Health Effects Institute

NUMBER 228
AUGUST 2025

RESEARCH REPORT

Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function

Lianne Sheppard, Magali N. Blanco, Annie Doubleday, Adam A. Szpiro, Jianzhao Bi, Christopher Zuidema, Si Cheng, Ali Shojaie, Sun-Young Kim, and Amanda Gassett

INCLUDES A COMMENTARY BY THE INSTITUTE'S IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL

www.healtheffects.org

Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function

Lianne Sheppard, Magali N. Blanco, Annie Doubleday, Adam A. Szpiro, Jianzhao Bi, Christopher Zuidema, Si Cheng, Ali Shojaie, Sun-Young Kim, and Amanda Gassett

with a Commentary by the HEI Improved Exposure Assessment Studies Review Panel

Research Report 228
Health Effects Institute
Boston, Massachusetts

Trusted Science · Cleaner Air · Better Health

Publishing history: This report was posted at *www.healtheffects.org* in August 2025.

Citation for report:

Sheppard L, Blanco MN, Doubleday A, Szpiro AA, Bi J, Zuidema C, et al. 2025. Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function. Research Report 228. Boston, MA: Health Effects Institute.

 Compositor: David Wade, Virginia Beach, VA. Library of Congress Catalog Number for the HEI Report Series: WA 754 R432.

ISSN: 2688-6855 (online)

CONTENTS

ABOUT HEI

The Health Effects Institute is a nonprofit corporation chartered in 1980 as an independent research organization to provide high-quality, impartial, and relevant science on the effects of air pollution on health. To accomplish its mission, the Institute

- identifies the highest-priority areas for health effects research
- competitively funds and oversees research projects
- provides intensive independent review of HEI-supported studies and related research
- integrates HEI's research results with those of other institutions into broader evaluations
- communicates the results of HEI's research and analyses to public and private decision-makers.

HEI typically receives balanced funding from the US Environmental Protection Agency and the worldwide motor vehicle industry. Frequently, other public and private organizations in the United States and around the world also support major projects or research programs. HEI has funded more than 390 research projects in North America, Europe, Asia, and Latin America, the results of which have informed decisions regarding carbon monoxide, air toxics, nitrogen oxides, diesel exhaust, ozone, particulate matter, and other pollutants. These results have appeared in more than 275 comprehensive reports published by HEI, as well as in more than 2,500 articles in peer-reviewed literature.

HEI's independent Board of Directors consists of leaders in science and policy who are committed to fostering the public–private partnership that is central to the organization. The Research Committee solicits input from HEI sponsors and other stakeholders and works with scientific staff to develop a Five-Year Strategic Plan, select research projects for funding, and oversee their conduct. The Review Committee or Panel, which has no role in selecting or overseeing studies, works with staff to evaluate and interpret the results of funded studies and related research.

All project results and accompanying comments by the Review Committee or Panel are widely disseminated through HEI's website (*www.healtheffects.org*), reports, newsletters, annual conferences, and presentations to legislative bodies and public agencies.

ABOUT THIS REPORT

Research Report 228, *Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function*, presents a research project funded by the Health Effects Institute and conducted by Dr. Lianne Sheppard at the University of Washington in Seattle and her colleagues. The report contains three main sections:

The HEI Statement, prepared by staff at HEI, is a brief, nontechnical summary of the study and its findings; it also briefly describes the Review Panel's comments on the study.

The Investigators' Report, prepared by Sheppard and colleagues, describes the scientific background, aims, methods, results, and conclusions of the study.

The Commentary, prepared by members of the Review Panel with the assistance of HEI staff, places the study in a broader scientific context, points out its strengths and limitations, and discusses the remaining uncertainties and implications of the study's findings for public health and future research.

This report has gone through HEI's rigorous review process. When an HEI-funded study is completed, the investigators submit a draft final report presenting the background and results of the study. Outside technical reviewers first examine this draft report. The report and the reviewers' comments are then evaluated by members of the Review Panel, an independent panel of distinguished scientists who are not involved in selecting or overseeing HEI studies. During the review process, the investigators have an opportunity to exchange comments with the Review Panel and, as necessary, to revise their report. The Commentary reflects the information provided in the final version of the report.

Although this report was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not reflect the views of the Agency; thus, no official endorsement by it should be inferred. The contents of this report also have not been reviewed by private party institutions, including those that support the Health Effects Institute and may not reflect the views or policies of these parties; thus, no endorsement by them should be inferred.

CONTRIBUTORS

RESEARCH COMMITTEE

David A. Savitz, Chair *Professor of Epidemiology, School of Public Health, and Professor of Obstetrics and Gynecology and Pediatrics, Alpert Medical School, Brown University, USA*

Benjamin Barratt *Professor, Environmental Research Group, School of Public Health, Imperial College London, United Kingdom*

David C. Dorman *Professor, Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, USA*

Christina H. Fuller *Associate Professor, School of Environmental, Civil, Agricultural and Mechanical Engineering, University of Georgia College of Engineering, USA*

Marianne Hatzopoulou *Professor, Civil and Mineral Engineering, University of Toronto, Research Chair in Transport Decarbonization and Air Quality, Canada*

Heather A. Holmes *Associate Professor, Department of Chemical Engineering, University of Utah, USA*

Marianthi-Anna Kioumourtzoglou* *Associate Professor of Environmental Health Sciences, Columbia Mailman School of Public Health, New York, USA*

Neil Pearce *Professor of Epidemiology and Biostatistics, London School of Hygiene and Tropical Medicine, United Kingdom*

Evangelia (Evi) Samoli *Professor of Epidemiology and Medical Statistics, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece*

Alexandra M. Schmidt *Professor of Biostatistics, School of Population and Global Health, McGill University, Canada*

Neeta Thakur *Associate Professor of Medicine, University of California San Francisco, USA*

Gregory Wellenius *Professor, Department of Environmental Health, Boston University School of Public Health and Director, BUSPH Center for Climate and Health, USA*

IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL

Jana Milford, Chair *Professor Emerita, Department of Mechanical Engineering and Environmental Engineering Program, University of Colorado Boulder, Colorado, USA*

Susanne Breitner-Busch *Senior Scientist, IBE Chair of Epidemiology, LMU Munich and Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany*

Anna Oudin *Associate Professor, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden, and Division of Occupational and Environmental Medicine, Lund University, Lund, Sweden*

John Volckens *Professor, Department of Mechanical Engineering, Walter Scott Jr. College of Engineering, Colorado State University, Fort Collins, Colorado, USA*

Shu Yang *Associate Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA*

HEI PROJECT STAFF

Allison P. Patton *Senior Scientist (Study Oversight)*

Hanna Boogaard *Consulting Principal Scientist (Report Review)*

Kristin C. Eckles *Senior Editorial Manager*

Tom Zaczkiewicz *Consulting Editor*

*Consultant.

PREFACE

HEI's Research to Assess Health Effects of Traffic-Related Air Pollution and to Improve Exposure Assessment for Health Studies

INTRODUCTION

Traffic emissions are an important source of urban air pollution and have been linked to various adverse health outcomes (Atkinson et al 2018; Health Canada 2016; HEI 2010; HEI 2022a; Huangfu and Atkinson 2020; US Environmental Protection Agency [US EPA] 2016). Over the last several decades, air quality regulations and improvements in vehicular emission control technologies have steadily decreased emissions from motor vehicles. As a result, ambient concentrations of several major traffic-related air pollutants have decreased in most high-income countries, even as vehicle miles traveled and economic activity have increased and older or malfunctioning vehicles have remained on the roads (HEI 2022a; US EPA 2023).

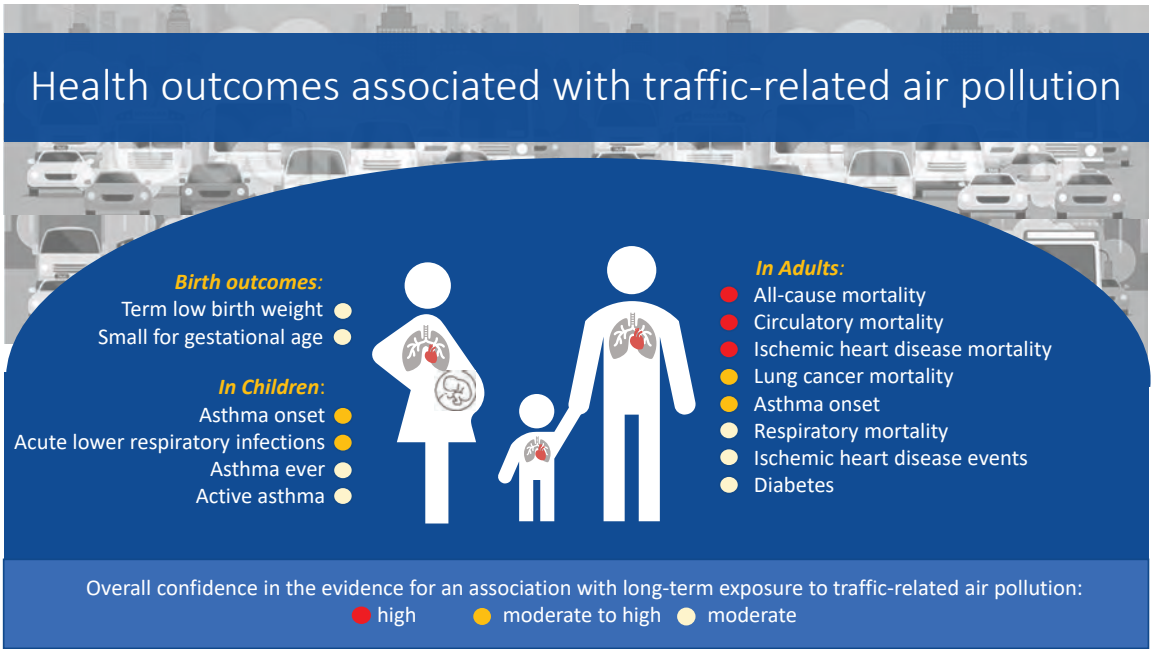
Following HEI's widely cited 2010 Report (HEI 2010), HEI published *Special Report 23*, a systematic review of more than 350 epidemiological studies on the health effects of long-term exposure to emissions of primary traffic-related air pollutants (HEI 2022a). The 2022 report found a high level of confidence that strong connections exist between traffic-related air pollution and early death due to cardiovascular diseases. A strong connection was also found between traffic-related air pollution and lung cancer mortality, asthma onset in children and adults, and acute lower respiratory infections in children (**Preface Figure**). The confidence in the evidence was considered moderate, low, or very low for the other selected outcomes, such as coronary events, diabetes, and adverse birth outcomes.

Although traffic-related emissions have decreased over the past decades, further research is warranted in several areas. Emerging evidence suggests that transportation can affect health through many intertwined pathways such as collisions, noise, climate change, temperature, stress, and the lack of physical activity and green space (Glazener et al. 2021). As tailpipe emissions from internal combustion engines decrease and electric vehicles increase market share, more studies are needed to quantify human exposures to nontailpipe particulate matter better and to assess the health effects associated with those exposures. Relatively few studies evaluate how influential factors such as green space, heat exposure, noise pollution, and physical activity interact with or modify air pollution health effects. Evaluation of those factors and exposures are critical because they reflect real-world conditions and might further advance our understanding of the implications of transportation activities on traffic-related air pollution and health (Khreis et al. 2020).

Moreover, better understanding is needed of the role of specific pollutants including nitrogen dioxide (NO_2) and ultrafine particles (UFPs), the health effects of short-term exposures versus long-term exposures, the effects on a broader range of health outcomes (such as neurological and birth outcomes) that have not been extensively examined, and the ways in which marginalized communities are affected. However, a challenge for exposure assessment of traffic-related air pollution is that traffic emits a complex mixture of pollutants in particulate and gaseous forms, many of which are also emitted by other sources. In addition, traffic-related air pollution is characterized by high spatial and temporal variability, with the highest concentrations occurring at or near major roads. Therefore, it has been difficult to identify an appropriate exposure metric that uniquely indicates traffic-related air pollution and to model the distribution of exposure at a sufficiently high degree of spatial and temporal resolution.

Various air quality models — such as dispersion, land use regression, and hybrid models — have been developed to estimate long-term exposure to air pollution (HEI 2022a; Hoek 2017; Jerrett et al. 2005). Recent developments in measurement technologies and approaches to modeling long-term exposure to air pollution have increasingly been used to provide air pollution estimates at fine spatial scales for epidemiological studies of large populations. Advances include novel air pollution sensors, mobile monitoring, satellite data, hybrid models, and machine-learning approaches (Hoek 2017).

Moreover, many improvements in exposure models have occurred over time with the advance of geographic information system approaches and the application of more sophisticated statistical methods; see, for example, several studies previously funded by HEI: Apte 2024, Barratt 2018,



Preface Figure. Overall confidence in the evidence for an association between long-term exposure to traffic-related air pollution and selected health outcomes. Health outcomes for which the overall confidence in the evidence was low to moderate, low, or very low are not in the figure. Reproduced from HEI 2022a.

Batterman 2020, Frey 2022, and Sarnat 2018. However, the usefulness of exposure estimates still depends on the model assumptions and input data quality, and there remain limitations and challenges when predicting air pollution exposure, particularly for such pollutants as UFPs, NO_2, and black carbon (BC) that vary highly in space and time. Few studies have compared the performance of different models and evaluated exposure measurement error and possible bias in health estimations.

Thus, HEI issued complementary requests for applications in 2017 (RFA 17-1) and 2019 (RFA 19-1) to evaluate traffic-related health effects in the context of spatially correlated factors — specifically traffic noise, socioeconomic status, and green space — and to improve exposure assessment for health studies.

OBJECTIVES OF THE RFAs

OBJECTIVES OF RFA 17-1

RFA 17-1, Assessing Adverse Health Effects of Exposure to Traffic-Related Air Pollution, Noise, and Their Interactions with Socioeconomic Status, solicited studies that sought to assess adverse health effects from exposure to traffic-related air pollution and to disentangle the effects from spatially correlated confounding or modifying factors — most notably, traffic noise, socioeconomic status, and the built environment, including green space. The RFA had five major objectives:

1. In the proposed health studies, develop, validate, and apply improved exposure assessment methods and models suitable for estimating exposure to traffic-related air pollution that take into account other air pollution sources in urban areas (such as airports, [sea]ports, industries, and other local point sources) and that would be able to distinguish between tailpipe and nontailpipe traffic emissions.
2. Propose ways in these studies to disentangle the relationship of adverse health effects of traffic-related air pollution and traffic noise.
3. Develop, evaluate, and apply indicators of socioeconomic status at the individual and community level in the proposed health studies; if such indicators are novel, compare with socioeconomic status indicators commonly used in the literature.
4. Explore the role of other factors that might confound or modify the health effects of traffic-related air pollution at the individual (e.g., age, smoking status, diet, physical activity, and health status) and community level (e.g., presence of green space, other factors related to the built environment, and walkability).
5. Investigate — to the extent that the measurements and patterns of a range of different indicators of traffic-related air pollution allow it (e.g., NO_2, UFPs, BC, and indicators of nontailpipe emissions) — whether one or more of them can be shown to have health effects independent of the other pollutants.

OBJECTIVES OF RFA 19-1

RFA 19-1, Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies, solicited studies to assess exposures to air pollution using new

and conventional exposure assessment approaches, to evaluate quantitatively exposure measurement error to determine the added value of the novel approaches, and to apply the exposure estimates in epidemiological analyses to evaluate the potential effect of exposure measurement error on chronic health estimates. The RFA had four major objectives:

1. Conduct a new monitoring campaign designed to determine long-term exposure to outdoor air pollutants with high spatial and temporal variability by using sensors, mobile monitoring, location tracking, or other approaches.
2. Develop several exposure assessment approaches suitable to estimate long-term exposure to air pollution at relevant spatial and temporal scales for use in an ongoing or future health study.
3. Quantify exposure measurement error by evaluating and comparing the performance of models of long-term air pollution exposure developed under this RFA to the performance of previous models.
4. Apply the various exposure estimates in an ongoing health study to evaluate the potential impact of exposure measurement error in health estimates or explain how the exposure assessments would be directly applicable to future health studies.

DESCRIPTION OF THE RESEARCH PROGRAM

Three 4-year studies were funded under RFA 17-1 and five 3-year studies were funded under RFA 19-1 to cover the various RFA objectives; they are summarized below (**Preface Table**). The study by Sheppard and colleagues described in this report (Research Report 228) is the fifth to be published.

STUDIES FUNDED UNDER RFA 17-1

HEI funded two studies in Europe and one study in the United States to evaluate various aspects of the association between long-term traffic-related air pollution and health by using existing cohorts (Denmark, USA) and a newly recruited cohort (Spain). Two studies focused on health outcomes during pregnancy (Dadvand) and childhood (Franklin), and one study focused on cardiometabolic outcomes in adults (Raaschou-Nielsen).

"Traffic-Related Air Pollution and Birth Weight: The Roles of Noise, Placental Function, Green Space, Physical Activity, and Socioeconomic Status (FRONTIER)," Payam Dadvand and Jordi Sunyer, Barcelona Institute for Global Health (ISGlobal), Spain Dadvand, Sunyer, and colleagues established a new cohort, named Barcelona Life Study Cohort (BiSC) of 1,080 healthy pregnant women in Barcelona, Spain, in 2018. They estimated exposure to various traffic-related pollutants by using hybrid models that included dispersion models, land use data, time-activity data, and personal and home-outdoor air pollution monitoring data. They linked the exposure to various birth outcomes including birth weight, small for gestational age, and fetal growth trajectories. They evaluated the role of traffic noise and green space and also took into account socioeconomic status and maternal stress (in press).

"Intersections as Hot Spots: Assessing the Contribution of Localized Non-Tailpipe Emissions and Noise on the Association between Traffic and Children's Respiratory Health," Meredith Franklin, University of Southern California, Los Angeles Franklin and colleagues developed novel exposure models of tailpipe and nontailpipe air pollutants and noise and applied those models to children's respiratory health in a large Southern California cohort that was also studied in a previous HEI-funded study led by Frank Gilliland; see *HEI Research Report 190*. They made use of the most recent Children's Health Study (CHS) cohort that was initiated in 2003 and included about 2,000 children in eight communities. Longitudinal data on asthma and lung function were collected at various time points (2008–2012) at ages 11 through 16. Air pollution models were supported by particulate matter filters at more than 200 locations in the eight Southern California communities (in press).

"Cardiometabolic Health Effects of Air Pollution, Noise, Green Space and Socioeconomic Status: The HERMES Study," Ole Raaschou-Nielsen, Danish Cancer Society Research Center, Copenhagen, Denmark Raaschou-Nielsen and colleagues evaluated effects of traffic-related air pollution, traffic noise, lack of green space, and other factors on myocardial infarction, stroke, diabetes, and related biomarkers in three cohorts, including an administrative cohort of about 2.6 million Danish adults in the period 2005–2017. They assessed traffic-related air pollution using a chemical transport model for various pollutants, including UFPs and NO_2. In addition, they assessed noise, household- and neighborhood-level socioeconomic status, and various residential green space exposure metrics (*Research Report 222*).

STUDIES FUNDED UNDER RFA 19-1

HEI funded five studies in North America and Europe to evaluate different aspects of improvements to exposure assessment and the application of different exposure assessment approaches to existing cohorts. Three studies focused on combining novel methods for measuring air pollution and diverse exposure assessment approaches to improve exposure assignment, including machine learning and mobile monitoring (Weichenthal and Hoek) and mobility (de Hoogh). Two studies tested the added value of incrementally more complex statistical modeling approaches to improving exposure assessment in London (Katsouyanni) and Seattle (Sheppard) and applied their findings to estimating health effects in epidemiological studies.

"Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada," Scott Weichenthal, McGill University, Montreal, Canada Weichenthal and colleagues estimated associations between long-term exposures to UFPs, BC, and other pollutants and mortality in Toronto and Montreal, Canada, using several exposure modeling approaches. They conducted mobile monitoring campaigns in both cities and used those newly collected data to develop various high-resolution exposure models, including land use regression and machine learning. They then evaluated how the effect estimates for nonaccidental and cause-specific mortality

in the Canadian Census Health and Environment Cohort (CanCHEC) are influenced by different exposure models (*Research Report 217*).

"Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models," Gerard Hoek, Utrecht University, The Netherlands Hoek and colleagues compared the performance of a suite of long-term exposure assessment methods in the Netherlands for four air pollutants. The predictions of the exposure models were compared at 20,000 random residential addresses in the Netherlands and tested on existing and new validation data over a 20-year period. They applied the various models to three major cohorts in the Netherlands — an administrative cohort of 10.8 million adults (DUELS), the European Prospective Investigation into Cancer and Nutrition Netherlands (EPIC-NL), and the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) birth cohort — to evaluate how they influence health effect estimates in epidemiological studies (*Research Report 226*).

"Accounting for Mobility in Air Pollution Exposure Estimates in Studies on Long-Term Health Effects," Kees de Hoogh, Swiss Tropical and Public Health Institute, Basel, Switzerland de Hoogh and colleagues used agent-based modeling to model mobility patterns in Switzerland and the Netherlands based on travel survey information. They used location tracking using a mobile phone application and GPS units for about 700 individuals in the Netherlands and Switzerland. They then compared exposure estimates accounting for mobility to those accounting only for residential exposures in association with health effects estimates in three major cohorts: the Study on Air Pollution and Lung Disease in Adults (SAPALDIA) in Switzerland, participants in the European Prospective Investigation into Cancer and Nutrition Netherlands (EPIC-NL), and the Swiss National Cohort (SNC) (in press).

"Investigating the Consequences of Measurement Error of Gradually More Sophisticated Long-Term Personal Exposure Models in Assessing Health Effects: The London Study (MELONS)," Klea Katsouyanni, Imperial College, United Kingdom Katsouyanni and colleagues evaluated whether successively more detailed estimates of long-term exposure to outdoor air pollution can be used to produce more accurate and realistic estimates of the health effects associated with exposure than are produced by other, less detailed approaches. They leveraged personal exposure data from four earlier studies in London. They compared predictions from various exposure models that accounted for exposure to indoor sources and mobility by using several types of air pollution models (dispersion, land use regression, machine learning, and hybrid models). Finally, exposures were applied to the London segment of the UK Biobank study with about 62,000 participants to evaluate associations with several health outcomes (*Research Report 227*).

"Optimizing Exposure Assessment with Application to Cognitive Function," Lianne Sheppard, University of Washington, Seattle Sheppard and colleagues compared the performance of different exposure assessment study design features on long-term exposure and health estimates in Seattle, for example fewer visits per site, fewer days of the week, fewer times of day, and fewer seasons. They leveraged detailed air pollution data, including a mobile monitoring campaign of UFPs, and cognitive function data from about 5,000 participants in the Adult Changes in Thought (ACT) Air Pollution study. The investigators used either the full dataset or subsets of measurements to develop annual average exposure estimates using a suite of models, including universal kriging, and machine learning models. In particular, they used statistical techniques to assess the bias and precision of health effect estimates and provided practical guidance on future mobile monitoring campaigns (current report).

FURTHER RESEARCH UNDERWAY

Given the large number of people exposed to traffic-related air pollution — both in and beyond the near-road environment — exposures to traffic-related air pollution remain an important public health concern and deserve greater attention from the public and from policymakers.

Although emissions from automobile exhaust systems have decreased in recent years, emissions from the use and wear of brakes, tires, and other nontailpipe sources now contribute a higher fraction of the particulate emissions. Therefore, HEI funded two ongoing studies funded under *RFA 21-1, Quantifying Real-World Impacts of Non-Tailpipe Particulate Matter Emissions*. The two studies involve measurements of mass and composition of ambient particles from nontailpipe motor vehicle sources to disentangle nontailpipe and tailpipe pollution and better understand how each effects human health. One study is measuring concentrations of nontailpipe particulate matter across Toronto, Canada, to determine how much nontailpipe pollution people might breathe in everyday life and how to improve measurement of these exposures in the future. The other study is a panel study in which asthmatic adults rode stationary bicycles on sidewalks in three different exposure environments in London, United Kingdom, to measure how exposure to traffic with different mixtures of nontailpipe and tailpipe emissions affects lung function.

Building on its prior and ongoing research and the recommendations from its systematic traffic review, HEI issued *RFA 23-1, Assessing Health Effects of Traffic-Related Air Pollution in a Changing Urban Transportation Landscape*. Investigators funded under RFA 23-1 will conduct epidemiological and health impact assessment studies to assess current and potential future population-level health effects and health burdens associated with current and future transportation systems and traffic-related air pollution. The studies began in late spring 2024. HEI also publishes reports on the State of Global Air to communicate the relationship between air quality and health

Preface Table. Key Characteristics of HEI's Research to Assess Health Effects of Traffic-Related Air Pollution and to Improve Exposure Assessment for Health Studies

Principal Investigator	Study Name	Location	Study Period	Study Population	Sample Size	Outcomes	Main Air Pollutants	Monitoring Data	Exposure Assessment
RFA 17-1, Assessing Adverse Health Effects of Exposure to Traffic-Related Air Pollution, Noise, and Their Interactions with Socioeconomic Status									
Dadvand	FRONTIER (BiSC)	Barcelona, Spain	2018–2022	Newborns	1,080	Birth weight, small for gestational age, fetal growth trajectories, and placental function	BC, NO_2, $PM_{2.5}$, Cu, Fe, and Zn	Personal, indoor, and outdoor home measurements	LUR, dispersion, and hybrid models
Franklin	CHS	Southern California	2008–2012	Children	2,000	Asthma and lung function	PM_{coarse}, $PM_{2.5}$, Cu, Fe, Zn, and many other elemental components	Outdoor home and school measurements and measurements near road intersections	Machine learning and LUR models
Raaschou-Nielsen	HERMES (DK-POP, DNHS, DCH–NG)	Denmark	2005–2017	Adults	2.9 million	Myocardial infarction, stroke, and diabetes	UFPs, EC, NO_2, $PM_{2.5}$	NA	Chemical transport model
RFA 19-1, Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies									
Weichenthal	CanCHEC	Montreal and Toronto, Canada	1991–2016	Adults	1.5 million	Mortality	UFPs, BC	Mobile	Machine learning and LUR models
Hoek	DUELS, EPIC-NL, PIAMA	Netherlands	1993–2019	Children and adults	10 million	Mortality, stroke, coronary events, lung function, and asthma	UFPs, NO_2, BC and $PM_{2.5}$	Mobile, outdoor low-cost sensors, regulatory monitors	LUR, dispersion, machine-learning, and hybrid models

continued on next page

Principal Investigator	Study Name	Location	Study Period	Study Population	Sample Size	Outcomes	Main Air Pollutants	Monitoring Data	Exposure Assessment
de Hoogh	EPIC-NL, SAPALDIA, SNC	Netherlands, Switzerland	1991–2018	Adults	3.5 million	Mortality, stroke, coronary events, lung function, and blood pressure	NO_2, $PM_{2.5}$	Personal measurements, location tracking, regulatory monitors	Agent-based, LUR, and machine-learning models
Katsouyanni	MELONS (BLW, COPE, DEMiSt, PASTA, London segment of UK Biobank)	London, UK	2006–2024	Adults	62,000	Mortality, asthma, COPD, myocardial infarction, stroke, and dementia	BC, NO_2, $PM_{2.5}$, O_3	Personal measurements, regulatory monitors	LUR, dispersion, machine learning, and hybrid models
Sheppard	ACT Air pollution study	Seattle	1994–2020	Older adults	5,400	Cognitive function	UFPs, $PM_{2.5}$, and NO_2	Mobile, outdoor low-cost sensors	Universal kriging, spatiotemporal models, and machine-learning models

ACT = Adult Changes in Thought; BiSC = Barcelona Life Study Cohort; BLW = Breathe London Wearables; CanCHEC = Canadian Census Health and Environment Cohort; CHS = Children's Health Study; COPE = Characterisation of COPD Exacerbations using Environmental Exposure Modelling; DCH–NG = Diet, Cancer and Health–Next Generations cohort; DEMiSt = Driver Diesel Exposure Mitigation Study; DK–POP = Danish Population cohort; DNHS = Danish National Health Survey; DUELS = Dutch Environmental Longitudinal Study; EPIC–NL = European Prospective Investigation on Cancer and Nutrition–Netherlands; PIAMA = Prevention and Incidence of Asthma and Mite Allergy; NA = not applicable; PASTA = Physical Activity through Sustainable Transport Approaches; SAPALDIA = Swiss Study on Air Pollution and Lung Disease in Adults; SNC = Swiss National Cohort.

around the world; see, for example, a recent report on cities and NO_2 (HEI 2022b).

Looking ahead, HEI continues to support improvements in exposure assessment via the use of new technologies, such as satellite remote sensing data. HEI held a *workshop* to discuss applications of high-quality satellite remote sensing data, which have opportunities for increased use in large epidemiological studies, studying the health effects of wildfires, and addressing environmental justice concerns. Challenges include the complexities of data assimilation and accessibility, and current data and algorithmic limitations. HEI recently issued RFA 25-1, *Advancing Satellite-Derived Air Quality Data and Approaches for Use in Health Studies*. The overarching goal of this RFA is to develop a resource for health research that links satellite-derived air quality products to quantified uncertainties and strengthens the understanding of the implications of such uncertainties for exposure, epidemiological, and health assessment research.

REFERENCES

Apte JS, Chambliss SE, Messier KP, Gani S, Upadhya AR, Kushwaha M, et al. 2024. Scalable Multipollutant Exposure Assessment Using Routine Mobile Monitoring Platforms. Research Report 216. Boston, MA: Health Effects Institute.

Atkinson RW, Butland BK, Anderson HR, Maynard RL. 2018. Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies. Epidemiology 29:460–472, *doi:10.1097/EDE.0000000000000847*.

Barratt B, Lee M, Wong P, Tang R, Tsui TH, Cheng W, et al. 2018. A Dynamic Three-Dimensional Air Pollution Exposure Model for Hong Kong. Research Report 194. Boston, MA: Health Effects Institute.

Batterman S, Berrocal VJ, Milando C, Gilani O, Arunachalam S, Zhang KM. 2020. Enhancing Models and Measurements of Traffic-Related Air Pollutants for Health Studies Using Dispersion Modeling and Bayesian Data Fusion. Research Report 202. Boston, MA: Health Effects Institute.

Frey HC, Grieshop AP, Khlystov A, Bang JJ, Rouphail N, Guinness J, et al. 2022. Characterizing Determinants of Near-Road Ambient Air Quality for an Urban Intersection and a Freeway Site. Research Report 207. Boston, MA: Health Effects Institute.

Glazener A, Sanchez K, Ramani T, Zietsman J, Nieuwenhuijsen MJ, Mindell JS, et al. 2021. Fourteen pathways between urban transportation and health: A conceptual model and literature review. J Transport Health 21:101070, *https://doi.org/10.1016/j.jth.2021.101070*.

Health Canada. 2016. Human Health Risk Assessment for Ambient Nitrogen Dioxide. Ottawa, Ontario, Canada: Water and Air Quality Bureau.

HEI Panel on the Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. 2022a. Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute.

HEI Panel on the Health Effects of Traffic-Related Air Pollution. 2010. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. HEI Special Report 17. Boston, MA: Health Effects Institute.

HEI. 2022b. Air Quality and Health in Cities: A State of Global Air Report 2022. Boston, MA: Health Effects Institute. Available: *https://www.stateofglobalair.org/resources/health-in-cities*.

Hoek G. 2017. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. Curr Environ Health Rep 4:450–462, *doi:10.1007/s40572-017-0169-5*.

Hoek G, Bouma F, Janssen N, Wesseling J, van Ratingen S, Kerckhoffs J, et al. 2025. Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models. Research Report 226. Boston, MA: Health Effects Institute.

Huangfu P, Atkinson R. 2020. Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: A systematic review and meta-analysis. Environ Int 144:105998, doi:10.1016/j.envint.2020.105998.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. 2005. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 15:185–204, *doi:10.1038/sj.jea.7500388*.

Katsouyanni K, Evangelopoulos D, Wood D, Barratt B, Zhang H, Walton H, et al. 2025. Investigating the Consequences of Measurement Error of Gradually More Sophisticated Long-Term Personal Exposure Models in Assessing Health Effects: The London Study (MELONS). Research Report 227. Boston, MA: Health Effects Institute.

Khreis H, Nieuwenhuijsen M, Zietsman J, Ramani T (eds.). 2020. Traffic-Related Air Pollution (1st edition). Waltham, MA: Elsevier.

Raaschou-Nielsen O, Poulsen AH, Ketzel M, Frohn LM, Roswall N, Hvidtfeldt UA, et al. 2024. Cardiometabolic Health Effects of Air Pollution, Noise, Green Space, and Socioeconomic Status: The HERMES Study. Research Report 222. Boston, MA: Health Effects Institute.

Sarnat JA, Russell A, Liang D, Moutinho JL, Golan R, Weber RJ, et al. 2018. Developing Multipollutant Exposure Indicators of Traffic Pollution: The Dorm Room Inhalation to Vehicle Emissions (DRIVE) Study. Research Report 196. Boston, MA: Health Effects Institute.

US EPA. 2016. Integrated Science Assessment for Oxides of Nitrogen–Health Criteria. EPA/600/R-15/068. Washington, DC: US EPA.

US EPA. 2023. Our Nation's Air Trends Through 2022. Available: *https://gispub.epa.gov/air/trendsreport/2023/* [accessed 22 February 2024].

Weichenthal S, Lloyd M, Ganji A, Simon L, Xu J, Venuta A, et al. 2024. Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada. Research Report 217. Boston, MA: Health Effects Institute.

HEI STATEMENT

Synopsis of Research Report 228

Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function

BACKGROUND

There remain important limitations and challenges when assessing long-term exposure to ambient air pollution for use in epidemiological studies. In 2019, the Health Effects Institute issued Request for Applications 19-1 to develop and apply novel, scalable approaches to improve assessments of long-term exposures to outdoor air pollutants that vary widely in space and time.

Dr. Sheppard was one of the five investigators funded under this Request for Applications. Dr. Sheppard and colleagues proposed to advance the understanding of exposure assessment study design features; for example, to investigate the influence of sampling fewer visits per site, fewer days of the week, restricted hours of the day, and fewer seasons. They also included a comparison of health estimates derived from those features. They leveraged detailed air pollution and cognitive function data available at baseline (1994 or later) from the Adult Changes in Thought study in Seattle, which is a cohort study of about 5,400 individuals 65 years of age or older.

What This Study Adds

- The study compared the performance of different exposure assessment study design features on long-term exposure and health estimates in Seattle, Washington.
- It leveraged detailed air pollution data, including a mobile monitoring campaign, and cognitive function data from the Adult Changes in Thought study — a cohort study of older adults.
- The investigators used either the full air pollution dataset or subsets of measurements to develop annual average universal kriging models, machine learning models, and other advanced statistical models.
- The study found that a mobile monitoring study with roadside sampling of ultrafine particles with at least 12 visits per location optimized exposure model performance while limiting costs.
- The study provides practical guidance on future mobile monitoring campaigns, which addresses a clear research gap.

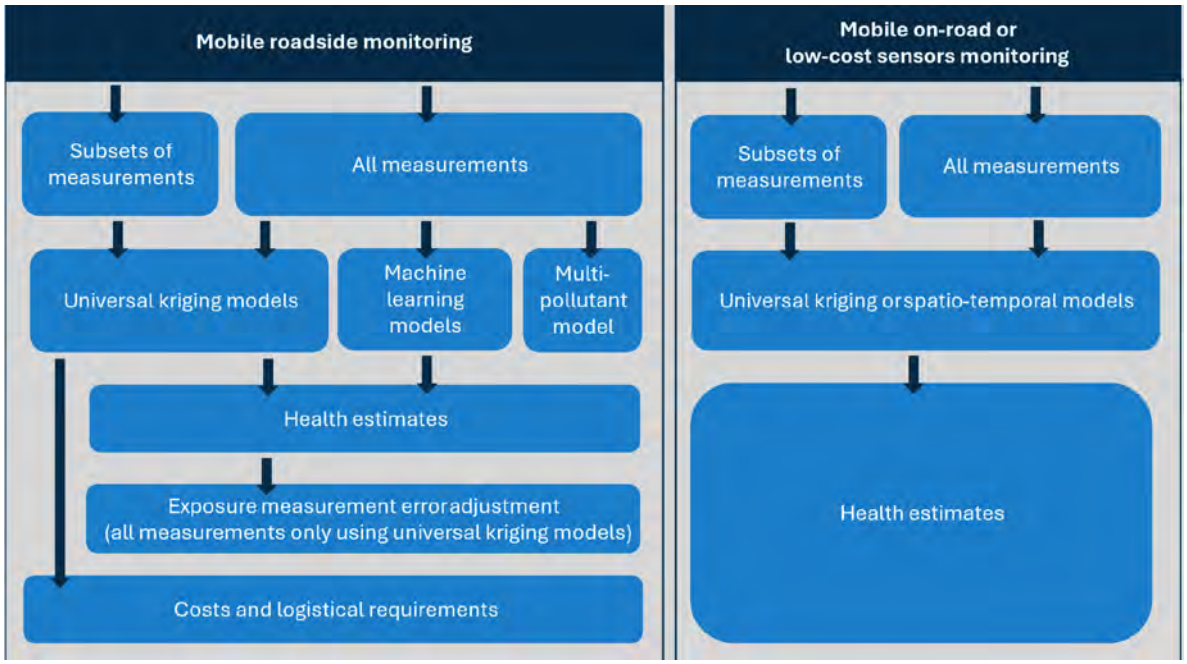
APPROACH

Most analyses focused on ultrafine particle data from a previously conducted mobile roadside monitoring data campaign in 2019–2020 that was designed to capture exposures for the Adult Changes in Thought cohort. In that campaign, short-term measurements were made from a parked vehicle at 309 roadside locations, with about 30 visits made to each site. For the present study, the investigators also conducted analyses using on-road measurements of ultrafine particles between roadside sites along predefined routes, for a total of 5,878 road segments. Each 100-meter segment was visited an average of 28 times. Ultrafine particles were measured with different instruments and size ranges captured, including a P-Trak and a NanoScan. Possible extreme values were replaced with a fixed percentile value for each visit (or segment), averaged across visits and log-transformed for subsequent exposure modeling.

Further analyses were conducted on fine particulate matter and nitrogen dioxide concentrations collected using low-cost sensors at about 115 fixed monitoring sites in 2017–2020, combined with regulatory monitoring data from a much longer time period.

The investigators subsampled measurement data to evaluate various exposure assessment study design features, such as fewer visits per site, fewer days of the week, restricted hours of the day, and fewer seasons. The investigators used either the full dataset or subsets of measurements to develop annual average exposure estimates using a suite of models, including universal kriging, spatiotemporal models, machine

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Lianne Sheppard at the University of Washington and colleagues. Research Report 228 contains the detailed Investigators' Report and a Commentary on the study prepared by the HEI Improved Exposure Assessment Studies Review Panel.



Statement Figure. Schematic overview of the study design.

learning, and other advanced statistical models (**Statement Figure**). They assessed the performance of each model using cross-validation (mobile monitoring data) or a combination of cross-validation and external validation (low-cost sensor data). The investigators reported several measures to test the performance of the exposure models, including the root mean squared error and the mean squared error-based explained variance. The investigators also explored the tradeoffs between universal kriging exposure model performance and logistical features (both cost and time) to identify optimal monitoring designs.

Each model was used to predict the 5-year average ultrafine particles or other pollutant exposures prior to the cognitive function measurement that was obtained at baseline (1994 or later). The investigators applied standard linear regression to assess the association between exposure to ultrafine particles from each exposure model and baseline cognitive function, adjusted for participant age, sex, education, and calendar year (confounder model 1). Some results were also adjusted for participant race and socioeconomic status (confounder model 2).

In all comparisons, the models using all data from the mobile roadside campaign or all the low-cost sensor data were taken as reference models. Because even the reference models contain exposure measurement errors, the investigators quantified the exposure measurement error using bootstrap methods and corrected the health effect estimates accordingly in the reference model for ultrafine particles.

KEY RESULTS

The universal kriging reference model using all ultrafine particle data from the mobile roadside campaign had a cross-validated mean squared error-based explained variance of 0.65 (NanoScan) and 0.77 (P-Trak). The universal kriging model typically performed slightly better than the various machine learning models.

The universal kriging models with restricted mobile roadside sampling of ultrafine particles almost always produced lower-performing exposure models compared to the reference model. The investigators found that a mobile monitoring study with roadside sampling of ultrafine particles with at least 12 visits per location optimized exposure model performance while limiting costs. Furthermore, the investigators noted that it is important that the exposure sampling in mobile monitoring campaigns covers all days of the week, most hours of the day (including early morning and late evening hours), and at least two seasons.

Exposure of ultrafine particles estimated using the reference model was negatively (adversely) associated with cognitive function at baseline when adjusted for participant age, sex, education, and calendar year (confounder model 1). Cognitive function's association with ultrafine particles was –0.020 (95% confidence interval: –0.036 to –0.004) per increase of 1,900 particles/cm^3. The reduced sampling designs led to similar findings in terms of negative (adverse) associations between ultrafine particle exposure and cognitive function at baseline. However, the strength (magnitude) of the observed negative associations sometimes differed

substantially, especially for the business and rush hours designs, which decreased associations by up to 60%. The investigators reported that the observed negative associations using confounder model 1 were affected more by features of the mobile monitoring design than by accounting for exposure measurement error in the reference exposure model for ultrafine particles.

Notably, the observed negative association in the reference model disappeared in health models that also adjusted results for race and socioeconomic status (confounder model 2). The null findings from the reference model using confounder model 2 hampered the assessment of the influence of sampling design on health estimates using different exposure estimates for ultrafine particles.

Using mobile on-road data for ultrafine particles, the investigators found that most comparisons identified the same design features or elements as important, although with a few notable differences.

The addition of low-cost sensor data to data from regulatory monitors and other research-grade monitors improved exposure modeling for fine particulate matter. Increasing the number of low-cost sensor locations and repeated measurements resulted in better exposure model performance. In contrast, in most comparisons for nitrogen dioxide, the addition of low-cost sensor data improved exposure estimates only slightly. Largely null findings were reported between fine particulate matter, nitrogen dioxide, and cognitive function for the various exposure models with and without low-cost sensor data. The null results hampered the assessment of the influence of adding low-cost sensor data for health estimates.

INTERPRETATION AND CONCLUSIONS

In its independent review, the HEI Review Panel thought the study was well-motivated and appreciated that it leveraged detailed air pollution and cognitive function data from the Adult Changes in Thought study in Seattle. The study provides practical guidance on future mobile monitoring campaigns, which addresses a clear research gap. The extensive year-long mobile monitoring campaign and the evaluation of various exposure assessment study design features were strengths of the study. Another strength was the extensive air pollution exposure modeling and rigorous evaluation of their performance. The Panel was also impressed by the large number of publications resulting from the work.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results. Many of the analyses focused on ultrafine particles, and there were few comparisons across pollutants, although more information is presented in the Additional Materials and other publications. For the evaluation of exposure design features using mobile roadside and mobile on-road monitoring data of ultrafine particles, findings were presented in different stand-alone chapters, and different monitoring instruments were selected for various analyses. This makes a direct comparison between the two monitoring approaches difficult. An analysis investigating how the removal of possible extreme values affected the subsequent exposure models was missing from the report, but was included in a paper resulting from this work.

The Panel recommended caution in interpreting the findings from the health analyses and thought some carefully designed simulations would have complemented the real-world health study. The health analyses were considered limited, particularly because most of the exposure models used were based on measurements conducted up to 25 years later than the health outcome. In addition, some analyses lacked adjustment for important confounding variables — most notably socioeconomic status. Further research in other cities and pollutants would be helpful to assess the generalizability of the specific findings related to exposure assessment design.

The comprehensive report includes many findings that will be of broad interest and value to a wide audience.

INVESTIGATORS' REPORT

Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function

Lianne Sheppard[1], Magali N. Blanco[1], Sun-Young Kim[4], Annie Doubleday[2], Si Cheng[3], Christopher Zuidema[1], Jianzhao Bi[1], Amanda Gassett[1], Ali Shojaie[1], and Adam A. Szpiro[1]

[1]University of Washington, Seattle, Washington, USA; [2]Department of Health, Olympia, Washington, USA; [3]Netflix, Inc., Los Gatos, California, USA; [4]National Cancer Center Graduate School of Cancer Science and Policy, Gyeonggi-do, South Korea

ABSTRACT

Introduction Epidemiological studies often make use of exposure data that is collected in opportunistic and logistically convenient ways. And, while exposure assessment is fundamental to environmental epidemiology, little is known about what exposure assessment study designs are optimal for health inference. The objective of this project was to advance our understanding of the design of exposure assessment measurement campaigns and evaluate their impact on estimating the associations between long-term average air pollution exposure and cognitive function. This feeds into the broader goal of advancing understanding of air pollution exposure assessment design for application to epidemiological inference.

Methods We leveraged data from the Adult Changes in Thought (ACT*) Air Pollution study (ACT-AP) to characterize exposures for over 5,000 participants from the ongoing ACT cohort. This is a population-based cohort of urban and suburban elderly individuals in the greater Puget Sound region drawn from Group Health Cooperative, now Kaiser Permanente, starting in 1994. Participants were routinely followed with routine biennial visits until dementia incidence, drop-out, or death. Extensive health, lifestyle, biological, and demographic data were also collected. The outcome measure used in this report is cognitive function at baseline based on the Cognitive Abilities Screening Instrument derived using Item Response Theory (CASI-IRT). The IRT transformation of the CASI score improves score accuracy, measures cognitive change with less bias, and accounts for missing test items. Health association analyses were based on 5,409 participants with both a valid CASI score and who had lived in the mobile monitoring region during at least 95% of the 5 years prior to baseline. We used 5-year average exposures that accounted for residential history.

Exposure data came from two distinct exposure assessment campaigns carried out by the ACT-AP study: a campaign using low-cost sensors (2017+) that supplemented existing regulatory monitoring data for fine particles ($PM_{2.5}$, 1978+) and nitrogen dioxide (NO_2, 1996+), and a year-long multipollutant mobile monitoring campaign (2019–2020). The evaluation of the added value of low-cost sensor data relied on a combination of regulatory monitoring data and other high-quality data from research studies, calibrated 2-week low-cost sensor measurements from over 100 locations, which were mostly ACT cohort residences, and a snapshot campaign that measured NO_2 using Ogawa samplers. Predictions were at a 2-week average time scale, used a suite of ~200 geographic covariates, and were obtained from a spatiotemporal model developed at the University of Washington. The Seattle mobile monitoring campaign collected a combination of stationary roadside and on-road measurements of ultrafine particles (UFPs, four instruments), black carbon (BC), NO_2, carbon dioxide (CO_2), and $PM_{2.5}$. Visits were temporally balanced over 288 drive days such that all sites were visited during all seasons, days of the week, and most hours of the day (5 a.m. to 11 p.m.) approximately 29 times each. For the on-road measurements, we divided the driving route into 100-meter segments and assigned all measurements to the segment midpoint. Predictions used the same suite of geographic covariates in a spatial model fit using partial least squares (PLS) dimension reduction with universal kriging (UK-PLS) to capture the remaining spatial structure. We reported model performance metrics for both the spatial and spatiotemporal models as root mean squared error (RMSE) and mean squared error (MSE)-based R^2. The reference observations for the spatiotemporal model were low-cost sensor measurements at home locations (with performance metrics averaged over their entire measurement period to approximate spatial contrasts), and for the spatial

This Investigators' Report is one part of Health Effects Institute Research Report 228, which also includes a Commentary by the Improved Exposure Assessment Studies Review Panel and an HEI Statement about the research project. Correspondence concerning the Investigators' Report may be addressed to Dr. Lianne Sheppard, Department of Environmental and Occupational Health Sciences, University of Washington, Box 351618, Seattle, WA 98195; email: *sheppard@uw.edu*. No potential conflict of interest was reported by the authors.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and might not necessarily reflect the views of the Agency; no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it might not reflect the views or policies of these parties, and no endorsement by them should be inferred.

* A list of abbreviations and other terms appears at the end of this volume.

model, the reference observations were the all data long-term averages at stationary roadside locations.

Using various approaches to sample data from these two exposure monitoring campaigns, we determined the impact on exposure prediction and estimates of health associations using two confounder models and 5-year average exposure predictions for cohort members at baseline developed from the alternative campaigns. For the low-cost sensor data, we evaluated temporally or spatially reduced subsets of low-cost sensors, as well as a comparison of the low-cost sensor versus snapshot campaigns for NO_2. For the mobile monitoring data, we considered designs focused on the stationary roadside and on-road data separately. We reduced the stationary roadside data temporally by restricting seasons, times of day, or days of week for the campaign, while also considering a reduced number of visits using balanced sampling, as well as a set of unbalanced visit designs. We also reduced the on-road data spatially and temporally to assess the importance of spatially or temporally balanced data collection. In addition, we considered the impact of incorporating temporal adjustment to account for temporally unbalanced sampling, as well as plume adjustment to account for on-road sources. For each design, we evaluated prediction model performance using the all data stationary roadside observations (mobile campaign) or the measurements at homes (low-cost sensor campaign) as reference observations to ensure consistency in reported performance metrics. We also used long-term average exposures estimated from these alternative campaigns in health association analyses under two different confounder models that were adjusted by potentially confounding variables: Model 1 adjusted for age, calendar year, sex, and educational attainment; Model 2 included all Model 1 variables with the addition of race and socioeconomic status. Furthermore, using the stationary roadside data, we applied parametric and nonparametric bootstrap methods to account for Berkson-like and classical-like exposure measurement error for the UFP exposure in confounder model 1.

In a separate methods-focused aim, we developed and applied advanced statistical methods using the stationary roadside mobile monitoring data. To evaluate possible improvements in exposure model performance, we applied tree-based machine learning algorithms that also account for residual spatial structure, and compared these to UK-PLS. This led to the development of a variable importance metric that uses a leave-one-out approach to evaluate the change in predictions across various user-specified quantiles. The variable importance metric produces covariate-specific averages that reflect how the predictions, on average, vary across different quantiles of each covariate. This serves as an intuitive measure of the contribution of this covariate to the predicted outcome. A key idea in this variable importance approach is to reuse the trained mean model across all locations and to refit the covariance model in a leave-one-out manner. In separate work to address dimension reduction for multipollutant prediction, we extended classical principal component analysis (PCA) and a recently developed predictive PCA approach to optimize performance by balancing the representativeness in classical PCA with the predictive ability of predictive PCA. We called the new method representative and predictive PCA, or RapPCA.

Finally, we characterized the various exposure assessment campaigns in terms of the value of their information as quantified by cost. We calculated costs, focused predominantly on staff days of effort, for various exposure assessment designs and compared these to exposure model performance statistics.

Results We found that air pollution exposure assessment design is critical for exposure prediction, and also impacts health inference. We showed that a mobile monitoring study with stationary roadside sampling that has at least 12 visits per location in a balanced and temporally unrestricted design optimizes exposure model performance while also limiting costs. Relative to weaker alternatives, a balanced and temporally unrestricted design has improved accuracy and reduced variability of health inferences, particularly for confounder model 1. To address temporal balance, it is important that the exposure sampling in mobile monitoring campaigns cover all days of the week, most hours of the day, and at least two seasons. The popular temporally restricted business-hours sampling design had the poorest performance, which was not improved by adjusting for the temporally unbalanced sampling approach. We found similar patterns using on-road data, though the findings were weaker overall.

For the alternative exposure campaign that supplemented regulatory monitoring data with low-cost sensor data, while the exposure prediction model performances improved with the inclusion of the low-cost sensors, there was little notable impact on the health inferences, and the costs were steep. Given that the supplementary exposure assessment data were sparse relative to the existing regulatory monitoring data, and that the low-cost sensor data collection used a rotating approach due to the limited number of sensors (i.e., low-cost sensor measurements were not collected using a balanced design), it was much more challenging to develop deep insights from this exposure assessment approach.

Finally, we found that leveraging spatial ensemble-learning methods for prediction did not improve exposure prediction model performances or alter health inferences. The new multipollutant dimension-reduction we developed, RapPCA, had the best predictive performance and also minimized the prediction error in comparison with both classical and predictive PCA.

Conclusions This project has shown that there should be greater attention to the design of the exposure data collection campaigns used in epidemiological inference. Based on the multiple investigations conducted, many of which focused on UFPs, we found that exposure predictions with better performance statistics resulted in health association estimates that were generally more consistent with those obtained using the "best" exposure model predictions (the model with all data included), although the pattern of health estimates was often

less conclusive than the pattern of prediction model performances. Furthermore, we found that it is possible to design air pollution exposure assessment studies that achieve good exposure prediction model performance while controlling their relative cost.

We developed strong recommendations for mobile monitoring campaign design, thanks to the well-designed and comprehensive Seattle mobile monitoring campaign. Insights from supplementing regulatory monitoring data with low-cost sensor data were less compelling, driven predominantly by a data structure with sparse and temporally unbalanced supplementary data that may not have been sufficiently comprehensive to demonstrate the impacts of alternative designs. Broadly speaking, better exposure assessment design leads to better exposure prediction model performance, which in turn can benefit estimates of health associations.

We did not find that leveraging advanced statistical methods (specifically spatial ensemble-learning methods for prediction) improved exposure prediction model performances. This finding is not consistent with the conclusions reached by other investigators, and may have been due to the already sophisticated UK-PLS approach we used by default, and in particular its application in conjunction with the large number of covariates that we considered in the PLS model, such that the contribution of any single covariate was approximately linear. In other words, it is reasonable to believe that in the presence of the large set of covariates we considered, each can contribute an approximately linear association with the pollutant being modeled, such that the potential added value of the spatial Random Forest approach is not observed in the model fit. Other settings with a smaller number of possible covariates available may lead to different conclusions and suggest greater added value of the application of a spatial Random Forest approach.

We based our approach on leveraging the extensive air pollution exposure assessment and outcome data available from the ACT-AP study. Thus, we sampled from the existing air pollution data to evaluate exposure assessment designs that were subsets of those data. Then, conditional on each of these designs, we evaluated subsequent health inferences, which focused on cognitive function at baseline using the CASI-IRT outcome. The magnitude and uncertainty of these health association estimates were dependent upon the associations evident in the ACT cohort, and the insights we were able to develop are conditional on the strengths and weaknesses of these data. Specifically, while we observed some larger impacts on health association estimates of more poorly performing exposure models relative to the complete all data exposure model, such as the business-hours design from a mobile monitoring campaign, many of the differences were small and did not deviate meaningfully from the health association estimate obtained from the "best" exposure model. The degree of impact on the epidemiological inference depended on the magnitude of the health association estimate from the "best" exposure model and the width of its confidence interval. Future investigations should replicate and expand upon these findings in other settings, including application to new cohorts and exposure assessment data, as well as in simulation studies, which provide an alternative approach to using real-world data to evaluate a constellation of exposure models. However, while knowledge of the assumed underlying truth is an important strength of simulation studies, it is challenging to capture real-world complexity meaningfully in simulation studies.

Our foray into applying advanced machine-learning methods to improve exposure predictions produced the surprising result that our default UK-PLS approach for spatial prediction produced similar performance metrics to spatial ensemble-learning methods. Future evaluations that assess smaller subsets of exposure covariates will allow determination of the relative exposure model performance benefits of UK-PLS versus spatial ensemble-learning methods, and provide insights into the possible reason that our conclusions differ from others in the literature.

CHAPTER 1: INTRODUCTION

Exposure assessment is fundamental to environmental epidemiology. Determining the most effective, feasible, and cost-effective approaches to improve exposure assessment for air pollution cohort studies will greatly enhance the quality of possible inferences about health effects. In particular, quality exposure assessment is important for studies of brain health, where evidence increasingly indicates that air pollution, particularly fine particles ($PM_{2.5}$) and traffic-related air pollution (TRAP), may be associated with brain health outcomes such as cognitive function (Delgado-Saborit et al. 2021; Peters et al. 2019). Some of the pollutants of greatest emerging interest, such as black carbon (BC) and ultrafine particles (UFPs), are not routinely measured at most government monitoring sites and cannot be used as exposures in epidemiological cohort studies without the adoption of special, and often costly, exposure assessment campaigns. Although many strategies can be employed to measure air pollution levels in communities, exceedingly little is known about the comparability of newer versus more traditional methods, or how different choices influence epidemiological inferences. Thus, it is important to investigate novel approaches to exposure assessment and articulate the key design and data analysis features of these studies that will improve inference in epidemiological cohort studies. To truly guide future studies, these improvements must also be considered within the context of the value of the exposure information, namely, cost and logistical feasibility.

The goal of this research is to improve the inferential strength of air pollution cohort studies by creating a diverse set of high-quality exposure metrics developed from different types of monitoring campaigns, and then assessing their value in the context of exposure assessment study design, epidemiological inference, cost, and feasibility. To accomplish this goal, we focused on a specific dataset and complemented this with extensive evaluations of alternative exposure assessment designs by subsampling our exposure data. We leveraged data from the NIH-funded Adult Changes in Thought Air Pollution (ACT-AP) cohort study, which in turn relied on the long-standing Adult Changes in Thought (ACT) cohort study. ACT began collecting data in 1994. ACT-AP collected exposure measurements using low-cost sensors beginning in 2017, and then, in 2019–2020, under supplemental funding from the National Institute on Aging, conducted mobile monitoring using a campaign specifically designed for this cohort. We also created exposure estimates of $PM_{2.5}$ and nitrogen dioxide (NO_2) using a spatiotemporal exposure prediction approach, as well as spatial predictions of UFPs, BC, and NO_2 from the mobile data. Finally, we conducted inference about cognitive function and dementia incidence for the primary purpose of more deeply understanding the link between air pollution and the aging brain, some of which have been published (Blanco et al. 2024; Shaffer et al. 2021a, 2021b). In this report, we leverage the extensive data provided by ACT and ACT-AP to determine how to best conduct exposure assessment for epidemiological inference by focusing on cross-sectional analyses of cognitive function as the outcome.

CHAPTER 2: SPECIFIC AIMS AND OVERARCHING APPROACH

Our scientific objective is to compare and contrast the scientific and logistical benefits of different exposure measurement campaigns and sampling design choices for air pollution exposure assessment in environmental epidemiology. Our broad aim is to identify those design and exposure modeling choices that optimize estimates of long-term average outdoor air pollution exposures to use in cohort study inference, focusing on (1) criteria air pollutants typically measured at multiple locations in a study region ($PM_{2.5}$ and NO_2), and (2) other pollutants often unavailable in cohort studies, specifically BC, UFP, and multipollutant mixtures. See **Table 2.1**. Our goal is to identify exposure assessment study designs that don't require extensive additional study-sponsored sampling and achieve good health association estimates, ideally for the lowest possible cost. We evaluated this broad objective by considering exposure assessment design in Aim 1, statistical methods to improve exposure prediction in Aim 2, and health association estimates in Aim 3. Finally, we put these insights together with cost estimates to understand the value of information in Aim 4. The **Research Roadmap** table elaborates on these specific aims and provides an overview of the methods.

We describe the data used in this report in Chapter 3. This includes the two exposure datasets briefly summarized in Table 2.1 as well as the baseline cognitive function data from the ACT cohort. We also describe the spatial and spatiotemporal exposure prediction approaches and the basic structure of our health models for cognitive function, as these are shared across multiple chapters. The Research Roadmap gives further elaboration of the specific investigations conducted for each aim and the chapters where these are presented.

Table 2.1. Pollutants in This Project and the Chapters Where They Are Presented

	Supplementary Low-Cost Sensors at >100 Locations (and for NO_2: Ogawa Snapshots at 110 Locations)		Mobile Monitoring Campaign with 309 Stationary Roadside Locations And >1,400 Hours of Mobile Monitoring	
Pollutant	**Included?**	**Chapter**	**Included?**	**Chapter**
$PM_{2.5}$	X	7, 9	Dropped for most analyses[a]	8
BC			X	8
UFP – P-Trak			X	4, 6, 8, 9
UFP – NanoScan			X	4, 5, 8
UFP – DiSCmini			X	8
NO_2	X	7, 9	X	8
Multipollutant			X	8

[a] Spatial model predictions of $PM_{2.5}$ from the mobile campaign data were, on average, lower than and were poorly correlated with the annual average predictions from the spatiotemporal model over the same time period, suggesting that the nephelometer on our mobile platform did not adequately characterize $PM_{2.5}$. See Chapter 3's Additional Materials for additional details.

Research Roadmap

Aims and Research Conducted	Methods Description
Aim 1: Identify key design choices to improve long-term average exposure prediction using: Mobile monitoring campaigns (Aim 1a) and stationary networks of low-cost sensors (Aim 1b). We subsampled exposure data to evaluate various exposure assessment study designs. These are covered in Chapter 4 for stationary roadside mobile monitoring data, Chapter 6 for on-road mobile monitoring data, and Chapter 7 for low-cost sensor data.	We considered two distinct datasets (Table 2.1) and exposure prediction approaches. We predicted exposures using a spatial model for the Seattle mobile monitoring campaign (Aim 1a), focusing on UFPs from the NanoScan (Chapter 4) and P-Trak (Chapter 6), and a spatiotemporal model for the low-cost sensor data (Aim 1b), focusing on $PM_{2.5}$ and NO_2. We compared exposure model prediction performances relative to the all data campaigns using reference observations.
Aim 2: Develop annual average TRAP exposure predictions from a mobile platform using novel statistical methods not previously employed. We applied spatial ensemble-learning methods in Aim 2a, explored leveraging the road network combined with spatial proximity in Aim 2b, and developed a multipollutant dimension reduction approach for improving predictions in Aim 2c. These are discussed in Chapter 8.	In Aim 2a we applied recently developed spatial ensemble learning methods by replacing the linear mean model in universal kriging with a spatial random forest method and evaluated this for all pollutants measured in the mobile monitoring campaign. The results led us to develop a new variable importance metric applicable to additive models with a correlated error term for comparing machine learning tools. In Aim 2c we developed a representative and predictive principal component analysis and applied this to the mobile monitoring data to produce a lower-dimensional summary of multipollutant data.
Aim 3: Determine the impact on inference of different predictions based on sampling designs and analysis approaches. We considered applied analyses by (1) plugging in the alternative predicted exposures developed in Aims 1 and 2 (Aim 3a), as well as (2) addressing the impact of measurement error correction methods (Aim 3b). The Aim 3a results are presented along with the exposure predictions in Chapters 4, 6, 7, and 8. The Aim 3b measurement error corrections are presented separately in Chapter 5.	We focused on a cross-sectional analysis of the ACT cohort's cognitive function outcome at baseline: the score from the Cognitive Abilities Screening Instrument using Item Response Theory (CASI-IRT). We conducted the same inferential analyses across a wide range of exposure assessment designs and statistical approaches, which were developed in Aims 1 and 2, and reported these along with the exposure prediction performance statistics. Additionally, we quantified the bias and variance of the health association estimates for UFPs from the NanoScan by applying bootstrap measurement calculations to stationary data from the mobile monitoring campaign.
Aim 4: Address the overall value of incorporating novel exposure data collection and modeling by comparing the logistical features (cost and time) of using different sampling designs and analysis choices. We present our recommendations for the design of mobile campaigns in Chapter 9, and comment on strategies for supplementing regulatory monitoring data with low-cost sensors in Chapter 9's Additional Materials.	We calculated costs for collecting exposure data divided into three types and compared these costs, along with exposure prediction model performances, across various exposure assessment designs. We focused on UFPs from the P-Trak from the Seattle mobile monitoring campaign and addressed low-cost sensor designs for NO_2 and $PM_{2.5}$ in the Additional Materials.

CHAPTER 3: OVERVIEW OF THE DATA USED IN THIS REPORT

INTRODUCTION

The purpose of this chapter is to describe the data used throughout this report. As summarized in Chapter 2, we cover three categories of data: mobile monitoring exposure data, low-cost sensor with regulatory monitoring exposure data, and health outcome data. The health outcome data includes a description of the ACT cohort, which underlies all the work in this project. In the following sections, we briefly summarize the exposure data and models, followed by the health outcome data and models. First, we briefly describe the Seattle mobile monitoring campaign and data, which were used in all aims, as well as our general approach to spatial exposure prediction modeling and validation. Then we describe the supplemental monitoring campaign with low-cost sensors that was part of the NIH-funded ACT-AP study. We further discuss the low-cost sensor and related agency data, along with the spatiotemporal prediction modeling approach that we leveraged for addressing questions about the added value of low-cost sensor data in Aims 1, 3, and 4. Before concluding our exposure-focused presentation, we make a brief note about the reference data we used in the exposure model performance estimates to highlight an important nuance. Finally, we give a brief overview of the ACT cohort, the cognitive function outcome, the model we used to address Aim 3, and summarize health estimates for primary exposure models considered in later chapters.

SEATTLE MOBILE MONITORING CAMPAIGN

Mobile Monitoring Data Collection, Quality Control, and Distillation for Analysis

We used data from the Seattle mobile monitoring campaign, a multipollutant mobile monitoring campaign that characterized TRAP exposure levels for the ACT cohort (see cohort details below) to address Aims 1–4 (Blanco et al. 2022, 2023a). To the best of our knowledge, this was one of the most extensive mobile monitoring campaigns conducted in terms of the pollutants measured, the spatial coverage and resolution, and the campaign duration and sampling frequency. The campaign and exposure model development are summarized below. Refer to the publications cited above for extensive additional details.

The Seattle mobile monitoring campaign was conducted between March 2019 and March 2020 in a 1,200 km^2 (463 mi^2) land area within the greater Seattle, WA area (Blanco et al. 2022). The map in Figure S3.1 (see Additional Materials on the HEI website) shows the size of the monitoring region with participant locations marked in red within the state of Washington. This region is a subset of the larger spatiotemporal modeling region (see the next section) with participant locations shown in black.

The mobile monitoring region was composed of census tracts where most of the ACT cohort (87% of the 11,904 locations) had historically resided between 1989 and 2018. This large region fell in western King County and southwest Snohomish County, and it included a variety of urban and rural areas with various land uses, including residential, industrial, commercial, and downtown areas. We used the Location-Allocation tool in ArcMap (ArcGIS v. 10.5.1) (Esri, 2019) to select 304 stationary roadside sites (also referred to as "roadside sites") within the monitoring region that were representative of the ACT cohort (approximately one monitoring site per 33 participant locations; see Blanco et al. 2022 Supplementary Information). Roadside sites were spatially distributed so that they covered all parts of the monitoring region. The exact sites selected were meant to minimize the distance between the monitoring and cohort locations. Five additional sites were co-locations at nearby agency air quality monitoring sites measuring pollutants similar to our platform. In total, there were 309 stationary roadside sites, most of which were along A4 (local and neighborhood roads; $n = 282$, 91%) and A3 (county and single-lane state highways; $n = 27$, 9%) roads. The average (standard deviation [SD]) distance between a cohort location and the nearest monitoring site was 611 (397) meters. The monitoring sites and cohort locations had similar distributions of various TRAP-related covariates (e.g., proximity to roadways, airport, railyard), indicating good spatial compatibility (see Blanco et al. 2022).

