sampling site were predicted. According to the sampling date, 24 h and 72 h of back trajectories were analyzed in 1 h increments. The possible past routes were tracked using the Global Data Assimilation System (GDAS) 1-degree meteorological data. The HYSPLIT version 5.0 and PySPLIT, which is a Python-compatible package (Warner, 2018), were used. The potential source contribution function (PSCF) was calculated based on the results of the backward trajectory analysis. The PSCF model indicates the conditional probability of air coming from an area (Ashbaugh et al., 1985) and is represented by Eq. 3.4.

$$\text{PSCF} = m_{ij}/n_{ij} \qquad \text{Eq. 3.4}$$

where m_{ij} is the total number of trajectory endpoints that exceed the threshold concentration in the i, jth grid cell; and n_{ij} is the total number of trajectory endpoints that pass the i, jth grid cell. In this study, the threshold concentration for m_{ij} was in the 70th percentile.

The weighted PSCF (WPSCF) value can lead to more reliable results because the PSCF value can have high uncertainty in some cases (Polissar et al., 2001). Therefore, the WPSCF was calculated using Eq. 3.5. In addition, visualization was performed using WPSCF(n_{ij}) at each grid and interpolated by Kriging. The results and discussion of the combined analysis with meteorological data is also provided.

$$WPSCF(n_{ij}) = \begin{cases} 1.0 \times PSCF(n_{ij}) & (n_{ij} > 3n_{avg}) \\ 0.7 \times PSCF(n_{ij}) & (3n_{avg} > n_{ij} > 1.5n_{avg}) \\ 0.4 \times PSCF(n_{ij}) & (1.5n_{avg} > n_{ij} > n_{avg}) \\ 0.2 \times PSCF(n_{ij}) & (n_{avg} > n_{ij}) \end{cases} \quad \text{Eq. 3.5}$$

3.2.3 Health risk assessment

Using the species concentration for each source obtained through PMF modeling, the health risk was calculated following the guidelines established by the US EPA (2013, 2009). We evaluated only the substances with toxicity values, similar to previous studies on health risks of air pollution (Choi et al., 2011a; Fu et al., 2021; Hu et al., 2012; Yang et al., 2013; Zhao et al., 2021). Therefore, the health risk results of this study did not reflect the ion components, OC, EC, and $PM_{2.5}$ itself. The health risk was assessed only for toxic elements in $PM_{2.5}$.

As inhalation is the predominant pathway for human exposure to $PM_{2.5}$ bound toxic elements, we considered only the inhalation pathway for carcinogenic (As, Cr, Ni, and Pb) and non-carcinogenic (As, Cr, Cu, Ni, Pb, V, and Mn) risk estimations. For Cr, because its hexavalent and trivalent forms generate different levels of health impacts, the ratio of hexavalent to trivalent was set to 3:7 by referring to the abundance ratio in the PM of other industrial cities (Torkmahalleh et al., 2013; Widziewicz et al., 2016).

The average daily dose of $PM_{2.5}$ bound trace elements via inhalation (ADD_{inh}) was calculated using Eq. 3.6 (US EPA, 2009).

$$ADD_{inh}\ (\mu g/m^3) = \frac{C \times ET \times EF \times ED}{AT} \quad \text{Eq. 3.6}$$

where C represents the mean concentration of a pollutant in the air ($\mu g/m^3$) over the

sampling period, and ET is the exposure time (h/d). EF is the frequency of exposure (365 d/y), ED is the exposure duration (y), and AT is the average time in h ($ED \times 365 \times 24$).

The health risk assessment was based on adults residing in Korea. The exposure parameters used in the cancer and non-cancer risk assessments and their sources are listed in Table 3.2.

Table 3.2. Exposure parameters and input variables used in health risk calculation

Factor	Definition	Unit	Value	Source
C	The concentration of the metal in Ambient air	$\mu g/m^3$	Median, 95 percentile values	This study
ET	Exposure Time	hours/day	6	Fan et al., 2021
EF	Exposure Frequency	day/year	350	This study
ED	Exposure Duration	year	63.7	Korean average (NIER, 2019)
AT	Average time	hours	558,012	

To estimate the carcinogenic risk by inhalation of $PM_{2.5}$ bound trace elements, the incremental lifetime cancer risk (ILCR) was calculated following the risk assessment guidelines established by the US EPA (2009, 2013). The $ILCR_{inh}$ was calculated using Eq. 3.7 (US EPA, 2009).

$$ILCR_{inh} = ADD_{inh} \times IUR \quad \text{Eq. 3.7}$$

where IUR is the inhalation unit risk ($m^3/\mu g$).

According to the US EPA(1998, 2013), an ILCR lower than 1×10^{-6} is regarded as negligible, an ILCR above 1×10^{-4} is likely to be harmful to human beings, and an ILCR value between 1×10^{-6} and 1×10^{-4} indicates a tolerable risks, but needing risk reduction plans. The IUR values were based on credible values from the US EPA's Integrated Risk Information System (IRIS), and the Office of Environmental Health Hazard Assessment, (OEHHA) from the US EPA (2021), depending on the element. Table 3.3 shows the IUR values of each element, their sources, and the calculation results of health effects.

Table 3.3. Toxicological data and carcinogenic risk of $PM_{2.5}$ in Siheung

Chemical	IUR ($m^3/\mu g$)	Critical effect*	Source**	ILCR	
				Using median concentrations	Using 95 percentile concentrations
As	4.3.E-03	Lung irritation, decreased production of both red blood cells and white cells, deoxyribonucleic acid (DNA) damage	IRIS	4.47E-06	1.17E-05
Cr^{6+}	1.2.E-02	Liver and kidney disease, lung cancer	IRIS	2.04E-06	4.17E-06
Ni	2.4.E-04	Lung embolisms, lung and nasal cancer	IRIS	7.07E-08	1.30E-07
Pb	1.2.E-05	Renal impairment, encephalopathic signs	OEHHA	6.92E-08	1.72E-07

* Critical effects indicated the major carcinogenic effects on humans listed in the literature (Briffa et al., 2020)
** The sources listed were the original reference of the value, and the values were downloaded from US-EPA (https://www.epa.gov/risk/regional-screening-levels-rsls-generic-tables, last access: 10 August 2021

The hazard quotient (HQ) and hazard index (HI) indicating the non-carcinogenic risk from $PM_{2.5}$ bound toxic elements were calculated using Eq. 3.8 and Eq. 3.9, respectively (US EPA, 2009).

$$HQ = ADD_{inh}/(RfC_i \times 1{,}000\ \mu g/mg) \quad \text{Eq. 3.8}$$

$$HI = \sum HQ_i \quad \text{Eq. 3.9}$$

where RfC_i is the inhalation reference concentration (mg/m^3) and i is the target element.

HI is a cumulative metric for HQs for individual toxic elements and exposure pathway. An HI value > 1 indicates the presence of non-carcinogenic risk, whereas values ≤ 1 indicate a negligible non-carcinogenic effect. The RfC_i values were determined according to the OEHHA, IRIS, and additional references (Agency for Toxic Substances and Disease Registry, ATSDR; Michigan Department of Environmental Quality, MDEQ; California Environmental Protection Agency, CalEPA) from the US EPA (2021).

The health risks calculated in Siheung were compared to those in Seoul and Daebudo, of which measured data were obtained from the literature (Kim et al., 2018; Park et al., 2019). Median values and the same exposure parameters were used in the health risk estimation for the comparison using consistent manners. The period of available data was 2013- 2014 for Seoul, 2019- 2020 for Siheung, and 2016 for Daebudo.

3.3. Results and discussion

3.3.1 $PM_{2.5}$ mass concentration and chemical speciation

The average mass concentration of $PM_{2.5}$ over the sampling period (11/16/2019 to 10/02/2020) was 23.5 ± 13.9 μg/m^3. A time series plot is shown in Fig. 3.3 to compare the $PM_{2.5}$ concentration data obtained in this study and those provided from a national monitoring station (https://www.airkorea.or.kr/, last access: August 10, 2021).

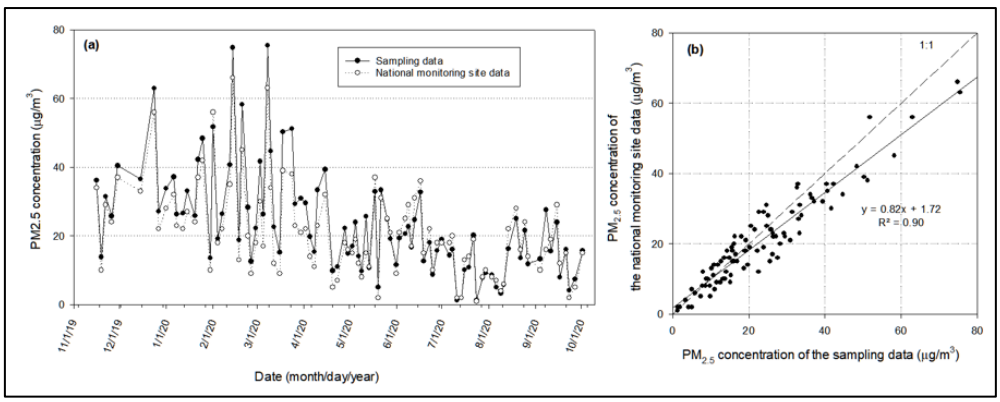


Fig. 3.3. $PM_{2.5}$ mass concentration comparisons between the sampled filter and the nearest national monitoring station. (a): time-series plot, and (b) 1:1 plot

Both time series presented a similar trend, which confirmed the validity of our data acquisition. High concentrations (over the Korean daily standard of 25 μg/m^3) were observed in 37 of the 95 samples, primarily in winter and spring (35 cases from November to May). The detailed concentrations of $PM_{2.5}$ and chemical species (29 species) are summarized in Table 3.4.

Table 3.4. $PM_{2.5}$ species concentrations in Siheung, Korea during the entire sampling period (11/16/2019 to 10/2/2020)

Species	Arithmetic mean (ng m^{-3})	25th percentile (ng m^{-3})	Median (ng m^{-3})	75th percentile (ng m^{-3})	Maximum (ng m^{-3})
$PM_{2.5}$	23,500	13,500	20,600	31,200	74,800
NO_3^-	5,160	993	2,590	7,740	27,200
SO_4^{2-}	3,580	1,800	3,260	4,380	14,100
NH_4^+	2,910	1,330	2,710	4,100	12,100
K^+	166	58.9	139	239	525
Na^+	165	104	144	188	604
Cl^-	366	59.2	168	477	2,490
OC	5,830	3,760	5,330	7,370	15,400
EC	649	406	561	826	1,908
Na	187	136	172	222	536
Mg	41.0	27.5	34.8	49.7	159
Al	84.1	44.7	72.0	113	265
Si	222	107	185	296	665
S	1,850	1,130	1,740	2,310	6,200
Cl	505	113	248	772	2,560
K	233	108	196	328	766
Ca	51.4	28.2	43.3	66.6	233
Ti	7.41	4.38	6.37	9.93	20.1
V	0.396	0.196	0.319	0.531	1.41
Cr	2.43	1.21	2.25	3.14	8.25
Mn	16.4	10.5	16.2	21.4	44.5
Ba	6.25	3.01	4.45	7.33	30.9
Fe	188	124	171	239	458
Ni	1.26	0.788	1.14	1.65	3.38
Cu	7.13	1.98	4.77	10.3	45.0
Zn	73.5	42.4	60.6	98.8	226
As	4.74	1.90	3.34	6.61	27.3
Se	1.63	0.881	1.56	2.22	3.82
Br	13.6	5.99	9.78	14.9	168
Pb	25.7	11.8	21.3	31.6	111

The $PM_{2.5}$ concentration levels in Siheung and other cities are shown in Fig. 3.2. The average daily $PM_{2.5}$ concentration in Siheung was similar to that in Seoul and higher than those in Yeosu and Ulsan, which are industrial cities in South Korea.

Seoul and Siheung are cities located in the northwest of South Korea and are known to be affected by long-range transport of $PM_{2.5}$ from China (Bae et al., 2019; Kumar et al., 2021). The contribution of long-range transport from China to $PM_{2.5}$ in Seoul was estimated ranged from 41% to 44% between 2012 and 2016 (Bae et al., 2019), approximately 20% in August, and approximately 60% in January and February (Kumar et al., 2021). In comparison to industrial cities of other countries, the average $PM_{2.5}$ concentration in Siheung was higher than those in Hamburg and Kassel, in Germany, and lower than those in Beijing and Shanghai in China. This suggests that source apportionment coupled with health risk assessment in Siheung may be an example of a small and medium-sized industrial city with moderate $PM_{2.5}$ pollution.

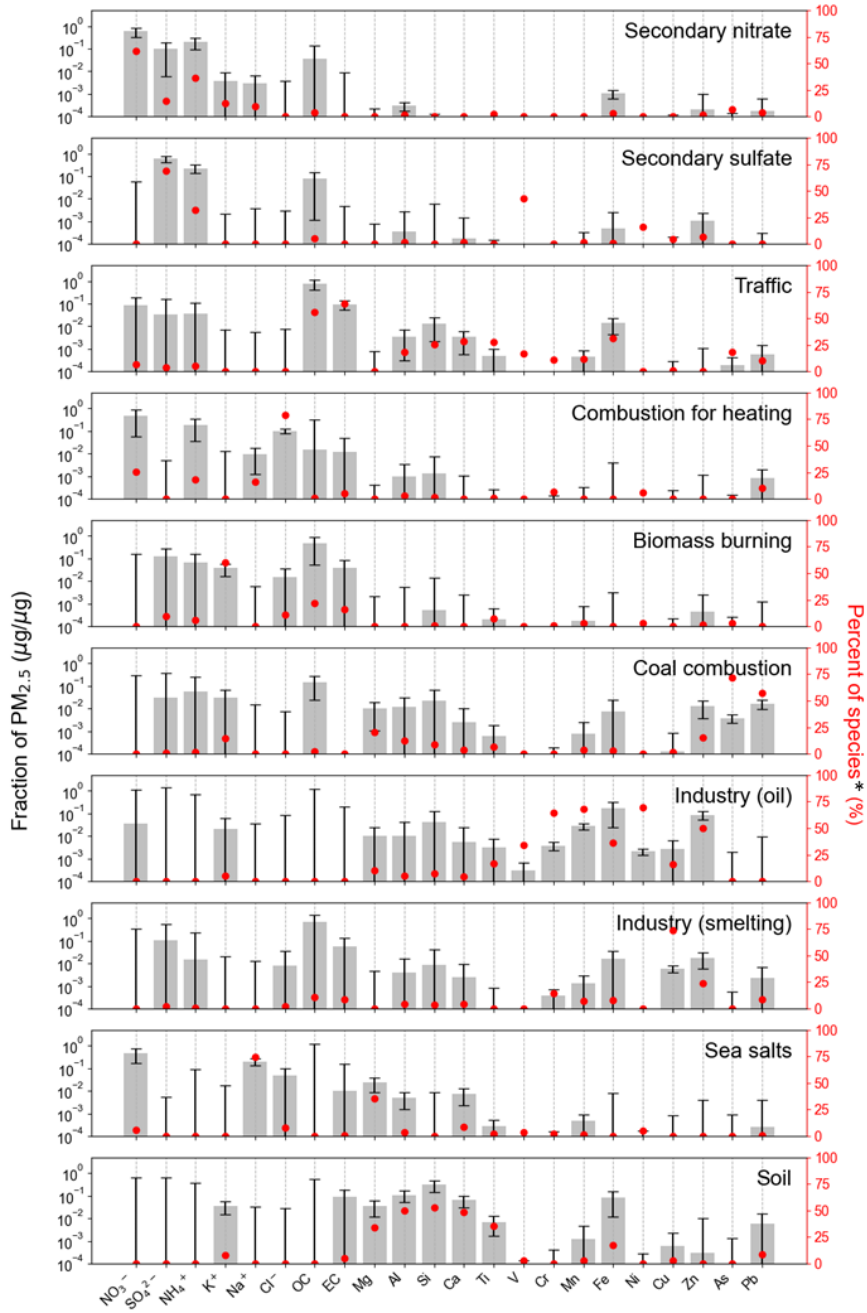
As the measurement and analysis period of this study included the COVID-19 lockdown or social distancing period in neighboring countries and Korea, we evaluated possible interferences. A previous study on air quality change in Seoul under COVID-19 social distancing reported that the monthly average $PM_{2.5}$ concentration (from 29 February to 29 March 2020) decreased by 10.4% in 2020, which was contrary to the average increase of 23.7% over the corresponding periods in the previous five years (Han et al., 2020). Je et al. (2021) also reported that the mean $PM_{2.5}$ level in 2020 decreased by 16.98 $\mu g/m^3$ nationwide in Korea compared to 2019, which represented a decrease of 45.45% ($p < 0.001$). However, significant reductions in $PM_{2.5}$ were observed in Korea even before social distancing owing to the changes in transboundary $PM_{2.5}$ concentration (Kim and Lee, 2018). In China, the average $PM_{2.5}$ concentration during the lockdown period (January to February 2020) was 18 $\mu g/m^3$, which represented a reduction of 30–60% in most regions (Bai et al., 2021).

Although there may be a gap between present results and previous ones, comparison with previous data is essential to obtain detailed information on $PM_{2.5}$ pollution. A comparison of average concentrations of $PM_{2.5}$ bound chemicals obtained in this study and those by Park et al. (2019) in Seoul indicated that Siheung had a higher concentration of Cr than Seoul. The average concentrations of As, Pb, Cr, Mn, Ni, Cu, Zn, and V, which are major toxic elements, were 4.74, 25.74, 2.43, 16.37, 1.26, 7.13, 73.55, and 0.40 ng/m^3 in Sheung, and 5.53, 38.11, 1.74, 16.93, 2.11, 7.92, 100 and 4.30 ng/m^3 in Seoul (Park et al., 2019) respectively. The concentrations of toxic elements except Cr were higher in Seoul than in Siheung. However, further research is required to determine the impacts of reduced concentrations attributed to the effects of the COVID-19. When comparing the concentrations of elements in Siheung and Seoul during the sampling period of this study, the mean concentrations of Pb, Cr, Mn, Ni, Cu, Zn, and V in Siheung were 1.6, 3.0, 2.2, 4.0, 2.8, 2.2, and 1.4 times higher than those in Seoul (Korea Ministry of Environment and National Institute of Environmental Research, 2022), respectively. These results might indicate that Siheung has a high concentration of Cr and other elements because the concentrations were high even during the COVID-19 lockdown period. This was suggested because these elements are considered chemical markers of combustion and traffic sources (Farahani et al., 2021), which were reduced during the lockdown period. In Beijing, the mean concentrations of $PM_{2.5}$-bounded As, Pb, Cr, Mn, Ni, Zn, and V during the winter of 2018 were 4, 44, 15, 34, 8, 110, and 7 ng/m^3 (Fan et al., 2021), respectively, which are overall higher than those obtained in Siheung. The concentrations of the clean case presented in the literature showed similar results to those of Siheung. In Quebedo, Portugal (Silva et

al., 2020), the concentrations of As, Cr, and Zn were 0.44, 3.55, and 11.0 ng/m^3, which were lower than those in Siheung, Korea.

3.3.2 Source apportionment of $PM_{2.5}$ by PMF modeling

The source profile and the time series of PMF factor contribution are shown in Fig. 3.4 and Fig. 3.5, respectively.



* Percent of species: the percentage concentration of each chemical species contributing to each of the sources (i.e., the sum of the percent of species values for each element from all sources is 100)

Fig. 3.4. Source profile results of PMF modeling with DISP errors (The black bar corresponds to the left axis, and the red dot corresponds to the right axis)

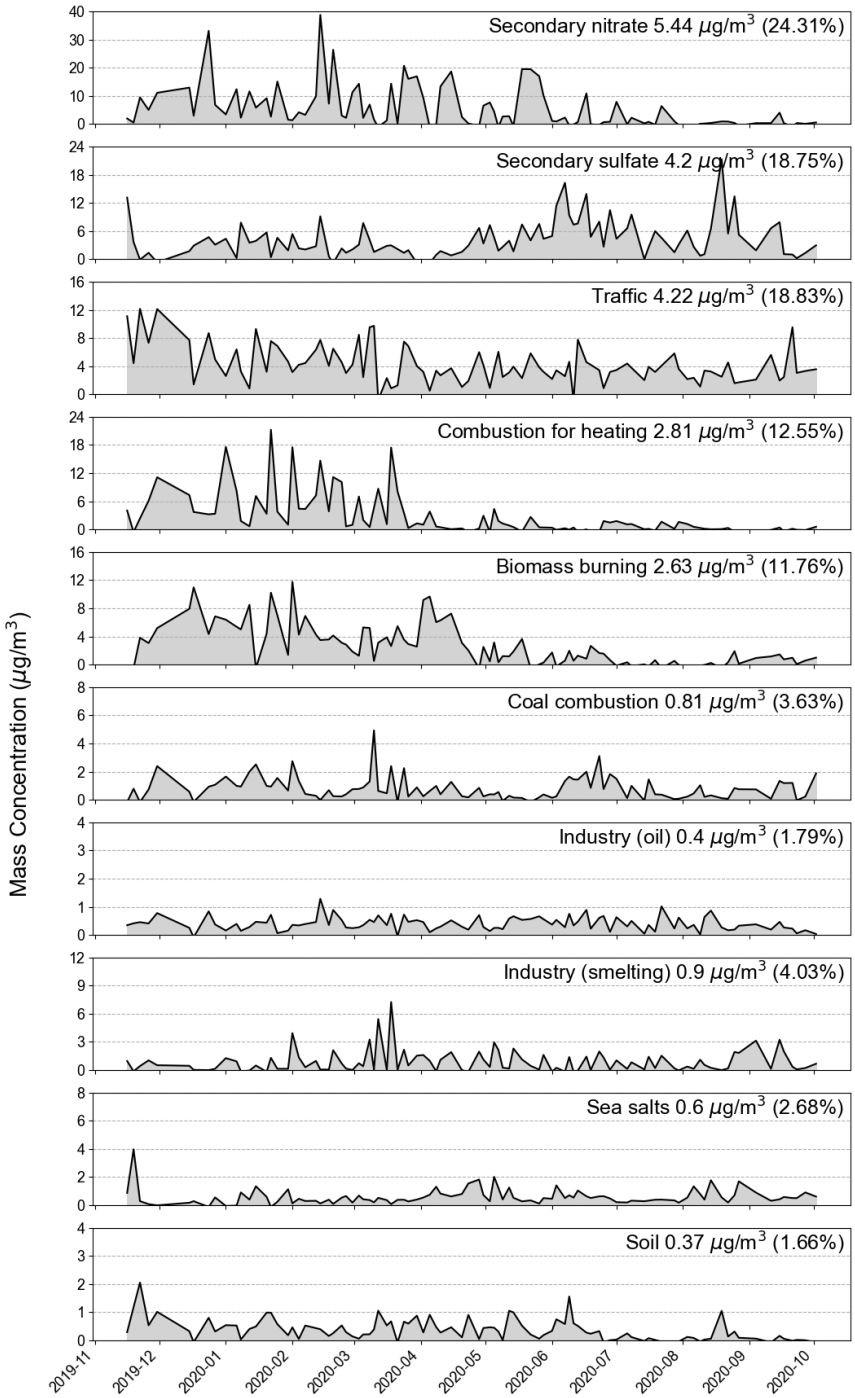


Fig. 3.5. Source contribution time-series plot of $PM_{2.5}$ in Siheung, Republic of Korea

Total 10 sources of $PM_{2.5}$ were identified, and all major species of the sources were within the DISP intervals (Fig. 3.4). The R^2 between observed and predicted $PM_{2.5}$ concentrations for the best solution was 0.92, indicating a reasonable modeling result. The 10 sources included secondary nitrate, secondary sulfate, traffic, combustion for heating, biomass burning, coal combustion, heavy oil industry, smelting industry, sea salts, and soil. The sources with the highest contributions were the secondary-generated particles (secondary nitrate and sulfate) (Fig. 3.5).

Secondary nitrate had an average contribution of 24.3% to $PM_{2.5}$ mass concentration. The concentration of secondary nitrate was relatively high in the winter when the temperature was low (Fig. 3.5). The main species of secondary nitrate are NH_4^+ and NO_3^-, which are formed in urban air primarily through gas-particle partitioning (Shi et al., 2019). This occurs because nitrogen oxide and ammonia gas, which are gaseous precursors in spring and winter, easily react in the atmosphere producing particulate nitrate (Choi et al., 2013; E. H. Park et al., 2020). Secondary sulfate (18.8%) was identified by the high concentrations of SO_4^{2-} and NH_4^+ (E. H. Park et al., 2020). The contribution of secondary sulfate tended to increase primarily in the summer. This is considered to reflect the formation of sulfate in the atmosphere that becomes active when both temperature and humidity are high (Heo et al., 2009).

Traffic was identified as a source using OC and EC as major indicator components, and it contributed to 18.8% of the $PM_{2.5}$. The high component ratio of carbon species exhibited the characteristics of automobile pollutants. Fe is also considered as an indicator of traffic resuspension as it is emitted from the brake wear of gasoline and diesel-powered engines (Belis et al., 2013).

Combustion for heating as a pollution source was characterized by the high Cl^- content (Tian et al., 2020), and it presented a high contribution from November 2019 to March 2020. This period coincided with the heating periods in Korea and northern China. The combustion for heating contributed to 12.6% of the $PM_{2.5}$.

Biomass burning contributed to 11.8% of $PM_{2.5}$, with K^+ as its major component (Andreae, 1983). Its contribution was identified by the high load of OC and the medium load of EC (Liu et al., 2017; Moon et al., 2008). In addition, biomass burning exhibited seasonal characteristics with a high contribution in the winter (Shi et al., 2014), which is consistent with the increase in the use of wood fire for domestic heating (Choi et al., 2013).

Coal combustion contributed to 3.6% of $PM_{2.5}$, and As and Pb were considered its major indicator components. The contribution of coal combustion did not exhibit any distinct seasonal fluctuations, which was consistent with the characteristics of local sources. For example, Arsenic is known as a major marker of coal combustion pollution (Duan and Tan, 2013), and it is known to be largely emitted from fossil fuel burning.

Industrial sources were divided into heavy oil- and smelting-related sources. The high ratio of V and Ni was considered a characteristic of heavy oil-based industrial sources (Jang et al., 2007). For industrial smelting sources, the major indicators were heavy metal components such as Cu, Cr, Mn, Pb, and Zn (Dai et al., 2015). The industrial contributions did not show significant seasonal fluctuations.

Sea salt sources were identified by high concentrations of Na, Mg, and K (E. H. Park et al., 2020). The source was referred to as a fresh seal salt because of the relatively high concentration of chlorine ions (Han et al., 2017). Its

concentrations exhibited seasonal characteristics, and the highest contributions were observed during the winter. Finally, soil sources were identified by the existence of representative crustal components such as Mg, Al, Si, Ca, and Ti (Liu et al., 2017; Thorpe and Harrison, 2008) and they contributed to 1.7% of $PM_{2.5}$.

Park et al. (2020) performed PMF modeling in Seoul in 2014–2015 and isolated 9 sources. The contributions of secondary sources and traffic sources in Seoul were 6.3 and 5.3 ug/m^3 higher than those in Siheung, respectively. Unlike in the study of Seoul (E. H. Park et al., 2020), the industrial smelting source was extracted in this study probably due to non-ferrous smelter sources in the near national industrial complex. The existence of a smelting source was also observed in a PMF modeling study in Daebudo (Kim et al., 2018), near Siheung. In the literature, Cu, Zn, and Pb have been designated as major markers of industrial smelting sources (Kim et al., 2018).

3.3.3 Carcinogenic and non-carcinogenic health risks

The uncertainty of health risk estimates coupled with PMF modeling results was calculated. The difference between the health risks using the measured values and the health risks coupled with PMF model results was within 10% (data not shown). The calculated carcinogenic health risks by elements were shown in Table 3.3.

The obtained carcinogenic health risks indicated that both the median and 95 percentile concentrations of As and Cr^{6+} exceeded the ILCR value of 1E-06, whereas the ILCR values of Ni and Pb did not exceed the reference value (Table 3.3).

These results suggest that air pollution management in Siheung should be based on pollution sources, focusing on As and Cr sources. This can also be confirmed in Table 3.5, which presents the health risk assessment results by element and source.

Table 3.5. Estimated carcinogenic risk in Sihueng (median elemental concentrations used)

Source	Toxic element in $PM_{2.5}$				Sum of incremental cancer risk by source
	As	Cr^{6+}	Ni	Pb	
Secondary nitrate	2.90E-07	-	-	2.86E-09	**2.93E-07 (4.4%)**
Secondary sulfate	-	-	1.14E-08	-	**1.14E-08 (0.2%)**
Mobile	8.34E-07	2.30E-07	-	7.07E-09	**1.07E-06 (16.0%)**
Combustion for heating	-	1.32E-07	4.51E-09	7.17E-09	**1.44E-07 (2.1%)**
Biomass burning	1.52E-07	2.12E-08	2.19E-09	-	**1.75E-07 (2.6%)**
Coal combustion	**3.24E-06**	-	-	4.02E-08	**3.28E-06 (48.9%)**
Industry (oil)	-	**1.32E-06**	4.93E-08	-	**1.37E-06 (20.4%)**
Industry (smelting)	-	3.02E-07	-	6.26E-09	**3.08E-07 (4.6%)**
Sea salts	-	5.11E-08	3.53E-09	4.61E-10	**5.51E-08 (0.8%)**
Soil	-	-	2.60E-10	6.13E-09	**6.39E-09 (0.1%)**
Sum of incremental cancer risk by element	**4.52E-06 (67.2%)**	**2.06E-06 (30.7%)**	**7.12E-08 (1.1%)**	**7.02E-08 (1.0%)**	**6.71E-06 (100%)**

According to the estimated health risks from $PM_{2.5}$ sources using the median concentrations, the sources with high health risk potentials were coal combustion, oil industries, and traffic, which accounted for 48.9%, 20.4%, and 16.0% of the total ILCR value, respectively (Table 3.5). The concentration of portioned As and Cr had

the greatest influence on the health risk values of each source. However, the absolute contributions of them to $PM_{2.5}$ mass concentrations, were 3.6%, 1.8%, and 18.8%, respectively (Fig. 3.5). Fig. 3.6 shows annual average contributions of sources to $PM_{2.5}$ mass concentrations and to cumulative cancer risk, and of elements to cumulative cancer risks.

