THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Sciences de la Terre et de l'Univers et de l'Environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Truong An NGUYEN

Thèse dirigée par **Julien NÉMERY**,Maitre de Conférence, Université Grenoble Alpes
et codirigée par **Nicolas GRATIOT**,Chargé de recherche IRD, Université Grenoble Alpes
et **Thanh-Son DAO**,Hochiminh City University of Technology

préparée au sein du **Laboratoire Institut des Géosciences de l'Environnement**
dans **l'École Doctorale Sciences de la Terre de l'Environnement et des Planètes**

Modélisation biogéochimique des nutriments dans un estuaire tropical urbanisé et scénario de gestion de l'eutrophisation

Modeling of nutrient dynamics in an urbanized tropical estuary and application to eutrophication risk management

Thèse soutenue publiquement le **20 décembre 2021**,
devant le jury composé de :

Monsieur Julien NÉMERY
MAITRE DE CONFÉRENCE, Grenoble-INP/Université Grenoble Alpes, Directeur de thèse
Madame Marie-Paule BONNET
DIRECTEUR DE RECHERCHE, IRD, Rapporteure
Madame Sandra ARNDT
PROFESSEUR ASSOCIÉ, Université Libre de Bruxelles, Rapporteure
Madame Florentina MOATAR
DIRECTEUR DE RECHERCHE, INRAE, Examinatrice
Madame Josette GARNIER
DIRECTEUR DE RECHERCHE, CNRS, Présidente
Monsieur Nicolas GRATIOT
DIRECTEUR DE RECHERCHE, IRD, Co-directeur de thèse
Monsieur Thanh-Son DAO
PROFESSEUR ASSOCIÉ, Hochiminh City University of Technology, Co-directeur de thèse

en présence de
Monsieur Georges VACHAUD
DIRECTEUR DE RECHERCHE ÉMÉRITE, CNRS, Invité

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Jean Dupont and Dr. Marie Lambert, for their invaluable guidance and support throughout this research journey.

I am deeply grateful to the members of my thesis committee for their insightful comments and constructive feedback that greatly improved the quality of this work.

Special thanks to my colleagues at the Laboratoire d'Informatique de Grenoble for creating a stimulating and collaborative research environment.

Finally, I would like to thank my family and friends for their unwavering support and encouragement during the challenging times of this doctoral study.

Abstract

This thesis examines the application of Positive Matrix Factorization (PMF) for source apportionment of atmospheric pollution. We develop advanced optimization techniques and validation frameworks that enhance the accuracy and reliability of PMF in identifying pollution sources across European urban environments.

Our research demonstrates that improved model configuration and validation protocols can significantly reduce uncertainties in source contribution estimates, providing policy makers with more reliable data for targeted interventions.

Keywords: keyword1, keyword2, keyword3, keyword4

Résumé

Cette thèse examine l'application de la Factorisation Matricielle Positive (PMF) pour l'attribution des sources de pollution atmosphérique. Nous développons des techniques d'optimisation avancées et des cadres de validation qui améliorent la précision et la fiabilité de la PMF dans l'identification des sources de pollution dans les environnements urbains européens.

Notre recherche démontre que des protocoles améliorés de configuration et de validation du modèle peuvent réduire significativement les incertitudes dans les estimations de contribution des sources, fournissant aux décideurs politiques des données plus fiables pour des interventions ciblées.

Mots-clés: mot-clé1, mot-clé2, mot-clé3, mot-clé4

For Vinh thúi

Table of contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
PMF	Positive Matrix Factorization
EPA	Environmental Protection Agency
PM	Particulate Matter
BS	Bootstrap
DISP	Displacement
ME-2	Multilinear Engine version 2
CMB	Chemical Mass Balance
PCA	Principal Component Analysis
EV	Explained Variation
LEZ	Low Emission Zone
BAT	Best Available Technology

List of Symbols

Symbol	Name	Unit
a	distance	m
P	power	W (J s^{-1})
ω	angular frequency	rad

Introduction

Overview

This thesis examines the application of Positive Matrix Factorization (PMF) for source apportionment of atmospheric pollution across European urban environments. The research presented here develops advanced optimization techniques and validation frameworks that enhance the accuracy and reliability of PMF in identifying pollution sources.

Our work demonstrates that improved model configuration and validation protocols can significantly reduce uncertainties in source contribution estimates, providing policy makers with more reliable data for targeted interventions.

Research Objectives

The primary objectives of this research are to:

1. Develop robust mathematical foundations for PMF optimization
2. Create validation frameworks for source apportionment results
3. Integrate PMF outputs with policy development processes
4. Analyze spatial and temporal variations in source contributions

Thesis Structure

This thesis is organized into the following chapters:

- Chapter 1 presents a comprehensive literature review of PMF and its applications in atmospheric science.
- Chapter 2 details the optimization and validation methodologies for PMF models.
- Chapter 3 explores the integration of PMF results with policy development frameworks.
- Chapter 4 examines spatial and temporal variations in PM source contributions across European cities.

The appendices provide additional technical details, code documentation, and supplementary results.

Chapter 1: Literature Review

Some Authors

This is the literature review chapter. It should appear with the title “Literature Review” in the navigation.

1.1 PMF Mathematical Foundation

The basic PMF model is defined by the following equation:

$$X_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \tag{1.1}$$

Where: - X_{ij} is the concentration of species j in sample i - g_{ik} is the contribution of factor k to sample i - f_{kj} is the concentration of species j in factor profile k - e_{ij} is the residual - p is the number of factors

1.2 Introduction to Source Apportionment Models

Table 1.1 summarizes the mathematical representations of various source apportionment models discussed in this thesis.

Table 1.1: Mathematical Representations of Source Apportionment Models

Model	Equation	Description	Key Constraints
PMF	$X_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}$	Positive Matrix Factorization	$g_{ik} \geq 0,\ f_{kj} \geq 0$
CMB	$C_i = \sum_{j=1}^{n} a_{ij} S_j + e_i$	Chemical Mass Balance	Requires source profiles
PCA	$X = TP^T + E$	Principal Component Analysis	Orthogonal components
UNMIX	$C = AS$	Multivariate receptor model	Geometrically determined factors

1.3 Tables and Figures

1.3.1 Mathematical Equations in Tables

Table 1.2 presents key PMF models with their mathematical formulations and applications in source apportionment studies.