A hybrid vehicle outfitted with equipment visited all stationary roadside sites approximately 29 times throughout the campaign and collected 2-minute measurements (visits) with various instruments, including those for measuring particle number concentration (PNC) (TSI P-Trak 8525, 20–1,000 nm particles — two instruments, one with a diffusion screen; TSI NanoScan 3910, 10–420 nm particles; Testo DiSCmini 10–700 nm), BC (AethLabs microAeth MA200), $PM_{2.5}$ (Radiance Research Nephelometer), carbon dioxide (CO_2) (Li-Cor LI-850), and NO_2 (Aerodyne Research Inc. CAPS). The use of a hybrid vehicle meant that the vehicle was operating by battery with the engine off during roadside short-term measurement periods, thus reducing the possibility of self-contamination. Measurements were also collected on-road between stationary sites along fixed routes that were designed to minimize travel on A1 roads and capture additional exposure information relevant to cohort locations. This design also allows for analysis of mobile on-road data and comparison with the stationary roadside data. **Figure 3.1** depicts the nine monitoring routes, which were driven in both the forward and backward direction, along with the stationary roadside locations and ACT cohort residential locations. Visits were temporally balanced over 288 drive days such that all sites were visited during all seasons, days of the week, and most hours of the

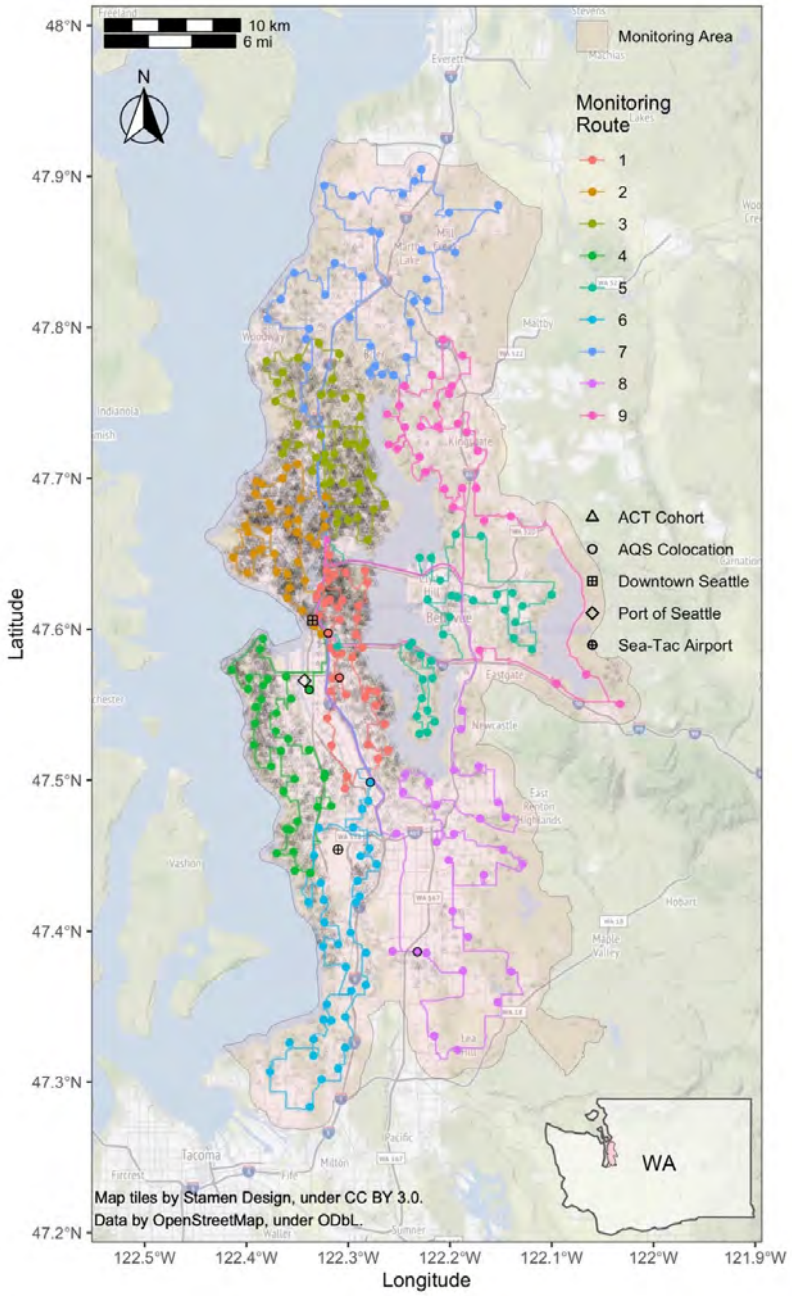


Figure 3.1. Mobile monitoring routes from the Seattle mobile monitoring campaign. There were 309 stationary roadside sites along 9 routes and 10,330 unique jittered ACT cohort locations. The inset map shows the monitoring area within Washington (WA) state. Reprinted with permission from Blanco and colleagues (2022), Copyright 2022 ACS.

day (5 a.m. to 11 p.m.). As detailed in Chapter 3's Additional Materials, we decided not to use the $PM_{2.5}$ predictions from this campaign for most of the analyses in this report because these predictions had limited variability, had poorer model performances, and were poorly correlated with predictions from the spatiotemporal model from the same time period.

We conducted various quality assurance and quality control activities throughout the study period to ensure the reliability and integrity of our data. Activities included calibrating gas instruments, checking particle instruments for zero concentration responses, assessing co-located instruments for agreement, inspecting time-series data for concentration pattern anomalies, and dropping readings associated with instrument error codes or those outside the instrument measurement range. See Blanco et al. 2022 Supplementary Information for additional details.

For the stationary roadside measures, we calculated the median pollutant concentration of each 2-minute stationary visit because most instruments used in this study collected measurements every 1–10 seconds. These visit concentrations were winsorized at the site level such that visit concentrations below the 5th and above the 95th quantile were set to those thresholds to reduce the influence of extreme observations on annual average site concentrations. This approach produces annual average site concentrations that are influenced by high concentrations (which may reflect a true phenomenon) while being robust to the most extreme observations, which may be overly influential in a limited data setting (i.e., mobile monitoring). We previously showed that winsorizing prior to averaging can slightly improve some pollutant models due to the reduction of influential points (Blanco et al. 2022). On the other hand, using approaches completely robust to influential points (i.e., medians) can at times produce worse-performing models in settings with limited spatial variability (e.g., CO_2 exposure surfaces). For consistency across pollutants, we winsorized all pollutant measures and log-transformed site averages before modeling.

For the on-road measurements, we divided the driving route into 100-meter segments and assigned all measurements to the segment midpoint (i.e., on-road "sites"). We excluded A1 roads (interstates and highways with restricted access) because these are not representative of residential exposures, segments with fewer than a median of five 1-second measurements per visit, and segments with fewer than 23 repeat visits. This yielded 5,887 100-meter road segments, as detailed in Doubleday and colleagues (2023; note S1). We averaged the UFP measurements to 10-second periods, calculated the median UFP concentrations across all 10-second measurements from each segment and visit, winsorized these across visits at the segment level, and calculated mean visit concentrations per road segment. This resulted in a median of 28 (interquartile range [IQR]: 27–28) visit concentrations per road segment. Data were adjusted to

reduce the influence of high on-road plume concentrations (Doubleday et al. 2023). In summary, this method applies an absolute principal component score model to pollutants measured alongside UFPs (BC, NO_2, CO_2, and all UFP measurements from the P-Trak and DiSCmini) to identify on-road source components, to estimate adjusted UFP concentrations by reducing the contributions from these on-road factors, and to use stationary location measurements that are not plume impacted in an iterative process to select the final adjustment.

Exposure Prediction Modeling Using Data from the Mobile Monitoring Campaign

For modeling, we used a suite of geographic covariates estimated at all monitoring and participant locations. These were generated using PostGIS (ver. 2.4.4, *http://postgis.net*) and include over 800 proximity and buffer measurements. We preprocessed these to eliminate those that were potentially too influential or inadequately informative, using standard procedures (Keller et al. 2015; M. Wang et al. 2015). This resulted in under 200 geographic predictors offered to each model. Table S3.2 provides a summary of selected covariates; they include characterization of land use, roadway proximity, population density, and other geographic predictors of TRAP.

These data were used to develop what we refer to in this study as "all data" annual averages for each stationary roadside and nonstationary on-road location and treated as gold-standard reference estimates for model evaluation for the results presented in Chapters 4, 5, and 6. We built annual average universal kriging (UK) with partial least squares (PLS; i.e., UK-PLS) pollutant exposure models. The dependent variables in these models were the log-transformed annual average site concentrations. The independent variables were the first two PLS components, which summarized hundreds of geographic covariate predictors (**Equation 3.1**). The models were

$$Ln(Conc_i) = \alpha + \sum_{m=1}^{M} \theta_m Z_{mi} + \varepsilon_i \quad (3.1)$$

where *Conc* is the pollutant concentration at the i^{th} location, are the first two PLS principal component scores ($M = 2$), and are model coefficients, and is the residual term with mean zero and a modeled geostatistical structure.

We evaluated the predictive performance of each campaign using fivefold cross-validation (CV). Sites were randomly selected into one of five groups (folds), and for each campaign, five models were built, each time using site concentrations from four of the folds to build a model and predict concentrations at sites from the remaining fold. Model predictions were evaluated by comparing cross-validated site model predictions (on the native scale) to the respective all data annual average site estimates (our best estimates) using root mean squared error (RMSE) and mean squared error (MSE)-based R^2. RMSE and MSE-based R^2 are based on the MSE using observations from the reference data (our best estimates of the true annual average):

$$MSE_{ref} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,ref} - \hat{y}_{i,campaign}\right)^2 \quad (3.2)$$

where $\hat{y}_{i,campaign}$ is the prediction from a campaign for a given design, $y_{i,ref}$ is the reference all data annual average at a given site unless otherwise noted, and n is the total number of sites. RMSE is defined as the square root of MSE:

$$RMSE_{ref} = \sqrt{MSE_{ref}} \quad (3.3)$$

and MSE-based R^2 is defined as

$$R^2_{MSE} = \max\left(0, 1 - \frac{MSE_{ref}}{\frac{1}{n}\sum_{i=1}^{n}\left(y_{i,ref} - \bar{y}_{ref}\right)^2}\right) \quad (3.4)$$

where $\bar{y}_{ref}$ is the average all data annual average estimate across all n sites. Also see the brief discussion in the subsection "A Note on Model Performance Assessment" below.

MSE-based R^2 was used instead of (and sometimes reported along with) the more common regression-based R^2 based on the squared Pearson correlation coefficient because it evaluates whether predictions and observations are the same (i.e., are near the one-to-one line), rather than merely correlated. As such, it assesses both bias and variation around the one-to-one line. In contrast, regression-based R^2 solely assesses whether pairs of observations are linearly associated, regardless of whether observations are the same or not. MSE-based R^2 performs similarly to or worse than regression-based R^2.

Ultrafine Particle Data Used in Chapters 4, 5, 6, 8, and 9

As noted previously, our mobile campaign is unique in that it collected UFPs as PNCs using four different instruments: two P-Traks (one with a diffusion screen and one without), a NanoScan, and a DiSCmini. The NanoScan collected data on a 1-minute time scale; the remaining instruments recorded on a 1-second time scale. The particle size ranges for the instruments are 20–1,000 nm for the unscreened P-Trak, 36–1,000 nm for the screened P-Trak, 10–420 nm for the NanoScan, and 10–700 nm for the DiSCmini. As discussed by Blanco and colleagues (2022), "PNC measures from different instruments were strongly correlated with each other, and they produced broadly similar spatial surfaces, strengthening our confidence in the quality of our measurements. Differences in the reported PNC levels across instruments, however, can be attributed to multiple factors, including differences in technology, each technology's unique particle size detection efficacy, and built-in calibration (if present), all of which impact the reported particle size ranges and concentrations of each instrument."

With data from the mobile monitoring campaign, we focused our initial analyses on the P-Trak for the following reasons: this instrument has been a common choice in previous mobile monitoring campaigns, it produces high-quality

data at the 1-second time scale, and we found broadly similar spatial surfaces across UFP instruments. However, because the NanoScan has a lower size cut and thus captures the smallest particles that may be of most interest, we prioritized this instrument for our primary analyses of the stationary roadside data in Chapters 4 and 5. To assess comparability of findings, we also conducted sensitivity analyses using the P-Trak as well as the NanoScan using bins in the 10–100 nm size range; these demonstrated similar patterns of results as the primary analyses based on the NanoScan (Chapter 4's Additional Materials). We prioritized the unscreened P-Trak in all additional single-pollutant analyses of UFPs, including the variable importance analyses in Chapter 8, the value of information analyses in Chapter 9, and the on-road analyses in Chapter 6. Note that due to the NanoScan's longer 1-minute measurement duration, we were unable to use it for analyses of on-road mobile data. Finally, we used all UFP measurements, including individual bin size data from the NanoScan, in the multipollutant analyses reported in Chapter 8.

LOW-COST SENSOR AND RELATED REGULATORY MONITORING DATA

Parts of this section have been reprinted (adapted) with permission from (1) Bi et al. 2024, Copyright 2024 Elsevier, and (2) Zuidema et al. 2024, Copyright 2024 Nature.

This section covers the data used as well as the prediction modeling approach for leveraging low-cost sensor data in exposure models for application to epidemiological cohorts (i.e., the results described in Chapter 7). Our Aim 1 study questions considered how models changed with the inclusion of some or all of the low-cost sensor data collected under the auspices of ACT-AP (described below). The geographic region for our modeling is shown as the red and black participant residences on the map in Figure S3.1. For modeling, we also used a subset of the over 800 geographic predictors that were discussed above and summarized in Table S3.2.

The data for evaluation of the added value of low-cost sensor data relied on the following:

1. A combination of regulatory monitoring data from the Washington Department of Ecology and the Puget Sound Clean Air Agency (PSCAA), and other high-quality data from research studies conducted at the University of Washington (At times, we refer to all these sources as "agency" data.)
2. Calibrated 2-week low-cost sensor measurements from over 100 locations collected as part of the ACT-AP study (These were collected as part of the Remote Air Data campaign.)
3. The ACT-AP snapshot campaign that measured NO_2 using Ogawa samplers and was focused on roadway gradients at three distinct time periods

Agency Data Description and Distillation for Analysis

As noted in number 1 above, we used data from the Washington Department of Ecology, the PSCAA, and previous University of Washington research that includes a number of studies and independent monitoring campaigns. Agency data are regular (e.g., hourly in recent years), generally complete over long periods of time (e.g., 10+ years), and have relatively few missing data points. For $PM_{2.5}$, the reference-grade monitors included the Federal Reference Method (FRM) monitors (N of monitors = 19), Federal Equivalent Method (FEM) monitors (N = 7) using the tapered element oscillating microbalance and beta attenuation monitoring methods, and nephelometers (N = 27). While some ambient monitoring in the Puget Sound began as early as 1978 using nephelometers, most of the $PM_{2.5}$ analyses discussed in this report were restricted to 2010 onward.

Complete and continuous agency NO_2 data began in 1996, which we used as the starting year for modeling. The NO_2 data were measured by chemiluminescent FRM or FEM instruments (e.g., Teledyne API, San Diego, CA; model 200 EU) with a limit of detection of 1 ppb and a data quality objective that requires that the bias and percentage coefficient of variation be within (±) 15% (PSCAA 2020). Despite these positive attributes, there was a limited number of agency locations measuring NO_2 in the Puget Sound (N = 11). Of these 11 monitors, three met our criteria for inclusion in the model's calculation of long-term time trends (monitors that were not seasonal, operated for at least 6 years, and with at least 100 2-week average measurements; further details below). This low number of NO_2 monitors in the Puget Sound is a situation not uncommon in many US metropolitan areas.

Supplementary monitoring data that were also treated as agency data in the modeling included: the Diesel Exhaust Exposure in the Duwamish Study (DEEDS) (Schulte et al. 2015), the Panel Study on Income Dynamics (PSID) (Liu et al. 2003), the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) Pilot (MAP) (Wilton et al. 2010), and the Yesler Terrace study (Yesler) (Wong, 2010). Details of these supplementary monitoring campaigns are provided in Chapter 7's Additional Materials. The NO_2 analyses used all these sources, while the $PM_{2.5}$ analysis only used the DEEDS data.

We averaged all data to the 2-week timescale, using the same inclusion criteria as considered in our previous research (Keller et al. 2015). For NO_2, we omitted one outlying 2-week agency average equal to 75 ppb; the next highest 2-week NO_2 agency measurement was 33 ppb.

ACT-AP Snapshot Campaign

The ACT-AP snapshot campaign was primarily a roadway gradient monitoring campaign. NO_2 was measured with Ogawa passive samplers (Ogawa & Co., USA, Inc., Pompano Beach, FL) deployed on telephone poles. Ogawa passive samplers are convenient measurement devices with a high level

of performance; for example, one study observed an absolute difference of 1.2 ppb NO_2 and R^2 = 0.95 compared to FRMs (Sather et al. 2007). For the ACT-AP snapshot, samples were taken in the Puget Sound at 110 locations in 17 clusters around major roads, most of which consisted of six sensors varying from <50 meters up to 350 meters away from the major road (with a mean ± SD = 4.2 ± 2.5 sensors per cluster). Sampling was conducted during three seasons with samplers deployed for 2 weeks in April 2018, August 2018, and February 2019. The ACT-AP snapshot measurements were an important data input into the NO_2 spatiotemporal model because of the limited number of agency locations available to contribute geographic information at locations with NO_2 measurements.

Low-Cost Sensor Campaign: Remote Air Data Collection, Quality Control, and Distillation for Analysis

We conducted supplemental sampling with low-cost sensors between April 2017 and September 2020 with 25 low-cost sensor instruments. 20 were deployed at a rotating set of over 100 community locations for two seasons each, and five sampled continuously at co-located regulatory monitoring sites (Note: for most of the analyses in Chapter 7, only the 20 rotating instruments were evaluated; the focus of the remaining five co-located monitors was quality control and calibration, rather than prediction at cohort locations). The monitoring locations were primarily ACT participant volunteer homes selected to represent the distribution of ACT participant locations in both geographic and covariate space (specifically population density, elevation, and proximity to major roads). The low-cost sensors had sensors for $PM_{2.5}$, nitric oxide (NO), NO_2, ozone (O_3), and carbon monoxide (CO), and were built by the Seto lab at the University of Washington; an example low-cost sensor is shown in **Figure 3.2**. Each sensor box had two pairs of $PM_{2.5}$ sensors (Plantower PMS A003 and Shinyei PPD42NS), plus a set of Alphasense B4 gas sensors to monitor NO, NO_2, O_3, and CO (specifically NO-B4, NO_2-B43F, OX-B431, CO-B4), as well as temperature and relative humidity sensors. The gas sensors rely on a chemical reaction that produces an electrical current; this is linearly proportional to the fractional volume (i.e., it measures parts per volume, rather than mass). Data were transmitted wirelessly every 5 minutes to a secure server.

During the low-cost sensor campaign, routine quality control was carried out by scrutiny of an automated weekly report that flagged problems with data completeness, outside range values, and overall variability; this report also considered correlation of the duplicate low-cost sensors and nearby network monitors. Broken sensors were replaced when identified by this routine screening.

After quality control screening, low-cost sensors were calibrated on the daily time scale with regression models developed from co-located low-cost sensors and agency monitors. The low-cost sensors had over 7,500 monitor-days of co-location with reference-grade monitors during the deployment period, and the results of these calibrations are

Figure 3.2. Low-cost sensor developed by the Seto lab and used in the low-cost sensor campaign.

described in detail elsewhere (Zuidema et al. 2021; Zusman et al. 2020). For NO_2, the calibrations had CV RMSE = 3 ppb, and CV-R^2 = 0.79 (Zuidema et al. 2021). The low-cost sensor calibration performance for $PM_{2.5}$ varied slightly depending upon the cross-validation approach and regulatory monitors included (i.e., FRM only or FRM plus FEM; see Table 4 of Zusman et al. 2020). Due to data quality concerns, only the Plantower measurements were used and reported. The CV RMSE ranged between 1.02 and 2.13 μg/m^3, depending on the cross-validation approach and regulatory monitors included, while the CV MSE R^2 ranged between 0.86 and 0.97 (Zusman et al. 2020).

As with the agency data, we averaged the low-cost sensor data to the 2-week time scale, using the same inclusion criteria as considered in our previous research (Keller et al. 2015). The $PM_{2.5}$ and NO_2 analyses based on the low-cost sensor data used slightly different subsets of data. The $PM_{2.5}$ analysis used 112 locations, none of which were co-located with agency sites, while the NO_2 analysis used 117 locations, including five co-located with $PM_{2.5}$ agency sites. Most of the locations were ACT-AP participant and volunteer homes (N = 99); there were an additional 12 community sites. **Figure 3.3** shows the study region for the $PM_{2.5}$ analyses. A related map for the NO_2 modeling is shown in Figure S7.4. Summary statistics for the data included in the analysis are provided in Tables S7.1 and S7.2 for $PM_{2.5}$ for the 2010–2020 and 1978–2021 time periods, respectively, and Table S7.5 for NO_2 for the 1996–2020 time period. Data density figures are shown in Figures S7.1 and S7.2 for $PM_{2.5}$, and Figure S7.3 for NO_2.

Spatiotemporal Exposure Modeling Approach and Model Selection

For all the exposure modeling based on these data, we used our published spatiotemporal model, which is designed to accommodate imbalanced data, with some locations providing long time series while other locations provide

only a few measurements (Lindström et al. 2014; Szpiro et al. 2010). The long-term measurements are used to establish the smoothed temporal trends of pollution, which are known as *time basis functions*. The short-term measurements at locations with high monitoring density help to determine the spatial coefficients necessary to create spatially varying linear combinations of the smoothed temporal trends. The framework can be expressed as

$$y(s,t) = \mu(s,t) + v(s,t) \quad (3.5)$$

where $y(s,t)$ denotes the log-transformed 2-week average $PM_{2.5}$ or NO_2 concentration at location s and time t; $\mu(s,t)$ denotes the spatiotemporal mean surface; and $v(s,t)$ denotes the spatiotemporal residual variation. The spatiotemporal mean surface can be broken down as

$$\mu(s,t) = \beta_0(s) + \sum_{i=1}^{I} \beta_i(s) f_i(t) \quad (3.6)$$

where $\beta_0(s)$ denotes the long-term mean (intercept) at location s, and $f_i(t)$ denotes smoothed temporal trends (i.e., time-basis functions) derived by singular-value decomposition (Guttorp et al. 2007; Sampson et al. 2011). $\beta_i(s)$ denotes spatially varying coefficients for the i^{th} temporal trend based on universal kriging (UK):

$$\beta_i(s) \sim N[\boldsymbol{X}_i(s)\alpha_i, \sum_i(\phi_i, \sigma_i, \tau_i)],\ i = 0, 1, \ldots, I \quad (3.7)$$

where $\boldsymbol{X}_i(s)$ denotes reduced-dimension summaries of geographic covariates by partial least squares (PLS) (Abdi, 2010) at location s, and α_i denotes UK coefficients to be estimated. $\sum_i$, the covariance structure for $\beta_i(s)$, is a spatial smoothing model with covariance functions parameterized by a range ϕ_i, partial sill σ_i, and nugget τ_i. I is the number of singular-value decomposition-derived smoothed temporal trends. Finally, $v(s,t)$, the spatiotemporal residual is assumed to have a mean of zero and a spatial correlation structure that is independent in time with covariance structure Σ_t:

$$v(s,t) \sim N[0, \sum_t(\phi_t, \sigma_t, \tau_{tj}, \epsilon_j)],\ t = 1, 2, \ldots, T,\ j = monitor\ type \quad (3.8)$$

Both the $PM_{2.5}$ and NO_2 models accommodated indexing by monitor type to distinguish between the agency and the low-cost sensors. The spatiotemporal modeling and prediction procedures were conducted using the R (ver. 3.6.2) package "Spatiotemporal" (ver. 1.1.7) (Lindstrom et al. 2023).

To derive the PLS scores, we used different data sources for the $PM_{2.5}$ and NO_2 models due to the small number of long-term sites available in the NO_2 model (three sites; see

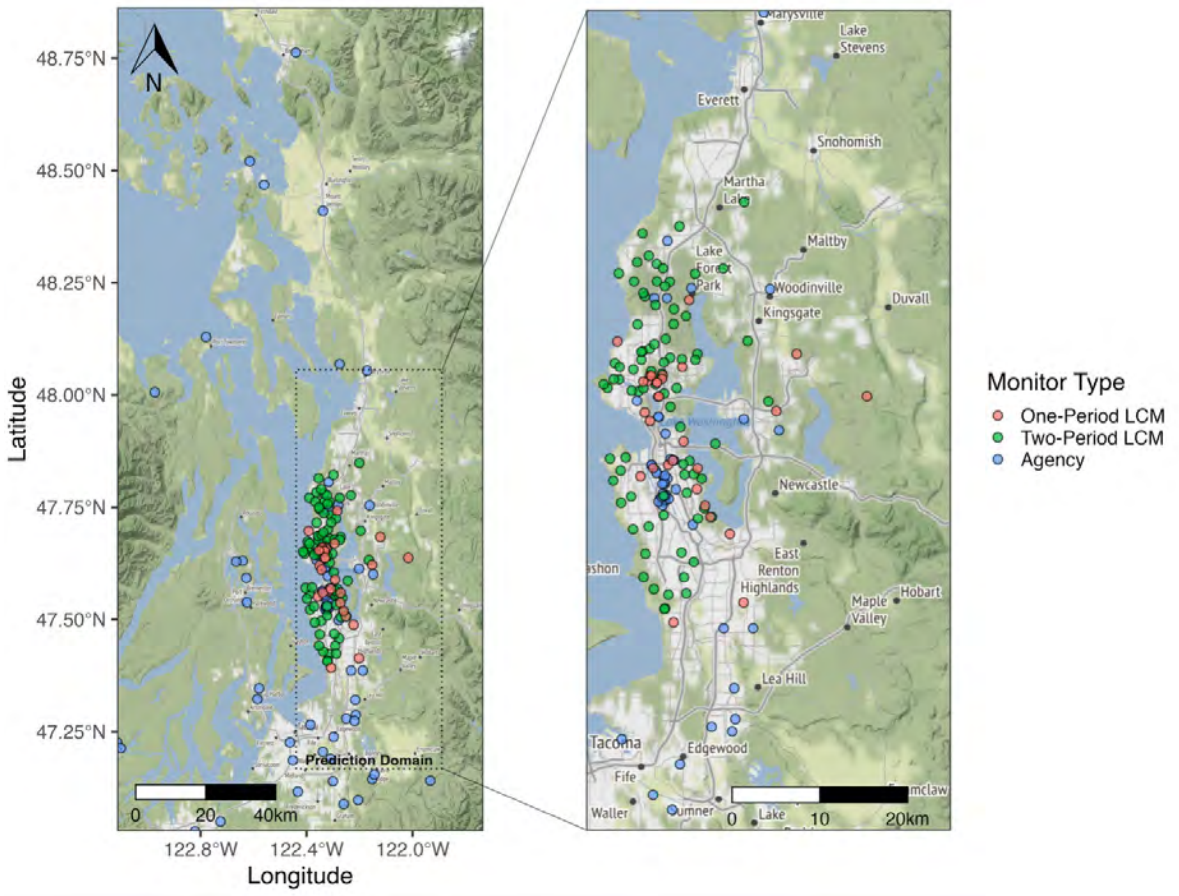


Figure 3.3. The $PM_{2.5}$ modeling study region with locations of $PM_{2.5}$ sensors. The dashed box shows the prediction domain for model assessment in the $PM_{2.5}$ analyses. This is the Puget Sound region of Washington state. The right inset map shows the greater Seattle area prediction domain. Reprinted with permission from Bi et al. 2024. LCS = low-cost sensor.

Table S7.5). The $PM_{2.5}$ model uses our research team's typical approach (e.g., Keller et al. 2015) to develop scores from "outcomes", which were regression coefficient estimates obtained at each of the long-term site locations from regressing the location-specific 2-week average $PM_{2.5}$ time series against the smoothed time trends. For NO_2, we averaged the three seasonal NO_2 measurements at each of the 110 snapshot locations, creating temporally averaged measurements. Next, we conducted PLS regression at each of the ACT-AP snapshot campaign locations with the average NO_2 concentration as the outcome and the vector of geographic covariates as the predictors. PLS scores were then calculated at all other locations (e.g., MAP, PSID, and agency locations) using the geographic covariates at those locations and the PLS scores defined by the regression with the ACT-AP snapshot averages. See Bi and colleagues (2024) and Zuidema and colleagues (2024) for additional details on these approaches; note that for the 1978–2021 time period data, the spatiotemporal model was based on detrended data (Bi et al. 2024, supplemental information).

We employed cross-validation (CV) to identify the optimal model parameter values and number of terms, aiming to create a parsimonious model that displayed satisfactory prediction performance (the description of the CV performance assessment follows in the next subsection). In particular, for $PM_{2.5}$, we established $f_i(t)$ as the single temporal trend from singular-value decomposition (I = 1; over two or three temporal trends). The temporal trend was generated based on reference-grade monitors with over ten 2-week measurements during the modeling period from January 2010 to September 2020. For the UK process related to β_i, the first two principal components from PLS were utilized (over one or three principal components). The covariance functions of $\sum i$ were represented by exponential functions with ranges, partial sills, and nuggets (over an independent covariance structure). The covariance function of $v(s,t)$ was fitted to an exponential structure with a range, partial sill, nugget, and random effect. This random effect, ϵ_j, is a constant variance added to the covariance Σ_t, which can be interpreted as a partial sill with infinite range. The random effect enabled the model to account for temporally uncorrelated deviations from the smoothed time trends. For NO_2, the final model had one time trend, $f_i(t)$, three PLS components to summarize the geographic covariates for the time trend, exponential covariance structure for β_0 (long-term temporal mean concentration), independent covariance structure for β_i (coefficients for spatially varying time trend), and exponential covariance structure for $v(s,t)$ (spatiotemporal residual variation). For this final model, in addition to $v(s,t)$'s nugget with random effect, we considered varying the nugget by measurement type (agency FRM, supplemental data collected with Ogawa passive samplers, and low-cost sensors), but this resulted in a negligible change in performance. Additional details about the model selection are in the published papers (Bi et al. 2024; Zuidema et al. 2024).

Model Performance Assessment

We carried out cross-validation procedures on various observations and predictions, summarizing the model performance with CV statistics *RMSE*, MSE R^2, and R^2_{reg}. For CV *RMSE* and CV MSE R^2, abs_i is the observed concentration (in ppb for NO_2 and μg/m³ for $PM_{2.5}$) and $pred_i$ is the predicted concentrations for the observed time points at location i (Bergen et al. 2013):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i)^2} \tag{3.9}$$

$$R^2 = \max\left(0, 1 - \frac{RMSE^2}{Var(obs)}\right) \tag{3.10}$$

Note that *Var(obs)* is scaled by n rather than the usual n-1.

Compared to the traditional regression-based coefficient of determination (R^2_{reg}), which describes the goodness of fit to the regression line, CV MSE R^2 represents a measure of fit to the 1:1 line of the observed versus predicted concentrations. The CV MSE R^2 therefore describes the accuracy of the predictions — the value of interest we are most interested in characterizing — and is often lower than CV–R^2_{reg} (Keller et al. 2015). The correlation-based R^2, which assesses fit to the regression line between measurements and predictions, has been widely used in air pollution model evaluation (Diao et al. 2019; Sorek-Hamer et al. 2020). This correlation-based metric, however, inflates the R^2 value when the regression line deviates from the 1:1 line. In comparison, the MSE-based R^2 fits the 1:1 line and is a more trustworthy indicator of a model's prediction accuracy because it takes both systematic bias and uncertainty into account.

In cross-validation, the training dataset comprises all the data except for the data included in a single cross-validation group. The model is refitted on the training dataset, and predictions are obtained for the observations in the cross-validation group omitted from the training dataset. The procedure is repeated until predictions are obtained for all observations across all cross-validation groups. Because of the complexity of the spatiotemporal model and the variety of ways that cross-validations could be done, most cross-validation analyses we report refit spatiotemporal models, which include all the data of certain types (e.g., the "agency" data), and only cross-validated subsets of a specific type (e.g., low-cost sensor data at residential locations or Ogawa snapshot locations). Cross-validation is a type of external validation procedure because predictions are evaluated on data that were not included in the training dataset. In some of the external validations described in Chapter 7 and the Additional Materials for Chapter 9, we use a combination of cross-validation and pure external validation because some (or all) of the data we are validating were left out of the model under consideration. For additional details on the combined cross and pure external validation approach, see the $PM_{2.5}$ methods description in

Chapter 7. For additional discussion of various approaches to cross-validating this type of spatiotemporal model, see Lindstrom and colleagues (2014).

We predicted 2-week concentrations on the log scale and then exponentiated these predictions to back-transform them to the original concentration scale. We summarized the cross-validated models at the 2-week (or *spatiotemporal validation*), 1-year (NO_2 only), and whole modeling period (or *spatial validation*) timescales for various data subsets. For NO_2, the spatial validation represents site averages for the 1996–2020 period, whereas for our primary report of $PM_{2.5}$ results, it represents 2010–2020. The NO_2 ACT-snapshot and spatiotemporal low-cost sensor cross-validations were summarized at the 2-week timescale.

In the Chapter 7 methods sections, we report the pollutant-specific consideration of the added value of the low-cost sensors, as well as performance summary statistics. We report some additional analyses of the added value of NO_2 low-cost sensor data in Chapter 9's Additional Materials.

Health Association Reporting

For estimating the health associations, we used 5-year average predictions from the spatiotemporal model, either for the entire modeling period (spatiotemporal) or restricted to 2019 as purely spatial contrasts. These spatial predictions are from the same spatiotemporal models, just over the restricted 2019 time frame to make them more comparable to the Seattle mobile monitoring campaign, which was largely conducted in 2019. Note that while the spatiotemporal model estimates reflect time-varying air pollution and the spatial estimates do not, both take into account time-varying address history over the 5 years prior to study enrollment.

A NOTE ABOUT THE OBSERVATIONS USED IN MODEL PERFORMANCE ASSESSMENT

As described above in two different subsections, we use CV RMSE and CV MSE R^2 to characterize model performance from the spatial models used for the mobile monitoring campaign and the spatiotemporal models used with the low-cost sensor data. As an aside, for some performance evaluations of the spatiotemporal models in Chapter 7, we report pure out-of-sample statistics or a combined performance using both CV and external data. Broadly speaking, mean squared error is the average squared difference between observations and predictions. Typically, the observations come from the same dataset that is used to generate the predictions. This is the approach used for evaluating the low-cost sensor data in Chapter 7, with the caveat that some of the evaluations are pure out-of-sample. In contrast, when evaluating performance statistics for the mobile data in this report, we use the complete all data observations as the reference observations, even when predictions are developed from a different subset of the mobile monitoring data. This is done to facilitate "apples to apples" comparisons between mobile monitoring designs, and because it provides a more objective assessment of the performance of interest. We and others have shown the importance of validating model predictions against unbiased observations, and how comparisons against biased, unstable campaign measurements (e.g., from restricted sampling designs) produce noisy and misleading conclusions (Blanco et al. 2023a, 2023b; Kerckhoffs et al. 2016; Messier et al. 2018). See Blanco and colleagues (2023b) for further details.

ASSOCIATION BETWEEN COGNITIVE FUNCTION AND AIR POLLUTION IN THE ADULT CHANGES OF THOUGHT COHORT

For this report, we are focusing on the CASI-IRT outcome at baseline in the entire cohort of individuals recruited into the ACT study between baseline and March 2020. The following sections address the ACT study design, the CASI-IRT cognitive function outcome, ACT cohort characteristics including a summary of the data used in this report, a description of the inferential analyses that will be presented later in the report, and health association results from the all data exposure models presented in later chapters of this report.

This study was approved by the University of Washington Institutional Review Board with IRB ID STUDY00009108.

ACT Study Design

We obtained the data for the health effects inference (Aim 3) from the ACT study, under the previously funded ACT-AP study. The ACT cohort is a population-based cohort of urban and suburban elderly individuals drawn from Group Health Cooperative (now Kaiser Permanente), a health maintenance organization (HMO) administered by the Kaiser Permanente Washington Health Research Institute. Participants were at least 65 years of age at the time of recruitment and, on average, had been part of the HMO for two decades prior to enrollment. Recruitment began in 1994, with 2,581 individuals enrolled between 1994 and 1996. A second recruitment wave was conducted between 2000 and 2002 to expand the cohort, adding 811 individuals. Then, beginning in 2005, there has been continuous enrollment in ACT to maintain over 2,000 active at-risk person-years in each calendar year (Kulick et al. 2020).

ACT study visits are scheduled every 2 years; in-person measures include the Cognitive Abilities Screening Instrument (CASI) for cognitive function (further details below), clinical measures such as blood pressure, and questionnaire data for additional risk factors, clinical, and demographic information, including residential address. ACT has an exemplary Completeness of Follow-up Index (95.6%), and it has over 20 years of cognitive function data on many participants (Clark et al. 2002; Kukull et al. 2002).

ACT participants are typically stably enrolled in the HMO, with more than half of the ACT participants completing at least 23 years of HMO membership preceding ACT enrollment.

Residence history information was available from ACT records as well as HMO billing records starting in 1989. To create the initial address histories of participants, we retrieved 28,385 addresses from historic HMO billing records going back to 1989, 7,933 addresses from ACT study records going back to 2007, and 651 records from LexisNexis for the 20 most recent addresses for participants who have died or developed dementia. We developed an address history protocol that provided guidance on how best to organize the address history maintained for the project: which sources were used to collect data (e.g., Kaiser Permanente historical billing records, LexisNexis, or telephone interviews with ACT participants) and how frequently the data were updated; how to prioritize conflicting address data; a summary of information linked to geocoding; and procedures for updating addresses over time. We geocoded addresses using ArcGIS Business Analyst. Due to the nature of billing records, healthcare enrollment in the United States, and collating of multiple sources, not all participants had continuous address data over time. Many of these participants lived at the same location before and after the gap in address records, and some merely had short gaps in records. We addressed these gaps with some realistic simplifying assumptions, which are documented in the address history protocol.

Participants are prospectively followed with routine biennial visits until dementia incidence, drop-out, or death. Extensive health, lifestyle, biological, and demographic data are also collected.

Cognitive Function Outcome Measure

Cognitive function is measured using the CASI score, which ranges from 0 to 100, with scores below 86 triggering a standardized cognitive evaluation in ACT with examination by a study physician and neuropsychological tests. CASI is a 40-item global cognitive test that assesses a broad range of cognitive domains (Teng et al. 1998). These include attention, concentration, orientation, short-term memory, long-term memory, language abilities, visual construction, list-generating fluency, abstraction, and judgement (Li et al. 2017). The CASI inherently has relatively few difficult items and many easy items, resulting in curvilinear scoring, particularly when participants start with different baseline cognitive abilities (Crane et al. 2016). However, item difficulties are not considered in standard scoring (Teng et al. 1998). Thus, the final cognition score and outcome measure for all health association estimates in this report are derived using Item Response Theory (CASI-IRT) to improve score accuracy, measure cognitive change with less bias, and to account for missing test items (Crane et al. 2008; Ehlenbach et al. 2010; Li et al. 2017). IRT scoring has been found to diminish the impact of differential item functioning, which occurs when different groups of individuals have different probabilities of correctly answering an item, even after controlling for overall ability level. IRT scores also allow for greater sensitivity at higher levels of cognitive function, which is especially important for the detection of early cognitive deficits (Crane et al. 2016; Kulick et al. 2020). ACT uses Parscale to generate modern psychometric scores from item-level CASI data (Gibbons, 2015). CASI-IRT scores are normalized, with values less than 0 indicating lower cognitive function and scores above 0 indicating greater cognitive function than average.

In this report, we consider the CASI-IRT measurement only when obtained at baseline. We chose this outcome because previous literature has shown a link between air pollution exposure and cognitive function (Delgado-Saborit et al. 2021; Peters et al. 2019). Furthermore, by considering a cross-sectional study design with a continuous outcome measurement, the analyses we report can focus on the role of exposure assessment study design and exposure measurement error in health association estimates, without the added complexity that would accompany analysis of a longitudinal study design or a binary or survival outcome.

ACT Cohort Characteristics

As of March 2020, the original cohort consisted of 5,763 participants. All analyses were restricted to baseline and included 5,409 (94%) participants with both a valid CASI score and who had lived in the mobile monitoring region during at least 95% of the prior 5 years (Figure S3.3 shows the inclusion flow chart). ACT participants have excellent residential histories and air pollution coverage (Blanco et al. 2022; Shaffer et al. 2021a). On average, this analytic cohort lived in the monitoring area >99% of the time, had exact residential addresses 98% of the time, and had imputed addresses 5% of the time (i.e., from residential gaps). **Table 3.1** describes the baseline analytic cohort characteristics. Participants were on average (SD) 74 (6) years old, slightly more likely to be female, about half had at least a college education, and had an average (SD) CASI-IRT score of 0.33 (0.71).

Table 3.1. ACT Cohort Characteristics at Baseline

	Overall (*N* = 5,409)
Visit Age (Years)	
Mean (SD)	74 (6)
Median [Min, Max]	73 [65, 101]
Sex	
Male	2,259 (42%)
Female	3,150 (58%)
Degree	
None	422 (8%)
GED/High School	1,997 (37%)
Bachelor's	1,274 (24%)
Master's	872 (16%)
Doctorate	330 (6%)
Other	514 (10%)
Cognitive function CASI-IRT	
Mean (SD)	0.34 (0.71)
Median (Min, Max)	0.37 (−2.12, 1.75)

Inferential Analyses of Cognitive Function and Air Pollution We assessed the association between air pollution and baseline cognitive function in ACT using multiple linear regression. We used data for air pollutant exposures (indicated as X_j in **Equation 3.11**), including UFPs (measured as PNC; from various exposure models detailed in later chapters), $PM_{2.5}$, and NO_2. Each model was adjusted for age, calendar year (categorical 2-year bins), sex, and educational attainment (categorical) (indicated as W_j in Equation 3.11) to make confounder model 1. Calendar year was included in the model, as in our prior work (Shaffer et al. 2021a), to account for trends in both air pollution and dementia over time. Chapter 4 (stationary roadside mobile monitoring) and Chapter 6 (on-road mobile monitoring) report an additional confounder model 2 that is further adjusted for race (White, People of Color) and socioeconomic status (SES) based on the Neighborhood Disadvantage Index at a participant's longest-lived address at or prior to baseline. The Neighborhood Disadvantage Index is a validated indicator composed of Census tract-level variables from the American Community Survey (Miles et al. 2016). These underlying models can be written as:

$$Y_j = \alpha + \beta X_j + \delta W_j + e_j \quad (3.11)$$

where the index j represents each participant. Y_j is the measured CASI-IRT score at baseline; α is the intercept; X_j is the mean pollutant concentration at baseline; W_j refers to the vector of cross-sectional covariates measured at baseline; and e_j is residual error. The model quantifies the relationship between exposure and cognitive function; β is our primary parameter of interest, and the cross-sectional association between cognitive function and air pollutant exposures before the baseline exam. Because our focus was on comparison of exposures given a single model for a health outcome, we did not address additional considerations of the epidemiological inferential model, although Chapters 4 and 6, which address the design of stationary roadside and on-road mobile monitoring respectively, include analyses that adjust for additional potential confounding using confounder model 2 described above. We discuss this topic further in Chapter 10.

We compared the health parameter (β) estimated from reference all data exposure prediction models with the parameters estimated from alternative exposure models (e.g., reduced mobile monitoring sampling).

The estimated associations between air pollution and CASI-IRT from confounder model 1 with various "all data" pollutant predictions are summarized in **Table 3.2**. Air pollution associations are contextualized in terms of month-equivalents based on each model's estimated age association. In other words, the health associations are expressed as the equivalent of aging a certain number of months. For example, the first row shows that exposure to an IQR increase in UFP concentration is equivalent to aging by 7.5 months; this means that, adjusted for potential confounding, individuals with higher UFP exposure (per 1,900 pt/cm^3 increment) performed similarly to peers who were 7.5 months older but had lower exposures.

Figure S3.1 shows a map of Washington State with ACT participant residential locations marked in red, black, or gray. The red represents residence locations within the mobile monitoring region, the black represents those within the spatiotemporal modeling region, and the gray represents those outside either modeling region. All inferential analyses (Chapters 4, 5, 6, 7, 8) were restricted to participants residing within the mobile monitoring region; the locations are shown in red.

Given the temporal misalignment of the cognitive measurements in ACT (1994+) and the available mobile monitoring air pollution measurements (2019–2020 vs. the true exposure period of interest), in the analyses with the mobile data and with the purely spatial version of the spatiotemporal model, we assume that air pollution has remained constant over time such that recent exposures are surrogates for longer-term exposures (Blanco, 2021; Kim et al. 2017; Levy et al. 2015; Meng et al. 2019; Molter et al. 2010; Y. Wang et al. 2011). We are unable to evaluate the accuracy of this assumption given the lack of historical UFP data. Nonetheless, the goal of these analyses was to estimate the potential *magnitude* of bias and variability when applying restricted sampling campaigns relative to intentionally designed, more spatially and temporally complete campaigns, and not necessarily to estimate causal associations between UFPs and cognitive function. Any remaining potential confounding unaccounted for in our analyses was constant across monitoring designs, thus allowing us to capture differences driven by changes in monitoring design. The exact degree of bias will depend on the true underlying associations between an exposure and the outcome of interest and on the exposure monitoring design.

Using real-world data is a strength of these analyses as they provide a direct connection to the real-world implications of our exposure assessment study design findings. Linking real cohort data with air pollution exposures, for example, captures the subtle effects of air pollution in complex environments where confounding and interactions may exist — aspects that can be challenging to realistically capture in a simulation study.

Table 3.2. Estimated Associations Between Air Pollutants and Baseline Cognitive Function Based on Reference Exposure Models and Adjusted for Age, Calendar Year, Sex, and Education (Confounder Model 1) (*N* = 5,409)

Chapter	Model	Pollutant	Air Pollution Association per IQRb (95% CI)	Age Association (years)	Air Pollution Aging Equivalent (months)
4–5	All-data roadside, NanoScan	UFP	–0.020 (–0.036, –0.004)	–0.032	7.5 (1.5, 13.6)
4 Sensitivity Analyses, 6, 8	All-data roadside, P-Trak	UFP	–0.021 (–0.039, –0.003)	–0.032	7.9 (1.1, 14.8)
7	Full Spatial model with low-cost sensors	$PM_{2.5}$	–0.023 (–0.075, 0.028)	–0.032	8.8 (–10.7, 28.2)
7	Full Spatiotemporal model with low-cost sensors	$PM_{2.5}$	–0.009 (–0.036, 0.17)	–0.032	3.5 (–6.5, 13.5)
7	Full Spatial model with low-cost sensors	NO_2	–0.030 (–0.057, –0.002)	–0.032	11.3 (0.86, 21.8)
7	Full Spatiotemporal model with low-cost sensors	NO_2	0.001 (–0.028, 0.030)	–0.031	–0.4 (–11.7, 10.8)

[a] Also shown are the age association with cognitive function in years and the air pollution aging equivalent in months. The air pollution aging equivalent associations are scaled relative to the association between cognitive function and age to give context to the air pollution estimates.
[b] IQR increment for UFP is 1,900 pt/cm^3, for $PM_{2.5}$ is 1 μg/m^3, and for NO_2 is 3 ppb.

CHAPTER 4: USING STATIONARY DATA FROM MOBILE MONITORING STUDIES: EXPOSURE ASSESSMENT DESIGN AND HEALTH INFERENCE*

Lead Authors: Magali Blanco, Lianne Sheppard

INTRODUCTION

Short-term mobile monitoring campaigns, which are the collection of repeated short-term air samples at selected sites, are increasingly being used as an efficient and cost-effective approach for estimating multilocation long-term air pollution averages (Kim et al. 2023). While many strategies have been employed, however, little is known about how monitoring network design features, including the number of stops and sampling temporality, impact exposure assessment models. Most campaigns, for example, collect limited data at each site consisting of approximately 1–5 repeat visits, sample during restricted time periods such as business hours, and/or have short monitoring durations lasting under a few months. The resulting exposure prediction models generally have moderate or poor coefficients of determination (R^2), suggesting limited prediction accuracy and potentially less value for epidemiological inference.

A few studies have gained valuable insight regarding the impact of the number of sites and repeat site visits on the quality of the resulting exposure prediction models (Hatzopoulou et al. 2017; Messier et al. 2018; Saha et al. 2019). However, these findings have been largely focused on nonstationary on-road measurement campaigns in relatively smaller geographic areas and have not demonstrated whether exposure data from nonstationary, on-road campaigns are representative of residential human exposure levels (Alexeeff et al. 2018; Kerckhoffs et al. 2016). Furthermore, cohort study applications often span large geographic areas and are interested in long-term exposures. Little is known about how short-term, temporally restricted sampling campaigns impact long-term exposure prediction models or subsequent epidemiological inferences.

We previously investigated the impact of sampling design using year-round measurements for oxides of nitrogen (NO_x) from 69 fixed monitoring sites in California (Blanco et al. 2023b). In that work, we collected 28 random 1-hour samples following various sampling designs (30 times each), and used the resulting data to develop annual average exposure prediction models using PLS. Sampling designs included collecting 1-hour samples in a year-round balanced design during all seasons, all days, and all or most (5 a.m. to 11 p.m.) hours of the day; and collecting samples during weekday rush hours (7–10 a.m. and 3–6 p.m.) or business hours (9 a.m. to 5 p.m.) during two seasons (e.g., winter and summer, with a 2-week period per season). We evaluated cross-validated exposure predictions to the true long-term averages at each site, which are unknown in true mobile monitoring campaigns. We found that temporally balanced designs that sample during most hours generally produced unbiased annual averages. Note that some late-night hours were excluded for driver safety and logistical reasons. Common sampling designs like weekday business and rush hours regularly produced more biased annual averages. Importantly, we found that traditional model assessment approaches that compare model predictions to observations collected during the sampling campaign do not clearly reveal the inferior performance of rush hours and business hours designs (compared to the balanced design), and at times, they even appeared to perform better. This evaluation approach does not consider that "reference" observations from restricted sampling designs may themselves be biased.

In the current study, we investigated the impact of monitoring design in subsequent work (Blanco et al. 2023a), leveraging the extensive, multipollutant Seattle mobile monitoring campaign described in Chapter 3. We used Monte Carlo sampling to collect subsamples and investigate how the mobile monitoring design features, including the number of stops (sites × visits) and sampling temporality (seasons, days, hours), impacted the resulting UK-PLS exposure prediction models. Predictions from the all data campaign performed well, with CV MSE R^2s of 0.51–0.77. We found similar model performances (85% of the all data campaign CV MSE R^2) with ~1,000–3,000 randomly selected stops for NO_2, UFPs, and BC, and ~4,000–5,000 stops for $PM_{2.5}$ and CO_2 (**Figure 4.1**). Exposure model performances continued to improve as more total stops were achieved, although at slower rates. It did not make a meaningful difference whether the total stop count was achieved by visiting more sites the same number of times (increased spatial coverage) or the same number of sites a greater number of times (increased temporal coverage), so long as total stops increased (i.e., on average, both spatial and temporal coverage can improve exposure assessment in limited sample settings). Moreover, as with our prior work in California (Blanco et al. 2023b), this investigation showed that repeated, short-term sampling campaigns with additional temporal restrictions (e.g., business hours, rush hours, weekdays, or fewer seasons) had reduced exposure model performances when compared to temporally unrestricted campaigns with the same number of total stops. These campaigns produced different spatial surfaces, leading to some areas being substantially over- or under-predicted (**Figure 4.2**).

Chapter 4 has two objectives. The first is to investigate the degree to which a stationary roadside monitoring design impacts exposure assessment models. The second is assessing the impact of these monitoring designs on subsequent epidemiological inferences. We conduct a case study of UFP exposures and late-life cognitive function, an area of growing interest (Brugge and Fuller 2020; HEI 2013; US EPA 2019),

*Some of this work has been reprinted (adapted) with permission from Blanco et al. 2023a. Copyright 2023 American Chemical Society.

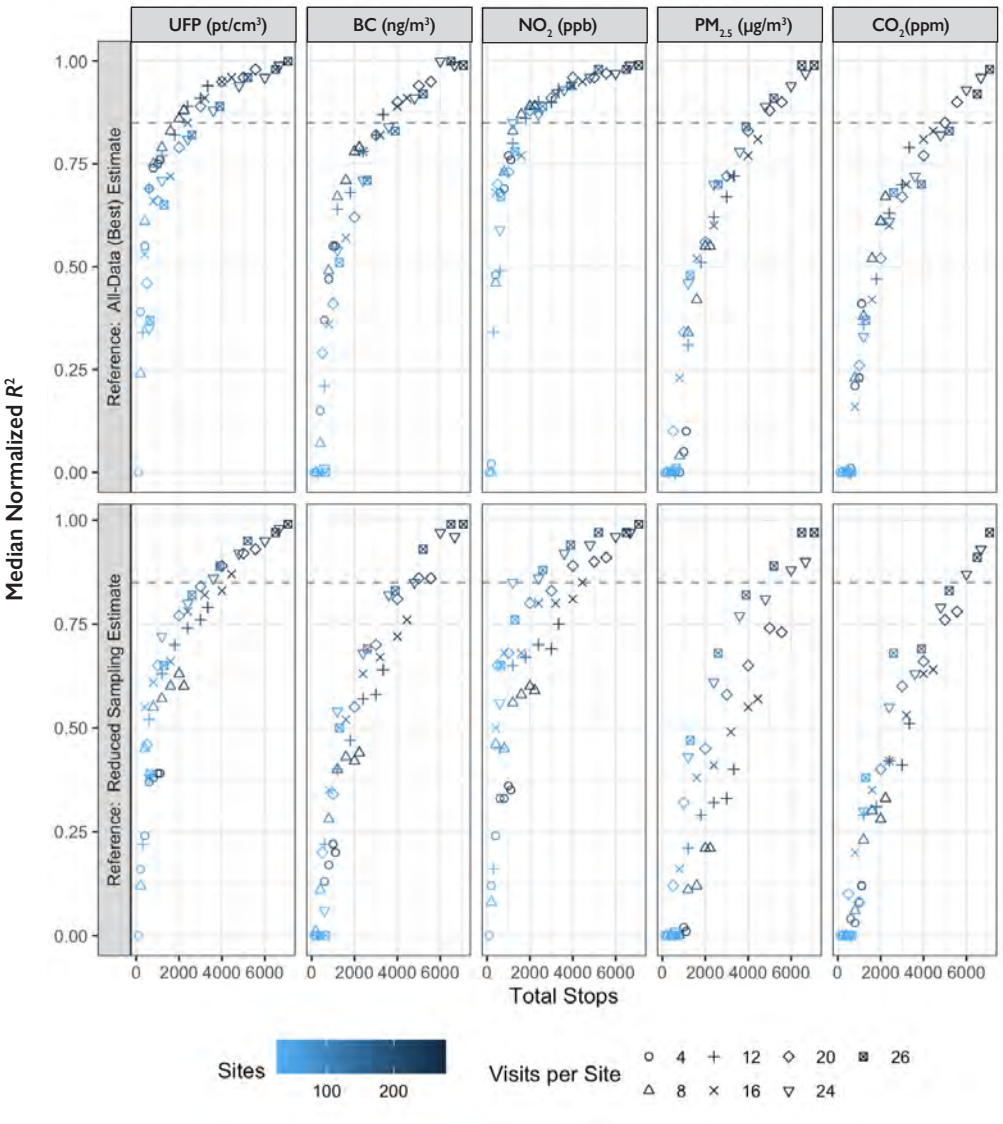


Figure 4.1. Median normalized MSE-based R^2 for fewer total stops designs using the stationary roadside data from the Seattle mobile monitoring campaign. MSE R^2 is calculated by comparing cross-validated predictions to annual average estimate references from either the all data campaign (top) or the reduced sampling campaigns (bottom) and normalized (divided by) the R^2 from the all data campaign. Normalized R^2 values below one indicate worse performance than the all data campaign. Median performance parameters are each based on 30 campaigns. The dashed line is at .85. Reprinted with permission from Blanco et al. 2023a. Copyright 2023 American Chemical Society (ACS).

by leveraging the Seattle mobile monitoring campaign and the ACT cohort (both described in Chapter 3). The monitoring campaign was specifically designed to assess unbiased annual average UFP exposures for the ACT cohort study of the aging brain (Kukull et al. 2002). We follow common mobile monitoring designs to sample the Seattle mobile monitoring campaign dataset, develop design-specific UFP exposure assessment models, use these models to assess participant exposures, and run health analyses of the estimated association between UFP exposure and cognitive function in ACT. We end this chapter by providing guidance on mobile monitoring study design features that should be prioritized if the goal is to develop exposure assessment models for epidemiological applications.

METHODS

Figure 4.3 illustrates the analytic approach for this analysis, as detailed below. In summary: UFP measurements from the Seattle mobile monitoring campaign (described in Chapter 3) were randomly sampled following common sampling designs; the resulting data were used to develop UFP exposure prediction models; and these predictions were used to evaluate UFP exposure model performance and to assess the estimated association between UFP exposures and cognitive function in the ACT cohort.

Cohort and Cognitive Assessments

This study was conducted in the ACT cohort (Kukull, 2001) under the auspices of the ACT-AP study. As described

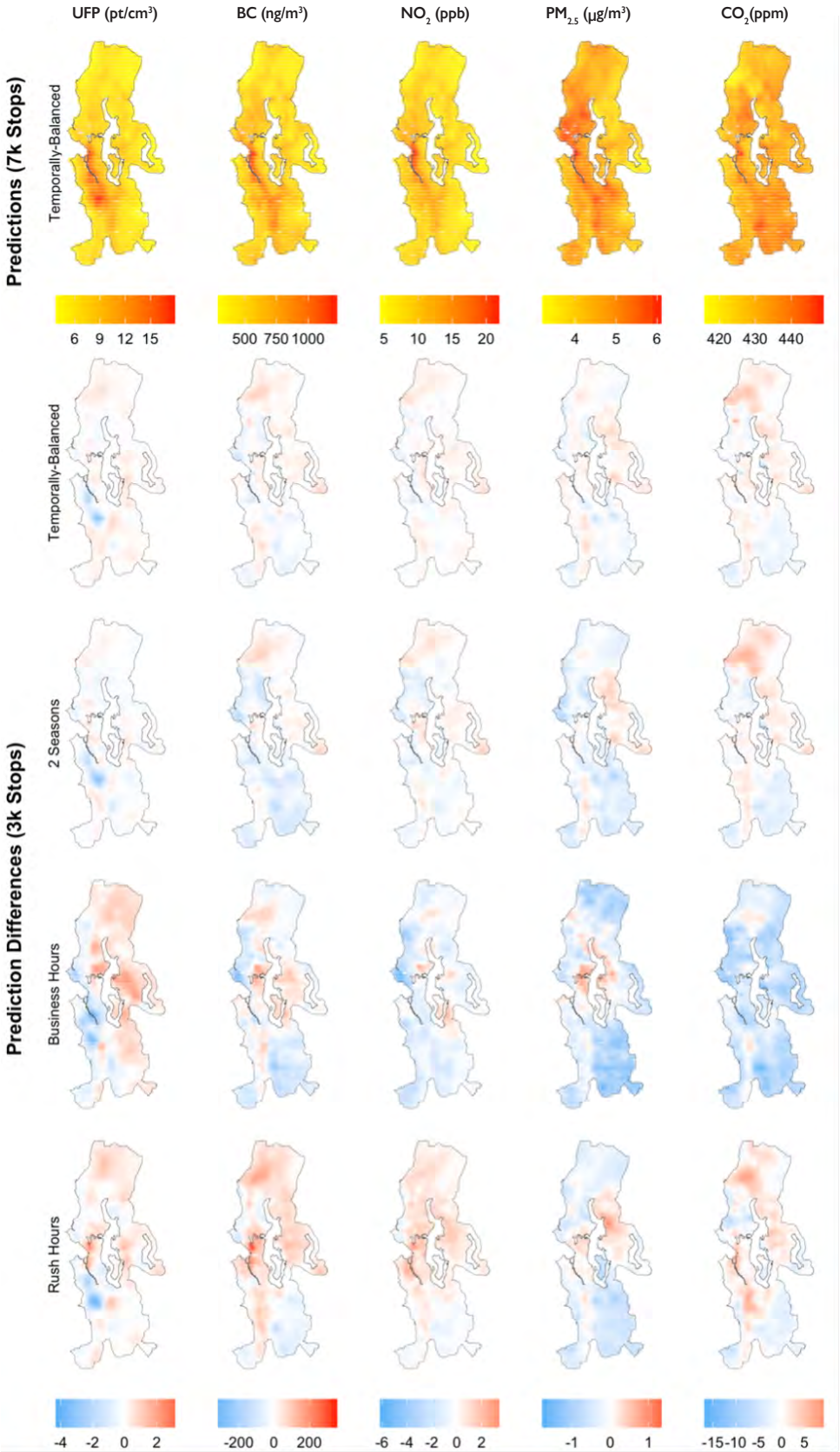


Figure 4.2. Comparison of the exposure surface from the all data campaign with the median prediction difference from some example reduced sampling campaigns. The all data campaign had over 7,000 stops (top) while the reduced sampling campaigns had ~3,000 stops using stationary roadside data from the Seattle mobile monitoring campaign. The ~3,000 temporally-balanced stops are from the fewer total stops design that randomly selects from the all data campaign with no time restrictions and can serve as a reference for the other reduced sampling campaigns. Reprinted with permission from Blanco et al. 2023a. Copyright 2023 American Chemical Society (ACS).

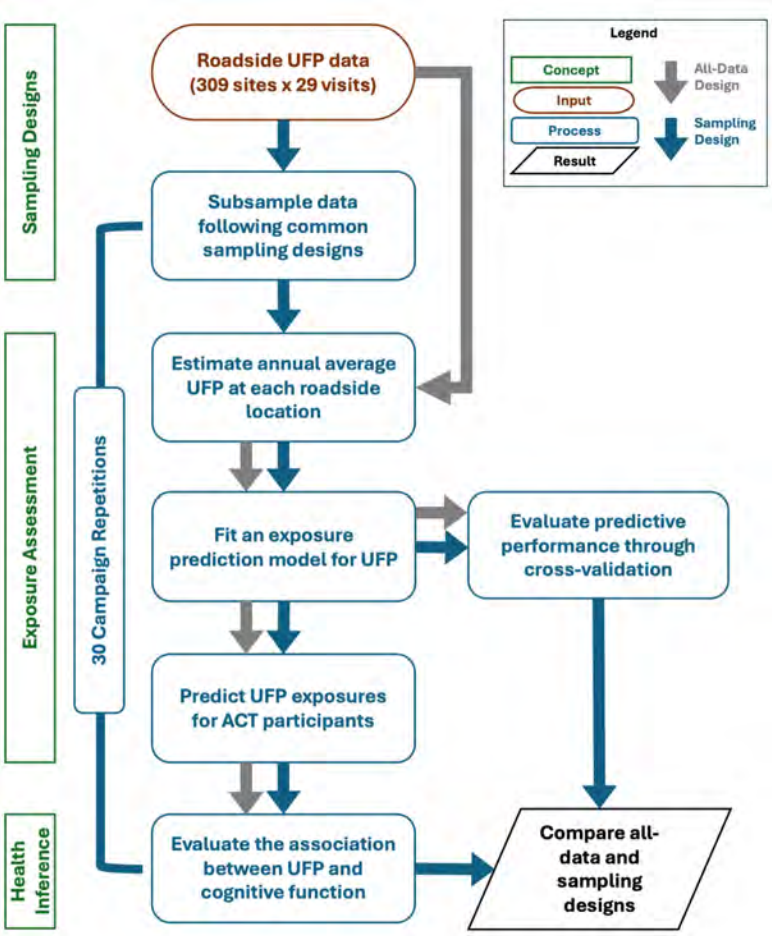


Figure 4.3. Summary of analytic approach for alternative exposure assessment designs derived from the Seattle mobile monitoring campaign stationary roadside data. See Table 4.1 for a summary of the specific sampling designs.

in Chapter 3, the analytic cohort consisted of 5,409 older (65+) participants who were dementia-free at baseline (see Table 3.1). We used baseline cognitive function (measured by CASI-IRT score) as the main outcome of interest in this study.

Exposure Assessment from Mobile Monitoring Campaigns

We leveraged the stationary roadside UFP measurements from the Seattle mobile monitoring campaign described in Chapter 3 (Blanco et al. 2022, 2023a). These data consisted of approximately 29 (IQR: 29–29, range: 26–35) temporally balanced visit measurements that were collected from 309 roadside sites. We use total (10–420 nm) PNC from the NanoScan as our primary UFP measurement, and in sensitivity analyses, we looked specifically at 10–100 nm PNC from the NanoScan (UFPs are commonly defined as ≤ 100 nm) and at 20–1,000 nm PNC from the P-Trak instrument (a common UFP monitoring instrument that does not capture the smallest particles that may be of most interest). These data were used to estimate "all data" site annual averages and were treated as gold-standard reference estimates, as described below.

We subsampled all of the UFP data (309 sites × ~29 visits each) with replacement, following four common restricted sampling designs (30 campaigns each) (**Table 4.1**).

In the first design, we sampled fewer visits per site (n = 4, 6, and 12) with no additional temporal restrictions. In the second design, we restricted sampling to fewer (1–3) seasons and collected 12 visits from each site, balanced across sampling seasons (e.g., 4 samples per season for a three-season campaign).

In the third design, we sampled fewer visits per site (n = 12) during weekday business (9 a.m. to 5 p.m.) or rush (7–10 a.m. and 3–6 p.m.) hours. The fewer-visit design with 12 visits per site was a reference for these designs; it collected the same number of visits per site without temporal restrictions. We used business and rush hour visit samples both as-is (unadjusted) and temporally adjusted — a common approach for addressing known biases resulting from restricted sampling campaigns that do not sample during the full exposure period of interest (e.g., excluding weekends or nighttime) when the goal is to estimate an annual average (Eeftens et al. 2012; Klompmaker et al. 2015; Montagne et al. 2015; van de Beek et al. 2021; van Nunen et al. 2017). This temporal adjustment approach generally entails using an air monitoring site with continuous monitoring (typically a "background" or low-concentration site); calculating time-specific adjustment factors, based most commonly on the difference between a time-specific (e.g., hourly) measurement and the site's long-term average; and applying these adjustment factors to the measured concentrations. Our approach to approximating this general strategy is detailed in Note S4.1. In summary, our temporal approach consisted of (1) simulating long-term UFP monitoring at an urban background site (Beacon Hill; continuous measures were unavailable for the entire mobile monitoring study period) from periodic UFP measures, co-located NO_2 measures, and temporal indicators; (2) generating adjustment factors, defined as the difference between the predicted hourly UFP concentrations and the long-term average UFP concentrations at Beacon Hill; and (3) applying these adjustment factors to the mobile monitoring data collected under business and rush hours designs.

In our fourth design, we evaluated a strategy characterized by unbalanced visits, a practice employed by nearly all field campaigns. We sampled based on predicted site

Table 4.1. Reduced Sampling Designs Using the Seattle Mobile Monitoring Campaign Data Consisting of 309 Stationary Roadside Sites, Each with Approximately 29 Visits[a]

Design[b]	Versions	Number of Versions	Total Visits[c]	Visits per Site	Campaign Repetitions
All-data	All-data	1	8,969	29[d]	1
Fewer Visits (no temporal restrictions)	4, 6, 12 visits per site	3	1,236, 1,854, 3,708	4, 6, 12	30
Fewer Seasons[e]	1–4 seasons	4	3,708	12	30
Fewer Hours	Weekday business or rush hours, unadjusted or temporally adjusted	4	3,708	12	30
Unbalanced Visits	High (H) and low (L) variability sites receive the following visits: H2 L22, H6 L18, H12 L12 (all receive 12 visits), H18 L6, H22 L2	5	3,708	2–22[e]	30

[a] Each reduced sampling design has 30 campaign repetitions.
[b] The all data campaign is a reference for all other reduced sampling designs.
[c] Total visits is 309 sites times the number of visits at each site.
[d] Mean and median: 29; IQR: 29–29; range: 26–35.
[e] Samples were distributed evenly across the randomly selected seasons (e.g., 12 site visits/3 seasons = 4 site visits/season).
[f] Mean: 12.

variability, defined by PLS regression where we regressed site-specific UFP IQR (median [range] of these IQRs: 7,183 [2,834–22,625] pt/cm^3 based on ~29 visits per site) against the first two PLS components summarizing hundreds of geographic covariate predictors (Table S3.2). The in-sample model R^2 was 0.46. We used this model to predict in-sample site-specific IQR and ordered these such that 129 (42%) sites were treated as medium variability sites, and visits continued to be fixed to 12. The remaining sites were split into high or low variability (n = 90 [29%] each). Figure S4.4 shows the distribution of IQRs used for the variability group. We incorporated more visits for high-variability sites (14 to 22 visits) and fewer visits for low-variability sites (10 to 2), and vice versa.

The same sampling campaigns (i.e., exact visit samples) were used for sensitivity analyses of 10–100 nm and 20–1,000 nm particles for all designs other than the business and rush hour designs, where a different set of 30 campaigns was randomly sampled. This should not be a source of bias because all campaigns were randomly selected.

In total, there were 480 candidate sampling campaigns and subsequent exposure models for our primary analysis using 10–420 nm PNC from the NanoScan in addition to the all data exposure model.

Health Inference

As outlined in Chapter 3, for each campaign, we calculated annual average site concentrations and developed UK-PLS exposure models that we evaluated using fivefold cross-validation. We used these to evaluate the time-weighted average UFP exposure for each participant at baseline based on their prior 5-year residential history in confounder model 1 (adjusting for age, calendar year, sex, and educational attainment) to assess the adjusted association between UFP exposure and baseline cognitive function (CASI-IRT), as given in Equation 3.11. We also fit confounder model 2 that further adjusted these analyses for race (White, People of Color) and SES based on the Neighborhood Disadvantage Index at a participant's longest-lived address at or prior to baseline. The Neighborhood Disadvantage Index is a validated indicator composed of Census tract-level variables from the American Community Survey (Miles et al. 2016).

RESULTS

Cohort Characteristics

Table 4.2 describes the baseline analytic cohort characteristics. On average (SD), participants were 74 (6) years old, slightly more were female, about half had at least a college education, and the average (SD) CASI-IRT score of 0.34 (0.71). See Chapter 3 for additional cohort information.