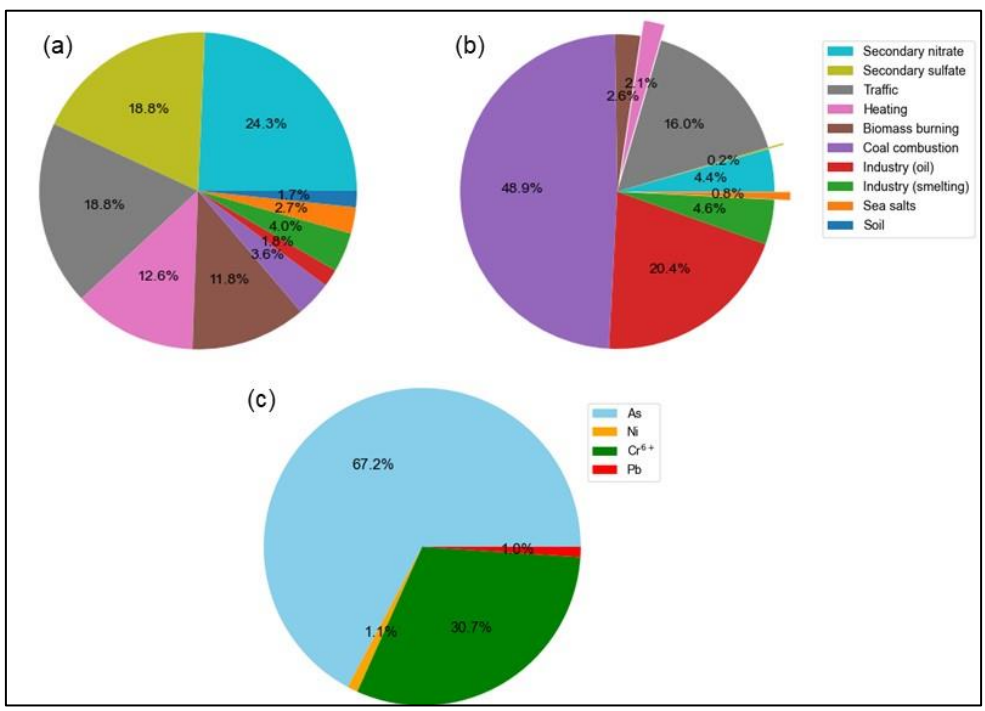


Fig. 3.6. Annual average contributions (a) of sources to $PM_{2.5}$ mass concentrations, (b) of sources to cancer risks, and (c) of elements to cancer risks

The contributions of sources to $PM_{2.5}$ mass concentration and to health risks were very different. Therefore, the contribution of $PM_{2.5}$ sources might not be representative of health risks, which supports the argument that to manage $PM_{2.5}$ with a focus on health risks, the concentration of toxic metal elements should be

considered rather than total mass concentration. (Farahani et al. 2021).

The concentrations of As and Cr that must be reduced to achieve negligible health effects were calculated. The results indicate that to reduce the health risks of As to below 1E-06, the As concentration should be reduced to 1 ng/m^3 or less, which represents a reduction of at least 75% compared to the current level. For Cr, the required concentration reduction was at least 50%. Therefore, there is a need for a significant reduction in coal combustion, which is the main source of As pollution, and in emissions from the oil industry, which are the main sources of Cr. In addition, as the seasonal differences in ILCR were not significant (data not shown), an overall reduction is necessary, instead of a specific-season reduction plan.

Strengthening the control of pollutants emitted from industrial sources is an important environmental and public health issue. Therefore, the industrial emission sources of As and Cr in cities such as Siheung need to be managed, and efforts to reduce ambient concentrations need to be taken. Owing to the COVID-19 pandemic, industrial activity and traffic were likely restricted compared to usual rates during this study. This is supported by Dai et al. (2021), who reported that human activities, such as industry and transportation, declined during the epidemic outbreak and spread. Therefore, it is possible that the health risks assessed in this study were underestimated. Therefore, further studies beyond the pandemic period are needed for an accurate estimation of health risks.

The calculated ILCR values for Siheung (2019–2020), Seoul (2013–2014), and Daebudo (2016) are shown in Table 3.6.

Table 3.6. Estimated carcinogenic and non-carcinogenic risks of $PM_{2.5}$ in Siheung, Seoul, and Daebudo, Korea (median concentration of each element used)

Toxic elements in $PM_{2.5}$	Siheung, Korea (2019.11 – 2020. 10)		Seoul, Korea* (2013 – 2014)		Daebudo, Korea** (2016)	
	ILCR	**HQ**	**ILCR**	**HQ**	**ILCR**	**HQ**
As	4.52E-06	7.01E-02	5.70E-06	8.84E-02	2.89E-06	4.47E-02
Cr^{6+}	2.06E-06	3.42E-02	1.50E-06	2.50E-02	8.63E-08	1.44E-03
Cr^{3+}	-	3.99E-03	-	2.92E-03	-	1.68E-04
Cu	-	8.63E-04	-	9.49E-04	-	1.07E-03
Ni	7.12E-08	2.12E-02	1.21E-07	3.61E-02	5.75E-09	1.71E-03
Pb	7.02E-08	3.90E-02	1.10E-07	6.09E-02	4.43E-08	2.46E-02
V	-	8.74E-04	-	1.03E-02	-	2.73E-02
Mn	-	7.84E-02	-	8.12E-02	-	3.84E-02
Sum	**6.71E-06**	**2.49E-01**	**1.35E-05**	**5.85E-01**	**3.02E-06**	**1.39E-01**

The results of Seoul were calculated from the data of Park et al. (2019), and the results of Daebudo were calculated from the data of Kim et al. (2018). The health risk from As in Siheung (4.52E-06) was lower than those in Seoul (1.35E-05) and Daebudo (3.02E-06). This result might have been obtained because the Siheung data reflected an underestimation of the decrease in human activity owing to the COVID-19 pandemic. The health risk values in Nanjing (Hu et al., 2012) and Beijing (Fan et al., 2021) in China were 9.04E-06 and 1.67E-06, respectively, which were similar to the value Siheung. These results indicate that As presents a health risk even at low concentrations (ng/m^3). This is consistent with previous studies suggesting that the presence of As in the atmosphere is a major public concern for human health (Widziewicz et al., 2016). Nevertheless, the health risk of Cr^{6+}, Siheung, and Seoul also exceeded 1E-06, and Siheung presented the highest value (2.06E-06); therefore, Cr pollution in Siheung should be carefully managed. A similar observation of Cr-

dominated carcinogenic risk from industrial and traffic sources has been reported in Delhi, India (Khillare and Sarkar, 2012). Hu et al. (2012) and Fan et al. (2021) reported that the carcinogenic risks of Cr for adults from $PM_{2.5}$ in Nanjing and Beijing were 8.70E-05; and 2.2E-05, respectively, which are approximately 20.9 and 5.3 times the value in Siheung. The industries were identified as Cr sources in this study (Fig. 3.4). Accordingly, Fan et al. (2021) identified the metal smelting industry as the main source of Cr.

The non-carcinogenic health risks of all elements were less than 0.1 for both average and 95 percentile concentrations. Moreover, the HI value was 0.55, which did not exceed 1, thereby indicating a negligible toxic risk for all elements (Table 3.7). The maximum HQ value was 0.18 for As when the 95 percentile concentration was used. The calculations using median concentrations indicated that the pollutants with high toxicity values were the oil industry, coal combustion, and traffic (Table 3.8), which accounted for 37.4%, 30.5%, and 12.2% of the total HQ value, respectively. In contrast, according to the absolute contributions to $PM_{2.5}$ concentration, their contributions accounted for 1.8%, 3.6%, and 18.8%, respectively (Fig. 3.5). According to the HQ results, Seoul had a higher non-carcinogenic health risk (at 0.585, which did not exceed 1) than Siheung and Daebudo. This was consistent with the results of a similar study in China (Hu et al. 2012), in which the calculated HI was less than 1 for adults, so that the non-carcinogenic health risks were considered of relatively low importance.

Table 3.7. Toxicological data and non-carcinogenic risk in $PM_{2.5}$ of Siheung

Chemical	RfC_i (mg/m^3)	Critical effect*	Source**	HQ	
				Using median concentrations	Using 95 percentile concentrations
As	1.5.E-05	Heart problems, brain damage	OEHHA	6.9E-02	1.8.E-01
Cr^{6+}	5.00E-06	Allergic contact dermatitis and eczema, gingivitis	IRIS	3.4E-02	7.0.E-02
Cr^{3+}	1.0.E-04	DNA lesions (rarely toxic compared to hexavalent form)	ATSDR,2012	4.0E-03	8.1.E-03
Cu	2.0.E-03	Insomnia, anxiety, restlessness	MDEQ, 2009***	8.5E-04	2.4.E-03
Ni	1.4E-05	Asthma, allergic reactions, heart disorders	CalEPA	2.1E-02	3.9.E-02
Pb	1.5.E-04	Hypertension, miscarriages, stillbirth	IRIS	3.8E-02	9.6.E-02
V	1.0.E-04	Throat pain, headaches, impairment to the nervous system	ATSDR	8.7E-04	1.7.E-03
Mn	5.00E-05	Hypotension, pneumonia, sperm damage	IRIS	7.8E-02	1.5.E-01
		HI (Summation)		0.25	0.55

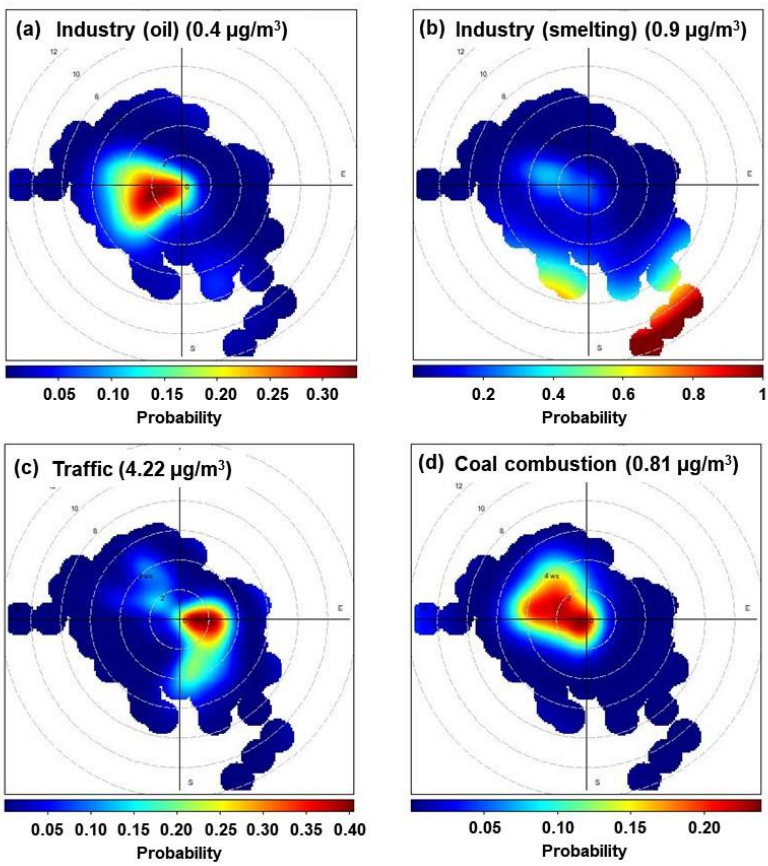
* Critical effects indicated the major non-carcinogenic effects on humans listed in the literature (Briffa et al., 2020)
** The sources listed were the original reference of the value, and the values were downloaded from US-EPA (https://www.epa.gov/risk/regional-screening-levels-rsls-generic-tables, last access: 10 August 2021)
*** The value from MDEQ was accessed in the chemical update worksheet of the State of Michigan website (https://www.michigan.gov/documents/deq/deq-rrd-chem-CopperDatasheet_527899_7.pdf, last access: last access: 10 August

Table 3.8. Estimated non-carcinogenic risk in Siheung (median elemental concentrations used)

Source	Toxic element in $PM_{2.5}$								Sum of incremental cancer risk by source
	As	Cr^{6+}	Cr^{3+}	Cu	Ni	Pb	V	Mn	
Secondary nitrate	4.50E-03	-	-	-	-	1.59E-03	-	-	**6.1E-03 (2.5%)**
Secondary sulfate	-	-	-	4.16E-05	3.39E-03	-	3.75E-04	1.08E-03	**4.9E-03 (2.0%)**
Mobile	1.29E-02	3.84E-03	4.47E-04	9.91E-06	-	3.93E-03	1.44E-04	9.14E-03	**3.0E-02 (12.2%)**
Combustion for heating	-	2.20E-03	2.56E-04	-	1.34E-03	3.98E-03	1.14E-06	-	**7.8E-03 (3.1%)**
Biomass burning	2.35E-03	3.54E-04	4.13E-05	-	6.51E-04	-	-	2.35E-03	**5.7E-03 (2.3%)**
Coal combustion	5.03E-02	-	-	1.29E-05	-	2.23E-02	-	3.07E-03	**7.6E-02 (30.5%)**
Industry (oil)	-	2.19E-02	2.56E-03	1.35E-04	1.47E-02	-	2.96E-04	5.35E-02	**9.3E-02 (37.4%)**
Industry (smelting)	-	5.03E-03	5.87E-04	6.36E-04	-	3.48E-03	-	5.68E-03	**1.5E-02 (6.2%)**
Sea salts	-	8.52E-04	9.94E-05	6.70E-07	1.05E-03	2.56E-04	3.06E-05	1.40E-03	**3.7E-03 (1.5%)**
Soil	-	-	-	2.70E-05	7.74E-05	3.40E-03	2.73E-05	2.24E-03	**5.8E-03 (2.3%)**
Sum of incremental cancer risk by element	**7.01E-02 (28.2%)**	**3.42E-02 (13.8%)**	**3.99E-03 (1.6%)**	**8.63E-04 (0.4%)**	**2.12E-02 (8.5%)**	**3.90E-02 (15.7%)**	**8.74E-04 (0.4%)**	**7.84E-02 (31.6%)**	**0.25 (100%)**

3.3.4 Probable source areas or directions

The probable emission locations were estimated for coal combustion, industries, and traffic sources, which presented a relatively high carcinogenic risk in the health risk assessment. The CPF results are shown in Fig. 3.7, and the PSCF results calculated through 24-h and 72-h back trajectory HYSPLIT analysis are shown in Fig. 3.8.



* The center of each figure is the measurement site
** The scale of the circle shows the wind speed (m/s)

Fig. 3.7. The CPF results of (a) industry (oil), (b) industry (smelting), (c) traffic, and (d) coal combustion sources

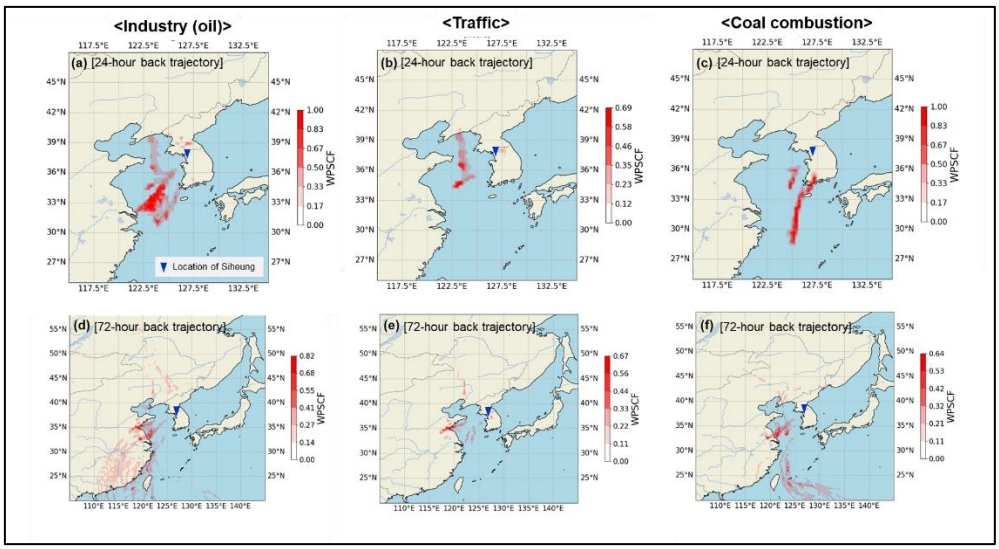


Fig. 3.8. PSCF results of $PM_{2.5}$ sources in Siheung, Republic of Korea, 24-hour back trajectory of (a) Industry (oil); (b) Traffic; (c) Coal combustion, 74-hour back trajectory of (d) Industry (oil); (e) Traffic; (f) Coal combustion

The CPF results for industrial sources indicated that the contribution of oil industries increased when the southwest winds of less than 4 m/s, that of smelting industries increased with southeast winds of 6 m/s or more. The results of the back trajectory analysis showed that the contribution of industries was widely distributed in southwest areas, from the Shandong Peninsula of China to the Taiwan region. According to Kim et al. (2018), the CPF of non-ferrous smelter sources pointed to the southeast of Daebudo, which was consistent with CPF results for smelting industry sources in this study. There are 4,632 high-tech manufacturing companies such as metal processing and machinery located in the national industrial complex of Siheung (as of 2019, Korea Statistical geographic information service, https://sgis.kostat.go.kr/, last access: last access: 10 August 2021), and more than 240,000 people are working in related industries. PM emitted from such industrial complexes was presumed to be industry (smelting) sources.

Coal combustion presented the highest contribution for northwest winds of approximately 2−6 m/s in the CPF plot (Fig. 3.7). In the 72-h back trajectory analysis (Fig. 3.8), PSCF was distributed along the Chinese coast from the west coast of Korea to the southwest of Korea. The results suggested that coal combustion sources presented high emissions from internal sources. Coal-fired power plants, petrochemical complexes, and Incheon ports are located around Siheung, so it was assumed that the influence of various sources was mixed. However, it was difficult to identify the specific locations, as there were various influencing factors in the vicinity. Long-term studies are required.

The CPF of traffic source showed that the contribution increased with slow winds of 3 m/s or less (Fig. 3.7). Siheung City has much traffic because of its proximity to Seoul and Incheon ports and it is presumed that this trend was well-reflected. The wind direction pattern also showed a result that was generally consistent with the arrangement of highways around the target area. The probability of the western sea of Korea was also high in the back-trajectory analysis (Fig. 3.8).

3.4. Summary

Ten types of $PM_{2.5}$ emission sources were derived using a PMF model in Siheung, South Korea. Based on the sources derived, the carcinogenic and non-carcinogenic health risks due to $PM_{2.5}$ inhalation were estimated. For coal combustion, heavy oil industry, and traffic sources, the contribution to $PM_{2.5}$ mass concentration was low but exceeded the benchmark carcinogenic health risk value (1E-06). The carcinogenic risk from $PM_{2.5}$ inhalation in Siheung was similar to or lower than that of Seoul, Republic of Korea and Nanjing, China, and Beijing, China. Therefore, countermeasures on the $PM_{2.5}$ emission sources are better to be performed not only based on the $PM_{2.5}$ mass concentration but also based on the health risks. In order to manage the effects of $PM_{2.5}$ on human health in industrial cities, it is necessary to reduce the concentration of major toxic elements (especially As and Cr) and manage the emission sources. The methodology used in this study, which combines PMF modeling and health impact assessment, can be used to derive source types and calculate health impacts by source in other cities.

References

Anderson, H.R., 2009. Air pollution and mortality: A history. Atmos. Environ. 43, 142–152. https://doi.org/10.1016/j.atmosenv.2008.09.026

Andreae, M.O., 1983. Soot carbon and excess fine potassium: Long-range transport of combustion-derived aerosols. Science (80-.). 220, 1148–1151. https://doi.org/10.1126/science.220.4602.1148

Ashbaugh, L.L., Malm, W.C., Sadeh, W.Z., 1985. A residence time probability analysis of sulfur concentrations at grand Canyon National Park. Atmos. Environ. 19, 1263–1270. https://doi.org/10.1016/0004-6981(85)90256-2

Bae, C., Kim, B.U., Kim, H.C., Yoo, C., Kim, S., 2019. Long-Range Transport Influence on Key Chemical Components of PM2.5 in the Seoul Metropolitan Area, South Korea, during the Years 2012–2016. Atmos. 2020, Vol. 11, Page 48 11, 48. https://doi.org/10.3390/ATMOS11010048

Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2012.11.009

Briffa, J., Sinagra, E., Blundell, R., 2020. Heavy metal pollution in the environment and their toxicological effects on humans. Heliyon 6, e04691. https://doi.org/10.1016/j.heliyon.2020.e04691

Brown, S.G., Eberly, S., Paatero, P., Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Sci. Total Environ. 518–519, 626–635. https://doi.org/10.1016/j.scitotenv.2015.01.022

Carslaw, D., 2015. The openair manual open-source tools for analysing air pollution data. King's Coll. London 287.

Cassee, F.R., Héroux, M.E., Gerlofs-Nijland, M.E., Kelly, F.J., 2013. Particulate matter beyond mass: Recent health evidence on the role of fractions, chemical constituents and sources of emission. Inhal. Toxicol. 25, 802–812. https://doi.org/10.3109/08958378.2013.850127/SUPPL_FILE/IIHT_A_850127_SM0004.PDF

Choi, E., Choi, K., Yi, S.M., 2011. Non-methane hydrocarbons in the atmosphere of a Metropolitan City and a background site in South Korea: Sources and health risk potentials. Atmos. Environ. 45, 7563–7573. https://doi.org/10.1016/j.atmosenv.2010.11.049

Choi, E., Muk, S., Young, Y., Lee, S., Jo, H., Ok, S., Jong, B., Heo, B., 2022. Sources of airborne particulate matter - bound metals and spatial - seasonal variability of health risk potentials in four large cities , South Korea. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-021-18445-8

Choi, J. kyu, Heo, J.B., Ban, S.J., Yi, S.M., Zoh, K.D., 2013. Source apportionment of PM2.5 at the coastal area in Korea. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2012.12.047

Cohen, D.D., Crawford, J., Stelcer, E., Bac, V.T., 2010. Characterisation and source apportionment of fine particulate sources at Hanoi from 2001 to 2008. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2009.10.037

Dai, Q., Ding, J., Song, C., Liu, B., Bi, X., Wu, J., Zhang, Y., Feng, Y., Hopke, P.K., 2021. Changes in source contributions to particle number concentrations after the COVID-19 outbreak: Insights from a dispersion normalized PMF. Sci.

Total Environ. 759. https://doi.org/10.1016/j.scitotenv.2020.143548

Dai, Q., Liu, B., Bi, X., Wu, J., Liang, D., Zhang, Y., Feng, Y., Hopke, P.K., 2020. Dispersion normalized PMF provides insights into the significant changes in source contributions to PM2.5 after the CoviD-19 outbreak. Environ. Sci. Technol. 54, 9917–9927. https://doi.org/10.1021/acs.est.0c02776

Dai, Q.L., Bi, X.H., Wu, J.H., Zhang, Y.F., Wang, J., Xu, H., Yao, L., Jiao, L., Feng, Y.C., 2015. Characterization and source identification of heavy metals in ambient PM10 and PM2.5 in an integrated Iron and Steel industry zone compared with a background site. Aerosol Air Qual. Res. 15, 875–887. https://doi.org/10.4209/aaqr.2014.09.0226

Du, X., Yang, J., Xiao, Z., Tian, Y., Chen, K., Feng, Y., 2021. Source apportionment of PM2.5 during different haze episodes by PMF and random forest method based on hourly measured atmospheric pollutant. Environ. Sci. Pollut. Res. 2021 1–12. https://doi.org/10.1007/S11356-021-14487-0

Duan, J., Tan, J., 2013. Atmospheric heavy metals and Arsenic in China: Situation, sources and control policies. Atmos. Environ. 74, 93–101. https://doi.org/10.1016/J.ATMOSENV.2013.03.031

Fan, M.Y., Zhang, Y.L., Lin, Y.C., Cao, F., Sun, Y., Qiu, Y., Xing, G., Dao, X., Fu, P., 2021. Specific sources of health risks induced by metallic elements in PM2.5 during the wintertime in Beijing, China. Atmos. Environ. 246, 118112. https://doi.org/10.1016/j.atmosenv.2020.118112

Fang, C., Wang, L., Gao, H., Wang, J., 2020. Analysis of the PM2.5 emission inventory and source apportionment in Jilin City, Northeast of China. Environ. Sci. Pollut. Res. 2020 2730 27, 37324–37332. https://doi.org/10.1007/S11356-

020-07605-X

Farahani, V.J., Soleimanian, E., Pirhadi, M., Sioutas, C., 2021. Long-term trends in concentrations and sources of PM2.5–bound metals and elements in central Los Angeles. Atmos. Environ. 253, 118361. https://doi.org/10.1016/j.atmosenv.2021.118361

Fu, S., Yue, D., Lin, W., Hu, Q., Yuan, L., Zhao, Y., Zhai, Y., Mai, D., Zhang, H., Wei, Q., He, L., 2021. Insights into the source-specific health risk of ambient particle-bound metals in the Pearl River Delta region, China. Ecotoxicol. Environ. Saf. 224, 112642. https://doi.org/10.1016/J.ECOENV.2021.112642

Fushimi, A., Nakajima, D., Furuyama, A., Suzuki, G., Ito, T., Sato, K., Fujitani, Y., Kondo, Y., Yoshino, A., Ramasamy, S., Schauer, J.J., Fu, P., Takahashi, Y., Saitoh, K., Saito, S., Takami, A., 2021. Source contributions to multiple toxic potentials of atmospheric organic aerosols. Sci. Total Environ. 773, 145614. https://doi.org/10.1016/J.SCITOTENV.2021.145614

Han, B.-S., Park, K., Kwak, K.-H., Park, S.-B., Jin, H.-G., Moon, S., Kim, J.-W., Baik, J.-J., 2020. Air Quality Change in Seoul, South Korea under COVID-19 Social Distancing: Focusing on PM2.5. Int. J. Environ. Res. Public Heal. 2020, Vol. 17, Page 6208 17, 6208. https://doi.org/10.3390/IJERPH17176208

Han, F., Kota, S.H., Wang, Y., Zhang, H., 2017. Source apportionment of PM2.5 in Baton Rouge, Louisiana during 2009–2014. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2017.01.189

Hannigan, M.P., Busby, W.F., Cass, G.R., 2005. Source contributions to the mutagenicity of urban particulate air pollution. J. Air Waste Manag. Assoc. 55, 399–410. https://doi.org/10.1080/10473289.2005.10464633

Heo, J.-B., Hopke, P.K., Yi, S.-M., 2009. Source apportionment of PM $_{2.5}$ in Seoul, Korea. Atmos. Chem. Phys. 9, 4957–4971. https://doi.org/10.5194/acp-9-4957-2009

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. https://doi.org/10.1080/10962247.2016.1140693

Hu, X., Zhang, Y., Ding, Z., Wang, T., Lian, H., Sun, Y., Wu, J., 2012. Bioaccessibility and health risk of arsenic and heavy metals (Cd, Co, Cr, Cu, Ni, Pb, Zn and Mn) in TSP and PM2.5 in Nanjing, China. Atmos. Environ. 57, 146–152. https://doi.org/10.1016/j.atmosenv.2012.04.056

Jang, H.N., Seo, Y.C., Lee, J.H., Hwang, K.W., Yoo, J.I., Sok, C.H., Kim, S.H., 2007. Formation of fine particles enriched by V and Ni from heavy oil combustion: Anthropogenic sources and drop-tube furnace experiments. Atmos. Environ. 41, 1053–1063. https://doi.org/10.1016/j.atmosenv.2006.09.011

Ju, M.J., Oh, J., Choi, Y.H., 2021. Changes in air pollution levels after COVID-19 outbreak in Korea. Sci. Total Environ. 750, 141521. https://doi.org/10.1016/J.SCITOTENV.2020.141521

Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., Amann, M., 2015. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087

Khillare, P.S., Sarkar, S., 2012. Airborne inhalable metals in residential areas of Delhi, India: Distribution, source apportionment and health risks. Atmos. Pollut. Res. 3, 46–54. https://doi.org/10.5094/APR.2012.004

Kim, I., Lee, K., Lee, S., Kim, S.D., 2019. Characteristics and health effects of PM2.5 emissions from various sources in Gwangju, South Korea. Sci. Total Environ. 696, 133890. https://doi.org/10.1016/j.scitotenv.2019.133890

Kim, I., Park, K., Lee, K.Y., Park, M., Lim, H., Shin, H., Kim, S.D., 2020. Application of various cytotoxic endpoints for the toxicity prioritization of fine dust (PM2.5) sources using a multi-criteria decision-making approach. Environ. Geochem. Health 42, 1775–1788. https://doi.org/10.1007/s10653-019-00469-2

Kim, K.H., Kabir, E., Kabir, S., 2015. A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005

Kim, S., Kim, T.Y., Yi, S.M., Heo, J., 2018. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. J. Environ. Manage. 214, 325–334. https://doi.org/10.1016/j.jenvman.2018.03.027

Kim, Y.P., Lee, G., 2018. Trend of Air Quality in Seoul: Policy and Science. Aerosol Air Qual. Res. 18, 2141–2156. https://doi.org/10.4209/AAQR.2018.03.0081

Korea Ministry of Environment, National Institute of Environmental Research, 2022. 2020 Annual Report of Intensive Air Quality Monitoring Station.

Kumar, A., Chauhan, A., Arora, S., Tripathi, A., Alghanem, S.M.S., Khan, K.A., Ghramh, H.A., Özdemir, A., Ansari, M.J., 2020. Chemical analysis of trace metal contamination in the air of industrial area of Gajraula (U.P), India. J. King Saud Univ. - Sci. 32, 1106–1110. https://doi.org/10.1016/j.jksus.2019.10.008

Kumar, N., Park, R.J., Jeong, J.I., Woo, J.H., Kim, Y., Johnson, J., Yarwood, G.,

Kang, S., Chun, S., Knipping, E., 2021. Contributions of international sources to PM2.5 in South Korea. Atmos. Environ. 261, 118542. https://doi.org/10.1016/J.ATMOSENV.2021.118542

Leogrande, S., Alessandrini, E.R., Stafoggia, M., Morabito, A., Nocioni, A., Ancona, C., Bisceglia, L., Mataloni, F., Giua, R., Mincuzzi, A., Minerba, S., Spagnolo, S., Pastore, T., Tanzarella, A., Assennato, G., Forastiere, F., 2019. Industrial air pollution and mortality in the Taranto area, Southern Italy: A difference-in-differences approach. Environ. Int. 132, 105030. https://doi.org/10.1016/j.envint.2019.105030

Li, H., Qian, X., Wang, Q., 2013. Heavy metals in atmospheric particulate matter: A comprehensive understanding is needed for monitoring and risk mitigation. Environ. Sci. Technol. 47, 13210–13211. https://doi.org/10.1021/es404751a

Liu, B., Wu, J., Zhang, J., Wang, L., Yang, J., Liang, D., Dai, Q., Bi, X., Feng, Y., Zhang, Y., Zhang, Q., 2017. Characterization and source apportionment of PM2.5 based on error estimation from EPA PMF 5.0 model at a medium city in China. Environ. Pollut. 222, 10–22. https://doi.org/10.1016/j.envpol.2017.01.005

Long, L., He, J., Yang, X., 2021. Characteristics, emission sources and health risk assessment of trace elements in size-segregated aerosols during haze and non-haze periods at Ningbo, China. Environ. Geochem. Health 1–19. https://doi.org/10.1007/s10653-020-00757-2

Lv, L., Chen, Y., Han, Y., Cui, M., Wei, P., Zheng, M., Hu, J., 2021. High-time-resolution PM2.5 source apportionment based on multi-model with organic tracers in Beijing during haze episodes. Sci. Total Environ. 772, 144766.

https://doi.org/10.1016/j.scitotenv.2020.144766

Moon, K.J., Han, J.S., Ghim, Y.S., Kim, Y.J., 2008. Source apportionment of fine carbonaceous particles by positive matrix factorization at Gosan background site in East Asia. Environ. Int. 34, 654–664. https://doi.org/10.1016/j.envint.2007.12.021

National Institute of Environmental Research, 2019. Korean exposure factors handbook. Republic of Korea.