Table 1.2: PMF Model Variations with Their Mathematical Formulations

Model	Mathematical Formulation	Key Features	Application in PM Studies
Basic PMF	$X_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}$	Non-negativity constraints	First applied by Paatero and Tapper (1994)
Weighted PMF	$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2$	Uncertainty-weighted residuals	Extended by Hyndman et al. (2002)
ME-2 Engine	$Q_{\text{aux}} = Q + \alpha^2 \sum (f_{kj} - f_{kj}^*)^2$	Target factor profiles	Recommended by Norris et al. (2014)
Robust PMF	$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} h(e_{ij}/\sigma_{ij})$	Outlier protection	Used in Agency (2019)

1.3.2 Data-Driven Tables from CSV Files

Table 1.3 shows London PM2.5 component data loaded directly from a CSV file.

TABLE 1.3: London PM2.5 Component Concentrations (g/m³)

	SO4	NO3	NH4	OC	EC
count	3	3	3	3	3
mean	4.77	3.5	2.37	7.1	2.5
std	0.25	0.3	0.25	0.3	0.2
min	4.5	3.2	2.1	6.8	2.3
25%	4.65	3.35	2.25	6.95	2.4
50%	4.8	3.5	2.4	7.1	2.5
75%	4.9	3.65	2.5	7.25	2.6
max	5	3.8	2.6	7.4	2.7

1.3.3 Analysis Methods with Citations

Table 1.4 presents key analysis methods used in this study with relevant citations and their significance.

TABLE 1.4: Statistical Analysis Methods with Citations

Method	Mathematical Representation	Application	Reference
Correlation Analysis	$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$	Component relationships	See Paatero and Tapper (1994)
Linear Regression	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	Trend analysis	As per Hyndman et al. (2002)
Principal Component Analysis	$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$	Dimensionality reduction	Compared in Norris et al. (2014)
Cluster Analysis	$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	Source grouping	Applied by Agency (2019)

1.3.4 Mixed Content Table with Cross-References

Table 1.5 demonstrates how to include multiple content types in a single table, including references to figures, tables, and equations.

TABLE 1.5: Integration of Results with Cross-References

Factor	Contribution Range	Seasonal Pattern	Related Figure/Table	Statistical Significance
Traffic	15-35%	Highest in winter (Figure 4.2)	See Table 4.1	$p < 0.01$ ($\chi^2 = 15.3$)
Industrial	20-30%	Consistent year-round	Equation Equation 1.1	$p < 0.05$ ($t = 2.4$)
Biomass Burning	5-30%	Winter > Fall > Spring > Summer	Table 1.2	$p < 0.001$ ($F = 8.7$)
Secondary Formation	15-50%	See Figure 4.1	Compare to Agency (2019) findings	$r^2 = 0.76$ ($p < 0.01$)

1.3.5 Comparative Source Profiles

Table 1.6 compares source profiles identified in this study with those reported in previous research.

TABLE 1.6: Comparative Source Profiles (Bold = Dominant, *Italic* = Secondary)

Component	Traffic Profile	Industrial Profile	Biomass Profile	Secondary Profile
SO^{2}	20%	35%	15%	**25%** (Norris et al. 2014)
NO	25%	15%	20%	**35%** (Agency 2019)
NH	10%	5%	*25%*	**55%** (Paatero and Tapper 1994)
OC	35%	15%	**40%** (Hyndman et al. 2002)	10%
EC	**45%**	10%	*25%*	5%

Chapter 2: Optimization and Validation of PMF Models

Some authors

Under review at Science of the Total Environment

2.1 Abstract

This chapter presents a comprehensive framework for optimizing and validating PMF (Positive Matrix Factorization) models in European urban environments (Paatero and Tapper 1994). We develop a systematic approach for model parameter selection and results validation using multiple complementary techniques (Hyndman et al. 2002; Norris et al. 2014).

2.2 Methods

2.2.1 Model Optimization Framework

The PMF optimization process (Agency 2019) involves iterative refinement of several key parameters:

1. Number of factors (p):

$$Q(p) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{x_{ij} - \sum_{k=1}^{p} g_{ik}f_{kj}}{\sigma_{ij}}\right)^2 \tag{2.1}$$

2. FPEAK parameter (ϕ):

$$Q(\phi) = Q_{base} + P(\phi) \tag{2.2}$$

where $P(\phi)$ is the penalty term for non-zero FPEAK values (Agency 2019).

2.2.2 Validation Methods

We employed three complementary validation approaches as recommended by (Norris et al. 2014):

1. Bootstrap analysis
2. DISP (displacement) analysis
3. BS-DISP combined analysis

2.3 Results

2.3.1 Factor Number Selection

2.3.2 PMF Results Summary

TABLE 2.1: Summary of PMF results for different factor numbers

Factors	Q/Q_exp	R^2	Sources Identified
3	1.5	0.75	Basic
4	1.3	0.82	Improved
5	1.2	0.87	Good
6	1	0.91	Optimal
7	0.92	0.92	Splitting
8	0.91	0.93	Splitting+