Exposure Assessment and Model Performances

The median (IQR) site UFP concentration for the primary analysis from the all data campaign was 9,742 (8,412–11,199) pt/cm^3 (Figure S4.5). Sampling designs had similar but slightly more variable annual average site estimates. Sensitivity analyses resulted in lower site concentration estimates; 20–1,000 nm PNC from the P-Trak had the smallest concentrations (Figure S4.5).

Table 4.2. Baseline ACT Cohort Characteristics by UFP Tertile Using NanoScan Stationary Roadside Data (10–420 nm) from The Seattle Mobile Monitoring Campaign[a]

	Low UFPs (*N* = 1,785)	Medium UFPs (*N* = 1,785)	High UFPs (*N* = 1,839)	Overall (*N* = 5,409)
Visit Age (Years)				
Mean (SD)	73.6 (6.03)	74.0 (6.40)	74.4 (6.48)	74.0 (6.31)
Median (Min, Max)	72.0 (65.0, 98.0)	73.0 (65.0, 96.0)	73.0 (65.0, 101)	73.0 (65.0, 101)
Sex				
Male	767 (43.0%)	742 (41.6%)	750 (40.8%)	2,259 (41.8%)
Female	1,018 (57.0%)	1,043 (58.4%)	1,089 (59.2%)	3,150 (58.2%)
Degree				
None	128 (7.2%)	136 (7.6%)	158 (8.6%)	422 (7.8%)
GED/High School	657 (36.8%)	607 (34.0%)	733 (39.9%)	1,997 (36.9%)
Bachelor's	423 (23.7%)	443 (24.8%)	408 (22.2%)	1,274 (23.6%)
Master's	288 (16.1%)	319 (17.9%)	265 (14.4%)	872 (16.1%)
Doctorate	110 (6.2%)	122 (6.8%)	98 (5.3%)	330 (6.1%)
Other	179 (10.0%)	158 (8.9%)	177 (9.6%)	514 (9.5%)
Race				
White	1,666 (93.3%)	1,602 (89.7%)	1,551 (84.3%)	4,819 (89.1%)
People of Color	116 (6.5%)	181 (10.1%)	285 (15.5%)	582 (10.8%)
Missing	3 (0.2%)	2 (0.1%)	3 (0.2%)	8 (0.1%)
Neighborhood Disadvantage Index (NDI)				
Mean (SD)	–0.978 (0.608)	-0.818 (0.627)	–0.341 (0.753)	–0.710 (0.719)
Median (Min, Max)	–1.10 (–2.59, 1.20)	–0.907 (–2.66, 1.97)	–0.242 (–2.45, 2.64)	–0.851 (–2.66, 2.64)
Missing	30 (1.7%)	35 (2.0%)	53 (2.9%)	118 (2.2%)
Cognitive function CASI-IRT				
Mean (SD)	0.37 (0.69)	0.37 (0.72)	0.28 (0.71)	0.34 (0.71)
Median (Min, Max)	0.41 (–1.96, 1.75)	0.40 (–1.98, 1.75)	0.30 (–2.12, 1.75)	0.37 (–2.12, 1.75)
Residential UFP (pt/cm^3) Exposure				
Mean (SD)	8,760 (647)	10,100 (311)	12,500 (2,080)	10,500 (2,020)
Median (Min, Max)	8,890 (5,930, 9,570)	10,100 (9,570, 10,700)	11,700 (10,700, 22,100)	10,100 (5,930, 22,100)

[a] Low, medium, and high UFP tertile is based on the predicted UFPs from the all data exposure model.

The all data campaign UFP exposure models had a cross-validated R^2_{MSE} value of 0.65 (**Figure 4.4**). Almost all sampling designs with restricted sampling had lower-performing exposure models. Performances were incrementally worse for campaigns with fewer visits per site, fewer seasons, and more restricted sampling days and times (business and rush hours). Performance was worse when temporal adjustments were applied (see Figure S4.6 for paired comparisons of adjusted and unadjusted campaign model performances), and when there was an unbalanced number of visits sampled across sites, particularly when high variability sites had very few visits. Despite collecting the same number of total visits (309 sites × 12 visits), campaigns with one season duration, those conducted during business hours (adjusted and unadjusted), and those with few visits to high variability sites (even when more visits were collected from lower variability sites) performed worse than otherwise unrestricted 12 visit designs. Sensitivity analyses for 10–100 nm and 20–1,000 nm PNC showed similar patterns (Figure S4.8).

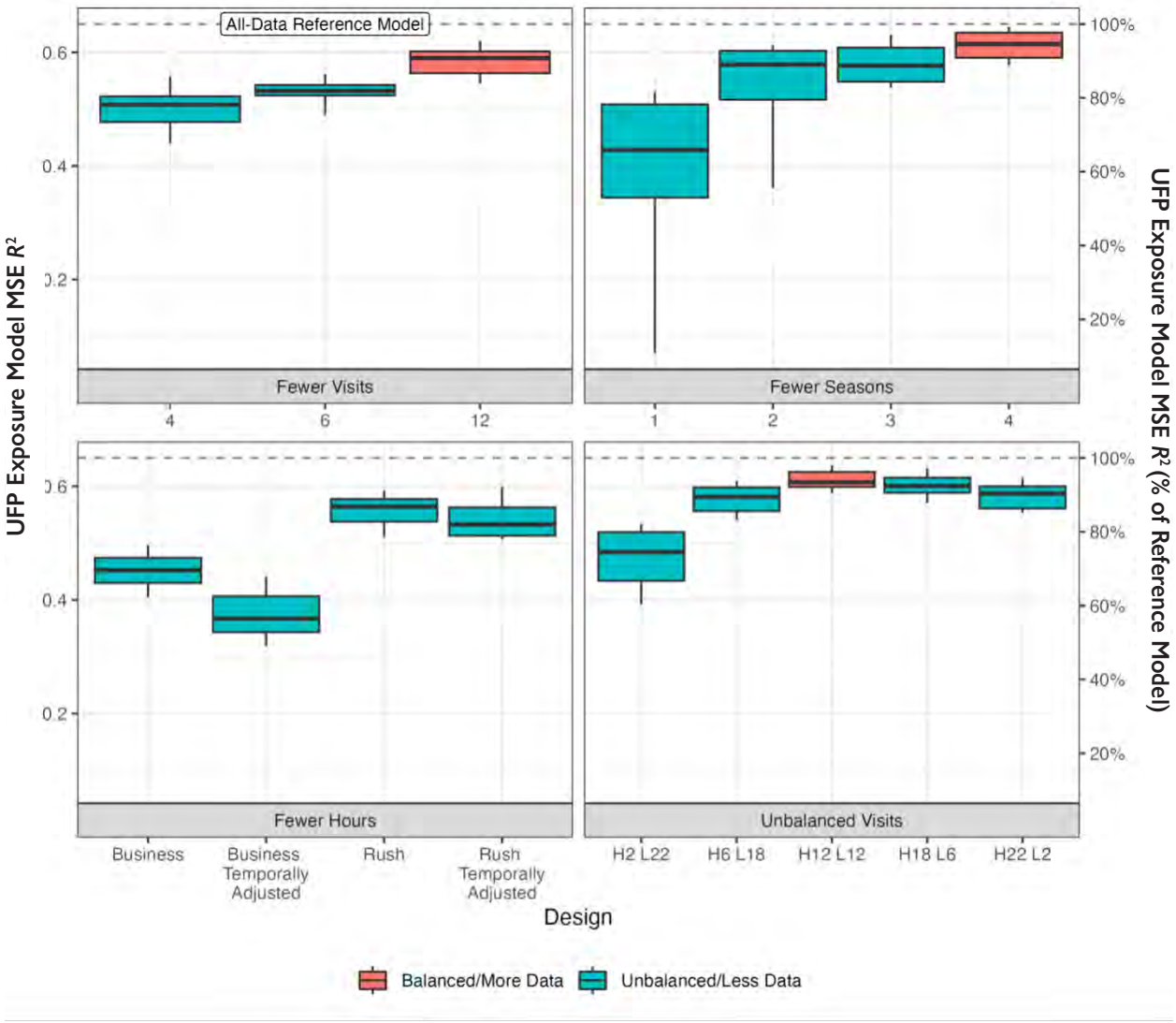


Figure 4.4. Cross-validated UFP model performances (N = 30 campaigns per design) using stationary roadside data from the Seattle mobile monitoring campaign. The dashed lines indicate the all data campaign performance. Red design reference box plots indicate the least restrictive or most balanced campaigns; any of these can serve as a reference for the business and rush hours designs. Business and rush hour designs produce annual average site estimates from unadjusted and temporally-adjusted visits. UFP models are for a total of 10–420 nm particles (pt/cm^3) from the NanoScan instrument. The dashed line indicates the MSE R^2 from the reference all data model, which is 0.65.

The median (IQR) predicted UFP concentration for participants was 10,100 (9,200–11,100) pt/cm^3 and ranged from 5,930–22,100. Exposure predictions for sampling designs varied across campaigns (Figure S4.9). The business hour design tended to underpredict high exposures relative to the all data exposure model, while the rush hour design overpredicted high exposures. Designs with few visits to high variability sites (and more visits to low variability sites, with H2 L22 being the most extreme case) had highly variable predictions across campaigns, particularly for high concentrations. We saw similar patterns in sensitivity analyses of 10–100 nm and 20–1,000 nm PNC.

Predictions from most designs were highly correlated with predictions from the all data campaign (median Pearson correlations [R] > 0.85), although the business hour design was consistently lower than all other designs (R ~0.77–0.78; Figure S4.10). Lower correlations indicate differences in exposure surfaces (predictions) for mobile monitoring designs with fewer visits per site, those with fewer seasons measured, those limited to business hours, and those with fewer visits at high-variability sites. All designs had one or more atypical campaigns (i.e., outliers not visualized in the Figure S4.10 box plots) that had a meaningfully lower correlation with the all data campaign than the majority of other similarly designed

campaigns, indicating potentially meaningful variability and lower exposure model performances across campaign iterations.

Inferential Analyses

Using the all data campaign exposure model, the adjusted mean baseline CASI-IRT score from confounder model 1 was lower by –0.020 (95% confidence interval [CI]: –0.036, –0.004) for every increment of 1,900 pt/cm^3 (Table 3.2). **Figure 4.5** summarizes the point estimates across sampling campaigns and their percentage difference relative to the health association obtained from the all data exposure model. The health associations for the fewer-visits and fewer-season designs are similar, with campaigns with more visits and longer durations being most similar and associated with more consistent (less variable) results across campaigns. Designs with the highest variability in the estimated health associations across campaigns (indicating less consistent results) were those with fewer visits (4 visits per site), one season, and poor spatial balance, where high-variability sites receive few visits while low-variability sites receive many (H2 L22). Business and rush hour designs, on the other hand, produce biased, attenuated health associations that are about 60% and 40% different from the all data estimate, respectively. Temporally adjusting these designs was associated with slightly more accurate health associations, at least for the rush hours design. Campaigns with balanced designs where all sites receive the same number of visits (12) produce health associations that are the most consistent with the all data exposure model. Figure S4.11 shows similar results for exposure sensitivity analyses, with 20–1,000 nm PNC P-Trak exposure models showing greater differences closer to 70% for business hour designs. Figure S4.12 further details the point and 95% CI for selected campaigns for primary and sensitivity exposure analyses.

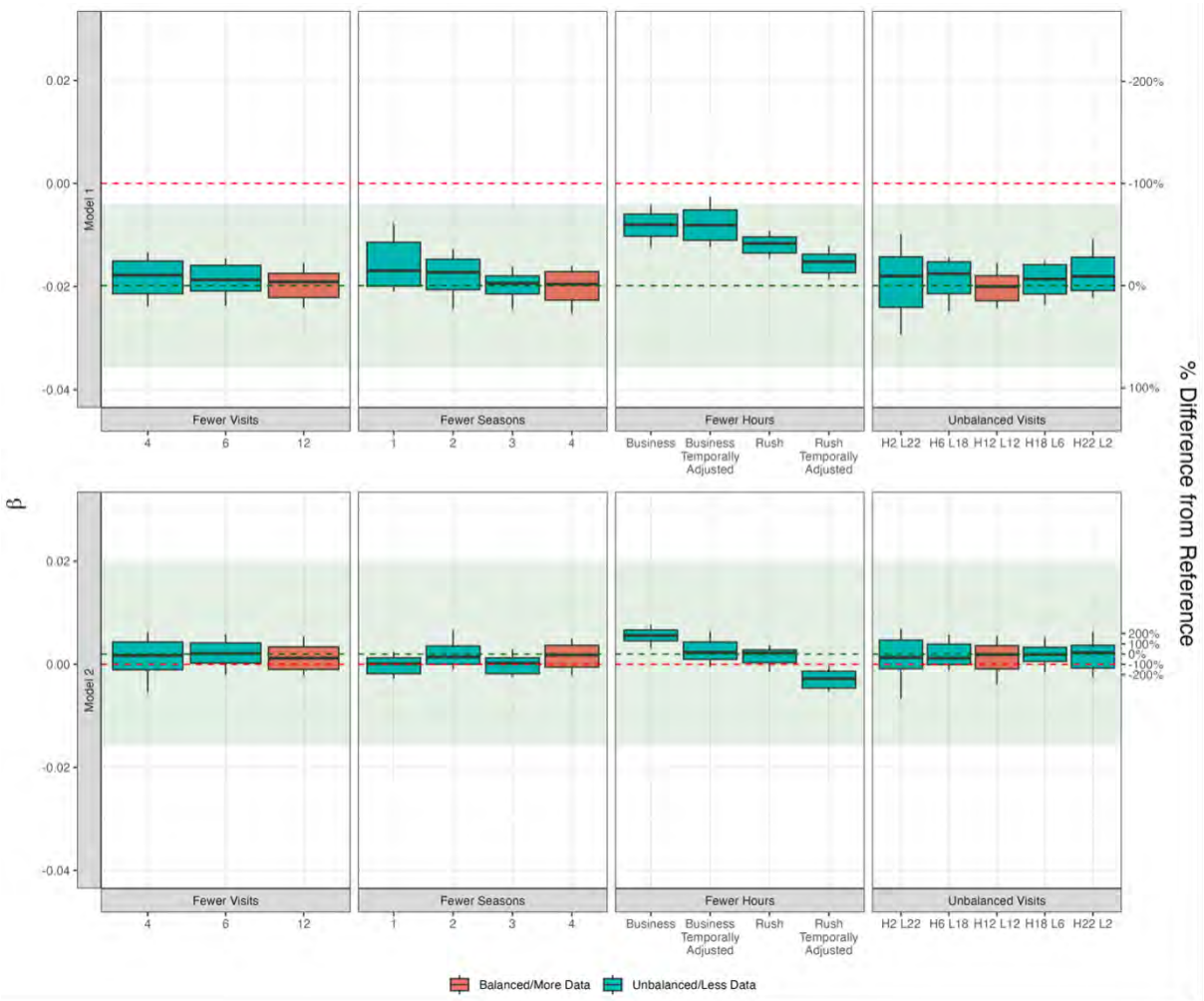


Figure 4.5. Estimated association between UFPs (per 1,900 pt/cm^3) and cognitive function adjusted for age, calendar year, sex, and education (confounder model 1). Confounder model 2 was further adjusted for race and SES using stationary roadside data from the Seattle mobile monitoring campaign. The dashed green lines and shaded areas indicate the estimated point and 95% CIs from the all data exposure model, which are –0.020 (95% CI: –0.036, –0.004) in confounder model 1 and 0.002 (–0.016, 0.020) in confounder model 2. The dashed red line indicates no association. Box plots show the results when using exposure estimates from reduced mobile monitoring sampling campaigns (N = 30 estimates per box plot). Boxes show the median and IQR, whiskers show the 10th and 90th percentiles. Percentages on the y-axis show the estimated association relative to using the all data exposure model.

Confounder model 2 with the all data exposure model produced an inconsistent (nonstatistically significant) association between baseline CASI-IRT and each 1,900 pt/cm^3 increment in UFP concentration (0.002 [95% CI: −0.016, 0.020]). As before, most reduced sampling designs produced generally similar health inferences except for business and rush hour designs. Business hours designs produced larger health estimates that are in the wrong hypothesized direction (i.e., suggesting "protective" air pollution effects). Temporal adjustment slightly reduced these estimates closer to the reference estimates. Rush hour designs, on the other hand, generally produced unbiased health estimates. Temporal adjustment, in this case, moved away from the reference estimate, albeit in the hypothesized direction.

DISCUSSION

Mobile monitoring campaigns to assess traffic pollutants, including UFPs, are being used around the globe to address monitoring gaps (Kim et al. 2023). Many campaigns now aim to develop exposure assessment models to be used in epidemiological applications. This application generally necessitates capturing long-term, off-road (generally residential) exposures and is different from commuter exposures or hotspot identification studies, among others. Still, guidance on mobile monitoring study design for epidemiological applications has been largely absent from the literature (Blanco et al. 2022, 2023b). As a result, there is substantial variability in how mobile monitoring campaigns are designed and implemented. We previously showed that monitoring design features like the number of sites, number of visits, campaign duration, and choice of sampling days and sampling hours can greatly impact the predictive performance of exposure assessment models (Blanco et al. 2023b). Here, we further assess additional monitoring approaches and how these differences in exposure model choices impact epidemiological inferences.

We found that, when compared to an extensive mobile monitoring campaign intentionally designed for epidemiological application (the all data exposure model), campaigns with fewer visits (~4–12) but no temporal restrictions, fewer seasons (~2–3), and a balanced number of visits across sites (12 in this case) had only slightly worse exposure model performances (Figure 4.4). Similarly, those designs had highly correlated (MSE R^2 values mostly greater than 0.85) participant exposure assessments (Figure S4.10), and health inferences that were similar to the reference (Figure 4.5). As expected, shorter campaigns (e.g., one season) and those with fewer repeat site visits generally had worse-performing models with predictions that were less correlated to those from the all data campaign. Rush hour and especially business hour designs, on the other hand, had much worse exposure model performances and more variable participant exposure assessments, although their predicted exposures were moderately (business) and highly (rush) correlated with the all data exposure model predictions. Health associations using confounder model 1 for the temporally restricted rush and business hours designs were less consistent with the reference model point estimate, although generally within its confidence interval. More generally, the health associations using confounder model 1 from these monitoring campaigns were noticeably different from all other designs, including much shorter (e.g., one season vs. year-round) campaigns and those with fewer samples (e.g., 4 vs. 12 visits per site), suggesting that capturing temporal variability through extended hours designs (e.g., sampling weekends and extending the sampling hours) is important for capturing long-term annual average exposures. Notably, these reduced day and hour campaigns are most common in the field because an operator is required to operate a vehicle and monitor instrumentation throughout the sampling period.

Interestingly, business and rush hour designs with temporal adjustments were sometimes associated with worse exposure model performances compared to temporally unadjusted designs (Figure 4.4, Figure S4.6, Figure S4.7), and the resulting health estimates varied in whether they were closer to or further from the reference estimates (Figure 4.5). Sites within a region can have different temporal patterns (Blanco et al. 2023b) related to major nearby sources (or their absence), such as airports, highways, or industrial sites. Applying temporal adjustments from a single site may incorrectly or insufficiently adjust exposure estimates. The temporal adjustment is also likely to introduce more classical-like measurement error due to the estimation of new temporal adjustment parameters. The temporal adjustment may also affect Berkson-like measurement error by improving or worsening the temporal alignment of the exposure, depending upon the reference information used and the approach to the temporal adjustment. These features will impact the interplay between Berkson-like and classical-like measurement error, such that temporally adjusted estimates could result in inferences that are more or less biased than the corresponding design without temporal adjustment. For instance, in situations where inferences were improved, it is possible that while the added complexity of the time adjustment introduced a form of classical-like measurement error that adversely impacted prediction accuracy, the improved temporal alignment decreased the impact of Berkson-like error, which was responsible for the dominant health association bias from the unadjusted exposure estimates (Szpiro et al. 2011b; Szpiro and Paciorek 2013a). Prior work has also shown that exposure measurement error can bias health inferences, although the direction of the bias is unclear and not always toward the null (Brenner and Loomis 1994; Szpiro and Paciorek 2013a). Exposure monitoring design thus matters. Our findings suggest that deviating from balanced designs can produce biased or inconsistent health inferences that may be challenging to correct with existing temporal adjustment approaches.

A feature of our temporal adjustment approach was that we based it on a simulated UFP monitoring site, as described in the Methods, from co-located UFP and highly temporally correlated NO_2 observations along with other temporal indicators. This approach was associated with good UFP

predictions and captured much of the temporal variation in UFPs (Figures S4.3–S4.5), suggesting it was a reliable source for estimating temporal adjustment factors. In the literature, various adjustment approaches (e.g., "difference" or "ratio" approaches) and sites have been used to temporally adjust mobile monitoring readings. These approaches have not been validated with data and themselves produce fluctuating adjustment factors. Finally, we used hourly adjustment factors to adjust 2-minute mobile monitoring site visits. We have previously shown that these measurements are highly correlated (Blanco et al. 2022).

While confounder model 1 did not extensively adjust for potential confounding, we observed similar patterns in confounder model 2 analyses that further adjusted for race and neighborhood SES. These analyses did not result in a strong association between UFPs and cognitive function. While this may be a result of confounding by race and SES, it is also possible that adjusting for these strongly correlated factors (Chambliss et al. 2021; Hajat et al. 2015; Saha et al. 2022) lowers air pollution variability for a given race and SES level, thus reducing study power. Moreover, the goal of the analyses in this chapter was to estimate the potential *magnitude* of bias and variability when applying restricted sampling campaigns relative to intentionally designed, spatially and temporally balanced campaigns. Any remaining potential confounding unaccounted for in our analyses is constant across monitoring designs, thus allowing us to capture trends driven by differences in exposure assessment design.

There are different ways field sampling is conducted that lead to unbalanced sampling, whereby some locations receive more visits than others. This may result from having nonfixed driving routes, logistical constraints that make it challenging to visit some sites while others along common driving routes are naturally oversampled, intentionally oversampling sites anticipated to have high variability while deprioritizing sites with low variability (e.g., suburban areas), etc. We present one approach whereby sampling is influenced by the anticipated site concentration variability across time. We did not see an appreciable benefit to exposure model performance from oversampling high variability sites, although performance was lower when we dramatically undersampled high variability sites (Figure 4.4, H2 L22). Schemes with balanced samples across sites (12 visits each) had the least biased and least variable health estimates (Figure 4.5). These findings suggest using a balanced sampling design whenever feasible. If traveling to sites with low anticipated variability presents a significant logistical challenge, however, our results suggested that strategies characterized by somewhat fewer visits to these sites may be a reasonable choice.

For evaluating unbalanced sampling, we used predicted, in-sample IQR based on PLS regression analysis rather than "true" IQR to classify sites. This adds some error to site classifications despite being in-sample predictions, which can produce overfitted or optimistic results. Nonetheless, true site concentrations and variability are largely unknown prior to conducting in-field mobile monitoring, adding uncertainty to monitoring decisions. Moreover, defining target sites becomes more challenging for multiple pollutants because spatial and temporal patterns may vary across pollutants, such that a site could have high variability for one pollutant and low variability for another.

Overall, our findings suggest that strategic monitoring design can be implemented to optimize the accuracy of health inferences and the anticipated consistency of these results across campaigns (i.e., generally narrower box plots in Figure 4.5), while keeping in mind the logistical constraints unique to mobile monitoring. We recommend prioritizing sampling during extended periods outside of rush and business hours, as these times were sometimes linked with less accurate health inferences, and standard temporal adjustment methods produced inconsistent results. Beyond that, collecting data over at least two seasons if the goal is to estimate an annual average, collecting a balanced (fixed) number of visits across locations, and collecting a higher number of visits per location may improve health inference accuracy and/or reduce variability across campaigns, depending on the exact analysis. In terms of generalizability, seasonal sampling requirements may be impacted by seasonal variability across geographical locations and over time, and may thus differ across locations. Moreover, sampling design impacts some pollutants more than others. This variability will likely translate to subsequent health inferences to varying degrees.

Spatial and temporal compatibility (i.e., similarity in distributions) between monitoring and cohort locations is an important feature for minimizing the impact of measurement error and consequently optimizing health inferences (Keller et al. 2017; Szpiro and Paciorek 2013a). This is particularly relevant for air pollution epidemiology, where health associations can be subtle, and biases or lower levels of precision can easily obscure meaningful associations. Our study is inherently spatially aligned because our extensive mobile monitoring campaign was specifically designed to capture exposures for the ACT cohort (Blanco et al. 2022). It is notable that most campaigns select monitoring locations based on geographic features or sources (e.g., major roads, industry, airports) and do not explicitly set out to capture exposures based on the geographical spread of study participants. A focus on spatial compatibility may instead prioritize monitoring at locations near or representative of cohort locations and ensure that monitoring covers the specific geographical variability of the cohort (regional and cohort geographic variability may or may not be aligned). One limitation of this work, however, is the use of outdoor (ambient) exposures at residential locations to evaluate personal exposures, which primarily occur indoors and can also occur away from home (i.e., time–activity patterns). This approach is common in large cohort studies due to monitoring challenges, although it adds exposure measurement error that can affect the estimated health associations.

In this case study of UFPs and cognitive function, we observed an association between our exposure data and the outcome of interest when we used confounder model 1. While mobile monitoring inherently results in missing observations, the all data and other similar designs (e.g., three seasons) estimated annual average exposure levels that are close to the true annual average. The day- and time-restricted designs, however, sample during times that are temporally misaligned with the longer-term exposures of interest. These designs contribute to bias from Berkson-like error, which is the difference between the true annual average exposure surface and the more limited part captured by the modeling process (Szpiro and Paciorek, 2013a). The ideal way to eliminate the bias from temporal misalignment is by modifying the sampling design, but if this is not possible, an alternative is to introduce a spatiotemporal model that fully captures the complexity of the underlying exposure surface. Our use of temporal adjustment affected the amount of bias differently for the two confounder models (see Figure 4.5), but evidence of residual bias remained. Future research can explore the question of whether the available data are sufficiently rich to support a full spatiotemporal model that will more fully capture the underlying exposure surface, and thus eliminate bias from Berkson-like error. Another possibility is to reweigh the data to achieve temporal compatibility. However, it is not clear whether either of these approaches will be successful with the rush-hour or business-hour designs because key information about what happens during the noncovered hours is completely unavailable.

More generally, our findings are conservative with respect to published studies. Many mobile monitoring campaigns incorporate multiple features that might limit their applicability to long-term population exposure studies; for example, campaigns that last less than a year, collect fewer repeat visits per site (median ~4), sample only during weekday business hours, and collect unbalanced numbers of visits per site (Kim et al. 2023). We anticipate that these designs will produce biased health associations like those that we observed for the business hours design, if not more severe. Moreover, most mobile monitoring campaigns explicitly collect nonstationary, on-road data. While nonstationary designs achieve higher spatial coverage than stationary roadside designs like the one used in this study, they measure on-road concentrations that are typically higher than those captured by stationary, roadside locations (which are more similar to residential exposures); collect less data per road segment (seconds vs. minutes) making for more unstable estimates; and the resulting exposure models are associated with poorer performance (Doubleday et al. 2023; Kim et al. 2023). Most nonstationary campaigns also do not adjust on-road data to minimize the influence of air pollution plumes (spike concentrations) common on roads, but less so at residential locations. These design features are unique to nonstationary mobile monitoring, and their potential impact on health inferences should be investigated in future work. Chapter 6 summarizes our investigation of nonstationary mobile monitoring designs.

We conducted this study using data from the long-standing, community-based ACT cohort. While we could have simulated health outcome data to conduct this analysis, we chose to use this data source to reflect real-world impacts and incorporate aspects that might not be included in a simulation study. As such, this approach may be more illuminating of real-world implications compared to a simulation study. ACT has consistently collected measurements over time, including cognitive function, demographics, and lifestyle factors. ACT's extensive participant residential histories allowed us to assess UFP exposures for most participants. Because we used fixed annual average 2019 UFP exposure surfaces to assess exposures, there is inherent exposure assessment error in these analyses, including the all data campaign, and this likely was higher for earlier time periods. Historical UFP data are rare, and we assumed that the exposure surface was constant over time (Blanco 2021; Kim et al. 2017; Levy et al. 2015; Meng et al. 2019; Molter et al. 2010; Y. Wang et al. 2011). More generally, our inferential models in this analysis were not necessarily meant to characterize causal effects between UFPs and cognitive function. Such analysis could consider additional confounding adjustments and address potential selection biases that may have resulted, for example, from conducting complete case analyses. The goal of this analysis was to characterize how mobile monitoring design choices may impact the estimated health associations of air pollution.

Intentional monitoring design that ensures that the exposure data collected is aligned with the planned application of the exposure assessment, while potentially less critical in some settings, can support the validity of inference in epidemiological analyses. To maximize exposure and health inference accuracy, we recommend extending sampling beyond typical weekday business or rush hours, collecting data over at least two seasons if the goal is to estimate an annual average, collecting a balanced (fixed) number of visits across locations, and collecting more visits per location. Our future work will investigate how monitoring design more specifically impacts nonstationary, on-road data exposure models and health inferences. Some of this is summarized in Chapter 6.

CHAPTER 5: CHARACTERIZATION OF AND ADJUSTMENT FOR MEASUREMENT ERROR IN HEALTH INFERENCE

Lead authors: Magali Blanco, Adam Szpiro, Lianne Sheppard

INTRODUCTION

Adjusting for exposure measurement error, the discrepancy between the surrogate and true exposure, is critical for environmental epidemiology (Katsouyanni and Evangelopoulos 2022; Keller et al. 2017; Wei et al. 2022). Measurement error has the potential to modify the estimated health associations and their uncertainty estimates in ways that may be unclear. It often (but not always) biases health estimates toward the null, with the exact impact depending on the correlation between the exposure and confounders (Katsouyanni and Evangelopoulos, 2022; Wei et al. 2022). Biased health associations and their accompanying uncertainty estimates misinform us of the true population risks and may impact policy efforts to understand and control potentially hazardous exposures. Despite this, exposure measurement error has received limited attention in the literature. Measurement error is particularly relevant for air pollution health studies where predicted ambient (rather than personal observed) pollution exposures are almost always used to assess exposures, and where the estimated health associations can be subtle.

In these analyses, we investigate the impact of exposure measurement error on epidemiological inferences in a case study of UFPs and cognitive function in ACT. The exposure and outcome data are the same as those used in Chapter 4, which focused on documenting the sensitivity of health associations obtained from varying exposure monitoring campaign designs. This chapter focuses on characterizing the bias and uncertainty (in the all data campaign) that can be directly attributed to using an exposure prediction model for inference about health, as summarized in **Figure 5.1**. Operating in the spatial measurement error framework developed by Szpiro and colleagues (Bergen et al. 2016; Keller et al. 2017; Szpiro et al. 2011a, 2011b; Szpiro and Paciorek 2013a, 2013b), we decomposed the error into two components. Classical-like error is the excess variability in exposure predictions induced by random selection of monitoring times and locations, stochastic variation in measurements, and local fluctuations in air pollution levels. All these sources of randomness are propagated through the exposure prediction model (e.g., universal kriging, spatial random forest), affecting the variability of the exposure model parameters, and can induce bias and inflate standard errors of the estimated health associations. Berkson-like error is the difference between the true exposure surface and the smooth exposure surface that would be reconstructed "on average" given the sampling design and choice of exposure model. Berkson-like error can bias health estimates. It can also inflate standard errors in health estimates, but this contribution is largely captured by model-robust standard calculations that do not specifically account for measurement error.

We use the bootstrap to address both of these sources of error. We take a nonparametric bootstrap approach to capture health inference bias resulting from classical-like measurement error that's associated with the uncertainty in the surface and participants. This approach involves (1) resampling monitor locations to reflect variation in the predicted exposure surface derived from different monitor locations and sampling times, and (2) further resampling participants to capture additional variability in the epidemiological inferences due to sampling different participants. We also take a parametric bootstrap approach to capture bias in the health association resulting from Berkson-like measurement error due to using predicted rather than observed exposure. We

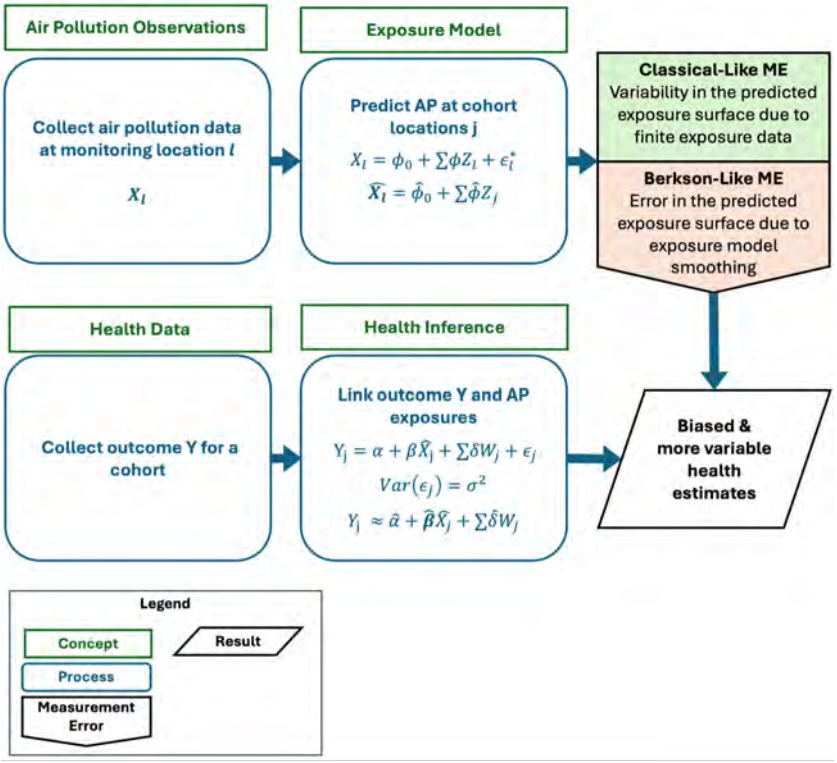


Figure 5.1. Sources and types of measurement error in air pollution epidemiology and its impacts on health inference bias and variability. ME = measurement error.

discuss the impact of measurement error on epidemiological inferences and compare it to the impacts of the various mobile monitoring study designs described in Chapter 4.

METHODS

Study Population and Cognitive Function Outcome

This study was conducted in the ACT cohort (Kukull 2001, Kukull et al. 2002) under the auspices of the ACT-AP study. As described in Chapter 3, the analytic cohort consisted of 5,409 older (65+) participants who were dementia-free at baseline (see Table 3.1). We used baseline cognitive function (measured by CASI-IRT score) as the main outcome of interest in this study.

Air Pollution Data and Exposure Assessment

Exposure to UFPs (TSI NanoScan, 10–420 nm particles) was assessed using stationary, roadside measurements from the Seattle mobile monitoring campaign described in Chapter 3 (Blanco et al. 2022). As previously described, we used these measurements to develop annual average universal kriging–partial least squares (UK-PLS) exposure prediction models. UFPs were log-transformed and regressed against the first two PLS components, which summarized hundreds of geographic covariates predictive of traffic-related air pollution (TRAP) (e.g., land use, roadway proximity, population density). Participants were assigned UFP exposures using this model based on their residential histories over the 5 years prior to baseline.

Epidemiological Inference

We fit a least squares linear regression model to assess the association between baseline CASI-IRT score and UFPs, adjusted for age, calendar year (categorical), sex, and education (categorical) (confounder model 1) using the all data (i.e., gold standard) exposure assessment model (~29 temporally balanced visits per site distributed across all four seasons, days of the week, and most hours of the day). The model is

$$Y_j = \alpha + \beta X_j + \delta W_j + e_j \quad (5.1)$$

which is similar to Equation 3.11 with an additional index m, where Y_j is CASI-IRT for participant j, W_j is predicted UFP exposure from model m for participant j, W_j refers to the vector of cross-sectional covariates measured at baseline, and e_j is residual error. β is the health parameter for UFP exposure and CASI-IRT for a given exposure model; it is the primary parameter of interest. See Chapter 3 for additional details about the health model, outcome, and covariates.

Nonparametric Bootstrap: Health Inference Bias from Classical-Like Measurement Error and Variability from both Classical-Like and Berkson-Like Measurement Error

Figure 5.2 summarizes our approach for estimating the impact of classical-like measurement error on health inference bias and both classical-like and Berkson-like measurement error on health inference variability. Classical-like measurement error results from variability in exposure predictions due to having finite exposure data. This variability and the implications for health inference can be quantified by a nonparametric bootstrap in which exposure data at monitoring locations (e.g., stationary roadside locations in the mobile campaign) are resampled with replacement, and then used to refit the exposure model and derive a new health estimate. There are two aspects to this nonparametric bootstrap approach: quantifying the bias and then further quantifying the variability.

Health Inference Classical-Like Measurement Error Bias Resulting from Variability in the Exposure Surface For the bias, that is, the bias result in Figure 5.2, the health data are retained in their original form throughout this process, without resampling. The bias in health estimates can be derived from the bootstrap samples by comparing the mean of the bootstrap samples ($\overline{\hat{\beta}_{CL}}$) to the estimated health association ($\hat{\beta}$) from the original all data exposure assessment model dataset. Specifically, we generated 500 annual average UFP exposure models as follows:

1. **Resample air pollution data:**
    a. For each iteration b, 309 locations (l) were sampled with replacement from the original 309 set of locations (e.g., the dataset is expected to contain some

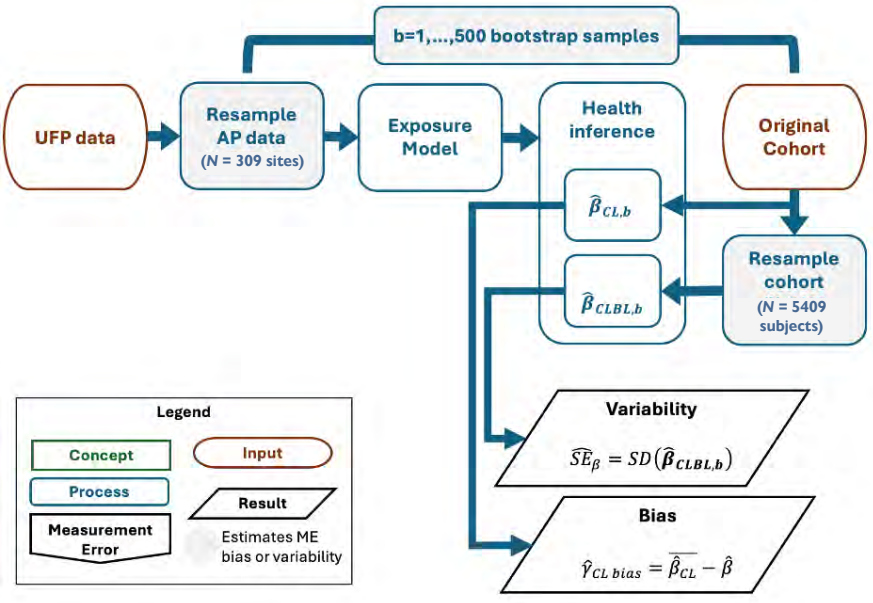


Figure 5.2. Overview of the estimation of health inference bias from classical-like measurement error and health inference variability from classical-like and Berkson-like measurement error. ME = measurement error; AP = air pollution.

duplicated locations and to have some locations not included in the sample).

   b. For each of these locations in the b^{th} iteration, n stops were sampled from the n stops available, also with replacement. One annual average was calculated for each location based on the n stop measurements sampled.

2. **Exposure model:** For the b^{th} iteration, a UK-PLS model was generated, as described previously, based on the dataset generated above, and predictions for participant locations were generated. Monitoring location latitudes and longitudes were jittered to a small degree (8th decimal place) to avoid numerical errors associated with having duplicate monitoring locations.

3. **Health inference in the original cohort:** We assessed the adjusted association between baseline cognitive function in the original analytic cohort (not resampled) and predicted 5-year UFP exposure for each model (Equation 5.1) and obtained for $\widehat{\beta}_{CL,b}$ for $b = 1, \ldots, 500$.

We defined bias from classical-like measurement error ($\hat{\gamma}_{CL\,bias}$) as the difference in the mean health inference from the 500 nonparametric bootstrap approaches $\overline{\widehat{\beta}_{CL}} = \frac{1}{500}\sum_{b=1}^{500}\widehat{\beta}_{CL,b}$ and $\widehat{\beta}$ from the all data exposure model (Equation 5.1), as well as the percent bias as

$$\hat{\gamma}_{CL\,bias} = \overline{\hat{\beta}_{CL}} - \hat{\beta} \tag{5.2}$$

$$\hat{\gamma}_{CL\,percent\,bias} = \frac{\hat{\gamma}_{CL\,bias}}{\overline{\beta}_{bias\,corrected}} x100 \tag{5.3}$$

where $\widehat{\beta}_{bias\,corrected}$ is defined in (**Equation 5.7**).

Health inference for both classical-like and Berkson-like measurement error variability resulting from variability in the exposure surface and in study participants

As shown by Szpiro and Paciorek (2013a), we can quantify the full variability in the estimated association induced by the finite population of the cohort, and classical-like and Berkson-like measurement error with the standard error of bootstrap health estimates that also includes resampling the cohort. For a corrected standard error (SE), i.e., the variability result in Figure 5.2, the health data are resampled by extending the nonparametric bootstrap. After completing steps 1–2 above:

4. **Health inference in the resampled cohort:**

   a. For each iteration b, we resampled the analytic cohort with replacement (n = 5,409).

   b. For each iteration b (now with the new cohort sample), we assessed the adjusted association between baseline cognitive function and predicted 5-year UFP exposure (Equation 5.1), where the estimate of the parameter of interest is $\widehat{\beta}_{CLBL,b}$ here, and obtained for $\widehat{\beta}_{CLBL,b}$ for $b = 1, \ldots, 500$.

We used the standard deviation (SD) of the $\widehat{\beta}_{CLBL,b}$ from the data in step 4b to estimate the health inference SE that accounts for classical-like and Berkson-like measurement error ($\widehat{SE}_{\beta}$).

Parametric Bootstrap: Health Inference Bias from Berkson-Like Measurement Error

Figure 5.3 summarizes our approach for estimating the impact of Berkson-like measurement error on health inference bias (i.e., the bias result in Figure 5.3). We expected that applying predicted rather than observed exposures to our health analyses would primarily manifest in the Berkson-like component of measurement error and be a source of bias in the estimation of the health association. Unlike classical-like error, Berkson-like error is not identifiable from the stochastic behavior of the exposure model as calculated by a nonparametric bootstrap. The key observation is that we can estimate the Berkson-like error at monitor locations because true and predicted exposures are available there. For quantitative outcomes, we have shown that this approach can be used to derive an analytic approximation to the bias from Berkson-like error by numerically integrating the estimated Berkson-like error (Bergen and Szpiro, 2015). When applying the parametric bootstrap, the difference (averaged over many bootstrapped samples) between these two health estimates (i.e., those estimated using the observed and predicted

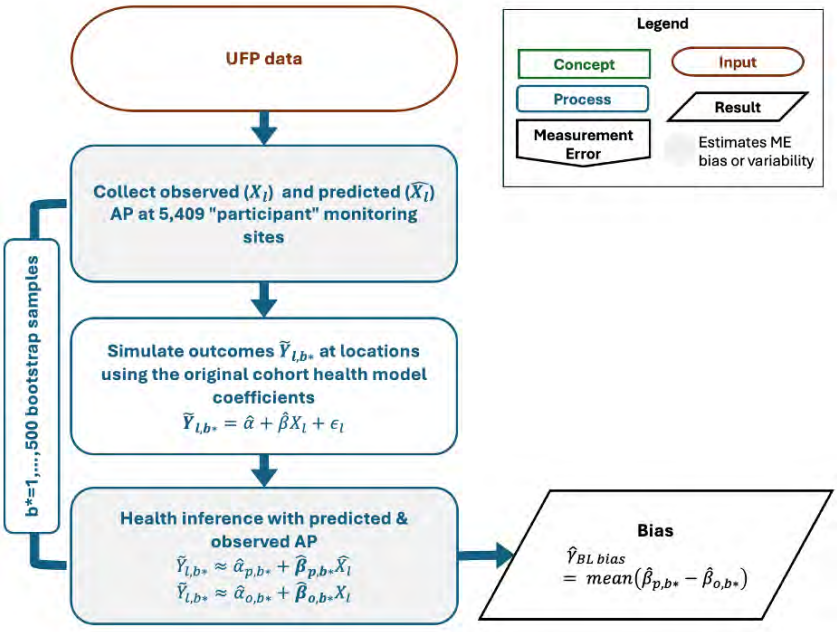


Figure 5.3. Overview of the estimation of health inference bias from Berkson-like measurement error. ME = measurement error; AP = air pollution.

exposures based on the simulated outcomes conditional on the observed exposures) provides an estimate of bias in the original uncorrected analysis.

Specifically, for each of $b^* = 1, \ldots, 500$ times, we

1. **Collect observed and predicted air pollution**: Sampled monitoring sites (n = 309) with replacement until we had 5,409 simulated "participant" locations (l).
2. **Simulate outcomes:** Simulated CASI-IRT outcomes at all 5,409 "participant" monitoring sites, $\tilde{Y}_{l,b*}$:

$$\tilde{Y}_{l,b*} = \hat{\alpha} + \hat{\beta} X_l + \epsilon_l \quad (5.4)$$

using the estimated intercept ($\hat{\alpha} \approx 0.0354$) and the UFP health association estimate ($\hat{\beta} = -0.0198$) from the original data fit (Equation 5.1), along with added statistical noise from this same model fit (Simulated as normally distributed errors with μ = 0, SD = 0.625, which was the SD of the residuals), and observed air pollution levels at a given location, AP_l.

3. **Health inference:** Fit least squares linear regression models for the association between CASI-IRT and predicted, , or observed, $\widehat{X}_l$, or observed, X_l, air pollution:

$$\tilde{Y}_{l,b*} \approx \hat{\alpha}_{p,b*} + \hat{\beta}_{p,b*} \widehat{X}_l \quad (5.5a)$$

$$\tilde{Y}_{l,b*} \approx \hat{\alpha}_{o,b*} + \hat{\beta}_{o,b*} X_l \quad (5.5b)$$

$\hat{\alpha}_{p,b*}$ and $\hat{\alpha}_{o,b*}$ are the estimated intercepts from predicted and observed air pollution, respectively. $\hat{\beta}_{p,b*}$ and $\hat{\beta}_{o,b*}$ are the estimated coefficients for the associations between CASI-IRT and predicted and observed air pollution, respectively, and are our parameters of interest.

We defined bias from Berkson-like measurement error ($\hat{\gamma}_{BL\ bias}$) as the difference in the mean health inference from the 500 parametric bootstrap approaches with the observed and predicted exposures:

$$\hat{\gamma}_{BL\ bias} = mean\left(\hat{\beta}_{p,b*} - \hat{\beta}_{o,b*}\right) \quad (5.6)$$

$$\hat{\gamma}_{BL\ percent\ bias} = \frac{\hat{\gamma}_{BL\ bias}}{\hat{\beta}_{bias\ corrected}} \times 100 \quad (5.7)$$

Health Inference Bias Correction and SE Calculation

We adjusted the all data health association estimate, $\hat{\beta}$ (Equation 5.1), for classical-like and Berkson-like measurement error absorbed in the estimate as shown in **Equation 5.8**. We used the SE estimated from the nonparametric approach described above in 4b (Health inference in the resampled cohort) to replace the asymptotic SE from the all data exposure model to account for the extra variability from classical-like and Berkson-like measurement error ($\widehat{SE}_\beta$ in Figure 5.2). Note that this approach misses the variability from doing the bias corrections, which were not feasible to conduct in this study. All analyses were conducted in R (v. 4.2.2) (R Core Team, 2023).

$$\hat{\beta}_{bias\ corrected} = \hat{\beta} - \hat{\gamma}_{CL\ bias} - \hat{\gamma}_{BL\ bias} \quad (5.8)$$

RESULTS

The estimated adjusted association between CASI-IRT and UFPs from the all data exposure model was –0.0198 (95% CI: –0.0356 to –0.0040; SE: 0.0081) per 1,900 pt/cm^3. **Table 5.1** summarizes the findings from the original (reference) analysis in the first row and the bootstrap analyses (nonparametric and parametric bootstrap) in the remaining rows. The first two columns summarize the information provided with respect to the type of measurement error present or being corrected in that row and the bootstrap approach description, which also refers to the relevant figure. The remaining two columns show the health estimate and its SE and estimated bias. The overall bias is given in the first row, the classical-like and

Table 5.1. Health Estimates (Betas) and Their Standard Errors for the Adjusted Association Between UFPs (per 1,900 pt/cm^3) and Cognitive Function from Parametric and Nonparametric Bootstrapped Approaches (sampled 500 times) using Stationary Roadside NanoScan data from the Seattle Mobile Monitoring Campaign

Measurement Error	Description	Health Estimate (SE)	Bias (percentage of bias-corrected health estimate)
Classical-like + Berkson-like present	Unadjusted reference from the all data exposure model	$\hat{\beta}$ = –0.0198 (SE: 0.0081)	$\hat{\gamma}_{CL\ bias} + \hat{\gamma}_{BL\ bias}$ = 0.0013 (6.3%)
Classical-like bias correction	Nonparametric approach: Different monitor locations and sampling times (Figure 5.2)	$\hat{\beta} - \hat{\gamma}_{CL\ bias}$ = –0.0188	$\hat{\gamma}_{CL\ bias}$ = 0.0010 (4.8%)
Classical-like + Berkson-like variability correction	Nonparametric approach: Nonparametric + different participants (Figure 5.2)	(SE_β = 0.0092)	
Berkson-like bias correction	Parametric approach: Predicted UFP at simulated cohort (i.e., bootstrapped monitoring) locations (Figure 5.3)	$\hat{\beta} - \hat{\gamma}_{BL\ bias}$ = –0.0195	$\hat{\gamma}_{BL\ bias}$ = 0.0003 (1.5%)
Corrected for Classical-like + Berkson-like	Adjusted reference	$\hat{\beta}_{bias\ corrected}$ = –0.0212 (SE_β: 0.0092)	

Berkson-like corrections are provided in rows two through four, and the bias-corrected estimate and SE are given in the last row. The effect of measurement error in this analysis was modest, with a mean estimated bias of 0.0010 (4.8% of the bias-corrected health estimate) from classical-like and 0.0003 (1.5%) from Berkson-like measurement error, with an approximately 13% increased variability of health estimates after classical-like and Berkson-like corrections (SE from 0.0081 to 0.0092). Altogether, the corrected, adjusted association between CASI-IRT and UFPs was –0.0212 (SE: 0.0092, 95% CI: -0.0392, -0.0031) per 1,900 pt/cm^3.

DISCUSSION

In this work, we investigated the impact of exposure measurement error on the estimated association between UFPs and cognitive function in the ACT cohort. We leveraged the methods developed by Szpiro and colleagues (Bergen et al. 2016; Keller et al. 2017; Szpiro et al. 2011a, 2011b; Szpiro and Paciorek 2013a, 2013b) to decompose the error into classical-like and Berkson-like components known to impact health inference bias and SE. In a nonparametric bootstrap approach, we captured health inference bias resulting from classical-like measurement error associated with the uncertainty in the exposure surface, and variability resulting from classical-like and Berkson-like measurement error associated with uncertainty in both the exposure surface and participants. In a parametric bootstrap approach, we captured health inference bias resulting from Berkson-like error associated with using predicted rather than observed exposure. We note that a critical assumption needed for the measurement error correction approach was met because the Seattle mobile monitoring campaign's stationary locations were selected to be spatially aligned, or spatially compatible with the target ACT cohort. Specifically, spatial compatibility is the assumption that the distribution function of participants' locations is the same as the distribution function of monitoring locations (referred to as $G(\times)$ and $H(\times)$, respectively, in Szpiro and Paciorek (2013a) and follow-on papers).

We found that the original health association between UFPs and cognitive function had been impacted by both classical-like and Berkson-like error to a small degree, with a total bias of 6% relative to the bias-corrected health estimate, and a slightly smaller variability relative to the error-corrected health estimate (13% smaller SE before rounding). We corrected the original health estimate from –0.020 (SE: 0.008) to –0.021 (0.009).

Putting these results into context, the health inferences were more meaningfully impacted by features of the mobile monitoring design as described in Chapter 4. In that work, we leveraged the same UFP measurements from the Seattle mobile monitoring campaign in this analysis (described in Chapter 3). We simulated less robust monitoring designs that involved subsampling the monitoring data following common field designs, including sampling fewer visits, more restricted seasons or hours, and collecting an unbalanced number of visits across sites. We developed annual average UFP exposure models with the resulting data and ran health analyses to estimate the adjusted association between 5-year UFP exposure and baseline cognitive function in the ACT cohort. We found that more restricted mobile monitoring designs generally produced poorer-performing exposure models. These, in turn, produced health inferences from confounder model 1 with median health associations that were approximately 25% to 60% different from the health inference produced by the all data (gold-standard reference) exposure model. These median health associations were all attenuated. Furthermore, as described in Chapter 4, in confounder model 2, we observed higher percentage differences, though these large percentages were driven by the small, nonstatistically significant health association estimate from the all data exposure model. Notably, the biases in confounder model 2 were both positive and negative, making them difficult to anticipate and correct. Note that the Chapter 4 health analyses were not corrected for measurement error as in this chapter.

It is noteworthy that the resulting bias-adjusted health estimate still contains sources of exposure measurement error that we did not address in this study. We used predicted ambient air pollution at participant residences as a surrogate for long-term personal exposures. This approach is common in large population settings, given the logistical challenges of collecting long-term personal exposures. Outdoor–indoor infiltration rates, time–activity patterns, and indoor sources of air pollution (e.g., cooking), however, are known to impact true exposures (Allen et al. 2003; Jung et al. 2011; Klepeis et al. 2001; Vardoulakis et al. 2020). This adds exposure assessment error, although we anticipate that the retiree ACT population spends a large portion of their time at home, and indoor and outdoor air pollution levels are often correlated. Moreover, we used exposure surfaces from 2019 measurements as surrogates for the long-term exposure surface. This adds some error to earlier time periods of assessments, although our prior work has shown that the spatial variability of UFPs is greater than the within-year temporal variability (Blanco et al. 2022), and that long-term UFP trends have likely been stable for over a decade (Blanco et al. 2024). Future work in this area to better quantify the degree to which these approaches impact health inferences would be a valuable contribution to the literature.

One of the limitations we faced in conducting this measurement error analysis was that we had to transfer exposure data to the Kaiser Permanente Washington Health Research Institute to be linked to participants' address histories. The ideal methodology for this analysis would have involved generating $500^2 = 250{,}000$ exposure surfaces based on two levels of nonparametric resampling (500 samples with replacement to calculate the variability, and 500 subsamples of each of those samples to separately calculate a bias correction for each of the samples). This would have permitted us to use an integrated approach to calculate the measurement-error-corrected health estimate and 95% CI because we could have calculated the standard deviation of bias-corrected health

estimates. Instead, we calculated the two types of bias and the measurement-error-corrected SE as semi-independent steps, using a much smaller number of exposure surfaces, which were only based on one level of resampling of monitoring locations. We note that exposure predictions generated by the various resamples were highly correlated, and the biases were very small, so we have no reason to believe that the more computationally intensive method would have produced substantially different results.

Overall, we found that while measurement error can impact the validity of the estimated health associations, the monitoring study's overall design can be a much more important consideration. Our findings hold in a setting where the spatial compatibility assumption is met. In this case, we can conclude that improving monitoring design should thus be prioritized, including in cases where error correction is not possible because nonideal monitoring sampling design can introduce bias (Chapter 4) that may be difficult to correct using available analytic or bias-correction methods. Further investigation is needed to extend these results to spatially misaligned exposure data, which are more common in practice.

CHAPTER 6: USING ON-ROAD DATA FROM MOBILE MONITORING STUDIES: EXPOSURE QUANTIFICATION, DESIGN, AND EPIDEMIOLOGICAL INFERENCE

Lead authors: Magali Blanco, Annie Doubleday, Lianne Sheppard

INTRODUCTION

Mobile monitoring, which is the collection of repeated short-term samples from an area of interest, is increasingly used to characterize air pollutants with high spatial variability (Kim et al. 2023). No standard monitoring protocols exist, however, and monitoring designs vary widely. Most campaigns collect on-road samples while driving a vehicle, with fewer collecting stationary roadside samples. The sampling approaches vary in their temporal and spatial coverage, with most sampling during weekday business hours and collecting approximately 1–40 (median 4) samples per location. While mobile measurements may result in more spatial coverage, stationary measurements may be more stable and representative of residential exposures. Several studies have reported elevated air pollution concentrations resulting from on-road measurements when compared to stationary measurements (Chambliss et al. 2020; Doubleday et al. 2023; Kerckhoffs et al. 2016, 2017; Minet et al. 2018). Approaches to address these concerns have been developed, although most campaigns have not yet implemented them. Methods include plume detection approaches (Kerckhoffs et al. 2016), those that leverage stationary, roadside measurements (Yuan et al. 2022), and others that leverage both stationary and multipollutant measurements to detect plumes as described in Chapter 3 and published work (Doubleday et al. 2023). Importantly, it is unclear how on-road monitoring choices may impact the resulting exposure assessment models, and whether some approaches may be better suited for epidemiological applications.

A few on-road monitoring studies have gained valuable insights into the number of measurements required to produce robust exposure assessment models. These have reported robust, well-performing exposure models relative to all data models when a subset of measurements are collected; for example, approximately 4–8 (Messier et al. 2018) or 5–15 (Kerckhoffs et al. 2024) repeat visits per road segment in Oakland, CA (depending on the modeling approach), and 10–12 repeat visits in Montreal, Canada (Hatzopoulou et al. 2017). No on-road mobile monitoring studies have investigated whether temporally or spatially balanced sampling is necessary for estimating unbiased long-term averages, the degree to which plume adjustment approaches may improve both balanced and unbalanced sampling campaigns, or how monitoring design may impact subsequent health inferences. Moreover, relying exclusively on exposure data and prediction quality, as prior studies have done, does not directly address the suitability of these campaigns for epidemiology.

We previously conducted a spatially and temporally extensive mobile monitoring campaign in the greater Seattle area to capture annual average traffic-related air pollution (TRAP) concentrations for the Adult Changes in Thought (ACT) study, as described in Chapter 3 and our published work (Blanco et al. 2022). The campaign collected stationary roadside and on-road measurements of various traffic-related air pollutants, including ultrafine particles (UFPs, measured as a particle number concentration [PNC] with a TSI P-Trak 8525 instrument [20–1,000 nm particles]), at locations representative of the cohort across all seasons, days of the week, and most hours of the day. Chapter 4 details some of the insights we gained from investigating sampling design with the stationary roadside measurements. We found that increasing the total number of stops (sites × visits per site) is strongly associated with increased exposure model performance, with diminishing returns for UFPs above approximately 309 sites and 12 visits per site (Blanco et al. 2023a). Furthermore, we documented the most accurate and consistent exposure models and health estimates when monitoring campaigns were designed to collect UFP measurements across three to four seasons, during all days and most hours, and in a spatially balanced way to ensure that all sites received the same number of visits.

In a follow-up study focused on the P-Trak data from the mobile monitoring campaign, we documented the differences between stationary and on-road data with the goal of better characterizing on-road data for epidemiology (Doubleday et al. 2023). We showed that on-road measurements approaching and departing from stationary locations were much more similar to the adjacent stationary roadside measurements than nonadjacent on-road measurements that were otherwise comparable in terms of their location characteristics. We further showed that on-road data were systematically higher on average than stationary roadside data due to on-road sources. Thus, we developed a plume adjustment approach to decrease the impact of on-road sources. This approach is summarized in Chapter 3. The adjusted on-road predictions were more consistent with predictions from the stationary data than the unadjusted on-road predictions.

The objective of this chapter is to evaluate how on-road mobile monitoring design features impact the resulting exposure model performances and subsequent health inferences in a case study of UFP exposures and cognitive function in ACT. We leverage our prior Seattle mobile monitoring campaign to assess designs with a smaller number of visits per road segment, restricted sampling during weekday business hours, spatially unbalanced sampling where some locations receive more visits than others, and the application of a plume adjustment to the resulting data. We provide guidance on key design features that on-road mobile monitoring campaigns can leverage to improve exposure model performance and support future epidemiological applications.

METHODS

Cohort and Cognitive Assessments

We assessed cognitive function in the ACT cohort using baseline CASI-IRT scores, as described in Chapters 3 and 4 (Table 3.1, Figure S3.3, Table 4.1).

Ultrafine Particle Data

We leveraged measurements from the Seattle mobile monitoring campaign described in Chapter 3 (Blanco et al. 2022, 2023b; Doubleday et al. 2023), which included both on-road and stationary measurements. We distilled these measurements as described in Chapter 3. This study focuses on the UFP measurements collected from the TSI P-Trak 8525, which measured the total PNC for 20–1,000 nm particles every second.

The stationary visit data were used to develop site annual averages and were treated as a gold-standard external validation dataset for the on-road prediction models, as described below and in Chapter 3.

Mobile Monitoring Sampling Designs

We simulated hypothetical sampling campaigns by sampling with replacement from the full dataset, using segment-level drive-pass median concentrations in the following process (**Figure 6.1**):

- **Visits per location**: For each road segment, we collected either 4 or 12 visits.
- **Spatial balance**:
    - Spatially unbalanced sampling: We collected visits using an unbalanced approach guided by a lognormal distribution (e.g., median: 4, IQR: 3–7.5, range: 1–28).
        - Road segments: We assumed some road segments were visited more frequently than others, possibly allowing for large discrepancies in the number of visits between neighboring segments.
        - Spatial clusters: We sampled certain areas more heavily by establishing spatial clusters, each composed of an average of 93 100-meter segments. Road segments within a cluster were sampled the same number of times. Specifically:

            Random clusters: Clusters were randomly assigned the number of visits.

            Sensible clusters: Clusters with higher all data annual average UFP concentrations received more visits.

            Nonsensible cluster: Clusters with lower all data annual average UFP concentrations received more visits.

            Road type: We sampled road segments based on the predominant road type, with larger, more heavily trafficked roads receiving more visits.
    - Spatially balanced sampling: We sampled visits using a balanced approach, ensuring that all locations received the same number of visits.
        - Road Segments: Each road segment was visited the same number of times. As before, we allowed for independent segment visits (e.g., different days), a common approach that does not account for the temporal correlation integrated into mobile monitoring, as neighboring segments are sampled together.
- **Time selection**:
    - All hours: Visits were conducted during all days and most hours.
    - Business hours: Visits were restricted to weekday business hours (9 a.m. to 5 p.m.).
- **Temporal adjustment**:
    - None: No temporal adjustment was applied to the data.
    - Fixed-site temporal adjustment: We followed a common approach of using a fixed-site background monitor to develop time-specific adjustment factors, defined as the difference between a measured short-term concentration (e.g., 1 hour in the study period) and the longer-term average concentration of interest (Eeftens et al. 2012; Hoek et al. 2002; Montagne et al. 2015; van Nunen et al. 2017). Chapter 4 details this approach for NanoScan PNC measurements.
- **Plume adjustment**:
    - None: We used the raw, possibly plume-impacted data.
    - Plume adjustment: We applied the plume adjustment method developed by Doubleday and colleagues (2023), which is summarized in Chapter 3.

We calculated annual average segment concentrations from each sampling campaign.

Ultrafine Particle Exposure Assessment

We used annual average segment UFP concentrations from each sampling campaign to develop universal kriging–partial least squares (UK-PLS) exposure prediction models, as described in Chapter 3. We evaluated each model at the 309 stationary roadside locations by comparing the predicted annual average UFP concentrations to estimated annual average concentrations using both traditional regression-based R^2 (R^2_{reg}) and mean squared error (MSE) -based R^2 (R^2_{MSE}) (Chapter 3 differentiates between these two measures of performance).

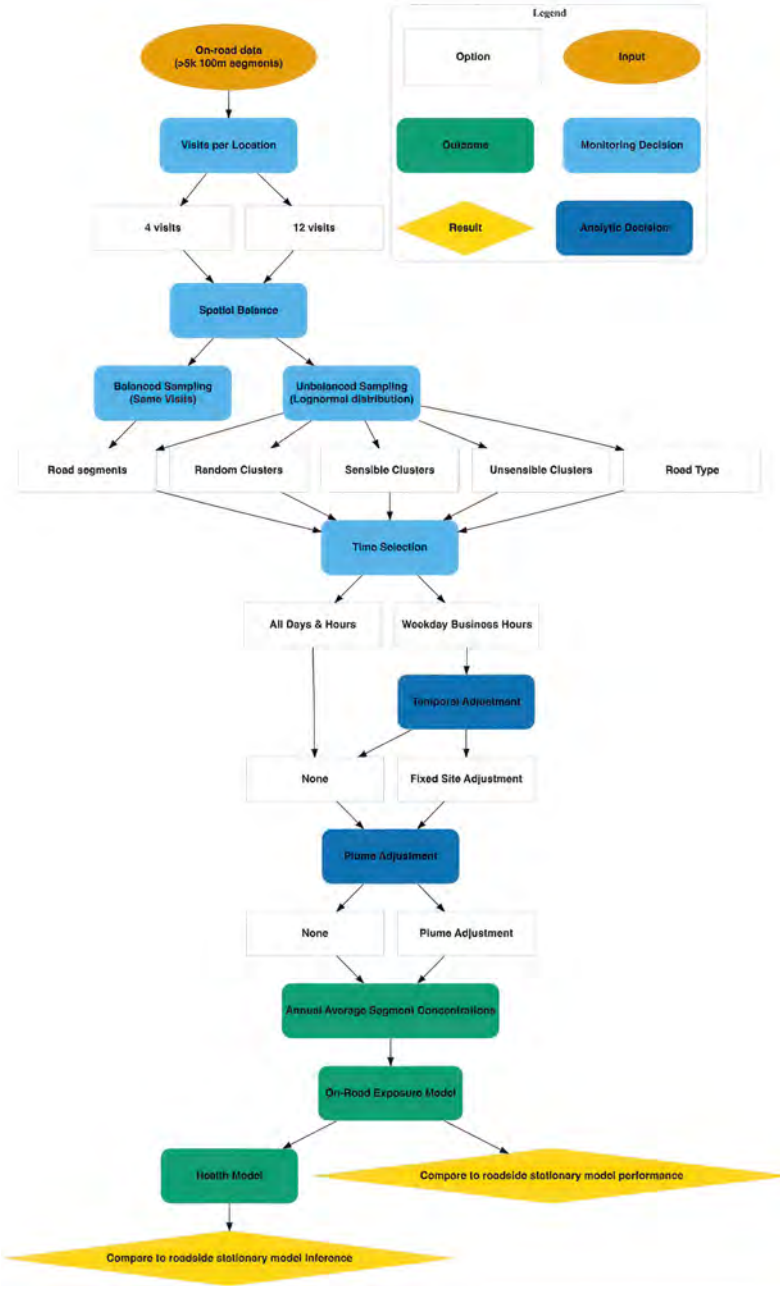


Figure 6.1. On-road mobile monitoring sampling designs. See Methods — Mobile Monitoring Sampling Designs for details. There are 30 campaigns for each sampling approach ($N = 24$ design paths × 30 campaigns = 720 total campaigns) in the main analyses.

We used each campaign model to predict time-weighted average UFP exposures for each ACT participant at baseline based on their prior 5-year residential history.

Inferential Analyses

As outlined in Chapter 3, we assessed the adjusted association between UFP exposure using each exposure model and baseline cognitive function (CASI-IRT), as given in Equation 3.11 for confounder models 1 and 2. We compared the estimated health parameter from each UFP exposure model to the health parameter estimated when using the all data stationary exposure model.

We conducted all analyses in R (v. 4.2.2) (R Core Team, 2023).

RESULTS

Cohort Characteristics

The cohort characteristics are described in Table 3.1 and Table 4.1.

Exposure Assessment and Model Performances

The all data stationary reference campaign produced a median (IQR) observed UFP concentration of 6,665 (5,667–

8,029) pt/cm³ and cross-validated predicted UFP concentration of 6,588 (5,817–7,898) pt/cm³ at the 309 stationary sites. Exposure models from on-road campaigns produced similar, slightly elevated predictions at the stationary locations, with plume adjustments slightly reducing these inflated predictions (Figure S6.2). Cohort predictions from the all data stationary UFP model had a median (IQR) of 7,008 (6,212–7,995). As before, on-road sampling campaigns generated UFP predictions that were slightly more elevated, while plume adjustment slightly reduced these inflated predictions (Figure S6.3).