Nazarenko, Y., Pal, D., Ariya, P.A., 2021. Air quality standards for the concentration of particulate matter 2.5, global descriptive analysis. Bull. World Health Organ. 99, 125–137. https://doi.org/10.2471/BLT.19.245704

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemom. Intell. Lab. Syst. 37, 23–35. https://doi.org/10.1016/S0169-7439(96)00044-5

Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126. https://doi.org/10.1002/env.3170050203

Park, M. Bin, Lee, T.J., Lee, E.S., Kim, D.S., 2019. Enhancing source identification of hourly PM2.5 data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). Atmos. Pollut. Res. 10, 1042–1059. https://doi.org/10.1016/j.apr.2019.01.013

Park, E.H., Heo, J., Kim, H., Yi, S.-M., 2020. Long term trends of chemical constituents and source contributions of PM2.5 in Seoul. Chemosphere 251, 126371. https://doi.org/10.1016/j.chemosphere.2020.126371

Park, E.S., Hopke, P.K., Kim, I., Tan, S., Spiegelman, C.H., 2018. Bayesian Spatial

Multivariate Receptor Modeling for Multisite Multipollutant Data. Technometrics 60, 306–318. https://doi.org/10.1080/00401706.2017.1366948

Park, E.S., Lee, E.K., Oh, M.S., 2021. Bayesian multivariate receptor modeling software: BNFA and bayesMRM. Chemom. Intell. Lab. Syst. 211, 104280. https://doi.org/10.1016/j.chemolab.2021.104280

Park, E.S., Oh, M.S., 2015. Robust Bayesian multivariate receptor modeling. Chemom. Intell. Lab. Syst. 149, 215–226. https://doi.org/10.1016/j.chemolab.2015.08.021

Polissar, A. V., Hopke, P.K., Harris, J.M., 2001. Source regions for atmospheric aerosol measured at Barrow, Alaska. Environ. Sci. Technol. 35, 4214–4226. https://doi.org/10.1021/es0107529

Riojas-Rodríguez, H., Da Silva, A.S., Texcalac-Sangrador, J.L., Moreno-Banda, G.L., 2016. Air pollution management and control in Latin America and the Caribbean: Implications for climate change. Rev. Panam. Salud Publica/Pan Am. J. Public Heal. 40, 150–159.

Samara, C., Kouimtzis, T., Tsitouridou, R., Kanias, G., Simeonov, V., 2003. Chemical mass balance source apportionment of PM10 in an industrialized urban area of Northern Greece. Atmos. Environ. 37, 41–54. https://doi.org/10.1016/S1352-2310(02)00772-0

Shende, P., Qureshi, A., 2022. Burden of diseases in fifty-three urban agglomerations of India due to particulate matter ($PM_{2.5}$) exposure. Environ. Eng. Res. 27, 210042–0. https://doi.org/10.4491/EER.2021.042

Shi, G.L., Liu, G.R., Tian, Y.Z., Zhou, X.Y., Peng, X., Feng, Y.C., 2014. Chemical characteristic and toxicity assessment of particle associated PAHs for the short-

term anthropogenic activity event: During the Chinese New Year's Festival in 2013. Sci. Total Environ. 482–483, 8–14. https://doi.org/10.1016/j.scitotenv.2014.02.107

Shi, X., Nenes, A., Xiao, Z., Song, S., Yu, H., Shi, G., Zhao, Q., Chen, K., Feng, Y., Russell, A.G., 2019. High-Resolution Data Sets Unravel the Effects of Sources and Meteorological Conditions on Nitrate and Its Gas-Particle Partitioning. Environ. Sci. Technol. 53, 3048–3057. https://doi.org/10.1021/acs.est.8b06524

Shiraiwa, M., Ueda, K., Pozzer, A., Lammel, G., Kampf, C.J., Fushimi, A., Enami, S., Arangio, A.M., Fröhlich-Nowoisky, J., Fujitani, Y., Furuyama, A., Lakey, P.S.J., Lelieveld, J., Lucas, K., Morino, Y., Pöschl, U., Takahama, S., Takami, A., Tong, H., Weber, B., Yoshino, A., Sato, K., 2017. Aerosol Health Effects from Molecular to Global Scales. Environ. Sci. Technol. 51, 13545–13567. https://doi.org/10.1021/ACS.EST.7B04417

Silva, A.V., Oliveira, C.M., Canha, N., Miranda, A.I., Almeida, S.M., 2020. Long-Term Assessment of Air Quality and Identification of Aerosol Sources at Setúbal, Portugal. Int. J. Environ. Res. Public Health 17, 1–23. https://doi.org/10.3390/IJERPH17155447

Thorpe, A., Harrison, R.M., 2008. Sources and properties of non-exhaust particulate matter from road traffic: A review. Sci. Total Environ. 400, 270–282. https://doi.org/10.1016/j.scitotenv.2008.06.007

Tian, Y., Zhang, Y., Liang, Y., Niu, Z., Xue, Q., Feng, Y., 2020. PM2.5 source apportionment during severe haze episodes in a Chinese megacity based on a 5-month period by using hourly species measurements: Explore how to better conduct PMF during haze episodes. Atmos. Environ. 224, 117364.

https://doi.org/10.1016/j.atmosenv.2020.117364

Torkmahalleh, M.A., Yu, C.-H., Lin, L., Fan, Z. (Tina), Swift, J.L., Bonanno, L., Rasmussen, D.H., Holsen, T.M., Hopke, P.K., 2013. Improved Atmospheric Sampling of Hexavalent Chromium. J. Air Waste Manag. Assoc. 63, 1313.

US-EPA, 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide. Environ. Prot. Agency Off. Researc Dev. Publushing House Whashington, DC 20460 136.

US EPA, 2021. Regional Screening Levels (RSLs) Tables [WWW Document]. URL https://www.epa.gov/risk/regional-screening-levels-rsls-generic-tables (accessed 8.18.21).

US EPA, 2013. Users' guide and background technical document for US EPA region 9's preliminary remediation goals (PRG) table [WWW Document]. URL https://semspub.epa.gov/work/02/103453.pdf (accessed 8.18.21).

US EPA, 2009. Risk Assessment Guidance for Superfund Volume I: Human Health Evaluation Manual (Part F, Supplemental Guidance for Inhalation Risk Assessment). Off. Superfund Remediat. Technol. Innov. Environ. Prot. Agency I, 1–68.

Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., Yu, J.Z., 2018. Source Apportionment of PM2.5 Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-Resolution Influence. J. Geophys. Res. Atmos. 123, 5284–5300. https://doi.org/10.1029/2017JD027877

Wang, S., Ji, Y., Zhao, J., Lin, Y., Lin, Z., 2020. Source apportionment and toxicity assessment of PM2.5-bound PAHs in a typical iron-steel industry city in

northeast China by PMF-ILCR. Sci. Total Environ. 713, 136428. https://doi.org/10.1016/j.scitotenv.2019.136428

Wang, Y., Hopke, P.K., Xia, X., Rattigan, O. V., Chalupa, D.C., Utell, M.J., 2012. Source apportionment of airborne particulate matter using inorganic and organic species as tracers. Atmos. Environ. 55, 525–532. https://doi.org/10.1016/j.atmosenv.2012.03.073

Warner, M.S.C., 2018. Introduction to PySPLIT: A python toolkit for NOAA ARL's HYSPLIT model. Comput. Sci. Eng. 20, 47–62. https://doi.org/10.1109/MCSE.2017.3301549

WHO, 2005. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005 1–21. https://doi.org/10.1016/0004-6981(88)90109-6

Widziewicz, K., Rogula-Kozłowska, W., Loska, K., 2016. Cancer risk from arsenic and chromium species bound to PM2.5 and PM1 – Polish case study. Atmos. Pollut. Res. 7, 884–894. https://doi.org/10.1016/J.APR.2016.05.002

Wu, X., Vu, T. V., Shi, Z., Harrison, R.M., Liu, D., Cen, K., 2018. Characterization and source apportionment of carbonaceous PM2.5 particles in China - A review. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2018.06.025

Yang, L., Cheng, S., Wang, X., Nie, W., Xu, P., Gao, X., Yuan, C., Wang, W., 2013. Source identification and health impact of PM2.5 in a heavily polluted urban atmosphere in China. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2013.04.058

Zhang, L., Xu, H., Fang, B., Wang, H., Yang, Z., Yang, W., Hao, Y., Wang, X., Wang, Q., Wang, M., 2020. Source Identification and Health Risk Assessment

of Polycyclic Aromatic Hydrocarbon-Enriched PM2.5 in Tangshan, China. Environ. Toxicol. Chem. 39, 458–467. https://doi.org/10.1002/etc.4618

Zhao, X., Liu, Y., Han, F., Touseef, B., Yue, Y., Guo, J., 2021. Source profile and health risk assessment of PM2.5 from coal-fired power plants in Fuxin, China. Environ. Sci. Pollut. Res. 28, 40151–40159. https://doi.org/10.1007/s11356-020-11378-8

Zhao, Z., Lv, S., Zhang, Y., Zhao, Q., Shen, L., Xu, S., Yu, J., Hou, J., Jin, C., 2019. Characteristics and source apportionment of PM2.5 in Jiaxing, China. Environ. Sci. Pollut. Res. 2019 268 26, 7497–7511. https://doi.org/10.1007/S11356-019-04205-2

Zong, Z., Wang, X., Tian, C., Chen, Y., Qu, L., Ji, L., Zhi, G., Li, J., Zhang, G., 2016. Source apportionment of PM 2.5 at a regional background site in North China using PMF linked with radiocarbon analysis: insight into the contribution of biomass burning. Atmos. Chem. Phys 16, 11249–11265. https://doi.org/10.5194/acp-16-11249-2016

Chapter 4. Feature extraction and prediction of $PM_{2.5}$ chemical constituents using machine learning models

4.1. Introduction

There is growing interest in the measurement, management, and reduction of $PM_{2.5}$ ever since there have been reports on adverse health effects of exposure to airborne $PM_{2.5}$ (particulate matter with a diameter of ≤ 2.5 μm) (Hopke et al., 2020; Kim et al., 2015; Lee et al., 2022). In recent years, the hourly mass concentration of $PM_{2.5}$ are measured in many countries, and these values are made available by the World Air Quality Index project (https://aqicn.org/data-platform/register/). In addition to determining the total mass concentration of $PM_{2.5}$, monitoring stations to determine real-time $PM_{2.5}$ chemical constituents with different characteristics in terms of origin, conversion, and health effects, are increasing globally (Park et al., 2019; Wang et al., 2018). Accordingly, the quantification of $PM_{2.5}$ chemical composition with high spatial and temporal resolutions is an area of active research (Hopke, 2016; Shi et al., 2019), and $PM_{2.5}$ compositional data are gradually acquiring the characteristics of big data. As of 2021, 10 national monitoring stations in South Korea could measure the mass concentration of $PM_{2.5}$ and its chemical constituents on an hourly basis in real-time.

Complete and reliable data are not always available despite the high cost and time required to obtain $PM_{2.5}$ chemical composition. Missing values are one of the most prevalent impediments to data interpretation, making the appropriate use of the data challenging (Khan and Hoque, 2020). For example, the data of $PM_{2.5}$ chemical constituents measured in Seoul, South Korea, had an average missing ratio

of 9.43% from 2018 to 2020 (Table 4.1). However, it has been challenging to impute missing values of the chemical constituents because of the complexity of the chemical composition of $PM_{2.5}$. Researchers have responded by employing various fragmentary methods, such as excluding samples with any constituent missing values or replacing them with mean values (Kim et al., 2018; Park et al., 2019; Shi et al., 2021). These methods can reduce the data accuracy and the reliability of the modeling results, such as for source apportionment, relying on such input data.

Prediction of $PM_{2.5}$ components may be appropriate to attempt with nonlinear regression models because of their complexity(Baker and Foley, 2011). For nonlinear regression modeling of complex data, deep neural network (DNN) works excellently and has been widely used in many fields such as computer vision, behavior prediction, language process, and marketing to extract useful features from datasets(Jordan and Mitchell, 2015). However, little attention has been posed to predict $PM_{2.5}$ components using DNN models because DNN has recently begun to attract attention in the field of atmospheric environment(Gil et al., 2021).

Therefore, this study aimed to evaluate the applicability of the feature extraction using machine learning models to predict the chemical composition of $PM_{2.5}$. Four ML models were employed in this study: generative adversarial imputation network (GAIN), fully connected deep neural network (FCDNN), RF, and k-nearest neighbor (kNN). The prediction accuracy of each model was compared to evaluate the applicability of the models according to the stepwise increase of input data and changes in targeted components for prediction. Additionally, the effect of missing ratios and the available period of input data on prediction accuracy by models were examined. The present study findings can help expand the scope of ML

model-based interpretation of air pollution.

Table 4.1. Missing ratio and median values of $PM_{2.5}$ chemical speciation data (2018-2020)

Species	Missing ratio* (%)			Median (ng m^{-3})		
	BR	Seoul	Ulsan	BR	Seoul	Ulsan
$PM_{2.5}$	5.26	1.88	2.45	14,000	19,000	13,000
SO_4^{2-}	18.43	9.18	12.14	2,840	2,300	2,660
NO_3^-	18.43	9.18	12.14	1,070	2,370	1,670
Cl^-	18.7	9.26	12.37	140	140	200
Na^+	18.43	9.94	12.62	140	30	80
NH_4^+	18.43	9.37	12.21	1,340	1,840	2,030
K^+	18.43	11.44	24.25	70	50	40
Mg^{2+}	18.43	9.42	12.66	10	10	10
Ca^{2+}	18.48	9.8	13.39	30	30	20
OC	21.95	10.14	11.65	1,510	2,855	2,140
EC	21.95	10.23	11.67	358	730	420
S	10.74	9.16	4.07	1,704	1,548	4,296
K	10.76	9.22	4.07	80	80	80
Ca	10.77	9.2	4.09	32	43	32
Ti	10.78	9.17	4.07	6	6	6
V	10.77	9.16	4.07	2	2	2
Cr	10.75	9.17	4.07	1	1	1
Mn	10.77	9.16	4.07	5	7	10
Fe	10.75	9.17	4.07	86	148	140
Ni	13.44	9.17	4.07	0.90	0.40	0.84
Cu	16.46	9.27	4.07	4.76	5.17	4.94
Zn	10.78	9.27	4.07	19.6	31.1	35.3
As	10.97	9.16	4.07	2.14	2.23	1.86
Se	13.43	9.16	4.07	0.6	0.6	0.6
Br	10.75	9.17	4.07	3.43	4.42	5.53
Pb	10.81	9.16	4.07	8.26	12.63	9.51
Average (except $PM_{2.5}$)	14.58	9.43	7.85	-	-	-

* Total n = 26,305 at respective site

4.2. Materials and methods

4.2.1. Study Sites and Data Collection

The mass concentrations and chemical constituents of $PM_{2.5}$ are measured at 1-h intervals by the Air Quality Research Centers (Korea Ministry of Environment and National Institute of Environmental Research, 2021), which are operated by The Korean Ministry of Environment. The data used in this study were measured at Baengnyeong Island (BR, 37°57'52.9"N, 124°38'02.4"E), Seoul (Seoul, 37°36'35.3"N, 126°56'05.3"E), and Youngnam (Ulsan, 35°34'52.0"N, 129°19'27.0"E) from 2018 to 2020, and represent remote, metropolitan, and industrial areas, respectively (Fig. 4.1).

Mass concentrations of $PM_{2.5}$ were measured using BAM 1020 (Continuous Particulate Monitor by Met One Instruments, Inc., USA) employing the β-ray absorption method. Organic carbon (OC) and elemental carbon (EC) were measured by SOCEC (South Orange County Economic Coalition's Sunset Laboratory Inc., USA) using the thermal-optical transmittance method. Ionic species (NO_3^-, SO_4^{2-}, Cl^-, K^+, and NH_4^+) were measured using URG-9000D (Ambient Ion Monitor by Thermo Fisher Scientific Corp., USA) employing the ion chromatography analysis method. $PM_{2.5}$ elemental species concentrations were measured using Xact™ 620 (Ambient Trace Elements Monitor by Cooper Environmental Services [CES], USA) (S. S. Park et al., 2014) via X-ray fluorescence spectrometry (XRF), a non-destructive analysis method. The guideline for the installation and operation of the national air pollution monitoring network includes quality assurance/quality control

(QA/QC) for $PM_{2.5}$ component analysis (Korea Ministry of Environment and National Institute of Environmental Research, 2022).

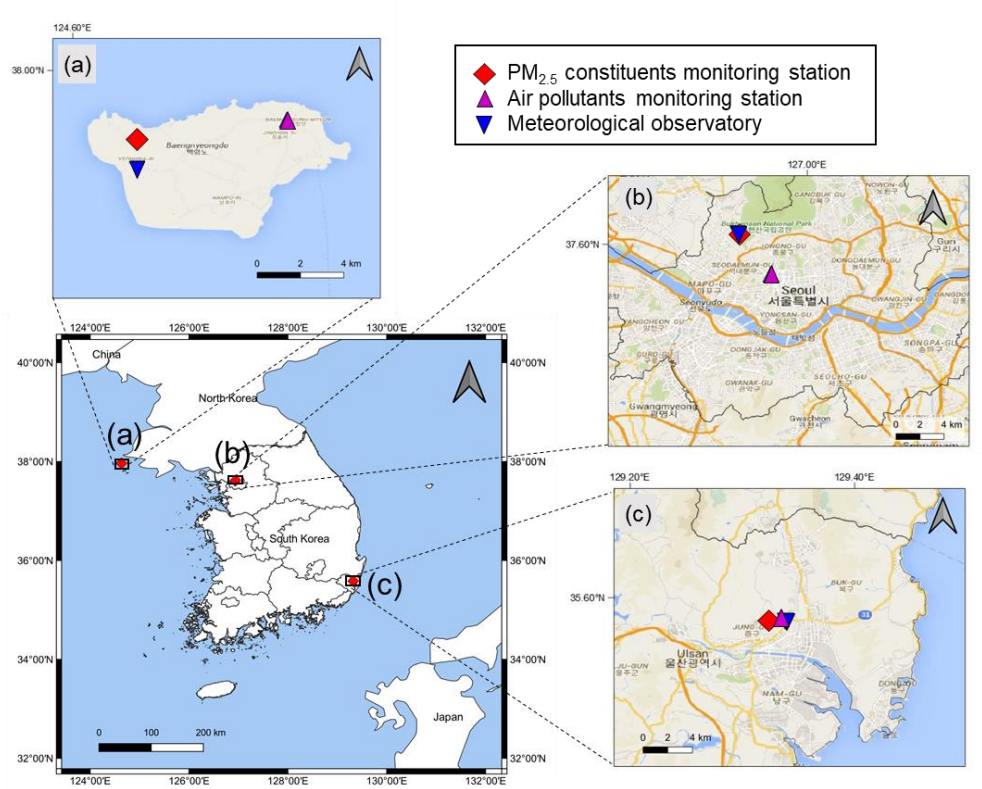


Fig. 4.1. Study sites: (a) Baengnyeong (BR), (b) Seoul, and (C) Ulsan

The chemical species (CS) of $PM_{2.5}$ were used as part of the input data to predict missing values using four ML models (Table 4.2). Table 4.1 provides a list of the types, total number, missing ratio, and median of the chemical composition data used. Additionally, three groups of input data were used for feature extraction: time information (TI), air pollutants (AP), and meteorological data (MD) (Table 4.2). TI included the hour, month, and weekday of $PM_{2.5}$ constituent data. The hourly concentrations of AP (i.e., $PM_{2.5}$, PM_{10}, SO_2, CO, O_3, and NO_2) measured at the AP national monitoring station closest to each $PM_{2.5}$ component monitoring station were

obtained from the AirKorea website. The national climate data center operated by Korea Meteorological Administration provided the MD measured at the automated synoptic observing system nearest to each $PM_{2.5}$ chemical constituents monitoring station. All input data were min-max normalized for each characteristic data prior to model training, and each parameter was then subjected to inverse normalization after modeling.

Table 4.2. Input variables

Classification	Variable
Chemical species (CS)	$PM_{2.5}$, Ion species (SO_4^{2-}, NO_3^-, Cl^-, Na^+, NH_4^+, K^+, Mg_2^+, Ca_2^+), Carbons (OC, EC), Trace elements (S, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Se, Br, Pb)
Time information (TI)	Weekdays, hours, months
Air pollutants (AP)	$PM_{2.5}$, PM_{10}, SO_2, CO, O_3, NO_2
Meteorological data (MD)	Temperature, rainfall, wind speed, wind direction, relative humidity, vapor, dew point, pressure, sunshine, snowfall, cloudiness, visibility

4.2.2. Machine Learning Models and Hyperparameter Optimization

Four ML models (GAIN, FCDNN, RF, and kNN), extensively used for regression analysis or missing value replacement, were applied to predict $PM_{2.5}$ chemical constituents. Two of the four ML models, GAIN and FCDNN, are further categorized into deep learning models, which use a complex structure of algorithms called multi-layered artificial neural networks. Additionally, GAIN and kNN are unsupervised learning models, whereas FCDNN and RF are supervised learning models requiring separate training and testing. All the models were implemented using Python 3.8 (Python Software Foundation, USA), while the input pipelines for the two deep learning models were built using Tensorflow 2.2 (Google Developers, USA). All codes used for the four ML models in this study are accessible (see Code Availability at the end of the manuscript).

The GAIN ML model is a missing-value processing model based on a generative adversarial network, in which a generator and discriminator compete to learn and improve accuracy (Li et al., 2019; Nazábal et al., 2020; Yoon et al., 2018). The discriminator is trained to accurately distinguish between real and fake data in a generated dataset, while the generator, in turn, learns to make it difficult for the discriminator to distinguish real from fake data (Yoon et al., 2018). In this study, the GAIN model was constructed as a long-short term memory (LSTM) network suitable for time series data. The model was separately trained and predicted on each division, and the results were then concatenated after dividing the data into 10-day period datasets. The hyperparameter settings that achieved the highest accuracies were found by manual search. Table 4.3 provides a list of the hyperparameter search ranges and optimized values.

Table 4.3. Hyperparameter searching range and optimized values

Model used	Hyperparameter searching range	Optimized Hyperparameter
GAIN	Number of hidden layers: 0 – 4	Number of hidden layers: 2
	Number of units in a layer: 500 – 400	Number of units in a layer: 52 – 200
	Learning rate: 1E-04 – 1E-02	Learning rate: 5E-04
	Hint rate: 0.7 – 0.9	Hint rate: 0.8
	Sequence: 120 – 720	Sequence: 240
	alpha: 10 – 100	alpha: 10
FCDNN	Activation function: ReLU, tanh, LeakyReLU (alpha=0.1)	Activation function: LeakyReLU (alpha=0.1)
	Number of hidden layers: 2 – 20	Number of hidden layers: 4 – 8
	Number of units in a layer: 32 – 2,048 (increment: 32)	Number of units in a layer: 1,300 – 2,000
	Learning rate: 1E-06 – 1E-02	Learning rate: 5E-05 – 1E-04
	Dropout rate: 0.10 – 0.20 (increment: 0.01)	Dropout rate: 0.10 – 0.15
RF	n_estimators: 1 – 2,000	n_estimators: 1,300 – 2,000
	max_depth: 1 – 30	max_depth: 13 – 30
	min_samples_leaf: 1 – 30	min_samples_leaf: 1 – 2
	min_samples_split: 2 – 30	min_samples_split: 2 – 4

kNN	k: 2 – 20	k: 3

The FCDNN is specialized in reducing dimensionality and performing feature extraction through hidden layers and is one of the most widely used neural network models in regression (Hinton and Salakhutdinov, 2006; Hwangbo et al., 2021). FCDNN model is trained by adjusting the weights and biases of the hidden layer neurons to correspond to each input and output data. Overfitting avoidance and optimization of hyperparameters are important for FCDNN models to have high prediction accuracy not only with training data but also with actual application data (Montavon et al., 2018). In this study, the latest technique for auto-optimization of hyperparameters, Keras-tuner (Asim et al., 2021), was used, and the hyperparameters with the highest R^2 were derived after more than 100 repetitions using both Hyperband and Bayesian search. The search and optimized ranges of hyperparameters are listed in Table 4.3. The number of training epochs was 200.

RF is a widely employed ensemble model for multi-dimensional classification and regression problems (Breiman, 2001). Various decision trees in RF models are trained using input data for feature extraction, which helps enhance model performance (Tella et al., 2021). The hyperparameters of the RF model were automatically optimized using the Hyperopt module (Bergstra et al., 2015). In this study, RF modeling used RandomForestRegressor in the scikit-learn package (Pedregosa et al., 2011).

kNN is a non-parametric model for classification and regression, wherein the prediction object is calculated as the average of k values closest to the prediction point (Tella et al., 2021; Yao and Ruzzo, 2006). Euclidean distance for the judgment of the nearest neighbor is used to achieve distance calculations in kNN. Through a

preliminary analysis adjusting k from 2 to 20, it was set at 3, which produced the highest prediction accuracy (Table 4.3). KNNImputer in the scikit-learn package was used for kNN modeling calculation.

4.2.3. Prediction Scenarios

Two scenarios were applied to compare the prediction accuracy of $PM_{2.5}$ constituents by four ML models, as described in Table 4.4. In Scenario #1, stepwise increase in four groups of input data (ID) and seven combinations of the prediction target component (PC) were applied. The ID groups were categorized from ID#1 to ID#4, wherein the larger the number, the more input data were used for prediction, starting with more accessible variables. The three study sites in Scenario #1 used three years (2018–2020) of hourly data with a fixed missing ratio of 20%.

In Scenario #2, four periods (1-month, 3-month, 12-month, and 36-month) with four missing ratios (20%, 40%, 60%, and 80%) were applied to the Seoul site to compare the prediction accuracies by four ML models according to the changes in the period and the missing ratio of ID. The missing ratios were determined by referring to the missing ratio of the actual data. In Scenario #2, the same four ML models as in Scenario #1 were applied; however, only ID#4 and PC#7 were used in Scenario #2 (Table 4.4). The number of iterations, n, was set to check the variations in prediction results. For Scenarios #1 and #2, a total of 2,400 model predictions were made. the difference in prediction accuracy between the ID periods identified one-way ANOVA with Tukey's honestly significant difference (HSD) test.

Table 4.4. Scenarios used for the prediction of $PM_{2.5}$ chemical composition

Classification	Scenario #1		Scenario #2	
	No. of case	Case	No. of case	Case
Period	1	3 years (2016–2018)	4	1 month (2018.12), 3 months (2018.10–12), 12 months (2018) 36 months (2016–2018)
Missing ratio (%)	1	20	4	20, 40, 60, 80
Input data	4	ID#1: CS; ID#2: CS and TI; ID#3: CS, TI, and AP; ID#4: CS, TI, AP, and MD	1	ID#4
Location	3	Baengnyeong (BR), Seoul, Ulsan	1	Seoul
Model	4	GAIN, FCDNN, RF, kNN	4	GAIN, FCDNN, RF, kNN
Prediction components	7	PC#1: ions; PC#2: carbons; PC#3: trace elements; PC#4: ions and carbons; PC#5: ions and trace elements; PC#6: carbons and trace elements; PC#7: ions, carbons, and trace elements	1	PC#7
Iteration	6	-	6	-
Total number of predictions	2,016		384	

4.2.4. Model Validation and Error Estimation

Fixed seed numbers (322, 777, and 1,004) were intentionally used following the randomized sampling methods in Pandas to ensure the reproducibility of the modeling results. There was no data duplication for the training and test. Model training for FCDNN and RF was performed using 80% of the entire data. The remaining 20% of the isolated data were compared with the prediction results. Comparing the observed values (isolated test data) with the predicted values allowed for model validation and error estimation. The coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) were used for error estimation. These values are commonly used indices for verifying the accuracy of regression models. Their formulas and mathematical backgrounds can be found in the literature (Chicco et al., 2021). The main text presents the R^2 value, the most insightful error estimation parameter of the three (Chicco et al., 2021), and compares it to the other indices to demonstrate the accuracy of the predictions made by the four ML models.

4.3. Results and discussion

4.3.1. Hyperparameter Optimization

Table 4.3 provides a list of the four ML models' optimized hyperparameters. The derived hyperparameters are expressed as ranges because there were variations in the hyperparameter results depending on prediction constituents and iterations. The number of hidden layers for deep learning models had optimal ranges with only a single-digit value. 4–8 hidden layers in FCDNN models and 2 hidden layers in GAIN were derived as optimized hyperparameters, as shown in Table 4.3. Similarly, a previous study predicted air quality response to emission changes using a convolutional neural network with three hidden layers (Xing et al., 2020). An ensemble model developed with CMAQ predictions using the FCDNN model applied four hidden layers (Lyu et al., 2019). It implies that single-digit hidden layers can lead to acceptable prediction accuracy in deep learning models when multi-AP data are used as ID.

The optimized hyperparameters (e.g., the learning rate of 5E-05–1E-04; dropout rate of 0.10–0.15; the number of units of 1,300–2,000 for the FCDNN model) derived for the four ML models in this study (Table 4.3) can be used for starting points for designing an AP prediction model. These values, however, are not absolute standards, and for better prediction results, the hyperparameters must be independently optimized based on the prediction target and available data.

4.3.2. Prediction Results for Scenario #1

Table 4.5 shows the prediction accuracies of $PM_{2.5}$ constituents for all prediction cases of Scenario #1 at the Seoul site. The prediction results of BR and Ulsan sites were similar to those of Seoul, according to the stepwise increase of ID # and PC # (Table 4.6). Additionally, as shown in Tables S4 and S5, respectively, the trend of the RMSE and MAE results reflected that of R2. The following comparison of prediction outcomes uses only R^2 values.

4.3.2.1. Overall prediction accuracy by the four ML models

The prediction accuracy of the four models for Scenario #1 varied from 0.071 to 0.947 in R^2 (Table 4.5). The highest R^2 was found in the case of predicting ions (PC#1) with ID#3 by the GAIN model, and the lowest in the case of predicting all components in $PM_{2.5}$ (PC#7) with ID#1 by the kNN model. Out of the seven PCs, the GAIN had the highest R^2 in the six PCs, and FCDNN had the highest R^2 in the one PC (PC#2) (marked by a superscript "a" in Table 4.5). The highest R^2 values for the seven PCs were greater than 0.875, which indicated that $PM_{2.5}$ chemical composition can be predicted with high accuracy using three-year CS, TI, air quality, and meteorological data with a 20% missing ratio.

These predicted accuracies can be compared indirectly with other studies that predicted missing values of $PM_{2.5}$ concentrations since there is no study on predicting missing values of $PM_{2.5}$ chemical components (Hadeed et al., 2020; Quinteros et al., 2019). Quinteros et al. (2019) predicted the missing values of $PM_{2.5}$ concentrations at monitoring stations in Chile with an accuracy of 0.37–0.91 of R^2. Hadeed et al. (2020) predicted missing values of short-term $PM_{2.5}$ measurements

(<24 h) in households with an accuracy of 0.32–0.65 of R^2 when the missing ratio was 20%. Additionally, the prediction accuracy in literature (Liu et al., 2019) that predicted trace elements in $PM_{2.5}$ through Weather Research and Forecasting (WRF) and CMAQ models in China was 0.35–0.91 in R (not R^2). If the highest prediction accuracy for each PC was selected in Table 4.5, R^2 range from 0.875 to 0.947, indicating considerably higher prediction accuracy than that of other studies.

The four ML models used in this study, as shown in Tables 3 and S6, appeared to be more accurate than an existing method for addressing missing values, which substitutes mean values for the missing values. The R^2 value between observation and prediction is calculated to be zero when the missing values in the same test data set are replaced by the mean values of each $PM_{2.5}$ component concentration (Table 4.9 and Fig. 4.2). Therefore, missing values can be more effectively compensated using the ML models by feature extraction from ID (Alpaydin, 2020). Additionally, expectation-maximization (EM) algorithm and multiple imputation (MI), which are utilized for completely random missing values in statistics, was used to compare the results of ML models (Fig. 4.7, Fig 4.8, and Fig. 4.9).