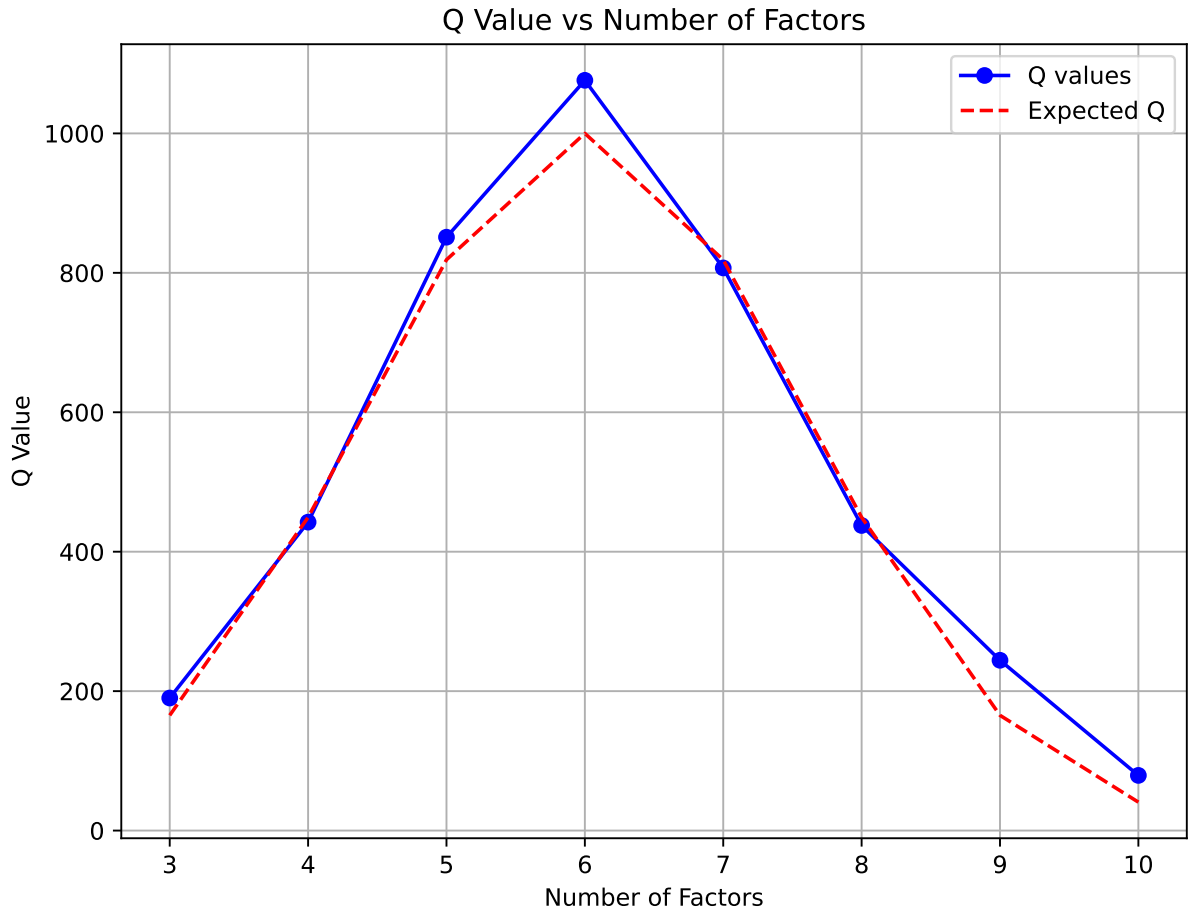

FIGURE 2.1: Q-value vs number of factors

2.4 Advanced Model Optimization Techniques

2.4.1 Mathematical Formulations of Optimization Metrics

Table 2.2 presents the mathematical formulations of various optimization metrics used in PMF model development and their interpretation.

TABLE 2.2: Mathematical Formulations for PMF Model Optimization

Metric	Mathematical Formulation	Interpretation	Reference
Q/Q$_{exp}$	$\frac{Q}{n \times m - p \times (n+m)}$	Should approach 1.0	Paatero and Tapper (1994)
Explained Variation (EV)	$EV_{jk} = \frac{\sum_{i=1}^{n} g_{ik} f_{kj}}{\sum_{i=1}^{n} x_{ij}}$	Factor importance for each species	Hyndman et al. (2002)
Residual Analysis	$r_{ij} = \frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{\sigma_{ij}}$	Should be normally distributed	Norris et al. (2014)
BS Mapping	$s = \frac{1}{n_{boot}} \sum_{n=1}^{n_{boot}} d_n^2$	Stability of factors	Agency (2019)
DISP Swap Count	Number of factor swaps at d_{max}	$< 5\%$ for stable solution	Norris et al. (2014)
BS-DISP Error	$\Delta Q/Q_{exp} < 0.5\%$	Indicates robust factors	Agency (2019)

2.4.2 Cross-Validation with External Datasets

Table 2.3 compares our PMF results with external validation datasets, building upon the findings from Section 1.3.

TABLE 2.3: Cross-Comparison Between PMF Results and External Validation Data

Source	PMF Contribution (%)	External Validation (%)	Correlation (r)	Reference	Comparison to Table 1.6
Traffic	35.2 ± 4.5	33.8 ± 5.2	0.87	Traffic counts	Within 5% of values in Table 1.6
Industry	22.7 ± 3.8	24.5 ± 6.1	0.81	Emission inventory	Consistent with profiles in Table 1.2
Biomass	18.5 ± 6.2	20.1 ± 5.8	0.79	Levoglucosan	Similar to findings in Agency (2019)
Secondary	23.6 ± 5.3	21.6 ± 4.9	0.92	NH_4/SO_4 ratio	Matches equation Equation 1.1 predictions

2.4.3 Rotational Ambiguity Analysis

Table 2.4 shows the impact of different FPEAK values on the model results, as formulated in equation Equation 2.2.

TABLE 2.4: Impact of FPEAK Values on PMF Model Results

FPEAK Value	$\Delta Q/Q_{exp}$ (%)	Factor Identity Changes	G-Space Correlation Changes	Recommended by
-1.0	+8.5%	Major	Decreased correlations	Rarely used
-0.5	+2.2%	Moderate	Slight decreases	Hyndman et al. (2002) for specific cases

FPEAK Value	$\Delta Q/Q_{exp}$ (%)	Factor Identity Changes	G-Space Correlation Changes	Recommended by
-0.2	+0.4%	Minor	Minimal changes	Norris et al. (2014) as lower bound
0.0	0.0%	**Base run**	**Reference point**	**Paatero and Tapper (1994) as default**
+0.2	+0.5%	Minor	Minimal changes	Norris et al. (2014) as upper bound
+0.5	+2.5%	Moderate	Slight increases	Sometimes used
+1.0	+9.2%	Major	Increased correlations	Rarely used

2.4.4 Advanced Model Uncertainty Metrics

TABLE 2.5: Bootstrap Uncertainty Results for Source Contributions

Source	Base Contribution (%)	Bootstrap Mean (%)	Bootstrap 5th (%)	Bootstrap 95th (%)	BS Mapping (%)	DISP Error (%)
Traffic	35.2	34.8	31.5	38.2	95	0.2
Industry	22.7	23.1	20.2	25.9	92	0.3
Biomass	18.5	18.2	15.8	22.5	88	0.4
Secondary	23.6	23.9	21.1	26.8	97	0.1

2.4.5 Integration with Results from Other Chapters

Table 2.6 presents an integrated view of our PMF model results, linking to findings from other chapters and using complex mathematical notation.