The all data stationary P-Trak UFP model had a cross-validated R^2_{MSE} of 0.77. On-road campaigns had similar model performances in balanced campaigns with either four or 12 visits per location (**Figure 6.2**). Four and 12-visit campaigns produced similar results, with 12-visit campaigns producing more stable estimates across campaigns. Business-hours campaigns produced the worst-performing exposure models (~25% of the all data reference). Temporally adjusting these campaigns produced more variable results and, at times, improved exposure model performances. Plume adjustment improved business-hours exposure models to a larger degree than temporal adjustment. Temporally adjusting plume-adjusted models produced relatively noisier models. Plume adjusted all-hours campaigns produced the best-performing exposure models.

Spatial balance generally had a minimal impact on exposure model performances. Campaigns with unbalanced road segment sampling (vs at the cluster level) produced similar results (not shown). Figure S6.4 shows regression-based R^2_{reg}, a more common metric reported in the field. R^2_{reg} is always higher than R^2_{MSE}, and it differentiates between campaign performances less well than R^2_{MSE}.

Inferential Analyses

Using the reference all data stationary exposure model, the adjusted mean baseline CASI-IRT score differed by –0.021 (95% confidence interval [CI]: –0.039 to –0.003) in confounder model 1 and 0.007 (–0.013 to 0.027) in confounder model 2 per each 1,900 pt/cm³. In confounder model 1, compared to the reference estimate, health associations from on-road UFP exposure models were most similar for campaigns that sampled during all hours, with plume adjustment marginally improving results (**Figure 6.3**). Median health estimates were attenuated (i.e., closer to 0) by approximately 50% to 75% when on-road monitoring campaigns were restricted to business hours. Spatially balanced sampling had a relatively small impact on the estimated health inferences. The benefit of monitoring design was less clear in confounder model 2, where there was no association detected between UFPs and cognitive function. Results were similar for 4- and 12-visit campaigns, with 4-visit campaigns producing noisier estimates (Figure S6.5).

DISCUSSION AND CONCLUSIONS

Mobile monitoring campaigns are increasingly used to characterize air quality in an area. The monitoring designs vary, however, and it is unclear what the key design features are for campaigns aiming to develop exposure assessment models for epidemiological applications. Only a few other studies have investigated monitoring design, and these have mostly focused on the amount of data collected (Blanco et al. 2023b; Hatzopoulou et al. 2017; Kerckhoffs et al. 2024; Messier et al. 2018).

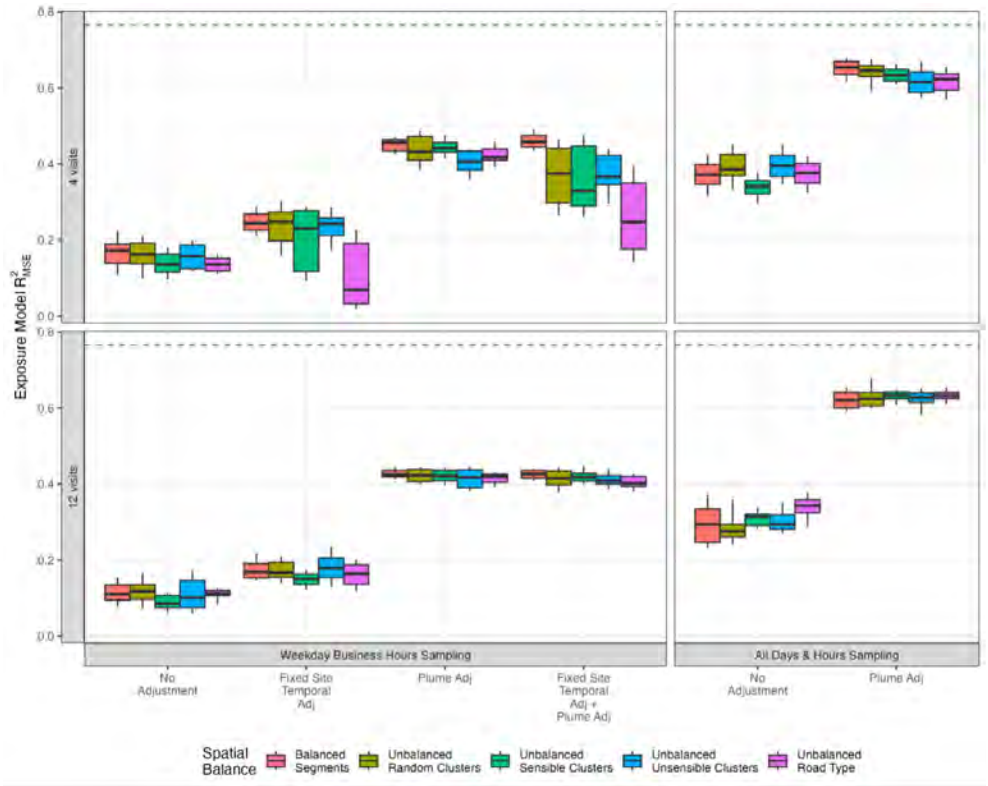


Figure 6.2. Out-of-sample UFP (pt/cm³) exposure model performances for on-road campaigns (*N* = 30 campaigns per combination (i.e., box plot). R_{MSE} is based on a comparison of the predicted PNC at 309 stationary locations and the annual average site estimate from stationary roadside measures. UFP models are for total particles (20–1,000 nm) from the unscreened P-Trak instrument and the Seattle mobile monitoring campaign. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles. The dashed line indicates the R^2 from the reference all data stationary model, which is 0.77.

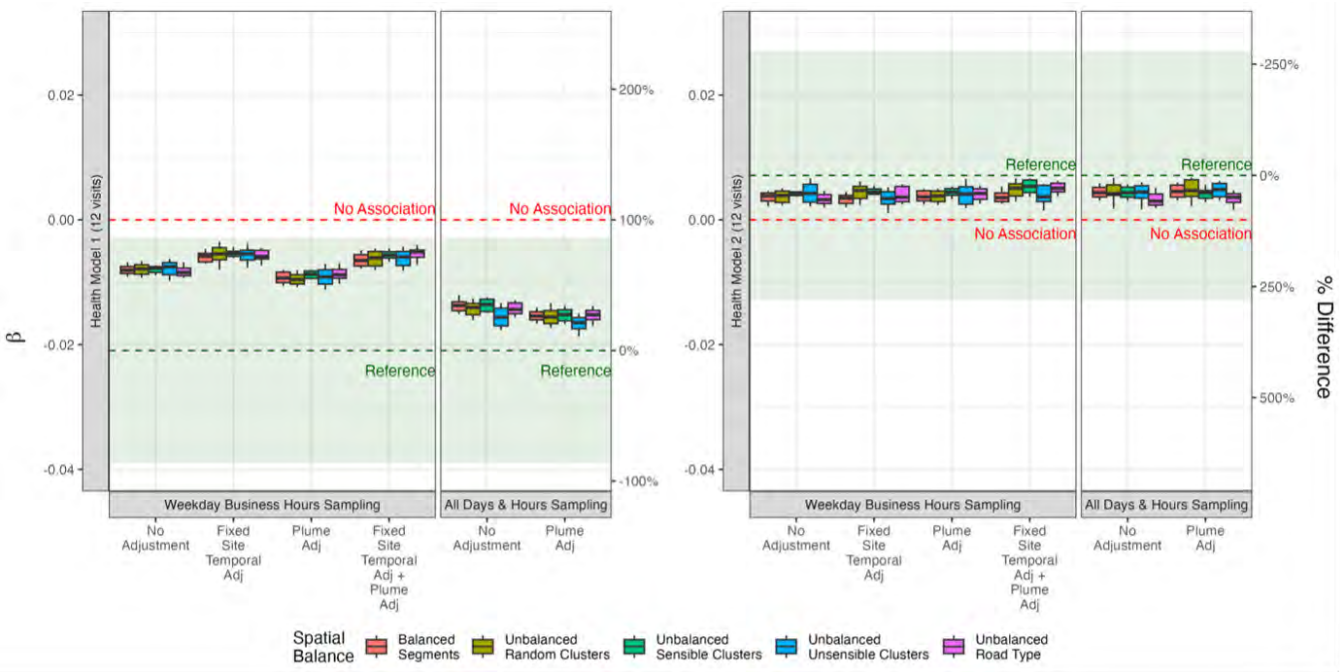


Figure 6.3. Estimated association between UFPs (1,900 pt/cm³) and cognitive function at baseline using confounder models 1 and 2. UFP exposures are predicted from on-road monitoring campaigns using the 12-visit campaigns. The dashed green lines and shaded areas indicate the estimated point and 95% CIs from the all data roadside exposure model, which are –0.021 (95% CI: –0.039 to –0.003) in confounder model 1 and 0.007 (95% CI: –0.013 to 0.027) in confounder model 2. The dashed red line indicates no association. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

We leveraged an extensive mobile monitoring campaign and a long-standing prospective cohort study to gain insight into the impact of common on-road monitoring and analytic decisions on both exposure prediction performance and health inference. The reference all data stationary campaign produced well-performing exposure models (R^2_{MSE} = 0.77). In confounder model 1, which utilized the roadside reference P-Trak exposure model, the adjusted mean baseline CASI-IRT score was reduced by 0.021 (95% CI: –0.039 to –0.003) per each 1,900 pt/cm³. Conversely, confounder model 2 produced an estimate that barely overlapped with confounder model 1, was not statistically significant (0.007 [95% CI: –0.013 to 0.027]), and was not in the hypothesized direction (Figure 6.3). Plume-adjusted all-hours designs yielded results most consistent with the confounder model 1 all data association estimates, which reported a stronger association, whereas other designs produced more attenuated results. Temporal and plume adjustments typically improved exposure model performance statistics to varying degrees (Figure 6.2), but they did not meaningfully improve health inference estimates (Figure 6.3), suggesting that temporally balanced sampling is critical. We saw similar patterns when evaluating campaigns with four and 12 visits, with four-visit designs producing slightly attenuated, more variable results. This may be a result of the increased temporal representation captured by a larger number of visits. Collecting spatially balanced samples had a relatively small effect on the results, indicating that some degree of spatial imbalance (e.g., due to logistical constraints) may not meaningfully affect the resulting conclusions. In confounder model 2, where the estimated associations were smaller and not statistically significant, there was considerable overlap across designs.

There is some overlap in the findings from this on-road campaign study and our prior work with stationary roadside monitoring (Chapter 4). Stationary data are more precise than mobile data in terms of the accuracy of the geographic location and the duration of sampling. They are also less likely to be contaminated by on-road sources, making them potentially more representative of residential cohort exposures (Doubleday et al. 2023). They often represent fewer monitoring locations than mobile data, however (19 times fewer in this case, with 309 stops vs. 5,878 road segments), and they are less common in the literature. Despite these differences, we previously also found that collecting temporally unrestricted visits had the largest impact on exposure and health models. We previously discussed these results in the context of temporal misalignment of the data relative to the longer-term exposures of interest (Chapter 4). Designs with temporally unbalanced or incomplete exposure data result in higher levels of Berkson-like error than other designs and are associated with bias in the resulting health association (Szpiro et al. 2011b; Szpiro and Paciorek, 2013a). We previously also found that the resulting exposure model performances and health inferences are worsened when collecting fewer visits

per location, although this impact is smaller than the impact from sampling during restricted days and hours (i.e., business or rush hours).

In our prior work with stationary data, we reported health inferences that were slightly more variable and biased when locations with more concentration variability were prioritized (visited more) or deprioritized (visited less). In these analyses using on-road data, however, we observed no clear pattern. This may be due to the proximity of neighboring segments when compared to stationary locations, which may make models appear to perform better. These findings may reflect the ability of geostatistical models to borrow information from nearby locations. Future analyses should consider route structure to evaluate the impact of temporal correlation and visit frequency in realized on-road monitoring campaigns. Furthermore, work characterizing the degree to which geostatistical models alleviate unbalanced sampling may be valuable.

Interestingly, adjusting on-road data for plume concentrations improved exposure models, but minimally improved health inferences, and this improvement was not sufficient to counteract the attenuating effects of business hours sampling in confounder model 1. Similarly, temporally adjusting business hours designs at times improved exposure models, but their impacts on health estimates were inconsistent. In our prior work using stationary data, we also found that temporally adjusting data produced inconsistent results. These findings suggest that collecting temporally balanced data could be a reliable approach for both roadside and on-road campaigns, as existing temporal adjustment methods may not adequately address the inconsistent effects of temporally restricted designs. Future work to identify more effective adjustment approaches would be valuable.

There are likely other monitoring design features that are important to consider in future work that we did not address in this study. The overlap in some of the findings between on-road campaigns discussed in this chapter and roadside campaigns (e.g., all hours, more visits), discussed both in Chapter 4 and by Blanco and colleagues (2023a), indicates that some of our earlier conclusions may also be relevant for on-road campaigns. For example, while we focused this study on UFPs — a pollutant that is commonly measured with mobile monitoring — we previously found similar trends across pollutants, with some variability based on a pollutant's spatial and temporal variability (Blanco et al. 2023a). Moreover, while we did not investigate campaign duration in this study, we previously found that campaigns with samples collected across three to four seasons produced the most robust annual average exposure assessment models and health inferences. Nonetheless, there may be sampling strategies beyond location concentration or variability (the latter based on stationary data) that may allow for less sampling at some locations without negatively impacting exposure predictions or epidemiological inferences. For example, selectively adding more visits to poorly predicted locations has not been evaluated. Finally, results will vary for different health outcomes and association estimates. In this study, based on the sizes of the coefficient estimates, we found a more striking impact on health inferences by exposure design when considering confounder model 1, and smaller and less consistent impacts on inferences across designs when considering confounder model 2. More work is needed to investigate other relevant hypotheses.

There are a few features of this analysis that should be considered for the generalizability of our findings. The use of real data from the long-standing ACT cohort is a strength of this study. These data naturally incorporate aspects that might not be included in a simulation study, thus strengthening the real-world implications of our findings. A feature of this analysis was the use of 2019 air pollution measurements. The use of these data to assess longer-term exposures adds some degree of exposure assessment error, particularly to earlier time periods. Nonetheless, our analyses should have still captured the impacts of mobile monitoring design *changes*. Moreover, it is notable that we evaluated exposure model performances at out-of-sample monitoring locations where we had high confidence in the accuracy of the annual average site concentrations. These data are not generally available in practice, resulting in analyses that compare predicted exposures against the collected campaign observations. We have previously shown that this assessment approach can produce unclear or misguided results when certain common designs (e.g., business hours) are implemented (Blanco et al. 2023b). We also primarily reported R^2_{MSE} in these analyses, which is less commonly reported than R^2_{reg} and, as we observed, performs worse. It is also notable that R^2_{reg} was less able to differentiate between different monitoring designs.

In conclusion, on-road monitoring campaign design can be leveraged to improve exposure models, although its impact on health inferences may be subtle, with resulting estimates potentially attenuated. Exposure models benefit from data collected using temporally balanced and unrestricted approaches. While plume and temporal adjustments can sometimes improve exposure models, they can also introduce statistical noise. Temporally unrestricted on-road campaigns may yield more accurate health inferences than temporally restricted on-road campaigns. Increasing the number of visits per location generally leads to more consistent findings across campaigns. These design principles can be applied by on-road monitoring campaigns to support epidemiological applications.

CHAPTER 7: ADDED VALUE OF LOW-COST SENSORS AND OTHER NONREGULATORY MONITORING DATA FOR EXPOSURE PREDICTION AND HEALTH INFERENCE*

Lead authors: Jianzhao Bi, Chris Zuidema, Dustin Burnham, Magali Blanco, Lianne Sheppard

INTRODUCTION

Historically, associations between ambient air pollution and health effects have been studied at the ecological or regional level using regulatory agency monitoring data, such as data collected by the US Environmental Protection Agency (US EPA) Air Quality System. For air pollutants such as fine particulate matter ($PM_{2.5}$; particles that are 2.5 µm or less in aerodynamic diameter) and nitrogen dioxide (NO_2), regulatory measurements are typically taken with Federal Reference Monitors (FRMs) or Federal Equivalent Methods (FEMs) and provide high-quality pollutant data, but the monitoring locations have sparse geographical distribution. This geographic sparseness poses challenges for environmental epidemiological studies of air pollution, as pollutant measurements may not adequately characterize individual-level spatial exposure variability. Past exposure assessment approaches relied on intercity differences (Dockery et al. 1993) or nearest monitor assignment (Miller et al. 2007); however, these approaches have limitations and may lead to low exposure variability and exposure measurement error or misclassification, and may obscure relationships between air pollution exposure and disease outcomes, or lead to spurious associations. Determining the most feasible and cost-effective approaches to improving $PM_{2.5}$ and NO_2 exposure assessment will considerably enhance the quality of epidemiological inferences about their health effects.

Recent advances in microprocessor platforms, open-source software, and low-cost sensors have enabled the proliferation of low-cost sensor networks to measure air pollution. These networks, both custom-made and commercialized, detect particulate matter, hazardous gases, or both, although the availability of sensors tends to reflect pollutants of regulatory importance (e.g., $PM_{2.5}$, PM_{10}, carbon monoxide, NO_2, and ozone). Examples of sensor networks described in the literature are growing (Datta et al. 2020; English et al. 2017; Gao et al. 2015; Heimann et al. 2015; Ikram et al. 2012; Jiang et al. 2016; Jiao et al. 2016; Malings et al. 2019; Mead et al. 2013; Moltchanov et al. 2015; Zimmerman et al. 2018), and offer insights into the calibration, deployment, practical issues, and quality of data that low-cost sensor networks provide. Compared to reference-grade FRMs and FEMs, low-cost sensors may provide data of lower quality but are orders of magnitude less expensive, which facilitates their deployment in geographically dense networks. Several researchers have suggested that a key potential of low-cost sensor networks is to supplement traditional regulatory monitoring (Kumar et al. 2015), particularly in exposure assessment for epidemiological studies (Jerrett et al. 2017).

The use of low-cost sensors, however, may be coupled with high personnel costs with regard to sensor deployment, construction, and data management. It is unclear what type of monitoring design is spatially and temporally optimal, such that a minimum number of monitoring locations and timepoints can maximize the quality of $PM_{2.5}$ and NO_2 exposure modeling at specific locations of interest (e.g., residential locations of an epidemiological cohort). Some previous studies investigated the sensitivity of $PM_{2.5}$ modeling to the spatial and temporal coverage of the pollutant's measurements (Bi et al. 2022a; Geng et al. 2018; Huang et al. 2019), implying that spatiotemporally denser measurements can improve the modeling quality. Nonetheless, very few, if any, efforts have been made to evaluate the impacts of monitoring design on $PM_{2.5}$ and NO_2 modeling in the context of using low-cost sensors to improve exposure assessment for epidemiological inference.

This work had two objectives. The first was to determine whether low-cost sensors can improve the predictions of $PM_{2.5}$ and NO_2 in spatiotemporal air pollution models, particularly at residential locations, as these are the locations of interest for epidemiology. The second was to determine the impact of the resulting predictions on health association estimates. We addressed these objectives by considering different features of low-cost sensor designs. Because the available data and challenges in our study associated with modeling $PM_{2.5}$ and NO_2 were different, some aspects of our approach varied by pollutant. In this chapter, we rely extensively on the Chapter 3 data description and its presentation of the spatiotemporal modeling approach to provide much of the essential background. Within the methods and results subsections, we first present $PM_{2.5,}$ followed by NO_2. The $PM_{2.5}$ methods section focuses on the low-cost sensor monitoring designs we evaluated. The results section covers both the impact on $PM_{2.5}$ predictions for different amounts of short-term low-cost sensor data included, as well as the impact on inference about cognitive function across low-cost sensor designs. The $PM_{2.5}$ methods and results are a distillation of the recently published $PM_{2.5}$ exposure modeling paper (Bi et al. 2024); it additionally covers the impact of $PM_{2.5}$ prediction models on $PM_{2.5}$ health association estimates. The NO_2 campaign and analysis presentation mirrors that of the $PM_{2.5}$ presentation. Similar to the $PM_{2.5}$ presentation, additional details are in the published NO_2 exposure modeling paper (Zuidema et al. 2024). The discussion and conclusion section addresses each pollutant separately before considering the combined lessons learned from the NO_2 and $PM_{2.5}$ findings.

*Parts of this section have been reprinted (adapted) with permission from (1) (Bi et al. 2024) Copyright 2024 Elsevier, and (2) (Zuidema et al. 2024) Copyright 2024 Nature.

METHODS

Introduction and Brief Data Overview

Chapter 3 gives a brief overview of the types of data used for this analysis and presents the general spatiotemporal modeling approach for producing predictions from these data. Chapter 3 also provides an overview of the ACT cohort, the CASI-IRT cognitive function outcome, and the model used for the health association estimates.

Table S7.1 gives a summary of all the available $PM_{2.5}$ data used for the primary modeling evaluations described in this chapter (which focus on the 2010–2020 time period), including the type of monitor, the number of monitors, the number of measurements, the time period each monitoring type was active for, and some descriptive statistics by monitor type. Figure S7.1 shows plots for the monitoring data for each monitor, and Figure 3.3 is a map with the locations of the monitors. Table S7.2 and Figure S7.2 give the same summaries for the 1978–2021 time period. These are the data used for the health analyses.

Table S7.5 gives a summary of all the available NO_2 data, including the type of monitor, the number of monitors, and some descriptive statistics by monitor type. The supplementary monitoring data descriptions at the beginning of Chapter 7's Additional Materials give the time period for the supplementary monitoring data. Figure S7.3 shows plots for the monitoring data for each monitor, and Figure S7.4 is a map of the monitoring region with the locations of the monitors.

Methods — $PM_{2.5}$ Modeling

To evaluate the model performance for various low-cost sensor campaign designs, we considered reduced temporal coverage as well as reduced spatial coverage. We conducted our primary modeling evaluations over the period from January 2010 to September 2020. We selected this time period because the short-term low-cost sensor campaigns were carried out between April 2017 and September 2020, and we determined that a 10-year period for the long-term $PM_{2.5}$ exposure assessment was sufficient (Note: We only used low-cost sensor data between June 2017 and May 2019 in our analyses due to the availability of two-period measurements). We also report a similar but less comprehensive analysis using all $PM_{2.5}$ data beginning in 1978 and ending in 2021 (the full time period dataset). Our primary spatiotemporal model had all reference-grade and two-period low-cost sensor measurements included (*N* of monitors = 160, *N* of measurements = 6,688; Table S7.1) as our "*All data*" comparator. Refer to Chapter 3 for details on the modeling approach.

We examined five variants of temporally reduced models, and all of them used all the data from all reference-grade monitors. Each used different amounts of temporally reduced low-cost sensors, including the following:

1. **"First" —** Only the first measurement from each low-cost sensor within its monitoring period;
2. **"Last" —** Only the last measurement from each low-cost sensor within its monitoring period
3. **"Separate" —** Both the first and last measurements from each low-cost sensor within its monitoring period
4. **"Adjacent"—** The first two temporally consecutive measurements from each low-cost sensor
5. **"No" —** No low-cost sensor measurements

We also examined variants of models with three sets of spatially reduced low-cost sensors incorporated:

1. The first set, with five models, included measurements from randomly selected subsets (10%, 25%, 50%, 75%, and 90%) of low-cost sensors.
2. The second set, with four models, included measurements from low-cost sensors falling into the quadrants of long or short distance to major roads (defined as A1 and A2 roads based on the Census Feature Class Code classification; cutoff at the median distance of 1,086.5 meters) combined with high or low population density (defined as ambient population density within a radius of 3,000 meters of a location of interest; cutoff at the median value of 48,508.5 people).
3. The third set, with seven models, included measurements from low-cost sensors within one of the seven ACT-AP participant recruitment regions.

We describe the results based on an out-of-sample validation scheme that is a combination of cross-validation (CV) and external validation for the subset of two-period low-cost sensor locations (i.e., the low-cost sensors used in the models; *N* of monitors = 82, *N* of measurements = 502; Table S7.1). We refer to this validation as the "*combined validation*" hereinafter. The CV had a 10-fold leave-monitors-out design. For models with only a subset of low-cost sensor measurements (i.e., temporally and spatially reduced designs), we conducted CV for measurements included in the model and external validation for the remaining measurements. Importantly, for temporally reduced designs, because all 82 low-cost sensors were always included in the models, the external validation followed the CV grouping and was conducted in a leave-monitors-out manner based on the 10 CV groups; for spatially reduced designs, in contrast, the external validation was a pure independent validation for low-cost sensors that were not included in the model (i.e., the external validation was conducted once based on the model with all other low-cost sensors). While the 10 CV monitor groups were different across designs due to different low-cost sensors included, the combination of CV and external validations provided a consistent validation dataset (i.e., measurements from all reference-grade monitors and low-cost sensors) for all designs. Thus, this combined validation method ensured a fair comparison among designs.

We validated the models in two ways: *spatial validation* based on monitor-specific, temporally averaged $PM_{2.5}$ concentrations throughout the whole monitoring period, and *spatiotemporal validation* based on individual 2-week $PM_{2.5}$

concentrations. Note that some temporal variability remained in the spatial validation given the short duration of low-cost sensor sampling (Figure S7.1). We used MSE-based R^2 and root mean squared error (RMSE) to evaluate the model's performance (Keller et al. 2015; Young et al. 2016).

Finally, we predicted 5-year average exposures at ACT cohort residential locations for the 5 years prior to the baseline visit. We used these in Equation 3.11 using confounder model 1 to estimate associations with cognitive function, as quantified by CASI-IRT, for each of these exposures. In these analyses, we evaluated the various prediction model results using either or both time-varying spatiotemporal predictions and purely spatial predictions for the year 2019. For results reported based on both the spatiotemporal and spatial predictions, we used the full monitoring period dataset; for results reported for only the spatial predictions, we used the 2010–2020 dataset. Participants were restricted to those residing at residential locations within the mobile monitoring region (red locations in Figure S3.1). See the description in Chapter 3 for details of the cohort, outcome, and inferential model. Table S7.4 matches the descriptions of the models listed above to the labels in the results figure and indicates the time period (1978–2021 or 2010–2020) of the predictions for the health association results reported in **Figure 7.1**.

Methods — NO_2 Modeling

We evaluated the model performance for NO_2 concentration predictions without low-cost sensor data included in the model development. Given the three different data sources (agency, snapshot, and the Remote Air Data campaign), we conducted separate CV evaluations focused on each of the sources. For the agency data, we used a leave-one-site-out approach. For the ACT-AP snapshot data, we considered a leave-one-cluster-out approach as well as a ninefold CV with equally sized groups and, finally, a leave-one-campaign-out assessment. For the low-cost sensor data, we considered a 10-fold CV. Note that for the assessment of the low-cost sensor data, when they were omitted from the spatiotemporal model, the performance evaluation was a pure out-of-sample evaluation (rather than a CV).

The general approach we followed was that for each candidate spatiotemporal NO_2 model, we removed NO_2 observations at locations according to the procedures described below, fit the model with locations removed, predicted NO_2 concentrations for those locations, and repeated until predictions were made for all locations.

The agency leave-one-site-out cross-validation, carried out on 11 sites, describes how the model performs over longer

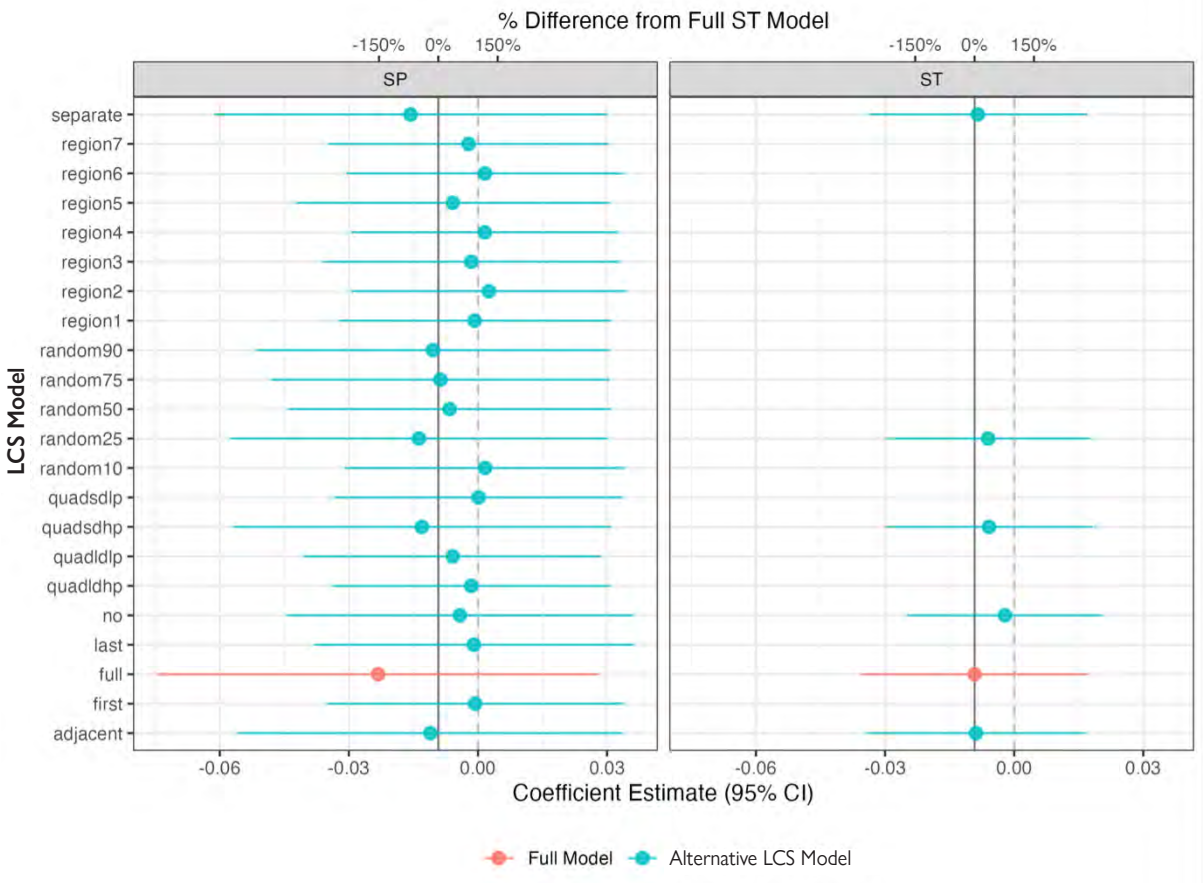


Figure 7.1. Estimated association (95% CI) between $PM_{2.5}$ (1 µm/m³) and cognitive function at baseline for different exposure models. The associations are adjusted for age, calendar year, sex, and education (confounder model 1). See Chapter 3 for the analysis approach and Table S7.4 for model descriptions. LCS = low-cost sensor; SP = spatial model; ST = spatiotemporal model.

periods of time, compared to other data sources, which comprise a shorter amount of time. Agency measurements are longer-term than other data sources available. While agency measurements are of the highest quality and represent "true" concentrations, they do not reflect locations of interest within the study region (i.e., residential locations of ACT-AP study participants) because their monitoring goals are generally to capture regional concentrations.

The primary cross-validation conducted on the ACT-AP snapshot measurements was leave-one-sampling-cluster-out, resulting in 26 cross-validation groups (17 groups with clusters and nine groups with single locations). We also conducted ninefold cross-validation, by site, with similarly sized groups where we accounted for clustering among measurements, resulting in cross-validation groups of 9–13 sampling locations. We kept sampling clusters in the same cross-validation groups to enable fairer evaluation because of the correlation among geographically close measurement locations. With the large amount of spatial variability across ACT-AP snapshot locations, the interurban and regional performance of the NO_2 model can be assessed. We also investigated leaving one season (i.e., sampling campaign) of three out to assess if the model predictions were sensitive to season.

We performed 10-fold cross-validation, by site, of the low-cost sensor measurements, summarizing results at ACT-AP participant, community, and volunteer residential locations where low-cost sensors were sited to provide a metric of the NO_2 model performance at residential locations (N = 111). After choosing the best-performing model from the candidates, we repeated the performance evaluation on a dataset omitting low-cost sensor data to assess the low-cost sensor data's contribution to the final model's performance. To create a similar performance metric for the final spatiotemporal model without low-cost sensor data, we generated out-of-sample predictions at low-cost sensor locations and compared those predictions to low-cost sensor measurements.

For health analyses, we used an approach similar to the one described for $PM_{2.5}$. We applied confounder model 1 (Equation 3.11) to the 5-year average exposures prior to each participant's baseline visit, according to their address history during that 5-year period. We considered exposure models with and without low-cost sensors included. We present spatiotemporal results for time-varying exposures and 2019 spatial results for time-constant exposures (see Chapter 3 for details). Because the earliest NO_2 exposure data were from 1996, we assumed location-specific predicted exposures before 1996 were equal to exposures in 1996.

RESULTS — $PM_{2.5}$ EXPOSURE AND HEALTH INFERENCE

Summary Statistics — 2010–2020 Time Period

As shown in Table S7.1, the reference-grade monitors had mean 2-week $PM_{2.5}$ concentrations around 5–7 μg/m³, with a minimum of 0.4 μg/m³ and a maximum of 100.8 μg/m³. The low-cost sensors had similar mean 2-week concentrations around 6 μg/m³ with smaller concentration ranges from 2.6 to 17.0 μg/m³.

Temporally Reduced Designs — 2010–2020 Time Period

Table 7.1 shows the combined validation performance of the $PM_{2.5}$ exposure models built on temporally reduced monitoring designs. All monitoring designs resulted in decreased spatiotemporal performance compared to the *All data* model (MSE R^2 [RMSE] = 0.84 [0.88 μg/m³]). The spatiotemporal performance was the worst when low-cost sensor measurements were entirely excluded in the *No* model (MSE R^2 [RMSE] = 0.69 [1.20 μg/m³]).

Table 7.1. Combined Validation Performance for Prediction of $PM_{2.5}$ at the Two-Period Low-Cost Sensor Locations for Temporally Reduced Designs

Model	*N* of Reference-Grade Measurements for Modeling (*N* of sensors)	*N* of Low-Cost Measurements for Modeling (*N* of sensors)	*N* of Combined Validation Measurements (*N* of sensors)	Spatiotemporal Validation		Spatial Validation	
				MSE R^2	RMSE (μg/m³)	MSE R^2	RMSE (μg/m³)
All data	6,186 (78)	502 (82)	502 (82)	0.84	0.88	0.57	0.57
First		82 (82)		0.76	1.06	0.37	0.69
Last		82 (82)		0.77	1.04	0.43	0.66
Separate		164 (82)		0.79	0.99	0.48	0.63
Adjacent		164 (82)		0.77	1.03	0.41	0.67
No		0 (0)		0.69	1.20	0.11	0.82

MSE R^2 = mean squared error coefficient of determination; RMSE = root mean squared error.

All versions of the model that included all low-cost sensor locations but with reduced temporal coverage underperformed the *All data* model, but outperformed the *No* model. Among these, the model that had the greatest time interval between the two low-cost sensor measurements (*Separate*, i.e., with the first and last measurements) had the highest spatiotemporal performance across designs (MSE R^2 [RMSE] = 0.79 [0.99 μg/m^3]). In comparison, the two models with a single (first or last) measurement had lower spatiotemporal performance (MSE R^2 [RMSE] around 0.77 [1.05 μg/m^3]); and the model with two consecutive measurements (*Adjacent*) had similar spatiotemporal performance (MSE R^2 [RMSE] = 0.77 [1.03 μg/m^3]) as the single-measurement models.

Likewise, all monitoring designs had worse spatial performance compared to the *All data* model, which had an RMSE of 0.57 μg/m^3. The *No* model had the worst spatial performance with an increased RMSE of 0.82 μg/m^3. Among the remaining models, the performance of the *Separate* model was closest to that of the *All data* model, with an RMSE of 0.63. It should be noted that although the MSE R^2 for spatial validation was below 0.6, this metric was extremely sensitive to a few samples that deviated from the 1:1 line. For full appreciation of the models' performance, note that the low RMSEs suggested good performance.

Figure 7.2 shows the mean $PM_{2.5}$ predictions between June 2017 and May 2019 (i.e., the period of the short-term low-cost sensor campaigns) across designs. The *Separate* model had the smallest differences from the *All data* predictions, while the model without low-cost sensor measurements had the largest deviations. When compared to the *All data* predictions, the *Adjacent* model tended to overestimate $PM_{2.5}$ in urban areas and underestimate it in suburban and natural areas. Noticeable differences in estimated $PM_{2.5}$ concentrations were also observed for the models with only one low-cost sensor measurement, either the first or the last, as compared to the *All data* model.

Spatially Reduced Designs — 2010–2020 Time Period

Table 7.2 shows the combined validation performance of the $PM_{2.5}$ exposure models built on spatially reduced monitoring designs. When more randomly selected low-cost sensors were included in the model, both spatiotemporal and spatial performance improved, with larger R^2 values and smaller RMSEs. Notably, when a model included 90% of the low-cost sensors (74 out of 82), its spatiotemporal and spatial performances were nearly identical to the *All data* performances.

Low-cost sensors with shorter distances to major roads contributed more to model improvement compared to the model with road-distant low-cost sensors, while population density was a less important factor. Also, the model improvement did not necessarily correlate with the number of low-cost sensors included, as the "road-proximate and low-population" model with 15 low-cost sensors outperformed the "road-distant and low-population" model with 26 low-cost sensors.

As expected, incorporating low-cost sensors from a single participant recruitment region resulted in worse spatiotemporal and spatial performance as compared to using all low-cost sensors.

Model Performance for the 1978–2021 Time Period

For completeness, in Chapter 7's Additional Materials, we include two tables and one figure for the 1978–2021 time period: the descriptive statistics for the data from the entire time period (Table S7.2), the data availability plot for the entire time period (Figure S7.2), and the combined validation performance for a selected subset of the models discussed above (Table S7.3). The full time period results reported in Table S7.3 are generally consistent with the results focused on the 2010–2020 time period.

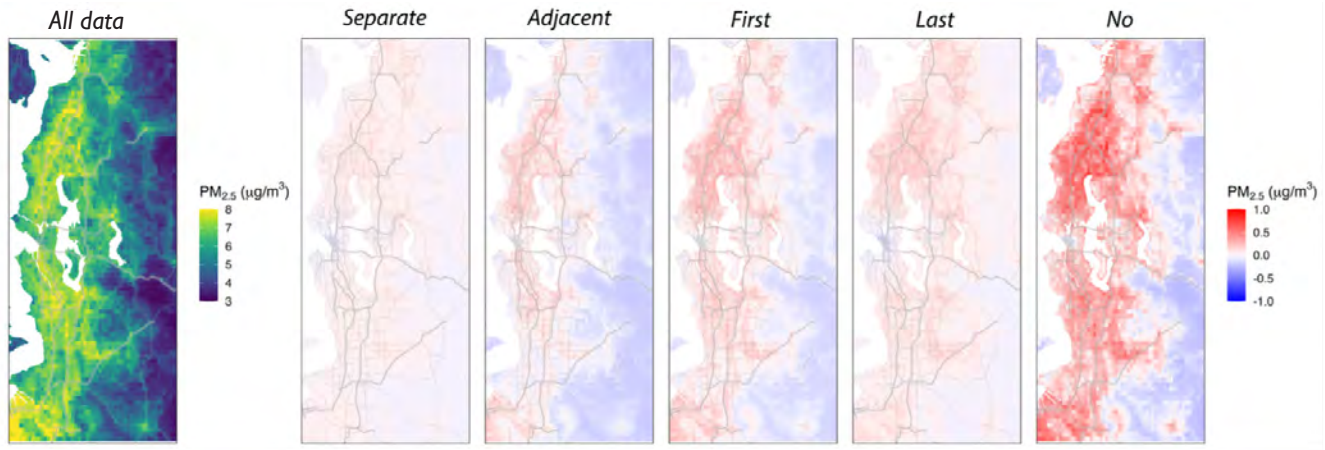


Figure 7.2. Mean $PM_{2.5}$ predictions for the all data model and the prediction differences for other monitoring designs. The data is from June 2017 to May 2019 (the period of the short-term low-cost sensor campaign). All data is the subtrahend for the other designs. Reprinted with permission from Bi et al. 2024.

Table 7.2. Combined Validation Performance for Prediction of $PM_{2.5}$ for the Whole Time Period (Spatial Validation) and the Two-Week Time Scale (spatiotemporal validation) at the Two-Period Low-Cost Sensor Locations for Various Spatially Reduced Designs

Scheme	Model	*N* of Reference-Grade Measurements for Modeling (*N* of sensors)	*N* of Low-Cost Measurements for Modeling (*N* of sensors)	*N* of Combined Validation Measurements (*N* of sensors)	Spatiotemporal Validation		Spatial Validation	
					MSE R^2	RMSE (µg/m³)	MSE R^2	RMSE (µg/m³)
All data		6,186 (78)	502 (82)	502 (82)	0.84	0.88	0.57	0.57
Random Selection	10%		61 (9)		0.74	1.11	0.28	0.74
	25%		131 (21)		0.77	1.04	0.39	0.69
	50%		250 (41)		0.8	0.98	0.45	0.65
	75%		382 (62)		0.81	0.94	0.5	0.62
	90%		458 (74)		0.84	0.88	0.57	0.57
Distance to Road and Population Density	Long Distance to Road & High Population Density		92 (15)		0.72	1.15	0.19	0.78
	Long Distance to Road & Low Population Density		161 (26)		0.74	1.1	0.27	0.75
	Short Distance to Road & High Population Density		154 (26)		0.78	1.01	0.4	0.68
	Short Distance to Road & Low Population Density		95 (15)		0.76	1.06	0.32	0.72
Participant Recruitment Region	Region 1		88 (13)		0.7	1.18	0.11	0.82
	Region 2		79 (14)		0.72	1.15	0.22	0.77
	Region 3		87 (14)		0.76	1.06	0.32	0.72
	Region 4		56 (8)		0.71	1.17	0.14	0.81
	Region 5		80 (13)		0.78	1.03	0.4	0.68
	Region 6		51 (9)		0.73	1.12	0.26	0.75
	Region 7		61 (11)		0.73	1.12	0.22	0.77

MSE R^2 = mean squared error coefficient of determination; RMSE = root mean squared error.

Health Analyses

Figure 7.1 shows the health estimates and their 95% confidence intervals for the association of a 1-µg/m³ increase in $PM_{2.5}$ with cognitive function using confounder model 1 (Equation 3.11) for analysis, while Table S7.4 defines its labels and also indicates the time period from which the model predictions were obtained. The figure shows predictions from both the full spatiotemporal model as well as those restricted in time to 2019 (purely spatial, with residential history based on the 5 years prior to enrollment). The health association estimates were –0.023 and –0.009 per 1-µg/m³ increase in $PM_{2.5}$ using the full spatiotemporal and spatial model, respectively, with wide confidence intervals that are consistent with no association.

Overall, all estimates ranged from negative to slightly positive and had wide and highly overlapping confidence intervals. While the primary spatial *All data* estimate (labeled "full" in the figure) was among the more negative estimates, there is little distinction between them. The confidence intervals are slightly wider for spatial predictions from the full time period model, and those health estimates tend to be negative.

RESULTS — NO_2 MODELING

Summary Statistics

The data available for the NO_2 model over the study period are shown in Figure S7.3 and summarized in Table S7.5. There

were 11 agency monitoring locations over this period (with three meeting our criteria for inclusion in the computation of long-term time trends) and several supplementary monitoring campaigns (described in Chapter 7's Additional Materials). Despite a relatively limited number of agency locations, there were 479 unique locations with NO_2 measurements among the supplementary monitoring campaigns that contributed information on the spatial distribution of NO_2 to the model. Our focus on the 117 low-cost sensors represents 24% of the locations of the dataset, and we do not assess the contribution of the other data sources to the model. Individual low-cost sensors varied in their deployment and their subsequent contribution to the modeling data; they had a range of 1–83 observations per location, and a median of seven 2-week observations per deployment location (Table S7.5). Another feature of the modeling data was that of the 110 ACT-AP snapshot locations, 13 had data for only two of the three seasons measured.

The NO_2 concentrations among the agency monitors included in the model's time trend were greater than those not included in the trend (location mean ± SD = 13.5 ± 7.3 ppb versus 11.9 ± 7.1 ppb), likely because they were generally located in more urban areas (Figure S7.4, Table S7.5). ACT-AP snapshot concentrations (11.3 ± 3.8 ppb) were comparable to the low-cost sensors (11.9 ± 3.1 ppb), even though the ACT-AP snapshot campaign was designed to capture pollutant gradients along roadways, while the low-cost sensors reflected residential locations. The Yesler Terrace study's NO_2 concentrations were the highest of all the model's data sources (26.0 ± 4.1 ppb), consistent with its sampling location adjacent to an interstate highway. The DEEDS and PSID studies also had relatively higher NO_2 concentrations, likely as a result of the study design (DEEDS sought to capture emissions from diesel combustion) and time period (the PSID study took place earlier in the modeling period when NO_2 concentrations tended to be higher).

Added Value of Low-Cost Sensors

The results of cross-validation procedures on final NO_2 models fitted with and without the low-cost sensor data are shown in **Table 7.3**. These measures describe the accuracy of the average NO_2 concentration while considering both the temporal and spatial components of the predictions. The inclusion of low-cost sensor data improved the spatial agency leave-one-site-out cross-validation results slightly, with the CV RMSE decreasing from 2.8 to 2.5 ppb and the CV MSE R^2 increasing from 0.82 to 0.85.

The impact of low-cost sensors on the various ACT-AP snapshot cross-validation results was generally small, with little to no change in the CV performance statistics. One exception to that, however, was the leave-one-campaign-out CV MSE R^2, which decreased from 0.50 to 0.23 with the inclusion of low-cost sensor data (CV RMSE changed little from 3.2 ppb without low-cost sensors to 3.9 ppb with). Finally, the results of the evaluation conducted on the low-cost sensors markedly improved with the inclusion of low-cost sensors, from RMSE = 3.8 ppb and MSE R^2 = 0.08 (without low-cost sensors) to CV RMSE = 2.8 ppb and CV MSE R^2 = 0.51 (with low-cost sensors).

Health Analyses

The primary focus of this analysis is whether association estimates change with the omission of relevant spatial exposure

Table 7.3. NO_2 Spatiotemporal Model CV Performance Statistics Comparing Predictions with Observations Across Several CV Designs, Over Short and Long Time Periods, and With and Without Low-Cost Sensor Data Included in the Model Development

	RMSE (ppb)		MSE R^2		Regression R^2	
CV Description	**Without Low-Cost Sensors**	**With Low-Cost Sensors**	**Without Low-Cost Sensors**	**With Low-Cost Sensors**	**Without Low-Cost Sensors**	**With Low-Cost Sensors**
Agency Leave One Site Out						
Spatial (whole modeling period)	2.8	2.5	0.82	0.85	0.84	0.85
1-year	2.6	2.5	0.82	0.83	0.85	0.85
Spatiotemporal (2-week)	3.5	3.4	0.72	0.74	0.76	0.76
ACT-AP snapshot						
Leave-one-cluster-out	2.7	2.7	0.61	0.61	0.63	0.69
Ninefold, equally sized groups	2.7	2.7	0.64	0.64	0.66	0.72
Leave-one-campaign out	3.2	3.9	0.50	0.23	0.51	0.50
Low-cost sensor 10-fold	**3.8** [a]	**2.8**	**0.08** [a]	**0.51**	**0.39** [a]	**0.60**

[a] Performance statistics for low-cost sensor cross-validation without low-cost sensor data are out of sample, not CV.

information provided by the low-cost sensors. **Figure 7.3** shows the health association estimates for exposures with and without low-cost sensor data and their 95% confidence intervals for the association of a 3-ppb increase in NO_2 with cognitive function using confounder model 1 (Equation 3.11) for analysis. The figure shows columns for predictions from the full spatiotemporal model (ST, right column), which are time-varying with exposures before 1996 assumed to be equal to the 1996 levels, as well as those restricted in time to 2019 while still retaining the time-varying address history (SP, left column). Overall, the mean associations of NO_2 with cognitive function ranged from negative to slightly positive. The estimated association using the spatial exposure model with low-cost sensor data is the strongest and most negative (–0.0030; 95% CI: –0.057 to –0.002), and the same exposure model without low-cost sensor data is closer to zero. Furthermore, both of the estimates from the spatiotemporal exposure models were approximately zero (e.g., 0.001; 95% CI: –0.028 to 0.030 for the full model), and their confidence intervals suggest that they are consistent with a wide range of effects.

DISCUSSION AND CONCLUSIONS

This chapter considered the added value of low-cost sensor data in predicting $PM_{2.5}$ and NO_2 for an application to epidemiology. Differences in the underlying data for these two pollutants contributed to differences in the two assessments. In the text that follows, we first discuss lessons learned from each pollutant evaluation separately, with an emphasis on lessons learned from the exposure modeling, before providing unified comments on the findings.

Insights from the $PM_{2.5}$ Assessment

This study investigated fixed-site low-cost sensor designs with varying sampling durations, repeated measurements across periods, and sampled locations that affect the spatiotemporal prediction of ambient $PM_{2.5}$ exposure. Our goal was to assess practical considerations for optimizing spatiotemporal coverage of low-cost sensors in $PM_{2.5}$ exposure modeling and gain insight into its application for long-term epidemiological studies. The exposure model performance assessments demonstrated that the inclusion of low-cost sensor measurements improved $PM_{2.5}$ exposure modeling and that increasing the number of low-cost sensor locations and times resulted in better performance statistics and more accurate $PM_{2.5}$ predictions. We also demonstrated that the same number of low-cost sensor measurements, if temporally separated, could result in better model performance as compared to temporally consecutive measurements. Thus, it is key to conduct repeated low-cost sensor measurements while ensuring adequate temporal separation between the measurements; maintain an adequate number of spatial monitoring locations to effectively capture the spatial variability of $PM_{2.5}$ concentrations; and incorporate near-roadway monitoring locations that provide valuable insights into the impact of traffic-related emissions on $PM_{2.5}$ concentrations. These findings agree with our hypothesis regarding the overall value of incorporating low-cost sensor data into $PM_{2.5}$ exposure modeling and show that certain design strategies can reduce the resources invested in low-cost sensor deployment while ensuring good exposure modeling. However, when we used predictions from these various models in our estimates

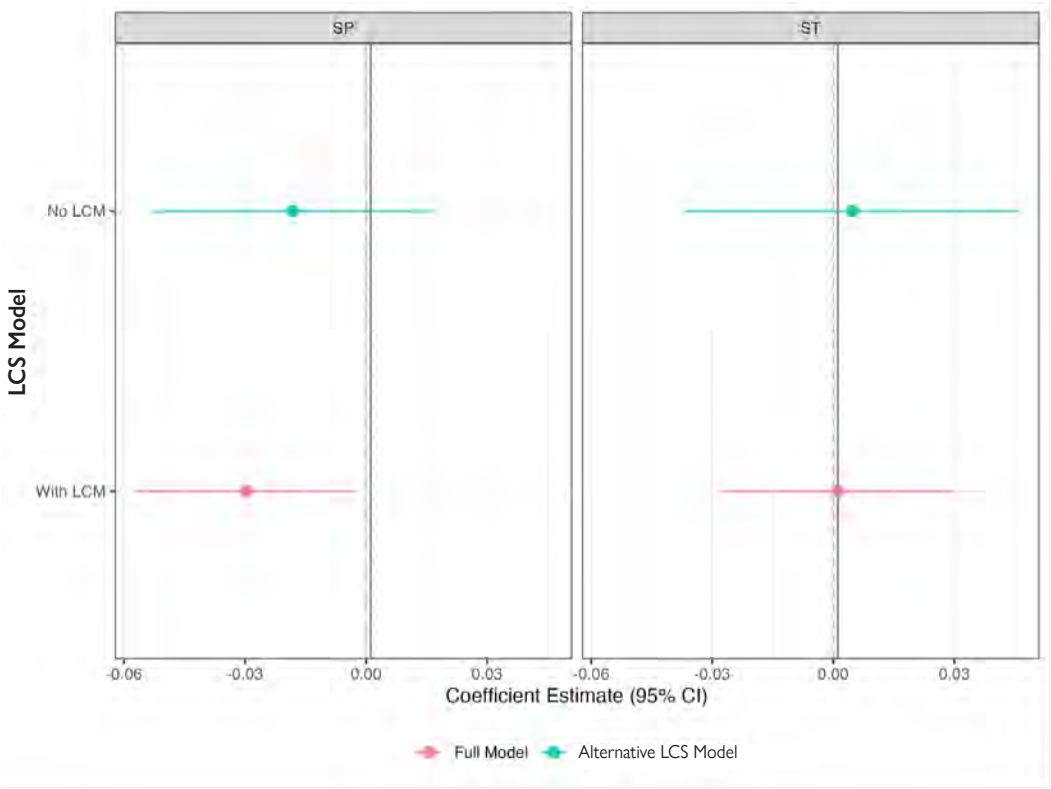


Figure 7.3. Estimated association (95% CI) between NO_2 (3 ppb) and baseline cognitive function for exposure models with and without low-cost sensor data. The associations are adjusted for age, calendar year, sex, and education (confounder model 1). LCS = low-cost sensor; SP = spatial model; ST = spatiotemporal model.

of association with cognitive function, we found that the estimates from all the various models were overlapping, and there was little meaningful difference between them.

Returning to the exposure prediction results, we found that all temporally reduced monitoring designs resulted in decreased performance when compared to the *All data* model that used all available low-cost sensor measurements. However, the model with the greatest time interval between the two low-cost sensor measurements had the highest performance among designs, while the model with temporally consecutive measurements had a comparable performance to the models with only one measurement during a single period. While only a subset of the results was included in Chapter 7's Additional Materials, we still observed differences in performance after extending the model to a multidecadal period of 1978–2021, although the extent of the performance changes was smaller than those observed in our primary modeling period of 2010–2020. These findings emphasize the importance of characterizing longer-term temporal variations of $PM_{2.5}$ during the monitoring to optimally support the exposure modeling, as opposed to making repeated measurements over a shorter time period. If logistical considerations require fewer sampling periods, ensuring reasonably long time intervals between deployment campaigns (e.g., ensuring a temporal spread of more than one season) can enhance model performance for the same amount of measurement data in the exposure modeling.

As expected, our findings suggest that incorporating more low-cost sensor locations increases model performance. However, the performance was found to reach its optimum relative to all the low-cost sensor data when 90% of the low-cost sensors were included. This result may suggest that further incorporation of low-cost sensors beyond a certain threshold does not result in meaningful model improvements, although this hypothesis must be treated with caution, given that we had a fixed maximum of low-cost sensors to evaluate.

Our study indicates that a low-cost sensor's distance from major roads has a noticeable impact on the model's performance, with low-cost sensors located closer to major roads contributing more to improved model performance compared to those farther away. This is likely due to siting criteria for PM monitoring that avoid near-road locations (US EPA 2008). This finding suggests that increased availability of low-cost sensors that measure $PM_{2.5}$ concentrations near roadways can contribute to a more comprehensive characterization of $PM_{2.5}$ pollution patterns in proximity to roads, thus enhancing the model's performance. (Bi et al. 2022b; Zhu et al. 2002) On the other hand, we found that population density is a less critical factor, as there were no substantial differences in performance when comparing models including low-cost sensors in areas restricted to either high or low population density. This could be because the ACT-AP low-cost sensors were deployed in areas with relatively high population densities. Specifically, all low-cost sensors were at locations with a population density that is above the median population density in the Puget Sound region. These findings highlight the importance of carefully siting low-cost sensors and considering their geographical and demographic characteristics to improve the performance of $PM_{2.5}$ models.

The primary limitation of our study is that the low-cost sensors at the residential locations of the ACT-AP cohort covered relatively short time periods, and most of them were deployed during only two seasonal periods. Nonetheless, our study presented clear evidence that different temporal designs can impact $PM_{2.5}$ exposure modeling. To conduct a more comprehensive temporal analysis, we recommend using low-cost sensors with more repeated measurements over more extended deployment periods. This would also facilitate estimation of the model's performance because the spatial summaries over a long time period would more accurately represent pure spatial variation. Additionally, the ACT-AP low-cost sensors were deployed at staggered times and for varying durations (to manage the deployment cost effectively, we rotated the monitors at different residential locations). We expect that a more balanced deployment design would allow more extensive evaluation of low-cost sensor sampling designs (e.g., to evaluate alternative methods for selecting any two consecutive measurements, rather than being limited to the first two consecutive measurements of a monitor). Furthermore, the external validation was restricted to a relatively small number of one-period low-cost sensor locations, which constrained a holistic evaluation of the model's performance. In the future, we plan to use a considerably larger dataset of $PM_{2.5}$ measurements from the openly accessible low-cost air monitoring network, PurpleAir, to further our investigation. This will allow us to better answer our research question on the most optimal low-cost sensor campaign design for epidemiological inference that requires the least amount of time and money invested in monitor deployment.

Low-cost sensors may still have residual uncertainties despite calibration, impacting our performance evaluation. However, this impact was likely to be consistent across low-cost sensor campaign designs, ensuring a valid interdesign comparison. Finally, our analysis was conducted during relatively low $PM_{2.5}$ concentration ranges, with the limitation of excluding extreme measurements during wildfire weeks. It would be worthwhile to undertake additional modeling efforts aimed at effectively analyzing and confirming our findings under high-concentration conditions.

Our health analyses gave overlapping results from the various exposure estimates, such that it is difficult to develop deeper insights from the findings of this single study.

Insights from the NO_2 Assessment

Our primary exposure modeling goal was accurate NO_2 exposure prediction at ACT-AP participant residences, particularly the long-term averages, for input into epidemiological models examining associations with cognitive function. Ideally, we would have high-quality measurements over a

long period of time at many participants' residences to both calculate accurate long-term average exposures and evaluate predictions from our NO_2 exposure model. Data limitations, however, presented challenges to this goal. As a compromise, we evaluated cross-validated predictions for several data sources, including agency data at various timescales and supplementary monitoring campaigns; striking a balance between data quality and availability, the locations of various measurement types, and the relevance of the timescale.

The highest quality supplementary data (i.e., ACT-AP snapshot) was designed to capture NO_2 road gradients, not to measure residential exposures. The ACT-AP snapshot cross-validation primarily characterizes the added value of additional spatial data. These measurements, taken at the same 110 sites over three seasons, do not contribute substantial amounts of temporal information, but were collected at many locations, and are of high quality (i.e., they have good agreement with FRM methods). The low-cost sensors, on the other hand, were deployed at residences, but the measurements were of lower quality. In hindsight, it may have been beneficial to deploy Ogawa Passive Samplers — perhaps even co-located with low-cost sensors — in residential locations as an additional supplementary data source.

The inclusion of low-cost sensor data into the spatiotemporal NO_2 model generally increased the prediction concentrations and decreased the variability across all time periods, although neither meaningfully (Zuidema et al. 2024). Furthermore, the results of cross-validation procedures showed only modest improvement of the model predictions with the addition of low-cost sensor data in most of the evaluations considered (Table 7.3), even though the low-cost sensor data increased the number of spatial locations by 32%. The low-cost sensor data improved the Agency's leave-one-site-out cross-validation performance slightly, and had mostly no impact on the various ACT-AP snapshot cross-validation results, except for the leave-one-campaign-out cross-validation, where the performance decreased. The August 2018 ACT-AP snapshot measurement campaign occurred during a period that was highly impacted by wildfire smoke, with potentially negative effects on the cross-validation in this fold. The low-cost sensors were calibrated while excluding wildfire-impacted periods (Zuidema et al. 2021) nitric oxide (NO; NO-B4, which may explain the decreased prediction accuracy of the model during this period. Despite observing these differences, the magnitude of the performance increase or decrease was not large, so we caution against over-interpreting these results.

While low-cost sensors did not definitively alter the quality of the predictions as evaluated through cross-validation at nonresidential locations, out-of-sample predictions at residential locations showed substantial improvement. At residential locations (low-cost sensor sites), the performance improvement with low-cost sensors gave RMSE = 3.8 ppb NO_2 and MSE R^2 = 0.08 (without low-cost sensors), compared to CV RMSE = 2.8 ppb NO_2 and CV R^2 = 0.51 (with low-cost sensors). Note that the prediction performance assessment at low-cost sensor sites without low-cost sensor data is completely out-of-sample and not cross-validated.

The cross-validation results observed in this study are generally comparable to those observed in the low-cost NO_2 sensor calibration (CV RMSE = 3 ppb, and CV R^2_{reg} = 0.79), indicating that the model performs as well as the sensors, and suggesting that the low-cost sensor data may not lower the performance of the spatiotemporal NO_2 model (Zuidema et al. 2024). As others have observed, low-cost sensor data with lower data quality (e.g., lower accuracy or precision) may be more obviously detrimental to models of air pollution (Bi et al. 2020; Zuidema et al. 2021). In this study, however, the NO_2 sensors were specifically calibrated in the Puget Sound region for this modeling purpose, potentially reducing the amount of measurement error present. Furthermore, the model we developed had measures of cross-validation performance consistent with similarly constructed spatiotemporal models of NO_2 in six US metropolitan areas (Keller et al. 2015). The incorporation of low-cost sensor data, which contributed to a moderate 24% increase in the number of spatial locations with data used to fit the model, generally increased the predicted NO_2 concentrations.

Our epidemiological analyses were limited to two exposure models — one with and one without low-cost sensors. The health association estimates from the time-constant (spatial) and time-varying (spatiotemporal) exposure models with low-cost sensors imply different conclusions, with the spatial models suggesting an adverse association between NO_2 and baseline cognitive function, while the spatiotemporal models suggest no association. In contrast, association estimates for exposures without low-cost sensor data were more comparable with each other and had wider confidence intervals.

Combined Discussion and Overall Conclusions

For both $PM_{2.5}$ and NO_2, we found that the addition of low-cost sensor data improved exposure prediction performance at residential locations such that exposure models with an absence or smaller fractions of low-cost sensor data all performed worse than models with all available low-cost sensor data. Furthermore, while different predictions across exposure models with no, all, or subsets of low-cost sensor data produced different health association estimates, these differences were small. The health association 95% confidence intervals overlapped across exposure models; these results did not suggest favoring one exposure model over another.

While in principle the time-varying exposures (spatiotemporal) more accurately reflect individuals' exposures at baseline, evidence in the literature suggests that the spatial distribution of exposure is stable over time (Cesaroni et al. 2012; Eeftens et al. 2011; Wang et al. 2013), such that a time-constant exposure estimate (e.g., spatial exposures predicted from the mobile campaign) should be an adequate

replacement for the more temporally aligned time-varying exposures. Furthermore, the association model controlled for calendar year to remove potential bias from residual temporal confounding. We found that while the association estimates for spatial and spatiotemporal exposures differed somewhat, these differences were not substantial, and most of the health estimates overlapped. The largest difference between spatiotemporal and spatial exposure metrics was for the NO_2 full model with low-cost sensor data. The estimates of the full models were the only pair of estimates where the mean of one confidence interval wasn't contained within the other's 95% CI. These results suggest that reliance on the time-constant exposure estimates may not substantially impact association estimates.

Overall, it was challenging to develop deep insights into the utility of collecting low-cost sensor data to supplement regulatory monitoring due to the structure of the data. As seen in Figures S7.1, S7.2, and S7.3, these space-time datasets are very sparse and unbalanced over time, with more missing observations than available data. These sparse and unbalanced features pertain to the added low-cost sensor data as well because the 20 low-cost sensor instruments had to be deployed in a rotating time-unbalanced design. These features don't negate the added value of the low-cost sensor campaign which provided a substantial increase in the number of spatial locations, with 82 (2010–2020) and 117 new spatial locations used in the spatiotemporal models for $PM_{2.5}$ and NO_2; these represent a 105% and 32% respective increase in the number of locations used. However, even with the large increase in the number of spatial locations, while it did improve prediction model performance at the target residential locations, we did not observe substantial impacts on the health association estimates, as noted above. Furthermore, the lower precision of the low-cost sensor measurements, particularly for NO_2, may also have contributed to these results.

CHAPTER 8: APPLICATION OF ADVANCED STATISTICAL METHODS FOR EXPOSURE PREDICTION USING THE MOBILE DATA

Lead authors: Si Cheng, Magali Blanco, Lianne Sheppard, Ali Shojaie, Adam Szpiro

INTRODUCTION

The research described in this chapter addresses our efforts to apply novel statistical methods to develop traffic-related air pollution (TRAP) exposure predictions with the ultimate goal of improving health inferences. For all applications, we used the stationary roadside data from the mobile monitoring campaign described in Chapter 3. We summarize three projects, and then develop the first and last in more detail in the following subsections. We conclude with a discussion of the insights developed.

These projects were as follows:

1. **Application of spatial ensemble-learning methods and resulting variable importance metrics:** We applied spatial ensemble-learning methods to replace our universal kriging (UK) approach using a linear mean derived from a partial least squares model with modern ensemble machine learners (e.g., Random Forest) with spatial structures that efficiently identify nonlinear relationships from high-dimensional geographic predictors. Below, we describe these methods and the new variable importance metric we developed, with our results concluding that the ensemble learning methods did not perform better than the universal kriging with partial least squares approach (UK-PLS).
2. **Leveraging driving distance to improve spatial prediction:** We worked to determine whether taking into account additional information about the road network would improve our predictions. We leveraged driving distances along the road network and considered their added value, both by creating a spatial model using driving distance alone and by adding driving distance to the spatial model based on Euclidean distance (as determined by longitude and latitude). We implemented a network graph approach with the edges assigned weights equal to the inverse driving distances between stationary locations. Because driving distance is not a well-defined mathematical metric, we identified resulting computational issues. Furthermore, even when these computational challenges were absent, we found that incorporating the driving distance information did not result in better performance than leveraging Euclidean distance. Thus, we dropped further study of this idea.
3. **Multipollutant dimension reduction, a tool to be applied prior to the prediction of spatial data:** We developed a principal component analysis (PCA) approach that balances prediction and approximation accuracy, called representative and predictive PCA or RapPCA. We briefly summarize RapPCA below and show how it compares with classical and predictive PCA in the Seattle mobile monitoring dataset. In future work, we will apply this method to compare health inference using a variety of exposure mixture prediction approaches.

APPLICATION OF SPATIAL ENSEMBLE-LEARNING METHODS AND RESULTING VARIABLE IMPORTANCE METRICS

Methods — Introduction

We used the mobile monitoring data at the 309 stationary roadside locations described in Chapter 3 for analysis. We considered five pollutants: UFPs, BC, NO_2, CO_2, and $PM_{2.5}$. For UFPs, we focused on the data from the unscreened P-Trak (20–1,000 nm particles) instrument. The methods, results, and discussion are a distillation from a published preprint (Cheng et al. 2024b).

Methods — Spatial Prediction Approach

Our initial work focused on the prediction of annual average pollutant concentrations with various machine learning methods, including the following:

UK-PLS — A two-step procedure that first extracts the top partial least squares (PLS) components from the covariates (where the number of components is determined by cross-validation within the training set, to maximize prediction R^2), and then fits a UK model via maximum likelihood with the selected components as covariates and an exponential covariance (Additional details of the UK-PLS approach are provided in Chapter 3.)

SpatRF — A spatial random forest algorithm proposed by Wai and colleagues (2020), where the tree-building algorithm selects each split of the tree adjusting for spatial correlation via thin plate regression splines (TPRS)

We selected hyperparameters by cross-validation with a grid search. We considered two optimization approaches: the pseudo-likelihood optimization introduced by Wai and colleagues (2020) (SpatRF-PL) and a nonparametric optimization approach (SpatRF-NP), also discussed by Wai and colleagues (2020).

RF — A standard random forest algorithm implemented by the randomForest R package (Breiman et al. 2022), ignoring spatial correlation

TPRS — Spatial smoothing via thin plate regression splines implemented by the mgcv R package (Wood 2003, 2017)

$X^*_{m\times p}$.

RF-TPRS — A two-step procedure that first runs RF and then conducts TPRS spatial smoothing on the residuals from RF

TPRS-RF — A two-step procedure that first runs TPRS and then applies RF on the residuals from TPRS

Our primary models were UK-PLS, the standard on which all of our other work is based, and SpatRF-PL, the recently developed spatial random forest with pseudo-likelihood optimization. We also evaluated the remaining models as benchmarks for comparison. We modeled the pollutants on the log scale and converted these to the native scale for model assessment. We assessed performance accuracy using 10-fold cross-validated MSE R^2 as described in Chapter 3. Finally, using the baseline cognitive function outcome and confounder model 1 described in Chapter 3, we used each of these predicted exposures for UFPs from the unscreened P-Trak to estimate the cognitive function health association using Equation 3.11.

Following the assessment of prediction accuracy, we proposed an intuitive variable importance metric for spatial machine learning models that can be applied to a wide variety of models with separable mean and correlation components. We applied these to the UK-PLS and SpatRF-PL models. The variable importance metric assesses the difference in predicted values for each datapoint when each predictor is fixed at specific quantile levels and averages them as the overall contribution of each predictor to the outcome. We provide technical details of this metric in the next subsection.