Table 4.5. Prediction accuracy (R^2) of PC#1 to PC#7 by four machine learning models for Seoul

	Coefficient of determination (R^2)						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.938	0.921	0.867	0.934	0.880	0.869	0.880
ID#2	0.939	0.921	0.866	0.935a	0.885	0.871	0.886
ID#3	0.947[a]	0.928	0.882	0.943	0.896[a]	0.885[a]	0.897[a]
ID#4	0.937	0.923	0.875[a]	0.934	0.895	0.880	0.895
FCDNN							
ID#1	0.898	0.936	0.808	0.898	0.417	0.791	0.403
ID#2	0.929	0.940	0.850	0.926	0.717	0.857	0.571
ID#3	0.933	0.943	0.856	0.934	0.859	0.865	0.832
ID#4	0.933	0.945a	0.860	0.933	0.869	0.867	0.861
RF							
ID#1	0.788[b]	0.897	0.725	0.803[b]	0.426	0.739	0.407
ID#2	0.822	0.902	0.765	0.830	0.644	0.763	0.549
ID#3	0.831	0.912	0.769	0.832	0.736	0.777	0.733
ID#4	0.839	0.912	0.782	0.834	0.773	0.789	0.785
kNN							
ID#1	0.812	0.899	0.702[b]	0.820	0.258[b]	0.709[b]	0.071[b]
ID#2	0.875	0.899	0.807	0.868	0.656	0.789	0.458
ID#3	0.902	0.915	0.833	0.900	0.817	0.832	0.801
ID#4	0.832	0.860[b]	0.746	0.831	0.748	0.747	0.744

* The standard deviation of all predicted values was within 5% and omitted for brevity.
** The number of datasets for train and test was 15,618 and 3,904 in ID#1 and ID#2; 14,602 and 3,516 in ID#3; 13,976 and 3,494 in ID#4.
[a] Values denote the largest R^2 in the respective PC
[b] Values denote the smallest R^2 in the respective PC

Table 4.6. Prediction accuracy (R^2) in BR, Seoul, and Ulsan by machine learning models

	BR*						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.904	0.940	0.896	0.907	0.893	0.899	0.886
ID#2	0.902	0.937	0.898	0.911	0.890	0.899	0.894
ID#3	0.919	0.948	0.914	0.925	0.915	0.918	0.912
ID#4	0.895	0.938	0.903	0.904	0.898	0.905	0.900
DNN							
ID#1	0.872	0.954	0.850	0.877	0.497	0.847	0.422
ID#2	0.884	0.956	0.884	0.896	0.761	0.882	0.610
ID#3	0.905	0.958	0.910	0.912	0.874	0.902	0.850
ID#4	0.891	0.958	0.904	0.890	0.879	0.896	0.875
RF							
ID#1	0.778	0.919	0.798	0.800	0.501	0.796	0.425
ID#2	0.810	0.932	0.823	0.830	0.704	0.834	0.612
ID#3	0.824	0.935	0.840	0.841	0.794	0.850	0.790
ID#4	0.805	0.934	0.842	0.842	0.821	0.857	0.824
kNN							
ID#1	0.805	0.922	0.786	0.818	0.405	0.747	0.143
ID#2	0.827	0.930	0.826	0.838	0.654	0.818	0.514
ID#3	0.875	0.932	0.863	0.883	0.824	0.863	0.805
ID#4	0.800	0.901	0.794	0.816	0.774	0.800	0.773

* The number of datasets for train and test of BR was 11,222 and 2,805 in ID#1 and ID#2; 10,547 and 2,636 in ID#3; 10,014 and 2,503 in ID#4.

Table 4.6. Prediction accuracy (R^2) in BR, Seoul, and Ulsan by machine learning models (continued)

	Seoul**						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.938	0.921	0.867	0.934	0.880	0.869	0.880
ID#2	0.939	0.921	0.866	0.935	0.885	0.871	0.886
ID#3	0.947	0.928	0.882	0.943	0.896	0.885	0.897
ID#4	0.937	0.923	0.875	0.934	0.895	0.880	0.895
DNN							
ID#1	0.898	0.936	0.808	0.898	0.417	0.791	0.403
ID#2	0.929	0.940	0.850	0.926	0.717	0.857	0.571
ID#3	0.933	0.943	0.856	0.934	0.859	0.865	0.832
ID#4	0.933	0.945	0.860	0.933	0.869	0.867	0.861
RF							
ID#1	0.788	0.897	0.725	0.803	0.426	0.739	0.407
ID#2	0.822	0.902	0.765	0.830	0.644	0.763	0.549
ID#3	0.831	0.912	0.769	0.832	0.736	0.777	0.733
ID#4	0.839	0.912	0.782	0.834	0.773	0.789	0.785
kNN							
ID#1	0.812	0.899	0.702	0.820	0.258	0.709	0.071
ID#2	0.875	0.899	0.807	0.868	0.656	0.789	0.458
ID#3	0.902	0.915	0.833	0.900	0.817	0.832	0.801
ID#4	0.832	0.860	0.746	0.831	0.748	0.747	0.744

** The number of datasets for train and test of Seoul was 15,618 and 3,904 in ID#1 and ID#2; 14,602 and 3,516 in ID#3; 13,976 and 3,494 in ID#4.

Table 4.6. Prediction accuracy (R^2) in BR, Seoul, and Ulsan by machine learning models (continued)

	Ulsan***						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.898	0.829	0.711	0.875	0.767	0.720	0.758
ID#2	0.896	0.822	0.713	0.879	0.771	0.727	0.771
ID#3	0.907	0.835	0.748	0.890	0.792	0.757	0.787
ID#4	0.893	0.825	0.689	0.881	0.759	0.708	0.760
DNN							
ID#1	0.852	0.856	0.623	0.834	0.312	0.625	0.284
ID#2	0.867	0.870	0.702	0.867	0.569	0.736	0.441
ID#3	0.891	0.864	0.754	0.890	0.758	0.763	0.750
ID#4	0.897	0.866	0.750	0.890	0.779	0.762	0.775
RF							
ID#1	0.753	0.803	0.597	0.762	0.320	0.596	0.290
ID#2	0.780	0.809	0.631	0.781	0.546	0.640	0.467
ID#3	0.793	0.817	0.691	0.773	0.651	0.687	0.663
ID#4	0.821	0.823	0.674	0.808	0.667	0.671	0.668
kNN							
ID#1	0.697	0.795	0.569	0.700	0.127	0.531	0.000
ID#2	0.740	0.820	0.685	0.732	0.526	0.667	0.362
ID#3	0.838	0.821	0.726	0.833	0.718	0.732	0.707
ID#4	0.753	0.757	0.593	0.751	0.621	0.599	0.623

*** The number of datasets for train and test of Ulsan was 14,065 and 3,516 in ID#1 and ID#2; 13,144 and 3,285 in ID#3; 12,328 and 3,082 in ID#4.
**** The standard deviation of all predicted values was within 5% and omitted for brevity.

Table 4.7. Prediction accuracy (RMSE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3)

	BR*						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.655	0.365	0.130	0.620	0.442	0.184	0.478
ID#2	0.666	0.368	0.131	0.616	0.491	0.186	0.462
ID#3	0.636	0.342	0.126	0.602	0.399	0.169	0.446
ID#4	0.724	0.372	0.131	0.664	0.455	0.181	0.454
DNN							
ID#1	0.675	0.302	0.099	0.656	0.834	0.175	0.978
ID#2	0.638	0.292	0.099	0.643	0.569	0.158	0.720
ID#3	0.548	0.287	0.091	0.503	0.400	0.155	0.442
ID#4	0.546	0.293	0.089	0.575	0.418	0.150	0.426
RF							
ID#1	0.869	0.409	0.118	0.792	0.850	0.200	0.969
ID#2	0.804	0.379	0.118	0.721	0.634	0.185	0.728
ID#3	0.734	0.377	0.116	0.678	0.541	0.183	0.551
ID#4	0.755	0.375	0.121	0.679	0.483	0.181	0.485
kNN							
ID#1	0.825	0.406	0.118	0.775	0.909	0.216	1.305
ID#2	0.847	0.394	0.136	0.805	0.729	0.218	0.822
ID#3	0.751	0.395	0.129	0.697	0.534	0.193	0.565
ID#4	0.929	0.489	0.181	0.881	0.649	0.251	0.658

* The number of datasets for train and test of BR was 11,222 and 2,805 in ID#1 a nd ID#2; 10,547 and 2,636 in ID#3; 10,014 and 2,503 in ID#4.

Table 4.7. Prediction accuracy (RMSE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3) (continued)

	Seoul**						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.502	0.408	0.090	0.486	0.334	0.170	0.352
ID#2	0.504	0.409	0.090	0.490	0.326	0.170	0.343
ID#3	0.476	0.393	0.087	0.459	0.320	0.162	0.323
ID#4	0.525	0.403	0.094	0.503	0.326	0.169	0.341
DNN							
ID#1	0.517	0.369	0.081	0.508	0.705	0.189	0.778
ID#2	0.453	0.349	0.070	0.446	0.464	0.158	0.585
ID#3	0.399	0.339	0.065	0.426	0.322	0.145	0.375
ID#4	0.407	0.328	0.068	0.391	0.293	0.137	0.320
RF							
ID#1	0.670	0.498	0.107	0.650	0.703	0.215	0.770
ID#2	0.593	0.481	0.092	0.591	0.488	0.201	0.582
ID#3	0.561	0.468	0.092	0.571	0.462	0.196	0.486
ID#4	0.556	0.457	0.091	0.570	0.423	0.191	0.423
kNN							
ID#1	0.682	0.491	0.105	0.657	0.779	0.226	0.898
ID#2	0.636	0.484	0.105	0.624	0.558	0.220	0.662
ID#3	0.578	0.446	0.103	0.559	0.418	0.193	0.442
ID#4	0.815	0.587	0.152	0.796	0.554	0.259	0.577

** The number of datasets for train and test of Seoul was 15,618 and 3,904 in ID# 1 and ID#2; 14,602 and 3,516 in ID#3; 13,976 and 3,494 in ID#4.

Table 4.7. Prediction accuracy (RMSE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3) (continued)

	Ulsan***						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
GAIN							
ID#1	0.506	0.497	0.322	0.527	0.431	0.355	0.456
ID#2	0.516	0.508	0.315	0.519	0.419	0.349	0.430
ID#3	0.458	0.512	0.288	0.484	0.383	0.324	0.405
ID#4	0.534	0.490	0.320	0.532	0.424	0.347	0.434
DNN							
ID#1	0.489	0.428	0.210	0.509	0.934	0.266	1.009
ID#2	0.486	0.409	0.196	0.465	0.533	0.243	0.634
ID#3	0.413	0.439	0.179	0.415	0.412	0.233	0.434
ID#4	0.377	0.409	0.177	0.396	0.347	0.224	0.371
RF							
ID#1	0.601	0.526	0.225	0.598	0.940	0.294	1.002
ID#2	0.569	0.518	0.220	0.571	0.567	0.280	0.640
ID#3	0.545	0.543	0.224	0.557	0.536	0.287	0.534
ID#4	0.513	0.489	0.231	0.527	0.482	0.277	0.496
kNN							
ID#1	0.661	0.538	0.239	0.664	1.060	0.311	1.329
ID#2	0.592	0.531	0.270	0.594	0.596	0.320	0.700
ID#3	0.556	0.550	0.284	0.557	0.478	0.329	0.503
ID#4	0.731	0.612	0.380	0.717	0.582	0.421	0.590

*** The number of datasets for train and test of Ulsan was 14,065 and 3,516 in ID#1 and ID#2; 13,144 and 3,285 in ID#3; 12,328 and 3,082 in ID#4.
***** The standard deviation of all predicted values was within 5% and omitted for brevity.

Table 4.8. Prediction accuracy (MAE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3)

MAE	BR*						
	PC#1	PC#2	PC#3	PC#4	PC#5	PC#6	PC#7
GAIN							
ID#1	0.220	0.198	0.026	0.220	0.100	0.048	0.117
ID#2	0.226	0.199	0.025	0.218	0.101	0.048	0.109
ID#3	0.201	0.182	0.024	0.199	0.089	0.043	0.103
ID#4	0.236	0.202	0.026	0.229	0.101	0.048	0.110
DNN							
ID#1	0.222	0.170	0.024	0.232	0.226	0.049	0.281
ID#2	0.217	0.166	0.022	0.239	0.151	0.046	0.213
ID#3	0.195	0.164	0.020	0.184	0.107	0.045	0.128
ID#4	0.186	0.168	0.019	0.205	0.100	0.042	0.113
RF							
ID#1	0.286	0.228	0.025	0.282	0.230	0.055	0.279
ID#2	0.258	0.208	0.024	0.254	0.160	0.049	0.209
ID#3	0.249	0.206	0.024	0.247	0.134	0.049	0.150
ID#4	0.255	0.204	0.025	0.240	0.121	0.048	0.133
kNN							
ID#1	0.250	0.210	0.025	0.256	0.249	0.059	0.374
ID#2	0.257	0.209	0.027	0.261	0.189	0.059	0.250
ID#3	0.222	0.199	0.025	0.225	0.120	0.049	0.144
ID#4	0.308	0.261	0.036	0.305	0.150	0.066	0.166

* The number of datasets for train and test of BR was 11,222 and 2,805 in ID#1 a nd ID#2; 10,547 and 2,636 in ID#3; 10,014 and 2,503 in ID#4.

Table 4.8. Prediction accuracy (MAE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3) (continued)

MAE	Seoul**						
	PC#1	PC#2	PC#3	PC#4	PC#5	PC#6	PC#7
GAIN							
ID#1	0.182	0.247	0.019	0.196	0.080	0.047	0.096
ID#2	0.183	0.247	0.019	0.197	0.079	0.047	0.094
ID#3	0.169	0.231	0.018	0.180	0.075	0.044	0.088
ID#4	0.184	0.239	0.020	0.196	0.078	0.046	0.093
DNN							
ID#1	0.198	0.227	0.019	0.212	0.205	0.054	0.254
ID#2	0.167	0.212	0.017	0.183	0.129	0.045	0.192
ID#3	0.151	0.202	0.016	0.175	0.087	0.042	0.114
ID#4	0.151	0.200	0.016	0.164	0.079	0.040	0.096
RF							
ID#1	0.261	0.309	0.025	0.279	0.203	0.062	0.250
ID#2	0.225	0.296	0.022	0.252	0.137	0.058	0.187
ID#3	0.216	0.285	0.022	0.247	0.126	0.056	0.149
ID#4	0.207	0.278	0.022	0.241	0.112	0.055	0.126
kNN							
ID#1	0.241	0.285	0.023	0.260	0.224	0.063	0.300
ID#2	0.223	0.284	0.023	0.250	0.153	0.063	0.219
ID#3	0.199	0.257	0.021	0.215	0.104	0.053	0.128
ID#4	0.285	0.344	0.031	0.304	0.133	0.071	0.157

** The number of datasets for train and test of Seoul was 15,618 and 3,904 in ID# 1 and ID#2; 14,602 and 3,516 in ID#3; 13,976 and 3,494 in ID#4.

Table 4.8. Prediction accuracy (MAE) in BR, Seoul, and Ulsan by machine learning model (unit: μg/m^3) (continued)

MAE	Ulsan***						
	PC#1	PC#2	PC#3	PC#4	PC#5	PC#6	PC#7
GAIN							
ID#1	0.197	0.273	0.059	0.217	0.114	0.086	0.133
ID#2	0.199	0.275	0.058	0.216	0.111	0.084	0.125
ID#3	0.174	0.257	0.051	0.195	0.099	0.077	0.115
ID#4	0.202	0.269	0.058	0.218	0.111	0.083	0.124
DNN							
ID#1	0.192	0.247	0.044	0.220	0.273	0.074	0.318
ID#2	0.198	0.232	0.041	0.208	0.164	0.066	0.214
ID#3	0.164	0.232	0.036	0.179	0.119	0.063	0.137
ID#4	0.151	0.234	0.037	0.176	0.102	0.062	0.117
RF							
ID#1	0.240	0.303	0.046	0.259	0.275	0.082	0.311
ID#2	0.225	0.297	0.044	0.247	0.175	0.077	0.215
ID#3	0.218	0.295	0.045	0.242	0.158	0.076	0.168
ID#4	0.200	0.280	0.045	0.227	0.142	0.076	0.159
kNN							
ID#1	0.248	0.300	0.049	0.271	0.308	0.087	0.415
ID#2	0.231	0.288	0.054	0.256	0.181	0.086	0.236
ID#3	0.204	0.274	0.052	0.222	0.128	0.079	0.148
ID#4	0.280	0.339	0.075	0.296	0.160	0.108	0.178

*** The number of datasets for train and test of Ulsan was 14,065 and 3,516 in ID#1 and ID#2; 13,144 and 3,285 in ID#3; 12,328 and 3,082 in ID#4.
***** The standard deviation of all predicted values was within 5% and omitted for brevity.

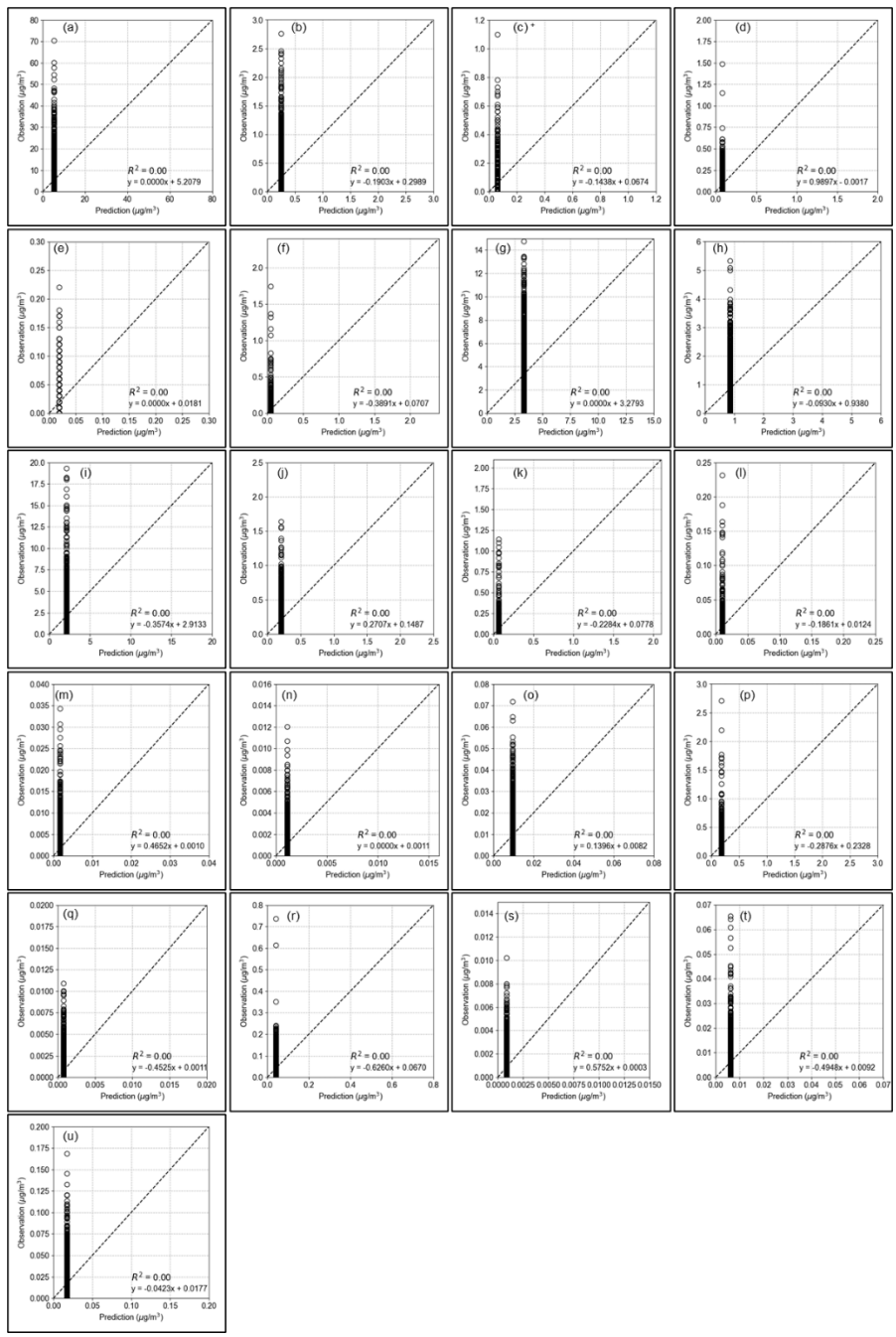


Fig. 4.2. Comparisons of observations and predictions of mean substitution in Seoul (ID#4, PC#7): (a) NO_3^-; (b) Cl^-; (c) Na^+; (d) K^+; (e) Mg^{2+}; (f) Ca^{2+}; (g) OC; (h) EC; (i) S; (j) K; (k) Ca; (l) Ti; (m) V; (n) Cr; (o) Mn; (p) Fe; (q) Ni; (r) Zn;

(s) Se; (t) Br; (u) Pb

Table 4.9. Prediction accuracy (R^2, RMSE, and MAE) of mean substitution

	Seoul						
	PC#1	**PC#2**	**PC#3**	**PC#4**	**PC#5**	**PC#6**	**PC#7**
R^2							
ID#1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ID#4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RMSE ($\mu g/m^3$)							
ID#1	2.926	1.587	0.528	2.712	1.778	0.737	1.763
ID#4	2.949	1.581	0.521	2.731	1.790	0.731	1.774
MAE ($\mu g/m^3$)							
ID#1	1.235	1.053	0.116	1.199	0.505	0.226	0.549
ID#4	1.238	1.056	0.115	1.201	0.506	0.226	0.550

4.3.2.2. Effect of input variables on prediction accuracy by models

Table 4.5 shows the improvement or change in R^2 by the stepwise increase of input variables. It indicates that the ML models can extract the features of the association between $PM_{2.5}$ constituent concentrations and input variables such as TI, AP, and MD. Thus, the influence of time, air quality, or meteorological conditions on the complex mechanism to form atmospheric $PM_{2.5}$ constituents can be explained by ML models. Notably, when only TI (time, day, month) was added in ID#1, resulting in ID#2, the R^2 increased in most prediction cases. The exceptions were PC#2 and PC#3 by GAIN and PC#2 by kNN (Table 4.5). Even in the case of predicting all chemical constituents in $PM_{2.5}$ (i.e., PC#7), the R^2 was improved by adding only TI (e.g., R^2 from 0.071 to 0.458 in kNN), implying that the patterns of $PM_{2.5}$ constituent concentrations are affected by time (hour), weekday, and season (month) and learned by the ML models. Additionally, the prediction accuracy was improved in ID#3, where AP is folded into the models. R^2 at ID#3 was higher than that at ID#1 or ID#2 in all PCs by the four ML models. In the case of PC#7, where all chemical components of $PM_{2.5}$ are targets for prediction, the R^2 by GAIN, FCDNN, RF, and KNN models increased from 0.880, 0.403, 0.407, and 0.071 using ID#1, respectively, to 0.897, 0.832, 0.733, and 0.801 using ID#3. The R^2 at ID#3 was higher than that at ID#1 or ID#2 in all PCs (Table 4.5) even though the number of data samples used as known (e.g., training data) in ID#3 (n=14,602) was smaller than that in ID#1 or ID#2 (n=15,618). It suggested that smaller numbers of sample data may be supported by more characteristic input variables associated with the prediction target.

Meanwhile, the R^2 by GAIN and kNN models decreased in ID#4, where MD is added to the models (Tables 3; S3). However, it was found that the differences in the R^2 by the kNN model between ID#4 and ID#3 were much larger than those by the GAIN model. More input variables resulting in lower R^2 may be explained by the cumulative addition of input variables in this study. The difference in prediction accuracy by adding MD to CS as the second input variable was not examined and MD was added as the fourth input variable. Moreover, the influence of meteorology on air quality is reflected to some extent in the concentrations of CS (ID#1), TI (ID#2), and AP (ID#3) and thus, the effect of MD on the prediction accuracy cannot be compared with that of TI or AP, and needs to be studied more in future works. In contrast, the sharp decrease in the R^2 by the kNN model by the addition of MD information may be because of the characteristics of kNN, which do not belong to the deep learning model. Since it uses the simplest ML method, adding too many features will make the prediction difficult. Thus, the kNN model appears to be affected by the so-called "curse of dimensionality (Poggio et al., 2017)." The application of deep learning models may be more appropriate for $PM_{2.5}$ composition prediction when many input variables (i.e., chemical compositions in $PM_{2.5}$, TI, gaseous AP, and MD) are used as in this study. As the number of input variables rises, deep learning models suitable for regression utilizing high-dimensional data perform better than simple ML models (Alpaydin, 2020; Gao et al., 2017). Compared to the two ML models (RF, kNN), the two deep learning models (GAIN, FCDNN) had higher R^2 for all PCs with ID#4, which had the most input variables (Tables 3; S3). Owing to the high dimensionality of the supporting data, this study also implies that deep learning models have exceptional applicability to data on air pollution.

4.3.2.3. Comparisons of prediction accuracy by targeted components

Prediction accuracy by respective species is presented using the GAIN model in Fig. 4.3 and Fig. 4.4. The R^2 for ion species was higher than that for trace elements when all components were predicted (PC#7). Similar trends by GAIN were observed by all ML models. In the comparison between predicted and observed values for NH_4^+ and SO_4^{2-}, the slope was approximately 1.0, and the R^2 values were 0.97, which was higher than those of As (0.87) and Cu (0.78) (Fig. 4.3). High R^2 was obtained for NO_3^-, SO_4^{2-}, and NH_4^+ with R^2 values of 0.97, and thus, the concentration of secondary aerosols may be effectively predicted from $PM_{2.5}$ concentration, TI, AP, and MD. It seems that the secondary aerosol reaction mechanism is learned by the deep learning model using the SO_2, NO_2, and O_3 concentrations and meteorological and time information even though secondary aerosols are engaged in extraordinarily complex reactions including diffusion in the atmosphere and chemical reactions depending on the weather and gaseous substance supply (Liu et al., 2022). Additionally, the high contribution of secondary aerosols to total $PM_{2.5}$ mass concentrations (Kim et al., 2018; E. H. Park et al., 2020) may help estimate their concentrations effectively from the $PM_{2.5}$ information.

The prediction accuracy was low when trace elements were included in the prediction targets in all ML models while comparing the results for each PC (Table 4.5). For example, when the trace elements were included in the prediction target (i.e., PC#3, PC#5, PC#6, and PC#7) with ID#4, the R^2 of their prediction results ranked from bottom to the fourth out of seven PC results. This is presumed to be because the concentration of trace elements accounts for a very small proportion of the total $PM_{2.5}$ concentration, and the characteristics supporting the prediction of the

concentration of trace elements were less included as input variables. Trace elements are more affected by emission sources than chemical reactions in the atmosphere (Choi et al., 2022); however, data related to emission sources were not folded into the models in this study.

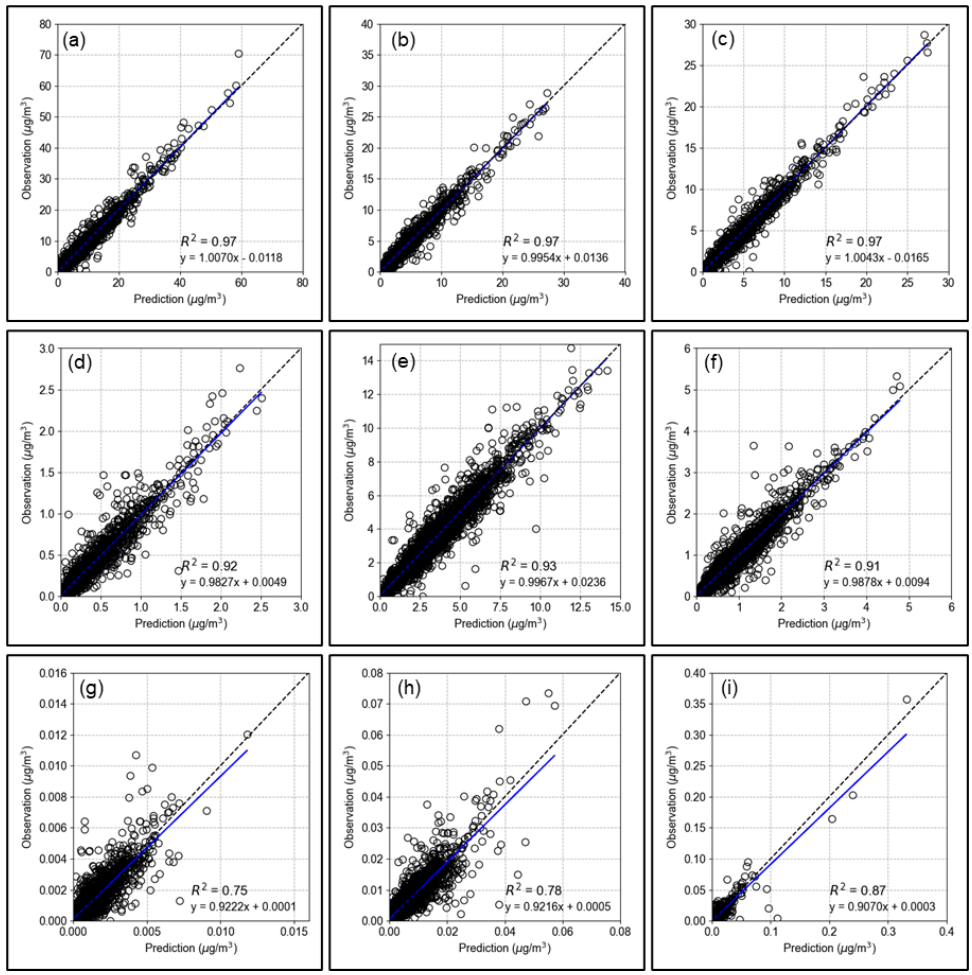


Fig. 4.3. Comparisons of observations and predictions by GAIN model prediction in Seoul (ID#4, PC#7): (a) NO_3^-, (b) SO_4^{2-}, (c) NH_4^+, (d) Cl^-, (e) OC, and (f) EC, (g) Cr, (h) Cu, and (i) As (No. of points = 3,494)

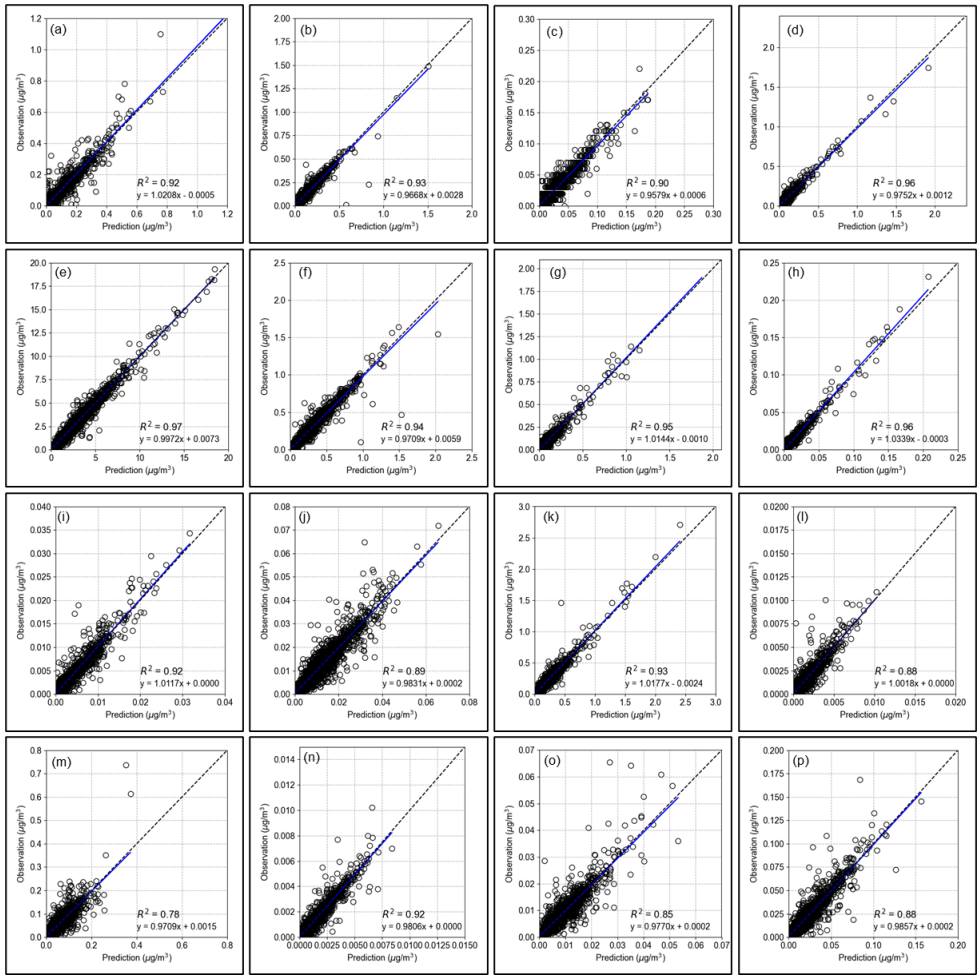


Fig. 4.4. Comparisons of observations and predictions by GAIN model in Seoul (ID#4, PC#7): (a) Na^+; (b) K^+; (c) Mg^{2+}; (d) Ca^{2+}; (e) S; (f) K; (g) Ca; (h) Ti; (i) V; (j) Mn; (k) Fe; (l) Ni; (m) Zn; (n) Se; (o) Br; (p) Pb

4.3.2.4. Variability in prediction accuracy of PC#6 among three cites

The variations in R^2 were characterized in three study sites, BR, Seoul, and Ulsan (Table 4.6; Fig.3). Fig. 4.5 shows the prediction results according to the ID combination and four ML models for PC#6 at the three sites. The R^2 decreased in the order of BR, Seoul, and Ulsan. The BR site with the highest R^2 has fewer anthropogenic emission sources, and the Ulsan site with the lowest R^2 has many anthropogenic sources. The emissions of $PM_{2.5}$ from industrial activity in 2019 were 77, 20,482, 1,197,173 kg in Baengnyeong, Seoul, and Ulsan, respectively (Air Pollutants Emission Inventory of Republic of Korea, 2019). The prediction accuracy was lower in cities with more $PM_{2.5}$ emission from industrial activity. The westernmost island in Korea, BR, is regarded as a remote place with the least impact from Korea's emission sources. Two industrial complexes are known to have a direct impact on Ulsan, a significant industrial city (Choi et al., 2011b; Lee and Hieu, 2011). It is inferred that lower R^2 in Ulsan in comparison to that in Seoul and BR for PC#6 corresponds with lower prediction accuracy for the trace element group of PC#3 (Table 4.6). The question of whether prediction accuracy for trace elements and/or in Ulsan may be increased by using additional input variables related to emission data from Ulsan industrial complexes remains to be explored in future investigation.