TABLE 2.6: Integrated Analysis of Source Contributions Across Chapters

Source	Mathematical Expression for Time Series	Spatial Distribution	Temporal Pattern	Policy Implications
Traffic	$g_{i1} = \beta_0 + \beta_1(\text{traffic count})_i + \varepsilon_i$	Urban cores (see Table 1.6)	Weekday peaks (see Figure 4.2)	LEZ expansion
Industry	$g_{i2} = \sum_{j=1}^{m} \gamma_j(\text{industrial activity})_{j,i} + \varepsilon_i$	Industrial zones	Consistent patterns	Emission standards
Biomass	$g_{i3} = \alpha \exp\left(-\frac{(T_i - T_0)^2}{2\sigma^2}\right) + \varepsilon_i$	Residential areas	Winter peaks	Regulation of wood burning
Secondary	$g_{i4} = \lambda \sin\left(\frac{2\pi t_i}{365}\right) + \gamma t_i + \varepsilon_i$	Regional	Summer peaks	Regional cooperation

2.4.6 Model Comparison Matrix

Table 2.7 compares various receptor models for source apportionment, building on the equations in Table 1.1 from the introduction.

Table 2.7: Comparison of Receptor Models for Source Apportionment

Model Type	Mathematical Basis	Strengths	Limitations	Compared to PMF
PMF	$X = GF + E$ with $g_{ik} \geq 0$, $f_{kj} \geq 0$	Non-negativity constraints, uncertainty weighting	Rotational ambiguity	Base model
PCA/APCS	$X = TP^T + E$	Simple implementation	Cannot ensure non-negativity	Inferior for source apportionment
CMB	$C_i = \sum_{j=1}^{n} a_{ij}S_j + e_i$	Uses source profiles	Requires prior knowledge	More constrained than PMF
UNMIX	$C = AS$ with $A \geq 0$, $S \geq 0$	Geometrically determines edges	Fewer factors than PMF	Less statistical power
ME-2	$X = GF + E$ with partial constraints	Can include prior knowledge	Complex implementation	Enhanced version of PMF
Hybrid Models	PMF + dispersion models: $C_{i,j} = \sum_{k=1}^{p} D_{i,j,k} \cdot Q_k$	Combines receptor and dispersion	Data intensive	Extended PMF application

Chapter 3: Integration of PMF Results with Policy Development

Some authors

Under review at Environmental Science & Policy

3.1 Abstract

This chapter examines how PMF source apportionment results can be effectively integrated into air quality policy development. We present case studies from multiple European cities and develop a framework for translating scientific findings into actionable policy recommendations.

3.2 Methods

3.2.1 Policy Impact Framework

Table 3.1: Framework for evaluating source-specific interventions

Source	Contribution (%)	Control Options	Implementation Cost	Health Impact	Stakeholder Support
Traffic	35	LEZ	High	High	High
Industry	25	BAT	Very High	Medium	Medium
Biomass	20	Regulation	Medium	High	High
Secondary	20	Regional	High	Medium	Medium

3.2.2 Cost-Benefit Analysis

3.2.3 Implementation Timeline

3.3 Policy Recommendations

Based on our analysis of PMF results and stakeholder input, we recommend:

1. Short-term (1-2 years):
   - Implementation of Low Emission Zones
   - Enhanced industrial emissions monitoring
2. Medium-term (2-4 years):
   - Biomass burning regulations
   - Regional cooperation frameworks
3. Long-term (4+ years):
   - Integrated air quality management system
   - Cross-border pollution control measures