Methods — Variable Importance Metric

To develop the variable importance framework, we consider a class of models where an outcome *Y(s)* is indexed by location $s \in \Omega$. This is modeled via an additive mean surface of the form

$$g\big(\mu(s)\big) = \sum_{k=1}^{K}[f_k\big(X(s)\big) + \upsilon_k(s)] \qquad (8.1)$$

where $g(\bullet)$ is a link function, $\mu(s)$ is the expected value of $Y(s)$, $X(s)$ represents the covariates, each $\upsilon_k(s)$ represents the correlated error term, and each $f_k(\bullet)$ is an unknown function within some function class F. The indexing k allows for application to multiple ensemble-learning methods. As an example, when g is the identity link and $\upsilon_k(s)$ is a correlated Gaussian process, $\mu(s)$ models the surface of a continuous outcome, for example, a UK model if f is linear.

Specifically, we describe how to apply this approach to the UK-PLS and then the SpatRF-PL models. Given the observed continuous outcome $Y_{n\times 1}$ (log-transformed), with identity link function $g(\bullet)$ and potentially high-dimensional covariates $X_{n\times p}$, the UK-PLS model has $k = 1$ and first conducts PLS to extract the first q $(1 \le q \le p)$ components of X by finding a decomposition of X

$$T_{n\times q} \coloneqq X_{n\times p}H_{p\times q}$$

such that the covariance between T and Y is maximized. The number q of components are selected by cross-validation. In the second step, a UK model is fit. It has the form

$$Y_{n\times 1} = T_{n\times q}\beta_{q\times 1} + \upsilon_{n\times 1}$$

$$\upsilon_{n\times 1} \sim \text{Normal}(0, \Sigma(\theta))$$

where θ are covariance parameters (e.g., the nugget, partial sill, and range, see (Cressie, 2015)), which can be solved jointly with β via maximum likelihood. With terms $\tilde{H}_{p\times q}, \hat{\beta}, \hat{\theta}$ estimated from the model for m new test locations with covariate values $X^*_{m\times p}$, the outcome Y^* can be predicted as

$$\hat{Y}^* = X^*\tilde{H}\hat{\beta} + \tilde{\Sigma}_{12}(\hat{\theta})\tilde{\Sigma}_{22}^{-1}(\hat{\theta})(Y - T\hat{\beta})$$

where $\tilde{\Sigma}_{(m+n)\times(m+n)}(\hat{\theta})$ is the covariance matrix induced by the distances between all n training and m test locations, partitioned as

$$\tilde{\Sigma}\big(\hat{\theta}\big) = \begin{bmatrix} \tilde{\Sigma}_{11}^{(m\times m)} & \tilde{\Sigma}_{12}^{(m\times n)} \\ \tilde{\Sigma}_{21}^{(n\times m)} & \tilde{\Sigma}_{22}^{(n\times n)} \end{bmatrix} \qquad (8.2)$$

The spatial random forest algorithm proposed by Wai (2020), which is solved by pseudo-likelihood (SpatRF-PL), is an ensemble model (i.e., $k > 1$) that specifies

$$\hat{\mu}(s) \coloneqq \sum_{k=1}^{K}\hat{\mu}(s) \coloneqq \sum_{k=1}^{K}\big[\hat{f}_k\big(X(s)\big) + \hat{\upsilon}_k(s)\big]$$

where each $\hat{f}_k(\bullet)$ is a regression tree, and each $\upsilon_k(s)$ could be modeled using common spatial smoothing methods such as kriging or regression splines (Friedman, 1991; Wood, 2003). With a kriging model, for each k, a spatially adjusted tree can be built by solving the optimization problem resulting from the profile likelihood, assuming normally distributed spatial error terms:

$$\arg\max_{\theta_k}\left[-\frac{1}{2}\log|\Sigma(\theta_k)| - \frac{1}{2}\Big(Y - \hat{f}_k(X|\Sigma(\theta_k)\Big)^T\Sigma^{-1}(\theta_k)\Big(Y - \hat{f}_k(X|\Sigma(\theta_k)\Big)\right]$$

$$\text{s.t. } \hat{f}_k(X|\Sigma(\theta_k) = \underset{f_k(X|\Sigma(\theta_k))}{\arg\min}\big(Y - f_k(X|\Sigma(\theta_k)\big)^T\Sigma^{-1}(\theta_k)(Y - f_k(X|\Sigma(\theta_k)))$$

Likewise, predictions can be made via

$$\hat{Y}^* = \sum_{k=1}^{K}\hat{f}_k(X^*) + \mathbb{E}_{\widehat{\theta_k}}(\upsilon^*|\upsilon) = \sum_{k=1}^{K}[\hat{f}_k(X^*) + \tilde{\Sigma}_{12}(\hat{\theta}_k)\tilde{\Sigma}_{22}^{-1}(\hat{\theta}_k)(Y - T\hat{\beta})]$$

with $\tilde{\Sigma}$ defined as above.

The variable importance model applies to additive models taking the form of Equation 8.1. This leave-one-out approach is based on the change in predictions across different

user-specified quantiles $q_1, q_2, \cdots, q_m$ for each covariate, evaluating it at each location $s_1, s_2, \cdots, s_n$ one at a time. Note that prediction at the test locations relies on evaluation of the trained covariance model $\hat{v}(s)$. This implies that when we permute or fix the values of the test set covariates X, as in many common variable importance analyses, the evaluation of the covariance model $\hat{v}(s)$ is also implicitly altered, and the distribution of the residuals at the test locations may no longer be well-fit by $\hat{v}(s)$. Therefore, the key idea in this approach is to reuse the trained mean model across all locations, but refit the covariance model in a leave-one-out manner. We write $\hat{F}_{X_j}(x_j), j = 1, \cdots, p$ as the empirical cumulative distribution function of the j^{th} covariate, and s_{-1} as the set of all locations except the i^{th} one.

Suppose we have trained a model

$$g(\hat{\mu}(s)) = \sum_{k=1}^{K} [\hat{f}_k(X(s)) + \hat{v}_k(s)]$$

from observations $\{X(s_i), Y(s_i)\}_{i=1}^n$. Then for the j^{th} covariate at the l^{th} quantile q_l of interest and within the k^{th} submodel $\hat{f}_k$, we replace each with $X_j(s_i)$ the sample q_l^{th} quantile and calculate the predicted mean as

$$\hat{\zeta}_k^{j,l}(s_i) \coloneqq \hat{f}_k\left(X_1(s_i), \cdots, F_{X_j}^{-1}(q_l), \cdots, X_p(s_i)\right)$$

for location i. In other words, this is the new predicted mean at s_i with the j^{th} covariate replaced by its q_l^{th} quantile across $s_1, s_2, \cdots, s_n$. Next, we refit the k^{th} error component with the new predicted means $\hat{\zeta}_k^{j,l}(s_{-i})$ along with observations $(X(s_{-i}), Y(s_{-i}))$, leaving out the i^{th} site. Denoting the resulting model as $\hat{v}_{(-1),k}^{j,l}(s)$, the leave-one-out approach yields the linear predictor

$$\hat{\eta}_k^{j,l}(s_i) \coloneqq \hat{\zeta}_k^{j,l}(s_i) + \mathbb{E}_{\hat{v}_{(-i),k}^{j,l}}\left[\hat{v}_{(-i),k}^{j,l}(s_i) | Y(s_{-i}), \hat{f}_k(X(s_{-i}))\right] \quad (8.3)$$

for location i, which is what the model would predict if the j^{th} covariate of all data points were replaced by the q_l^{th} quantile of its distribution, and if the error component were fit without the i^{th} data point, while keeping everything else intact. Refitting leads to updated covariance models that account for the implicit change in the error distribution caused by manipulating the covariates.

Computing this for all i, we obtain a sequence of linear predictors of the form Equation 8.3. Aggregating across each submodel k and location i finally leads to the average leave-one-out predictions at the q_l^{th} quantile for covariate j:

$$\bar{\mu}_{j,l} \coloneqq g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\eta}_k^{j,l}(s_i)\right) \quad (8.4)$$

For each j, the trajectory $\bar{\mu}_{j,1}, \cdots, \bar{\mu}_{j,m}$ reflects how the predictions, on average, vary across different quantiles of covariate X_j, which serves as an intuitive measure of the contribution of this covariate to the predicted outcome. This procedure could easily be parallelized to facilitate computation. To fix the ideas, Algorithm 1 describes the procedure in detail for the indices i, j, k, l, which corresponds to each monitoring site i, covariate of interest j, and submodel k that is fitted in an ensemble-learning method (including $K = 1$ models), and quantile level l, respectively.

Results — Spatial Prediction Approach

Table 8.1 summarizes the predictive performance of all methods for five pollutants. The relative performance of the different models reveals different sources of heterogeneity in pollutant concentrations. For UFPs and NO_2, covariate and spatial effects both account for part of the spatial heterogeneity, reflected by the reasonable performances of RF or TPRS alone, while accounting for both together in a joint (SpatRF-PL, SpatRF-NP) or two-step (RF-TPRS, TPRS-RF) manner leads to increased accuracy. The case is similar for BC and CO_2, although the covariate effects appear to be stronger than the spatial effects. $PM_{2.5}$, on the other hand, illustrates a scenario where spatial smoothing alone captures the major source of heterogeneity. UK-PLS and SpatRF-PL have the best overall performance for all pollutants. While neither shows clearly better or worse accuracy than the other, UK-PLS has slightly better performance for all pollutants other than BC.

For the remaining results in this section, we focus on UFP concentrations. Some results also prioritize the comparison between UK-PLS and SpatRF-PL. We note that the cross-validated MSE R^2s from these models were 0.81 and 0.78 for UK-PLS and SpatRF-PL, respectively. **Figure 8.1** displays the cross-validated prediction errors of UK-PLS and SpatRF-PL for UFPs at all stationary locations in the Seattle mobile monitoring campaign. We observe very similar spatial patterns in the distribution of prediction errors across the monitoring locations produced by UK-PLS and SpatRF-PL, despite their different nature: UK-PLS captures a linear trend in the mean model while SpatRF-PL allows for nonlinear effects; UK-PLS is a two-step procedure with explicit dimension reduction

Table 8.1. Cross-Validated MSE R^2 by Pollutant from the Seattle Mobile Monitoring Campaign Stationary Roadside Dataset and Model for Various Machine Learning Models [a]

	UK-PLS	RF	TPRS	RF-TPRS	TPRS-RF	SpatRF-PL	SpatRF-NP
UFP	0.81	0.75	0.76	0.79	0.80	0.78	0.78
BC	0.65	0.60	0.57	0.67	0.64	0.67	0.67
NO_2	0.77	0.70	0.70	0.76	0.74	0.75	0.74
CO_2	0.56	0.47	0.44	0.57	0.55	0.56	0.54
$PM_{2.5}$	0.76	0.66	0.73	0.72	0.73	0.74	0.71

[a] The machine-learning models are the universal kriging with partial least squares dimension reduction (UK-PLS), a standard random forest (RF) algorithm, thin plate regression splines (TPRS), RF followed by TPRS (RF-TPRS), TPRS followed by RF (TPRS-RF), and a spatial random forest algorithm with optimization done either using pseudo-likelihood (SpatRF-PL) or nonparametrically (SpatRF-NP).

Algorithm 1: Leave-one-out variable importance

Set quantile levels $q_1, \ldots, q_m \in [0,1]$, and of interest. Input the data $(X_{n\times p}, Y_{n\times 1})$ and a trained model

$$g\left(\hat{\mu}(s)\right) = \sum_{k=1}^{K}\left[\hat{f}_k(X(s)) + \hat{\nu}_k(s)\right]$$

for $j = 1, \ldots, p$ **do**
  for $l = 1, \ldots, m$ **do**
    for $i = 1, \ldots, n$ **do**
      for $k = 1, \ldots, K$ **do**
        Calculate the fitted mean at all but the ith location as $\hat{\zeta}_k^{j,l}(\boldsymbol{s}_{-i}) := \hat{f}_k(X(\boldsymbol{s}_{-i}))$
        Replace the ith observation of the jth covariate, X_{ij}, with the q_l-th sample quantile $\hat{F}^{-1}_{X_j}(q_l)$:

$$\tilde{X}_{i\cdot} := \left(X_1(s_i), \ldots, \hat{F}^{-1}_{X_j}(q_l), \ldots, X_p(s_i)\right)$$

        Calculate the predicted mean at the ith location as $\hat{\zeta}_k^{j,l}(s_i) := \hat{f}_k(\tilde{X}_{i\cdot})$
        Re-fit a covariance model with $(X_{-i,\cdot}, Y_{-i})$ conditioning on the mean model $\hat{\zeta}_k^{j,l}(s)$, denoted as $\hat{\nu}^{j,l}_{(-i),k}(s)$
        Evaluate the covariance term for location i from the fitted model $\hat{\nu}_k^{j,l}(s_i) := \mathbb{E}\left[\nu^{j,l}_{(-i),k}(s_i) \mid Y(\boldsymbol{s}_{-i}), \hat{f}_k\left(X(\boldsymbol{s}_{-i})\right)\right]$
        Calculate the linear predictor for location i as $\hat{\eta}_k^{j,l}(s_i) := \hat{\zeta}_k^{j,l}(s_i) + \hat{\nu}_k^{j,l}(s_i)$
        Calculate the averaged leave-one-out predictions $\bar{\mu}_{j,l} := g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\eta}_k^{j,l}(s_i)\right)$
      end
    end
  end
end

Result: Output averaged leave-one-out predictions $\bar{\mu}_{j,l}$ for each $j = 1, \ldots, p$, and $l = 1, \ldots, m$.

followed by spatial smoothing, while SpatRF-PL conducts implicit degree-of-freedom control and jointly accounts for the mean and covariance components. To further investigate the difference between the two models, **Figure 8.2** shows the gridded prediction maps over the Seattle TRAP study region. This is based on the evaluation of each model at an additional set of 2,815 gridded locations within the same study region. The difference map on the third panel reveals that predictions made by UK-PLS and SpatRF-PL, although very similar at the mobile monitoring locations, exhibit different spatial patterns when predicted at many more locations. These observations indicate that different spatial prediction models may appear to be very similar on some scales or when restricted to certain areas, and the true underlying differences between them may not be observed in some evaluations.

We also compared the various machine learning models in the context of the upcoming health association analyses. **Figure 8.3** shows the predicted 5-year average UFP exposures and compares the primary UK-PLS model to the various alternative machine learning models for the 5,409 participants in the health analysis. There is considerable overlap in all the predictions, and all are highly correlated with UK-PLS, with Pearson correlations (r) between 0.97–0.99. The greatest differences between exposure model predictions occur above 10,000 pt/cm^3, and are most easily observed by comparing the smoothers on the plot.

Figure 8.4 shows the health association estimates and their 95% confidence intervals for the various exposure models using confounder model 1 (Equation 3.11), the cognitive function outcome, and the analysis approach described in

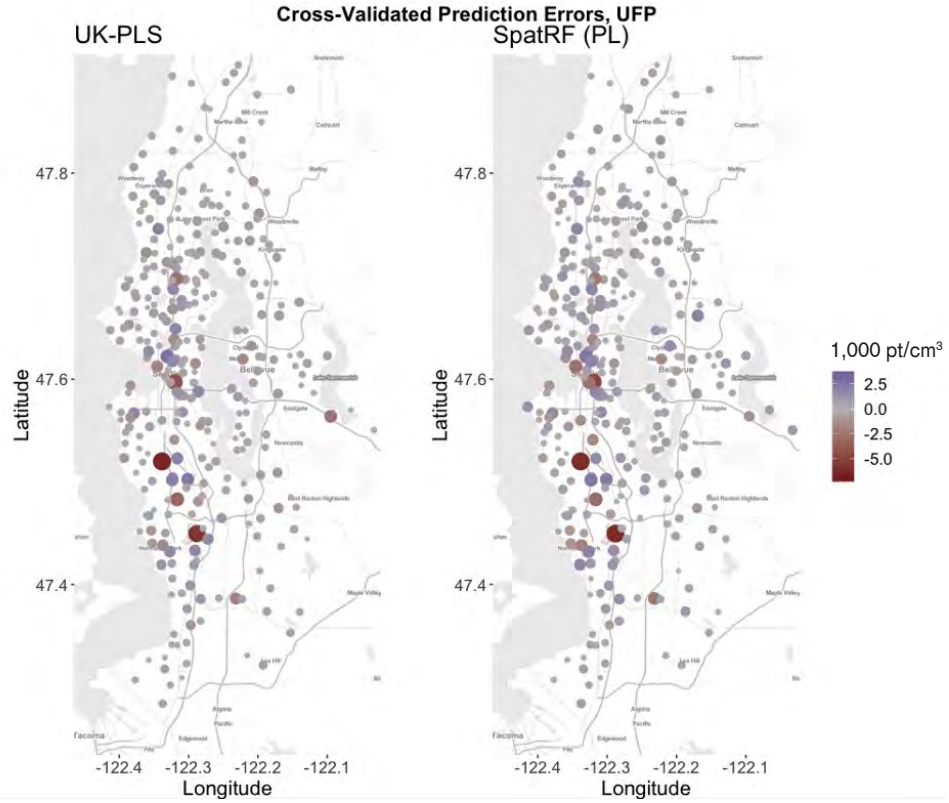


Figure 8.1. Cross-validated prediction errors of UFPs for UK-PLS and SpatRF-PL at each monitoring location for the Seattle stationary roadside mobile monitoring data. The shade, color, and size of the dots reflect the magnitude of the errors.

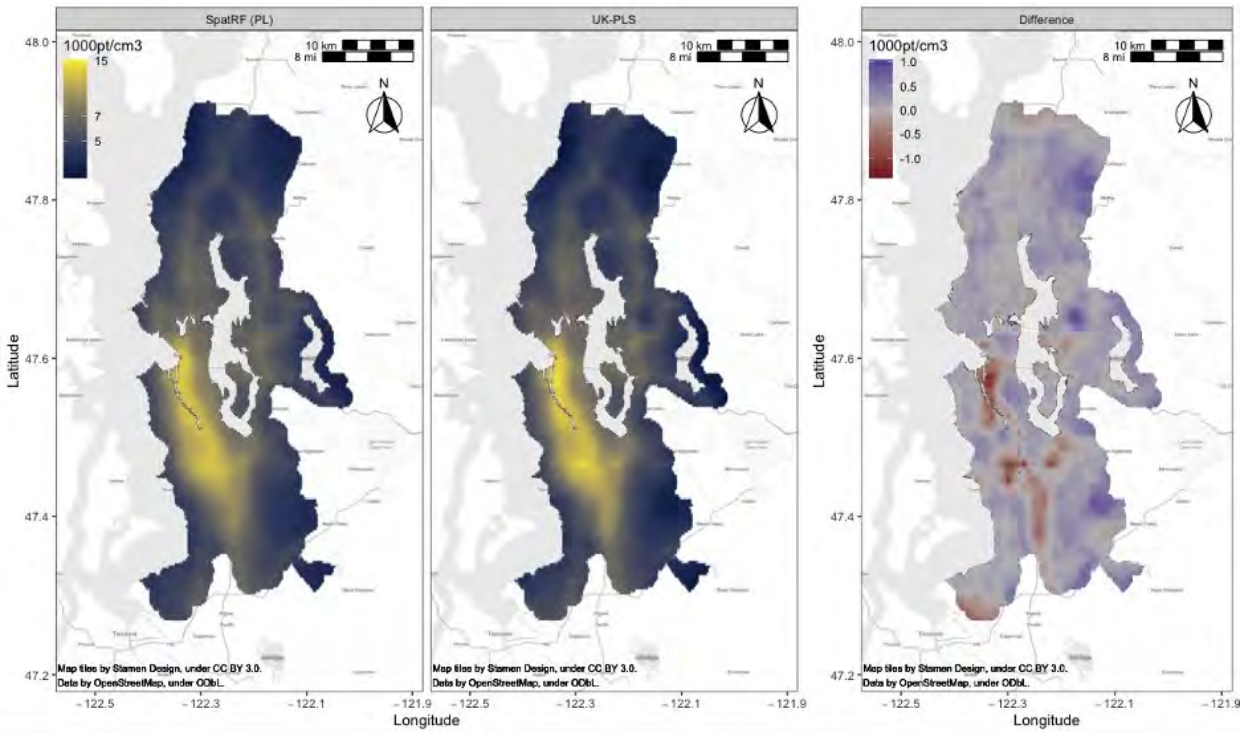
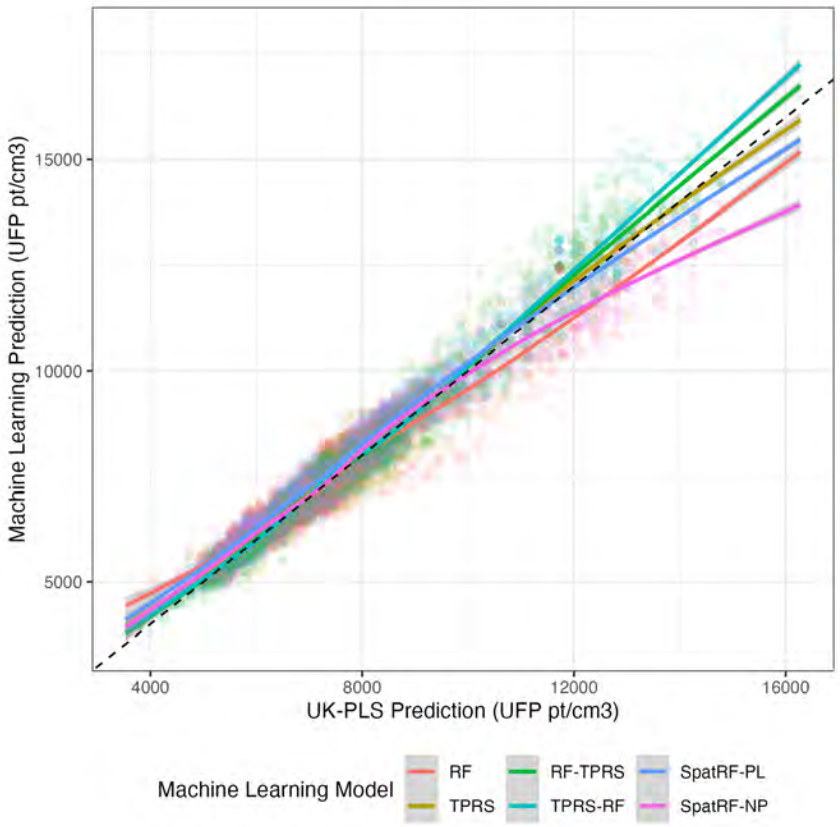


Figure 8.2. Predicted UFP concentration surfaces based on predictions at gridded locations derived from the Seattle mobile monitoring stationary roadside data. Predicted surfaces use SpatRF-PL (map 1), UK-PLS (map 2), and their difference (map 3; UK-PLS is the subtrahend).

Figure 8.3. Comparison of 5-year average UFP exposures for 5,409 participants in the health analyses predicted from the primary UK-PLS model and alternative machine learning models. Exposure models were developed from unscreened P-Trak instrument readings (20–1,000 nm particles) from the Seattle mobile monitoring stationary roadside data. All alternative machine learning model predictions were highly correlated with the main reference model (UK-PLS), with Pearson correlations between 0.97–0.99.



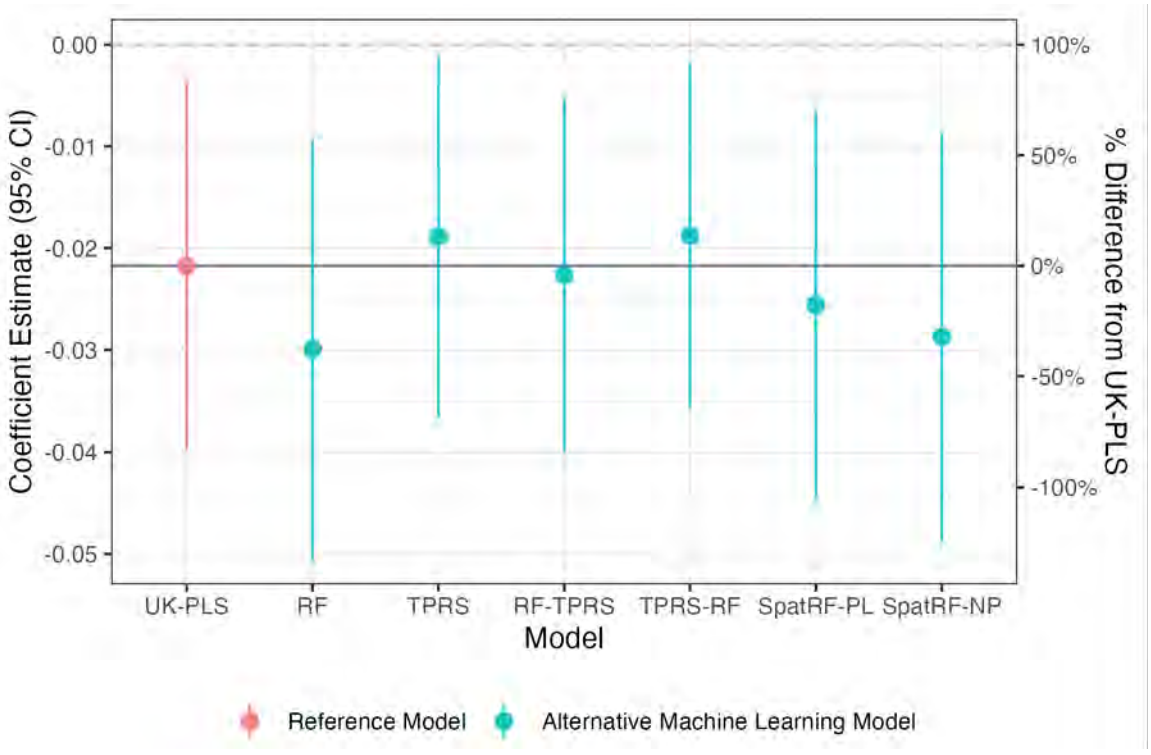


Figure 8.4. Estimated association (95% CI) between UFPs (1,900 pt/cm³) and cognitive function at baseline using various machine learning exposure assessment models. The associations are adjusted for age, calendar year, sex, and education (confounder model 1).

Chapter 3. The results from these various models were all very similar, consistent with their similar prediction model performances.

The similarity between the various spatial prediction models in this context motivates the evaluation of a variable importance metric that may facilitate investigation of the mechanisms that the various models capture, as well as aid in the selection and interpretation of models. This metric can be applied to a diverse class of prediction methods.

Results — Variable Importance Metric

We examined the proposed variable importance metric at each quartile of each predictor, that is, q_1, q_2, and q_3 are 0.25, 0.5, and 0.75, respectively. We compared the UK-PLS and SpatRF-PL models and looked at three contrasts to evaluate the contribution of each predictor: $\bar{\mu}_{j,2} - \bar{\mu}_{j,1}, \bar{\mu}_{j,3} - \bar{\mu}_{j,2}, \bar{\mu}_{j,3} - \bar{\mu}_{j,1}$. (The notation definition is given in Equation 8.4). **Figure 8.5** shows the contribution of predictors that had the greatest contribution to predicted UFP concentrations. Despite similar predicted maps between UK-PLS and SpatRF-PL, the plots highlight the difference in the mechanisms captured by each model. In particular, SpatRF-PL identifies the length of truck routes and closeness to major roads as major contributors to predicted UFP concentrations in the Seattle mobile monitoring campaign, while UK-PLS highlights the distance to a large airport as a more important contributor. As has been reported in prior studies, jet engine exhaust is a significant source of UFPs (Hudda et al. 2018), which suggests UK-PLS as a more sensible candidate predictive model in terms of scientific interpretation. In addition, the UK-PLS model is more consistently influenced by truck traffic and general traffic on large A1 roads than the SpatRF-PL predictions. It is also reasonable that a linear model based on such a large number of covariates would perform well in a relatively homogeneous and small area where relationships between sources and pollution levels are consistent across the domain. In other words, a linear approximation, even when the underlying relationship is nonlinear, is likely to perform well in this setting.

This variable importance metric also reveals the greedy nature of tree-building algorithms here. For instance, although both UK-PLS and SpatRF-PL find the length of truck routes within several buffer sizes to be important predictors of UFP concentrations, SpatRF-PL highlights only a few of these autocorrelated predictors, as indicated by the number of circles with a predictor contribution that deviates meaningfully from zero. This is in contrast to UK-PLS, which highlights all of them, as can be seen in Figure 8.5. Furthermore, the covariates and buffer sizes highlighted vary across quantile contrasts with SpatRF-PL, whereas this is more consistent across quantile contrasts for UK-PLS. This is related to the nonlinear property of SpatRF-PL (in contrast to the linear UK-PLS model); namely, the magnitude of effects of the same covariate could differ at different levels of its distribution.

The variable importance metric can also provide insight into how the performance of various fitted models differs on newly observed datapoints. We elaborate on this point in Chapter 8's Additional Materials and demonstrate figures that support the argument that models with similar behaviors on the training data (e.g., the monitoring sites) can have meaningful differences when extrapolated to new locations (e.g., the gridded locations). With the aid of our variable importance metric, the latter can be anticipated and captured by a variable importance analysis on just the training data.

MULTIPOLLUTANT PREDICTION FOR SPATIAL DATA

Introduction

The goal of multipollutant prediction is to predict an m-dimensional matrix of exposures $Y_{n\times m}$ given a potentially

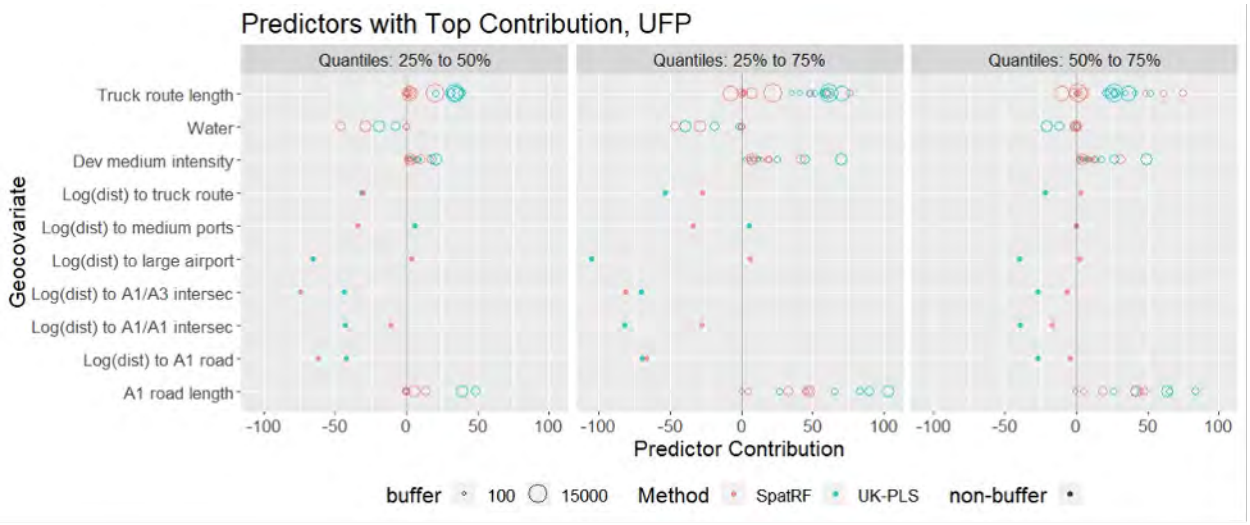


Figure 8.5. Variable importance plot for UFP concentration predictions, showing the predictors among the top five contributors for either method for at least one contrast. All buffer sizes were included if one of them was within the top five important predictors. Analyses based on the Seattle mobile monitoring stationary roadside data.

large-dimension matrix of covariates $X_{n\times p}$. In this case, m can be very large, and its components are often highly correlated. A typical approach is to use a low-dimensional approximation, $\tilde{Y}_{n\times r}$, such as the first r components of a principal component analysis (PCA). The sole focus of classical PCA is to prioritize representability, or minimal loss of information. However, in the context of spatial prediction analyses, predictability is very important. Jandarov and colleagues (2017) developed predictive PCA, which prioritizes the prediction of Y by X. The goal of this work is to balance representation and predictability. We propose a new metric, representative and predictive PCA, or RapPCA for short. We summarize the methods here, which are a distillation from a published preprint; see Cheng and colleagues (2024b) for additional details.

Methods

We refer to the following notation in this presentation:

- Y is a $n \times m$ matrix of exposure measurements,
- X is a $n \times p$ matrix of covariates,
- u is a $n \times 1$ vector of principal component scores,
- v is a $m \times 1$ vector of loadings,
- $\|\cdot\|_F$ represents a Frobenius norm for matrices,
- $\|\cdot\|_2$ represents a ℓ_2 norm for vectors, and
- α is a $p \times 1$ vector of parameters.

Note that while not explicitly indexed in the notation that follows, vectors u and v represent a sequence of vectors for multiple rank-1 optimizations to obtain multiple principal component decompositions. We focus on three principal components in the results.

The gist of this proposal is to combine into the proposed RapPCA two elements: classical PCA:

$$\min_{u,v}\|Y - uv^T\|_F^2 \quad \text{s.t.} \quad v^Tv = 1$$

and predictive PCA:

$$\min_{\alpha,v}\left\|Y - \left(\frac{\tilde{X}\alpha}{\|\tilde{X}\alpha\|_2}\right)v^T\right\|_F^2$$

where the columns of $\tilde{X}$ contain the covariates as well as thin-plate regression spline basis to capture the spatial effects. This gives the proposed RapPCA:

$$\min_{u,v,\alpha,\beta}\|Y - uv^T\|_F^2 + \gamma\|u - (K\alpha + B\beta)\|_2^2 + \lambda_1\alpha^TK\alpha + \lambda_2\beta^TQ\beta \quad \text{s.t.} \quad u = Yv, v^Tv = 1$$

where the four terms are for representability, predictability, and regularization (the last two terms), respectively. In this approach, the nonlinear covariate effects are modeled by the kernel matrix K: $K_{ij} = k(x_i, x_j)$ for a kernel function $k(\bullet,\bullet)$ and the spatial effects are captured by the spatial basis matrix B, with penalty matrix Q. Specifically, RapPCA interpolates between classical PCA and predictive PCA. It also includes classical PCA as a special case when $\gamma = \lambda_1 = \lambda_2 = 0$. To estimate the principal components (PCs), the vector Y is replaced by $Y - \tilde{u}\tilde{v}$ where $\tilde{u}$ and $\tilde{v}$ represent optimal solutions; this is repeated for multiple PCs. In the submitted methods paper (Cheng et al. 2024a), we provide the technical development of this method, including an analytical solution that attains the minimum despite the nonconvexity of the biconvex objective function. We also discuss three types of evaluation metrics: mean squared prediction error to quantify the prediction error, mean squared representation error to quantify the representation error, and the total mean squared error. These are computed on the training (*trn*) or test (*tst*) datasets, as indicated in the notation, and are given by

- **Mean squared prediction error:** $n_{tst}^{-1}\|(u_{tst}^* - \hat{u}_{tst})\tilde{v}^T\|_F^2$ where $u_{tst}^* = \arg\min_u\|Y_{tst} - u\tilde{v}^T\|_F^2$ is what the actual PC score on the test set would be, given the loadings $\tilde{v}$, if Y_{trn} were known. In other words, it characterizes the gap between the predicted u_{tst} and true PC scores and reflects the predictability of the PCs.
- **Mean squared representation error:** $n_{trn}^{-1}\|Y_{trn} - \tilde{u}_{trn}\tilde{v}^T\|_2^2$ measures the gap from approximating Y_{trn} in dimension reduction. Focusing on quantifying this on the training data allows us to assess the quality of the representation alone, without considering predictive performance.
- **Total mean squared error:** $n_{tst}^{-1}\|Y_{tst} - \hat{u}_{tst}\tilde{v}^T\|_F^2$ is the overall gap from both predicting $\tilde{u}$ and the dimension reduction of Y. This measures the discrepancy between the true data Y and the predicted scores that are transformed back to the original, higher-dimensional space.

Note that mean squared prediction error and mean squared representation error can be viewed informally as a decomposition of the overall total mean squared error.

To illustrate the approach, we considered measurements of UFPs, BC, and NO_2 from the 309 stationary roadside locations described in Chapter 3. Our goal was to use as much high-quality data as possible from the mobile campaign. Thus, because UFP concentrations were measured using multiple instruments and both UFPs and BC produced multiple measurements, we incorporated these multiple measurements into this analysis. We used 13 bin sizes from the NanoScan, the total particle concentration and median size estimates from the DiSCmini, and two P-Trak measurements: the total from the instrument with the diffusion screen (size range 36–1,000 nm) and the difference between the instruments with and without the diffusion screen (size range 20–36 nm). We used five wavelength measurements of BC from the micro-aethalometer: blue, green, infrared, red, and ultraviolet. We transformed the ultraviolet measurements to represent the difference between the ultraviolet and infrared ranges. We used the NO_2 concentration to normalize all other measurements, other than the median size measure from the DiSCmini. All variables were centered and scaled. This gave us $p = 23$ annual average measurements of pollutant concentrations at the 309 locations. We ran dimension reduction with each of the PCA algorithms to assess the spatial distri-

bution of the top three PC scores. For RapPCA, we selected the tuning parameters $\gamma, \lambda_1, \lambda_2$ that produced the optimal total mean squared error. We used 10-fold cross-validation to train spatial random forest prediction models (Wai et al. 2020) followed by spatial smoothing via TPRS, and then predicted the top three PC scores and evaluated them on a grid of 5,040 locations. In the results section, we show these smoothed PC scores on maps over the study region as well as the PC loadings, which reflect the contribution of each pollutant to the PC scores.

Results

Table 8.2 summarizes the overall prediction evaluation metrics for each of the three PCA approaches (classical PCA [PCA], predictive PCA [PredPCA], and representative and predictive PCA [RapPCA]). Comparisons should be made across the PCA approaches within each metric column. We observe that RapPCA has the best total predictive performance and also minimizes the prediction error. As expected, classical PCA performs slightly better than RapPCA in minimizing representation error.

Figure 8.6 shows the smoothed PC scores obtained by each method across the study region. We observe similar spatial patterns in the distribution of PC scores across the different methods, except for the south end of the study region for the third PC, where PCA identifies a stronger signal than either RapPCA or predictive PCA. In particular, all methods highlight regions near the airport and around major roads (the dark red area) for the first PC, indicating aircraft and road traffic emissions as a major source of the overall pollution level. This is consistent with the large contributions of BC as well as UFPs with small or moderate sizes. This is also reflected by the UFP and BC loadings in Figure 8.7.

Figure 8.7 shows the PC loadings for each pollutant measurement after applying the three different PCA methods. Roughly grouping the UFPs into small, moderate, and large particle size groups, we observe that RapPCA identifies one size group to be the main contributor to each of the PCs, while the other two groups have smaller or negligible loadings. Specifically, small particles contribute to PC1, moderately sized particles contribute to PC2, and the large ones contribute to PC3. In addition, BC mainly contributes to PC2 in RapPCA. This observation is consistent with findings from other studies where two key sources of TRAP were identified as aircraft emissions, which are primarily composed of the smallest UFPs, and road traffic, especially diesel exhaust, which is comprised of moderately sized UFPs and BC. In contrast, neither classical PCA nor predictive PCA distinguishes the contributions of different sizes of UFPs or BC as effectively.

DISCUSSION AND CONCLUSIONS

Application of Spatial Ensemble-Learning Methods and New Variable Importance Metric

We applied several spatial machine-learning algorithms, focusing most on UK-PLS and a recently developed spatial random forest approach, to predict exposure throughout the Seattle mobile monitoring region. We compared the results of these two approaches in terms of standard prediction model performance measures, health association estimates, and a new variable importance metric we developed in this project. We found that although the two models had comparable prediction performances and health association estimates, the variable importance metric indicated that some geographic features, specifically distance to large airport, distance to main roads, and length of truck routes, were differently important contributors to the predictions of UFP concentrations from UK-PLS and SpatRF-PL. Given that epidemiological cohort studies rely on predicted air pollution exposures for making inferences about health associations, the use of this variable importance metric has the potential to allow new insights into how seemingly similar exposure metrics may differ in practice. However, in this work, the association of cognitive function and UFPs from these various exposure metrics did not meaningfully differ.

The variable importance metric we developed is flexible, intuitive, and generally applicable to machine learning models that account for spatial correlation. This leave-one-out approach can be applied to additive models with separable mean and correlation components, including nonlinear, ensemble, or doubly stochastic spatial models. It provides

Table 8.2. Comparison of the Overall Metrics and Individual Prediction Mean Squared Errors (MSEs) for Each Principal Component (PC), Assessed by 10-Fold Cross-Validation[a]

	Overall Metrics			Individual MSEs		
	Total Mean Squared Error	Mean Squared Prediction Error	Mean squared Representation Error - Training	PC1	PC2	PC3
PCA	14.79	7.93	6.38	3.93	2.86	1.14
PredPCA	14.81	7.16	7.28	3.41	2.53	1.22
RapPCA	13.92	6.66	6.75	2.62	2.92	1.12

[a] These were developed using the Seattle mobile monitoring roadside data after normalization to NO_2 along with centering and scaling, and obtained from classical PCA (PCA), predictive PCA (PredPCA), and representative and predictive PCA (RapPCA), respectively.

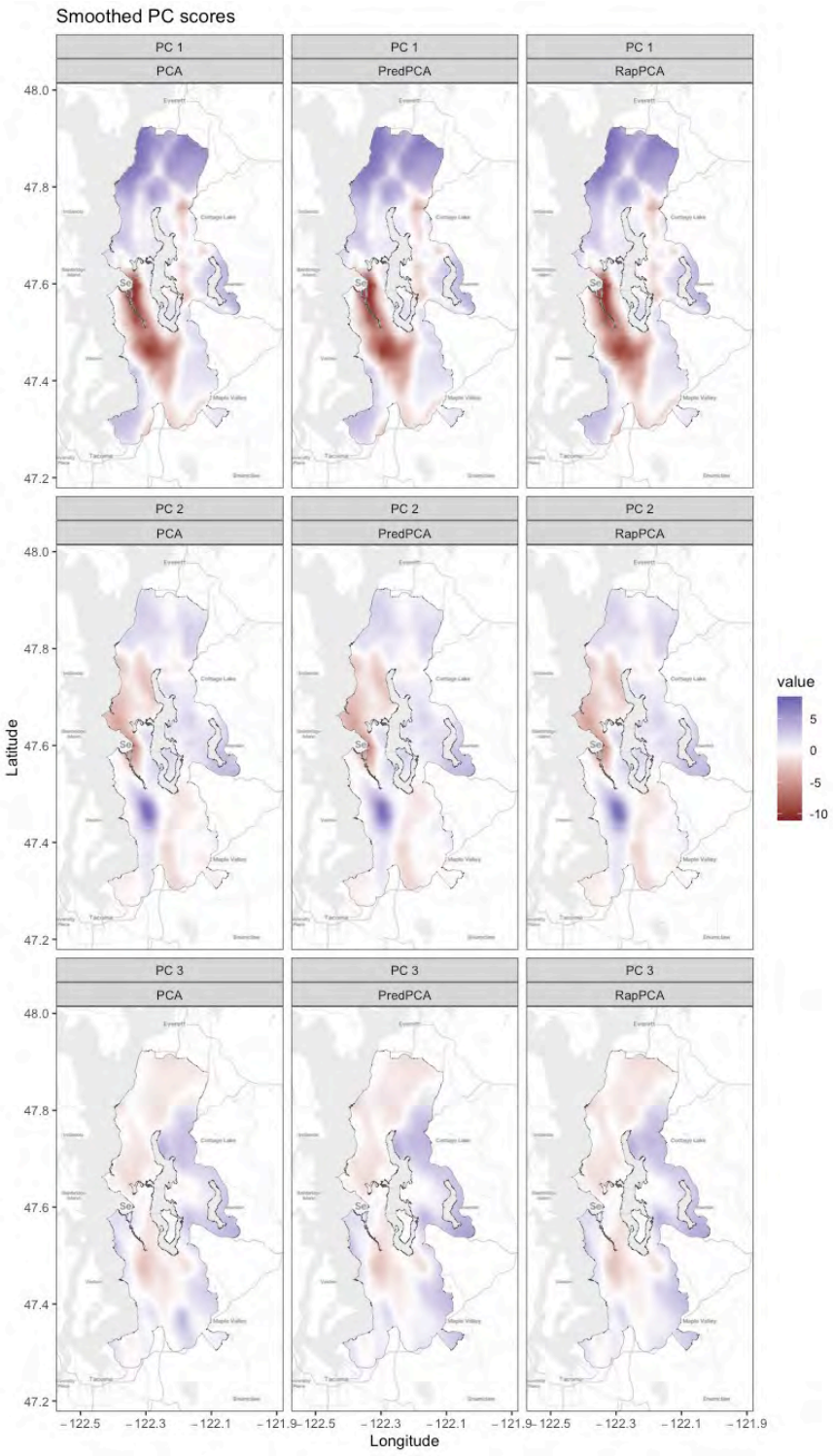


Figure 8.6. Smoothed principal component scores for the first three principal components. Components were developed on the Seattle mobile monitoring roadside data, and scores were determined from classical PCA (PCA), predictive PCA (PredPCA), and representative and predictive PCA (RapPCA), respectively.

a unifying notion of variable importance, which would otherwise be less comparable between different modeling approaches, and we have demonstrated that meaningful differences in the model structure can be found even for models producing similar predictions. An informative variable importance metric such as ours also facilitates a deeper understanding of complex prediction models from a methodological perspective. For instance, while our example illustrates the already well-understood greedy nature of tree-building algorithms, future applications of this variable importance metric may help reveal the behavior of more complex black-box machine learning models.

Multipollutant Dimension Reduction for Prediction

We presented a dimension reduction algorithm that allows flexible interpolation between the representability and predictability objectives of dimension reduction of multipollutant data for settings where there is a need to predict these data at unmeasured locations. This balance between prediction and representation is a generalized form of the bias-variance trade-off. While classical PCA minimizes the representation gap in the training data, it may also capture excessive noise by explaining too much of the variation in the data. In contrast, predictive PCA restricts the PC scores to fall within a certain model space — the linear span of the covariates, which is a form of regularization enforcing the smoothness of the PCs. The RapPCA seeks the optimal balance between these two aspects in an explicit and interpretable way. Our empirical evaluation in the Seattle mobile monitoring campaign data (and in another dataset described in the underlying paper [Cheng et al. 2024a]) suggests that striking this balance has the potential to improve, and won't diminish, the total mean squared error in comparison to existing alternatives.

To fully understand the implications of this newly developed mixture model approach, it will be important to apply it to a health analysis, both using a real-world example, such as the ACT cognitive function cross-sectional analysis, and using a simulated example where the true impact of the mixture on a health outcome is known. This is planned future research.

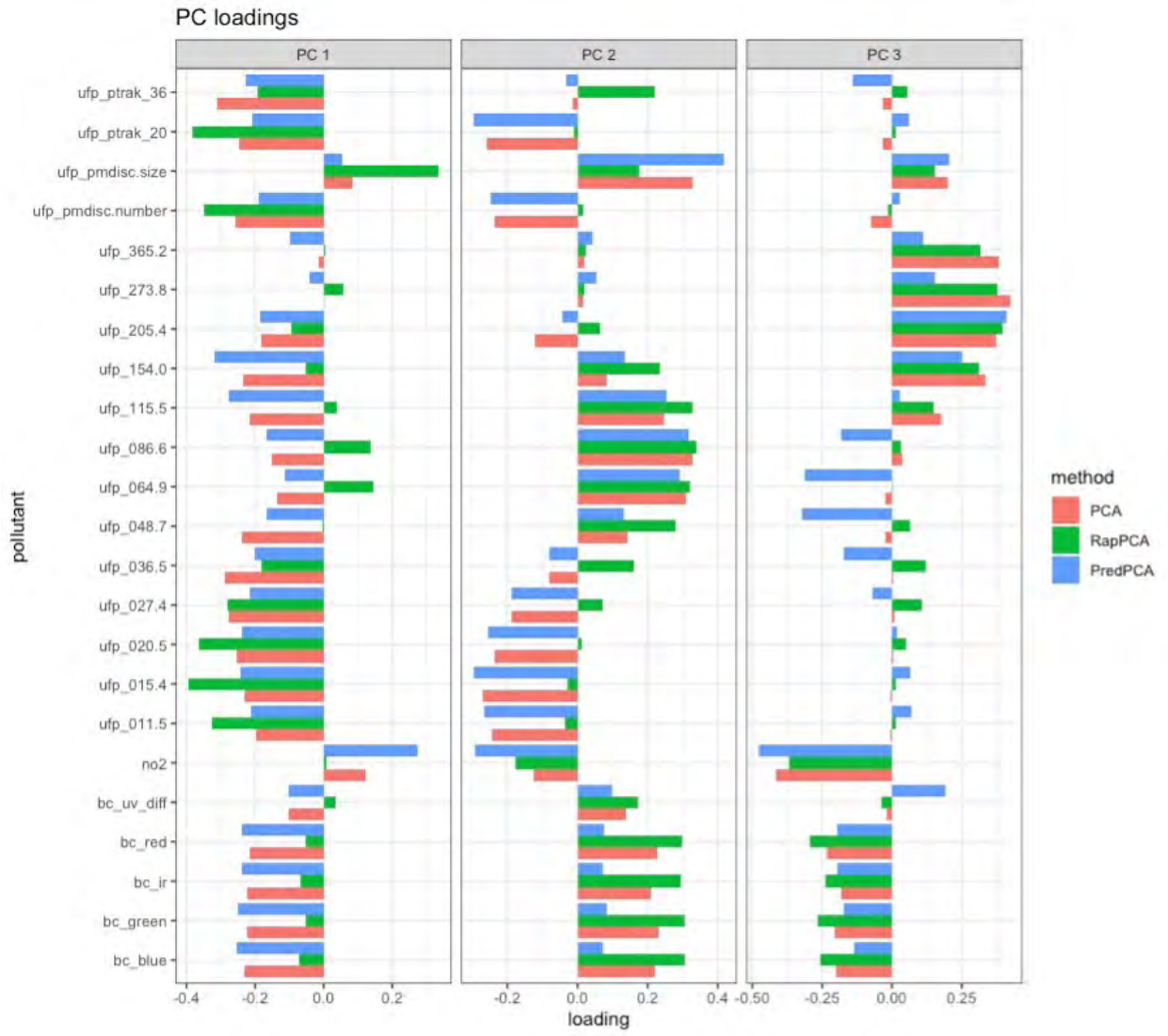


Figure 8.7. The first three PC loadings for each pollutant. These were developed on the Seattle mobile monitoring roadside data and obtained from classical PCA (PCA), predictive PCA (PredPCA), and representative and predictive PCA (RapPCA), respectively. There are three types of pollutants, where the suffix, if applicable, represents the properties of, or the instruments used to measure, each pollutant. In particular, the numeric suffix after ufp_ corresponds to the lowest value of the size range. Specifically, ufp_ptrak_36 represents P-Trak measurements with the diffusion screen (36–1000 nm) while ufp_ptrak_20 represents the difference between P-Trak measurements with and without the diffusion screen (20–36 nm). BC measurements at different wavelengths are labeled as: blue (bc_blue), green (bc_green), infrared (bc_ir), red (bc_red), and ultraviolet (bc_uv). The ultraviolet measurements were transformed to represent the difference (bc_uv_diff) between the ultraviolet and infrared ranges.

CHAPTER 9: EXPOSURE MONITORING STUDY DESIGNS FOR EPIDEMIOLOGY: COST AND PERFORMANCE COMPARISONS

Lead authors: Sun-Young Kim, Amanda J. Gassett, Magali N. Blanco, Lianne Sheppard

OVERVIEW

This chapter has two sections. The first section, *Cost and Performance Comparisons for Low-Cost Sensors Supplementing Regulatory Monitoring Data*, describes our work on the value of information in supplementary low-cost sensor campaigns. It is a summary of the write-up included in Chapter 9's Additional Materials. The second section*, Mobile Monitoring Study Designs for Epidemiology*, provides a more extensive presentation as it is a complete published manuscript (Kim et al. 2025).

COST AND PERFORMANCE COMPARISONS FOR LOW-COST SENSORS SUPPLEMENTING REGULATORY MONITORING DATA

We combined existing regulatory monitoring datasets with various subsets of additional low-cost sensor data to determine how cost-effective it is to collect additional low-cost sensor data as a supplement to regulatory monitoring data, to predict $PM_{2.5}$ and NO_2 concentrations for application to epidemiological cohort studies. Leveraging the data and modeling approaches summarized in Chapters 3 and 7 for these two pollutants, we considered designs that used various subsets of the supplementary data available in the full datasets. The NO_2 dataset had two supplementary datasets: the snapshot campaign using Ogawa samplers that deployed all Ogawa samplers simultaneously in each of three seasons, and the home-based low-cost sensor campaign that deployed a rotating set of monitors predominantly at participant homes for 2-week and longer periods. The $PM_{2.5}$ dataset only had the home-based low-cost sensor campaign. Chapter 9's Additional Materials describes the specific subsets of data we considered for these value of information analyses and the results as characterized by (1) differences in model prediction performances, and (2) differences in cost as quantified by the additional staff days required to deploy and take down instruments, as well as recruit participants, obtain consent, and schedule visits at home locations.

From an exposure model predictive performance perspective, results from both the $PM_{2.5}$ and NO_2 analyses indicate that home-based monitoring adds value to models developed for epidemiological purposes. For instance, for $PM_{2.5}$, there was considerable improvement of nearly all models with home data included, compared to a model with no home data (Table S9.7). For NO_2, all the models that considered home data had reasonably good performance when compared to the gold-standard model (Table S9.6).

However, home-based monitoring is also very expensive, as noted by the number of staff days required for sampling in Table S9.8. In addition to requiring more staff time, home-based monitoring also potentially involves different staff positions, because participants need to be (1) contacted for recruitment, (2) complete a consent process, and (3) be present when staff are at their home to provide access to their outdoor spaces for monitoring set-up and take-down. Home monitoring data are also subject to institutional review board (IRB) oversight, and these measurements cannot be publicly shared because that would violate participants' privacy. In previous studies, we have found that a team of two field technicians could reasonably plan to visit about five homes per workday. Thus, this sampling approach adds considerable staff time for every home location included.

In contrast, using a snapshot design with Ogawa passive samplers to measure NO_2, a two-person team can typically visit 40–60 utility pole locations in a day, making it possible to collect data from more locations much more quickly. As shown in Table S9.6 and focusing on performance statistics at the snapshot locations, we observed similar performances among the snapshot variations that included observations from all three seasons for each included location, though somewhat worse for the model with locations dropped at random rather than by cluster. We observed much larger reductions in performance for snapshot designs with fewer seasons, particularly for some season combinations. Performances for the snapshot campaigns were also worse when evaluated at home locations, particularly for the mean square error (MSE)-based R^2, which evaluates model fit around the 1:1 line. This poorer performance was driven more by bias than poor correlation, as the regression-based R^2s at the homes were much more similar to the MSE-based R^2s at snapshot locations. In other words, the limited seasonal information represented by the snapshot campaign's results in predictions that can be systematically offset from the 1:1 line, as is shown in the example in Figure S9.3.

Exposure assessments that included home monitoring required at least three times the number of staff days as the most time-intensive snapshot (Table S9.8). Thus, while snapshot campaign locations are not as spatially compatible with the target epidemiological cohort as those that can be selected in a home-based low-cost sensor campaign, this is a much more cost-effective approach. However, passive samplers such as the Ogawa badges that we used are only feasible for the collection of gases. For $PM_{2.5}$, there are a few good alternatives to home monitoring.

To summarize, in some settings, it may not be feasible to collect air pollution measurements at participant homes or on a rolling basis that allows measurement throughout the year. Particularly in multisite designs where monitoring is conducted by excursions from the research center to the

study sites, data collection designs are often logistically constrained. Snapshot designs are an appealing solution for measuring gases, but the snapshot designs in our study performed more poorly than anticipated. Based on the results presented in Table S9.6, we recommend reducing the number of NO_2 snapshot locations before reducing the number of snapshot monitoring seasons whenever possible. In contrast, the $PM_{2.5}$ results suggest that given a study design with home monitoring that covers the full temporal extent of the possible monitoring period, it is preferable to have fewer repeat visits to the same location rather than reducing the number of locations. It is also preferable to spread out those monitoring locations over space as much as is feasible.

MOBILE MONITORING STUDY DESIGNS FOR EPIDEMIOLOGY

Introduction

While mobile monitoring has been applied to assessing air pollution emitted from traffic or wildfires for decades (Austin et al. 2021; Boanini et al. 2021; Larson et al. 2007; Loeppky et al. 2013; Pirjola et al. 2012; Wagstaff et al. 2022), more recent mobile monitoring studies have focused on exposure assessment to investigate the association between long-term exposure to air pollution and human health (Kerckhoffs et al. 2016; Messier et al. 2018). Specifically, recent mobile monitoring campaigns aimed to achieve high-quality estimates of people's long-term average exposures to traffic-related air pollution (TRAP), given the increasing epidemiological evidence of adverse TRAP associations (Boogaard et al. 2022; Kerckhoffs et al. 2019; Weichenthal et al. 2016). To achieve this goal, investigators designed or leveraged existing mobile monitoring campaigns and subsequently used air pollution prediction models to provide estimates of individual-level long-term average exposures.

As mobile monitoring allows substantial flexibility with regard to the temporal frequency and timing of sampling, as well as the spatial domain, there are many available choices in designing a monitoring campaign. Some recent studies focused on quantifying the accuracy of individual long-term exposure estimates (Apte et al. 2017; Blanco et al. 2023a; Kerckhoffs et al. 2024; Klompmaker et al. 2015; Messier et al. 2018). Cost is another important factor that drives monitoring design, particularly given the high expense of payroll and equipment in mobile monitoring. Cost can vary dramatically as a function of logistical features; some designs optimized to achieve high accuracy in exposure estimates may be too expensive to implement. Thus, an investigation of the trade-off between the increase in cost and the improvement of prediction model performance can help investigators design their mobile monitoring campaigns in the future. Our previous study suggested that greater spatial and temporal coverage in mobile monitoring designs can improve prediction accuracy (Blanco et al. 2023a), as described in the current report. This study expands our focus to include cost to offer cost-effective design options to future air pollution mobile monitoring campaigns, and has been published (Kim et al. 2025).

Focusing on the goal of providing practical guidance, this study aims to compare the costs of input resources with the resulting exposure model performance by comparing and contrasting characteristics of mobile monitoring designs for application to environmental epidemiology. We assume that the ultimate exposure assessment goal in this setting is to produce high-quality predictions of annual average concentrations at study participants' residences to assess their health associations in a single geographically defined area. We specifically focus on UFPs quantified using a spatial prediction model. As UFP monitoring equipment can be very expensive and no regulatory monitoring networks include UFP measurements, a mobile platform carrying a few instruments has been widely considered to be a cost-effective option (Gozzi et al. 2016; Presto et al. 2021; Vallabani et al. 2023). Despite the increasing attention, there has been limited guidance on effective monitoring designs for epidemiological applications. To quantify costs and compare them with model performances, we leverage our experience in monitoring and modeling short-term stationary roadside measurements of UFPs that were produced under the auspices of the Adult Changes in Thought Air Pollution (ACT-AP) study.

Methods

ACT-AP Study Overview and Assumptions for the Present Study As a part of the ACT-AP study, an extensive mobile monitoring campaign has been conducted to characterize TRAP with high temporal and spatial resolution (Blanco et al. 2022). The mobile monitoring included stationary roadside and nonstationary (on-road) measurements of particulate and gaseous pollutants from multiple instruments loaded on a single vehicle that drove along nine fixed routes with a total length of 1,069 km in the 1,200 km^2 study area. Driving hours were between 5 a.m. and 11 p.m., including business and nonbusiness hours, weekdays and weekends, and all four seasons from March 2019 through March 2020. The monitoring platform did not operate from midnight to 5 a.m. because of logistical difficulties. We used the canonical calendar seasons: spring from March 20, 2019, to June 20, 2019, summer from June 21 to September 22, autumn from September 23 to December 20, and winter from December 21, 2019, to March 19, 2020 (with average seasonal temperatures of 8, 19, 12, and 5 °C, respectively). The stationary measurements, at 1 second to 1 minute depending on the pollutant, were collected at 309 roadside locations for 2 minutes at each site, with ~29 repeat visits per site. This study also carried out makeup sampling for missing data due to instrumentation, inclement weather, car trouble, driver illness, or driver error. As this campaign included most hours and all seasons and days of the week at each sampling site, we considered the annual average of measurements at each site without additional temporal adjustment as the representative annual average concentration. For particle number concentration (PNC) often used to

characterize UFPs, we computed the median concentration of each 2-minute visit, winsorized medians across visits within each site using the 5% threshold level, and computed averages at each site. The instrumentation details, including specifications and limits of quantification as well as quality control procedures, are provided in our previously published study and also described in Chapter 3 of the current report (Blanco et al. 2022).

Design Components Using the available mobile monitoring data, the present study quantified features of the complete and alternative (subsampled) monitoring campaign designs for UFPs sampled from an unscreened TSI P-Trak 8525. We choose a single instrument and location scenario to focus more directly on design components and omit from consideration differences in cost derived from instrument selection and travel to or from multiple study areas. We selected the P-Trak because it has been commonly used in previous mobile monitoring campaigns and produces high-quality data at a 1-second time scale, as we have documented previously (Blanco et al. 2022, 2023a; Doubleday et al. 2023).

We considered five design components in this investigation: number of sites, number of visits per site, days of the week, hours of the day, and number of seasons (**Table 9.1**). These five components indicate key spatial and temporal characteristics that contribute to assessing long-term average exposure of air pollution for cohort participants, but are commonly limited in previous monitoring campaigns given logistical and financial constraints. The complete all data design included 309 stationary roadside sites, ~29 visits per site, all 7 days of the week, most hours of the day (between 5 a.m. and 11 p.m.), and all four seasons. Alternative designs reduced the number of sites to 100, 150, 200, or 250 (i.e., spatially reduced), or reduced the number of visits per site to 4, 6, 12, 16, 20, or 24 (i.e., temporally reduced), to provide a wide range of values that could be used to visualize the shape of the association between amount of monitoring and model performance. In temporally restricted designs, we restricted the days of the week and 5 a.m. to 11 p.m. sampling hours to (1) most hours on weekdays with no additional temporal restriction, (2) weekday business hours from 9 a.m. to 5 p.m., and (3) weekday rush hours from 7 to 10 a.m. and 3 to 6 p.m. The alternative designs related to the season included two or three seasons with no other restrictions. The designs that restricted the numbers of seasons, days, or hours used consistent numbers of sites and visits per site, 309 and 12, respectively, to separate the effect of selective timing from the effect of the total amount of monitoring. We did not consider the following designs: weekends only, one season, or an unbalanced number of visits per site. The first two are unrealistic options for assessing long-term average exposure to air pollution, and the last, while common in the literature, is left for future work. To assess the representative model performance of each alternative design, we constructed 30 campaigns of each design by random sampling from the complete all data design.

Approach to Cost Estimation We assumed that the total cost of the mobile monitoring campaign is composed of two types: *up-front* and *per-drive-day* costs. Table S9.1 shows the characteristics and examples of up-front and per-drive-day costs. The up-front cost is defined as one-time or long-term costs mostly incurred before the beginning of regular monitoring. Examples include the purchase of instruments and various preparation efforts such as fabrication of the manifold, software development, protocol development, maintenance, and establishment of sampling locations. Together, these activities required about 40 workdays of 1 full-time and 3–5 part-time staff in the ACT-AP study. The actual number of days will vary depending on the experience and training status of staff as well as the presence of existing materials and protocols. Most of these elements did not vary by monitoring design. The exception to this was a small amount of incremental time needed to establish sampling locations, which included site-specific selection, review, and documentation, plus time needed for pilot drives of each route.

The per-drive-day cost includes the cost of staff time for planned driving, makeup sampling, related in-laboratory quality control activities, vehicle use, and data management. The per-day vehicle costs were based on our university's long-term or short-term rental, mileage, and fuel rates for fleet services and the monthly parking rate for our building's

Table 9.1. Alternative and Reference Roadside Mobile Monitoring Designs to Assess Model Performance and Costs

Design Component	Complete or Reduced Reference Design	Alternative Monitoring Design	Versions of Alternative Monitoring Design
Number of sites	309	Fewer sites	100, 150, 200, or 250
Number of visits	29[a] or 12	Fewer visits per site	4, 6, 12, 16, 20, or 24
Day of the week	All days	Fewer days	Weekdays only
Hour	Most hours for 5 a.m. to 11 p.m.	Fewer hours	Business or rush hours only[b]
Season	4 seasons	Fewer seasons	2 or 3 seasons

[a] Median with a range of 26–35.
[b] Monday to Friday only.

lab. Because the P-Trak provides direct real-time measurements, which were downloaded after each day of sampling, laboratory data management activities were limited to data review and cleaning as described in Table S9.1. We primarily assumed that the staff cost of a single workday was constant regardless of the hours of day or days of week worked, because we funded multiple full-time staff members instead of paying overtime. We quantified the staffing effort by using the number of workdays as a unit of cost measurement, which can also be expressed as a fraction or multiplier of a year of effort from a full-time staff person.

We estimated that a single full-time staff person should work 229 days in a year, based on the assumptions that a person works 5 days per week and has 3 weeks' paid vacation, 12 paid holidays, and 5 sick days. The calculations of required workdays for the full single-instrument single-city monitoring design based on ACT-AP are shown in **Table 9.2**. Given a single instrument, we estimated that 203 days of regular driving would be required to make 29 visits to all sites on seven routes. Each route was designed to be driven in 1 day (approximately 7.5 hours driving plus 15 minutes each for set-up and take-down) and included an average of 44 sites. We estimated 2 days per season for rescheduled drives to accommodate sick time, car trouble, inclement weather, or other unforeseen circumstances that prevent driving per the original plan. We also assumed that drivers spent 1 day in every 10 sampling days for in-lab quality control activities, including calibration, lab-based co-location, administrative tasks, meetings, and car or instrument repair. After aggregating all of these workdays in the complete design, we obtained a total between 232 and 244 workdays, or between 1.0 and 1.1 (232/229–244/229) field technicians or drivers needed for a year. For alternative designs, the number of workdays was reduced according to the difference from the complete design. Our cost estimation did not consider analysis work for exposure model development or collaboration efforts with expert groups, as these would vary depending on the scientific aims proposed for the use of the data.

We estimated monetary values corresponding to the number of workdays for alternative and complete designs based on actual expenditures in the ACT-AP study. Here, we have also highlighted monitoring elements that could incur additional costs: multiple instruments and the potential for a shift premium. For the multiple instruments scenario, we added two additional instruments for PNC (NanoScan and DiSCmini) and four instruments for black carbon (BC), nitrogen dioxide (NO_2), carbon dioxide, and carbon monoxide (microAeth MA200, CAPS NO_2, LI-850, and Langan, respectively). While additional PNC instruments are useful to verify PNC measurements, other pollutants can help adjust localized on-road plumes to assess representative residential exposure to UFPs as described earlier (Doubleday et al. 2023). A temporally varying shift-premium scenario illustrates the impact of a 30%, 50%, or 100% higher staff cost rate for weekend and evening driving than for weekday daytime driving.