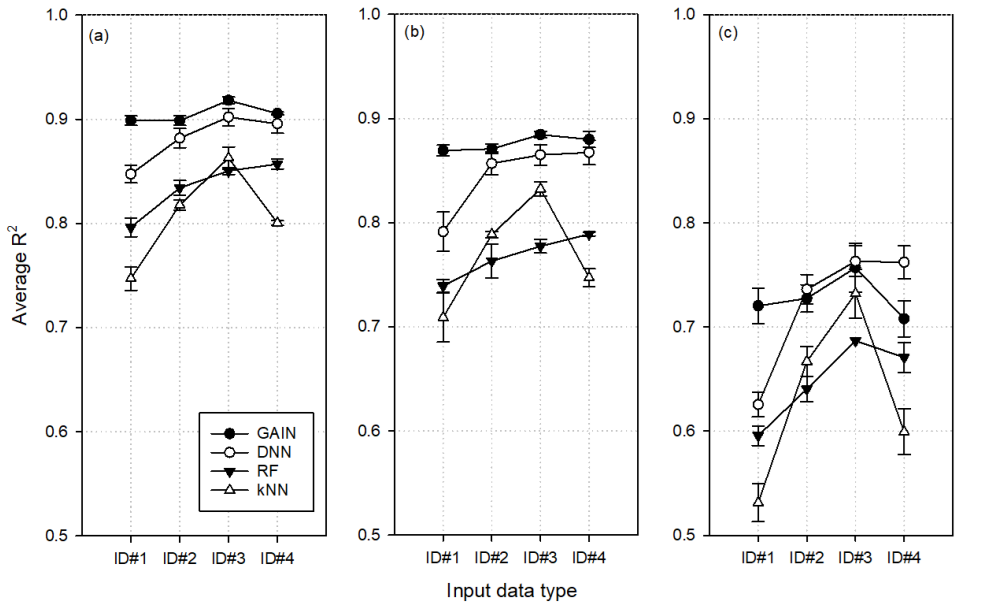


Fig. 4.5. Comparison of model accuracy by ID# (PC#6): (a) BR, (b) Seoul, and (c) Ulsan

4.3.3. Prediction Results for Scenario #2

In Fig. 4.6 and Table 4.10, the prediction accuracy for each model for the Seoul location is displayed with variations in the ID period (from 1 to 36 months) and missing ratios (from 20% to 80%). First, the R^2 was more than 0.8 by the two deep learning models at a missing ratio of 20% with data over 3 months, suggesting that a few months' data are sufficient to apply deep learning models in predicting the concentrations of $PM_{2.5}$ components. Second, when the period of ID increased, the R^2 of the deep learning models further increased. By contrast, RF and kNN did not show improvements in R^2 as the period of the ID increased than those of GAIN and FCDNN (Fig. 4.6). One-way ANOVA with Tukey's honestly significant difference (HSD) test identified that the longer the period of ID, the more the significant differences in prediction results between the models (Table 4.11). This shows that even though a substantial amount of data is used in this study, the two deep learning models used can successfully extract the features of data, as shown in earlier research. (Ciaburro and Iannace, 2021).

Table 4.10. Prediction accuracy (R^2) by model according to data input period and missing ratio (ID#4 and PC#7, Seoul)

Missing ratio	GAIN				DNN				RF				kNN			
	1-month	3-month	12-month	36-month	1-month	3-month	12-month	36-month	1-month	3-month	12-month	36-month	1-month	3-month	12-month	36-month
0.2	0.851	0.854	0.879	0.895	0.791	0.818	0.824	0.861	0.808	0.795	0.773	0.785	0.791	0.763	0.713	0.744
0.4	0.783	0.813	0.829	0.844	0.744	0.764	0.789	0.822	0.767	0.768	0.749	0.759	0.703	0.722	0.666	0.678
0.6	0.742	0.727	0.754	0.784	0.744	0.736	0.758	0.767	0.742	0.730	0.720	0.714	0.661	0.664	0.612	0.619
0.8	0.596	0.626	0.631	0.686	0.643	0.660	0.675	0.699	0.652	0.659	0.659	0.641	0.551	0.530	0.493	0.514

Table 4.11. One-way ANOVA with Tukey's honestly significant difference (HSD) test results among models according to data input period (ID#4, PC#7, Seoul, missing ratio 0.2)

Model		P-value			
		Input data period			
		1-month	3-month	12-month	36-month
GAIN	FCDNN	<0.001*	0.041*	0.003*	<0.001*
GAIN	RF	0.003*	0.002*	<0.001*	<0.001*
GAIN	kNN	<0.001*	<0.001*	<0.001*	<0.001*
FCDNN	RF	0.227	0.204	0.005*	<0.001*
FCDNN	kNN	0.900	0.005*	<0.001*	<0.001*
RF	kNN	0.227	0.041*	0.002*	<0.001*

* Significantly different (significance level of 0.05)

Prediction accuracy decreased as the missing ratio increased in all models (Table 4.10). This was anticipated because the learnable data itself decreased. However, notably, the decrease in accuracy was larger for GAIN and kNN as the missing ratio increased, compared to the other two models (Fig. 4.6 and Table 4.10). Since GAIN and kNN are unsupervised learning models that have the ability to create predictions that are plausible, it may be more challenging for these models to estimate accurate values in situations when there are inadequate reference data. In contrast, FCDNN and RF, which are supervised learning models whose training and testing are distinguished within the given data, were relatively less sensitive to the increases in the missing ratio. When the missing ratio was 80%, the FCDNN model had a higher prediction accuracy than that of the GAIN model in all periods (Fig. 4.6 and Table 4.10).

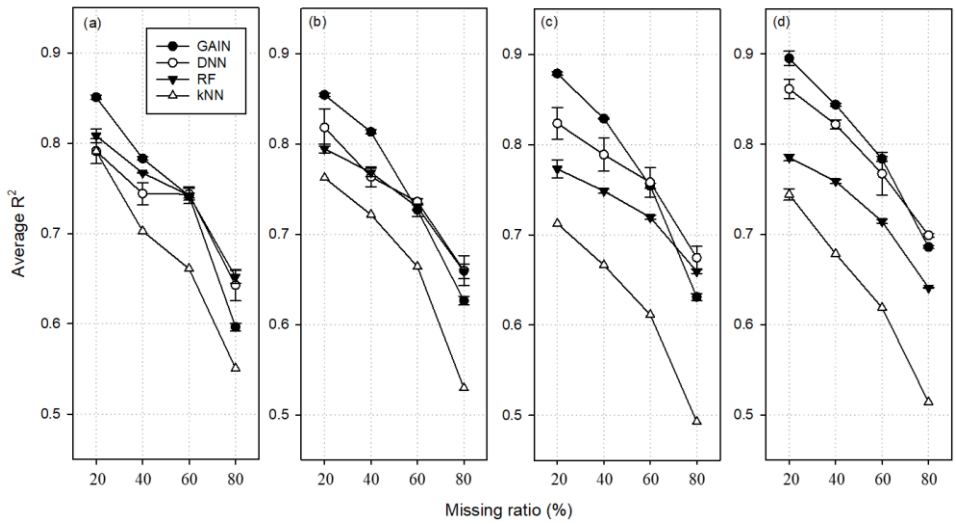


Fig. 4.6. Comparison of accuracy by model according to data input period and missing ratio (ID#4 and PC#7, Seoul): (a) 1-month, (b) 3-month, (c) 12-month, and (d) 36-month data usage

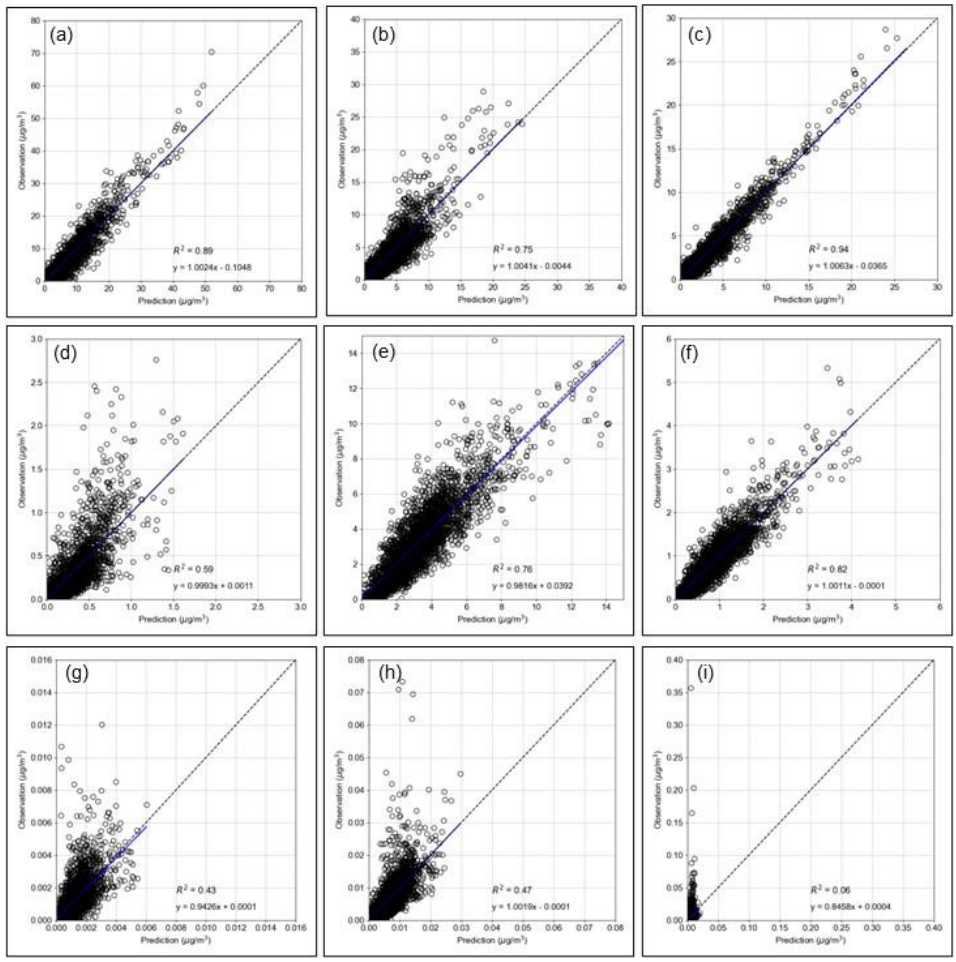


Fig. 4.7. Comparisons of observations and predictions by EM model prediction in Seoul (ID#4, PC#7): (a) NO_3^-, (b) SO_4^{2-}, (c) NH_4^+, (d) Cl^-, (e) OC, (f) EC, (g) Cr, (h) Cu, and (i) As (No. of points = 3,494)

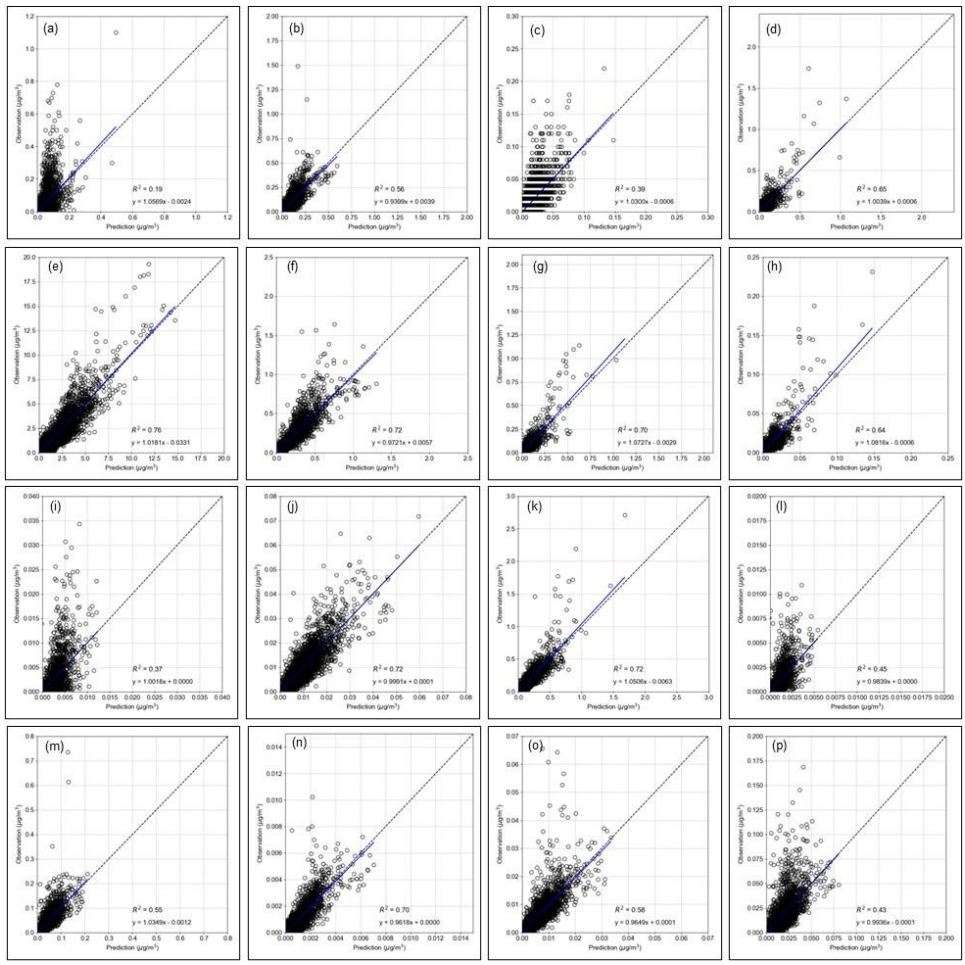


Fig. 4.8. Comparisons of observations and predictions by EM model in Seoul (ID#4, PC#7): (a) Na^+; (b) K^+; (c) Mg^{2+}; (d) Ca^{2+}; (e) S; (f) K; (g) Ca; (h) Ti; (i) V; (j) Mn; (k) Fe; (l) Ni; (m) Zn; (n) Se; (o) Br; (p) Pb

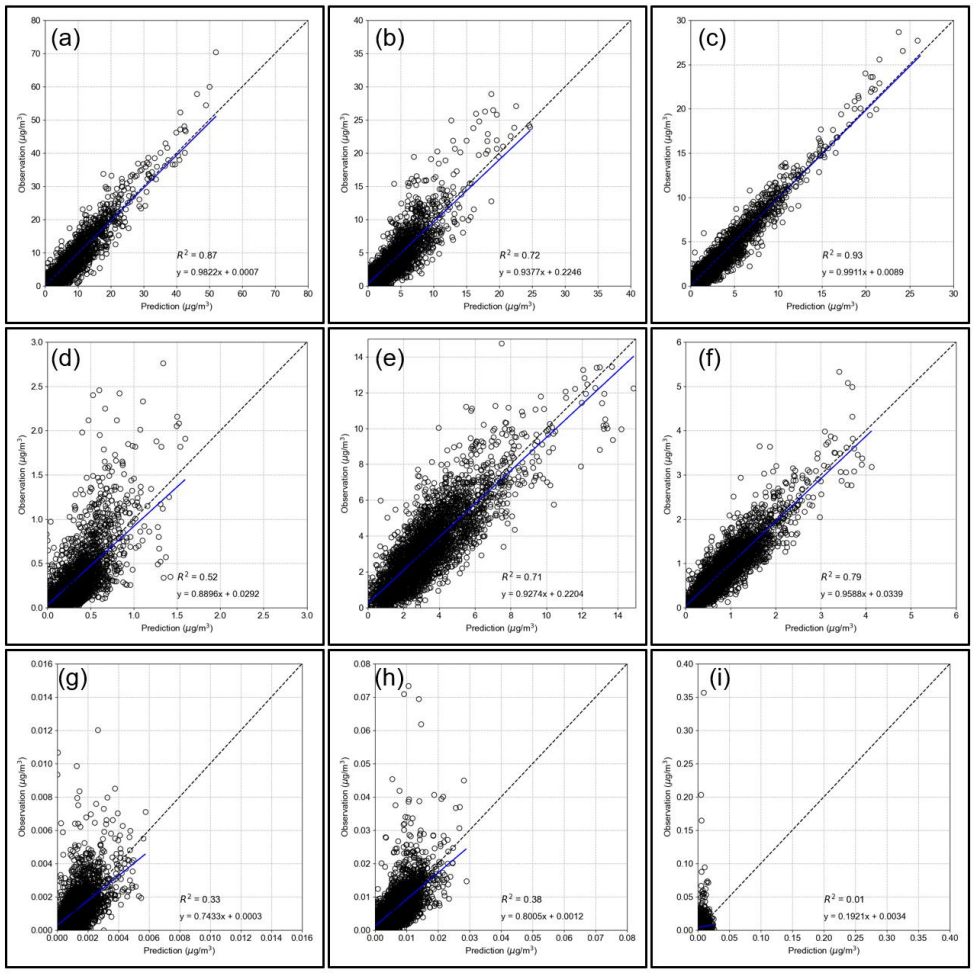


Fig. 4.9. Comparisons of observations and predictions by MI model prediction in Seoul (ID#4, PC#7): (a) NO_3^-, (b) SO_4^{2-}, (c) NH_4^+, (d) Cl^-, (e) OC, (f) EC, (g) Cr, (h) Cu, and (i) As (No. of points = 3,494)

4.3.4. Features and Performance of Four ML Models

The features and performance of the four ML models used in this study may be utilized for designing a predictive regression model in other studies. The prediction accuracy of the GAIN model was the highest (Table 4.5 and Fig. 4.5), indicating the best performance in predicting missing $PM_{2.5}$ constituent values. The decrease was also the least in the GAIN model even though the R^2 decreased when trace elements are predicted (e.g., PC#3, 5, and 6) in all models. These results may help to explain why the GAIN model has lately gained popularity and is being applied to a variety of domains for the processing of missing values. (Andrews and Gorell, 2020; Popolizio et al., 2021; Viñas et al., 2020).

However, prediction accuracy is not the only criterion for selecting a model as other qualitative pros and cons of the model should be considered. RF and kNN have the advantages of easy handling and simple algorithms although GAIN and FCDNN have higher prediction accuracies than RF and kNN and maintained high prediction accuracy values with an increase in input variables (Table 4.6).

The GAIN model undertakes unsupervised learning, with the main goal to make missing values appear similar to observed data by identifying hidden patterns in the data collected. Therefore, it may be difficult to interpret whether AP characteristics such as physicochemical reactions are learned by the model or not. Instead, it focuses on producing credible data. This is similar to kNN, which functions by finding the nearest points using the entire data. In contrast, because FCDNN and RF are supervised learning models, learning is separately completed from the training data sets, and the characteristics extracted from IDs can be used to

find the best weights and biases of the models. Therefore, examining the weights and biases learned by the FCDNN and RF models helps identify the most important characteristics for predicting each component of $PM_{2.5}$. With repeated usage of the trained models and a deeper comprehension of atmospheric chemical and physical processes, this aspect of supervised learning models may potentially offer benefits.

Interqurtile range of min-max normalized value of the data versus the prediction accuracy (R^2) of each model was shown in Fig. 4.10 to investigate the relevance between the prediction accuracy of ML models and the variability of data used. Moderate positive correlation in RF, EM, and MI ($0.5 < R < 0.7$). The higher the data variability, the higher the prediction accuracy of EM and MI model. It is suggested that the prediction accuracy is high due to the feature extraction performance of deep learning models, not the concentration variability of the chemical constituents of $PM_{2.5}$ itself.

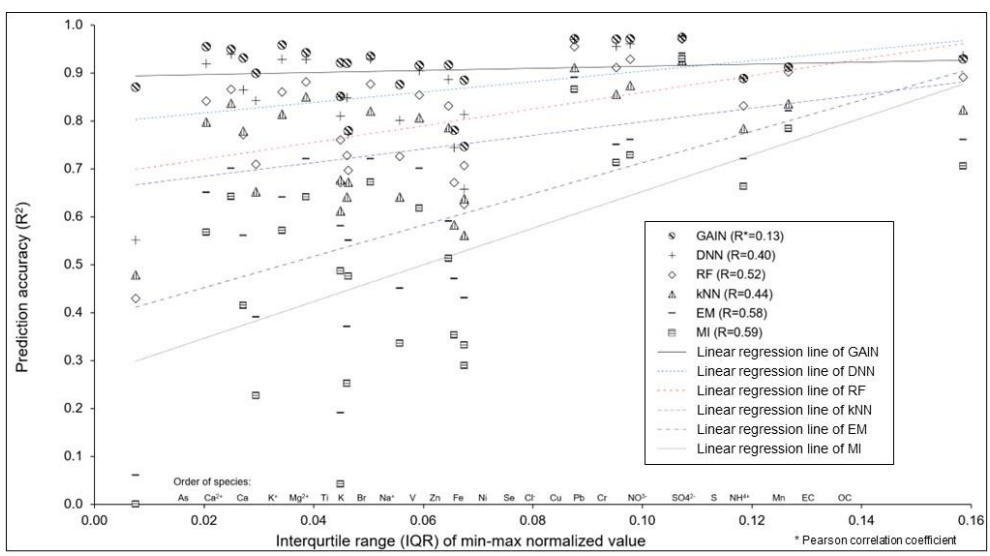


Fig. 4.10. Prediction accuracy (R^2) of each constituent by the variability of the data

4.4. Summary

In this study, the feature extraction capabilities of the four ML models to predict the chemical composition of $PM_{2.5}$ were assessed by comparing the prediction accuracy depending on input variables, target constituents for prediction, available period, missing ratios of input data, and study sites. The prediction accuracy identified by the coefficient of determination (R^2) between prediction and observation was highest in GAIN, followed by FCDNN and RF or kNN. As missing ratios (20%, 40%, 60%, 80%) of input data increased, prediction accuracy decreased in the four models and was more noticeable in GAIN and kNN, which are unsupervised models. As the period of input data increased, the two deep learning models (i.e., GAIN and DNN) had better applicability than the others (i.e., RF and kNN). In the comparison of prediction accuracy by city, the prediction accuracy was lower in cities with more particulate matter emission from industrial activity, resulting in the highest R^2 in BR island and lowest in Ulsan. Among the target constituent groups, the ions and trace elements were predicted with the highest and lowest R^2, respectively.

The high prediction accuracy of machine learning models means that features from data were extracted successfully with the suitable structure of the models (Alzubaidi et al., 2021). In terms of prediction accuracy, the ability to extract features from data, the ability to repeat tests following independent training, ease of use or convenience, and processing speed, each of the four models has strengths and weaknesses. This study can be used for reference in other studies to predict missing values of $PM_{2.5}$ chemical composition by selecting an appropriate model. The accuracy of prediction of missing values presented in this study was generally high

and was of a practically applicable level. Machine learning is a timely application that is ideal for data on air pollution that is growing high-dimensional and has more precise spatial and temporal needs. This study demonstrates that machine learning models can be extended for further air pollution studies depending on model features, required performance, and experimental conditions such as data availability and time constraint.

Data Availability

The AP data (e.g., PM_{10}, NO_2, and CO) are available on the AirKorea website (https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123).

The MD can be obtained from the automatic weather stations (https://data.kma.go.kr/data/grnd/selectAwsRltmList.do?pgmNo=56).

Code Availability

All the scripts used in the study for data processing and analysis are available in the form of .py or .ipynb files in the following GitHub repository: https://github.com/hadistar/hadistar, https://github.com/minjae960/GAIN_TF2

References

Alpaydin, E., 2020. Introduction to machine learning. MIT press.

Andrews, J., Gorell, S., 2020. Generating Missing Unconventional Oilfield Data using a Generative Adversarial Imputation Network (GAIN). OnePetro. https://doi.org/10.15530/urtec-2020-3014

Asim, M., Rashid, A., Ahmad, T., 2021. Scour modeling using deep neural networks based on hyperparameter optimization. ICT Express. https://doi.org/10.1016/j.icte.2021.09.012

Baker, K.R., Foley, K.M., 2011. A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM2.5. Atmos. Environ. 45, 3758–3767. https://doi.org/10.1016/J.ATMOSENV.2011.03.074

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., Cox, D.D., 2015. Hyperopt: A Python library for model selection and hyperparameter optimization. Comput. Sci. Discov. 8, 014008. https://doi.org/10.1088/1749-4699/8/1/014008

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Choi, E., Heo, J.B., Hopke, P.K., Jin, B.B., Yi, S.M., 2011. Identification, apportionment, and photochemical reactivity of non-methane hydrocarbon sources in Busan, Korea. Water. Air. Soil Pollut. 215, 67–82. https://doi.org/10.1007/s11270-010-0459-0

Choi, E., Yi, S.M., Lee, Y.S., Jo, H., Baek, S.O., Heo, J.B., 2022. Sources of airborne particulate matter-bound metals and spatial-seasonal variability of health risk potentials in four large cities, South Korea. Environ. Sci. Pollut. Res. 29, 28359–28374. https://doi.org/10.1007/s11356-021-18445-8

Ciaburro, G., Iannace, G., 2021. Machine Learning-Based Algorithms to Knowledge Extraction from Time Series Data: A Review. Data 2021, Vol. 6, Page 55 6, 55. https://doi.org/10.3390/DATA6060055

Gao, L., Song, J., Liu, X., Shao, Junming, Liu, J., Shao, Jie, 2017. Learning in high-dimensional multimedia data: the state of the art. Multimed. Syst. 23, 303–313. https://doi.org/10.1007/s00530-015-0494-1

Gil, J., Lee, M., Kim, J., Lee, G., Ahn, J., 2021. Simulation Model of Reactive Nitrogen Species in an Urban Atmosphere using a Deep Neural Network: RNDv1.0 2 3. Geosci. Model Dev. Discuss. https://doi.org/10.5194/gmd-2021-347

Hadeed, S.J., O'Rourke, M.K., Burgess, J.L., Harris, R.B., Canales, R.A., 2020. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Sci. Total Environ. 730, 139140. https://doi.org/10.1016/J.SCITOTENV.2020.139140

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science (80-.). 313, 504–507. https://doi.org/10.1126/SCIENCE.1127647

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. https://doi.org/10.1080/10962247.2016.1140693

Hopke, P.K., Dai, Q., Li, L., Feng, Y., 2020. Global review of recent source

apportionments for airborne particulate matter. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.140091

Hwangbo, S., Al, R., Chen, X., Sin, G., 2021. Integrated Model for Understanding N2O Emissions from Wastewater Treatment Plants: A Deep Learning Approach. Environ. Sci. Technol. 55, 2143–2151. https://doi.org/10.1021/acs.est.0c05231

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science (80-.). https://doi.org/10.1126/science.aaa8415

Khan, S.I., Hoque, A.S.M.L., 2020. SICE: an improved missing data imputation technique. J. Big Data 7, 1–21. https://doi.org/10.1186/s40537-020-00313-w

Kim, K.H., Kabir, E., Kabir, S., 2015. A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005

Kim, S., Kim, T.Y., Yi, S.M., Heo, J., 2018. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. J. Environ. Manage. 214, 325–334. https://doi.org/10.1016/j.jenvman.2018.03.027

Korea Ministry of Environment, National Institute of Environmental Research, 2022. 2020 Annual Report of Intensive Air Quality Monitoring Station.

Korea Ministry of Environment, National Institute of Environmental Research, 2021. Guidelines for Installation and Operation of National Air Pollution Monitoring Network.

Lee, B.K., Hieu, N.T., 2011. Seasonal Variation and Sources of Heavy Metals in Atmospheric Aerosols in a residential Area of Ulsan, Korea. Aerosol Air Qual. Res. 11, 679–688. https://doi.org/10.4209/aaqr.2010.10.0089

Lee, Y.S., Kim, Y.K., Choi, E., Jo, H., Hyun, H., Yi, S.-M., Kim, J.Y., 2022. Health risk assessment and source apportionment of PM2.5-bound toxic elements in the industrial city of Siheung, Korea. Environ. Sci. Pollut. Res. 1, 1–14. https://doi.org/10.1007/s11356-022-20462-0

Li, S.C.X., Marlin, B.M., Jiang, B., 2019. Misgan: Learning from incomplete data with generative adversarial networks, in: 7th International Conference on Learning Representations, ICLR 2019. International Conference on Learning Representations, ICLR.