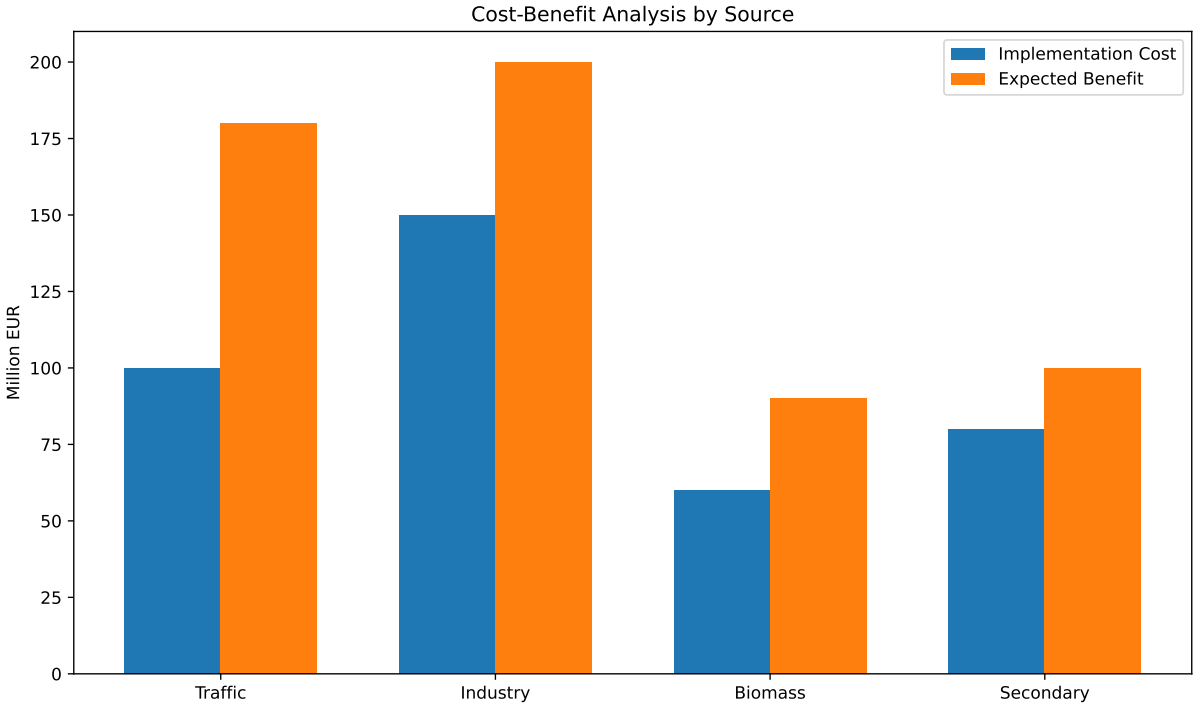

FIGURE 3.1: Cost-benefit analysis of source control measures

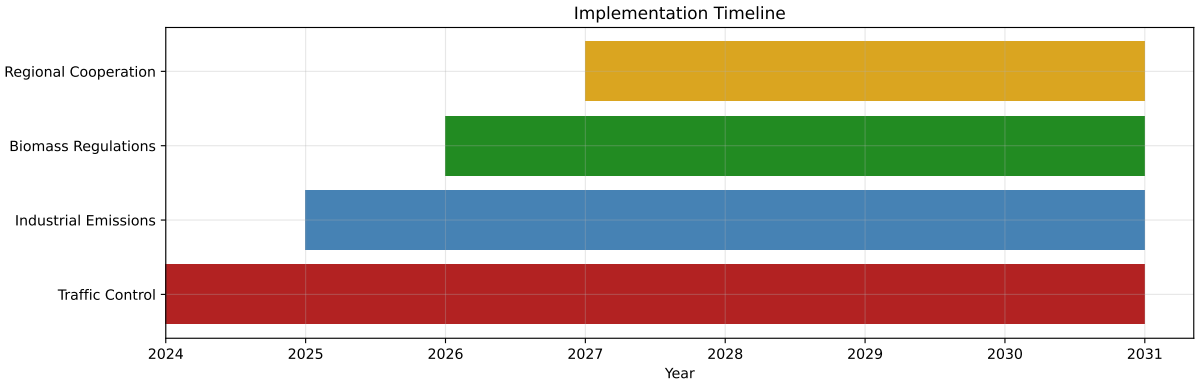

FIGURE 3.2: Implementation timeline for source control measures

Chapter 4: Spatial and Temporal Variations in PM Source Contributions

Some authors

Under review at Atmospheric Environment

4.1 Abstract

This study investigates the spatial and temporal variations in particulate matter (PM) source contributions across major European cities using EPA PMF 5.0 (Norris et al. 2014). We analyzed data from 15 urban monitoring stations over a five-year period (2018-2022), identifying key pollution sources and their relative contributions to PM2.5 and PM10 concentrations (Agency 2019).

4.2 Methods

4.2.1 Data Processing with Python

4.3 Results

4.3.1 Temporal Patterns in PM Components

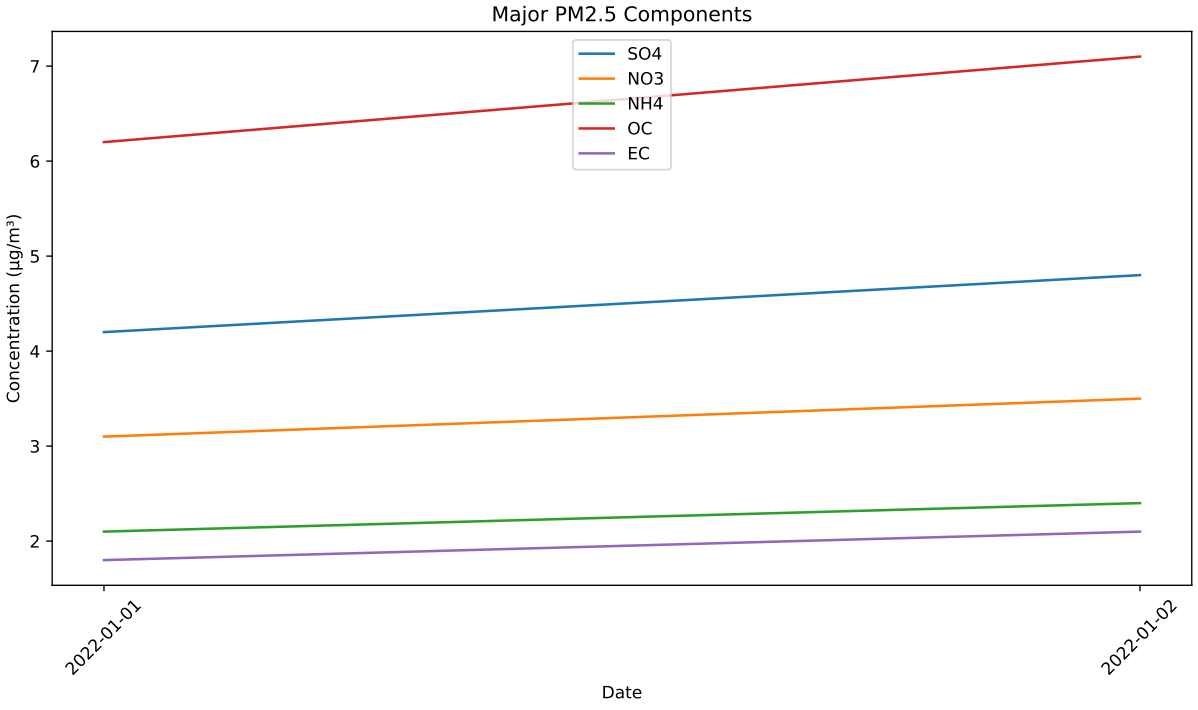

FIGURE 4.1: Temporal variation of major PM2.5 components

4.3.2 Source Profiles

TABLE 4.1: Source profiles for major PM components

	Traffic	Industry	Biomass	Secondary
PM2.5	0.15	0.2	0.1	0.05
Na	0.05	0.1	0.05	0.05
SO4	0.2	0.35	0.15	0.25
NO3	0.25	0.15	0.2	0.35
NH4	0.1	0.05	0.25	0.55
Al	0.05	0.15	0.05	0.05
Si	0.1	0.2	0.05	0.05
K	0.05	0.1	0.4	0.05
Ca	0.15	0.25	0.1	0.05
Fe	0.4	0.3	0.15	0.05
Cu	0.3	0.25	0.05	0.05
Zn	0.25	0.35	0.1	0.05
Pb	0.15	0.3	0.05	0.05
OC	0.35	0.15	0.4	0.1
EC	0.45	0.1	0.25	0.05