Exposure Model Performance We computed R^2 values from 10-fold cross-validation (CV) of the spatial prediction model for complete all data as well as alternative designs using the observations collected under the complete all data design of the ACT-AP study. The annual average UFP spatial prediction model was based on the previously published universal kriging with partial least squares (UK-PLS) framework that reduces the dimension of hundreds of geographic covariates to a few predictors estimated by partial least squares, followed by spatial smoothing based on universal kriging (Sampson et al. 2013). The details of model specification and implementation were previously provided in the report and the published paper (Blanco et al. 2023a). In our CV, we computed annual average predictions at all 309 sites for all alternative and complete designs, except for the fewer sites design. We divided the 309 sites in each alternative

Table 9.2. Calculation for the Expected Number of Workdays Needed for the Full Single-Instrument and Single-City Monitoring Design of the ACT-AP Study

	Type of Cost	Number of Working Days	Equation
Available working days (per person)		229	365 days – [52 weeks × 2 weekend days[a] + 12 paid holidays + 3 weeks × 5 paid vacation days + 5 sick days]
Required working days	Up-front	40[b]	
	Per-drive-day: total	232–244	
	Regular drives	203	7 routes × 29 visits
	Rescheduled drives	8–20	2–5 days × 4 seasons
	Quality control	21	(203+8) drive days × 1/10[c]

[a] Two mid-weekdays when weekend driving is performed.
[b] Protocol development, fabrication, and assembly of the platform, and pilot testing. Examples include writing standard operating procedures, developing the instrument set-up checklist, testing and troubleshooting the instruments and the set-up inside the vehicle, working with information technology to develop the data organization procedure, and other preparation activities.
[c] Quality control activities every ~10 days.

design into 10 groups, developed the prediction model using nine groups of sites after holding out one group as test sites, predicted annual-average UFP concentrations at test sites, and repeated this procedure for each of the remaining nine groups. For the fewer sites design, we obtained predictions using a combination of CV at the 100–250 sites included in that design and external validation for the rest of the sites not included. For example, in the 250-site design, we computed CV predictions at 250 sites and then applied the model based on all 250 sites to obtain external predictions at the remaining 59 sites. For CV statistics, we always calculated mean squared error (MSE)-based R^2 values for each alternative design from predictions at all 309 sites compared to the annual average observations from the complete all data design, instead of the observations from the corresponding alternative design. We considered these observations at 309 sites from the complete all data design as the gold standard to validate all alternative designs against. MSE R^2s evaluate whether predictions and observations are the same (i.e., are near to the one-to-one line); they are computed as 1 minus the ratio of MSE divided by data variance (scaled by n, not n – 1), as opposed to more common regression-based R^2s that are calculated as squared Pearson correlation coefficients to assess whether pairs of observations are linearly associated. Because we constructed each alternative design by resampling 30 campaigns, we presented median CV MSE R^2 and its 5th and 95th percentiles across 30 campaigns for each design.

Comparison of Costs and Model Performance Between Different Designs

We compared costs (computed as workdays) to model performances (presented as CV MSE R^2s) across various alternative designs (**Table 9.3**). Specifically, the comparison focused on the decrease of CV MSE R^2s and the decrease of workdays in alternative designs compared to the relevant reference design. This comparison indicates the trade-off between the deterioration in model performance and the cost reduction. For the reference design, we used the complete all data design (309 sites and ~29 visits per site) to compare to alternative designs that were spatially reduced with fewer sites (100, 150, 200, or 250) or temporally reduced with fewer visits per site (4, 6, 12, 16, 20, or 24). For all the other alternative designs with temporal restriction(s), as in those with fewer days (weekday only), hours (business or rush hours only), or seasons (two or three seasons), the reference design was a reduced design with 12 visits per site (309 sites and 12 visits) as the "complete all data" design of temporally restricted designs. We selected this reference based on both feasibility as well as reasonably good model performance in our previous study (Blanco et al. 2023a). For example, two seasons only have 114 days (229 available workdays/2), and it would not be physically possible to drive 7 routes more than 12 times over two seasons using a single instrument platform (12 repeat visits × 7 routes regular driving + 8 makeup driving + 21 quality control = 113 days). Therefore, a realistic two-season design inherently requires fewer repeat visits with a realistic maximum of 12 visits compared to the complete all data design, or it would require much higher up-front costs to purchase more equipment.

All statistical analyses, including exposure model performance and comparison to cost, were performed in R (v. 4.1.2) (R Core Team, 2023).

Results

Table 9.3 shows the summary of design components, costs, and model performances. The complete all data design showed the highest CV MSE R^2 of 0.77. CV MSE R^2s decreased as the number of sites, visits, seasons, days, and hours decreased, showing the largest declines for designs with limited repeat visits and hours (median CV MSE R^2 = 0.60 for 4 visits per site and 0.55 for business hours only). The number of workdays was also the largest in the complete all data reference design (232 days), with linear decreases for fewer sites or visits designs (73 and 40 days for 100 sites and 4 visits per site, respectively).

The relationships between cost and model performance varied across design components (Table 9.3, Table S9.2, **Figure 9.1**). In particular, we found a contrast in this relationship for the spatially and temporally reduced versus temporally restricted designs. As shown in Figure 9.1, cost was continuously higher and model performance was better as designs included more sites or visits, whereas costs were constant despite improving model performance when removing temporal restrictions by adding more seasons, days, or hours. Compared to the complete all data design, the number of workdays linearly decreased as the number of sites or visits decreased. The model performances, however, showed different rates of decline based on the number of sites or visits per site. Model performance CV MSE R^2s were lower by 3–7 percentage points compared to the complete all data reference design, until the number of sites or visits dropped to 200 sites or 12 repeat visits with all other design components identical to those of the complete all data design (median MSE R^2s of 0.70–0.74 vs. 0.77) (Table 9.3 and Table S9.2). When the monitoring design had fewer than 200 sites or 12 visits, CV MSE R^2s markedly decreased by more than 7–17 percentage points (median MSE R^2s of 0.60–0.70 vs. 0.77). In contrast, the restriction in the number of seasons, days, or hours resulted in decreased CV MSE R^2s, even though the number of workdays was roughly the same as the reduced 12-visit reference design. When we included only two or three seasons in the monitoring campaign, CV MSE R^2s were lower by up to 3 percentage points compared to the reduced reference design (median MSE R^2s of 0.67–0.69 vs. 0.70). The fewer-days design restricted to weekdays gave about a 5 percentage point lower CV MSE R^2 (median MSE R^2s of 0.65 vs. 0.70). The fewer-hours designs showed the most notable decrease in CV MSE R^2 down to a median of 0.55, when the monitoring was restricted to business hours only.

The estimated monetary values for up-front and per-drive-day costs combined showed similar relationships with

Table 9.3. Model Performance and Estimated Per-Drive-Day Costs Across Different Alternative Sampling Designs Based on the Data from the ACT-AP Roadside Mobile Monitoring Campaign, Given a Single Instrument for UFP Monitoring (P-Trak) in a Single Study Area

			Design Components				Cost (workdays for per-drive-day costs)				Model Performance	
Design		Version	*N* of Sites	*N* of Visits per Site[a]	*N* of Routes[b]	*N* of Seasons	Regular Drive	Rescheduled Drive[c]	Quality Control	Total	CV MSE R^2[d]	Reference Design[e]
Complete	**All**[f]		**309**	**29**	**7**	**4**	**203**	**8**	**21**	**232**	**0.77**	CR
Alternative	Fewer sites	250	250	29	6	4	174	8	18	200	0.76	
		200	200	29	5	4	145	8	15	168	0.74	
		150	150	29	3	4	87	8	10	105	0.70	
		100	100	29	2	4	58	8	7	73	0.67	
	Fewer visits	24	309	24	7	4	168	8	18	194	0.72	
		20	309	20	7	4	140	8	15	163	0.72	
		16	309	16	7	4	112	8	12	132	0.71	
		12	309	12	7	4	84	8	9	101	0.70	RR
		6	309	6	7	4	42	8	5	55	0.63	
		4	309	4	7	4	28	8	4	40	0.60	
	Fewer days[g]	Weekday	309	12	7	4	84	8	9	101	0.65	
	Fewer hours[h]	Business	309	12	7	4	84	8	9	101	0.55	
		Rush	309	12	7	4	84	8	9	101	0.64	
	Fewer seasons[g,i]	3	309	12	7	3	84	6	9	99	0.69	
		2	309	12	7	2	84	4	9	97	0.67	

[a] Maximum number of visits feasible given the restriction of reduced designs in a single mobile platform.
[b] A single route as a packet of driving covering an average of 35 sites that provides the number of driving days.
[c] Make-up driving for at least 2 days per season.
[d] Median cross-validated MSE R^2 across campaigns ($N = 30$) for each design using predictions of annual UFP averages at all 309 sites; For the fewer-site designs, predictions were obtained from cross-validation for 100 to 250 sites and external validation for the rest of the sites not included in a given campaign.
[e] Application of the complete reference design (CR — red cell) to alternative designs with spatial reduction with fewer sites or temporal reduction with fewer visits (orange cells), and the reduced reference design with 12 visits (RR — blue cell; 309 sites but 12 visits instead of 29 visits) to alternative designs including fewer days, hours, or seasons (light blue cells).
[f] 309 sites, ~29 visits per site, all days, most hours for 5 a.m. to 11 p.m., and four seasons.
[g] We did not consider the designs of weekend-only and 1 season, as they are unrealistic options.
[h] Business hours: weekdays 9 a.m. to 5 p.m.; rush hours: weekdays 7 to 10 a.m. and 3 to 6 p.m.
[i] Specific combination, such as summer + winter and summer + spring.

model performance as those for workday costs. Although CV MSE R^2s decreased in the temporally restricted alternative designs with fewer seasons, days, or hours (median MSE R^2s of 0.55–0.69 vs. 0.70), the costs were similar to those for the reduced 12-visit reference design (155,000–162,000 vs. 162,000 USD) (Table S9.2). However, the spatially or temporally reduced alternative designs with fewer sites or visits showed much lower costs (124,000–205,000 USD) compared to the complete all data reference design (219,000 USD), as well as a reduction in model performance (median MSE R^2s of 0.60–0.76 vs. 0.77). When we compared these estimated costs from our primary monitoring scenario with a single instrument and fixed staff cost to those from additional monitoring scenarios, the total cost of the complete design slightly increased with a temporally varying shift premium and increased by more than 80% with multiple instruments (232,000–264,000 and 398,000 USD vs. 219,000 USD) (Table S9.3). However, consistent patterns were showing a large cost decrease in spatially and temporally reduced designs as well as a roughly constant cost across temporally restricted designs (Figures S9.1 and S9.2).

Discussion

Our study investigated the tradeoffs between model performance and cost in alternative mobile monitoring designs, and attempted to identify optimal monitoring designs to satisfy the goal of good quality predictions of air pollution along with reduced costs. Our alternative designs represent mobile monitoring designs applicable in practice with reduced spatial or temporal coverage. Designs that restrict the number of seasons, days of the week, or hours of the day adversely impact model performance without reducing cost

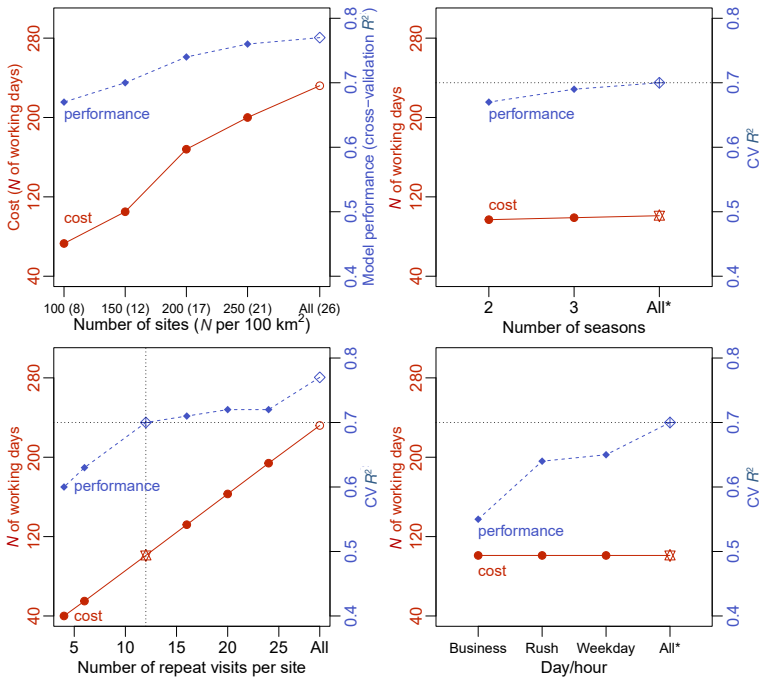


Figure 9.1. Relationships between the number of workdays and median CV MSE R^2s across complete and alternative mobile monitoring designs according to the number of sites, visits, seasons, days, and hours. "All" and "All*" indicate all sites, seasons, days, or hours from the complete all data and reduced reference designs, respectively. Open circles and diamonds represent workdays and CV MSE R^2s, respectively, from the complete reference design that serves as a reference for spatially reduced alternative designs with fewer sites or temporally reduced alternatives with fewer visits. ✡ and ⌖ display workdays and CV MSE R^2s, respectively, from the reduced reference design that is equivalent to the complete all data design, except reduced to 12 visits instead of ~29, and is used as the "complete" design for temporally-restricted alternative designs with fewer seasons, days, or hours. The black dotted horizontal and vertical lines in the plot for the number of repeat visits per site highlight the CV MSE R^2 for the reduced reference design. The designs in the left column are for spatially and temporally reduced alternatives for sites and visits, respectively, as opposed to the designs in the right column that show temporally restricted alternatives based on the number of seasons, or days/hours.

compared to designs requiring the same number of workdays. The optimal design for 12 visits per site includes at least two seasons, both weekdays and weekends, and most hours of the day, including business, early morning, and nighttime hours (5 a.m. to 11 p.m.). Furthermore, while cost increases are driven by the number of sites or visits, the cost-performance tradeoff becomes less favorable above 12 repeat visits due to the relatively slower increase in performance. This finding remained when we extended the monitoring campaign scenarios to multiple instruments and temporally varying staff costs (i.e., shift premium).

Our study extends and improves on previous studies that considered monitoring design guidance by considering cost and leveraging a year-long monitoring campaign with balanced sampling over all days of the week and most hours of the day. The literature includes a few recent studies of mobile monitoring designs that provide insights into the relationship between design components and model performance with the ultimate goal of epidemiological application (Table S9.4). Like our work, these studies sampled from their original campaign to create new campaigns with specific designs, and compared the performance of prediction models developed from these subsets to that of the complete design. Two studies in Oakland, California, USA, leveraged a campaign that collected BC and nitric oxide (NO) during weekday business hours in 2015–2017. Using land-use regression models to predict from annual medians on all 19,149 30-meter road segments, the first study showed that at least 4 repeat visits gave comparable model performance with the complete data of 10–41 visits, depending on the road segment (Messier et al. 2018). The second study applied a mixed-effects model to predict average concentrations using the same data, and reported CV R^2s of 0.8 and 0.9 for 5 and 15 repeat visits per road segment, respectively (Kerckhoffs et al. 2024). In Montreal, Canada, another mobile monitoring study collected UFP data for 34 days over three seasons and showed that reduced campaigns with 12 repeat visits per site gave a median adjusted R^2 close to 0.70, which is lower than 0.74 for the complete data of at least 16 visits (Hatzopoulou et al. 2017; Levy et al. 2014). As with our findings, all these studies suggested the import-

ant role of repeat visits in model performance, although these studies used nonstationary on-road data as opposed to our use of stationary roadside data. In contrast to our work, their insights into the optimal monitoring designs were based on temporally limited monitoring campaigns over less than 40 days or during the daytime on weekdays, which may limit their generalizability (Kim et al. 2023). A recent review also pointed out the limited temporal coverage of mobile monitoring studies, noting that only 15% of studies lasted for at least 1 year. It also reported a median of 10 repeat visits from 41 studies that developed spatial regression models (Wang et al. 2023).

Our focus on costs provides a different optimization of monitoring designs from those that focus exclusively on prediction accuracy. Our previous study (described in the current report) explored the role of different design components and model performance by using the same monitoring data, and showed improved model performance with more sites and repeat visits in a design that covered most hours and days (Blanco et al. 2023a). In this study, we showed that while costs increase linearly with increasing numbers of sites or visits, performance increases much more steeply from 4 to 12 visits per site than from 12 to 26 visits per site in a design that had ~29 total visits per site. We also showed that designs that include multiple different seasons, weekdays and weekends, and most day and night hours can improve prediction accuracy with no cost increase, assuming the per-drive-day cost does not vary by time of day or day of week. Thus, leveraging the temporally and spatially resolved monitoring data and extending experiments using various subsampled datasets, we can conclude that temporally balanced designs with 12 visits per site are the most cost-effective.

For this study, we considered a simple monitoring environment by assuming a single-city, pollutant, and instrument condition. This allowed us to focus on the changes in incremental costs derived from per-drive-day sampling. Application of monitoring to more than one city results in travel costs, including car rentals, flights, additional travel days, housing and per diem, and shipping of instruments, when the primary institution mainly takes charge of monitoring in multiple cities. These additional costs can vary widely depending on the target city. Multicity campaigns also need to plan for additional coordination time to find lab space at a partner institution and possibly subcontract costs to that institution. Additional instruments require more set-up and take-down time each drive day, more data management and quality control, and more up-front time to develop protocols. This additional workload associated with more instruments will likely limit the length of the driving routes and thus the number of locations that can be visited in 1 day. For example, the ACT-AP study involved multiple particulate and gaseous instruments that required an hour to set up and take down at the beginning and end of each driving day, and thus data collection was limited to 6 hours of driving (compared to 7.5 hours in our primary monitoring scenario with a single instrument), resulting in more routes to visit all 309 sites (9 vs. 7), fewer sites on each route (35 vs. 44 on average), and more total driving days (261 vs. 233). In addition, each instrument has unique up-front, consumable, and per-drive-day costs. The P-Trak, chosen in this study, is a 7,000 USD instrument that requires low-cost consumables and does not need special calibration. Another instrument used for UFP monitoring in the ACT-AP study, the NanoScan, costs 30,000 USD up-front, which is approximately the equivalent of 56 days of per-drive-day costs for the simplified single-instrument design. The manufacturer also recommends factory maintenance and calibration once per year, costing about 2,000 USD for service and shipping. Although this additional cost results in much higher total costs compared to those of the single instrument scenario, the cost increase is mostly derived from up-front costs, which are constant across different designs, as shown in Table S9.3. The relationship between per-drive-day costs and design components in the multi-instrument scenario was consistent with our findings based on a single instrument (Figures S9.1 and S9.2).

We employed the simplifying assumption of identical staffing costs for business hours versus nonbusiness hours sampling as well as weekday versus weekend sampling, which resulted in roughly constant costs between all designs requiring the same number of workdays. We based this simplification on our experience using full-time staff who are paid a constant staff cost rate across different days of the week or hours of the day. However, this could vary between institutions and countries. When we applied higher staff costs to evening and weekend sampling, the estimated cost was slightly higher, as shown in Figure S9.1. However, the increases were relatively small and do not affect our overall interpretation that the temporally unrestricted designs are the most cost-effective (Table S9.3, Figures S9.1 and S9.2).

Our study has several limitations, and future studies could develop alternative cost estimation approaches to represent the complex reality of mobile monitoring campaigns. There are different costs and logistical considerations associated with funding graduate students versus staff, and full versus part-time workers. For example, students might need to accommodate class schedules or may not be available for 8-hour shifts. A variable shift schedule could also require extra up-front planning to accommodate the design with the available staff or additional time to hire staff willing and able to work an unusual schedule. This staffing plan is also more vulnerable to staff turnover or disruption if field technicians develop unexpected obligations outside of work, such as family duties, potentially causing delays or data loss. In addition, it is important to note that all our comparisons were constrained by the full dataset available for analysis. Had we collected additional data, for example, at more sites or more visits per site, the changes in performance statistics may have been different for the number of sites and visits per site that we considered. Finally, all our exposure assessments focused on ambient exposure, which could affect exposure misclassification and health inference, although consideration of indoor air quality or personal exposure was outside the scope of our study.

We focused on mobile-monitoring stationary roadside sampling with 2-minute measurements and excluded nonstationary sampling with 1-second measurements, to characterize residential rather than on-road pollution levels. Our previous analysis of minute-level data from two regulatory monitoring sites in Seattle showed that multiple visits of a 1-minute duration minimized the error in estimating the long-term average concentration of NO and NO_2 (Figure S2 in Blanco et al. 2022). This finding supports our focus on 2-minute stationary roadside data when our goal is long-term average prediction at residential locations. In addition, we have shown that on-road data requires additional sampling and analysis costs to characterize residential off-road exposures. Specifically, additional pollutants must be measured to account for on-road sources such as tailpipe pollution (Doubleday et al. 2023). Our multi-instrument scenario indicates a possible and realistic attempt to account for on-road plumes to assess representative residential exposures to UFPs by using other gaseous pollutants. Furthermore, nonstationary measurements could be highly correlated given adjacent sampling locations. These unique features of nonstationary data could result in inconsistent long-term average concentrations compared to those from stationary roadside data, when combined for developing a single exposure prediction model. Future studies should examine both nonstationary on-road data alone as well as on-road data combined with stationary roadside data to provide guidance for exposure model development of combined data.

Some of the performance statistics in our results may be influenced by randomly resampling the same observations in designs with restricted sites, visits, seasons, days, and hours. We summarized the model performance using the median across 30 campaigns to avoid the impact of reporting extreme performance statistics. In addition, our resampling from the temporally balanced data could also reduce the impact. The ACT-AP study scheduled its drive days for operational and technical feasibility as seen in other monitoring campaigns, but rotated the monitoring schedule on a biweekly basis to accomplish its overall temporally balanced data collection. Specifically, ACT-AP consistently monitored all routes during a specific 8-hour time slot in the morning, midday, or evening over a 2-week period before transitioning to a new 8-hour time slot over the next 2-week period. The monitoring over 8-hour time blocks helped reduce technician burden compared to a less consistent schedule and accommodated battery charging, which takes at least 8 hours. Furthermore, ACT-AP's temporally balanced data collection allowed all performance statistics to be evaluated against a consistent set of observations, which facilitated comparisons between designs (Blanco et al. 2023b).

It is possible that mobile monitoring designs for characterizing UFPs that are supplemental to longer-term fixed site monitoring campaigns or emerging monitoring options could be even less expensive. More research is needed to establish best practices for integrating different monitoring data sampled by different instruments as well as monitoring designs. New monitoring technology, such as low-cost sensors, commonly identified as devices costing <2,500 USD, can also reduce data collection costs for $PM_{2.5}$ and gases, but to date, there are no validated low-cost sensors for UFPs (Morawska et al. 2018). Different mobile monitoring platforms based on walking, biking, and public transportation, as performed in some previous studies, may also be able to reduce costs (Farrell et al. 2016; Hankey and Marshall, 2015; Mueller et al. 2016).

Although the present study focused on the impact of design choices on exposure assessment, our findings also provide some insight into the impact on health inference. In related work, we applied UFP predictions obtained from the same mobile monitoring data and similar alternative mobile monitoring designs to those discussed here, to cognitive function in the ACT cohort, and compared health estimates (Chapter 4). Broadly, while the pattern of health estimates was less conclusive than the prediction model performances, we found similar disadvantages with the temporally restricted designs as we found for cost. Whereas there were small differences in health estimates from the fewer visits or seasons designs compared to the reference designs, most health estimates from the business- and rush-hour-only designs were less comparable to the reference health estimates. We observed many biased and some unbiased health estimates for these temporally restricted designs, with biased health estimates in both directions. Temporally restricted monitoring designs that exclude nonbusiness hours have been common choices in previous mobile monitoring campaigns. Our findings indicate that future studies can improve the accuracy in exposure prediction as well as health estimation with little additional cost when they use a design with temporally unrestricted monitoring.

In summary, despite emerging interest in and increasing deployment of mobile monitoring campaigns to assess long-term exposure to air pollution for epidemiological applications, few studies have provided guidance on monitoring designs. The few studies that have considered the question of how much sampling is enough have focused exclusively on the accuracy of air pollution prediction and have not also considered cost, which is a significant determinant of mobile monitoring design given its high expense and flexible design options. By focusing on the trade-off between cost and predictive exposure model performance, we have provided practical guidance to help future mobile monitoring studies optimize their designs to assess the health associations with traffic-related air pollution.

CHAPTER 10: SYNTHESIS, INTERPRETATION, AND IMPLICATIONS OF FINDINGS

Lead authors: Lianne Sheppard and Magali Blanco

SUMMARY OF FINDINGS

This project aimed to determine how to optimally design air pollution exposure assessment and model air pollution exposure data for application to epidemiological cohort studies where the inferential goal is to estimate the association between long-term average air pollution exposures and health outcomes. Overall, we can conclude that air pollution exposure assessment design is critical for exposure prediction performance and also impacts the estimation of health associations. Based on the multiple investigations conducted, many of which focused on UFPs, we found that exposure predictions with better performance statistics result in health association estimates that are generally more consistent with those obtained using the "best" exposure model predictions (the model with all data included), although the pattern of health estimates was often less conclusive than the pattern of prediction model performances. While there were some larger impacts on health association estimates of more poorly performing exposure models relative to the complete all data exposure model, such as the business hours design from a mobile monitoring campaign, many of the differences were small and did not deviate meaningfully from the association estimate obtained from the "best" exposure model. The degree of impact on the epidemiological inference depended on the magnitude of the health association estimate from the "best" exposure model and the width of its confidence interval. Specifically, the impact on health inference was stronger in the confounder model 1 results that used the mobile monitoring data. This model estimated an adverse association of UFPs with cognitive function.

Our findings were clearest and most dramatic for the design of mobile monitoring studies that collect stationary roadside data. In that setting, we can recommend that future studies use stationary roadside sampling during most hours of the day, at least two seasons, and all days of the week. The implications of adding low-cost sensors to regulatory monitoring data were more challenging to discern due to the inherently sparse and unbalanced data available for this investigation. We did not find that leveraging advanced statistical methods (specifically, spatial ensemble-learning methods for prediction) improved exposure prediction model performances. This may have been due to the already sophisticated UK-PLS approach we used by default, and in particular its application in conjunction with the large number of covariates that we considered in the PLS model, such that the contribution of any single covariate was approximately linear. In other words, it is reasonable to believe that in the presence of such a large set of covariates considered, each is approximately linearly associated with the pollutant being modeled, such that the potential added value of the spatial random forest approach is not observed in the model fit. Other settings with a smaller number of possible covariates available may lead to different conclusions and suggest greater added value from the application of a spatial random forest approach.

Our value of information analyses demonstrated that consideration of the cost versus exposure prediction performance tradeoff can identify designs that control exposure monitoring cost with relatively little impact on prediction model performance. Thus, for instance, 12 repeat visits per site versus the complete all data mobile monitoring design with ~29 repeat visits per site reduces cost considerably, with relatively less reduction in exposure model performance. Furthermore, for the same cost, some designs perform much better than others and should be prioritized. Specifically, given 12 visits per site: sampling most hours of the day, at least two seasons, and all days of the week has the same cost as temporally restricted reduced hours designs, such as the common business hours design, with meaningfully better model performance.

We elaborate further in the following subsections after first addressing a fundamental point about the health inference models and approach we used in this report.

ANALYSIS APPROACH AND CHOICE OF HEALTH INFERENCE MODELS

We based our approach on leveraging the extensive air pollution exposure assessment and outcome data available from the ACT-AP study. Thus, we sampled from existing air pollution data to evaluate exposure assessment designs that were subsets of those data. Then, by considering each of these designs, we evaluated subsequent health inferences, which focused on baseline cognitive function using the CASI-IRT outcome. The magnitude and uncertainty of these health association estimates were dependent on the associations evident in the ACT cohort. The insights we were able to develop are conditional on the strengths and weaknesses of these data. It will be important to replicate our approach in the future with a new cohort and data from a new air pollution exposure assessment campaign.

Most of the health associations included in this report focus on confounder model 1, which is less completely adjusted than model 2 (details in Chapter 3). For the UFP exposure, this model yielded a statistically significant association with cognitive function in the hypothesized direction. Confounder model 1 was not adjusted for the spatially varying potential confounders of race and neighborhood disadvantage index. While an inferential model adjusting for these variables (e.g., confounder model 2) is arguably the better model to focus on from an epidemiological inference perspective, our decision to focus initially on model 1 was based on the goals of the project. Our primary goal was to

understand how aspects of exposure assessment affect epidemiological inference. Because the association of cognitive function with UFPs using the mobile monitoring data was estimated to be adverse in confounder model 1, whereas in confounder model 2 it was nearly zero and consistent with a wide range of effects, the model 1 estimate provided a much clearer demonstration of the impact of exposure assessment design on inference. Using confounder model 1, we were able to demonstrate how various mobile monitoring exposure assessment designs impacted inference for an association that was statistically significant under the "best" exposure model. Results from various exposure assessment designs applied to confounder model 2 were fairly similar to the all data exposure model result and thus did not allow the same clear demonstration of this impact. For the spatiotemporal model, the "best" exposure models for $PM_{2.5}$ and NO_2 were consistent with a wide range of effects using confounder model 1, and there was little impact on inference in comparison with the alternative exposure modeling campaigns.

An alternative to using real-world data to evaluate various exposure models would be to conduct simulation studies. Simulation studies are experiments where the investigator assumes underlying distributions for the exposures and outcome, an exposure measurement campaign structure, and an association of the health outcome given known exposures. The investigator simulates realized values of the exposures, outcomes, and measurements from the exposure assessment campaign, and evaluates the impact of the exposure assessment design on the epidemiological estimates obtained from relating predicted exposures to the simulated outcome. Knowledge of the assumed underlying truth is an important strength of simulation studies. However, it is challenging to capture real-world complexity meaningfully in simulation studies, a challenge that can be overcome partially but never completely by developing a large number of different simulation scenarios. Further, many investigator-hypothesized experiment-driven scenarios are likely to be more clear-cut and less challenging than real-world evaluations. While there is an important role for simulation study evaluations, they should not lull researchers into believing that they alone will reveal important real-world insights.

MOBILE MONITORING DESIGN – INSIGHTS FROM THE STATIONARY ROADSIDE DATA

The clearest and most compelling findings from this project came from our evaluation of stationary roadside UFP data obtained from mobile monitoring. For these analyses, we relied on a Seattle mobile monitoring campaign that was designed specifically for epidemiological inference in the ACT cohort (Blanco et al. 2022). As described in detail in Chapter 3, there were 309 stationary roadside locations distributed across a 1,200 km^2 land area that were selected to be representative of ACT cohort residential locations. Short-term stationary roadside stops were visited for 2 minutes, approximately 29 times over a year. The 288 sampling days were temporally balanced over most hours of the day (5 a.m. to 11 p.m.), all days of the week, and all seasons. Thus, for the stationary roadside dataset, we analyzed ~1 hour of data for every ~6 hours of sampling in a day. Our focus on temporally balanced sampling within sites meant that all primary analyses could rely on the design to obtain observed annual averages, rather than adjusting for temporal factors in the analysis. This approach is distinct from most other published mobile monitoring studies; we addressed temporally adjusting mobile monitoring data in some of the analyses reported in Chapter 4 and discuss our insights further below.

As reported in Chapter 4, we found that exposure model performances from poorer or less comprehensive designs were attenuated, including fewer visits per site; fewer seasons, restricted temporal sampling, specifically business or rush hours weekday sampling; and unbalanced sampling with unequal numbers of visits per site. We considered UFPs in depth in that chapter, and also provided a summary of exposure model performance for other pollutants (BC, NO_2, $PM_{2.5}$, and CO_2) from our published work (Blanco et al. 2023a). However, in contrast to the exposure model performances, there were smaller impacts on UFP health association estimates for many of these designs (excluding the business and rush hours designs) relative to the inference obtained from the all data design. Across the 30 campaigns (i.e., subsamples of the data, each representing a hypothetical realized mobile monitoring campaign) for each design, there was no overall health association bias for 12 rather than 29 visits per site, sampling in all four seasons, and balanced sampling. There tended to be additional biases, some of which were small, in the median campaign health association estimates for fewer visits (4 or 6 per site) or fewer seasons, and unbalanced sampling. Accompanying the biases in the medians across the 30 campaigns, we also observed greater variability in health association estimates across campaigns than for the designs that did not show an overall bias. The impact on variability of health association estimates across realizations of specific designs is a feature that future researchers will need to grapple with (Blanco et al. 2023a).

In contrast to the fewer-visit designs, the common temporally restricted designs showed larger biases in the health association estimates for UFPs across all 30 campaigns, with some inconsistency in the direction and magnitude of biases relative to the reference health estimates. Temporal adjustment of the exposure monitoring data, while showing an overall decrease in exposure model performance (Figure 4.4), showed similar (confounder model 1 for business hours), slightly improved (confounder model 2 for business hours, confounder model 1 for rush hours), or more biased (confounder model 2 for rush hours) health estimates relative to predictions from the unadjusted exposure data (Figure 4.5). Interestingly, while the health association conclusions differed by confounder model, the pattern in the box plots across the four exposures considered (business and rush hours, with and without temporal adjustment) was similar across confounder models (Figure 4.5). An explanation for

this contrasting pattern of performances for the exposure model predictions versus the health association estimates is that there is a trade-off between biases driven by classical-like and Berkson-like measurement error, which results in different amounts and possibly direction of biases in various health association estimates, while the temporally adjusted exposure predictions are derived from noisier exposure data and thus show worse prediction model performance. See Chapter 4 for additional discussion of this topic.

Many of the designs for UFPs considered in Chapters 4 and 6 focused on fewer visits per site and not on the reduction in the number of sites. As we have published (Blanco et al. 2023a) and briefly summarized in Chapter 4 (see Figure 4.1), we found that it is the total number of stops that matters most in a stationary roadside campaign for all pollutants considered (UFPs, BC, NO_2, $PM_{2.5}$, and CO_2). The number of stops is the sum of the number of visits per site across all sites. Hence, in this report, we focused on fewer visits rather than changing the number of sites. While we didn't consider health analyses for fewer sites, given the general consistency between the exposure and health results for most designs (even though the pattern of health estimates was less conclusive), we can speculate that the health association estimates for fewer sites would be similar to our findings with fewer visits, for designs with a comparable number of total stops. Future research should verify this speculative conclusion.

Our overall recommendations for increasing the accuracy and precision of exposure and health models from stationary roadside sampling from mobile monitoring campaigns are as follows:

- Reductions in the number of visits should be made randomly rather than systematically to avoid spatially or temporally restricted or unbalanced designs. For example, weekday business hours or an unbalanced number of visits across locations should be avoided.
- Use spatially balanced sampling with an equal number of visits per site. In other words, collecting an unbalanced number of samples across locations, where some locations systematically receive more or fewer visits, can worsen exposure predictions. This translates into a bias of health association estimates on average, with larger variability in these health association estimates anticipated across campaigns.
- Use temporally unrestricted sampling by sampling all days of the week and most hours of the day.
- Temporally adjusting data from temporally restricted sampling campaigns (e.g., weekday business hours) may add statistical noise to exposure assessment models, and the temporal adjustment is not guaranteed to improve health inferences due to a trade-off between biases caused by classical-like and Berkson-like error, as discussed in Chapter 4.
- Mobile monitoring campaigns should sample during two or more seasons. Note that seasonal sampling requirements may be impacted by seasonal variability across geographical locations and over time (meteorological conditions, unique sources), and may differ across geographic locations.
- Plan for at least 12 total visits per site, recognizing that while designs with more than 12 visits per site will have better exposure model performances, they are less cost-effective, as discussed in Chapter 9.

As reported in Chapter 5, when we quantified the impact of Berkson-like and classical-like measurement error in the complete all data stationary roadside mobile monitoring dataset, we estimated a 6% bias and a 13% increase in variability by correcting for these errors. The magnitude of these biases was much smaller than the most poorly performing mobile campaign designs (i.e., the business hours and rush hours designs), while in the same ballpark as the magnitude of the median biases from fewer visits, fewer seasons, or unbalanced sampling, at least for confounder model 1 with a statistically significant health estimate. Thus, we believe that for mobile monitoring studies, exposure assessment design is a more important consideration than exposure measurement error when considering the impact on health association estimates in cohort studies. However, the design comparisons discussed in Chapter 4 did not address inflation of the standard error due to measurement error, which we estimated to be 13%.

One of the reasons that we did not show as much impact from the exposure measurement error corrections in Chapter 5, as from some of the more limited designs considered in Chapter 4, may be that the locations in our mobile campaign are spatially compatible with the target cohort. The spatial compatibility assumption is needed for the measurement error correction (Keller et al. 2017). Work has yet to be published to show how to correct for spatial incompatibility, though reweighting the distribution of monitoring locations to align with cohort locations should address this. Our research suggests that a new term should be introduced, that of *temporal compatibility* in mobile monitoring campaigns. We have found that balanced sampling that includes most hours of the day in the target long-term average is necessary. Interestingly, we did not find that reweighting the measurements (i.e., temporal adjustment) to align with the target distribution of times improved the inference. We believe that the additional statistical noise and possible bias due to misaligned reference location(s) outweigh the potential theoretical benefit of temporal adjustment.

While our careful study of both exposure assessment design and exposure measurement error with the Seattle mobile monitoring campaign's stationary data led to some important new insights, we must recognize that using predicted ambient air pollution at participant residences as a surrogate for long-term average personal exposures does not capture true personal exposure. This approach is common in large population settings, given the logistical challenges

of collecting long-term personal exposures. However, many other sources are known to impact true exposures, including outdoor–indoor infiltration rates, time–activity patterns (e.g., time spent outdoors), and indoor sources of air pollution (e.g., cooking) (Allen et al. 2003; Jung et al. 2011; Klepeis et al. 2001; Vardoulakis et al. 2020). This adds an exposure assessment error that we have not accounted for at all in our analyses, although we were able to address residential mobility by incorporating individual-level residential histories into our exposure estimates. Also, because indoor and outdoor air pollution levels are often correlated, there is less concern as we anticipate that the retiree ACT population spends a large portion of their time at home.

INSIGHTS ABOUT ULTRAFINE PARTICLES FROM ON-ROAD MOBILE MONITORING STUDIES

While our work using stationary roadside data from the Seattle mobile monitoring campaign gave us valuable insights about the impact of monitoring design on exposure and health models, most mobile monitoring campaigns collect on-road mobile rather than stationary roadside measurements. These measurements may be impacted by on-road sources of high air pollution (plumes) that may impact the application of these data for human exposure and health studies. We designed a plume adjustment approach to address these elevated concentrations and make them more comparable to stationary roadside data (Doubleday et al. 2023). Furthermore, we investigated key features of on-road monitoring designs by comparing the resulting exposure assessment models and health inferences to estimates from the reference all data stationary exposure model. A number of our findings can be applied to support future studies aiming to develop UFP exposure assessment models for epidemiology with mobile monitoring, specifically:

- Health association estimates were closer to zero for on-road compared to stationary roadside exposure data predictions (Figure 6.3). This was true whether we considered confounder model 1, which provided a statistically significant health estimate, or confounder model 2, which provided a statistically nonsignificant health estimate that was much closer to zero and in the opposite (positive) direction than hypothesized. It is important that this finding be replicated in future research before we draw strong conclusions.
- As with stationary measurements, on-road measurements should be collected in a temporally balanced way: measurements should be collected at all sites, a similar number of times, across all days of the week, and most hours of the day. Not doing so can worsen both exposure prediction performance and health association estimates in terms of bias and increased variability across campaigns (Figures 6.2 and 6.3).
- Adjusting on-road data for on-road sources not relevant to residential exposures (i.e., plume adjustment) can be done by leveraging multipollutant on-road and stationary measurements. We observed that the plume adjustment improved exposure model performances (Figure 6.2), but only slightly improved the health estimates in confounder model 1 and did not consistently show improvement in confounder model 2, where no association was found between UFPs and cognitive function (Figure 6.3). We conclude that temporally unrestricted monitoring designs are more critical than a plume adjustment.
- It is possible that key monitoring design features identified for stationary roadside monitoring campaigns (Chapter 4) apply to on-road campaigns. For example, collecting measurements across at least two seasons. Further investigation is needed to confirm this.
- Altogether, while conducting a fully on-road campaign without stationary measurements may be more time-efficient for the amount of spatial data collected, not collecting stationary measurements may limit the ability to adjust for on-road plumes. Hence, collecting stationary measurements may increase the accuracy of the estimated exposures, and possibly, in turn, the health associations. We consider our proposed plume adjustment approach to be preliminary, such that additional research is needed to determine whether refinements will be made to make on-road data more comparable to realistic residential exposures in a consistent fashion across multiple applications. Thus, future campaigns should consider developing alternative plume adjustment approaches. For example, future plume adjustment approaches may be able to leverage existing fixed-site regulatory monitoring data. Moreover, we used stationary measurements in this work, which tend to be more stable due to longer sampling periods and may better represent off-road residential exposures, to validate the on-road measurements. This approach may be particularly valuable for pollutants with high spatial variability and those not regularly monitored, such as UFPs.

COMMENTS ON ON-ROAD VERSUS STATIONARY SAMPLING IN MOBILE CAMPAIGNS

A crucial consideration for mobile monitoring campaigns is whether they should add stationary roadside sampling to the otherwise mobile on-road data collection, given the higher staff effort and relatively low number of locations visited for stationary sampling. As carried out in the Seattle mobile monitoring study, stationary roadside locations were selected to be representative of the residential locations of the target cohort to ensure spatial compatibility. Stationary stops lasted for 2 minutes to ensure there was a solid minute of data collected by the NanoScan instrument. Thus, for a 6-hour driving day, the vehicle collected about 5 hours of mobile on-road data and only 1 hour of stationary data, visiting between 28 and 45 sites per day along the nine fixed driving routes. There were 309 stationary roadside locations, and we had 19 times (n = 5,887) the number of on-road locations (based on 100-meter

road segments), after excluding segments on A1 highways, or with fewer than five 1-second measurements or fewer than 23 repeat visits. Based on our experience to date, we strongly recommend that mobile exposure assessment campaigns of UFPs intended for epidemiological applications include stationary data collection for the reasons described below. However, this recommendation should be investigated and affirmed by other investigators.

Table 10.1 gives our assessment of the impact of various design and analysis choices on exposure model performance and health inference; as this evaluation requires considerable judgment and can lead to somewhat different conclusions depending on which results are weighted more heavily, we encourage readers to make their own comparisons. Table 10.1 indicates our judgment of the improvement in each design choice given an a priori "better" design. The boxes are colored to indicate that the "better" design shows an improvement (green), no meaningful improvement (blue), inconsistent or unstable results (pink), or worse results (magenta). The key plots from Chapters 4 and 6 (Figures 4.4, 4.5, 6.2, and 6.3) are compiled in Chapter 10's Additional Materials to facilitate readers making their own comparisons. Our judgments are weighted toward the all-hours design choice for design choices that appear in multiple different combinations. Note that while the primary results in Chapter 4 are based on UFP concentrations from the NanoScan instrument, which measures particles in the 10–420 nm range, the primary results in Chapter 6 are based on the unscreened P-Trak, which measures particles in the 20–1,000 nm range. Chapter 4's Additional Materials presents comparisons of these two instruments for exposure model performances (Figure S4.8) and health estimates using confounder model 1 (Figure S4.11). While the details in the results across these instruments are different, the patterns are similar. Most notably, the complete all data exposure model for the NanoScan has a CV MSE R^2 of 0.65, while it is 0.77 for the P-Trak. The poorer performance of the NanoScan is driven by a few poorly predicted sites, as documented in Blanco and colleagues (2022).

Based on our conclusions summarized in Table 10.1, we note that while on-road campaigns collect considerably more data relative to stationary sampling, more design choices result in consistent and conclusive results (green boxes) with the stationary rather than on-road data. We have observed that exposures predicted from purely on-road data are more variable than those obtained from stationary roadside sampling, but have found it challenging to quantify to our satisfaction how much of this additional variability is real (i.e., capturing important exposure features not evident in the stationary data), rather than merely statistical noise (Doubleday et al. 2023). Furthermore, the design choice impacts exposure model results more clearly than it does subsequent health inferences, although many of the health inferences were improved for several design choices for the stationary roadside data in confounder model 1. Finally, while not obvious from Table 10.1 but clearer from Figures 4.5 and 6.3 (compiled adjacent to each other in Chapter 10's Additional Materials), while there are distinctions in the health estimate results across designs, there was considerable overlap in these estimates and nearly all campaigns yield health estimates that are within the 95% confidence interval for the reference all data health estimate.

In general, when considering each design choice in turn, more visits (focusing on 12 versus 4 per site) are better for exposure accuracy for roadside sampling, and have less variability for on-road sampling, in both exposure and health models. We only considered four versus fewer seasons with the stationary data and found that while it improved exposure models, the evidence was only compelling for the health estimates using confounder model 1. The benefit of spatial balance in sampling (i.e., sampling the same number of visits per location) is more discernible for stationary roadside sampling. We also investigated spatial clustering with the on-road data, but did not find consistent and conclusive patterns in the cases we studied. The most striking design choice, which was evident in the accuracy of exposure and health models for both roadside and on-road sampling, was time selection; that is, avoiding temporal restriction by sampling most hours versus restricting to business or rush hours. Notably, the pattern of box plots across the stationary data confounder models for the time-restricted designs (the fewer hours plots in Figure 4.5) was strikingly similar for health estimates using confounder models 1 and 2, but led to different accuracy conclusions in Table 10.1, with improved most hours designs for confounder model 1 and no consistent improvement for confounder model 2. In other words, for confounder model 1, the restricted hours designs, with and without temporal adjustment, led to attenuated health estimates such that the most hours design is clearly better. However, for confounder model 2, where the reference health estimate was slightly positive and near zero, the pattern resulted in biases in both directions. Despite efforts to correct (i.e., temporally adjust) for this temporally restrictive sampling, the most hours design remained better in most but not all comparisons. Finally, plume adjustment of the on-road data, leveraging the presence of multipollutant stationary roadside data, resulted in better exposure predictions, although this did not carry through to result in clearly improved health association estimates.

More research is needed to more deeply investigate and further digest the value of on-road versus short-term stationary data from mobile campaigns. It is important to recognize that on-road sampling occurs in conjunction with stationary roadside sampling, such that the on-road data are collected for "free" along with short-term stationary data. Furthermore, our study used fixed driving routes with temporally and spatially balanced sampling, and we prioritized our stationary sampling when we drove makeup routes to address the inevitable missed stops and equipment failures that occur in any mobile monitoring campaign. The design strategies that we prioritized have not been consistently employed in previously published mobile campaigns, so it may be challenging for other investigators to conduct similar design-related evaluations with existing mobile monitoring

datasets. Finally, we have assumed that the health estimates obtained from stationary roadside sampling are closest to the truth. Future work should investigate alternatives, including thorough application of simulation studies or leveraging of longer-term stationary data to produce more stable annual average estimates.

FINDINGS USING LOW-COST SENSORS TO SUPPLEMENT REGULATORY MONITORING DATA

Insights about the design of monitoring campaigns using low-cost sensors to supplement existing regulatory monitoring data for $PM_{2.5}$ and NO_2, while broadly consistent with the findings from the designs of mobile monitoring studies, were much less clear and compelling. As described in Chapter 3, the data we relied on included 2-week averages from "agency" locations, which included both regulatory monitoring locations and other high-quality research data, as well as low-cost sensor data from over 100 locations collected under the auspices of the ACT-AP study. The nature of the data leveraged in the spatiotemporal modeling used in this analysis is an important factor in these results. Generally, there are relatively few long-term regulatory monitoring locations that can be leveraged in any single location, such as the Puget Sound. The regulatory monitoring sites are also not usually selected to align with a specific epidemiological cohort, as was the case in the Seattle mobile monitoring campaign. The ACT-AP supplementary low-cost sensor data, while obtained from many additional spatial locations, was relatively short-term, and the low-cost sensor sampling was temporally unbalanced due to the need to rotate a limited number of monitoring devices (Figures S7.1, S7.2, S7.3). This is a common feature of similar supplementary sampling campaigns. Overall, the sparsity of the supplementary data available from the ACT-AP study limited the types of evaluations that we could conduct.

To predict exposure, we used our published spatiotemporal model, which is designed to accommodate unbalanced data. As noted in previous and recent publications (Lindström et al. 2014; Zuidema et al. 2024), quantifying exposure model performance statistics from this spatiotemporal model is a challenge. While our scientific interest is in long-term average exposures, due to the unbalanced sampling, it was impossible to report purely spatial model performance statistics, particularly at cohort residential locations, which are of primary scientific interest. To provide consistency across all assessments, we focused on reporting model performance of long-term "spatial" averages at residential locations, which were only represented by low-cost sensor data. Given the short-term and unbalanced sampling of low-cost sensor data, these spatial averages inherently also include some temporal information.

A strength of the low-cost sensor data in the ACT-AP study is that the majority of the measurement locations were at actual ACT participant residences and thus spatially compatible (Szpiro and Paciorek 2013a). However, because these data were sampled using low-cost sensors, they are less precise than regulatory monitoring data and require calibration before use in our analysis. Thus, there is some inherent additional uncertainty in the performance statistics focused on data at residential locations considered in the spatiotemporal models.

There were some important differences between the approaches we took to evaluate the added value of the low-cost sensor data in the $PM_{2.5}$ and NO_2 datasets. Some of these were driven by differences in the data available for analysis. In particular, the $PM_{2.5}$ dataset had multiple different kinds of regulatory monitors available, though the locations measured varied over the enormously long sampling period of 1978–2021. Thus, we focused the lion's share of our model performance analyses on the 2010–2020 time period when the regulatory monitoring locations were more stable. Findings for the longer time period yielded generally similar performance statistics, even though the modeling required some modifications to the approach for estimating the long-term time trend. There were small differences in the health estimates obtained from exposure models from the 2010–2020 versus 1978–2021 time periods (ST panel in Figure 7.1). For NO_2, we had only three long-term sites that were suitable for estimating a long-term time trend, which was insufficient for a typical spatiotemporal model application such as described by Keller and colleagues (2015). Thus, we modified the approach to rely more heavily on the snapshot campaign of 2-week Ogawa samplers measured in three seasons; this provided much of the spatiotemporal model foundation that is usually provided by long-term monitoring sites. However, while the Ogawa samplers were of higher quality than the NO_2 low-cost sensor data, they weren't placed at participant homes. We had also considered placing Ogawa samplers at participant homes, but we determined that it was unaffordable to absorb the extra cost of visiting each home and extra time on a fixed schedule to pick up the Ogawa sampler.

Despite these limitations, the low-cost sensor data added to the overall evidence base developed in this project. As discussed in Chapter 7, we found that $PM_{2.5}$ and NO_2 exposure model performances evaluated at residential locations were best when all supplementary low-cost sensor data were included. In our more in-depth study of low-cost sensor design for the $PM_{2.5}$ dataset, we found that while reducing the total number of repeat low-cost sensor samples within sites resulted in decreased model performance, if we reduced the number of low-cost sensor measurements while maximizing their temporal spread, this resulted in only a small reduction in model performance. In contrast, the prediction model performance was worse if we reduced the temporal sampling by the same number of repeat low-cost sensor measurements, but with less temporal separation between low-cost sensor measurements. We also found that we could reduce the number of low-cost sensor sites by 10% with little impact on model performance. However, this was based on a random selection of locations. The impact on predictions from preferentially excluding low-cost sensors varied by design. In particular, excluding those closest to major roads had the largest adverse

Table 10.1. Summary of Roadside and On-Road Mobile Monitoring Campaign Design Impacts on Exposure and Health Models for UFPs[a]

	Roadside Mobile Monitoring		On-Road Mobile Monitoring	
Design Choice	**Exposure Models**	**Health Models** (Using Confounder Model 1, 2)	**Exposure Models**	**Health Models** (Using Confounder Model 1, 2)
More Visits (12 vs. 4)				
Improved accuracy				
Reduced variability across campaigns				
More Seasons (All 4 vs. 1–3)				
Improved accuracy				
Reduced variability across campaigns				
Spatial Balance (vs. Unbalanced)				
Improved accuracy				
Reduced variability across campaigns				
Sampling Most Hours[b] (vs. Business/Rush Hours)				
Improved accuracy[c]				
Reduced variability across campaigns				
Temporal Adjustment of Business/Rush Hours[d] (vs. Most Hours)				
Improved accuracy				
Reduced variability across campaigns				
Plume Adjustment (vs. Same Design without Plume Adjustment)				
Improved accuracy				
Reduced variability across campaigns				

[a] *Improved accuracy* indicates that campaigns typically result in higher exposure model R^2 values and more accurate health inferences when compared to the reference all data stationary model results. *Reduced variability* across campaigns (i.e., narrower box plots) indicates more consistent results. Colors indicate improvement (green), no meaningful improvement (aqua — the default color when the conclusion doesn't clearly show an improvement across all evaluations), inconsistent/unstable (pink) or worse (magenta) results, with the extended efforts indicated by the design choice. Light gray boxes indicate designs that were not evaluated. Multiple boxes in health models indicate variable results across confounder models 1 and 2, respectively. For design choices that are evaluated in combinations, judgments are weighted toward the all days and most hours designs rather than averaged across all designs.

[b] Most hours refers to the Fewer Visit design with 12 visits, which does not have temporal restrictions.

[c] In confounder model 2, the two boxes represent the differing results for business and rush hours designs.

[d] Business hours only for the on-road data.

impact on performance statistic results. This is likely due to PM monitoring siting criteria that avoid near-road locations.

Regarding estimating health associations with cognitive function for $PM_{2.5}$ and NO_2, we found that the pollutant-specific health estimates were overlapping across models and that there were few meaningful differences between them. Except for the spatial "with low-cost sensors" full model for NO_2, which showed a decrease in cognitive function associated with an increase in NO_2, these health association estimates were consistent with no adverse effect of $PM_{2.5}$ or NO_2 on cognitive function. This is in contrast to some of the associations of UFPs with cognitive function reported in Chapters 4, 5, and 6. Thus, relative to the analysis of the mobile monitoring campaign data, it was challenging to develop deep insights into the utility of collecting low-cost sensor data to supplement regulatory monitoring from these findings.

INSIGHTS FROM APPLICATION OF ADVANCED STATISTICAL METHODS

The goal of the second aim of this project was to improve exposure prediction modeling and health association estimates by applying advanced statistical methods. We considered three subaims:

1. Application of spatial ensemble-learning methods
2. Leveraging driving distance to improve spatial prediction
3. Multipollutant dimension reduction for the prediction of spatial data

In applying spatial ensemble-learning methods, we considered alternatives to the universal kriging with dimension reduction using partial least squares (UK-PLS), which has been our standard spatial exposure prediction modeling approach and is summarized in Chapter 3. The primary alternative we compared was a recently developed spatial random forest approach (SpatRF-PL), which uses a tree-building algorithm adjusted for spatial correlation using thin plate regression splines (TPRS). For more in-depth comparisons, we also fit the Random Forest (RF) algorithm alone, TPRS alone, and each of these following the other (RF-TPRS and TPRS-RF). Interestingly, the cross-validated MSE R^2 statistics for UK-PLS versus SpatRF-PL were all within 0.03, with the direction depending on the specific pollutant. Predictably, the performances of RF and TPRS were worse, while RF-TPRS and TPRS-RF were mostly similar to UK-PLS and SpatRF-PL (Table 8.1). The similarity of performances may be attributed to the general linearity of predictors in this setting. Specifically, even when the true model is nonlinear, given the high dimensionality of the geographic covariates with p = ~200 predictors for n = 309 observations, and given the setting where these covariates are fixed (i.e., inference is conditional on them), any nonlinear model can be represented by a linear model in $p > n$ covariates (Bühlmann and Geer 2015).

Nonetheless, we did find some differences in predictions and prediction errors between UK-PLS and SpatRF-PL, as shown in the maps depicted in Figures 8.1 and 8.2. Thus, we developed a variable importance metric that can be used to highlight features that differ between the prediction models. This variable importance metric is flexible, intuitive, and generally applicable to additive models that account for spatial correlation. In applying this metric to the Seattle mobile monitoring campaign predictions, we found that the SpatRF-PL predictions tended to have larger predictions of UFPs for shorter truck route lengths, higher distance to the airport, lower population density, higher proportion of evergreen forest land, and higher normalized difference vegetation index. These were more evident when considering all the grid locations (Figure S8.1), rather than focusing on the ACT cohort locations (Figure S8.2). This is likely due to the spatial compatibility of the cohort locations with the mobile monitoring observations relative to the grid locations.

The second subaim of this project aimed to leverage driving distance information from the road network to improve predictions. However, computational challenges surfaced because driving distance is not a well-defined mathematical metric. Furthermore, even in the absence of these challenges, the performance did not improve. Thus, we abandoned further work on this subaim.

Finally, we developed a new approach to dimension reduction of multipollutant spatial data, called representative and predictive principal component analysis (RapPCA), that combines features of both classical (or representative) and predictive principal component analysis. In constructing dimension-reduced principal component scores, RapPCA seeks the optimal balance between prediction and representation in an explicit and interpretable way. Particularly in a setting where multipollutant data are collected simultaneously, as in the Seattle mobile monitoring campaign, this new dimension-reduction tool has the potential to advance our understanding of multipollutant exposures on health outcomes. In the future, we will evaluate predictions of these exposures on inference about health associations in the ACT cohort.

INSIGHTS ABOUT THE VALUE OF INFORMATION: THE TRADE-OFF BETWEEN COST AND EXPOSURE MODEL PERFORMANCE

Most studies that consider exposure assessment design for application to epidemiological cohorts focus on some aspect of exposure quantification, such as the performance of exposure prediction models. The innovation of this aim was to further characterize various exposure assessment designs by also considering cost. The primary analyses presented in Chapter 9 focused on stationary UFP data from the Seattle mobile monitoring campaign with a single instrument. Regarding the design versus cost trade-off, we found, based on our data, that the optimal design had 12 repeat visits per site in a balanced design that sampled all days of the week, all

seasons, and most hours of the day. When restricting attention to 12 visits per site and thus a constant overall cost in our primary scenario, there was a minimal drop in performance for the two or three-season alternatives to four-season sampling. There were much more dramatic reductions in performance when restricting sampling to weekdays only, especially with the common business and rush hours designs. Notably, the business hours design had the worst performance of all 12-visit designs we considered.

Our primary approach to cost estimation assumed all working days were interchangeable. Thus, we assumed that staffing a campaign to sample 12 visits per site costs the same whether the sampling was spread over all four seasons or restricted to only two. Similarly, we assumed the same sampling cost for weekdays, weekends, and nonbusiness hours. Depending on various factors for a specific campaign, this assumption may not be realistic in practice, and thus, we relaxed this assumption to allow for a temporally varying shift premium in some additional evaluations reported in Chapter 9. Overall, the exchangeable cost assumption did not have a large impact on the cost estimates. In contrast, the multiple instrument scenario, such as in ACT-AP, was much more expensive (Figure S9.1).

We also considered the cost versus model performance trade-off for the low-cost sensor sampling that supplemented the regulatory monitoring data available for the criteria air pollutants. We focused on NO_2 as well as $PM_{2.5}$ for this analysis. The NO_2 data were distinct from the $PM_{2.5}$ data because we also had the Ogawa sampler snapshots to consider. The Ogawa samplers record more reliable measurements of NO_2 than standard low-cost sensors, and were collected using a completely different design: instead of the rotating low-cost sensors at participant homes, the Ogawa samplers were mounted on telephone poles using a spatial snapshot design repeated in three seasons. Overall, we found that home monitoring adds value to exposure prediction models developed for epidemiological applications for both pollutants. However, this approach to monitoring is very resource-intensive and thus costly. In contrast, while the snapshot campaign involved considerably lower staff effort and is much less logistically challenging, we found that the exposure model performance was poor when evaluated at the home locations for the models that only supplemented the regulatory monitoring data with snapshot data. This finding needs further investigation to determine whether there are additional modifications that would improve the model performance at homes. Given that low-cost sensor NO_2 measurements are noisier than those for $PM_{2.5}$ (Zuidema et al. 2021; Zusman et al. 2020), performance statistics based on these low-cost sensor home observations may also be a factor. Furthermore, our results suggest that using only three snapshot campaigns may be insufficient to adequately model the temporal trends in space.

In summary, we considered the value of information for two distinct types of campaigns: mobile monitoring with stationary data, and low-cost sensors to supplement existing regulatory monitoring data. We developed insights into the cost versus exposure model performance trade-offs for each. A natural question is whether features of both campaigns can be combined into one unified exposure assessment. We did not consider this, as the data we used was distinct. There was only one pollutant that provided useful data in both campaigns — NO_2. Future investigations could address the relative costs of the two approaches to sampling for this pollutant, as well as other scientific considerations, such as the feasible duration of exposure sampling, the spatial extent of the monitoring data, the target cohort, and the primary scientific questions of interest.

LIMITATIONS AND FUTURE RESEARCH PLANS

Several features of this study may affect our conclusions. First, we relied on existing data that were collected from a single cohort in one geographic area over a fixed time period. Results may vary by geographic location, for example, due to differences in air pollution sources and meteorological conditions. While we did find consistency in results across different aspects of our data and with previous works of ours and other investigators, it is possible that some of the conclusions we reached may not generalize to other regions or time periods. Furthermore, the use of real data from the long-standing ACT cohort is a strength of this study. These data naturally incorporate aspects that might not be included in a simulation study, which strengthens the real-world implications of our findings. However, because we relied on this real-word dataset, we don't know the underlying "true" health effect, which is naturally included in a simulation setting based on a hypothetical scenario. Using the ACT data, we are only able to use the health association estimate from our best exposure model as a reference for alternative exposure models that were typically based on using less or weaker exposure data. Because every cohort study is different for a myriad of reasons that we could not evaluate in this project, replication in other locations and cohorts is warranted.

It is appealing to consider what more we could have learned from conducting health analyses based on simulated outcomes. Simulation studies may provide some useful insights into the patterns of health inferences, given that we did not observe completely consistent patterns of results across confounder models 1 and 2 (the first of which was statistically significant for UFPs while the other was not and provided a health estimate close to zero), particularly when we also compared and contrasted insights from the Seattle stationary versus on-road mobile monitoring data. We note that simulation studies are experiments in which the investigator hypothesizes all the relationships of interest and then evaluates the impact on inference when some conditions are changed (e.g., how the exposure data are sampled, as was the focus in this project). The challenge with simulation studies is that the real world is complex, and any hypothetical simulated scenario may not adequately represent a realistic, com-

plex real-world scenario. Furthermore, no single simulation study will be sufficient to adequately capture all the features that should be considered, necessitating that a large number of different scenarios be simulated. With these features in mind, future work to further solidify the insights developed in this project could be conducted to confirm or develop a deeper understanding of the implications of exposure assessment design on health inference. For instance, using the ACT cohort to anchor the simulation study design, one could assume that the regression parameter estimate from a health model conditional on the all data exposure model is the true exposure effect, and then simulate outcomes for pseudo-ACT participants with similar covariate distributions and unexplained variability in the outcome. Then, one could conduct the same health analyses conditional on different (alternative) exposure assessment designs and across multiple exposure assessment campaigns as we reported in this project. This would allow one to determine whether the same patterns in the estimated health parameters hold under a known exposure-health outcome relationship. One could extend this exercise by hypothesizing stronger and weaker true health parameters (or residual outcome variability) to ascertain whether the observed patterns change with the strength of the relationship. This might help to deepen understanding of the results we obtained across the stationary roadside and on-road mobile monitoring data for confounder models 1 and 2.

Our inferential analyses were limited, not only by exclusive consideration of the cognitive function outcome at baseline in a cross-sectional analysis, but also by the health analysis approach we used. The primary focus of this project was on how changes in exposure quantification — whether from changes in exposure assessment design or exposure modeling approach — affect health inference, rather than a focus on estimating associations or developing causal conclusions about the effect of air pollution on cognitive function. Thus, we did not focus our attention on multiple important observational study analysis considerations that would normally be incorporated into reports of evidence from inferential analyses, such as confounder model selection and the conduct of multiple sensitivity analyses of the health results. Nor did we frame our approach to use modern statistical methods that are more appropriate for reaching causal conclusions from observational study data. Furthermore, we restricted our attention to a single cross-sectional continuous outcome at baseline in our inferential analyses. A linear health model for a continuous health outcome is the most straightforward inferential analysis that can be considered. We made this choice to simplify the technical aspects of the inferential analyses and to allow us to focus on the exposure assessment considerations. Finally, as discussed in Chapter 3, due to the temporal misalignment between ACT cohort baseline outcomes (1994+) and air pollution measurements (2019–2020 for mobile monitoring; less dramatically temporally misaligned for the exposures developed from low-cost sensor data, particularly for $PM_{2.5}$), we assumed in many of our analyses that air pollution surfaces remained constant over time. This is another source of exposure assessment error, particularly for early ACT enrollees, which may in turn affect the resulting health estimates. However, because we applied this "constant over time" exposure surface assumption identically to all analyses based on the mobile monitoring data, its impact did not vary across exposure assessment designs. Thus, we do not believe that this limitation affects the conclusions we reached in this project about exposure assessment design and analysis. Regardless, future studies can advance our collective understanding by considering other health outcomes, confounder models, and other cohorts, as well as greater attention to sensitivity analyses and novel methods for confounding control for the health inferences. Some of the analyses leveraging the ACT data that are currently underway will allow greater insight into the associations of multiple air pollutant exposures with health outcomes. Future work will allow these to be extended to understanding the role of exposure assessment design and modeling for these health inferences.

This study focused on long-term average exposures and their health associations. The ACT cohort study, as well as the air pollution exposure assessment campaigns and models, were not designed to address the health impacts of other biologically relevant exposures, such as peak exposures. The considerations for modeling peak exposures and associating them with shorter-term health outcomes are inherently different.

GENERALIZABILITY OF FINDINGS TO OTHER SETTINGS

An important consideration is how generalizable our findings are to other settings. We have relied on one set of exposure monitoring campaigns in one geographic location linked to one long-standing cohort that is focused on brain health. Thus, it is important that the insights that we have developed be replicated in other settings, with other cohorts, and for other health outcomes. However, there are a few features of our approach that merit highlighting because they strengthen the potential generalizability of our conclusions. Most notably, we started from a solid foundation with the Seattle mobile monitoring campaign, given our temporally and spatially balanced design, selection of spatially compatible locations with the ACT cohort residences, fixed driving routes, and a large number of visits per site. We used 30 campaign realizations for each of the designs we considered in Chapters 4 and 6. This considerably strengthens the results as one can observe the collective pattern of findings from a specific design and understand how variable these results are. However, the datasets we studied are large and complex, such that there are many data analytic steps needed to produce the results we reported, and multiple opportunities for small differences in the analysis pipeline to possibly impact findings. Thus, independent replication is necessary. Another feature to highlight is that we used the reference observations from the complete all data design in our performance evaluations

rather than the observations for the specific realized design. This is not realistic in practice because real-world campaigns do not often have more complete observations. However, given our goal of comparing various designs, the use of the complete all data observations facilitates comparisons by always using the same reference observations, which we assume are the most complete and the least biased, and thus allows for deeper insights into design features. As we have published in other work where we used a similar design sampling idea based on existing regulatory monitoring data such that the true annual average concentration at the monitoring locations was known (Blanco et al. 2023b), using observations from the same campaign can lead to misleading performance statistics, particularly when the design being considered is biased for the target of modeling a long-term average concentration.

We conclude this section by listing the most generalizable findings from our study. We welcome a robust dialogue with other investigators to support or dispute our conclusions. The context for our conclusions is to strive for high-quality inference about long-term average air pollution exposures on relevant health outcomes measured in cohort studies. Regarding exposure assessment for air pollution epidemiology, we believe the following:

- Exposure assessment study design is critically important for high-quality epidemiological inference and is a generally underappreciated feature in the literature.
- The fundamental importance of using balanced designs without temporally restricted sampling in mobile monitoring campaigns should hold up in future investigations (Chapters 4 and 6), and this approach is no more costly than temporally restrictive sampling campaigns, which are often used because they are logistically more convenient (Chapter 9).
- Mobile monitoring study design is more important than exposure measurement error correction, at least for mobile campaigns that have been designed to be spatially compatible with the target cohort, because spatial compatibility is a key assumption in the measurement error approach we employed (Chapters 4 and 5).
- It is worth including short-term stationary sampling in mobile campaigns even though it adds time and thus cost, given its potential to support data quality by serving as a more reliable reference dataset, to allow for data adjustments to better approximate off-road residential exposures (e.g., plume adjustment), and to provide more directly representative residential exposures (Chapters 6 and 10).
- Supplementary exposure data collection (e.g., using low-cost sensors at participant homes) improves exposure model predictions but requires a large staff effort.
- When hundreds of geographic covariates are available, machine learning methods such as spatial random forest do not improve prediction relative to UK-PLS (Chapter 8).

CONCLUDING COMMENTS

Epidemiological studies often make use of exposure data that is collected in opportunistic and logistically convenient ways. This is particularly true in air pollution epidemiology, where routinely collected exposure data can often be leveraged, even though these data may not be fit for purpose. Furthermore, the collection of new exposure data is typically constrained by cost and logistics. As discussed in depth in this synthesis chapter, this project has shown that there should be greater attention to the design of the exposure assessment data collection for use in epidemiological inference. We developed strong recommendations for mobile monitoring campaign design, thanks to the comprehensive and well-designed Seattle mobile monitoring campaign. The supplementary low-cost sensor data we also leveraged from the Puget Sound-based ACT-AP study were much less comprehensive, and thus it was more challenging to develop similarly deep insights. Overall, while supplementary monitoring data improves exposure predictions even when from a low-cost platform (albeit with possibly minimal impact on health association estimates), it is expensive to deploy monitors at participant residences. Regarding mobile monitoring campaigns, we were able to show that extensive stationary roadside campaigns with good designs, such as the Seattle mobile monitoring campaign, can be modified to reduce cost. Specifically, we recommend balanced data collection with at least 12 visits per site covering all days of the week, most hours of the day, and at least two seasons. Finally, broadly speaking, better exposure assessment design leads to better exposure prediction model performance, which in turn may improve the health association estimates, although we found the impact on health inference to be less compelling than the impact on exposure model performance. In combination with cost considerations, it is possible to design air pollution exposure assessment studies for the best relative cost that achieve good exposure prediction models and potentially also benefit health association estimates.