Liu, S., Zhu, C., Tian, H., Wang, Y., Zhang, K., Wu, B., Liu, X., Hao, Y., Liu, W., Bai, X., Lin, S., Wu, Y., Shao, P., Liu, H., 2019. Spatiotemporal Variations of Ambient Concentrations of Trace Elements in a Highly Polluted Region of China. J. Geophys. Res. Atmos. 124, 4186–4202. https://doi.org/10.1029/2018JD029562

Liu, X., Fu, Y., Wang, Q., Bi, Y., Zhang, Li, Zhao, G., Xian, F., Cheng, P., Zhang, Luyuan, Zhou, J., Zhou, W., 2022. Unraveling the process of aerosols secondary formation and removal based on cosmogenic beryllium-7 and beryllium-10. Sci. Total Environ. 821, 153293. https://doi.org/10.1016/J.SCITOTENV.2022.153293

Lyu, B., Hu, Y., Zhang, W., Du, Y., Luo, B., Sun, X., Sun, Z., Deng, Z., Wang, Xiaojiang, Liu, J., Wang, Xuesong, Russell, A.G., 2019. Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved PM2.5 Exposure Fields in 2014-2017. Environ. Sci. Technol. 53, 7306–7315.

https://doi.org/10.1021/acs.est.9b01117

Montavon, G., Samek, W., Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. A Rev. J. 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nazábal, A., Olmos, P.M., Ghahramani, Z., Valera, I., 2020. Handling incomplete heterogeneous data using VAEs. Pattern Recognit. 107. https://doi.org/10.1016/j.patcog.2020.107501

Park, M. Bin, Lee, T.J., Lee, E.S., Kim, D.S., 2019. Enhancing source identification of hourly PM2.5 data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). Atmos. Pollut. Res. 10, 1042–1059. https://doi.org/10.1016/j.apr.2019.01.013

Park, E.H., Heo, J., Kim, H., Yi, S.-M., 2020. Long term trends of chemical constituents and source contributions of PM2.5 in Seoul. Chemosphere 251, 126371. https://doi.org/10.1016/j.chemosphere.2020.126371

Park, S.S., Cho, S.Y., Jo, M.R., Gong, B.J., Park, J.S., Lee, S.J., 2014. Field evaluation of a near–real time elemental monitor and identification of element sources observed at an air monitoring supersite in Korea. Atmos. Pollut. Res. 5, 119–128. https://doi.org/10.5094/APR.2014.015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., Liao, Q., 2017. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review.

Int. J. Autom. Comput. https://doi.org/10.1007/s11633-017-1054-2

Popolizio, M., Amato, A., Liquori, F., Politi, T., Quarto, A., Lecce, V. Di, 2021. The GAIN Method for the Completion of Multidimensional Numerical Series of Meteorological Data. IAENG Int. J. Comput. Sci. 48, 1–11.

Quinteros, M.E., Lu, S., Blazquez, C., Cárdenas-R, J.P., Ossa, X., Delgado-Saborit, J.M., Harrison, R.M., Ruiz-Rudolph, P., 2019. Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. Atmos. Environ. 200, 40–49. https://doi.org/10.1016/j.atmosenv.2018.11.053

Shi, G., Liu, J., Zhong, X., 2021. Spatial and temporal variations of PM2.5 concentrations in Chinese cities during 2015-2019. Int. J. Environ. Health Res. https://doi.org/10.1080/09603123.2021.1987394

Shi, X., Nenes, A., Xiao, Z., Song, S., Yu, H., Shi, G., Zhao, Q., Chen, K., Feng, Y., Russell, A.G., 2019. High-Resolution Data Sets Unravel the Effects of Sources and Meteorological Conditions on Nitrate and Its Gas-Particle Partitioning. Environ. Sci. Technol. 53, 3048–3057. https://doi.org/10.1021/acs.est.8b06524

Tella, A., Balogun, A.L., Adebisi, N., Abdullah, S., 2021. Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. Atmos. Pollut. Res. 12, 101202. https://doi.org/10.1016/J.APR.2021.101202

Viñas, R., Azevedo, T., Gamazon, E.R., Liò, P., 2020. Gene expression imputation with Generative Adversarial Imputation Nets. bioRxiv. https://doi.org/10.1101/2020.06.09.141689

Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., Yu, J.Z., 2018. Source Apportionment of PM2.5 Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-

Resolution Influence. J. Geophys. Res. Atmos. 123, 5284–5300. https://doi.org/10.1029/2017JD027877

Xing, J., Zheng, S., Ding, D., Kelly, J.T., Wang, S., Li, S., Qin, T., Ma, M., Dong, Z., Jang, C., Zhu, Y., Zheng, H., Ren, L., Liu, T.Y., Hao, J., 2020. Deep Learning for Prediction of the Air Quality Response to Emission Changes. Environ. Sci. Technol. 54, 8589–8600. https://doi.org/10.1021/acs.est.0c02923

Yao, Z., Ruzzo, W.L., 2006. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7, S11. https://doi.org/10.1186/1471-2105-7-S1-S11

Yoon, J., Jordon, J., Van Der Schaar, M., 2018. Supplementary materials GAIN: Missing data imputation using generative adversarial nets, in: 35th International Conference on Machine Learning, ICML 2018. PMLR, pp. 9052–9059.

Chapter 5. Bayesian spatial multivariate receptor modeling for spatiotemporal analysis of $PM_{2.5}$ sources

5.1. Introduction

Particulate matter less than 2.5 micrometers ($PM_{2.5}$) is a major pollutant of interest for clean air, and the demand for its reduction continues to increase (Hopke et al., 2019). Identifying major emission sources and assessing their contributions to the total $PM_{2.5}$ concentrations is crucial for developing more targeted enforcement strategies and effective management of $PM_{2.5}$, which can also be reflected in environmental health policies. Human health risks due to emission sources in certain areas can also be evaluated based on the estimated source contributions (Hopke et al., 2020; Wang et al., 2021). Furthermore, it would be beneficial if the spatio-temporal distribution of sources could be modeled simultaneously in that it can be used as an important reference for emission reduction measures or to identify high incidence areas. (Shi et al., 2021). Regardless of continuous attempts to derive scientific information about major $PM_{2.5}$ sources and their contributions, there remain many challenges because of the limitations, such as the requirement for high-resolution data, measurement uncertainty, and modeling and estimation uncertainty (Diao et al., 2019; Hopke, 2016; Hopke et al., 2020).

For source apportionment of $PM_{2.5}$, receptor models based on factor analysis, chemical mass balance (CMB), and principal component analysis (PCA) have been used over four decades (Hopke, 2016). Karagulian et al. (2015) summarized a total of 419 source apportionment studies conducted in 51 countries around the world and identified the 14 types of receptor models utilized from 1986 to 2012. Positive matrix

factorization (PMF), a type of factor analysis method, and CMB have been mostly used with 45% and 19% of usage, respectively (Karagulian et al., 2015). Especially, PMF, the most widely used method for source apportionment in recent decades, has its uncertainty evaluation capabilities for source compositions, such as bootstrap and displacement options, although not for source contributions, as well as producing source compositions and contribution estimates that are interpretable based on domain knowledge (Hopke, 2016; Paatero and Tapper, 1994; Polissar et al., 2001). More recently, advanced PMF methods such as dispersion normalized PMF and window PMF have also emerged to reduce the influence of meteorology on the source emission patterns (Dai et al., 2020a, 2020b; Hopke, 2021). Also, PMF modeling research using hourly data rather than daily data has increased the time resolution of source apportionment with the recent development of measurement techniques (Dai et al., 2020b; Park et al., 2019; Shi et al., 2019; Wang et al., 2018). In addition, ensemble approaches such as integrating multiple receptor models and chemical transport models (CTMs) have been employed to achieve better performance and improve the CTM forecast. (Hopke, 2016; Sokhi et al., 2021). However, the challenges such as the rotational ambiguity problem and the difficulty with incorporating spatial correlation in multi-site data into PMF estimates still exist (Hopke, 2021). Overall, the source apportionment methods capable of incorporating spatial correlations in multi-site data into estimation are very limited (Park et al., 2018).

More recently, there has been growing interest in Bayesian approaches in receptor modeling (Hopke, 2016). In fact, the Bayesian approaches are increasingly being used in all social science and engineering fields, including environmental

engineering, with the development of computational technology (M. H. Park et al., 2020). Bayesian factor analysis methodology was previously introduced into receptor modeling as a way to resolve a rotational ambiguity problem as well as handling the unknown number of sources and providing uncertainty estimates of source profiles and contributions (E. S. Park et al., 2014; Park et al., 2021, 2018, 2004, 2002; Park and Oh, 2018, 2015; Park and Tauler, 2020). In a Bayesian approach, each parameter is assumed to have its own probability distribution, called a prior distribution, of which variability reflects the uncertainty in prior information. With this important feature, any prior information about pollution sources from the domain knowledge, in addition to the data, can be incorporated into parameter estimation in a mathematically rigorous fashion in Bayesian source apportionment models (Park and Tauler, 2020), which is not possible in non-Bayesian source apportionment models.

Bayesian spatial multivariate receptor modeling (BSMSM), proposed by Park et al. (2018), is a Bayesian source apportionment approach that can incorporate spatial correlations in multi-site multipollutant data into parameter estimation and enable spatial prediction of source contributions at unmonitored sites. Furthermore, it can simultaneously deal with model uncertainty resulting from an unknown number of sources or rotational ambiguity. Therefore, BSMRM has the advantage of being able to account for both the uncertainty in source apportionment and spatial correlations in the data in prediction. The first application of BSMRM was based on 17 volatile organic compounds data collected from nine monitoring sites in Harris County, Texas, USA. The predicted source contributions for five major sources of the Harris County area were derived incorporating spatial correlations in the VOCs

data from multiple monitoring sites (Park et al., 2018).

With the recent surge in interest in $PM_{2.5}$ management in Korea, the number of nationally operated $PM_{2.5}$ speciation monitoring sites has been increasing. Accordingly, the level of demand for scientific source apportionment is high in terms of data utilization, and several PMF modeling and analysis results using national measurement data have been reported recently (Hwang et al., 2020; Jeong et al., 2017; Lee et al., 2019; Park et al., 2019). However, there have been no studies on spatial prediction of source-specific $PM_{2.5}$ pollution using multi-site $PM_{2.5}$ speciation data. A source apportionment study applying BSMRM to multi-site $PM_{2.5}$ speciation data in Korea is a timely new attempt. Source apportionment results with spatial prediction at unmonitored sites (cities) using $PM_{2.5}$ speciation data operated at the national level could be vital information for successful management of $PM_{2.5}$. As mentioned earlier, prediction of source-specific $PM_{2.5}$ pollution at any location can lead to developing an effective pollution control plan for a city with no $PM_{2.5}$ chemical speciation monitoring site.

This study aims to predict source-specific $PM_{2.5}$ pollution at unmonitored sites in regional scale by employing BSMRM, which models spatial correlation in multi-site $PM_{2.5}$ chemical speciation data to make spatial predictions. BSMRM will also be evaluated by verifying the model results based on the held-out test data not used for model development and estimation. Prediction of unobserved source contributions from BSMRM at an unmonitored site (a test site) will be compared with the source contributions estimated by a single-site source apportionment method, BNFA (Park et al., 2021), based on data from the test site. Finally, maps of source-specific pollution surfaces over Korea, constructed based on predicted values

from BSMRM, will also be presented.

5.2. Materials and methods

5.2.1 Air pollution data

The $PM_{2.5}$ chemical speciation data measured for 1/1/2020-12/29/2020 from 8 sites were used for the analysis. Fig. 5.1 shows the location of the monitoring stations from which the data were obtained. The concentrations of the chemical components of $PM_{2.5}$ for 7 out of those 8 sites (Baengnyeong, Seoul, Ansan, Daejeon, Gwangju, Ulsan, and Jeju) were obtained from intensive $PM_{2.5}$ monitoring stations operated by the Korean Ministry of Environment (NIER, 2016). Mass concentrations of $PM_{2.5}$ were measured by β-ray absorption method (BAM 1020, Met One Instruments, Inc., USA). The analysis methods for quantifying the chemical composition of $PM_{2.5}$ are as follows: Ionic species were measured by ion chromatography (URG-9000D ambient ion monitor, URG Corp.). Organic carbon (OC) and elemental carbon (EC) were measured by thermal-optical transmittance method (OC-EC Analyzer, Sunset Laboratory Inc., USA). Elemental concentrations were measured using an ambient elemental monitor (XactTM 620, Cooper Environmental Services, USA) which is analyzed by X-ray fluorescence spectrometry (XRF). QA/QC for $PM_{2.5}$ and its components data can be found in the guideline for installation and operation of national air pollution monitoring network (Korea Ministry of Environment and National Institute of Environmental Research, 2021).

The methods for collecting the data from the remaining one site (Siheung) are described in detail in Lee et al. (2022). Briefly, the mass concentrations of Siheung data were obtained by measuring the weight of a 24-hour dried Teflon filter

(PT47P, MTL, US) before and after sample collection, and then dividing the obtained value by the collected air volume. The analysis methods for quantifying the chemical composition of $PM_{2.5}$ are also described in detail in Lee et al. (2022).

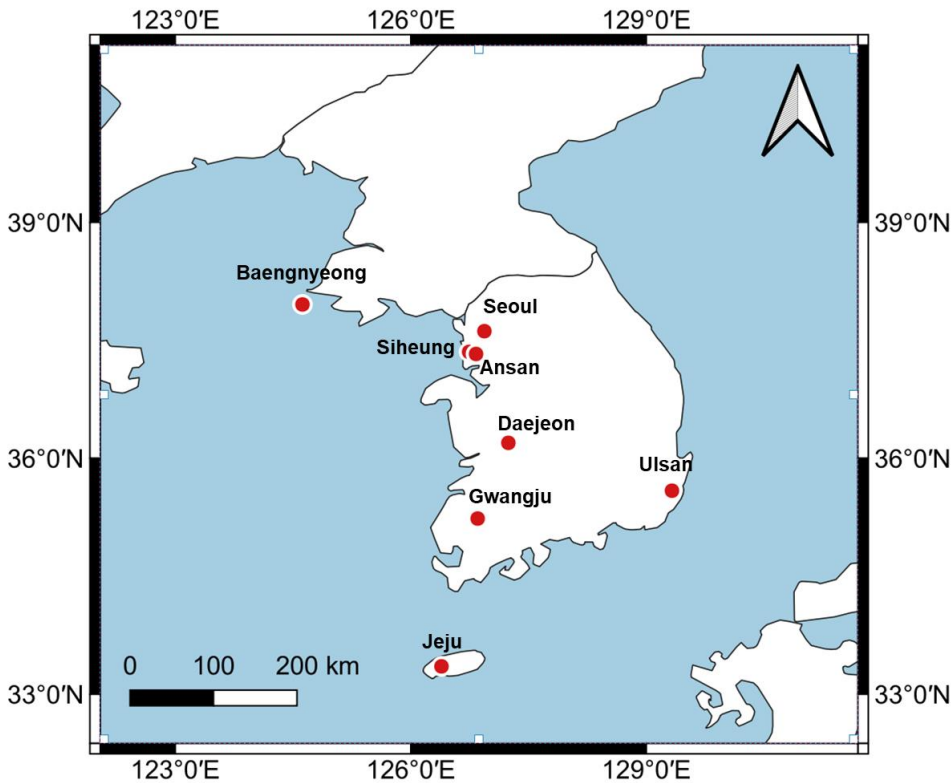


Fig. 5.1. Locations of $PM_{2.5}$ chemical speciation monitoring sites in South Korea.

Table 5.1 contains the summary statistics for 20 species used in this analysis and the total $PM_{2.5}$ mass concentration based on data from all 8 monitoring sites.

Table 5.1. Summary statistics for $PM_{2.5}$ and its chemical species

Species No.	Species	Mean	Median	Std. Dev.	Minimum	Maximum
	$PM_{2.5}$ ($\mu g/m^3$)	20.5	17.3	13.0	0.846	78.7
	Ion ($\mu g/m^3$)					
1	NO_3^-	4.62	2.55	5.10	0.021	29.1
2	SO_4^{2-}	3.47	2.86	2.49	0.035	17.8
3	Cl^-	0.359	0.230	0.388	0.000	2.93
4	Na^+	0.112	0.083	0.112	0.000	1.22
5	NH_4^+	2.84	2.17	2.24	0.007	15.1
6	K^+	0.110	0.071	0.120	0.000	0.912
7	Mg^{2+}	0.022	0.015	0.025	0.000	0.237
	Carbon ($\mu g/m^3$)					
8	OC	3.17	2.69	2.26	0.000	14.4
9	EC	0.652	0.569	0.452	0.000	2.94
	Trace Element (ng/m^3)					
10	Ca	50.6	40.0	37.5	0.702	283
11	Ti	7.79	6.38	6.40	0.029	91.9
12	V	0.426	0.218	0.621	0.000	6.87
13	Cr	1.23	0.907	1.29	0.000	11.4
14	Mn	10.6	8.39	9.31	0.000	79.5
15	Fe	162	145	105	2.48	738
16	Ni	0.978	0.754	0.853	0.000	5.13
17	Cu	6.83	4.33	9.24	0.038	93.2
18	Zn	43.14	35.3	33.6	0.615	226
19	As	4.81	3.03	6.07	0.000	72.6
20	Pb	15.5	11.2	14.7	0.000	111

5.2.2 Bayesian spatial multivariate receptor modeling (BSMRM)

In this study, we performed source apportionment of $PM_{2.5}$ for Korea based on speciated $PM_{2.5}$ data collected from multiple monitoring sites by using Bayesian spatial multivariate receptor modeling (BSMRM) proposed by Park et al. (2018). The main motivation of BSMRM was to account for spatial correlation in the air pollution data collected from multiple monitoring sites in modeling and estimation and predict source contributions at unmonitored sites. For completeness, BSMRM models are briefly described here again. Let N be the number of monitoring sites and T be the number of time points. The basic model for the rth monitoring site at time t is

$$X_t^r = A_t^r \boldsymbol{P} + E_t^r, \ t = 1, \cdots, T, \ r = 1, \cdots, N, \qquad \text{Eq. 5.1}$$

where **P** is a $q \times J$ source-composition matrix, $X_t^r = \left(X_{t1}^r, \cdots, X_{tJ}^r\right)$ is a vector of observed concentrations on J pollutants at monitoring site r at time t, $A_t^r = \left(A_{t1}^r, \cdots, X_{tq}^r\right)$ is a vector of contributions from q sources at monitoring site r at time t, and $E_t^r = \left(E_{t1}^r, \cdots, E_{tJ}^r\right)$ is a J-dimensional vector of errors associated with each observation at the rth monitoring site and time t. The elements of **P** are constrained to be nonnegative. Park et al. (2018) extended the model in Eq. 5.1 to incorporate spatial correlation in multi-site multipollutant data into multivariate receptor modeling by adapting the dynamic factor process convolution model of Calder (2007) based on the discrete process convolution approach, originally proposed by Higdon (1998), to modeling spatial data. The discrete process convolution approach

expresses the spatial process as a sum of the discrete underlying (latent) processes defined on *L* locations on a coarse grid $\{\omega_1, \omega_2, \cdots, \omega_L\}$, covering the spatial domain, smoothed by the kernel κ. Park et al. (2018) relaxed the assumption of the known number of factors and known identifiability conditions of Calder (2007) and handled uncertainty in the unknown number of factors and identifiability conditions simultaneously with parameter estimation. They also incorporated physically meaningful non-negativity constraints (that were not enforced in Calder 2007) for the source composition profile matrix and the source contribution matrix into the estimation.

BSMRM considers the following model for the multivariate air pollution data $\{\boldsymbol{X}(s_r, t), t = 1, \cdots, T\}$ collected from N spatial sites $\{s_1, s_2, \cdots, s_N\}$ over T time points:

$$\boldsymbol{X}(s_r, t) = \boldsymbol{K}(s_r)\boldsymbol{G_t}\boldsymbol{P} + \boldsymbol{E}(s_r, t) \qquad \text{Eq 5.2}$$

where s_r is the spatial location of the *r*th receptor ($r = 1, \cdots, N$), $\mathbf{G_t}$ represents q underlying processes located at *L* spatial locations $\{\omega_1, \omega_2, \cdots, \omega_L\}$ chosen from a coarse grid that covers a spatial domain, $G_t \sim N\left(0, \mathbf{I}_L, \Omega\right)$, $\boldsymbol{K}(s_r) = [\kappa(\omega_1 - s_r), \cdots, \kappa(\omega_L - s_r)]$, κ is a smoothing kernel, and $\boldsymbol{E}(s_r, t)$ is an *iid*, mean zero, Gaussian process on (s_r, t) with variance $\Sigma_J = diag\left(\sigma_1^2, \cdots, \sigma_J^2\right)$.

This spatially extended multivariate receptor model makes it possible to predict source contributions at any location as $\boldsymbol{K}(s)\boldsymbol{G_t}$ by plugging in the estimates for $\boldsymbol{G_t}$ and the corresponding values for $\boldsymbol{K}(s)$ where *s* is a new location.

Uncertainty in the number of major sources and identifiability conditions can be handled by considering marginal likelihood (model evidence given the data) for each model which can be viewed as a measure of model fit (the larger, the better). Estimation of model parameters and computation of marginal likelihoods can be performed by Markov chain Monte Carlo (MCMC) methods. The MATLAB codes for MCMC implementation of BSMRM are also freely available from the Supplementary Materials for Park et al. (2018).

5.2.3 Application of BSMRM to Korea $PM_{2.5}$ speciation data

We applied BSMRM to the $PM_{2.5}$ speciation data collected from seven monitoring sites (except for a test site denoted by a triangle) in Fig. 5.2 in Korea in 2020. The data for each of Daejeon City, Gwangju City, and Ansan City, are set aside as held-out data (test data) to use for validation of BSMRM. These 3 sites were selected to test model validity in inner regions among the monitoring stations to avoid extrapolation. There were a total of 103 days when $PM_{2.5}$ speciation measurements were made for most of the eight monitoring sites. The number of missing observations at any given site varies with species, ranging from 0 to 35 days. The missing observations were imputed by *k*-nearest neighbor imputation (Little and Rubin, 2014), namely, using the spatial average of pollutants from three nearest neighboring sites for each day.

Based on the previous (single-site) studies for different cities in Korea (Choi et al., 2022, 2013; Heo et al., 2009; Hwang et al., 2020; Kim et al., 2018; Park et al., 2019; E. H. Park et al., 2020), secondary sulfate, secondary nitrate, traffic, coal

combustion (including heating), oil combustion, industrial sources, biomass burning, soil, and sea salt were presumed to be potential candidate sources affecting the region. This prior knowledge was utilized in prespecification of zeros in the source-composition profile matrix to achieve model identifiability. Table 5.2 gives the major species for each of the candidate source types. Minor or absent species from each source type are candidates for preassigned zeros in source composition profiles.

Table 5.2. Major Species for Candidate Sources Considered in the Analysis

Source	Major species
Secondary nitrate	NO_3^- and NH_4^+
Secondary sulfate	SO_4^{2-} and NH_4^+
Traffic	OC, EC, and Cu, Fe
Coal combustion	As and Pb, Cl^- (heating)
Industry	V, Cu, Cr, Mn, Fe, Ni, Pb and Zn
Biomass burning	K^+, OC, and EC
Soil	Mg, Al, Si, Ca, Ti
Sea salts	Na, Mg, K

We constructed a range of different models, resulting from each combination of varying number of sources and identifiability conditions (prespecification of zeros), to be compared for the Korea $PM_{2.5}$ data. Based on previous studies on source identification and apportionment of $PM_{2.5}$ for South Korea and the NUMFACT procedure (Henry et al., 1999; Park et al., 2000), we presumed that the number of major sources was between 5 and 8. For candidate positions of zeros in **P** under each q-source model, we used the information on the

major sources of Korea from previous single-site studies aforementioned. Note that we use information from previous single-site studies only to find out the plausible sets of identifiability conditions (absent or minor species for each source type) under each q-source model. Other than that, the candidate models do not depend on the results from those previous studies. We compared eight candidate models with different numbers of sources (q = 5, 6, 7, 8) and different identifiability conditions (prespecification of zeros in **P**) in Table 5.3.

Table 5.3. Candidate Models for Korea $PM_{2.5}$ Data

Model #	q	Prespecification of zeros		logmD(*1.0e+04)
		Source No.	Species No. for preassigned zeros*	
1	5	1	2, 3, 8, 15	-2.1007
		2	1, 3, 8, 15	
		3	3, 5, 6, 15	
		4	4, 8, 9, 10	
		5	3, 8, 9, 15	
2	5	1	2, 3, 8, 15	-2.0952
		2	1, 3, 8, 15	
		3	3, 6, 10, 15	
		4	4, 7, 8, 9	
		5	6, 8, 9, 15	
3	6	1	2, 3, 4, 8, 15	-2.0986
		2	1, 3, 4, 8, 15	
		3	3, 5, 6, 10, 15	
		4	4, 7, 8, 9, 10	
		5	3, 6, 8, 9, 15	
		6	1, 2, 3, 5, 8	
4	6	1	2, 3, 4, 9, 15	-2.0974
		2	1, 3, 4, 10, 15	
		3	3, 5, 6, 7, 15	
		4	4, 5, 7, 9, 10	
		5	3, 5, 6, 8, 15	
		6	2, 3, 5, 8, 16	

5	7	1	2, 3, 4, 8, 10, 15	-2.1021
		2	1, 3, 4, 8, 10, 15	
		3	3, 4, 5, 6, 10, 15	
		4	4, 5, 7, 8, 9, 10	
		5	2, 3, 6, 8, 9, 15	
		6	1, 2, 3, 5, 8, 9	
		7	1, 2, 8, 9, 11, 15	
6	7	1	2, 3, 4, 9, 10, 15	-2.1015
		2	1, 3, 4, 8, 10, 15	
		3	3, 4, 5, 7, 10, 15	
		4	4, 5, 7, 8, 9, 10	
		5	3, 5, 6, 8, 9, 15	
		6	1, 2, 3, 5, 8, 16	
		7	1, 2, 8, 9, 15, 20	
7	8	1	2, 3, 4, 8, 10, 11, 15	-2.0973
		2	1, 3, 4, 8, 10, 15, 17	
		3	3, 4, 5, 6, 10, 12, 15	
		4	4, 5, 6, 7, 8, 9, 10	
		5	1, 2, 3, 6, 8, 9, 15	
		6	1, 2, 3, 5, 8, 10, 16	
		7	1, 2, 6, 9, 11, 15, 20	
		8	4, 5, 7, 15, 18, 19, 20	
8	8	1	2, 3, 4, 9, 10, 11, 15	-2.1040
		2	1, 3, 4, 9, 10, 15, 17	
		3	3, 4, 5, 7, 10, 12, 15	
		4	4, 5, 6, 7, 8, 9, 10	
		5	1, 3, 5, 6, 8, 9, 15	
		6	1, 2, 3, 5, 8, 9, 16	
		7	1, 2, 8, 9, 10, 15, 20	
		8	4, 5, 7, 16, 18, 19, 20	

* Numbers indicate Species Number in Table 5.1 (e.g., 1: NO_3^-, 2: SO_4^{2-}).

5.3. Results and discussion

5.3.1 Bayesian spatial multivariate receptor modeling (BSMRM) results

We fitted Bayesian spatial multivariate receptor models to the data consisting of 20 $PM_{2.5}$ species (measured in μg/m^3) given in Table 5.1 and estimated source-composition profiles and other model parameters along with marginal likelihood under each model. Note that Bayesian model comparison can be performed using the posterior model probability, which is proportional to the marginal likelihood under the indifference prior model probabilities. A model with a higher marginal likelihood (or posterior model probability) is thus preferred. Because concentrations of $PM_{2.5}$ species differed by two or three orders of magnitude and convergence problems may occur when elemental concentrations are on widely different scales, each element was scaled by its sample standard deviation before running MCMC. After the run, however, the individual elements of the estimated source profiles were multiplied by the corresponding sample standard deviations to bring them back to the original scale so that the relative amounts of species in each profile are physically interpretable. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1, \ldots, 17$), $c_0 = 0.5 \times 1_{p^+}$, and $C_0 = 100 \times \boldsymbol{I}_{p^+}$. Also, we set $\Omega = \mathbf{I}_q$ as a way to get around a scale invariance problem for these data. We modeled the underlying process at 9 locations (L=9) chosen to cover the spatial domain of interest shown in Fig.5.2 with the distance between adjacent location used as the standard deviation (σ_κ) of the Gaussian kernel used in model fitting (σ_κ =1.7379).

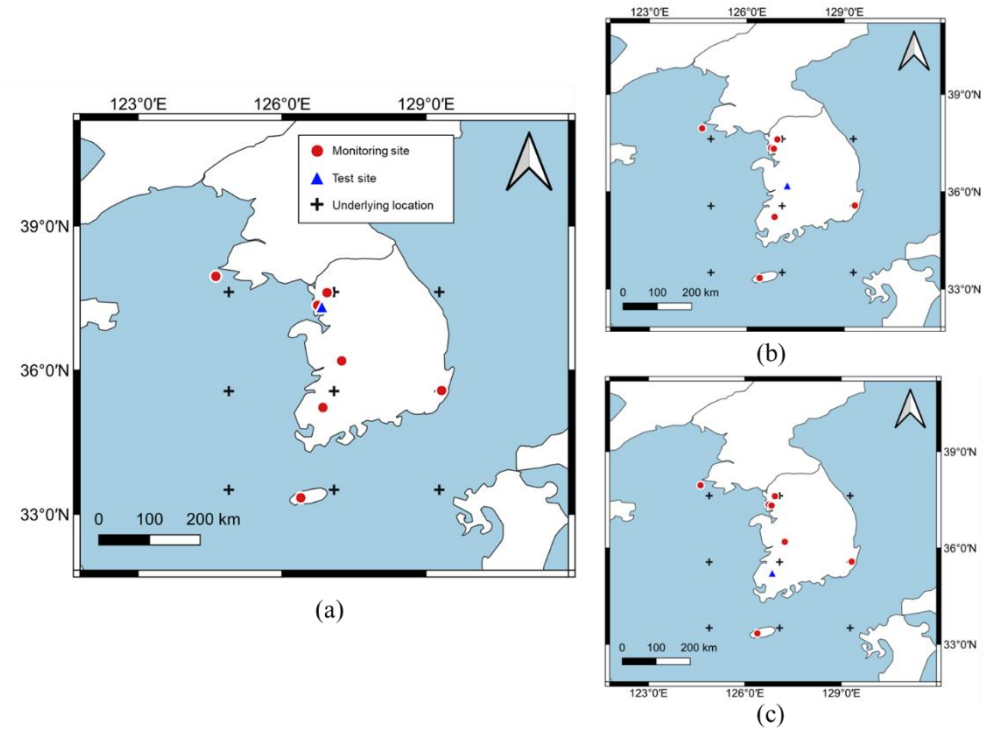


Fig. 5.2. Separation of locations for validation and underlying locations: Test site of (a) Ansan, (b) Daejeon, and (c) Gwangju

For model fitting, three different datasets (consisting of data from 7 remaining monitoring sites after excluding a test site) were used as follows: Dataset 1 excluding the Daejeon site, Dataset 2 excluding the Gwangju site, and Dataset 3 excluding the Ansan site. For each model fitted based on each dataset, an approximate posterior mode is obtained from a preliminary MCMC run, and this is used for $\theta^c = (G^c, \boldsymbol{P}^c, \Sigma^c)$ at which the marginal likelihood is calculated. An approximate posterior mode is obtained by evaluating the joint posterior density for 10,000 iterations after the first 10,000 draws are discarded. A main MCMC run is then started from $\theta^c = (G^c, \boldsymbol{P}^c, \Sigma^c)$, and the samples are collected for 10,000

iterations, without additional burn-in. The estimated marginal likelihood (in logs) for each model is also provided in Table 5.3. Model# 2 with 5 sources is selected as the best model because the marginal likelihood for Model# 2 is the highest among the candidate models considered.