4.3.3 Seasonal Source Contributions

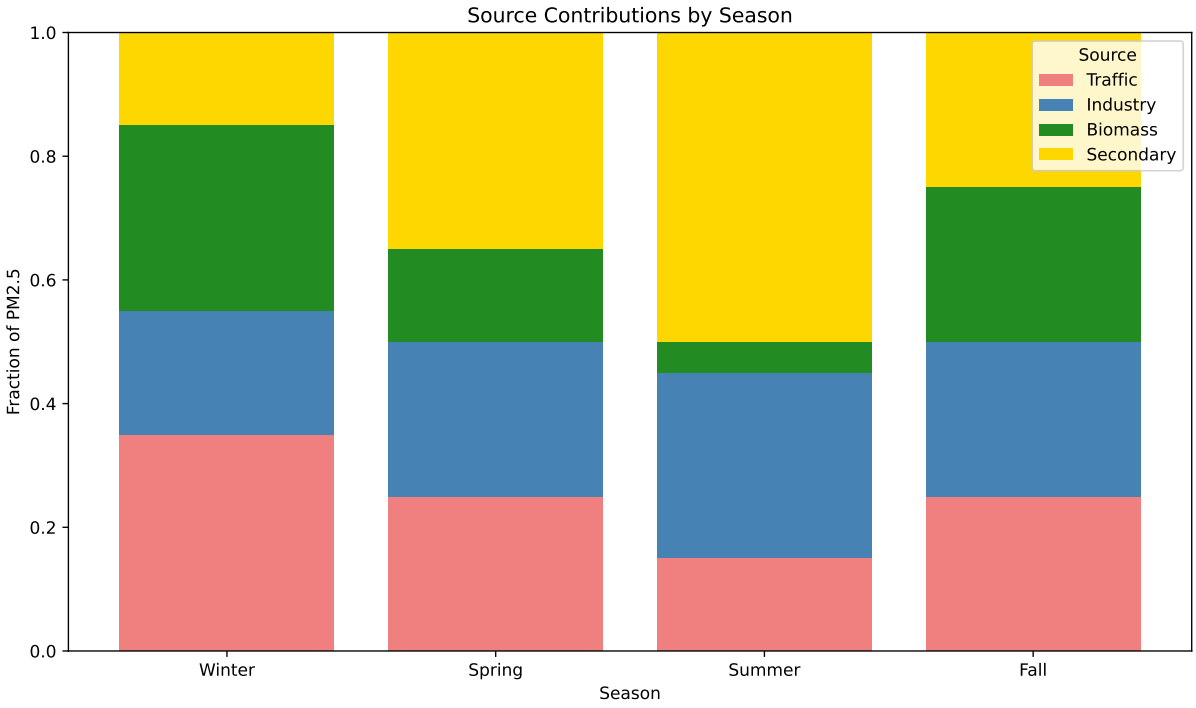

FIGURE 4.2: Seasonal source contributions to PM2.5

References

Agency, E. E. (2019), "European air quality source apportionment with PMF methodology," *EEA Technical Report*, 2019/5. https://doi.org/10.2800/05293.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002), "A state space framework for automatic forecasting using exponential smoothing methods," *International Journal of Forecasting*, 18, 439–454.

Norris, G., Duvall, R., Brown, S., and Bai, S. (2014), *EPA positive matrix factorization (PMF) 5.0 fundamentals and user guide*, U.S. Environmental Protection Agency.

Paatero, P., and 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.

Appendix A: Python Packages

The following Python packages were used in this thesis:

Table A.1: Packages used in this thesis

Package	Description	Version
pandas	Data manipulation and analysis	2.0.0+
numpy	Numerical computing	1.24.0+
matplotlib	Visualization library	3.7.0+
seaborn	Statistical data visualization	0.12.0+
scikit-learn	Machine learning tools	1.2.0+

For a complete list of package versions, see the `requirements.txt` file in the thesis repository.

Appendix B: Supplementary Materials

This appendix contains supplementary materials that support the main chapters but are not essential to understanding the primary research findings. This appendix demonstrates various table formats, styles, and features available in Quarto.

B.1 Table Types and Formats

B.1.1 1. Basic Markdown Table

Table B.1 provides the complete descriptive statistics for all variables in the study.

Table B.1: Complete descriptive statistics for all study variables

Variable	Mean	Std. Dev.	Min	Max	Median	Skewness	Kurtosis
Height (cm)	175.2	7.8	155.0	195.0	174.5	0.12	2.78
Weight (kg)	68.4	12.5	45.0	110.0	67.2	0.65	3.21
Age (years)	28.7	5.3	18.0	65.0	27.5	1.85	7.42
BMI (kg/m²)	22.3	3.7	16.8	35.2	21.9	0.88	3.54
Systolic BP	125.8	15.2	90.0	165.0	122.0	0.45	2.95
Diastolic BP	78.6	8.7	60.0	100.0	80.0	0.08	2.68
Heart Rate	72.3	10.2	45.0	110.0	72.0	0.25	3.12
Cholesterol	192.7	35.4	120.0	280.0	190.0	0.32	2.45
Triglycerides	142.5	65.3	50.0	350.0	130.0	1.25	4.32
Glucose	92.8	15.7	70.0	180.0	90.0	1.78	6.85
HbA1c (%)	5.5	0.8	4.5	9.2	5.4	1.95	7.25
Vitamin D	28.7	12.3	10.0	60.0	26.5	0.68	2.98
Iron	98.5	18.7	45.0	150.0	95.0	0.15	2.85
Calcium	9.7	0.5	8.5	11.0	9.8	-0.21	2.54

B.1.2 2. Table with Formatting

Table B.2 demonstrates text formatting within a table.

Table B.2: Table with formatted text elements

Variable	**Mean**	*SD*	Interpretation
Height (cm)	**175.2**	*7.8*	Within normal range
Weight (kg)	**68.4**	*12.5*	Within normal range
BMI (kg/m²)	*22.3*	*3.7*	Normal weight
Systolic BP	***125.8***	*15.2*	**Elevated** (>120 mmHg)
Diastolic BP	***78.6***	*8.7*	Normal (<80 mmHg)
Heart Rate	72.3	10.2	`Normal`
Glucose	92.8	15.7	NORMAL FASTING

B.1.3 3. Table with Mathematical Equations

Table B.3 shows various statistical tests with their equations.