DATA AVAILABILITY STATEMENT

AIR POLLUTION DATA AND EXPOSURE MODELS

The mobile monitoring campaign used to collect the data used throughout this report is described here:

> Blanco MN, Gassett A, Gould T, Doubleday A, Slager DL, Austin E, et al. 2022. Characterization of annual average traffic-related air pollution concentrations in the greater Seattle area from a year-long mobile monitoring campaign. Environ Sci Technol 56:11460–11472, *https://doi.org/10.1021/acs.est.2c01077*.

The stationary roadside mobile monitoring data and related documentation are publicly available through Zenodo: *https://zenodo.org/records/13761282*.

The on-road mobile monitoring data and related documentation will be made available through Zenodo upon publication or by request.

The spatiotemporal models presented in Chapter 7 are described in the following two papers:

Bi J, Burnham D, Zuidema C, Schumacher C, Gassett AJ, Szpiro AA, et al. 2024. Evaluating low-cost monitoring designs for $PM_{2.5}$ exposure assessment with a spatiotemporal modeling approach. Environ Pollut 343:123227, *https://doi.org/10.1016/j.envpol.2023.123227*.

Zuidema C, Bi J, Burnham D, Carmona N, Gassett AJ, Slager DL, et al. 2024. Leveraging low-cost sensors to predict nitrogen dioxide for epidemiologic exposure assessment. J Expo Sci Environ Epidemiol, *https://doi.org/10.1038/s41370-024-00667-w*.

Raw data were compiled as documented in the Data Organization and Operating Procedures from the MESA Air study: *https://deohs.washington.edu/sites/default/files/MESAAirDOOP_Rev12.pdf*. Raw data at public locations, predictions at public locations, and analytic code are available upon request to the MESA-Air Data Team. Access to information linked to participant residential locations requires additional IRB approval.

ANALYTIC CODE

The analytic code used to conduct the analyses in Chapters 4 to 6 and the health analysis sections in Chapters 7 and 8 is publicly accessible through GitHub: *https://github.com/magali17/hei_aim3a*.

The analytic code for the RapPCA methods developed in Chapter 8 is publicly accessible through GitHub: *https://github.com/chengs94/RapPCA*.

All other analytic code may be requested from the authors.

FINAL ANALYTIC DATASETS

The final analytic datasets, which include air pollution exposures linked to participant health outcomes, are not publicly available to maintain participant confidentiality. Access to the Adult Changes in Thought (ACT) cohort data requires approval through established data-sharing procedures. For more information on obtaining access to ACT cohort data, please visit *https://actagingresearch.org*.

ACKNOWLEDGMENTS

Research described in this report was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (US EPA) (Assistance Award No. CR-83998101) and certain motor vehicle and engine manufacturers. The contents of this report do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. In addition to the HEI funding, this research was partially supported by grant R01ES026187 jointly funded by the National Institute on Aging (NIA) and National Institute of Environmental Health Sciences (NIEHS) as well as grants T32ES007032 and T32ES015459 from NIEHS, the National Research Foundation of Korea (2022R1A2C2009971), and the National Cancer Center of Korea (NCC-2310220, NCC-24H1720). The authors would like to thank the ACT participants and community volunteers for the data they have provided and the many ACT investigators and staff who steward the ACT data (NIA U19AG066567). You can learn more about ACT at *https://actagingstudy.org*. We express additional gratitude for our fruitful cooperation with the Puget Sound Clean Air Agency and Washington State Department of Ecology; for the individuals who have written, updated and maintained the spatiotemporal package: Johan Lindström, Paul Sampson, Silas Bergen, Assaf Oron, Michael Young, and Victoria Knutson; for the individuals who collected and managed data for the Seattle mobile monitoring campaign: Brian High, Dave Slager, Timothy Gould, David Hardie, Jim Sullivan; and for the additional co-authors on the papers that form the backbone of this report. Many of these co-authors are included in the About the Authors section.

REFERENCES

Abdi H. 2010. Partial least squares regression and projection on latent structure regression PLS Regression. Wiley Interdiscip Rev Comput Stat 21:97–106, *https://doi.org/10.1002/wics.51*.

Alexeeff SE, Roy A, Shan J, Liu X, Messier K, Apte JS, et al. 2018. High-resolution mapping of traffic-related air pollution with Google Street View cars and incidence of cardiovascular events within neighborhoods in Oakland, CA. Environ Health 1738:1–13, *https://doi.org/10.1186/s12940-018-0382-1*.

Allen R, Larson T, Sheppard L, Wallace L, Liu LJS. 2003. Use of real-time light scattering data to estimate the contribution of infiltrated and indoor-generated particles to indoor air. Environ Sci Technol 37:3484–3492, *https://doi.org/10.1021/es021007e*.

Apte JS, Messier KP, Gani S, Brauer, M, Kirchstetter TW, Lunden MM, et al. 2017. High-resolution air pollution mapping with Google Street View cars: exploiting big data. Environ Sci Technol 5112:6999–7008, *https://doi.org/10.1021/acs.est.7b00891*.

Austin E, Xiang J, Gould TR, Shirai JH, Yun S, Yost MG, et al. 2021. Distinct ultrafine particle profiles associated with aircraft and roadway traffic. Environ Sci Technol 555:2847–2858, *https://doi.org/10.1021/acs.est.0c05933*.

Bergen S, Sheppard L, Kaufman JD, Szpiro AA. 2016. Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines. J Royal Stat Soc Series C Appl Stat 655:731–753, *https://doi.org/10.1111/rssc.12144*.

Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, et al. 2013. A national prediction model for $PM_{2.5}$ component exposures and measurement error–corrected health effect inference. Environ Health Perspect 121:1017–1025, *https://doi.org/10.1289/ehp.1206010*.

Bergen S, Szpiro AA. 2015. Mitigating the impact of measurement error when using penalized regression to model exposure in two-stage air pollution epidemiology studies. Environ Ecol Stat 223:601–631, *https://doi.org/10.1007/s10651-015-0314-y*.

Bi J, Burnham D, Zuidema C, Schumacher C, Gassett AJ, Szpiro AA, et al. 2024. Evaluating low-cost monitoring designs for $PM_{2.5}$ exposure assessment with a spatio-temporal modeling approach. Environ Pollut 343:123227, *https://doi.org/10.1016/j.envpol.2023.123227*.

Bi J, Carmona N, Blanco MN, Gassett AJ, Seto E, Szpiro AA, et al. 2022a. Publicly available low-cost sensor measurements for $PM_{2.5}$ exposure modeling: guidance for monitor deployment and data selection. Environ Int 158:106897, *https://doi.org/10.1016/j.envint.2021.106897*.

Bi J, Wildani A, Chang HH, Liu, Y. 2020. Incorporating low-cost sensor measurements into high-resolution $PM_{2.5}$ modeling at a large spatial scale. Environ Sci Technol 544:2152–2162, *https://doi.org/10.1021/acs.est.9b06046*.

Bi J, Zuidema C, Clausen D, Kirwa K, Young MT, Gassett AJ, et al. 2022b. Within-city variation in ambient carbon monoxide concentrations: leveraging low-cost monitors in a spatio-temporal modeling framework. Environ Health Perspect 1309:097008, *https://doi.org/10.1289/EHP10889*.

Blanco MN. 2021. Traffic-related air pollution and dementia incidence in a Seattle-based, prospective cohort study. Dissertation. Seattle, WA: University of Washington.

Blanco MN, Bi J, Austin E, Larson TV, Marshall JD, Sheppard, L. 2023a. Impact of mobile monitoring network design on air pollution exposure assessment models. Environ Sci Technol 571:440–450, *https://doi.org/10.1021/acs.est.2c05338*.

Blanco MN, Doubleday A, Austin E, Marshall JD, Seto E, Larson TV, Sheppard L. 2023b. Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. Journal of Exposure Science Environ Epidemiol 333:465–473, *https://doi.org/10.1038/s41370-022-00470-5*.

Blanco MN, Gassett A, Gould T, Doubleday A, Slager DL, Austin E, et al. 2022. Characterization of annual average traffic-related air pollution concentrations in the greater Seattle area from a year-long mobile monitoring campaign. Environ Sci Technol 5616:11460–11472, *https://doi.org/10.1021/acs.est.2c01077*.

Blanco MN, Shaffer RM, Li G, Adar SD, Carone M, Szpiro AA, et al. 2024. Traffic-related air pollution and dementia incidence in the Adult Changes in Thought Study. Environ Int 183:108418, *https://doi.org/10.1016/j.envint.2024.108418*.

Boanini C, Mecca D, Pognant F, Bo M, Clerico M. 2021. Integrated mobile laboratory for air pollution assessment: literature review and cc-TrAIRer design. Atmosphere 128:1004, *https://doi.org/10.3390/atmos12081004*.

Boogaard H, Patton AP, Atkinson RW, Brook JR, Chang HH, Crouse DL, et al. 2022. Long-term exposure to traffic-related air pollution and selected health outcomes: a systematic review and meta-analysis. Environ Int 164:107262, *https://doi.org/10.1016/j.envint.2022.107262*.

Breiman L, Cutler A, Liaw A, Wiener M. 2022. randomForest: Breiman and Cutler's random forests for classification and regression. *https://www.stat.berkeley.edu/~breiman/RandomForests/*.

Brenner H, Loomis D. 1994. Varied forms of bias due to nondifferential error in measuring exposure. Epidemiology 55:510–517.

Brugge D, Fuller CH, eds. 2020. Ambient Combustion Ultrafine Particles and Health. Hauppauge, NY: Nova Science Publishers.

Bühlmann P, Geer S van de. 2015. High-dimensional inference in misspecified linear models. Elect J Stat 91:1449–1473, *https://doi.org/10.1214/15-EJS1041*.

Cesaroni G, Porta D, Badaloni C, Stafoggia M, Eeftens M, Meliefste K, et al. 2012. Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. Environ Health 11:48, *https://doi.org/10.1186/1476-069X-11-48*.

Chambliss SE, Pinon CPR, Messier KP, LaFranchi B, Upperman CR, Lunden MM, et al. 2021. Local- and regional-scale racial and ethnic disparities in air pollution determined by long-term mobile monitoring. Proc Natl Acad Sci 11837:e2109249118, *https://doi.org/10.1073/pnas.2109249118*.

Chambliss SE, Preble CV, Caubel JJ, Cados T, Messier KP, Alvarez RA, et al. 2020. Comparison of mobile and fixed-site black carbon measurements for high-resolution urban pollution mapping. Environ Sci Tech 5413:7848–7857, *https://doi.org/10.1021/acs.est.0c01409*.

Cheng S, Blanco MN, Larson TV, Shojaie A, Szpiro AA. 2024a. Principal component analysis balancing prediction and approximation accuracy for spatial data. Available: *https://arxiv.org/abs/2408.01662*.

Cheng S, Blanco MN, Sheppard L, Shojaie A, Szpiro AA. 2024b. Variable importance measure for spatial machine learning models with application to air pollution exposure prediction. Available: *https://arxiv.org/abs/2406.01982*.

Clark TG, Altman DG, De Stavola BL. 2002. Quantification of the completeness of follow-up. Lancet 3599314:1309–1310, *https://doi.org/10.1016/S0140-67360208272-7*.

Crane PK, Gibbons LE, McCurry SM, McCormick W, Bowen JD, Sonnen J, et al. 2016. Importance of home study visit capacity in dementia studies. Alzheimers Dement 124:419–426, *https://doi.org/10.1016/j.jalz.2015.10.007*.

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. 2008. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol 6110:1018–1027, *https://doi.org/10.1016/j.jclinepi.2007.11.011*.

Cressie N. 2015. Statistics for Spatial Data. Hoboken, NJ: John Wiley & Sons.

Datta A, Saha A, Zamora ML, Buehler C, Hao L, Xiong F, et al. 2020. Statistical field calibration of a low-cost $PM_{2.5}$ monitoring network in Baltimore. Atmos Environ 242:117761, *https://doi.org/10.1016/j.atmosenv.2020.117761*.

Delgado-Saborit JM, Guercio V, Gowers AM, Shaddick G, Fox NC, Love S. 2021. A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population. Sci Total Environ 757:143734, *https://doi.org/10.1016/j.scitotenv.2020.143734*.

Diao M, Holloway T, Choi S, O'Neill SM, Al-Hamdan MZ, Van Donkelaar A, et al. 2019. Methods, availability, and applications of $PM_{2.5}$ exposure estimates derived from ground measurements, satellite, and atmospheric models. J Air Waste Manage Assoc 6912:1391–1414, *https://doi.org/10.1080/10962247.2019.1668498*.

Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, et al. 1993. An association between air pollution and mortality in six US cities. N Engl J Med 32924:1753–1759, *https://doi.org/10.1056/NEJM199312093292401*.

Doubleday A, Blanco MN, Austin E, Marshall JD, Larson TV, Sheppard L. 2023. Characterizing ultrafine particle mobile monitoring data for epidemiology. Environ Sci Technol 5726:9538–9547, *https://doi.org/10.1021/acs.est.3c00800*.

Eeftens M, Beelen R, Fischer P, Brunekreef B, Meliefste K, Hoek G. 2011. Stability of measured and modelled spatial contrasts in NO_2 over time. Occup Environ Med 6810:765–770, *https://doi.org/10.1136/oem.2010.061135*.

Eeftens M, Tsai MY, Ampe C, Anwander B, Beelen R, Bellander T, et al. 2012. Spatial variation of $PM_{2.5}$, PM_{10}, $PM_{2.5}$ absorbance and PM_{coarse} concentrations between and within 20 European study areas and the relationship with NO_2: results of the ESCAPE project. Atmos Environ 62:303–317, *https://doi.org/10.1016/j.atmosenv.2012.08.038*.

Ehlenbach WJ, Hough CL, Crane PK, Haneuse SJPA, Carson SS, Curtis JR, et al. 2010. Association between acute care and critical illness hospitalization and cognitive function in older adults. JAMA 3038:763–770, *https://doi.org/10.1001/jama.2010.167*.

English PB, Olmedo L, Bejarano E, Lugo H, Murillo E, Seto E, et al. 2017. The Imperial County Community Air Monitoring Network: a model for community-based environmental monitoring for public health action. Environ Health Perspect 1257:074501, *https://doi.org/10.1289/ehp1772*.

Esri. 2019. ArcGIS Desktop. Available: *https://www.esri.com/*.

Farrell W, Weichenthal S, Goldberg M, Valois M.-F, Shekarrizfard M, Hatzopoulou M. 2016. Near roadway air pollution across a spatially extensive road and cycling network. Environ Pollut 212:498–507, *https://doi.org/10.1016/j.envpol.2016.02.041*.

Friedman JH. 1991. Multivariate adaptive regression splines. Ann Stat 191:1–67, *https://doi.org/10.1214/aos/1176347963*.

Gao M, Cao J, Seto E. 2015. A distributed network of low-cost continuous reading sensors to measure spatio-temporal variations of $PM_{2.5}$ in Xi'an, China. Environ Pollut 199:56–65, *https://doi.org/10.1016/j.envpol.2015.01.013*.

Geng G, Murray NL, Chang HH, Liu Y. 2018. The sensitivity of satellite-based $PM_{2.5}$ estimates to its inputs: implications to model development in data-poor regions. Environ Int 121, 550–560, *https://doi.org/10.1016/j.envint.2018.09.051*.

Gibbons L. 2015. RUNPARSCALE: Stata module to run PARSCALE from Stata [computer software]. *https://ideas.repec.org/c/boc/bocode/s456724.html*.

Gozzi F, Della Ventura G, Marcelli A. 2016. Mobile monitoring of particulate matter: State of art and perspectives. Atmos Pollut Res 72:228–234, *https://doi.org/10.1016/j.apr.2015.09.007*.

Guttorp P, Fuentes M, Sampson PD. 2007. Using transforms to analyze space-time processes. Available: *https://www.researchgate.net/publication/254653615_Using_Transforms_to_Analyze_Space-Time_Processes*.

Hajat A, Hsia C, O'Neill MS. 2015. Socioeconomic disparities and air pollution exposure: a global review. Curr Environ Health Rep 24:440–450, *https://doi.org/10.1007/s40572-015-0069-5*.

Hankey S, Marshall JD. 2015. Land use regression models of on-road particulate air pollution particle number, black carbon, $PM_{2.5}$, particle size using mobile monitoring. Environ Sci Technol 4915:9194–9202, *https://doi.org/10.1021/acs.est.5b01209*.

Hatzopoulou M, Valois MF, Levy I, Mihele C, Lu G, Bagg S, et al. 2017. Robustness of land-use regression models developed from mobile air pollutant measurements. Environ Sci Technol 517:3938–3947, *https://doi.org/10.1021/acs.est.7b00366*.

HEI. 2013. Understanding the Health Effects of Ambient Ultrafine Particles. Perspectives 3. *https://www.healtheffects.org/publication/understanding-health-effects-ambient-ultrafine-particles*. Boston, MA: Health Effects Institute.

Heimann I, Bright VB, McLeod MW, Mead MI, Popoola OAM, Stewart GB, et al. 2015. Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. Atmos Environ 113:10–19, *https://doi.org/10.1016/j.atmosenv.2015.04.057*.

Hoek G, Meliefste K, Cyrys J, Lewné M, Bellander T, Brauer M, et al. 2002. Spatial variability of fine particle concentrations in three European areas. Atmos Environ 3625, 4077–4088, *http://dx.doi.org/10.1016/S1352-2310(02)00297-2*.

Huang, K, Bi, J, Meng, X, Geng, G, Lyapustin, A, Lane, K. J, et al. 2019. Estimating daily $PM_{2.5}$ concentrations in New York City at the neighborhood-scale: Implications for integrating non-regulatory measurements. Sci Total Environ 697:134094, *https://doi.org/10.1016/j.scitotenv.2019.134094*.

Hudda N, Simon MC, Zamore W, Durant JL. 2018. Aviation-related impacts on ultrafine particle number concentrations outside and inside residences near an airport. Environ Sci Technol 524, 1765–1772, *https://doi.org/10.1021/acs.est.7b05593*.

Ikram J, Tahir A, Kazmi H, Khan Z, Javed R, Masood U. 2012. View: Implementing low cost air quality monitoring solution for urban areas. Environ Syst Res 11:10, *https://doi.org/10.1186/2193-2697-1-10*.

Jandarov RA, Sheppard LA, Sampson PD, Szpiro AA. 2017. A novel principal component analysis for spatially misaligned multivariate air pollution data. J Royal Stat Soc Series C 661:3–28, *https://doi.org/10.1111/rssc.12148*.

Jerrett M, Donaire-Gonzalez D, Popoola O, Jones R, Cohen RC, Almanza E, et al. 2017. Validating novel air pollution sensors to improve exposure estimates for epidemiological analyses and citizen science. Enviro Res 158:286–294, *https://doi.org/10.1016/j.envres.2017.04.023*.

Jiang Q, Kresin F, Bregt AK, Kooistra L, Pareschi E, van Putten E, et al. 2016. Citizen sensing for improved urban environmental monitoring. J Sensors 2016:e5656245, *https://doi.org/10.1155/2016/5656245*.

Jiao, W, Hagler, G, Williams, R, Sharpe, R, Brown, R, Garver, D, et al. 2016. Community air sensor network CAIRSENSE project: evaluation of low-cost sensor performance in a sub-

urban environment in the southeastern United States. Atmos Meas Tech 911, 5281–5292, *https://doi.org/10.5194/amt-9-5281-2016*.

Jung KH, Bernabé K, Moors K, Yan B, Chillrud SN, Whyatt R, et al. 2011. Effects of floor level and building type on residential levels of outdoor and indoor polycyclic aromatic hydrocarbons, black carbon, and particulate matter in New York City. Atmosphere 22:96–109, *https://doi.org/10.3390/atmos2020096*.

Katsouyanni K, Evangelopoulos D. 2022. Invited perspective: impact of exposure measurement error on effect estimates: an important and neglected problem in air pollution epidemiology. Environ Health Perspect 1307:71302, *https://doi.org/10.1289/EHP11277*.

Keller JP, Chang HH, Strickland MJ, Szpiro AA. 2017. Measurement error correction for predicted spatio-temporal air pollution exposures. Epidemiology 283:338–345, *https://doi.org/10.1097/EDE.0000000000000623*.

Keller JP, Olives C, Kim S-Y, Sheppard L, Sampson PD, Szpiro AA, et al. 2015. A unified spatio-temporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. Environ Health Perspect 1234:301–309, *https://doi.org/10.1289/ehp.1408145*.

Kerckhoffs J, Hoek G, Messier KP, Brunekreef B, Meliefste K, Klompmaker, et al. 2016. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ Sci Technol 5023:12894–12902, *https://doi.org/10.1021/acs.est.6b03476*.

Kerckhoffs J, Hoek G, Portengen L, Brunekreef B, Vermeulen RCH. 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ Sci Technol 533:1413–1421, *https://doi.org/10.1021/acs.est.8b06038*.

Kerckhoffs J, Hoek G, Vermeulen R. 2024. Mobile monitoring of air pollutants: performance evaluation of a mixed-model land use regression framework in relation to the number of drive days. Environ Res 240:117457, *https://doi.org/10.1016/j.envres.2023.117457*.

Kerckhoffs J, Hoek G, Vlaanderen J, van Nunen E, Messier K, Brunekreef B, et al. 2017. Robustness of intra-urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ Res 1592017:500–508, *https://doi.org/10.1016/j.envres.2017.08.040*.

Kim S-Y, Blanco MN, Bi J, Larson TV, Sheppard L. 2023. Exposure assessment for air pollution epidemiology: a scoping review of emerging monitoring platforms and designs. Environ Res 223:115451, *https://doi.org/10.1016/j.envres.2023.115451*.

Kim S-Y, Gassett AJ, Blanco MN, Sheppard L. 2025. Ultrafine particle mobile monitoring study designs for epidemiology: Cost and performance comparisons. Environ Health Perspect 133:47010, *https://doi.org/10.1289/ehp15100*.

Kim S-Y, Olives C, Sheppard L, Sampson PD, Larson TV, Keller JP, et al. 2017. Historical prediction modeling approach for estimating long-term concentrations of $PM_{2.5}$ in cohort studies before the 1999 implementation of widespread monitoring. Environ Health Perspect 12538–12546, *https://doi.org/10.1289/EHP131*.

Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, et al. 2001. The National Human Activity Pattern Survey NHAPS: a resource for assessing exposure to environmental pollutants. J Expos Anal Environ Epidemiol 11:231–252, *https://doi.org/10.1038/sj.jea.7500165*.

Klompmaker JO, Montagne DR, Meliefste K, Hoek G, Brunekreef B. 2015. Spatial variation of ultrafine particles and black carbon in two cities: results from a short-term measurement campaign. Sci Total Environ 508:266–275, *https://doi.org/10.1016/j.scitotenv.2014.11.088*.

Kukull WA. 2001. The association between smoking and Alzheimer's disease: effects of study design and bias. Biol Psych 493:194–199, *https://doi.org/10.1016/S0006-32230001077-5*.

Kukull WA, Higdon R, Bowen JD, McCormick WC, Teri L, Schellenberg GD, et al. 2002. Dementia and Alzheimer disease incidence: a prospective cohort study. Arch Neurol 5911:1737–1746, *https://doi.org/10.1001/archneur.59.11.1737*.

Kulick ER, Wellenius GA, Boehme AK, Joyce NR, Schupf N, Kaufman JD, et al. 2020. Long-term exposure to air pollution and trajectories of cognitive decline among older adults. Neurology 9417:e1782–e1792, *https://doi.org/10.1212/WNL.0000000000009314*.

Kumar P, Morawska L, Martani C, Biskos G, Neophytou M, Di Sabatino S, et al. 2015. The rise of low-cost sensing for managing air pollution in cities. Environ Int 75:199–205, *https://doi.org/10.1016/j.envint.2014.11.019*.

Larson T, Su J, Baribeau A-M, Buzzelli M, Setton E, Brauer M. 2007. A spatial model of urban winter woodsmoke concentrations. Environ Sci Technol 417:2429–2436, *https://doi.org/10.1021/es0614060*.

Levy I, Levin N, Yuval Y, Schwartz JD, Kark JD. 2015. Back-extrapolating a land use regression model for estimating past exposures to traffic-related air pollution. Environ Sci Technol 496:3603–3610, *https://doi.org/10.1021/es505707e*.

Levy I, Mihele C, Lu G, Narayan J, Hilker N, Brook JR. 2014. Elucidating multipollutant exposure across a complex metropolitan area by systematic deployment of a mobile laboratory. Atmos Chem Phys 1414:7173–7193, *https://doi.org/10.5194/acp-14-7173-2014*.

Li G, Larson EB, Shofer JB, Crane PK, Gibbons LE, McCormick W, et al. 2017. Cognitive trajectory changes over 20 years before dementia diagnosis: a large cohort study. J Am Geriatr Soc 6512:2627–2633, *https://doi.org/10.1111/jgs.15077*.

Lindström J, Szpiro AA, Sampson PD, Bergen S, Oron AP, Young MT, et al. 2023. Spatio-temporal Version 1.1.7 Version 1.1.17 [computer software]. Available: *https://github.com/kaufman-lab/Spatio-temporal*.

Lindström J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson TV, et al. 2014. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ Ecol Stat 213:411–433, *https://doi.org/10.1007/s10651-013-0261-4*.

Liu L-J, Box M, Kalman D, Kaufman J, Koenig J, Larson T, et al. 2003. Exposure assessment of particulate matter for susceptible populations in Seattle. Environ Health Perspect 111: 909–918, *https://doi.org/10.1289/ehp.6011*.

Loeppky JA, Cagle AS, Sherriff M, Lindsay A, Willis P. 2013. A local initiative for mobile monitoring to measure residential wood smoke concentration and distribution. Air Qual Atmos Health 63:641–653, *https://doi.org/10.1007/s11869-013-0203-1*.

Malings C, Tanzer R, Hauryliuk A, Kumar SPN, Zimmerman N, Kara LB, et al. 2019. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. Atmos Meas Tech 122:903–920, *https://doi.org/10.5194/amt-12-903-2019*.

Mead MI, Popoola OAM, Stewart GB, Landshoff P, Calleja M, Hayes M, et al. 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos Environ 70:186–203, *https://doi.org/10.1016/j.atmosenv.2012.11.060*.

Meng J, Li C, Martin RV, van Donkelaar A, Hystad P, Brauer M. 2019. Estimated long-term 1981–2016 concentrations of ambient fine particulate matter across North America from chemical transport modeling, satellite remote sensing, and ground-based measurements. Environ Sci Technol 539:5071–5079, *https://doi.org/10.1021/acs.est.8b06875*.

Messier KP, Chambliss SE, Gani S, Alvarez R, Brauer M, Choi JJ, et al. 2018. Mapping air pollution with Google Street View cars: efficient approaches with mobile monitoring and land use regression. Environ Sci Technol 5221:12563–12572, *https://doi.org/10.1021/acs.est.8b03395*.

Miles JN, Weden MM, Lavery D, Escarce JJ, Cagney KA, Shih RA. 2016. Constructing a time-invariant measure of the socio-economic status of US census tracts. J Urban Health 931:213–232, *https://doi.org/10.1007/s11524-015-9959-y*.

Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, et al. 2007. Long-term exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med 3565:447–458, *https://doi.org/10.1056/NEJMoa054409*.

Minet L, Liu R, Valois MF, Xu J, Weichenthal S, Hatzopoulou M. 2018. Development and comparison of air pollution exposure surfaces derived from on-road mobile monitoring and short-term stationary sidewalk measurements. Environ Sci Technol 526:3512–3519, *https://doi.org/10.1021/acs.est.7b05059*.

Moltchanov S, Levy I, Etzion Y, Lerner U, Broday DM, Fishbain B. 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Sci Total Environ 502:537–547, *https://doi.org/10.1016/j.scitotenv.2014.09.059*.

Molter A, Lindley S, de Vocht F, Simpson A, Agius, R. 2010. Modelling air pollution for epidemiologic research – part II: predicting temporal variation through land use regression. Sci Total Environ 4091:211–217, *https://doi.org/10.1016/j.scitotenv.2010.10.005*.

Montagne DR, Hoek G, Klompmaker JO, Wang M, Meliefste K, Brunekreef B. 2015. Land use regression models for ultrafine particles and black carbon based on short-term monitoring predict past spatial variation. Environ Sci Technol 4914:8712–8720, *https://doi.org/10.1021/es505791g*.

Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, et al. 2018. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Environ Int 116:286–299, *https://doi.org/10.1016/j.envint.2018.04.018*.

Mueller MD, Hasenfratz D, Saukh O, Fierz M, Hueglin C. 2016. Statistical modelling of particle number concentration in Zurich at high spatio-temporal resolution utilizing data from a mobile sensor network. Atmos Environ 126:171–181, *https://doi.org/10.1016/j.atmosenv.2015.11.033*.

Peters R, Ee N, Peters J, Booth A, Mudway I, Anstey KJ. 2019. Air pollution and dementia: a systematic review. J Alzheimer Dis 70s1:S145–S163, *https://doi.org/10.3233/JAD-180631*.

Pirjola L, Lähde T, Niemi JV, Kousa A, Rönkkö T, Karjalainen P, et al. 2012. Spatial and temporal characterization of traffic emissions in urban microenvironments with a mobile laboratory. Atmos Environ 63:156–167, *https://doi.org/10.1016/j.atmosenv.2012.09.022*.

Presto AA, Saha PK, Robinson AL. 2021. Past, present, and future of ultrafine particle exposures in North America. Atmos Environ X 10:100109, *https://doi.org/10.1016/j.aeaoa.2021.100109*.

PSCAA. 2020. 2019 Air Quality Data Summary. Puget Sound Clean Air Agency. Available: *https://pscleanair.gov/DocumentCenter/View/4164/Air-Quality-Data-Summary-2019*.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Available: *https://www.R-project.org*.

Saha PK, Li HZ, Apte JS, Robinson AL, Presto AA. 2019. Urban ultrafine particle exposure assessment with land-use regression: influence of sampling strategy. Environ Sci Technol 53:7326–7336, *https://doi.org/10.1021/acs.est.9b02086*.

Saha PK, Presto AA, Hankey S, Marshall JD, Robinson AL. 2022. Racial-ethnic exposure disparities to airborne ultrafine particles in the United States. Environ Res Lett 1710:104047, *https://doi.org/10.1088/1748-9326/ac95af*.

Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, et al. 2013. A regionalized national universal kriging model using Partial Least Squares regression for estimating annual $PM_{2.5}$ concentrations in epidemiology. Atmos Environ 75:383–392, *https://doi.org/10.1016/j.atmosenv.2013.04.015*.

Sampson PD, Szpiro AA, Sheppard L, Lindström J, Kaufman JD. 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmos Environ 4536:6593–6606, *https://doi.org/10.1016/j.atmosenv.2011.04.073*.

Sather ME, Slonecker ET, Mathew J, Daughtrey H, Williams DD. 2007. Evaluation of Ogawa passive sampling devices as an alternative measurement method for the nitrogen dioxide annual standard in El Paso, Texas. Environ Monit Assess 1241:211–221, *https://doi.org/10.1007/s10661-006-9219-4*.

Schulte JK, Fox JR, Oron AP, Larson TV, Simpson CD, Paulsen M, et al. 2015. Neighborhood-scale spatial models of diesel exhaust concentration profile using 1-Nitropyrene and other nitroarenes. Environ Sci Technol 4922:13422–13430, *https://doi.org/10.1021/acs.est.5b03639*.

Shaffer RM, Blanco MN, Li G, Adar SD, Carone M, Szpiro AA, et al. 2021a. Fine particulate matter and dementia incidence in the adult changes in thought study. Environ Health Perspect 1298:087001, *https://doi.org/10.1289/EHP9018*.

Shaffer RM, Li G, Adar SD, Dirk Keene C, Latimer CS, Crane PK, et al. 2021b. Fine particulate matter and markers of Alzheimer's disease neuropathology at autopsy in a community-based cohort. J Alzheimers Dis 794:1761–1773, *https://doi.org/10.3233/JAD-201005*.

Sorek-Hamer M, Chatfield R, Liu Y. 2020. Review: strategies for using satellite-based products in modeling $PM_{2.5}$ and short-term pollution episodes. Environ Int 144:106057, *https://doi.org/10.1016/j.envint.2020.106057*.

Szpiro AA, Paciorek CJ. 2013a. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics 248:501–517, *https://doi.org/10.1002/env.2233*.

Szpiro AA, Paciorek CJ. 2013b. Rejoinder. Environmetrics 248:531–536, *https://doi.org/10.1002/env.2254*.

Szpiro AA, Paciorek CJ, Sheppard, L. 2011a. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology 225:680–685, *https://doi.org/10.1097/EDE.0b013e3182254cc6*.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 216:606–631, *https://doi.org/10.1002/env.1014*.

Szpiro AA, Sheppard L, Lumley T. 2011b. Efficient measurement error correction with spatially misaligned data. Biostatistics 124:610–623, *https://doi.org/10.1093/biostatistics/kxq083*.

Teng EL, Larson EB, Lin KN, Graves AB, Liu HC. 1998. The cognitive abilities screening instrument: short version CASI-Short. Clin Neuropsych 122:256–256.

US EPA (US Environmental Protection Agency). 2008. Ambient Air Monitoring Strategy for State, Local, and Tribal Air Agencies. Office of Air Quality Planning and Standards. Available: *https://www.epa.gov/sites/default/files/2020-09/documents/aams_for_slts_-_final_dec_2008.pdf*.

US EPA (US Environmental Protection Agency). 2019. Integrated science assessment ISA for particulate matter. US Environmental Protection Agency. Available: *https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=347534*.

Vallabani NVS, Gruzieva O, Elihn K, Juárez-Facio AT, Steimer SS, Kuhn J, et al. 2023. Toxicity and health effects of ultrafine particles: towards an understanding of the relative impacts of different transport modes. Environ Res 231:116186, *https://doi.org/10.1016/j.envres.2023.116186*.

van de Beek E, Kerckhoffs J, Hoek G, Sterk G, Meliefste K, Gehring U, et al. 2021. Spatial and spatio-temporal variability of regional background ultrafine particle concentrations in the Netherlands. Environ Sci Technol 552:1067–1075, *https://doi.org/10.1021/acs.est.0c06806*.

van Nunen E, Vermeulen R, Tsai M-Y, Probst-Hensch N, Ineichen A, Davey M, et al. 2017. Land use regression models for ultrafine particles in six European areas. Environ Sci Technol 516:3336–3345, *https://doi.org/10.1021/acs.est.6b05920*.

Vardoulakis S, Giagloglou E, Steinle S, Davis A, Sleeuwenhoek A, Galea KS, et al. 2020. Indoor exposure to selected air pollutants in the home environment: a systematic review. Int J Environ Res Public Health 1723:8972, *https://doi.org/10.3390/ijerph17238972*.

Wagstaff M, Henderson SB, McLean KE, Brauer M. 2022. Development of methods for citizen scientist mapping of residential woodsmoke in small communities. J Environ Manage 311:114788, *https://doi.org/10.1016/j.jenvman.2022.114788*.

Wai TH, Young MT, Szpiro AA. 2020. Random Spatial Forests arXiv:2006.00150. arXiv. Available: *https://doi.org/10.48550/arXiv.2006.00150*.

Wang A, Paul S, deSouza P, Machida Y, Mora S, Duarte F, et al. 2023. Key themes, trends, and drivers of mobile ambient air quality monitoring: a systematic review and meta-analysis. Environ Science Technol 5726:9427–9444, *https://doi.org/10.1021/acs.est.2c06310.*

Wang M, Keller JP, Adar SD, Kim S-Y, Larson TV, Olives C, et al. 2015. Development of long-term spatio-temporal models for ambient ozone in six metropolitan regions of the United States: the MESA air study. Atmos Environ 123:79–87, *https://doi.org/10.1016/j.atmosenv.2015.10.042*.

Wang R, Henderson S, Sbihi H, Allen R, Brauer M. 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmos Environ 64:312–319, *https://doi.org/10.1016/j.atmosenv.2012.09.056*.

Wang Y, Hopke PK, Chalupa DC, Utell MJ. 2011. Long-term study of urban ultrafine particles and other pollutants. Atmos Environ 4540:7672–7680, *https://doi.org/10.1016/j.atmosenv.2010.08.022*.

Wei Y, Qiu X, Yazdi MD, Shtein A, Shi L, Yang J, et al. 2022. The impact of exposure measurement error on the estimated concentration-response relationship between long-term exposure to $PM_{2.5}$ and mortality. Environ Health Perspect 1307:77006, *https://doi.org/10.1289/EHP10389*.

Weichenthal S, Ryswyk KV, Goldstein A, Bagg S, Shekkarizfard M, Hatzopoulou M. 2016. A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach. Environ Res 146:65–72, *https://doi.org/10.1016/j.envres.2015.12.016*.

Wilton D, Szpiro A, Gould T, Larson T. 2010. Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA, and Seattle, WA. Sci Total Environ 4085:1120–1130, *https://doi.org/10.1016/j.scitotenv.2009.11.033*.

Wong, S. 2010. A Spatial Model to Assess the Impact of Major Roadways on a Low-Income Seattle Neighborhood Using an Intensive NO_x Sampling Campaign. In Department of Environmental and Occupational Health Sciences: Vol. MS. Seattle, WA: University of Washington.

Wood SN. 2003. Thin plate regression splines. J Royal Stat Soc B 651:95–114, *https://doi.org/10.1111/1467-9868.00374*.

Wood SN. 2017. Generalized Additive Models: An Introduction with R. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Young MT, Bechle MJ, Sampson PD, Szpiro AA, Marshall JD, Sheppard L, et al. 2016. Satellite-based NO_2 and model validation in a national prediction model based on universal kriging and land-use regression. Environ Sci Technol 507:3686–3694, *https://doi.org/10.1021/acs.est.5b05099*.

Yuan Z, Kerckhoffs J, Hoek G, Vermeulen R. 2022. A knowledge transfer approach to map long-term concentrations of hyperlocal air pollution from short-term mobile measurements. Environ Sci Technol 5619:13820–13828, *https://doi.org/10.1021/acs.est.2c05036*.

Zhu Y, Hinds WC, Kim S, Shen S, Sioutas C. 2002. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. Atmospheric Environment, 3627, 4323–4335, *https://doi.org/10.1016/S1352-23100200354-0*.

Zimmerman N, Presto AA, Kumar SPN, Gu, J, Hauryliuk, A, Robinson ES, Robinson AL, et al. 2018. Calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech 111:291–313, *http://dx.doi.org/10.5194/amt-11-291-2018*.

Zuidema C, Bi J, Burnham D, Carmona N, Gassett AJ, Slager DL, et al. 2024. Leveraging low-cost sensors to predict nitrogen dioxide for epidemiologic exposure assessment. J Expo Sci Environ Epidemiol April 9. *https://doi.org/10.1038/s41370-024-00667-w*.

Zuidema C, Schumacher CS, Austin E, Carvlin G, Larson TV, Spalt EW, et al. 2021. Deployment, calibration, and cross-validation of low-cost electrochemical sensors for carbon monoxide, nitrogen oxides, and ozone for an epidemiological study. Sensors 2112:4214, *https://doi.org/10.3390/s21124214*.

Zusman M, Schumacher CS, Gassett AJ, Spalt EW, Austin E, Larson TV, et al. 2020. Calibration of low-cost particulate matter sensors: model development for a multi-city epidemiological study. Environ Int 134:105329, *https://doi.org/10.1016/j.envint.2019.105329*.

HEI QUALITY ASSURANCE STATEMENT

Westat staff conducted an independent audit of this study. The staff are experienced quality assurance (QA) auditors with expertise in statistical modeling, epidemiology, and exposure assessment. The Westat QA audit team consisted of Dr. Daniel Chacreton and Ms. Rebecca Birch. These staff members are highly experienced in quality assurance oversight across various relevant domains. The QA oversight program consisted of a remote audit of the final report and the data processing steps. Key details of the dates of the audit and the reviews performed are listed below.

Date: January 2025 – April 2025

Remarks: The Sheppard et al. study underwent an independent QA audit by two Westat auditors with quality assurance oversight experience and expertise relevant to exposure assessment, air quality monitoring and modeling, epidemiological methods, and statistical analysis. The Westat QA audit of the final Sheppard et al. report focused on adherence to the study protocol, appropriateness of the documentation of the study methods (e.g., data processing, exposure modeling, and statistical modeling), whether study assumptions and limitations were adequately addressed, and whether the investigators' conclusions were reasonable given the study findings and in consideration of the limitations. The QA team also evaluated whether the report was easy to understand.

The Westat QA audit team provided a written report to HEI and the study investigators. The Westat QA auditors concluded that the study was well conducted in accordance with the study protocol and Quality Assurance Project Plan. Methods are explicitly outlined with reference to sampling techniques, exposure modeling, and quality control procedures. The QC steps taken, particularly related to data handling, modeling strategies, and cross-validation processes, are clearly documented. The interpretations provided align clearly with the reported results. The auditors also provided HEI and the investigator team with specific recommendations for improvement, all considered minor. Areas of QA feedback included improved clarity and accessibility of some visual presentations in the main report and the supplementary material. Overall, the report is well written and provides good documentation, clear interpretations, and no major gaps were identified.

The Westat QA audit team attests that the final report appears to be representative of the study conducted.

Daniel Chacreton, PhD, Statistician, Quality Assurance auditor

Rebecca Jeffries Birch, MPH, Epidemiologist, Quality Assurance auditor

Date: April 8, 2025

ADDITIONAL MATERIALS ON THE HEI WEBSITE

Additional Materials for Chapters 3, 4, 6, 7, 8, 9, and 10 are available on the HEI website at *https://www.healtheffects.org/publications*.

ABOUT THE AUTHORS

Lianne Sheppard is the Rohm & Haas Endowed Professor in Public Health Sciences and interim chair of environmental and occupational health sciences at the University of Washington. She has a PhD in biostatistics from the University of Washington and an ScM in biostatistics from The Johns Hopkins University. She is a professor in two departments: Environmental and Occupational Health Sciences and Biostatistics. Her research focuses on understanding the health effects of environmental and occupational exposures, with particular emphasis on the implications of exposure assessment design and modeling on inference about health effects. She is an elected fellow of the American Statistical Association and former chair of the chartered Clean Air Scientific Advisory Committee. Sheppard is the principal investigator of this project.

Magali N. Blanco is a postdoctoral scholar and incoming assistant professor in the department of environmental and occupational health sciences at the University of Washington. She holds a PhD in environmental health and an MS in exposure science from the University of Washington. Her research focuses on quantitative air pollution exposure assessment for epidemiology, with a focus on brain health. Blanco contributed to all report chapters and is the lead author of several of the chapters.

Sun-Young Kim is a professor at the Graduate School of Cancer Science and Policy of the National Cancer Center, South Korea. She has a PhD in public health sciences with a biostatistics major from Seoul National University, South Korea. Her research focuses on the assessment of environmental exposures and their health effects, with a particular interest in air pollution and cancer. Kim is the lead author on Chapter 9, focusing on the value of information for mobile monitoring designs.

Annie Doubleday is an ambient air quality epidemiologist at the Washington State Department of Health. She holds a PhD and MPH in environmental health from the University of Washington. Her work focuses on air pollution and wildfire smoke epidemiology. Doubleday contributed to the analysis of on-road data presented in Chapter 6.

Si Cheng received her PhD in biostatistics from the University of Washington and her MS in biostatistics from Yale University. She is currently a data scientist at Netflix. Cheng's academic research focuses on statistical machine learning methodologies for spatial- and network-linked data, including applications to air pollution modeling, described in Chapter 8 of this project, where she is the lead author.

Christopher Zuidema is a clinical assistant professor in the department of environmental and occupational health sciences at the University of Washington. He received a PhD from The Johns Hopkins University and an MS from Harvard University. He is a certified industrial hygienist and a staff member of the University of Washington Field Research and Consultation Group. His research interests are in environmental exposure assessment and exposure assessment for environmental epidemiology. Zuidema led the spatiotemporal modeling of NO_2 and is one of the lead authors for Chapter 7.

Jianzhao Bi is an independent research scientist and former postdoctoral scholar in the department of environmental and occupational health sciences at the University of Washington. He earned his PhD in environmental health sciences from Emory University and his MS in atmospheric science from Tsinghua University in Beijing, China. His research focuses on spatial statistical and physical modeling of environmental exposures and natural hazards. Bi led the spatiotemporal modeling of $PM_{2.5}$ analysis and is one of the lead authors for Chapter 7.

Amanda Gassett is a staff scientist in the department of environmental and occupational health sciences at the University of Washington. She holds an MS in biostatistics from the University of Washington. She has provided support for environmental health research at the University of Washington for nearly two decades, focusing on traffic-related and ambient air pollution. Gassett was the project manager for the field data collection component of the ACT-AP study, contributed to Chapter 9, and is the lead author of the Chapter 9 Additional Materials, which are focused on the value of information for adding low-cost sensor data to existing regulatory monitoring datasets.

Ali Shojaie is a professor of biostatistics and statistics at the University of Washington and an associate chair of the department of biostatistics, founding director of the Summer Institute for Statistics in Big Data at the University of Washington, and lead of the Data Management and Statistics Core of the University of Washington Alzheimer's Disease Research Center. He earned his PhD in statistics from the University of Michigan. His research lies in the intersection of statistical machine learning, statistical network analysis, and applications in biology and social sciences. He is an elected Fellow of the American Statistical Association (ASA) and the Institute of Mathematical Statistics and a recipient of the Leo Breiman Award from the ASA Section on Statistical Learning and Data Science. Shojaie is a co-author on and provided essential methodological guidance to Chapter 8.

Adam A. Szpiro is a professor in the department of biostatistics at the University of Washington. He has a PhD in applied mathematics from Brown University and trained as a postdoctoral scholar in biostatistics at the University of Washington. His research focuses on developing and applying advanced statistical methods for environmental epidemiology, including spatiotemporal air pollution prediction, measurement error correction, machine learning methods for spatial and spatiotemporal statistics, unmeasured spatial and temporal confounding, and mixture analysis for air pollution and other environmental toxicants. Szpiro is a co-author on multiple report chapters and provided essential methodological guidance to Chapters 5 and 8.

OTHER PUBLICATIONS

This section has three parts: Publications Resulting from This Research, Summary of Publications by Aim, and Annotated Bibliography. Many of the publications in the Publications Resulting from This Research are also included in the Annotated Bibliography. A few also benefited from this funding but did not directly contribute to the results described in this report (Bi et al. 2022; Bramble et al. 2023; Kim et al. 2025). The Summary of Publications by Aim and Annotated Bibliography also describes several papers in preparation (not yet submitted).

PUBLICATIONS RESULTING FROM THIS RESEARCH

Bi J, Birnham D, Zuidema C, Schumacher C, Gassett AJ, Szpiro AA, et al. 2024. Evaluating low-cost monitoring designs for $PM_{2.5}$ exposure assessment with a spatio-temporal modeling approach. Environ Pollut 343:123227, *https://doi.org/10.1016/j.envpol.2023.123227*.

Bi J, Carmona N, Blanco MN, Gassett AJ, Seto E, Szpiro AA, et al. 2022. Publicly available low-cost sensor measurements for $PM_{2.5}$ exposure modeling: guidance for monitor deployment and data selection. Environ Int 158:106897, *https://doi.org/10.1016/j.envint.2021.106897*.

Blanco MN, Bi J, Austin E, Larson TV, Marshall JD, Sheppard L. 2023. Impact of mobile monitoring network design

on air pollution exposure assessment models. Environ Sci Technol 57:440–450, *https://doi.org/10.1021/acs.est.2c05338*.

Blanco MN, Doubleday A, Austin E, Marshall JD, Seto E, Larson TV, et al. 2022. Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. J Expo Sci Environ Epidemiol 33:465–473. *https://doi.org/10.1038/s41370-022-00470-5*.

Blanco MN, Gassett A, Gould T, Doubleday A, Slager DL, Austin E, et al. 2022. Characterization of annual average traffic-related air pollution concentrations in the greater Seattle area from a year-long mobile monitoring campaign. Environ Sci Technol 56:11460–11472, *https://doi.org/10.1021/acs.est.2c01077*.

Blanco MN, Shaffer RM, Li G, Adar SD, Carone M, Szpiro AA, et al. 2024. Traffic-related air pollution and dementia incidence in the Adult Changes in Thought Study. Environ Int 183:108418, *https://doi.org/10.1016/j.envint.2024.108418*.

Blanco MN, Szpiro AA, Crane PK, Sheppard L. 2025. Ultrafine particles and late-life cognitive function: Influence of stationary mobile monitoring design on health inferences. Environ Pollut 10;374:126222, *https://doi.org/10.1016/j.envpol.2025.126222*.

Bramble K, Blanco MN, Doubleday A, Gassett A, Hajat A, Marshall JD, et al. 2023. Exposure disparities by income, race and ethnicity, and historic redlining grade in the greater Seattle area for ultrafine particles and other air pollutants. Environ Health Perspect 131:77004, *https://doi.org/10.1289/ehp11662*.

Cheng Si, Blanco MN, Larson TV, Sheppard L, Szprio A, Shojaie A. 2024. Principal component analysis balancing prediction and approximation accuracy for spatial data. Available: *https://arxiv.org/abs/2408.01662*.

Cheng Si, Blanco MN, Sheppard L, Shojaie A, Szpiro A. 2024. Variable importance measure for spatial machine learning models with application to air pollution exposure prediction. Available: *https://arxiv.org/abs/2406.01982*.

Doubleday A, Blanco MN, Austin E, Marshall JD, Larson TV, Sheppard L. 2023. Characterizing ultrafine particle mobile monitoring data for epidemiology. Environ Sci Technol 57:9538–9547, *https://doi.org/10.1021/acs.est.3c00800*.

Kim S-Y, Blanco MN, Bi J, Sheppard L. 2023. Exposure assessment for epidemiology: a scoping review of emerging monitoring platforms and designs. Environ Res 223:115451, *https://doi.org/10.1016/j.envres.2023.115451*.

Kim S-Y, Gassett AJ, Blanco MN, Sheppard L. 2025. Ultrafine particle mobile monitoring study designs for epidemiology: cost and performance comparisons. Environ Health Perspect 133:47010, *https://doi.org/10.1289/ehp15100*.

Zuidema C, Bi J, Burham D, Carmona N, Gassett AJ, Slager D, et al. 2024. Leveraging low-cost sensors to predict nitrogen dioxide for epidemiologic exposure assessment. J Expo Sci Environ Epidemiol 35:169–179, *https://doi.org/10.1038/s41370-024-00667-w*.

SUMMARY OF PUBLICATIONS BY AIM

Aim 1a focused on the design of mobile monitoring campaigns. Our initial focus was on the stationary roadside data from the Seattle mobile monitoring campaign (i.e., locations where the mobile platform pulled over by the side of the road for 2 minutes at designated locations). Blanco and colleagues (2022), *Characterization of Annual Average Traffic-Related Air Pollution Concentrations in the Greater Seattle Area from a Year-Long Mobile Monitoring Campaign*, provided the foundation and basic data description of this dataset. Our general approach to exploring various exposure assessment designs by sampling monitoring data was first applied in Blanco and colleagues (2022), *Design and Evaluation of Short-Term Monitoring Campaigns for Long-Term Air Pollution Exposure Assessment.* This study was based on NO_x regulatory monitoring data from California. Our first results from the Seattle mobile monitoring campaign, conducted under this aim, were published as Blanco and colleagues (2023), *Design and Evaluation of Short-Term Monitoring Campaigns for Long-Term Air Pollution Exposure Assessment.* Further work on this aim is presented in Chapter 4 for the stationary data. We then turned to the assessment of the on-road data from this campaign. Doubleday and colleagues (2023), *Characterizing Ultrafine Particle Mobile Monitoring Data for Epidemiology,* describe the data and present predictions from the Seattle mobile monitoring campaign on-road data, as well as the methods we used for the plume adjustment. It also characterizes the on-road data just before arriving at and just after departing from a stationary location. (This analysis was planned in our proposal, but because it yielded less important insights than other on-road analyses, we do not present it in this report.) Further work on this aim is presented in Chapter 6 for the on-road data.

Aim 1b focused on the design of low-cost sensor campaigns. While the spatiotemporal modeling and model assessment build on a long history of previous work from the University of Washington (Keller et al. 2015; Lindström et al. 2014; Szpiro et al. 2010), the direct products from this project were Bi and colleagues (2024), *Evaluating Low-Cost Monitoring Designs for $PM_{2.5}$ Exposure Assessment with a Spatiotemporal Modeling Approach*, and Zuidema and colleagues (2024), *Leveraging Low-Cost Sensors to Predict Nitrogen Dioxide for Epidemiologic Exposure Assessment.* Both these papers relied on earlier published work by our group that described the calibration of the low-cost sensors (Zuidema et al. 2021; Zusman et al. 2020).

Aim 2 focused on statistical methods for predicting exposure. We developed a new variable importance metric presented in Cheng and colleagues (in review), *Variable Importance Measure for Spatial Machine Learning Models with Application to Air Pollution Exposure Prediction.* We also developed a new approach for exposure mixtures using principal component analysis in Cheng and colleagues (in review), *Principal Component Analysis Balancing Prediction and Approximation Accuracy for Spatial Data.*

Aim 3 focused on health inference. There are three papers in preparation or in review from this work, and presented in Chapters 4, 5, and 6. The work presented in Chapter 4 focused on the mobile monitoring of stationary roadside data in the 2024 preprint by Blanco and colleagues, *Impact of Roadside Mobile Monitoring Design on Epidemiologic Inference: A Case Study of Ultrafine Particles and Cognitive Function*. In Chapter 5, we applied an exposure measurement error adjustment to the inferential results presented for the all data model in Chapter 4. Blanco and colleagues (in preparation), *Implications of Exposure Measurement Error on Inference About Air Pollution and Cognitive Function.* In the last paper, we addressed health inference using exposure models developed from the on-road mobile data. This is presented in Blanco and colleagues (in preparation), *Epidemiologic Inference from On-Road Mobile Monitoring of Air Pollution: A Case Study of Ultrafine Particles and Cognitive Function*.

Aim 4 focused on the value of information. The first paper we published related to this aim was a scoping review of emerging monitoring platforms and designs. This was published as Kim and colleagues (2023), *Exposure Assessment for Air Pollution Epidemiology: A Scoping Review of Emerging Monitoring Platforms and Designs*. Further, we have submitted a cost and performance evaluation for mobile monitoring designs with Kim as the first author: *Ultrafine Particle Mobile Monitoring Study Designs for Epidemiology: Cost and Performance Comparison* (in review).

ANNOTATED BIBLIOGRAPHY

Aim 1a (4 papers)

1. Blanco MN, Gassett A, Gould T, Doubleday A, Slager DL, Austin E, et al. 2022. Characterization of annual average traffic-related air pollution concentrations in the greater Seattle area from a year-long mobile monitoring campaign. Environ Sci Technol 56:11460–11472, *https://doi.org/10.1021/acs.est.2c01077*.

   ABSTRACT. Growing evidence links traffic-related air pollution (TRAP) to adverse health effects. We designed an innovative and extensive mobile monitoring campaign to characterize TRAP exposure levels for the Adult Changes in Thought (ACT) study, a Seattle-based cohort. The campaign measured particle number concentration (PNC) to capture ultrafine particles (UFPs), black carbon (BC), nitrogen dioxide (NO_2), fine particulate matter ($PM_{2.5}$), and carbon dioxide (CO_2) at 309 roadside sites within a large, 1,200 land km^2 (463 mi^2) area representative of the cohort. We collected about 29 2-minute measurements at each site during all seasons, days of the week, and most times of day over a 1-year period. Validation showed good agreement between our BC, NO_2, and $PM_{2.5}$ measurements and monitoring agency sites (R^2 = 0.68–0.73). Universal kriging–partial least squares models of annual average pollutant concentrations had cross-validated mean squared error-based R^2 (and root mean squared error) values of 0.77 (1,177 pt/cm^3) for PNC, 0.60 (102 ng/m^3) for BC, 0.77 (1.3 ppb) for NO_2, 0.70 (0.3 µg/m^3) for $PM_{2.5}$, and 0.51 (4.2 ppm) for CO_2. Overall, we found that the design of this extensive campaign captured the spatial pollutant variations well, and these were explained by sensible land use features, including those related to traffic.

2. Blanco MN, Doubleday A, Austin E, Marshall JD, Seto E, Larson TV, et al. 2022. Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. J Expo Sci Environ Epidemiol 33:465–473, *https://doi.org/10.1038/s41370-022-00470-5*.

   ABSTRACT. Background: Short-term mobile monitoring campaigns to estimate long-term air pollution levels are becoming increasingly common. Still, many campaigns have not conducted temporally balanced sampling, and few have looked at the implications of such study designs for epidemiologic exposure assessment.

   Objective: We carried out a simulation study using fixed-site air quality monitors to better understand how different short-term monitoring designs impact the resulting exposure surfaces.

   Methods: We used Monte Carlo resampling to simulate three archetypal short-term monitoring sampling designs using oxides of nitrogen (NO_x) monitoring data from 69 regulatory sites in California: a year-around Balanced Design that sampled during all seasons of the year, days of the week, and all or various hours of the day; a temporally reduced Rush Hours Design; and a temporally reduced Business Hours Design. We evaluated the performance of each design's land use regression prediction model.

   Results: The Balanced Design consistently yielded the most accurate annual averages, while the reduced Rush Hours and Business Hours Designs generally produced more biased results.

   Significance: A temporally balanced sampling design is crucial for short-term campaigns such as mobile monitoring aiming to assess long-term exposure in epidemiologic cohorts.

3. Blanco MN, Bi J, Austin E, Larson TV, Marshall JD, Sheppard L. 2023. Impact of mobile monitoring network design on air pollution exposure assessment models. Environ Sci Technol 57:440–450, *https://doi.org/10.1021/acs.est.2c05338*.

   ABSTRACT. Short-term mobile monitoring campaigns are increasingly used to assess long-term air pollution exposure in epidemiology. Little is known about how monitoring network design features, including the number of stops and sampling temporality, impact exposure assessment models. We address this gap by leveraging an extensive mobile monitoring campaign conducted in the

greater Seattle area over the course of a year, during all days of the week and most hours. The campaign measured total particle number concentration (PNC; sheds light on ultrafine particulate number concentration), black carbon (BC), nitrogen dioxide (NO_2), fine particulate matter ($PM_{2.5}$), and carbon dioxide (CO_2). In Monte Carlo sampling of 7,327 total stops (278 sites × 26 visits each), we restricted the number of sites and visits used to estimate annual averages. Predictions from the all data campaign performed well, with cross-validated R^2s of 0.51–0.77. We found similar model performances (85% of the all data campaign R^2) with ~1,000–3,000 randomly selected stops for NO_2, PNC, and BC, and ~4,000–5,000 stops for $PM_{2.5}$ and CO_2. Campaigns with additional temporal restrictions (e.g., business hours, rush hours, weekdays, or fewer seasons) had reduced model performances and different spatial surfaces. Mobile monitoring campaigns wanting to assess long-term exposure should carefully consider their monitoring designs.

4. Doubleday A, Blanco MN, Austin E, Marshall JD, Larson TV, Sheppard L. 2023. Characterizing ultrafine particle mobile monitoring data for epidemiology. Environ Sci Technol 57:9538–9547, *https://doi.org/10.1021/acs.est.3c00800*.

ABSTRACT. Mobile monitoring is increasingly used to assess exposure to traffic-related air pollutants (TRAP), including ultrafine particles (UFPs). Due to the rapid spatial decrease in the concentration of UFPs and other TRAPs with distance from roadways, mobile measurements may be nonrepresentative of residential exposures, which are commonly used for epidemiologic studies. Our goal was to develop, apply, and test one possible approach for using mobile measurements in exposure assessment for epidemiology. We used an absolute principal component score model to adjust the contribution of on-road sources in mobile measurements to provide exposure predictions representative of cohort locations. We then compared UFP predictions at residential locations from mobile on-road plume-adjusted versus stationary measurements to understand the contribution of mobile measurements and characterize their differences. We found that predictions from mobile measurements are more representative of cohort locations after down-weighting the contribution of localized on-road plumes. Further, predictions at cohort locations derived from mobile measurements incorporate more spatial variation compared to those from short-term stationary data. Sensitivity analyses suggest that this additional spatial information captures features in the exposure surface not identified from the stationary data alone. We recommend the correction of mobile measurements to create exposure predictions representative of residential exposure for epidemiology.

Aim 1b (2 papers)

1. Bi J, Burnham D, Zuidema C, Schumacher C, Gassett AJ, Szpiro AA, et al. 2024. Evaluating low-cost monitoring designs for $PM_{2.5}$ exposure assessment with a spatiotemporal modeling approach. Environ Pollut 343:123227, *https://doi.org/10.1016/j.envpol.2023.123227*.

ABSTRACT. Determining the most feasible and cost-effective approaches to improving $PM_{2.5}$ exposure assessment with low-cost monitors (LCMs) can considerably enhance the quality of its epidemiological inferences. We investigated features of fixed-site LCM designs that most impact $PM_{2.5}$ exposure estimates to be used in long-term epidemiological inference for the Adult Changes in Thought Air Pollution (ACT-AP) study. We used ACT-AP collected and calibrated LCM $PM_{2.5}$ measurements at the 2-week level from April 2017 to September 2020 (N of monitors [measurements] = 82 [502]). We also acquired reference-grade $PM_{2.5}$ measurements from January 2010 to September 2020 (N = 78 [6186]). We used a spatiotemporal modeling approach to predict $PM_{2.5}$ exposures with either all LCM measurements or varying subsets with reduced temporal or spatial coverage. We evaluated the models based on a combination of cross-validation and external validation at locations of LCMs included in the models (N = 82), and also based on an independent external validation with a set of LCMs not used for the modeling (N = 30). We found that the model's performance declined substantially when LCM measurements were entirely excluded (spatiotemporal validation R^2 [RMSE] = 0.69 [1.2 μg/m^3]) compared to the model with all LCM measurements (0.84 [0.9 μg/m^3]). Temporally, using the farthest apart measurements (i.e., the first and last) from each LCM resulted in the closest model's performance (0.79 [1.0 μg/m^3]) compared to the model with all LCM data. The models with only the first or last measurement had decreased performance (0.77 [1.1 μg/m^3]). Spatially, the model's performance decreased linearly to 0.74 (1.1 μg/m^3) when only 10% of LCMs were included. Our analysis also showed that LCMs located in densely populated, road-proximate areas improved the model more than those placed in moderately populated, road-distant areas.

2. Zuidema C, Bi J, Burnham D, Carmona N, Gassett A, Slager DL, et al. 2024. Leveraging low-cost sensors to predict nitrogen dioxide for epidemiologic exposure assessment. J Expo Sci Environ Epidemiol 35:169–179, *https://doi.org/10.1038/s41370-024-00667-w*.

ABSTRACT. Background: Statistical models of air pollution enable intraurban characterization of pollutant concentrations, benefiting exposure assessment for environmental epidemiology. The new generation of low-cost sensors facilitates the deployment of dense monitoring networks and can potentially be used to improve intraurban models of air pollution.

Objective: Develop and evaluate a spatiotemporal model for nitrogen dioxide (NO_2) in the Puget Sound region of WA, USA, for the Adult Changes in Thought Air Pollution (ACT-AP) study, and assess the contribution of low-cost sensor data to the model's performance through cross-validation.

Methods: We developed a spatiotemporal NO_2 model for the study region, incorporating data from 11 agency locations, 364 supplementary monitoring locations, and 117 low-cost sensor locations for the 1996–2020 time period. Model features included long-term time trends and dimension-reduced land use regression. We evaluated the contribution of LCS network data by comparing models fit with and without sensor data using cross-validated (CV) summary performance statistics.

Results: The best performing model had one-time trend and geographic covariates summarized into three partial least squares components. The model, fit with LCS data, performed as well as other recent studies (agency cross-validation: CV-RMSE = 2.5 ppb NO_2; CV-R^2 = 0.85). Predictions of NO_2 concentrations developed with LCS were higher at residential locations compared to a model without LCS, especially in recent years. While LCS did not provide a strong performance gain at agency sites (CV-RMSE = 2.8 ppb NO_2; CV-R^2 = 0.82 without LCS), at residential locations, the improvement was substantial, with RMSE = 3.8 ppb NO_2 and R^2 = 0.08 (without LCS), compared to CV-RMSE = 2.8 ppb NO_2 and CV-R^2 = 0.51 (with LCS).

Significance: Our results suggest that low-cost sensors have the potential to improve models of air pollution exposure for environmental epidemiology by contributing observations in otherwise unmonitored locations.

Aim 2 (2 papers)

1. Cheng S, Blanco MN, Sheppard L, Shojaie A, Szpiro A. In review. Variable importance measure for spatial machine learning models with application to air pollution exposure prediction. Available on arXiv: *https://arxiv.org/abs/2406.01982*.

   ABSTRACT. Exposure assessment is fundamental to air pollution cohort studies. The objective is to predict air pollution exposures for study participants at locations without data to optimize our ability to learn about the health effects of air pollution. Researchers typically focus on generating the most accurate predictions possible to minimize exposure measurement error, but understanding the mechanism captured by the model that is fit to the data is also important. However, the latter may not be straightforward in all model settings due to the complex nature of machine learning methods as well as the lack of unifying notions of variable importance given the diversity of common models. This is further complicated in air pollution modeling by the presence of spatial correlation. We tackle these challenges in two datasets: sulfur (S) from regulatory United States national $PM_{2.5}$ subspecies data and ultrafine particles (UFPs) from a new Seattle-area traffic-related air pollution dataset. Our key contribution is a leave-one-out approach for variable importance that leads to interpretable and comparable measures for a broad class of models with separable mean and covariance components. We illustrate our variable importance measure with several spatial machine learning models, and such measure clearly highlights the difference in mechanisms captured by different models, even for those producing similar predictions. We leverage insights from this variable importance measure to assess the relative utilities of two exposure models for each of S and UFPs that have similar out-of-sample prediction accuracies but appear to draw on different types of spatial information to make predictions.

2. Cheng S, Blanco MN, Larson TV, Sheppard L, Szpiro A, Shojaie A. In press. Principal component analysis balancing prediction and approximation accuracy for spatial data. Available on arXiv: *https://arxiv.org/abs/2408.01662*.

   ABSTRACT. Dimension reduction is often the first step in statistical modeling or the prediction of multivariate spatial data. However, most existing dimension reduction techniques do not account for the spatial correlation between observations and do not consider the downstream modeling task when finding the lower-dimensional representation. We formalize the closeness of approximation to the original data and the utility of lower-dimensional scores for downstream modeling as two complementary, sometimes conflicting, metrics for dimension reduction. We illustrate how existing methodologies fall into this framework and propose a flexible dimension reduction algorithm that achieves the optimal trade-off. We derive a computationally simple form for our algorithm and illustrate its performance through simulation studies, as well as two applications in air pollution modeling and spatial transcriptomics.

Aim 3 (3 papers)

1. Blanco MN, Szpiro AA, Crane PK, Sheppard L. 2025. Ultrafine particles and late-life cognitive function: Influence of stationary mobile monitoring design on health inferences. Environ Pollut 10:374:126222; *https://doi.org/10.1016/j.envpol.2025.126222*.

   ABSTRACT. Growing evidence links ultrafine particles (UFPs) to neurotoxicity, but human studies remain limited. Various air pollution mobile monitoring approaches have been used to develop air pollution exposure models. However, whether design choices impact epidemiology, including for UFPs and cognitive function, remains unclear. We evaluated the adjusted association between 5-year average UFP number concentration (PNC) and late-life cognitive function (Cognitive Abilities Screening Instrument — Item Response Theory [CASI-IRT]) in the Adult Changes in Thought cohort (N = 5,283) by leveraging an extensive roadside mobile monitoring campaign specifically designed for epidemiology. To assess the impact of reduced monitoring approaches on this associ-

ation, we repeatedly subsampled UFP measures from the campaign, developed exposure models, and evaluated the degree to which associations were impacted. In the primary analyses, each 1,900 pt/cm³ increment in PNC was associated with an adjusted mean baseline CASI-IRT score that was 0.002 (95% CI: –0.016, 0.020) higher, which was not statistically significant. Point estimates were consistent across sampling designs with fewer visits per site (≤12), fewer seasons (1–3), and unbalanced visit frequency across sites. Sampling designs restricted to rush hours were more similar (median point estimate 0.002, IQR of point estimates: 0.000, 0.003) than business hour designs (0.006, IQR: 0.005, 0.007), but the opposite was true when temporal adjustments were applied (rush: –0.003, IQR: –0.005, –0.001; business: 0.002, IQR: 0.001, 0.004). We observed similar results in sensitivity and secondary analyses. We did not find evidence of an association between UFPs and cognitive function in fully adjusted models. Monitoring design had minimal impact on the inferential results in this setting, which may have been caused by the lack of association. Secondary analyses in a reduced model that is potentially confounded suggest that monitoring design might have a greater impact in other datasets. Further research is needed, particularly in contexts with robust, statistically significant health associations.

2. Blanco MN, Szpiro AA, Sheppard L. In preparation. Implications of exposure measurement error on inference about air pollution and cognitive function.

   ABSTRACT. Exposure measurement error (ME) is critical for environmental epidemiology but has received limited attention, particularly when also considering exposure assessment design. We investigate the impact of ME from air pollution exposure prediction modeling on inference in a case study of ultrafine particle (UFP) exposures and cognitive function. We leverage an extensive mobile monitoring campaign with repeated UFP measurements at temporary roadside locations, use measures to develop exposure prediction models, and evaluate the association between UFPs and late-life cognitive function (Cognitive Abilities Screening Instrument – Item Response Theory [CASI-IRT]) in the Adult Changes in Thought cohort. We apply nonparametric and parametric bootstrap approaches to decompose the ME and correct the estimated health inferences from: (a) classical-like (CL) error resulting from variability in the selected monitoring locations, sampling times, and cohort participants, and (b) Berkson-like (BL) error resulting from differences between the true and predicted exposure surfaces due to the sampling design and exposure modeling choices. CL and BL ME biased the health estimate by approximately 6% and increased the variability (SE) by 13%. We corrected the adjusted association between UFPs (per 1,900 pt/cm³) and CASI-IRT score from –0.0198 (SE: 0.0081) to –0.0212 (SE: 0.0092). In contrast, common air mobile monitoring sampling designs can produce health associations with median differences of up to ~75% different from reference exposure models, thus making exposure prediction ME a secondary consideration for addressing exposure ME relative to monitoring design.

3. Blanco MN, Doubleday A, Szpiro AA, Marshall JD, Crane PK, Sheppard L. In review. Influence of on-road mobile monitoring design on ultrafine particle exposure models and cognitive health inferences.

   ABSTRACT. On-road mobile monitoring is increasingly used to assess air pollution exposure, but the implications of various monitoring and analytic decisions on exposure and health inferences remain unclear. This study evaluated the impact of common on-road monitoring approaches in environmental epidemiology, focusing on ultrafine particle (UFP) exposures and late-life cognitive function. We used data from an on-road and roadside mobile monitoring campaign to develop UFP exposure models and assess associations with cognitive function, measured by the Cognitive Abilities Screening Instrument — Item Response Theory (CASI-IRT) score, in the Adult Changes in Thought cohort. UFP measures were subsampled using different strategies, including visit frequencies (4 vs. 12 visits per location), spatial balance, and sampling times. Temporal and plume adjustments were applied to develop exposure models, which were then used in health analyses. The robustness of these on-road monitoring approaches was evaluated by comparing the findings to those from a roadside reference campaign. Using the reference model, the adjusted mean baseline CASI-IRT score increased by 0.007 (95% CI: –0.013, 0.027) in confounder model 2 and decreased by 0.021 (95% CI: –0.039, –0.003) in the reduced model per 1,900 pt/cm³. Plume-adjusted, all-hours campaigns yielded models most consistent with these findings, particularly in reduced models where associations were stronger. Temporal and plume adjustments improved exposure model performance but did not meaningfully enhance health inferences, highlighting the importance of temporally balanced sampling. Campaigns with fewer visits per location produced more variable results. Monitoring and analytic decisions are crucial for on-road mobile monitoring studies aiming to support air pollution epidemiology. Temporally balanced sampling is essential for reliable exposure models and health inferences.

Aim 4 (2 papers)

1. Kim S-Y, Blanco MN, Bi J, Larson TV, Sheppard L. 2023. Exposure assessment for air pollution epidemiology: a scoping review of emerging monitoring platforms and designs. Environ Res 223:115451, *https://doi.org/10.1016/j.envres.2023.115451*.

   ABSTRACT. Background: Both exposure monitoring and exposure prediction have played key roles in assessing individual-level long-term exposure to air pollutants and

their associations with human health. While there have been notable advances in exposure prediction methods, improvements in monitoring designs are also necessary, particularly given new monitoring paradigms leveraging low-cost sensors and mobile platforms.

Objectives: We aim to provide a conceptual summary of novel monitoring designs for air pollution cohort studies that leverage new paradigms and technologies, to investigate their characteristics in real-world examples, and to offer practical guidance to future studies.

Methods: We propose a conceptual summary that focuses on two overarching types of monitoring designs, mobile and nonmobile, as well as their subtypes. We define mobile designs as monitoring from a moving platform, and nonmobile designs as stationary monitoring from permanent or temporary locations. We only consider nonmobile studies with cost-effective sampling devices. Then we discuss similarities and differences across previous studies with respect to spatial and temporal representation, data comparability between design classes, and the data leveraged for model development. Finally, we provide specific suggestions for future monitoring designs.

Results: Most mobile and nonmobile monitoring studies selected monitoring sites based on land use instead of residential locations, and deployed monitors over limited time periods. Some studies applied multiple designs and/or subdesign classes to the same area, time period, or instrumentation, to allow comparison. Even fewer studies leveraged monitoring data from different designs to improve exposure assessment by capitalizing on different strengths. To maximize the benefit of new monitoring technologies, future studies should adopt monitoring designs that prioritize residence-based site selection with comprehensive temporal coverage and leverage data from different designs for model development in the presence of good data compatibility.

Discussion: Our conceptual overview provides practical guidance on novel exposure assessment monitoring for epidemiological applications.

2. Kim S-Y, Gassett AJ, Blanco MN, Sheppard L. 2025. Ultrafine particle mobile monitoring study designs for epidemiology: cost and performance comparisons. Environ Health Perspect 133:47010, *https://doi.org/10.1289/ehp15100*.

ABSTRACT. Background: Given the difficulty of collecting air pollution measurements for individuals, researchers use mobile monitoring to develop accurate models that predict long-term average exposure to air pollution, allowing the investigation of its association with human health. Although recent mobile monitoring studies focused on predictive models' abilities to select optimal designs, cost is also an important feature.

Objectives: This study aimed to compare the costs to predictive model performance for different mobile monitoring designs.

Methods: We used data on ultrafine particle stationary roadside mobile monitoring and associated costs collected by the Adult Changes in Thought Air Pollution (ACT-AP) study. By assuming a single-instrument, local monitoring, and constant costs of equipment and investigator oversight, we focused on the incremental cost of staff workdays composed mostly of sampling drives and quality control procedures. The ACT-AP complete design included data collection from 309 sites, ~29 visits per site, during four seasons, every day of the week. We considered alternative designs by selecting subsets of fewer sites, visits, seasons, days of the week, and hours of the day. Then, we developed exposure prediction models from each alternative design and calculated cross-validation (CV) statistics using all observations from the complete design. Finally, we compared CV R^2s and the numbers of staff workdays from alternative designs to those from the complete design and demonstrated this exercise in a web application.

Results: For designs with fewer visits per site, the costs for the number of workdays were lower, and model performance (CV R^2) also worsened, but with a mild decline above 12 visits per site. The costs were also less for designs with fewer sites when considering at least 100 sites, although the reduction in performance was minimal. For temporally restricted designs that were constrained to have the same number of workdays and thus the same cost, restrictions on the number of seasons, days of week, and/or hours of the day adversely impacted model performance.

Discussion: Our study provides practical guidance to future mobile monitoring campaigns that have the ultimate goal of assessing the health effects of long-term air pollution. Temporally balanced designs with 12 visits per site are a cost-effective option that provides relatively good prediction accuracy with reduced costs.

Research Report 228, *Optimizing Air Pollution Exposure Assessment with Application to Cognitive Function*, by L. Sheppard et al.

INTRODUCTION

Outdoor air pollution is a major global public health concern. There is now broad expert consensus that exposure to ambient air pollution causes an array of adverse health effects, based on evidence from a large body of scientific literature that has grown exponentially since the mid-1990s (IARC 2016; US EPA 2016, 2019, 2022; WHO 2021).

The assessment of long-term exposure to ambient air pollution for epidemiological studies, however, remains challenging. Early cohort studies characterized exposure by assigning the average concentration measured at one or a few central sites within a city to each participant from the city (Dockery et al. 1993; Pope et al. 2002). Fixed-site networks — even those in North America and Western Europe — continue to have relatively limited spatial coverage in many areas, particularly in suburban and rural locations, and insufficient density to capture small-scale (within-city) variations of air pollution (Roque et al. 2025).

Recent developments in measurement technologies and modeling approaches have increasingly been used to estimate long-term air pollution exposure at finer spatial scales for epidemiological studies of large populations. Advances include novel air pollution sensors, mobile monitoring, satellite data, hybrid models, and machine-learning approaches (Hoek 2017). Even with those advances, important limitations and challenges remain when assessing long-term air pollution exposure, particularly for pollutants that vary widely across space and time.

In 2019, HEI issued Request for Applications 19-1, *Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies* (see Preface). Its goal was to develop and apply scalable novel approaches to improve assessments of long-term exposures to outdoor air pollutants that vary widely in space and time — such as ultrafine particles (UFPs*), black carbon (BC), and nitrogen dioxide (NO_2). Studies were intended to evaluate exposure measurement error quantitatively and determine how exposure assessment approaches might influence the health estimates.

Dr. Sheppard and colleagues proposed to advance the understanding of exposure assessment study design features, including a comparison of health estimates derived from those features. The HEI Research Committee recommended the study for funding because of the systematic evaluation of sampling designs to provide guidance to other researchers. They also appreciated the inclusion of UFPs and the application of exposure estimates to cognitive function, as it is an emerging health outcome.

This Commentary provides the HEI Improved Exposure Assessment Studies Review Panel's evaluation of the study. It is intended to aid the sponsors of HEI and the public by highlighting the study's strengths and limitations, and by placing the results presented in the Investigators' Report into a broader scientific and regulatory context.

SCIENTIFIC AND REGULATORY BACKGROUND

Traffic-related air pollution (TRAP) continues to be an important risk factor for poor health worldwide, with the highest exposures in urban settings and at residences near busy roadways (HEI 2022). TRAP is a complex mixture of gases and particles resulting from the use of motor vehicles. Motor vehicles emit various pollutants, including NO_2, BC, and UFPs (HEI 2022). Exposure assessment of those pollutants is challenging because they are characterized by high spatial and temporal variability.

Epidemiological studies have used different approaches to address those challenges. Researchers have increasingly used mobile monitoring in recent years by affixing monitoring instruments to vehicles and making measurements while systematically and repeatedly traveling a road network. Mobile monitoring strategies can involve mobile measurements made while driving predefined routes, or repeated short-term measurements made while in a vehicle parked at various roadside locations. Data collected through mobile monitoring have been used to develop land use regression models and other air pollution maps (e.g., Apte et al. 2024; Hatzopoulou et al. 2017; Kerckhoffs et al. 2016; Messier et al. 2018). Air pollution maps estimated from such monitoring are being increasingly applied in epidemiological studies (e.g., Alexeef et al. 2018; Downward et al. 2018).

Dr. Liane Sheppard's 3-year study, "Optimizing Exposure Assessment for Inference about Air Pollution Effects with Application to the Aging Brain," began in September 2020. Total expenditures were $800,000. The draft Investigators' Report from Sheppard and colleagues was received for review in January 2024. A revised report, received in September 2024, was accepted for publication in September 2024. During the review process, the HEI Improved Exposure Assessment Studies Review Panel and the investigators had the opportunity to exchange comments and clarify issues in the Investigators' Report and the Panel's Commentary. Review Committee member Sara D. Adar was not involved in the review of this report due to a conflict of interest.

* A list of abbreviations and other terms appears at the end of this volume.

In addition, low-cost sensors are increasingly being used in exposure assessment for health studies. They can be deployed on mobile platforms or can supplement fixed-site monitoring networks to develop exposure models, or they can enable simultaneous individual-level air pollution measurements to estimate personal exposure (e.g., Larkin and Hystad 2017, Morawska et al. 2018).

Exposure models are applied in epidemiological studies that underpin the air quality standards and guidelines. Governments in the United States and Europe have recently moved toward more stringent fine particulate matter ($PM_{2.5}$) annual standards — 9 and 10 µg/m³, respectively — which align more closely with the 2021 WHO Air Quality Guidelines of 5 µg/m³. A more stringent annual standard was also set in Europe for NO_2 (**Commentary Table 1**).

There are no specific ambient air quality standards or guidelines for UFPs and BC, and regulatory agencies do not commonly measure them. Hence, international or national standard methods to characterize them have not been established (HEI 2010; HEI Review Panel on Ultrafine Particles 2013). Although no air quality guidelines have been developed for UFPs and BC, the WHO has provided "good practice statements" for these pollutants geared toward additional monitoring, mitigation, and epidemiological research (WHO 2021).

As noted earlier, important limitations and challenges remain when predicting long-term air pollution exposure to pollutants that vary highly in space and time. The current study compared the performance of different exposure assessment study design features on long-term exposure estimates of UFPs, NO_2, and $PM_{2.5}$ in Seattle, Washington.

STUDY OBJECTIVES

The overarching aim of Dr. Sheppard's study was to advance the understanding of exposure assessment study design and analysis features for air pollution and health studies. The investigators specified the following four study aims:

1. Identify key design choices to improve long-term average exposure predictions using mobile monitoring campaigns and fixed-site networks of low-cost sensors
2. Develop annual average TRAP exposure predictions from mobile monitoring data using advanced statistical methods
3. Determine the impact of sampling designs and analytical approaches on the health estimates
4. Address the overall value of incorporating novel exposure data collection and modeling by comparing the logistical features (cost and time) of using different sampling designs and analysis choices

SUMMARY OF APPROACH AND METHODS

Dr. Sheppard and colleagues compared the performance of different exposure assessment study design features on long-term exposure estimates in Seattle, Washington. In a cohort study, the investigators evaluated how various approaches to air pollution sampling affected exposure prediction and health estimates. Most analyses focused on UFP data from a previously conducted mobile roadside monitoring campaign in 2019–2020. In that campaign, short-term measurements were made from a parked vehicle at 309 roadside locations, with about 30 visits made to each site. For the present study, the investigators also conducted analyses of UFPs using mobile on-road monitoring data for a total of 5,878 road segments. Each 100-meter segment was visited an average of 28 times. Further analyses were conducted on $PM_{2.5}$ and NO_2 concentrations collected using low-cost sensors at about 115 fixed monitoring sites in 2017–2020, combined with regulatory monitoring data from a much longer time period (**Commentary Table 2**).

The investigators used either the full dataset or subsets of measurements to develop annual average exposure estimates using a suite of models, including universal kriging, spatiotemporal models, machine learning, and other advanced statistical models (Commentary Figure 1). The study leveraged

Commentary Table 1. Annual NO_2 and $PM_{2.5}$ Standards in the US, EU, and WHO

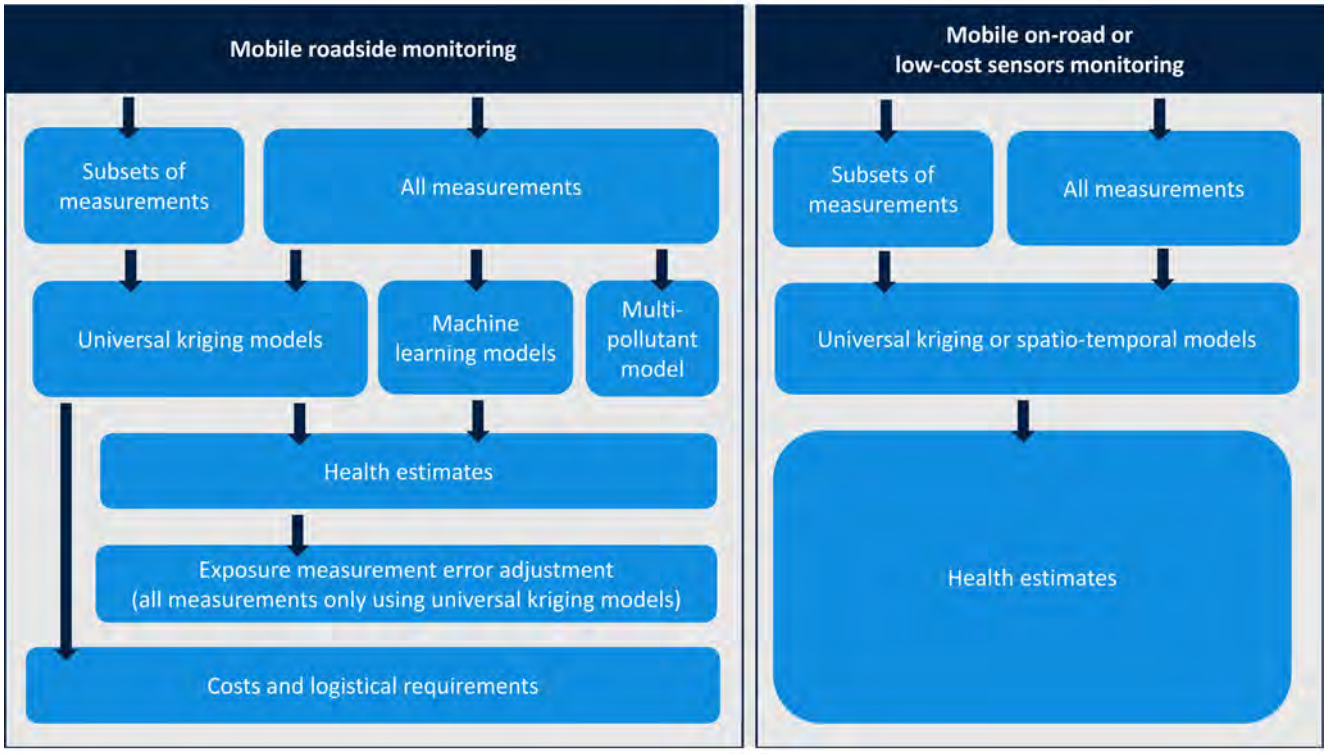
Organization	Annual $PM_{2.5}$ (µg/m³)	Annual NO_2 (µg/m³)	Notes
US EPA (2024)	9	100	NAAQS
US EPA (Previous)	12	100	Previous NAAQS
EU (2024)	10	20	Limit value for 2030
EU (Previous)	25	40	Previous limit value
WHO (2021)	5	10	Air Quality Guidelines
WHO (Previous)	10	40	Previous Air Quality Guidelines

NAAQS = National Ambient Air Quality Standards.

Commentary Table 2. Summary of Data and Analyses Conducted

Monitoring Data	Main Pollutants (UFP Device[a])	Exposure Model	Type of Analysis	Aim	Report Chapter
Mobile roadside	UFPs (NanoScan)	Universal kriging	Exposure model performance, health estimates	1, 3	4
Mobile roadside	UFPs (NanoScan)	Universal kriging	Health estimates corrected for exposure measurement error	3	5
Mobile roadside	UFPs (P-Trak)	Universal kriging	Costs and logistical requirements	4	9
Mobile roadside	UFPs (P-Trak)	Universal kriging and machine learning	Exposure model performance, health estimates	2, 3	8
Mobile roadside	UFPs (NanoScan, P-Trak, and DiSCmini), BC, NO_2, CO_2, $PM_{2.5}$	Universal kriging, machine learning, and multipollutant	Model performance	2	8
Mobile on-road	UFPs (P-Trak)	Universal kriging	Exposure model performance, health estimates	1, 3	6
Low-cost sensors	$PM_{2.5}$ and NO_2	Spatiotemporal	Exposure model performance, health estimates	1, 3	7

[a] The UFP instruments are TSI NanoScan 3910, TSI P-Trak 8525, and Testo DiSCmini.



Commentary Figure 1. Schematic overview of the study design.

detailed air pollution and cognitive function data available at baseline (1994 or later) from the Adult Changes in Thought (ACT) study in Seattle, which is a cohort study of about 5,400 individuals 65 years of age or older (Box 1).

AIR POLLUTION MONITORING

Mobile Monitoring Campaign of UFPs

UFPs were measured with three different instruments: a P-Trak, a NanoScan, and a DiSCmini. All instruments measure particle number concentration but differ in the measurement principle and size ranges captured. Different instruments were used for the analyses across the chapters (Commentary Table 2).

The investigators leveraged real-time mobile monitoring data of UFPs that were collected along nine predefined routes during a year-long campaign from March 2019 to March 2020 using a hybrid vehicle. A total of 309 sites off the side of the road were selected for short-term monitoring along the routes. These sites were intended to represent cohort participants' residential locations, which were on average 611 meters away from their closest roadside measurement site.

Air pollution measurements at the roadside mobile monitoring sites were made for 2 minutes while the vehicle was parked (visits). Visits were repeated on average 30 times throughout the monitoring campaign, and thus, about an hour of data was collected at each site. To obtain a representative annual average, the monitoring was temporally balanced to ensure that all sites were visited during all four seasons and all days of the week, including weekends, and at both early morning and late evening hours (from 5 a.m. to 11 p.m.).

The median of the real-time data was calculated for each visit. Then the investigators transformed the data to reduce the influence of possible extreme values by winsorizing. They winsorized the data across all visits for each site, which replaces values at the tails of the distribution with a fixed percentile value. Values below the 5th and above the 95th percentile were replaced with the values at those thresholds. The winsorized data were then averaged over all visits, log-transformed, and used for subsequent exposure modeling. The mobile roadside monitoring data were used in most of the analyses in the report (Chapters 4, 5, 8, and 9) to investigate their influence on exposure, measurement error, costs, and health estimates.

The investigators also conducted on-road measurements between roadside sites along the nine predefined routes, totaling 1,069 km in the 1,200 km^2 study area. The on-road measurements were aggregated to 100-meter road segments for a total of 5,878 road segments. Each segment was visited an average of 28 times. The investigators excluded some road segments to represent long-term residential exposure better, such as interstates and highways with restricted access, and road segments with fewer than 23 repeat visits.

Box 1: What Is the ACT Study?

The ACT study is a population-based cohort of older cognitively unimpaired adults (65 years or older) in the greater Seattle area. The goal of ACT is to study factors that affect brain aging. Recruitment for the study began in 1994, with a second recruitment wave in 2000, followed by continuous enrollment starting in 2005. Participants are randomly selected from Seattle area members of Group Health Cooperative (now Kaiser Permanente), which is a health maintenance organization. Participants are evaluated every 2 years, where a wealth of data is collected about participants' medical care, physical health, and cognitive health, including their memory and thinking abilities. For the current study, 5,409 participants were included with available information on cognitive function at baseline.

Cognitive function was measured with the Cognitive Abilities Screening Instrument (CASI), which is a 40-item cognitive test that assesses a broad range of cognitive domains (e.g., attention, memory) using scores ranging from 0 to 100. CASI scores were transformed using Item Response Theory (IRT), allowing for a more nuanced interpretation of cognitive function compared to a simple raw total score on the CASI test. For the air pollution analysis, the scores were normalized with values less than 0 indicating lower cognitive function and scores above 0 indicating greater cognitive function than average.

From 2017 to 2020, the ACT study added air pollution measurements from a mobile monitoring campaign and a low-cost fixed-site monitoring campaign specifically designed to represent outdoor exposures for the ACT cohort, which was used in the current study (ACT Air Pollution study).

For more information on the ACT study, go to *https://www.actagingresearch.org/index.php*.

The median of the real-time data was calculated for each 100-meter segment (equivalent to about 10 seconds of observations per visit). Similar to the mobile roadside data, the on-road data were winsorized at the segment level before averaging, and then log-transformed after averaging over all sampling days. On-road UFP data were further adjusted in some analyses to reduce the influence of localized on-road plumes to better represent concentrations at residential locations. Principal components analysis was used to partition simultaneous on-road measurements of multiple pollutants and identify those pollutants associated with localized plumes. An absolute principal component scores model was then used to adjust the UFP data in an iterative process that also leveraged the mobile roadside measurements, as described in detail in Doubleday and colleagues (2023). The mobile on-road monitoring data were used in Chapter 6 of the report to investigate their influence on exposure and health estimates.

Low-Cost Sensor Data for $PM_{2.5}$ and NO_2

The investigators leveraged outdoor $PM_{2.5}$ and NO_2 data from a low-cost sensor campaign conducted in 2017–2020 at about 115 fixed sites, most of them at ACT study participants' residences. $PM_{2.5}$ and NO_2 were measured with a Plantower sensor (PMS A003) and an Alphasense sensor (B43F), respectively. Detailed quality control was conducted. The low-cost sensor data were calibrated daily, using regression models from co-located regulatory measurements, which are described in detail elsewhere (Zuidema et al. 2021; Zusman et al. 2020).

Low-cost sensor data were used in models (see below) that combined them with data from regulatory monitors and other research-grade monitors from many additional locations from 1978 (1996 for NO_2) to 2020 to explore the added value of low-cost sensor data and its design features. In total, $PM_{2.5}$ and NO_2 data from those other data sources were available at 80 and 375 additional locations, respectively. In particular, a snapshot campaign of two or three 2-week measurements at 110 roadside locations conducted in 2018–2019 added substantial spatial information to the models for NO_2.

The investigators averaged all data into 2-week averages. The data were highly unbalanced, with some locations providing longer time series of data, and with only one or a few 2-week observations in other locations. The low-cost sensor data were used in Chapter 7 of the report to investigate their influence on exposure and health estimates.

AIR POLLUTION MODELING

Universal Kriging and Spatiotemporal Models

The investigators subsampled mobile monitoring data to evaluate various exposure assessment study design features, such as fewer visits per site, fewer days of the week, restricted hours of the day, and fewer seasons. The investigators evaluated the impact of those reduced features for both roadside and on-road mobile data. In addition, for the on-road data, they also tested a temporal adjustment for unbalanced diurnal sampling using a fixed-site background monitor and employed a plume correction method to reduce the influence of localized on-road plumes. Each alternative design was resampled 30 times, totaling 480 different design options that were tested for the primary analyses using mobile roadside data.

The investigators used either the full dataset or subsets of measurements to develop annual average exposure estimates using universal kriging models with dimension reduction using partial least squares (for brevity: universal kriging models). Separate models were developed for roadside and on-road mobile data. A suite of geographical covariates (about 200 variables) was available for inclusion in the models, including indicators of land use, roadway proximity, and population density.

Similarly, the investigators subsampled low-cost sensor data to evaluate designs with varying sampling durations, repeated measurements across periods, and sampled locations. Using the subsets of data, they developed up to 20 different spatiotemporal $PM_{2.5}$ models that were designed to accommodate highly imbalanced data. Note that all spatiotemporal models included all data from the regulatory monitors and other research-grade monitors.

They assessed the performance of each model using cross-validation (mobile monitoring data) or a combination of cross-validation and external validation (low-cost sensor data). They reported several measures of performance, including the root mean squared error (RMSE) and the MSE-based explained variance (MSE-R^2). The latter was used instead of the more common regression-based R^2 because it evaluates whether predictions and observations are the same (i.e., are near the one-to-one line), rather than merely correlated. As such, MSE-R^2 assesses both bias and random variation around the one-to-one line. Because of resampling each alternative design, the median of the performance measures and the distribution were reported in the analyses using mobile monitoring data. The models using all data from the mobile roadside campaign or all the low-cost sensor data were taken as reference models.

Machine Learning Models and Other Advanced Statistical Methods

Leveraging all data from the mobile roadside campaign — thus no subsets — and using all pollutants measured (UFPs, BC, $PM_{2.5}$, NO_2, and CO_2), the investigators explored the use of advanced statistical methods beyond the universal kriging model. They developed several machine learning models (six per pollutant), including spatial random forest, that allowed nonlinear and highly complex relationships between geographical predictor variables to be incorporated. They compared the performance of the different models to the earlier-mentioned universal kriging model, which was used as the reference model.

Moreover, because machine learning models are often considered a "black box," they developed a new "variable importance metric" in this study to aid in selecting and interpreting the models. They demonstrated the utility of this metric with the spatial random forest model.

All models developed up to this point were single-pollutant models. In one section of the report, the investigators describe the development and application of a spatial multipollutant model of multiple properties of UFPs (from all three instruments), BC, and NO_2. The model uses a new principal components-based dimension reduction algorithm that considers the spatial patterns in the data while optimizing extrapolation to locations without measurements. The report presents maps of the first three principal component scores for the Seattle area, which highlight the airport and major roads as multipollutant sources. The use of this approach for health estimates analyses of mixtures is identified as future work and is not further summarized in the Commentary.

Those advanced methods were used in one chapter of the report (Chapter 8).

HEALTH ESTIMATES

Each model, except for the multipollutant model, was used to predict the 5-year average UFPs or other pollutant exposures prior to the cognitive function measurement that was obtained at baseline (1994 or later). This exposure was assigned at the residential address level and accounted for residential mobility over the 5 years prior to baseline. The investigators then assessed the association between the 5-year average exposure from each exposure model and baseline cognitive function using standard linear regression. Each model was adjusted for participant age, sex, education, and calendar year (confounder model 1), and associations were expressed per interquartile range of exposure. Results presented in Chapters 4 (mobile roadside data for UFPs) and 6 (mobile on-road data for UFPs) were also adjusted for participant race and socioeconomic status (confounder model 2).

The adjusted associations using the all data exposure model from the mobile roadside campaign for UFPs or from the low-cost sensor data for $PM_{2.5}$ and NO_2 were taken as the reference models. To add context, the associations for cognitive function from the reference models were also expressed as the equivalent of cognitive decline due to aging for a certain number of months.

EXPOSURE MEASUREMENT ERROR ADJUSTMENT

The universal kriging model using all data from the mobile roadside campaign was taken as the reference model in many of the analyses, assuming that those estimates were the best representation of the true annual average. However, even this model contains exposure measurement errors that are, for example, related to the number of locations and times sampled or related to the level of smoothing when fitting the data. Hence, the investigators applied their previously developed bootstrap methods (e.g., Bergen et al. 2016; Szpiro et al. 2011a, 2011b) to quantify the exposure measurement error in the reference model for UFPs, and to correct the health effect estimates accordingly. A key assumption in the measurement error approach — spatial "compatibility" of monitoring sites and cohort locations — was met because the mobile monitoring campaign was specifically designed to capture exposures for the ACT cohort.

Note that the exposure measurement error was not quantified for the UFP models using subsets of mobile monitoring data or for $PM_{2.5}$ and NO_2 models using low-cost sensors.

COSTS AND LOGISTICAL REQUIREMENTS

Using similar exposure design features and universal kriging models as described earlier, the investigators explored the trade-offs between exposure model performance and logistical features (both cost and time) to identify optimal monitoring designs. The investigators used the data from the mobile roadside campaign for those comparisons, with the UFP comparisons only considering the P-Track instrument data. They conducted similar cost–performance analyses using $PM_{2.5}$ and NO_2 low-cost sensor data, but this analysis was described mostly in the Additional Materials and hence not summarized in this Commentary.

They included both up-front and per-drive day costs informed by their own monitoring experience and expenditures for the ACT Air Pollution study. Examples of up-front costs include the purchase of the P-Trak instrument, software development, and various preparation efforts. Most of these costs did not vary by monitoring design. The per-drive-day cost included the cost of staff time for driving, vehicle use, and data management. They expressed the cost as the number of working days, assuming fixed staff costs. They also explored the addition of multiple instruments and pollutants, as well as the addition of a premium for staff working evening and weekend hours. Note that the monetary analysis did not consider costs related to data management and analysis related to exposure model development, which is commonly performed after the monitoring campaign has been completed.

SUMMARY OF RESULTS

MOBILE MONITORING DATA

Performance of Exposure Models Using All Data

The universal kriging reference model using all UFP data from the mobile roadside campaign had a cross-validated MSE-R^2 value of 0.65 (NanoScan) and 0.77 (P-Trak). The cross-validated MSE-R^2 values were 0.65, 0.76, and 0.77 for BC, $PM_{2.5}$, and NO_2, respectively.

The universal kriging model performed slightly better — up to an increase of 0.10 in MSE-R^2 — than the various machine learning models for all pollutants except for BC, where similar performances were observed.

Performance of Exposure Models Using Subsets of Mobile Monitoring Data

The universal kriging models with restricted mobile roadside sampling of UFPs almost always produced lower-performing exposure models compared to the reference model. Models based on datasets with sampling restricted to only one season, sampling conducted only during business hours, and those with few visits to high variability sites had the lowest performance (MSE-R^2 of 0.43–0.48) (**Commentary Table 3**).

Predictions from most designs were highly correlated with predictions from the all data campaign (median Pearson correlations [R] > 0.85), although predictions from the business hour design were consistently less correlated than those from all other designs (R = 0.77).

Commentary Table 3. Performance of Various Exposure Assessment Study Design Features Using Mobile Roadside UFP Data and Its Impact on the Estimated Association Between UFPs and Cognitive Function Using Confounder Model 1

Design Choice	Performance of Exposure Models[a]	% Attenuation of the Association[b]
All-Data Reference Model	*0.65*	*Reference Association: –0.020 per 1,900 Particles/cm³*
Fewer visits		
12	0.59	5%
6	0.53	5%
4	0.51	10%
Fewer seasons for 12 visits in total		
4	0.61	0%
3	0.58	5%
2	0.58	15%
1	0.43	15%
Fewer hours		
Business hours	0.45	60%
Rush hours	0.56	40%
Spatial balance		
Balanced	0.61	0%
Low number of sites with high variability	0.48	10%
High number of sites with high variability	0.59	10%

[a] Median of cross-validated MSE-R^2 from universal kriging models using UFPs (NanoScan) from the mobile roadside campaign.

[b] % attenuation based on the median values, with associations adjusted for participant age, sex, education, and calendar year (confounder model 1).

Using mobile on-road UFP data, the investigators found that most comparisons identified the same design features or elements as important, although with a few notable differences. Spatial balance had a minimal impact on exposure model performances in the on-road models. Strikingly, on-road modeling results were generally similar when road segments were tested 4 times versus 12 times, although the latter produced more stable results.

Comparison of Health Estimates Using Different Exposure Estimates

Exposure of UFPs estimated using the reference model was negatively (adversely) associated with cognitive function at baseline when adjusted for participant age, sex, education, and calendar year (confounder model 1). The UFP association was –0.020 (95% CI: –0.036 to –0.004) per increase of 1,900 particles/cm³. This estimate is equivalent to accelerated aging of 7.5 months (on average) for cohort participants.

The reduced-sampling designs led to similar findings in terms of negative (adverse) associations between UFP exposure and cognitive function at baseline. However, the strength (magnitude) of the observed negative associations sometimes differed substantially, especially for the business and rush hours designs, which attenuated associations by up to 60% (Commentary Table 3).

Notably, the observed negative association in the reference model disappeared in health models that also adjusted results for race and socioeconomic status (confounder model 2). The null findings from the reference model using confounder model 2 hampered the assessment of the influence of sampling design on health estimates using different exposure estimates for UFPs. Hence, for the aims of their project, the investigators decided to focus on the findings from confounder model 1 throughout the report.

Exposure Measurement Error Adjustment

The observed negative associations using confounder model 1 were affected more by features of the mobile monitoring design than by accounting for exposure measurement error in the reference exposure model. The investigators reported only a modest influence on the observed negative associations when results were adjusted for exposure measurement error

in the reference model (6% on the association, 13% on the confidence intervals).

Costs and Logistical Requirements

The investigators found that a mobile monitoring study with roadside sampling of UFPs with at least 12 visits per location optimized exposure model performance while also limiting costs. Furthermore, the investigators noted that it is important that the mobile monitoring campaign covers all days of the week, most hours of the day (including early morning and late evening hours), and at least two seasons.

LOW-COST SENSOR DATA

The addition of low-cost sensor data improved $PM_{2.5}$ exposure modeling. The spatiotemporal model using all data had a cross-validated MSE-R^2 value of 0.84, whereas the model with only the regulatory monitors and other research-grade monitors had a value of 0.69. Furthermore, increasing the number of low-cost sensor locations and repeated measurements resulted in better exposure model performance.

In contrast, in most comparisons, the addition of low-cost sensor data improved the NO_2 estimates only slightly in models that combined data from regulatory monitors and other research-grade monitors. For example, the cross-validated MSE-R^2 value increased from 0.82 to 0.85 in a spatial comparison. Reasons may relate to the large amount of spatial information already in the model from many additional locations (375 in total), using other data sources that were not available for $PM_{2.5.}$

Largely null findings were reported between $PM_{2.5}$, NO_2, and cognitive function for the various spatiotemporal exposure models with and without low-cost sensor data using confounder model 1. The null results thus hampered the assessment of the influence of adding low-cost sensor data for health effect estimates.

HEI IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL'S EVALUATION

In its independent review, the HEI Review Panel thought the study was well-motivated and appreciated that it leveraged detailed air pollution and cognitive function data from the ACT study in Seattle. They thought the study was comprehensive, with thorough analyses and findings that will be of broad interest and value to a wide audience.

STRENGTHS OF THE STUDY

The Panel noted several strengths of the research. First, the Panel recognizes the benefits of an extensive year-long mobile monitoring campaign that includes both roadside and on-road sampling. The investigators leveraged a rich dataset on UFPs and other pollutants that covered various times of day between 5 a.m. and 11 p.m., weekdays, and weekends — thus including those times of day when people might be more likely to be at home — and all four seasons. Many other mobile monitoring campaigns have collected less data, sampled during more restricted periods, such as business hours only, or had short monitoring durations lasting only a few weeks or months. Some of those studies have used continuous measurements at a fixed reference site to account for temporal variation (Kim et al. 2023).

Second, using this detailed dataset, the investigators evaluated various exposure assessment study design features and provided practical guidance on future mobile monitoring campaigns. Developing this guidance addressed a clear research gap and should be of interest to a wide audience. The consideration of the study design costs in developing the guidance, as informed by their experience, was also appreciated.

Third, the extensive air pollution exposure modeling and rigorous evaluation of their performance are strengths of the study. The investigators developed a suite of models, including universal kriging, spatiotemporal models, machine learning, and other advanced statistical models. The large number of geographical covariates (about 200 variables) available for inclusion in the models was notable. The development of a novel variable importance metric that is applicable to machine learning methods such as spatial random forest may be an important contribution. Moreover, the investigators reported several measures of performance to test accuracy and possible bias — thereby providing an in-depth performance assessment.

Fourth, the Panel found that the analysis to adjust for exposure measurement error in the health analyses was a valuable contribution. Accounting for the inherent (spatially varying) uncertainty and biases in modeled estimates of air pollution remains largely an unresolved problem in air pollution epidemiology (Samoli and Butland 2017; Sheppard et al. 2012), and advances in this area are much needed.

Fifth, the Panel was impressed by the large number of publications resulting from the work, as nicely documented in the report's annotated bibliography.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results, as explained next.

FOCUS ON UFPs AND USE OF DIFFERENT INSTRUMENTS

Although the HEI-funded work encompassed analyses of multiple pollutants, most of the report is focused on UFPs from mobile monitoring or $PM_{2.5}$ and NO_2 from low-cost monitoring. Few comparisons across pollutants are included in the main report, although more information is presented in the Additional Materials and other publications. The Panel thought this limited the generalizability of the findings and that additional research is warranted for the other pollutants.

For the evaluation of exposure design features using mobile roadside and mobile on-road monitoring data of UFPs, findings were presented in different stand-alone chapters,

and different instruments were selected (a NanoScan and a P-Trak) for various analyses. This difference makes a direct comparison between the two monitoring approaches difficult.

Those instruments differ in the size range captured: 10–420 nm for the NanoScan and 20–1,000 nm for the P-Trak. The lower cut-off measurement is usually critical because most UFPs are smaller than 20 nm and not captured by the P-Trak (HEI Review Panel on Ultrafine Particles 2013). Even small differences in the lower cut point in the range below 20 nm can lead to substantial differences in the particle number concentration. The very small particles (< 20 nm) are also the particles that might exhibit the highest variability in space and time (HEI Review Panel on Ultrafine Particles 2013). Indeed, in the current study, the NanoScan measured concentrations that were roughly 50% higher than those of the P-Trak, with more variability (contrast).

The investigators did not report detailed information on particle size distributions, preventing an in-depth particle size distribution analysis. However, the investigators conducted sensitivity analyses using P-Trak for the mobile roadside analyses and reported similar results, alleviating the concern to some extent. Note that they could not conduct a similar sensitivity analysis using the NanoScan for the mobile on-road analyses because of the time resolution (60 seconds for NanoScan versus 1 second for the P-Trak). The investigators also added a qualitative comparison section of the two monitoring approaches in the synthesis chapter, based on the Panel's recommendation, which was appreciated.

REMOVING THE INFLUENCE OF POSSIBLE EXTREME VALUES

The investigators calculated medians for each visit (or segment), winsorized across all visits for each site, and then log-transformed the data for subsequent exposure modeling. All three approaches are meant to reduce the influence of possible extreme values and increase evenness in the data. A sensitivity analysis to investigate how each approach — in particular winsorizing the mobile monitoring data — would affect the exposure models was missing from the report. A more thorough examination of the influence of possible extreme values was, however, included in an earlier paper (Blanco et al. 2022).

The Panel concluded that winsorizing the data improved exposure model performance for some pollutants (e.g., BC, NO_2, $PM_{2.5}$) but not consistently for UFPs, which was the focus of most of the report. For example, in Blanco and colleagues (2022), the "median of medians" approach had improved out-of-sample model performance in terms of lower RMSE and near identical MSE-R^2 values compared to winsorizing, particularly for the NanoScan and P-Trak instruments used in Chapters 4 and 6. The investigators justified their approach to also winsorize UFPs for "consistency across pollutants," but as it turned out, only limited comparisons across pollutants were included in the report.

THE HEALTH ANALYSES WERE CONSIDERED LIMITED

The Panel members thought the study's main strengths lie in its contributions to methodological development regarding improved exposure assessment design rather than the evaluation of health estimates. Although the Panel appreciated the complexities involved, the health analyses were considered limited for three reasons.

First, most exposure models used were based on measurements conducted up to 25 years after the health outcome. The Panel thought the investigators could have used health outcome data from later years to better align with the 2019–2020 exposure models from the ACT study, which is an ongoing cohort with participants who are followed up every 2 years. This temporal mismatch between the period captured by the mobile measurements and the exposure window most relevant for epidemiological purposes is also apparent in some other cohort studies investigating UFPs (e.g., Alexeef et al. 2018; Bai et al. 2019; Downward et al. 2018; Pond et al. 2022; Weichenthal et al. 2017, 2024). However, the temporal mismatch in other studies is typically shorter (e.g., up to 10 years after the end of follow-up), and some of those studies (e.g., Weichenthal et al. 2024) applied a back-casting procedure using supplemental data to partially overcome the lack of air pollution data in earlier years. The investigators did not apply a back-casting procedure, although one was originally proposed, and assumed that air pollution exposure surfaces remained constant over all that time, which is a large assumption that they did not test in the report. The Panel thought this assumption is difficult to defend because air pollution concentrations have been declining over the past few decades in many high-income countries, due largely to successful air quality regulation and subsequent reductions in emissions from major air pollution sources, including transportation and power generation (Boogaard et al. 2024; US EPA 2016, 2022; WHO 2021).

Second, the health analysis was a cross-sectional analysis of one measure of cognitive function. In the original application, the investigators planned to use longitudinal data on cognitive decline and dementia incidence from the ACT cohort. Because of delays in accessing the health data and because the models they developed were more computationally intensive than expected, the investigators decided to conduct the current health analysis within the scope of the current project. The Panel would welcome the more advanced health analyses that the investigators are planning, as alluded to in the report.

Third, the Panel was concerned that residual confounding was likely in the analyses (confounder model 1) due to inadequate adjustment for characteristics that are correlated with air pollution and cognitive function, most notably socioeconomic status. Findings differed for models that adjusted for race and socioeconomic status (confounder model 2) compared to those that did not (confounder model 1), as documented in Chapters 4 and 6 of the report. The Panel thought the authors should have adjusted for socioeconomic status in all health analyses. The Panel also noted the lack of

information on potential individual lifestyle covariates, such as smoking.

For those reasons, the investigators avoided the use of causal inference language in the report, as supported by the Panel. Nevertheless, the Panel recommended caution when interpreting the findings of the health analyses.

USE OF REAL-WORLD DATA VERSUS SIMULATIONS

The Panel thought simulations would have complemented the study because of the limitations in the health analyses using real-world data. In particular, the null findings from the reference exposure model using confounder model 2 hampered the assessment of the influence of sampling design on health estimates using different exposure estimates. An advantage of using simulated data is that the underlying "true" health effects are known in that scenario, and one can systematically test one feature while holding all other conditions constant. The challenge with simulation studies is that they might not adequately represent the real world. Some carefully designed simulations, along the lines discussed in the report, could have shed light on the differing health results between confounder models. However, this would have increased the scope of the project beyond what the investigators originally proposed.

GENERALIZABILITY OF GUIDANCE ON MOBILE MONITORING CAMPAIGNS

The investigators provided practical guidance for the implementation of future mobile monitoring campaigns. However, the Panel had some concerns about the generalizability of the findings related to the improved exposure assessment design. The air pollution exposure estimates in the analyses were relatively low, and the variability (contrast) was limited. The average UFP concentrations were low (10,000 particles/cm^3 measured with the NanoScan), typical of urban background areas in North America, and lower than typical near-roadway locations. Concentrations of the other pollutants were similarly low; for example, the average $PM_{2.5}$ concentration measured using low-cost sensors was 6 $\mu g/m^3$, and the interquartile range was 1 $\mu g/m^3$. Also, Seattle has a temperate climate characterized by moderate temperatures with mild winters and dry summers with little extreme heat or cold.

The performance of the different reduced-sampling models was compared against the reference model, which included all roadside monitoring data. Results and recommendations regarding adequate or optimal numbers of sites and visits might be specific to the Seattle area and the time period of the study. Relatedly, it is important to mention that the mobile monitoring campaign was specifically designed to capture exposures for the ACT cohort (in other words, the monitoring sites and cohort locations were spatially "compatible"). In other studies, residential locations of interest might not be known in advance, so monitoring routes might need to be selected based on different criteria. Earlier campaigns have often selected monitoring locations based on maximizing air pollution variation by including different geographic features, land uses, or different sources of air pollution (Kim et al. 2023).

Thus, some caution is warranted in generalizing the findings, and further research in other cities would be helpful to assess the generalizability of the specific findings related to exposure assessment design. Note that the findings were not affected by the COVID-19 pandemic in 2020 because monitoring was completed before that time.

COMPARISON OF MOBILE MONITORING GUIDANCE WITH OTHER STUDIES

Regarding the guidance on future mobile monitoring campaigns, the number of repeated measurements (12 visits) per location that the investigators considered optimal aligns with an intensive mobile on-road monitoring study using Google Street View cars in Oakland, CA (Apte et al. 2017). Apte and colleagues (2017) documented that up to about 10 repeated driving days, the stability of the measured average concentration of BC, NO, and NO_2 increased, and after about 20 driving days, the stability of the average did not improve appreciably with further repeats (Apte et al. 2017). Note that UFPs were not measured in the Oakland study.

The investigators further emphasized the importance of including sampling beyond business hours, including extended times of the day and weekends. However, a recent analysis, which is part of the HEI report by Gerard Hoek and colleagues funded under the same RFA as the current study, did not identify the time of day as an important feature that would explain some of the heterogeneity of effect estimates observed (Hoek et al. 2025). Among many other exposure assessment approaches, Hoek and colleagues compared three mobile monitoring studies (Kerckhoffs et al. 2016, 2017, 2021) that excluded rush hours with a study using Google Street View cars that monitored from 8 a.m. to 10 p.m. on weekdays in the Netherlands (Kerckhoffs et al. 2022). However, they did not have access to a mobile monitoring study that covered 24 hours of the day and weekend days, which may partly explain the difference in findings from these two studies.

Furthermore, Sheppard and colleagues noted that it would be important that the exposure sampling in mobile monitoring campaigns include at least two seasons. Although the Panel generally agreed with the seasonal sampling requirement, the importance of season might be characterized by patterns in temperature, sunlight, humidity, precipitation, and wind (Pérez et al. 2020). Hence, the Panel emphasized that "seasons" depend on geographical location, and how many "seasons" exist varies by region.

SUMMARY AND CONCLUSION

The study by Dr. Sheppard and colleagues compared the performance of different exposure assessment study design features on long-term exposure estimates in Seattle, Washington. The investigators determined the impact on

exposure prediction and health effect estimates using various approaches to sample data collected from an earlier mobile monitoring campaign and a fixed-site monitoring campaign with low-cost sensors. The investigators used either the full dataset or subsets of measurements to develop annual average universal kriging models, machine learning models, and other advanced statistical models. The study leveraged detailed air pollution data and cognitive function data at baseline (1994) from the Adult Changes in Thought (ACT) Air Pollution study in Seattle, a cohort study of older adults.

The study provides practical guidance on future mobile monitoring campaigns, which addresses a clear research gap. The extensive year-long mobile monitoring campaign and the evaluation of various exposure assessment study design features were strengths of the study. Another strength was the extensive air pollution exposure modeling and rigorous evaluation of their performance. The Panel was also impressed by the large number of publications resulting from the work.

The investigators found that a mobile monitoring study with roadside sampling of UFPs with at least 12 visits per location optimized exposure model performance while limiting costs. Furthermore, the investigators noted that it is important that the exposure sampling in mobile monitoring campaigns covers all days of the week, most hours of the day (including early morning and late evening hours), and at least two seasons.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results. Many of the analyses focused on UFPs, and there were few comparisons across pollutants, albeit more information is presented in the Additional Materials and other publications. For the evaluation of exposure design features using mobile roadside and mobile on-road monitoring data of UFPs, findings were presented in different stand-alone chapters, and different monitoring instruments were selected for various analyses. This makes a direct comparison between the two monitoring approaches difficult. An analysis investigating how the removal of possible extreme values affected the subsequent exposure models was missing from the report, but was included in a paper resulting from this work.

The Panel recommended caution in interpreting the findings from the health analyses and thought some carefully designed simulations would have complemented the real-world health study. The health analyses were considered limited, particularly because most of the exposure models used were based on measurements conducted up to 25 years later than the health outcome. In addition, some analyses lacked adjustment for important confounding variables, most notably socioeconomic status. Further research in other cities and pollutants would be helpful to assess the generalizability of the specific findings related to exposure assessment design.

The comprehensive report includes many findings that will be of broad interest and value to a wide audience.

ACKNOWLEDGMENTS

The HEI Review Committee is grateful to the Improved Exposure Assessment Review Panel for their thorough review of the study. The Committee is also grateful to Allison Patton for oversight of the study, to Hanna Boogaard for assistance with review of the Investigators' Report and preparation of its Commentary, to Tom Zaczkiewicz for editing the Investigators' Report and its Commentary, and to Kristin Eckles for her role in preparing this Research Report for publication.

REFERENCES

Alexeeff SE, Roy A, Shan J, Liu X, Messier K, Apte JS, et al. 2018. High-resolution mapping of traffic-related air pollution with Google street view cars and incidence of cardiovascular events within neighborhoods in Oakland, CA. Environ Health 17:38.

Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, et al. 2017. High-resolution air pollution mapping with Google Street View cars: exploiting big data. Environ Sci Technol 51:6999–7008.

Apte JS, Chambliss SE, Messier KP, Gani S, Upadhya AR, Kushwaha M, et al. 2024. Scalable multipollutant exposure assessment using routine mobile monitoring platforms. Research Report 216. Boston, MA: Health Effects Institute.

Bai L, Weichenthal S, Kwong JC, Burnett RT, Hatzopoulou M, Jerrett M, et al. 2019. Associations of long-term exposure to ultrafine particles and nitrogen dioxide with increased incidence of congestive heart failure and acute myocardial infarction. Am J Epidemiol 188:151–159.

Bergen S, Sheppard L, Kaufman JD, Szpiro AA. 2016. Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines. J R Stat Soc Ser C-Appl Stat 65:731–753.

Blanco MN, Gassett A, Gould T, Doubleday A, Slager DL, Austin E, et al. 2022. Characterization of annual average traffic-related air pollution concentrations in the Greater Seattle Area from a year-long mobile monitoring campaign. Environ Sci Technol 56:11460–11472.

Boogaard H, Crouse DL, Tanner E, Mantus E, van Erp AM, Vedal S, et al. 2024. Assessing adverse health effects of long-term exposure to low levels of ambient air pollution: The HEI experience and what's next? Environ Sci Technol 58:12767–12783.

Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, et al. 1993. An association between air pollution and mortality in six US cities. N Engl J Med 329:1753–1759.

Doubleday A, Blanco MN, Austin E, Marshall JD, Larson TV, Sheppard L. 2023. Characterizing ultrafine particle mobile monitoring data for epidemiology. Environ Sci Technol 57:9538–9547.

Downward GS, van Nunen EJHM, Kerckhoffs J, Vineis P, Brunekreef B, Boer JMA, et al. 2018. Long-term exposure to ultrafine particles and incidence of cardiovascular and cerebrovascular disease in a prospective study of a Dutch cohort. Environ Health Perspect 126:127007.

Hatzopoulou M, Valois MF, Levy I, Mihele C, Lu G, Bagg S, et al. 2017. Robustness of land-use regression models developed

from mobile air pollutant measurements. Environ Sci Technol 51:3938–3947.

HEI 2010. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Special Report 17. Boston, MA: Health Effects Institute.

HEI Review Panel on Ultrafine Particles. 2013. Understanding the health effects of ambient ultrafine particles. Perspectives 3. Boston, MA: Health Effects Institute.

HEI. 2022. Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute.

Hoek G. 2017. Methods for assessing long-term exposures to outdoor air pollutants. Curr Environ Health Rep 4:450–462.

Hoek G, Bouma F, Janssen N, Wesseling J, van Ratingen S, Kerckhoffs J, et al. 2025. Comparison of long-term air pollution exposure from mobile and routine monitoring, low-cost sensors, and dispersion models. Research Report 226. Boston, MA: Health Effects Institute.

IARC (International Agency for Research on Cancer) Working Group on the Evaluation of Carcinogenic Risks to Humans. 2016. Outdoor air pollution. IARC Monogr Eval Carcinog Risks Hum 109:9–444.

Kerckhoffs J, Hoek G, Messier KP, Brunekreef B, Meliefste K, Klompmaker JO, et al. 2016 Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ Sci Technol 50:12894–12902.

Kerckhoffs J, Hoek G, Vlaanderen J, van Nunen E, Messier K, Brunekreef B, et al. 2017. Robustness of intraurban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ Res 159:500–508.

Kerckhoffs J, Hoek G, Gehring U, Vermeulen R. 2021. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ Int 154:106569.

Kerckhoffs J, Khan J, Hoek G, Yuan Z, Hertel O, Ketzel M, et al. 2022. Hyperlocal variation of nitrogen dioxide, black carbon, and ultrafine particles measured with Google Street View cars in Amsterdam and Copenhagen. Environ Int 170:107575.

Kim SY, Blanco MN, Bi J, Larson TV, Sheppard L. 2023. Exposure assessment for air pollution epidemiology: A scoping review of emerging monitoring platforms and designs. Environ Res 223:115451.

Larkin A, Hystad P. 2017. Towards personal exposures: how technology is changing air pollution and health research. Curr Environ Health Rep 4:463–471.

Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, et al. 2018. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environ Int 116:286–299.

Messier KP, Chambliss SE, Gani S, Alvarez R, Brauer M, Choi JJ, et al. 2018. Mapping air pollution with Google street view cars: efficient approaches with mobile monitoring and land use regression. Environ Sci Technol 52:12563–12572.

Pérez IA, García MÁ, Sánchez ML, Pardo N, Fernández-Duque B. 2020. Key points in air pollution meteorology. Int J Environ Res Public Health 17:8349.

Pond ZA, Saha PK, Coleman CJ, Presto AA, Robinson AL, Pope CA III. 2022. Mortality risk and long-term exposure to ultrafine particles and primary fine particle components in a national US Cohort. Environ Int 167:107439.

Pope CA 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287:1132–1141.

Roque NA, Andrews H, Santos-Lozada AR. 2025. Identifying air quality monitoring deserts in the United States. Proc Natl Acad Sci U S A 122:e2425310122.

Samoli E, Butland BK. 2017. Incorporating measurement error from modeled air pollution exposures into epidemiological analyses. Curr Environ Health Rep 4:472–480.

Sheppard L, Burnett RT, Szpiro AA, Kim SY, Jerrett M, Pope CA 3rd, et al. 2012. Confounding and exposure measurement error in air pollution epidemiology. Air Qual Atmos Health 5:203–216.

Szpiro AA, Paciorek CJ, Sheppard L. 2011. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology 22:680–685.

Szpiro AA, Sheppard L, Lumley T. 2011. Efficient measurement error correction with spatially misaligned data. Biostatistics 12:610–623.

US EPA (United States Environmental Protection Agency). 2016. Integrated science assessment for oxides of nitrogen–health Criteria. EPA/600/R-15/068. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency), 2019. Integrated science assessment for particulate matter. EPA/600/R-19/188. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency). 2022. Supplement to the 2019 integrated science assessment for particulate matter. EPA/600/R-22/028. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency). 2024. Final rule: Reconsideration of the national ambient air quality standards for particulate matter. Available: *https://www.epa.gov/pm-pollution/final-reconsideration-national-ambient-air-quality-standards-particulate-matter-pm* [accessed May 22, 2025].

Weichenthal S, Bai L, Hatzopoulou M, Van Ryswyk K, Kwong JC, Jerrett M, et al. 2017. Long-term exposure to ambient ultrafine particles and respiratory disease incidence in Toronto, Canada: A cohort study. Environ Health 16:64.

Weichenthal S, Lloyd M, Ganji A, Simon L, Xu J, Venuta A, et al. 2024. Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada. Research Report 217. Boston, MA: Health Effects Institute.

WHO (World Health Organization). 2021. WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide, and Carbon Monoxide. Geneva: World Health Organization.

Zuidema C, Schumacher CS, Austin E, Carvlin G, Larson TV, Spalt EW, et al. 2021. Deployment, calibration, and cross-validation of low-cost electrochemical sensors for carbon monoxide, nitrogen oxides, and ozone for an epidemiological study. Sensors 21:4214.

Zusman M, Schumacher CS, Gassett AJ, Spalt EW, Austin E, Larson TV, et al. 2020. Calibration of low-cost particulate matter sensors: model development for a multi-city epidemiological study. Environ Inter 134:105329.

ABBREVIATIONS AND OTHER TERMS

A1	feature class code road type – interstate highways (primary highway with limited access)
A2	feature class code road type – primary highway without limited access
A3	feature class code road type – secondary and connecting roads
A4	feature class code road type – local, neighborhood, and rural roads
ACT	Adult Changes in Thought
ACT-AP	Adult Changes in Thought Air Pollution
AD	Alzheimer's disease
BC	black carbon
CASI	Cognitive Abilities Screening Instrument
CASI-IRT	CASI transformed using item response theory
CO_2	carbon dioxide
CV	cross-validation
CV-R^2	cross-validation based R^2
CV-MSE R^2	cross-validated mean squared error R^2
CV-RMSE	cross-validated root mean squared error
DEEDS	Diesel Exhaust Exposure in the Duwamish Study (Schulte et al. 2015)
DiSCmini	instrument measuring UFP
FEM	Federal Equivalent Method
FRM	Federal Reference Method
HMO	health maintenance organization
IQR	interquartile range
IRB	institutional review board
MAP	MESA Air Pilot (Wilton et al. 2010)
ME	measurement error
MESA Air	Multi-Ethnic Study of Atherosclerosis and Air Pollution
MSE	mean squared error
NanoScan	instrument measuring UFP
NO_2	nitrogen dioxide
P-Trak	instrument measuring UFP
PC	principal component
PCA	principal component analysis
PLS	partial least squares
$PM_{2.5}$	particulate matter ≤2.5 μm in aerodynamic diameter
PNC	particle number concentration in pt/cm^3
PSCAA	Puget Sound Clean Air Agency
PSID	Panel Study on Income Dynamics (Liu et al. 2003)
R^2	coefficient of determination or explained variance
RAD	Remote Air Data
RapPCA	representative and predictive PCA
RF	random forest
RMSE	root mean squared error
SD	standard deviation
SE	standard error
SES	socioeconomic status
SpatRF	spatial random forest
SpatRF-NP	SpatRF optimized using a nonparametric approach
SpatRF-PL	SpatRF optimized using pseudo-likelihood
TPRS	spatial smoothing via thin plate regression splines
TRAP	traffic-related air pollution
UFP	ultrafine particles
UK	universal kriging
UK-PLS	universal kriging followed by PLS
US EPA	United States Environmental Protection Agency
Yesler	Yesler Terrace study (Wong 2010)

RELATED HEI PUBLICATIONS

Number	Title	Principal Investigator	Date
Research Reports			
227	Investigating the Consequences of Measurement Error of Gradually More Sophisticated Long-Term Personal Exposure Models in Assessing Health Effects: The London Study (MELONS)	K. Katsouyanni	2025
226	Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models	G. Hoek	2025
217	Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada	S. Weichenthal	2024
216	Scalable Multipollutant Exposure Assessment Using Routine Mobile Monitoring Platforms	J. Apte	2024
212	Mortality–Air Pollution Associations in Low-Exposure Environments (MAPLE): Phase 2	M. Brauer	2022
211	Assessing Adverse Health Effects of Long-Term Exposure to Low Levels of Ambient Air Pollution: Implementation of Causal Inference Methods	F. Dominici	2022
208	Mortality and Morbidity Effects of Long-Term Exposure to Low-Level $PM_{2.5}$, BC, NO_2, and O_3: An Analysis of European Cohorts in the ELAPSE Project	B. Brunekreef	2021
203	Mortality–Air Pollution Associations in Low-Exposure Environments (MAPLE): Phase 1	M. Brauer	2019
202	Enhancing Models and Measurements of Traffic-Related Air Pollutants for Health Studies Using Dispersion Modeling and Bayesian Data Fusion	S. Batterman	2020
200	Assessing Adverse Health Effects of Long-Term Exposure to Low Levels of Ambient Air Pollution: Phase 1	F. Dominici	2019
196	Developing Multipollutant Exposure Indicators of Traffic Pollution: The Dorm Room Inhalation to Vehicle Emissions (DRIVE) Study	J. Sarnat	2018
194	A Dynamic Three-Dimensional Air Pollution Exposure Model for Hong Kong	B. Barratt	2018
177	National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components	M. Lippman	2013
Perspectives			
3	Understanding the Health Effects of Ambient Ultrafine Particles	HEI	2013
Special Reports			
23	Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution	HEI	2022
17	Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects	HEI	2010

PDFs are available for free downloading at www.healtheffects.org/publications.

HEI BOARD, COMMITTEES, and STAFF

BOARD OF DIRECTORS

Richard A. Meserve, Chair *Senior of Counsel, Covington & Burling LLP; President Emeritus, Carnegie Institution for Science; former Chair, US Nuclear Regulatory Commission, USA*

Stephen Corman *President, Corman Enterprises, USA*

Martha J. Crawford *Operating Partner, Macquarie Asset Management, USA*

Ana V. Diez Roux *Dana and David Dornsife Dean and Distinguished University Professor of Epidemiology, Dornsife School of Public Health, Drexel University; Director, Drexel Urban Health Collaborative, USA*

Michael J. Klag *Dean Emeritus and Second Century Distinguished Professor, Johns Hopkins Bloomberg School of Public Health, USA*

Alan I. Leshner *CEO Emeritus, American Association for the Advancement of Science, USA*

Catherine L. Ross *Regents' Professor Emerita, City and Regional Planning and Civil and Environmental Engineering, Georgia Institute of Technology; Chairman of the Board of Directors of the Auto Club Group, American Automobile Association, USA*

Martha E. Rudolph *Environmental Attorney, Former Director of Environmental Programs, Colorado Department of Public Health and Environment, USA*

Karen C. Seto *Frederick C. Hixon Professor of Geography and Urbanization Science, Yale School of the Environment, Yale University, USA*

Jared L. Cohon *President Emeritus and Professor, Civil and Environmental Engineering and Engineering and Public Policy, Carnegie Mellon University, USA, In Memoriam 1947–2024*

RESEARCH COMMITTEE

David A. Savitz, Chair *Professor of Epidemiology, School of Public Health, and Professor of Obstetrics and Gynecology and Pediatrics, Alpert Medical School, Brown University, USA*

Benjamin Barratt *Professor, Environmental Research Group, School of Public Health, Imperial College London, United Kingdom*

David C. Dorman *Professor, Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, USA*

Christina H. Fuller *Associate Professor, School of Environmental, Civil, Agricultural and Mechanical Engineering, University of Georgia College of Engineering, USA*

Marianne Hatzopoulou *Professor, Civil and Mineral Engineering, University of Toronto, Research Chair in Transport Decarbonization and Air Quality, Canada*

Heather A. Holmes *Associate Professor, Department of Chemical Engineering, University of Utah, USA*

Neil Pearce *Professor of Epidemiology and Biostatistics, London School of Hygiene and Tropical Medicine, United Kingdom*

Evangelia (Evi) Samoli *Professor of Epidemiology and Medical Statistics, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece*

Alexandra M. Schmidt *Professor of Biostatistics, School of Population and Global Health, McGill University, Canada*

Neeta Thakur *Associate Professor of Medicine, University of California San Francisco, USA*

Gregory Wellenius *Professor, Department of Environmental Health, Boston University School of Public Health and Director, BUSPH Center for Climate and Health, USA*

continued on next page

HEI BOARD, COMMITTEES, and STAFF

REVIEW COMMITTEE

Sara D. Adar *Professor of Epidemiology and Global Public Health, Department of Epidemiology, University of Michigan School of Public Health, USA*

Kiros T. Berhane *Cynthia and Robert Citrone-Roslyn and Leslie Goldstein Professor and Chair, Department of Biostatistics, Mailman School of Public Health, Columbia University, USA*

Katherine B. Ensor *Noah G. Harding Professor of Statistics, Rice University, USA*

Ulrike Gehring *Associate Professor, Institute for Risk Assessment Sciences, Utrecht University, Netherlands*

Michael Jerrett *Professor, Department of Environmental Health Sciences, Fielding School of Public Health, University of California Los Angeles, USA*

Frank Kelly *Humphrey Battcock Chair in Community Health and Policy and Director of the Environmental Research Group, Imperial College London School of Public Health, United Kingdom*

Eric J. Tchetgen Tchetgen *University Professor and Professor of Biostatistics and Epidemiology, Perelman School of Medicine, and Professor of Statistics and Data Science, The Wharton School, University of Pennsylvania, USA*

John Volckens *Professor, Department of Mechanical Engineering, Walter Scott Jr. College of Engineering, Colorado State University, USA*

Scott Weichenthal *Professor, Department of Epidemiology, Biostatistics, and Occupational Health, School of Population and Global Health, McGill University, Canada*

STAFF AND CONSULTING SCIENTISTS

Elena Craft *President and CEO*

Ellen K. Mantus *Director of Science*

Donna J. Vorhees *Director of HEI Energy*

Thomas J. Champoux *Director of Science Communications*

Jacqueline C. Rutledge *Director of Finance and Administration*

Emily Alden *Corporate Secretary*

Daniel S. Greenbaum *President Emeritus, In Memoriam 1952–2024*

Robert M. O'Keefe *Vice President Emeritus*

Annemoon M. van Erp *Deputy Director of Science Emerita*

Amy Andreini *Science Communications Specialist*

Hanna Boogaard *Consulting Principal Scientist*

Jacki Collins *Senior Staff Accountant*

Dan Crouse *Senior Scientist*

Cloelle Danforth *Senior Scientist*

Gabriela Daza *Research Assistant*

Philip J. DeMarco *Compliance Manager*

Kristin C. Eckles *Senior Editorial Manager*

Hlina Kiros *Research Assistant*

Continues next page

HEI BOARD, COMMITTEES, and STAFF

(*Staff and Consulting Scientists, continued*)

Lissa McBurney *Senior Science Administrator*
Samantha Miller *Research Assistant*
Victor Nthusi *Consulting Research Fellow*
Pallavi Pant *Head of Global Initiatives*
Allison P. Patton *Senior Scientist*
Yasmin Romitti *Staff Scientist*
Anna Rosofsky *Senior Scientist and Community Health and Environmental Research Initiatives Lead*
Abinaya Sekar *Consulting Research Fellow*
Robert Shavers *Operations Manager*
Eva Tanner *Staff Scientist*
Alexis Vaskas *Digital Communications Manager*

RESEARCH REPORT

NUMBER 228
AUGUST 2025

Health Effects Institute
75 Federal Street
Suite 1400
Boston, Massachusetts, 02110, USA
+1-617-488-2300

www.healtheffects.org