Fig. 5.3 shows the estimated source composition profiles and contributions under Model 2 based on Dataset 1 (which excludes Ansan City of Fig. 5.2). Fig. 5. 3 (a) shows barplots for elements of estimated source profiles (for common major sources for the entire region) along with uncertainty estimates represented by error bars (lower and upper limits of 95% posterior intervals). Note that local sources that are specific to any single city may not be characterized by this regional modeling. Fig. 5.3 (b) contains the time-series plots of the predicted source contributions along with their uncertainty estimates (95% posterior intervals), at a held-out test site (Ansan City). Major species in the estimated source composition profiles of Fig. 5.3 (a) appear to be consistent with main elements of major $PM_{2.5}$ sources for South Korea identified by previous studies, namely, Secondary Nitrate, Secondary Sulfate, Motor Vehicles, Industry, and Sea Salt. The estimated yearly mean source contributions across 7 monitoring sites indicate that Secondary Nitrate, Secondary Sulfate, and Motor Vehicles play a major role in $PM_{2.5}$ emissions for the region, which agrees with previous studies based on the single-site data for each of the individual cities. Recall that the main purpose of this study is to predict major $PM_{2.5}$ source contributions at unmonitored locations (cities), and Bayesian spatial multivariate receptor modeling allows us to predict source contributions at any site (not just at monitoring sites).

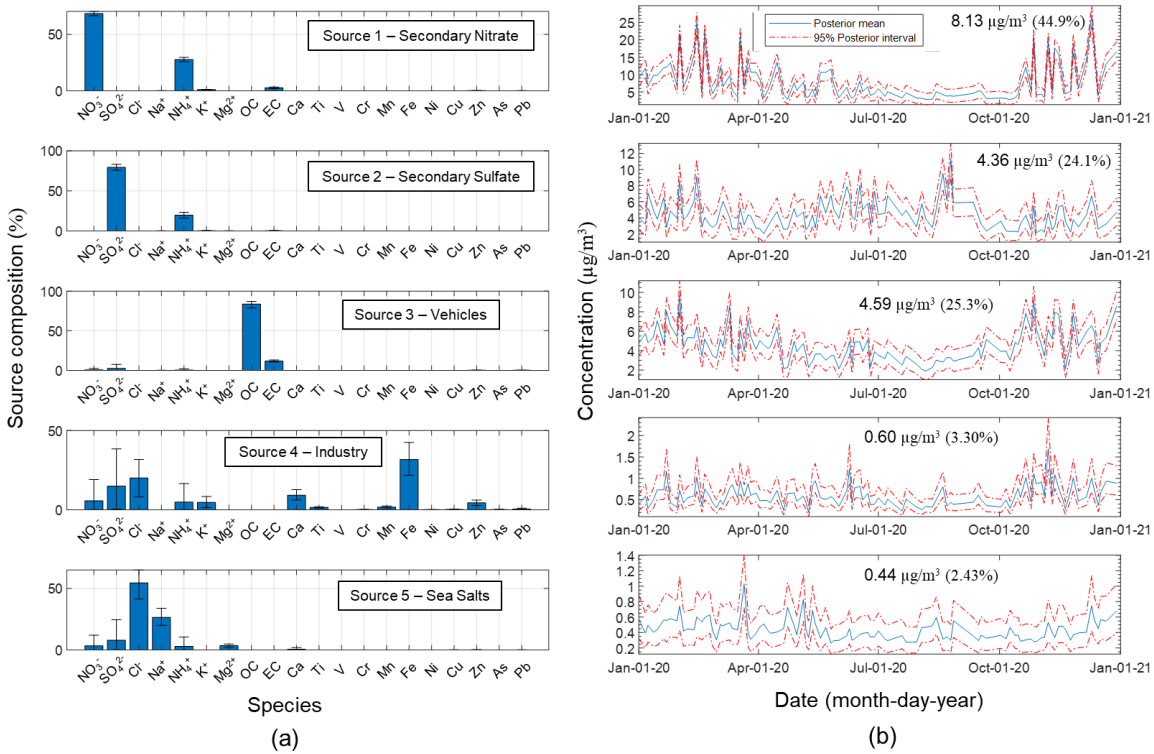


Fig. 5.3. BSMRM model fitting results: (a) Bar plots for the elements of the estimated source profiles along with error bars (lower and upper limits of 95% posterior intervals); (b) Time series plots of source contributions with 95% posterior intervals in Ansan City, predicted by BSMRM

Fig. 5.4 and Fig. 5.5 show the estimated source composition profiles and predicted source contributions for Daejeon City based on Dataset 2 and Gwangju City based on Dataset 3, respectively. Note that the estimated source composition profiles are similar across three cities, while predicted source contributions are different across those cities as expected.

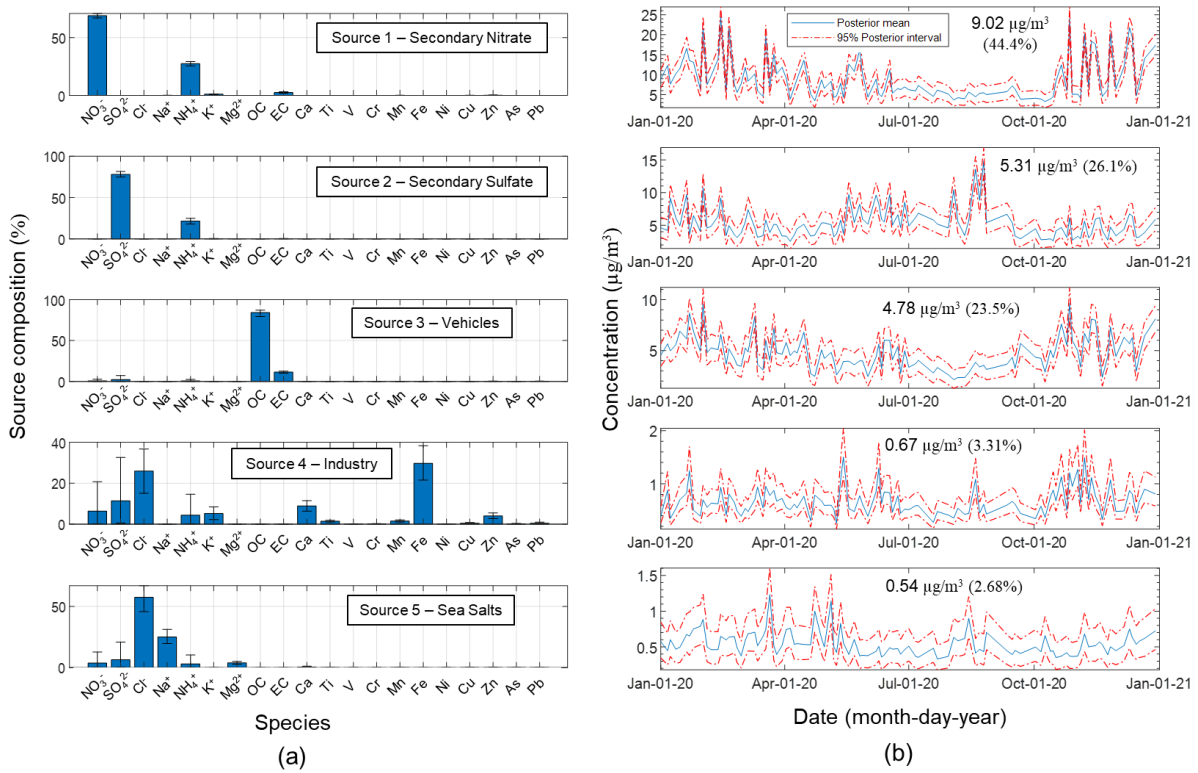


Fig. 5.4. BSMRM model fitting results: (a) Bar plots for the elements of the estimated source profiles along with error bars (lower and upper limits of 95% posterior intervals); (b) Time series plots of source contributions with 95% posterior intervals in Daejeon City, predicted by BSMRM

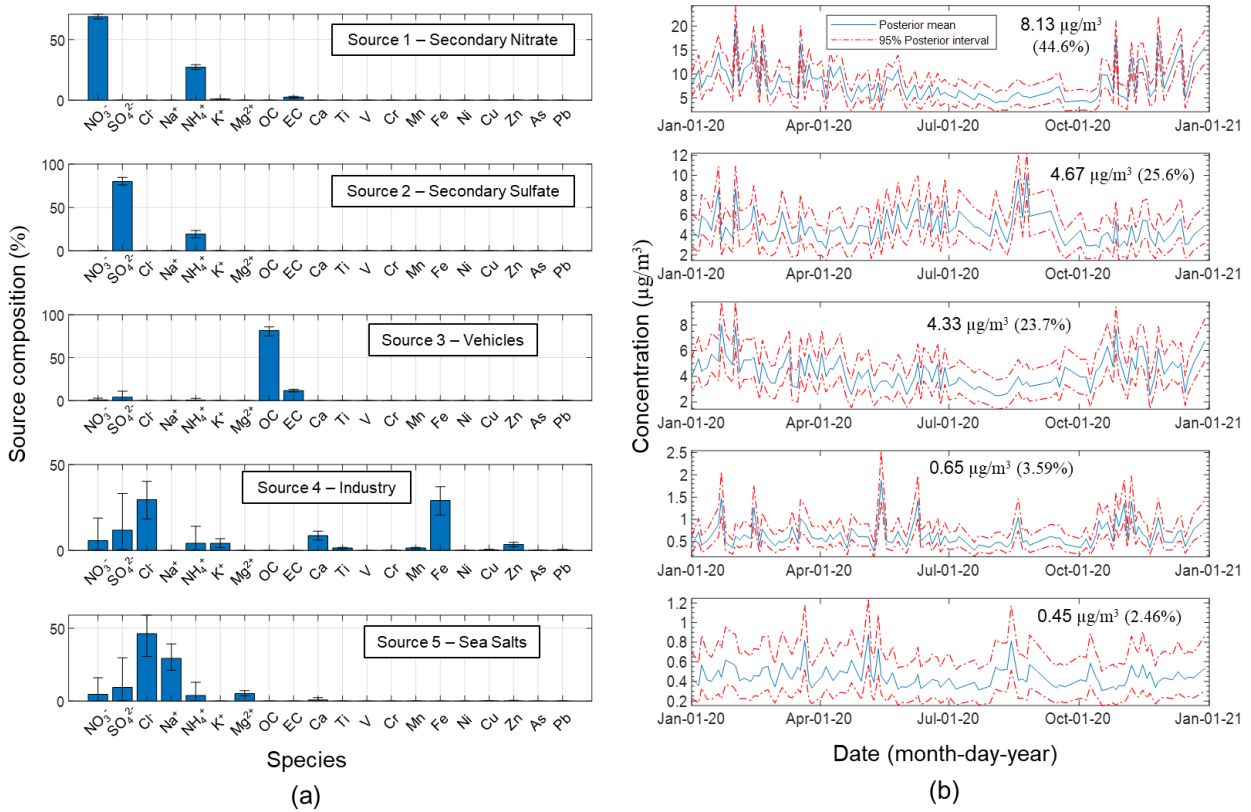


Fig. 5.5. BSMRM model fitting results: (a) Bar plots for the elements of the estimated source profiles along with error bars (lower and upper limits of 95% posterior intervals); (b) Time series plots of source contributions with 95% posterior intervals in Gwangju City, predicted by BSMRM

5.3.2 Model validation

For validation of the prediction by BSMRM, we estimated source contributions at a test site (each of Ansan, Daejeon, and Gwangju sites) using Bayesian multivariate receptor modeling for the single-site data. We performed source apportionment at each site by BNFA (Park et al. 2021). Fig. 5.6, Fig. 5.7, and Fig. 5.8 contain the time-series plots of the source contributions along with their uncertainty estimates (95% posterior intervals) and bar plots of source compositions, respectively, estimated using BNFA based on $PM_{2.5.}$ speciation data obtained from each of the Ansan, Daejeon, and Gwangju sites, respectively.

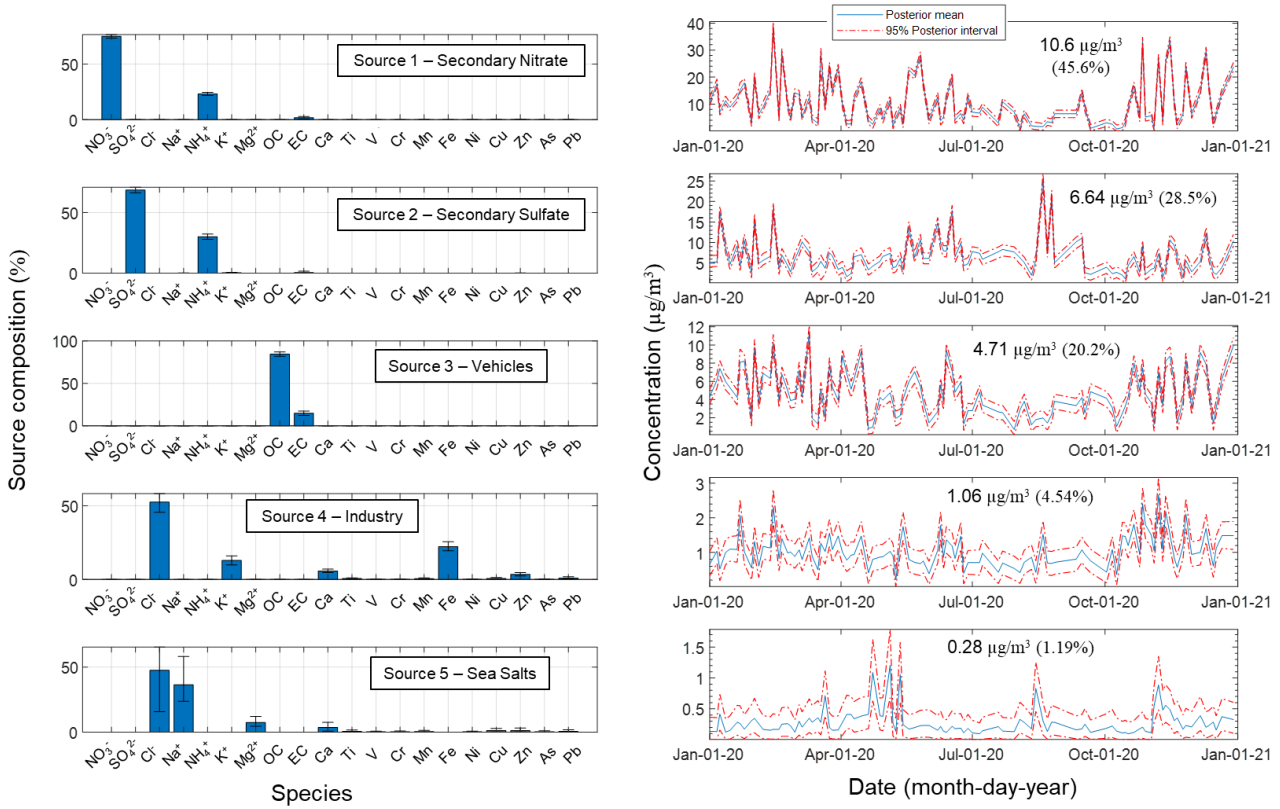


Fig. 5.6 Estimated source composition profiles and predicted source contributions by BNFA for Ansan City

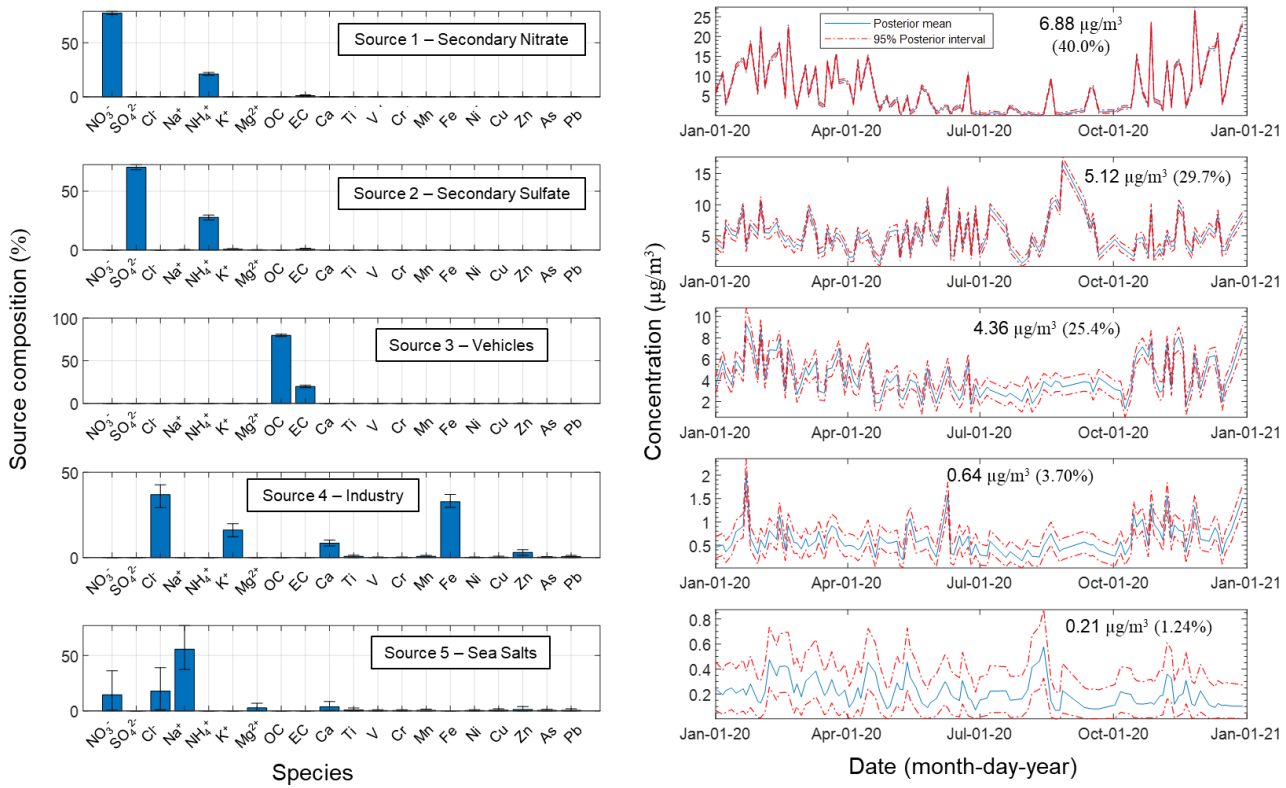


Fig. 5.7. Estimated source composition profiles and predicted source contributions by BNFA for Daejeon City

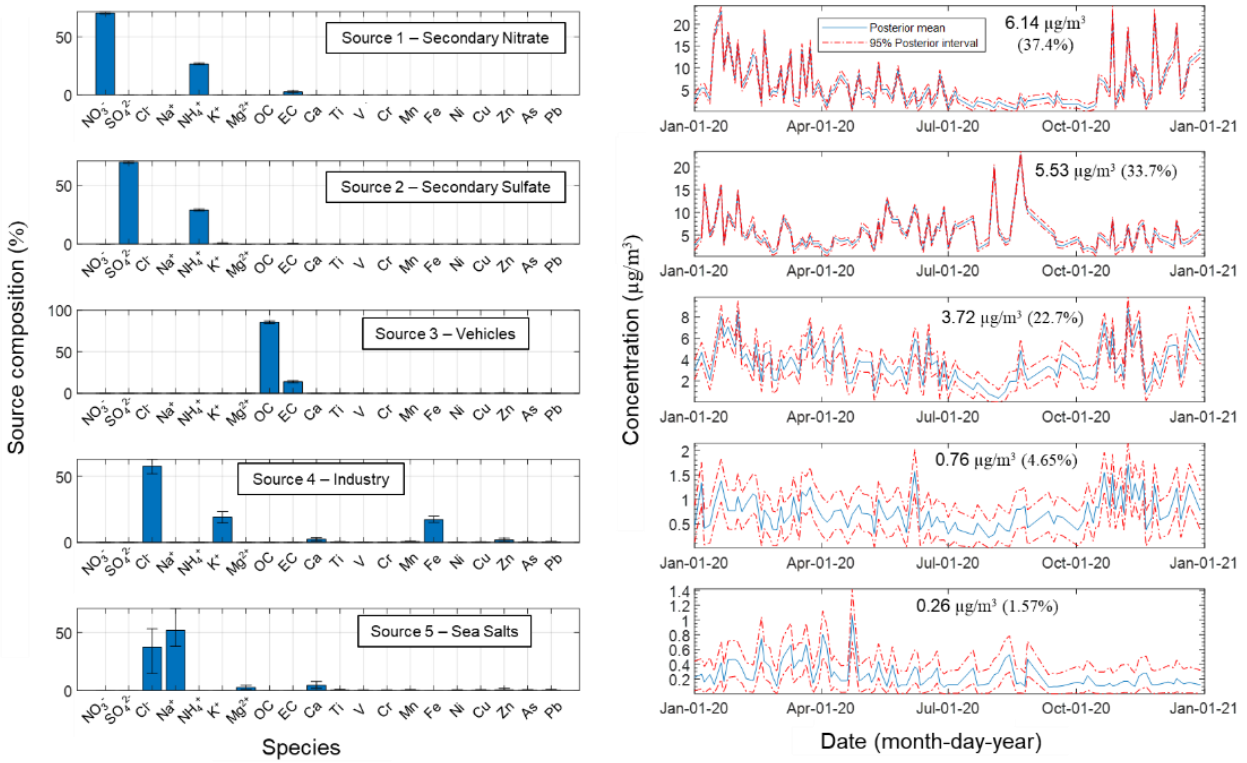


Fig. 5.8. Estimated source composition profiles and predicted source contributions by BNFA for Gwangju City

Figures from Fig. 5.9 to Fig. 5.11 show predicted source contributions by BSMRM overlaid with the predicted source contributions by BNFA at each test site. It can be observed from Fig. 5.3 to Fig. 5.8 that the overall patterns of the source contributions predicted by BSMRM and those estimated by BNFA are similar except peaks of the BNFA source contributions are more extreme, which seems to be a natural consequence of reflecting local conditions at the test site. It needs to be noted that, due to the sparsity of monitoring sites (only 7 monitoring sites are available for spatial prediction in this case), the predicted source contribution surface by BSMRM is supposed to be smoother than the true surface. As the number of monitoring sites increases, spatial prediction of local peaks will be improved. Other than prediction of those local peaks, the source contributions predicted by BSMRM appear to be consistent with the source contributions estimated by BNFA (which may be viewed as the surrogate for the true source contributions at each test site) and correlations (R) seem to be reasonable. Considering the very small number of monitoring sites, spatial prediction of source contributions at an unmonitored site by BSMRM is deemed to be satisfactory

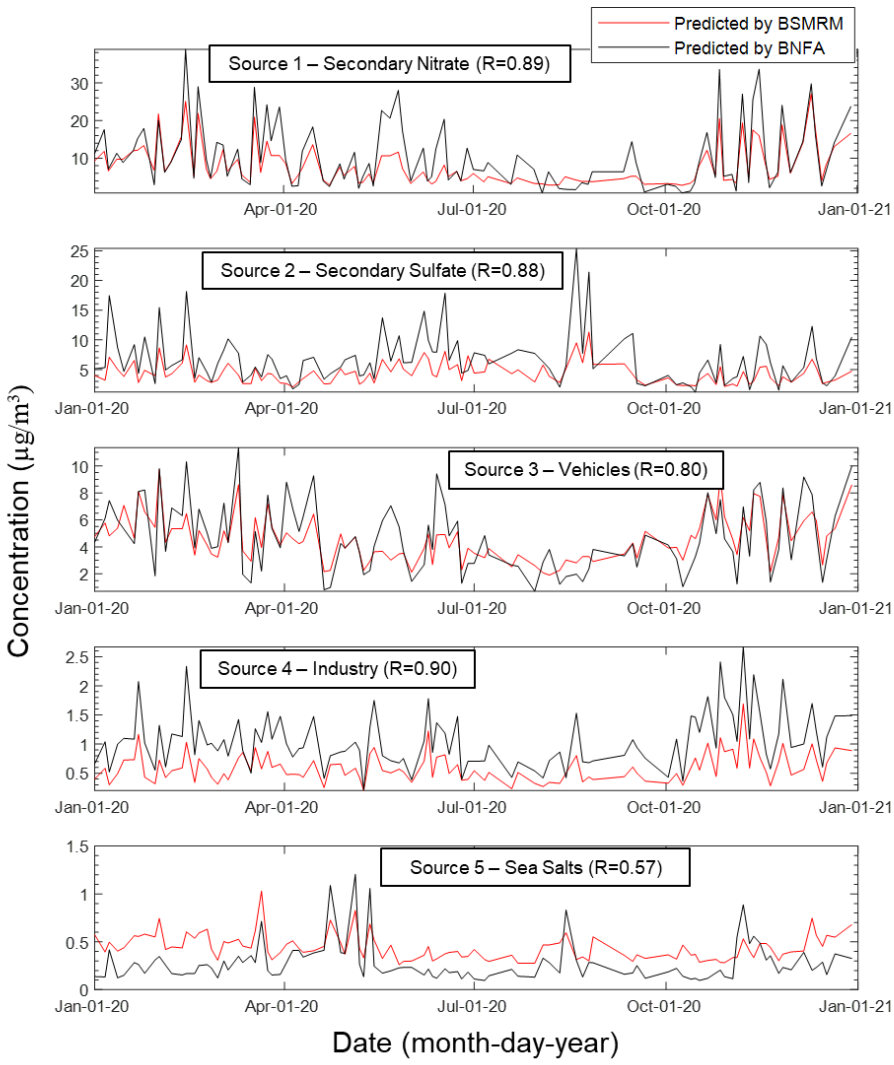


Fig. 5.9. Predicted source contributions by BSMRM (Model #2, red lines) and BNFA (Model #7, black lines) for Ansan City

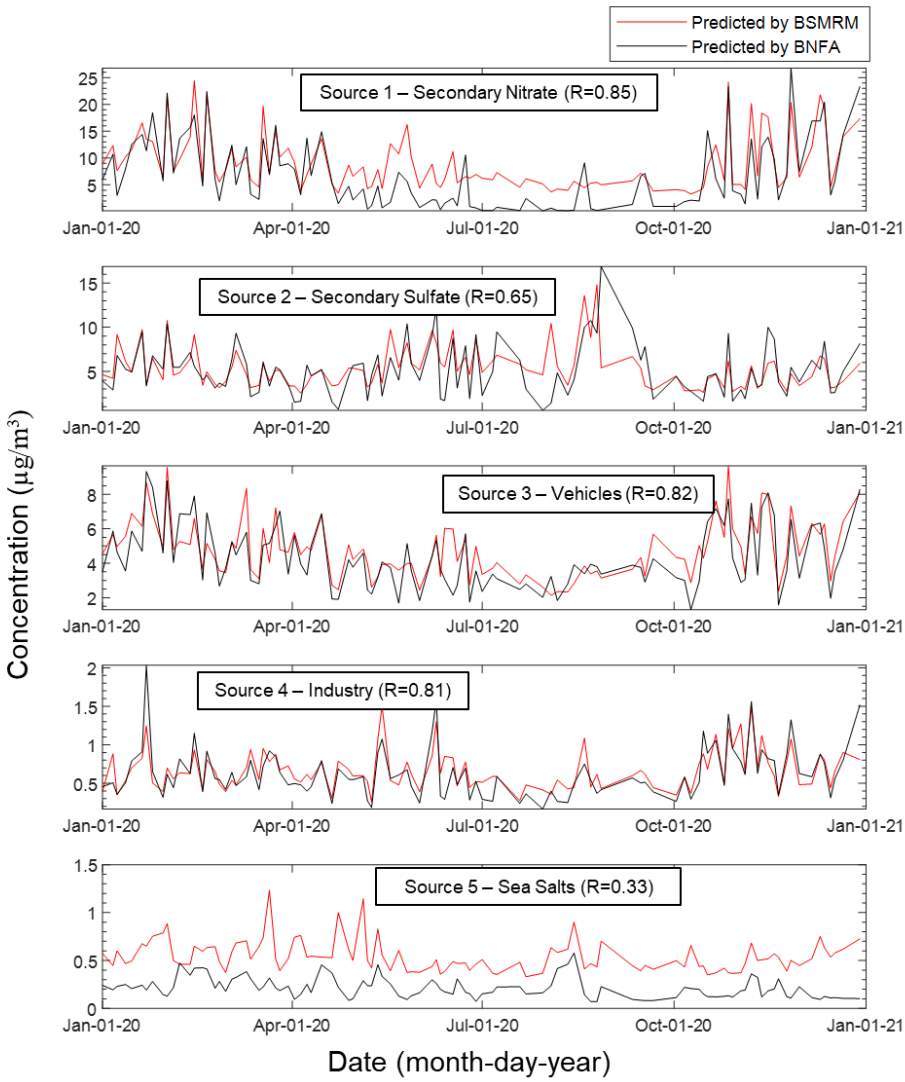


Fig. 5.10. Predicted source contributions by BSMRM (Model #2, red lines) and BNFA (Model #7, black lines) for Daejeon City

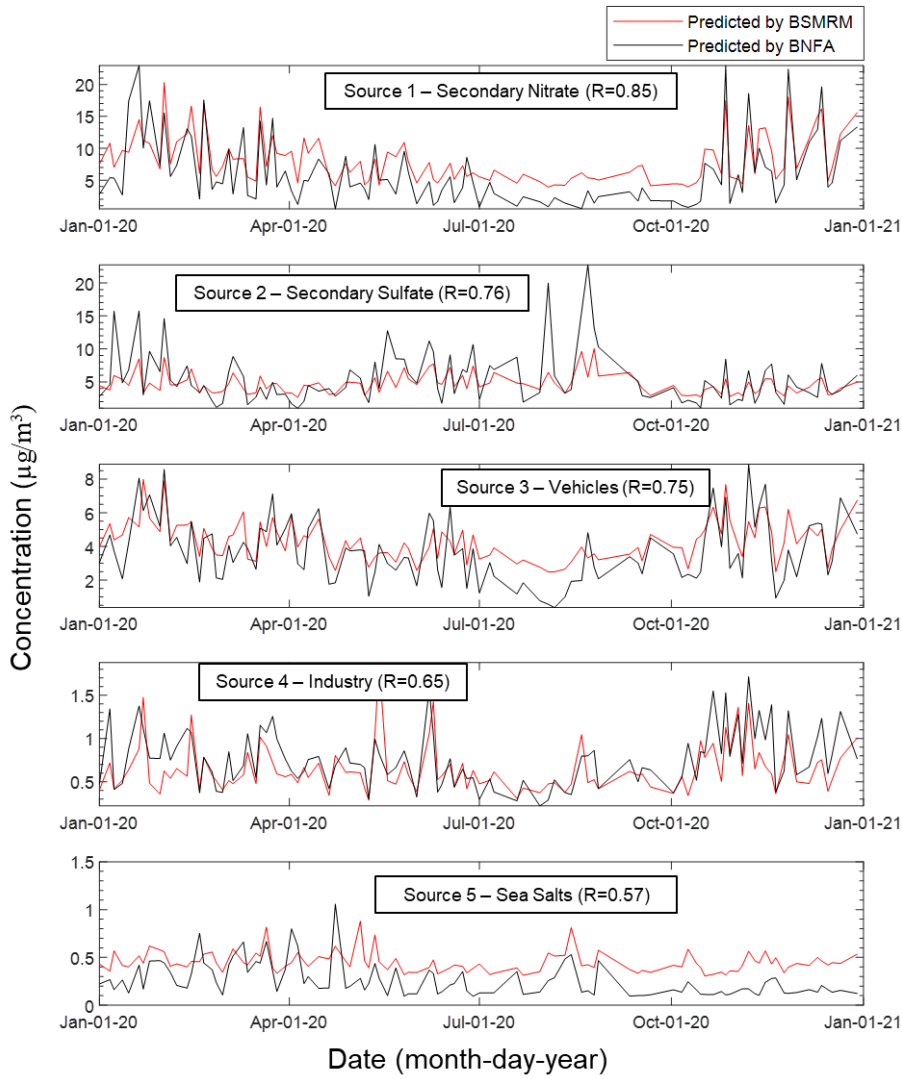


Fig. 5.11. Predicted source contributions by BSMRM (Model #2, red lines) and BNFA (Model #7, black lines) for Gwanju City

5.3.3 Spatial distribution of each source in South Korea

To examine spatial trends of source contributions in South Korea, we constructed the predicted source contribution surface maps using data from all of 8 monitoring sites. Fig. 5.12 and Fig. 5.13 shows the predicted source contribution surfaces for secondary nitrate and motor vehicle emissions for eight days in 2020, which show spatial and daily variations of contributions of secondary nitrate and motor vehicle emissions, respectively.

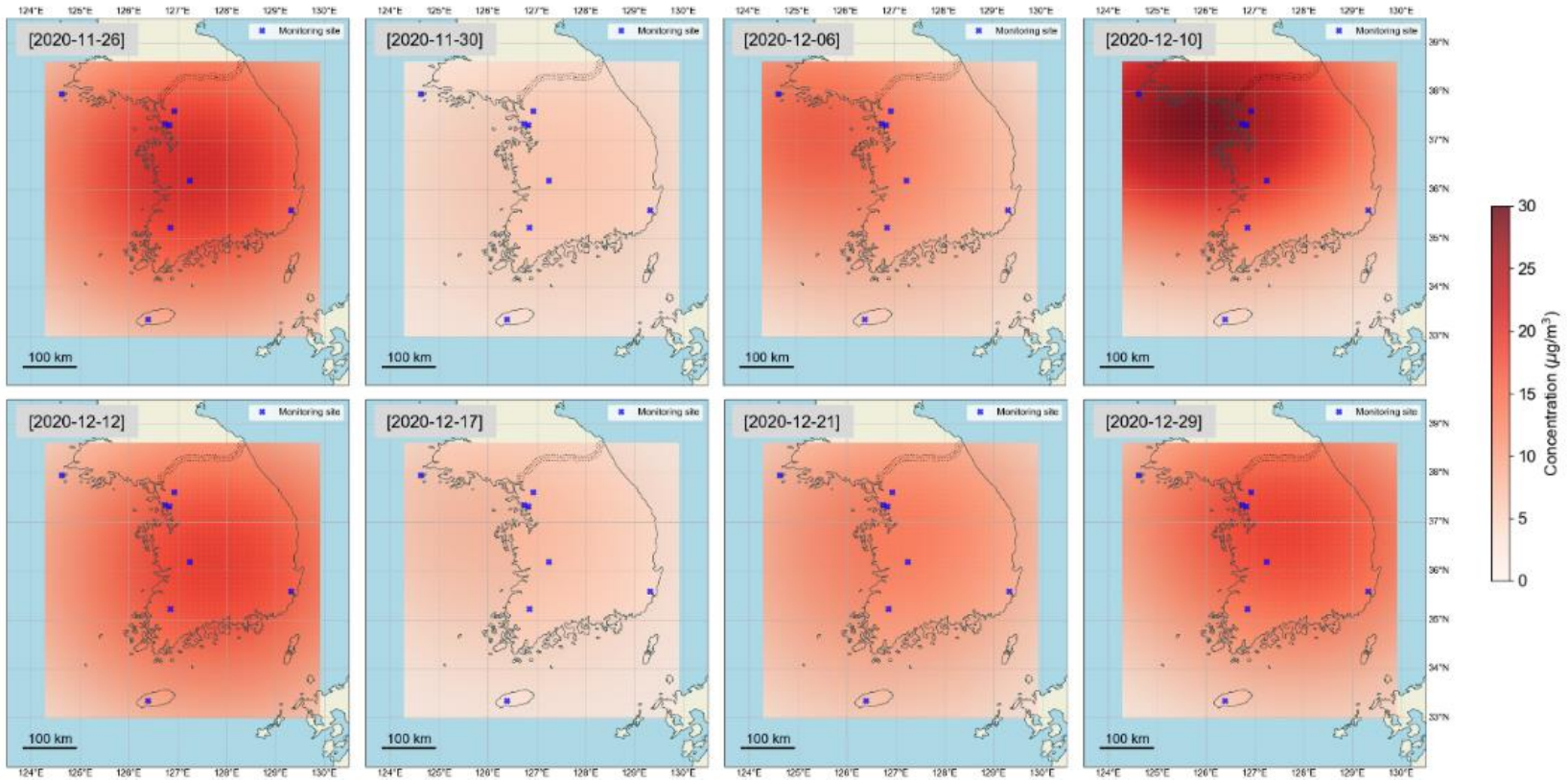


Fig. 5.12. Predicted source contribution surfaces of secondary nitrate for eight days

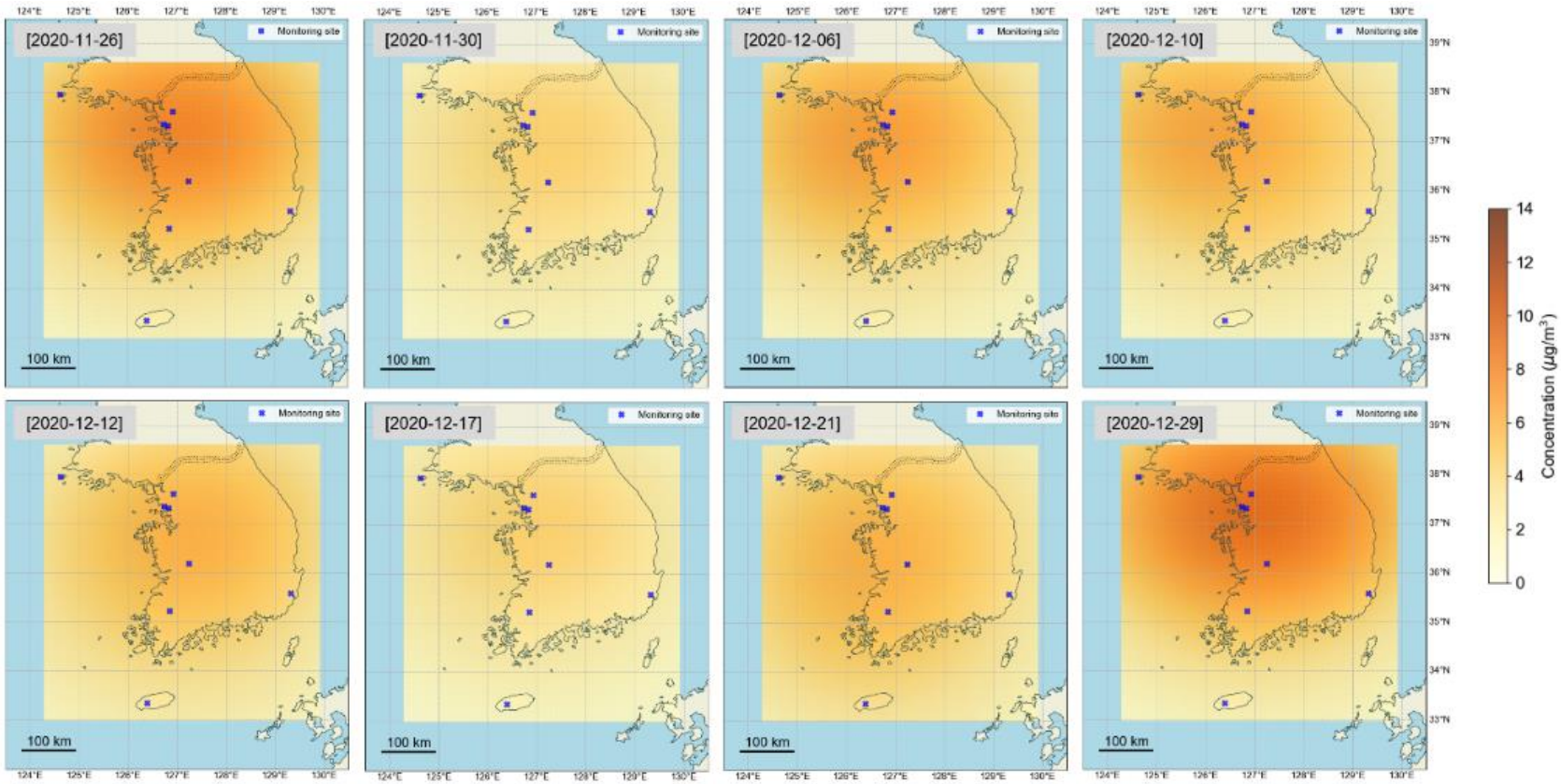


Fig. 5.13. Predicted source contribution surfaces of motor vehicle emission for eight days

5.4. Summary

In this paper, the source contributions for major sources of $PM_{2.5}$ in a regional scale were predicted and validated using BSMRM for the first time. We assessed the impact of major sources to ambient $PM_{2.5}$ concentrations in Korea and predicted source contribution surfaces using Bayesian spatial multivariate receptor modeling (BSMRM) based on multi-site $PM_{2.5}$ speciation data. Secondary Nitrate, Secondary Sulfate, Motor vehicle emission, Industry, and Sea Salt were determined to be significant contributors to ambient $PM_{2.5}$ concentrations in Korea. The source contributions predicted by BSMRM were also validated using the held-out data at a test site (using each of Ansan, Daejeon, and Gwangju, as a test site). Source contribution surface maps over the entire South Korea were also constructed. These predicted source contributions can greatly aid in developing effective $PM_{2.5}$ control strategies in cities where no speciated $PM_{2.5}$ monitoring stations are available. They can also be utilized as source-specific exposures in health effects studies even at the cities where no monitoring stations are available.

References

Calder, C.A., 2007. Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. Environ. Ecol. Stat. 14, 229–247. https://doi.org/10.1007/s10651-007-0019-y

Choi, E., Yi, S.M., Lee, Y.S., Jo, H., Baek, S.O., Heo, J.B., 2022. Sources of airborne particulate matter-bound metals and spatial-seasonal variability of health risk potentials in four large cities, South Korea. Environ. Sci. Pollut. Res. 29, 28359–28374. https://doi.org/10.1007/s11356-021-18445-8

Choi, J. kyu, Heo, J.B., Ban, S.J., Yi, S.M., Zoh, K.D., 2013. Source apportionment of PM2.5 at the coastal area in Korea. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2012.12.047

Dai, Q., Hopke, P.K., Bi, X., Feng, Y., 2020a. Improving apportionment of PM2.5 using multisite PMF by constraining G-values with a prioriinformation. Sci. Total Environ. 736. https://doi.org/10.1016/J.SCITOTENV.2020.139657

Dai, Q., Liu, B., Bi, X., Wu, J., Liang, D., Zhang, Y., Feng, Y., Hopke, P.K., 2020b. Dispersion normalized PMF provides insights into the significant changes in source contributions to PM2.5 after the CoviD-19 outbreak. Environ. Sci. Technol. 54, 9917–9927. https://doi.org/10.1021/acs.est.0c02776

Diao, M., Holloway, T., Choi, S., O'Neill, S.M., Al-Hamdan, M.Z., Van Donkelaar, A., Martin, R. V., Jin, X., Fiore, A.M., Henze, D.K., Lacey, F., Kinney, P.L., Freedman, F., Larkin, N.K., Zou, Y., Kelly, J.T., Vaidyanathan, A., 2019. Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models. J. Air Waste Manag. Assoc. 69, 1391–1414.

https://doi.org/10.1080/10962247.2019.1668498

Henry, R.C., Park, E.S., Spiegelman, C.H., 1999. Comparing a new algorithm with the classic methods for estimating the number of factors. Chemom. Intell. Lab. Syst. 48, 91–97. https://doi.org/10.1016/S0169-7439(99)00015-5

Heo, J.-B., Hopke, P.K., Yi, S.-M., 2009. Source apportionment of PM $_{2.5}$ in Seoul, Korea. Atmos. Chem. Phys. 9, 4957–4971. https://doi.org/10.5194/acp-9-4957-2009

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environ. Ecol. Stat. 5, 173–190. https://doi.org/10.1023/A:1009666805688

Hopke, P.K., 2021. Approaches to reducing rotational ambiguity in receptor modeling of ambient particulate matter. Chemom. Intell. Lab. Syst. 210, 104252. https://doi.org/10.1016/j.chemolab.2021.104252

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. https://doi.org/10.1080/10962247.2016.1140693

Hopke, P.K., Croft, D., Zhang, W., Lin, S., Masiol, M., Squizzato, S., Thurston, S.W., van Wijngaarden, E., Utell, M.J., Rich, D.Q., 2019. Changes in the acute response of respiratory diseases to PM2.5 in New York State from 2005 to 2016. Sci. Total Environ. 677, 328–339. https://doi.org/10.1016/J.SCITOTENV.2019.04.357

Hopke, P.K., Dai, Q., Li, L., Feng, Y., 2020. Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.140091

Hwang, I.J., Yi, S.M., Park, J., 2020. Estimation of Source Apportionment for Filter-

based PM2.5 Data using the EPA-PMF Model at Air Pollution Monitoring Supersites. J. Korean Soc. Atmos. Environ. 36, 620–632. https://doi.org/10.5572/KOSAE.2020.36.5.620

Jeong, J.H., Shon, Z.H., Kang, M., Song, S.K., Kim, Y.K., Park, J., Kim, H., 2017. Comparison of source apportionment of PM2.5 using receptor models in the main hub port city of East Asia: Busan. Atmos. Environ. 148, 115–127. https://doi.org/10.1016/j.atmosenv.2016.10.055

Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., Amann, M., 2015. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087

Kim, S., Kim, T.Y., Yi, S.M., Heo, J., 2018. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. J. Environ. Manage. 214, 325–334. https://doi.org/10.1016/j.jenvman.2018.03.027

Korea Ministry of Environment, National Institute of Environmental Research, 2021. Guidelines for Installation and Operation of National Air Pollution Monitoring Network.

Lee, T.-J., Park, M.-B., Kim, D.-S., 2019. Time Series Assessment of PM2.5 Source Contributions and Classification of Haze Patterns in Seoul. J. Korean Soc. Atmos. Environ. 35, 97–124. https://doi.org/10.5572/kosae.2019.35.1.097

Lee, Y.S., Kim, Y.K., Choi, E., Jo, H., Hyun, H., Yi, S.-M., Kim, J.Y., 2022. Health risk assessment and source apportionment of PM2.5-bound toxic elements in the industrial city of Siheung, Korea. Environ. Sci. Pollut. Res. 1, 1–14.

https://doi.org/10.1007/s11356-022-20462-0

Little, R.J.A., Rubin, D.B., 2014. Statistical Analysis with Missing Data, Wiley. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9781119013563

Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126. https://doi.org/10.1002/env.3170050203

Park, M. Bin, Lee, T.J., Lee, E.S., Kim, D.S., 2019. Enhancing source identification of hourly PM2.5 data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). Atmos. Pollut. Res. 10, 1042–1059. https://doi.org/10.1016/j.apr.2019.01.013

Park, E.H., Heo, J., Kim, H., Yi, S.-M., 2020. Long term trends of chemical constituents and source contributions of PM2.5 in Seoul. Chemosphere 251, 126371. https://doi.org/10.1016/j.chemosphere.2020.126371

Park, E.S., Guttorp, P., Kim, H., 2004. Locating major PM 10 source areas in Seoul using multivariate receptor modeling. Environ. Ecol. Stat. 11, 9–19. https://doi.org/10.1023/B:EEST.0000011361.33942.be

Park, E.S., Henry, R.C., Spiegelman, C.H., 2000. Estimating the number of factors to include in a high-dimensional multivariate bilinear model. Commun. Stat. - Theory Methods 29, 723–746. https://doi.org/10.1080/03610910008813637

Park, E.S., Hopke, P.K., Kim, I., Tan, S., Spiegelman, C.H., 2018. Bayesian Spatial Multivariate Receptor Modeling for Multisite Multipollutant Data. Technometrics 60, 306–318. https://doi.org/10.1080/00401706.2017.1366948

Park, E.S., Hopke, P.K., Oh, M.S., Symanski, E., Han, D., Spiegelman, C.H., 2014.

Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: A unified Bayesian approach. Biostatistics 15, 484–497. https://doi.org/10.1093/biostatistics/kxu004

Park, E.S., Lee, E.K., Oh, M.S., 2021. Bayesian multivariate receptor modeling software: BNFA and bayesMRM. Chemom. Intell. Lab. Syst. 211, 104280. https://doi.org/10.1016/j.chemolab.2021.104280

Park, E.S., Oh, M.S., 2018. Accounting for uncertainty in source-specific exposures in the evaluation of health effects of pollution sources on daily cause-specific mortality. Environmetrics 29. https://doi.org/10.1002/ENV.2484

Park, E.S., Oh, M.S., 2015. Robust Bayesian multivariate receptor modeling. Chemom. Intell. Lab. Syst. 149, 215–226. https://doi.org/10.1016/j.chemolab.2015.08.021

Park, E.S., Oh, M.S., Guttorp, P., 2002. Multivariate receptor models and model uncertainty. Chemom. Intell. Lab. Syst. 60, 49–67. https://doi.org/10.1016/S0169-7439(01)00185-X

Park, E.S., Tauler, R., 2020. Bayesian Methods for Factor Analysis in Chemometrics. Compr. Chemom. 355–369. https://doi.org/10.1016/B978-0-12-409547-2.14876-0

Park, M.H., Ju, M., Kim, J.Y., 2020. Bayesian approach in estimating flood waste generation: A case study in South Korea. J. Environ. Manage. 265, 110552. https://doi.org/10.1016/J.JENVMAN.2020.110552

Polissar, A. V., Hopke, P.K., Harris, J.M., 2001. Source regions for atmospheric aerosol measured at Barrow, Alaska. Environ. Sci. Technol. 35, 4214–4226.

https://doi.org/10.1021/es0107529

Shi, G., Liu, J., Zhong, X., 2021. Spatial and temporal variations of PM2.5 concentrations in Chinese cities during 2015-2019. Int. J. Environ. Health Res. https://doi.org/10.1080/09603123.2021.1987394

Shi, X., Nenes, A., Xiao, Z., Song, S., Yu, H., Shi, G., Zhao, Q., Chen, K., Feng, Y., Russell, A.G., 2019. High-Resolution Data Sets Unravel the Effects of Sources and Meteorological Conditions on Nitrate and Its Gas-Particle Partitioning. Environ. Sci. Technol. 53, 3048–3057. https://doi.org/10.1021/acs.est.8b06524

Sokhi, R.S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., Halenka, T., others, 2021. Advances in Air Quality Research--Current and Emerging Challenges. Atmos. Chem. Phys. Discuss. 1–133. https://doi.org/10.5194/ACP-2021-581

Wang, P., Shen, J., Zhu, S., Gao, M., Ma, J., Liu, J., Gao, J., Zhang, H., 2021. The aggravated short-term PM2.5-related health risk due to atmospheric transport in the Yangtze River Delta. Environ. Pollut. 275, 116672. https://doi.org/10.1016/J.ENVPOL.2021.116672

Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., Yu, J.Z., 2018. Source Apportionment of PM2.5 Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-Resolution Influence. J. Geophys. Res. Atmos. 123, 5284–5300. https://doi.org/10.1029/2017JD027877

Chapter 6. Conclusions and future work

6.1. Conclusions

The sources and chemical constituents of $PM_{2.5}$ pollution were thoroughly investigated. This study aimed to use source apportionment models and their spatiotemporal analysis for an effective $PM_{2.5}$ management strategy. Detailed objectives and a summary of the results are described in each chapter. The conclusions of this thesis corresponding to each goal are as follows:

(1) The sources of $PM_{2.5}$ and their contributions in a medium-sized industrial city, Siheung, South Korea, were identified using positive matrix factorization modeling. Ten sources were secondary nitrate (24.3%), secondary sulfate (18.8%), traffic (18.8%), combustion for heating (12.6%), biomass burning (11.8%), coal combustion (3.6%), heavy oil industry (1.8%), smelting industry (4.0%), sea salt (2.7%), and soil (1.7%). Based on the source apportionment results, health risks by inhalation of $PM_{2.5}$ were assessed for each source using the concentration of toxic elements mentioned. The estimated cumulative carcinogenic health risks from coal combustion, heavy oil industry, and traffic sources exceeded the benchmark 1E-06. Similarly, carcinogenic health risks from exposure to As and Cr exceeded 1E-05 and 1E-06, respectively, requiring a risk-reduction plan. The carcinogenic risk of $PM_{2.5}$ in Siheung was similar to or lower than that of mega-cities in Northeast Asia. The non-carcinogenic risk was lower than the hazard index of 1, implying a low potential for adverse health effects. The probable locations of sources with relatively high carcinogenic risks

were tracked. It is noteworthy that the mass contribution and health risks of each source in $PM_{2.5}$ were different. These results highlight the importance of $PM_{2.5}$ management focusing on health risks. This type of research evaluating $PM_{2.5}$ health risks from sources is rare in South Korea, and it is necessary to apply this method to other cities to evaluate their health risks from $PM_{2.5}$.

(2) The feature extraction capabilities of the four ML models to predict the chemical composition of $PM_{2.5}$ were assessed by comparing the prediction accuracy depending on input variables, target constituents for prediction, available period, missing ratios of input data, and study sites. The concentrations of $PM_{2.5}$, which are important and essential information for the identification of air pollution sources, were predicted at three sites (Seoul, Ulsan, and Baengnyeong) in South Korea between 2016 and 2018 using four machine learning (ML) models: generative adversarial imputation network (GAIN), fully connected deep neural network (FCDNN), random forest (RF), and k-nearest neighbor (kNN). Three $PM_{2.5}$ constituent groups were targeted for prediction, including eight ions, two carbons, and 15 trace elements. The latest hyperparameter optimization techniques were used to learn air pollution characteristics from ambient $PM_{2.5}$-related information, such as time, meteorology, and air pollutant concentrations. We compared the feature-extraction abilities of the four models. The prediction accuracy identified by the coefficient of determination (R^2) between the prediction and observation was highest in

GAIN, followed by FCDNN, RF, and kNN. Based on the availability of data on time, air pollutant concentrations, and meteorology, or all, 20% of the data of all $PM_{2.5}$ constituent groups were predicted, with R^2 = 0.897, 0.861, 0.785, and 0.801 by the GAIN, FCDNN, RF, and kNN, respectively. As the missing ratios (20, 40, 60, and 80%) of the input data increased, the prediction accuracy decreased in the four models and was predominantly more noticeable in GAIN and kNN. As the available period of data increased, the prediction accuracy increased for the GAIN and FCDNN. Trace elements were predicted to have the lowest R^2 among the target constituent groups in all the models. The study sites with more emission sources showed lower prediction accuracy, resulting in the highest R^2 in Baengnyeong Island and the lowest in Ulsan. The missing values of $PM_{2.5}$ chemical constituents could be predicted successfully using machine learning models.

(3) The source contributions for major sources of $PM_{2.5}$ on a regional scale, including unmonitored sites, were predicted and validated using Bayesian spatial multivariate receptor modeling (BSMRM) as the first study. The spatial distributions of five $PM_{2.5}$ sources in South Korea were estimated using BSMRM, which incorporates spatial correlation in data into modeling and estimation for spatial prediction of latent source contributions. Secondary nitrate, secondary sulfate, motor vehicle emissions, industry, and sea salt were identified as significant contributors to $PM_{2.5}$ concentrations in South Korea. The distribution of the daily average contribution for each

source in South Korea was derived from measurement data from the eight monitoring sites. The validity of the BSMRM results was also assessed based on the data from the test site (city), which were not used in model development and estimation as part of cross-validation. The results of the validation indicated that the use of the Bayesian spatial multivariate receptor model was appropriate, with high accuracy. In addition, the uncertainty of the source contributions was quantified, including unmonitored sites, which is not possible in other receptor models. The results of this study could be used to develop effective management strategies for $PM_{2.5}$.

6.2. Future work

Limitations or research needs for each chapter are suggested as future works.

(1) In Chapter 3, the health risks are evaluated only for $PM_{2.5}$ constituents toxicity data available (such as heavy metals). Therefore, the toxicity of some species, such as organic carbon and ionic constituents, was not considered. If the toxicity values of other constituents are reflected, the health effect of $PM_{2.5}$ will be estimated to be larger. This indicates that the health effects of $PM_{2.5}$ could be treated more seriously. Further studies that include other constituents are required.

(2) In Chapter 4, the increased usability of the missing-value corrected data using the methodology of this study was not evaluated. Missing-value corrected data can help improve the reliability of receptor models for source apportionment, such as PMF and Bayesian multivariate receptor models. Further research is needed to investigate the reliability of missing-value corrected data using machine learning models in source apportionment research.

(3) In Chapter 5, the predicted source contribution surface is assumed to be smoother than the true surface because of the sparsity of the monitoring sites (limitation of the data). As the number of available monitoring sites increases, the spatial prediction of local peaks improves. The number of sources that can be predicted also increased. Further research is required to increase the number of measurement sites and the period of data collection.

국문 초록(Abstract in Korean)

직경 2.5 μm 이하의 입자상 물질인 초미세먼지는 대기중에 존재하며, 건강에 미치는 악영향으로 인해 수십 년 동안 세계적으로 관심의 대상이 되고 있는 대기오염물질이다. 초미세먼지를 효과적으로 관리하기 위해서는 다양한 시간과 공간에 대해 초미세먼지의 오염원 유형을 파악하고, 각 유형별 기여도를 정량화하는 것이 중요하다. 따라서, 초미세먼지의 오염원 추정은 핵심 과제로 다뤄져 왔으며, 통계학적 방법론을 적용해 오염원을 추정하는 수용모델이 많이 활용되고 있다.

본 연구에서는 초미세먼지의 세부 특성을 파악하기 위해 오염원 추정과 추정된 오염원의 시공간 분석을 수행하였으며, 이를 통해 효과적인 초미세먼지 관리 방안 마련에 중요한 정보를 제공하는 것을 목적으로 하였다. 오염원 유형 추정 연구를 위해, 두 가지 모델링이 수행되었다. 첫번째는 양행렬 인자 분석(Positive matrix factorization, PMF) 모델링으로, 이는 한 장소에서 초미세먼지의 오염원 유형을 구체적으로 추정하기 위해 활용되었다. 두번째는 베이지안 다변량 수용모델링(Bayesian spatial multivariate receptor modelingm, BSMRM)으로, 이는 다수의 측정 지점으로부터 넓은 범위의 면적에 대해 주요 오염원 유형을 추정하기 위해 활용되었다. 또한, 기계학습 모델들을 활용하여 초미세먼지 오염원 유형 추정에 가장 중요한 자료로

활용되는 초미세먼지 화학성분 농도를 예측하였다. 기계학습 모델을 초미세먼지 화학성분 자료에 대해 활용가능한지를 검토하였고, 이를 통해 초미세먼지 화학성분 자료의 무결성을 향상시키고자 하였다.

PMF 모델링을 통해, 대한민국 시흥시의 초미세먼지 오염원 유형 10가지를 도출하였다. 이는 각각 2차 생성 질산염(24.3%), 2차 생성 황산염(18.8%), 이동 오염원(18.8%), 난방연소(12.6%), 생물체 연소(11.8%), 석탄 연소(3.6%), 중유 관련 산업 오염원(1.8%), 제련 관련 산업 오염원(4.0%), 해염 입자(2.7%), 토양(1.7%)였다. 도출된 오염원 유형별로, 초미세먼지 호흡에 따른 건강 영향을 평가하였다. 석탄 연소, 중유 관련 산업 오염원, 이동 오염원의 초미세먼지 기여도는 낮았지만, 이로 인한 발암 위해도는 10E-6 이상으로 나타났다. 따라서, 초미세먼지의 질량농도 감축 중심의 대응만이 아닌, 오염원별 건강영향 중심의 대응이 요구된다.

기계학습 모델의 초미세먼지 화학성분 예측 능력을 평가하기 위해 4가지 기계학습 모델에 대해 입력 자료 수준, 예측 대상 성분, 입력 자료 기간, 입력 자료의 결측 비율, 자료 대상 지역을 변화하며 예측 정확도를 비교 평가하였다. GAIN(Generative Adversarial Imputation Network), FCDNN(Fully Connected Deep Neural Network), Random forest(RF), kNN(k-nearest neighboring) 모델의 4가지 기계학습 모델을 한국의 3개 지역(서울, 울산, 백령)의 2016년부터 2018년까지의 초미세먼지 화학 성분 자료에 대해 적용하여

농도를 예측하였다. 예측값과 관측값 사이의 결정계수를 통해 정확도를 비교한 결과, 예측 정확도는 GAIN이 가장 높았고, FCDNN, RF 또는 kNN 순서로 나타났다. 입력 자료의 결측률이 20%에서 80%까지 증가함에 따라 예측 정확도는 모든 모델에서 감소하였으나, 비지도 기계학습 모델인 GAIN과 kNN에서 감소 폭이 더 크게 나타났다. 입력 자료의 기간이 길어질수록, 딥러닝 모델인 GAIN과 FCDNN이 다른 두 모델인 RF와 kNN보다 예측 정확도 증가 폭이 더 컸다. 예측 대상 지역별로는, 자체 배출원이 많은 울산의 경우가 예측 정확도가 가장 낮게 나타났고, 자체 배출원의 영향이 거의 없는 백령도의 경우 예측 정확도가 가장 높게 나타났다. 대상 성분별로는 이온 성분이 예측 정확도가 높게 나타났고, 미량원소 성분은 예측 정확도가 낮았다. 본 연구는 기계학습 모델의 예측 정확도를 다양한 실험 조건에 따라 평가하여 대기오염 분야에서의 기계학습 모델의 적용 가능성을 평가했다.

베이지안 다변량 수용 모델링(BSMRM)을 통해서는 8개의 관측 지점 자료를 통해 우리나라의 주요 초미세먼지 오염원 5가지를 도출하고, 각각 오염원 유형별 기여도를 우리나라 전체에 대한 공간 분포를 추정하였다. 5가지 오염원은 각각 2차 질산염, 2차 황산염, 자동차 배출, 산업 오염원, 해염 입자였다. 각 오염원 유형별 일평균 기여도 농도를 지도에 공간적으로 표현할 수 있었다. 또한, BSMRM을 통해 예측한 오염원 유형별 기여도의 타당성 검토를 위해 테스트 사이트(안산, 대전, 광주)의 자료는 각각 제외된 모델링을 수행하여

결과를 서로 비교하여 모델의 정확도를 확인하였다. 이처럼 공간적으로 추정된 오염원 유형 기여도는 초미세먼지 화학성분을 측정하지 않는 도시에서 초미세먼지 대응 방안을 수립하는데 큰 도움이 될 수 있다. 즉, 8개의 측정 자료만으로 우리나라 전체에 대해 예측한 결과를 통해, 측정 지점이 없는 모든 도시에 대해 추정이 가능하였으며, 이 결과는 건강 영향 평가와 같은 추가 연구에도 활용될 수 있다.

주요어 : 초미세먼지; 오염원 추정; 양행렬 인자분석; 기계학습 모델링; 초미세먼지 화학성분; 베이지안 수용모델

학 번 : 2019-32839