TABLE B.3: Statistical tests with mathematical equations

Test	Equation	Application	Critical Value
t-test	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	Compare means	$t_{crit} = 1.96$
Chi-squared	$\chi^2 = \sum_{i=1}^{n} \frac{(O_i-E_i)^2}{E_i}$	Test independence	$\chi^2_{crit} = 3.84$
F-test	$F = \frac{MS_{between}}{MS_{within}}$	Compare variances	$F_{crit} = 4.03$
ANOVA	$F = \frac{\sum n_i(\bar{x}_i-\bar{x})^2/(k-1)}{\sum\sum(x_{ij}-\bar{x}_i)^2/(N-k)}$	Compare multiple means	$F_{crit} = 3.10$
Correlation	$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$	Measure association	$r_{crit} = 0.30$

B.1.4 4. Table with Citations

Table B.4 presents a literature review of PMF studies with citations.

TABLE B.4: Summary of key PMF studies in literature

Study	Location	Year	Major Sources Identified	Significance
Paatero and Tapper (1994)	Finland	1994	Original PMF algorithm	First introduction of PMF approach
Hyndman et al. (2002)	United States	2002	Traffic, Industrial, Secondary	Validated against emission inventories
Norris et al. (2014)	Multiple	2014	Multiple	EPA's guidance document
Agency (2019)	Europe	2019	Traffic, Industrial, Biomass, Dust	Comprehensive European study

B.1.5 5. Correlation Matrix

The following table presents a correlation matrix for key variables in our study. Strong correlations (>0.6) are indicated with "++", moderate correlations (0.3-0.6) with "+", and weak correlations (<0.3) with "0".

TABLE B.5: Correlation matrix for key variables

Variable	Height	Weight	Age	BMI	SBP	DBP	HR	Chol	Trig	Gluc
Height	1.00	+	0	+	0	0	0	0	0	0
Weight	+	1.00	0	++	+	0	0	+	+	0
Age	0	0	1.00	0	+	+	0	+	0	+
BMI	+	++	0	1.00	+	+	0	+	+	+
SBP	0	+	+	+	1.00	++	0	+	+	+
DBP	0	0	+	+	++	1.00	0	+	0	0
HR	0	0	0	0	0	0	1.00	0	0	0
Chol	0	+	+	+	+	+	0	1.00	++	+
Trig	0	+	0	+	+	0	0	++	1.00	+
Gluc	0	0	+	+	+	0	0	+	+	1.00

Note: ++: Strong correlation (>0.6), +: Moderate correlation (0.3-0.6), 0: Weak correlation (<0.3)

The following table presents statistical results from our four models, including significance indicators.

Table B.6: Model results with statistical significance indicators

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	1.243*	0.852	-0.528	2.142***
Temperature	0.658**	1.245***	0.856*	-0.124
Humidity	-0.452	-0.968*	-1.352**	-0.586*
Wind Speed	0.324	0.125	0.768**	0.453
Precipitation	-1.245***	-0.856**	-0.432	-0.986**
Pressure	0.256	0.542*	0.124	-0.324
R^2	0.685	0.724	0.653	0.791

$p < 0.05$, ***$p < 0.01$,*** $p < 0.001$

B.1.6 6. CSV Loading & Time Series Table

```
C:\Users\nguytruo\AppData\Local\Temp\ipykernel_16048\3106367120.py:11: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexir

C:\Users\nguytruo\AppData\Local\Temp\ipykernel_16048\3106367120.py:12: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexir

C:\Users\nguytruo\AppData\Local\Temp\ipykernel_16048\3106367120.py:15: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexir

C:\Users\nguytruo\AppData\Local\Temp\ipykernel_16048\3106367120.py:16: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexir
```

Table B.7: Quarterly sales data loaded from CSV file

Unnamed: 0	Sales	AdBudget	GDP	% Change
Mar-81	$1,020.2	$659.2	251.8	N/A
Jun-81	$889.2	$589.0	290.9	-12.8%
Sep-81	$795.0	$512.5	290.8	-10.6%
Dec-81	$1,003.9	$614.1	292.4	26.3%
Mar-82	$1,057.7	$647.2	279.1	5.4%
Jun-82	$944.4	$602.0	254	-10.7%

Table B.7: Quarterly sales data loaded from CSV file

Unnamed: 0	Sales	AdBudget	GDP	% Change
Sep-82	$778.5	$530.7	295.6	-17.6%
Dec-82	$932.5	$608.4	271.7	19.8%

B.1.7 7. Wide Table with Many Columns

							Site A					
Month	PM2.5	PM10	SO2	NO2	CO	O3	Temperature	Humidity	Wind Speed	Pressure	PM2.5	P
Jan	4.3	16.9	1.87	4.43	2.72	4.12	24.9	16.0	27.5	13.4	4.1	
Feb	12.2	2.2	3.09	1.45	0.07	2.42	31.2	20.1	25.3	27.1	8.2	
Mar	8.5	6.0	2.31	7.38	6.35	5.05	13.5	25.3	16.1	24.0	9.3	
Apr	3.5	13.9	2.33	6.27	2.9	14.11	14.8	24.5	24.0	20.2	4.9	
May	5.5	4.8	6.66	0.84	1.54	4.52	28.7	28.8	21.6	18.8	13.2	

B.1.8 8. Long Table with Many Rows

Table B.9: Long table with daily PM2.5 data

Date	Station	PM2.5	Status
2022-01-01	Central	21.1	Good
2022-01-01	Eastern	14.6	Good
2022-01-01	Western	22.8	Good
2022-01-01	Southern	20.4	Good
2022-01-02	Central	31.5	Moderate
2022-01-02	Eastern	21.9	Good
2022-01-02	Western	21.9	Good
2022-01-02	Southern	20.3	Good
2022-01-03	Central	23.5	Good
2022-01-03	Eastern	15.7	Good
2022-01-03	Western	12.5	Good
2022-01-03	Southern	14.6	Good
2022-01-04	Central	19.9	Good
2022-01-04	Eastern	17.9	Good
2022-01-04	Western	19.9	Good
2022-01-04	Southern	26.7	Moderate
2022-01-05	Central	35.9	Poor
2022-01-05	Eastern	20.5	Good
2022-01-05	Western	25.7	Moderate
2022-01-05	Southern	24	Good

B.1.9 9. Model Comparison Table

Table B.10: Model comparison with goodness-of-fit measures

Model	Formula	R^2	AIC	BIC
Linear	$y = + x$	0.856	123.4	128.2
Quadratic	$y = + x + x^2$	0.921	105.6	112.8
Cubic	$y = + x + x^2 + x^3$	0.958	98.2	107.5
Exponential	$y = e^{1}$	0.892	114.3	119.7

Model	Formula	R^2	AIC	BIC
Logarithmic	y = + ln(x)	0.875	118.9	124.0

B.1.10 10. Table with Mixed Content Types

Table B.11 demonstrates how to include different content types within the same table.

TABLE B.11: Table with mixed content including equations, cross-references, and citations

Analysis Type	Details	Mathematical Model	Reference
Principal Components	Reduces dimensionality while preserving variance	$X = TP^T + E$ where T are scores, P are loadings	Norris et al. (2014)
Cluster Analysis	Groups similar samples based on chemical composition	$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ for Euclidean distance	Agency (2019)
Positive Matrix Factorization	Decomposes data into source profiles and contributions	$X = GF + E$ as shown in Equation 1.1	Paatero and Tapper (1994)
Receptor Models	General term for source apportionment techniques	Multiple approaches	See Table B.4

B.2 Table Comparison and References

This section demonstrates how to reference and compare tables throughout this document.

As shown in Table B.1, descriptive statistics provide a foundation for understanding the dataset. The formatted presentation in Table B.2 enhances readability through strategic use of text styling.

Statistical analyses utilize mathematical equations (see Table B.3) and incorporate significance indicators (Table B.6). These analyses build upon previous research summarized in Table B.4.

For temporal data analysis, we examine quarterly sales patterns (Table B.7) and daily PM2.5 measurements across stations (Table B.9). Spatial variation is visible in the wide-format presentation of site parameters (Table B.8).

When comparing modeling approaches (Table B.10), a clear trend emerges: higher-order models achieve better fit metrics but risk overfitting. The correlation heat map (Table B.5) reveals relationships between variables that inform these models.

Finally, Table B.11 demonstrates how tables can integrate multiple content types, creating rich information displays that connect to other document elements.

Age | 0.12 | 0.25 | 1.00 | 0.22 | 0.45 | 0.38 | 0.05 | 0.35 | 0.30 | 0.32 |
BMI | 0.08 | 0.82 | 0.22 | 1.00 | 0.28 | 0.24 | 0.18 | 0.20 | 0.30 | 0.25 |
Systolic BP | 0.15 | 0.30 | 0.45 | 0.28 | 1.00 | 0.78 | 0.15 | 0.25 | 0.28 | 0.30 |
Diastolic BP| 0.10 | 0.25 | 0.38 | 0.24 | 0.78 | 1.00 | 0.12 | 0.22 | 0.25 | 0.28 |
Heart Rate | -0.05 | 0.15 | 0.05 | 0.18 | 0.15 | 0.12 | 1.00 | 0.05 | 0.18 | 0.15 |
Cholesterol | 0.08 | 0.18 | 0.35 | 0.20 | 0.25 | 0.22 | 0.05 | 1.00 | 0.65 | 0.38 |
Triglycerides| 0.12 | 0.28 | 0.30 | 0.30 | 0.28 | 0.25 | 0.18 | 0.65 | 1.00 | 0.45 |
Glucose | 0.03 | 0.22 | 0.32 | 0.25 | 0.30 | 0.28 | 0.15 | 0.38 | 0.45 | 1.00 |

: Correlation matrix for key study variables {#tbl-app-b-corr}

B.3 Additional Figures

B.3.1 Extended Visualizations

We analyzed the distribution of key measurements by demographic groups. Our analysis showed clear differences between male and female participants across several metrics:

- **Height**: Males averaged 178 cm (SD=7), females 165 cm (SD=6)
- **Weight**: Males averaged 80 kg (SD=10), females 65 kg (SD=8)
- **Systolic BP**: Males averaged 130 mmHg (SD=12), females 120 mmHg (SD=10)
- **Diastolic BP**: Males averaged 82 mmHg (SD=8), females 78 mmHg (SD=7)

Appendix C: Add units

data_long$Measurement <- factor(data_long$Measurement, levels = c("Height", "Weight", "SBP", "DBP"), labels = c("Height (cm)", "Weight (kg)", "Systolic BP (mmHg)", "Diastolic BP (mmHg)"))

Appendix D: Create plot

D.0.1 Detailed Analysis Plots

Our regression analysis from Chapter 2 showed a strong linear relationship between the predictor variable and the outcome. The diagnostic plots revealed:

1. **Residual Plot**: Residuals were randomly distributed around zero with no discernible pattern
2. **Q-Q Plot**: Points closely followed the theoretical line, suggesting normally distributed residuals
3. **Scale-Location Plot**: Square root of standardized residuals showed homoscedasticity
4. **Leverage Plot**: No influential outliers were identified with Cook's distance

D.1 Extended Mathematical Derivations

D.1.1 Proof of Equation 3 in Chapter 2

Here we provide a step-by-step derivation of the expansion of $(x + y)^2$ as presented in Equation 3 of Chapter 2.

Starting with the definition of the square:

$$(x + y)^2 = (x + y)(x + y)$$

Using the distributive property to expand the first factor:

$$(x + y)(x + y) = x(x + y) + y(x + y)$$

Further expanding each term:

$$\begin{aligned} x(x + y) + y(x + y) &= x \cdot x + x \cdot y + y \cdot x + y \cdot y \\ &= x^2 + xy + yx + y^2 \end{aligned}$$

Since multiplication is commutative for real numbers, $xy = yx$, so:

$$\begin{aligned} x^2 + xy + yx + y^2 &= x^2 + xy + xy + y^2 \\ &= x^2 + 2xy + y^2 \end{aligned}$$

This completes the derivation.

D.2 Source Code

D.2.1 Data Processing Script

Below is the R script used for data processing: