NOVEL METHODS FOR OCCUPATIONAL AND NON-OCCUPATIONAL EXPOSURE ASSESSMENT FOR IMPROVED RISK ASSESSMENT AND DECISION MAKING

by
Andrew N. Patton

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
February 2021

ABSTRACT

Problem Statement

Human health risk assessments require accurate exposure assessments to be meaningful. Answering, or improving current answers, to exposure questions can require novel methodologies that provide additional utility to existing approaches. We demonstrate three such methods, one for data collection and two for data analysis. While the methods were developed specifically for this dissertation, they are scalable to other applications and can be generalized to similar research questions.

1) In a situation where there is a nearly 30-year gap in data collection, estimate benzene exposure and cancer risk to non-occupational and occupational groups from commercial gasoline station filling operations.
2) With an existing low-cost sensor network measuring ambient $PM_{2.5}$, utilize probabilistic machine learning models to improve on predictive accuracy of previously developed linear models and then use the output to conduct probabilistic exposure assessments.
3) Create a probabilistic machine learning calibration model for CO sensors deployed in an occupational low-cost sensor network and use the models to create probabilistic concentration hazard maps and American Industrial Hygiene Association exposure category hazard maps to assist with decision making.

CHAPTER SUMMARIES

Chapter II

Gasoline station exposures were measured using consumer self-sampling protocols where study participants were unsupervised and in control of the start and stop of their sampling period using whole air canisters to match the duration of their filling activities. Benzene exposures and cancer risks associated with gasoline station filling operations were found to be within accepted risk tolerance levels for both consumer and modeled occupational situations. Additionally, cancer risk from ambient benzene was found to be greater compared to pumping risk in all consumer scenarios and the majority of occupational scenarios.

Chapter III

The probabilistic GBDTs for calibrating $PM_{2.5}$ measurements from low-cost sensor networks were trained and tested on identical splits as an existing linear model for the same dataset. However, the GBDT used raw sensor data whereas the linear model used lab corrected data. The results of the GBDT were then spatially interpolated in a probabilistic manner to create exposure assessments based on administrative borders. The probabilistic gradient boosted decision trees were found to not only be more accurate than linear models at calibration at predicting reference concentrations, but directly used raw sensor data, not laboratory calibrated sensor data like the linear models.

Chapter IV

Utilizing a probabilistic calibration model for occupational CO exposures from a low-cost sensor network eliminated the need for laboratory calibrations and multiple linear calibration models. The model was then used to demonstrate how probabilistic predictions can create probabilistic hazard maps both of CO concentrations and AIHA

exposure ratings, an AIHA method for creating discrete exposure categories based on the uncertainty surrounding the 95th percentile of the sampling distribution. The AIHA exposure rating hazard maps were shown as a tool for improving decision making and also illustrated the weakness of traditional point predictions in terms of accurately estimating occupational exposures. These small sample size exposure rating calculations can be substantially underestimate or overestimate exposures, leading to incorrect industrial hygiene resource apportionment and potentially health and/or regulatory overexposures.

Conclusion

The novel exposure assessment methodologies presented, both in terms of data collection and data analysis are viable tools for capturing exposure information necessary for human health risk assessments.

Using self-sampling protocols in data collection for non-occupational groups allows for sampled individuals to complete the sampled task in as normal a manner as possible. While COVD-19 impacts reduced the sample numbers by over 60 percent from its intended size, the exposure data from this chapter filled a nearly 30-year gap in the literature and was used to demonstrate that gasoline pumping activities for consumers and occupational groups do not present an unacceptable amount of risk. With regards to low-cost sensor networks for ambient air pollution, both indoors and out, non-linear machine learning techniques can eliminate the need for time intensive lab calibration of sensor instruments and improve on linear approaches. In addition, probabilistic models can be used to create spatial exposure assessments that fit neatly into a probabilistic risk assessment framework for environmental exposures, or the American Industrial Hygiene Association exposure rating framework.

COMMITTEE OF THESIS READERS

Advisor: Kirsten Koehler, PhD
Associate Professor of Environmental Health and Engineering

Readers**:** Benjamin Zaitchik, PhD
Associate Professor of Earth and Planetary Science

Gurumurthy Ramachandran, PhD
Professor of Environmental Health and Engineering

Mary Fox, PhD
Assistant Professor of Health Policy and Management

Alternates: Fenna Sillé, PhD
Assistant Professor of Environmental Health and Engineering

Abhirup Datta, PhD
Assistant Professor of Biostatistics

RESEARCH COMMITTEE

Kirsten Koehler, PhD
Associate Professor of Environmental Health and Engineering

Mary Fox, PhD
Assistant Professor of Health Policy and Management

Gurumurthy Ramachandran, PhD
Professor of Environmental Health and Engineering

Abhirup Datta, PhD
Assistant Professor of Biostatistics

Fenna Sillé, PhD
Assistant Professor of Environmental Health and Engineering

ACKNOWLEDGEMENTS

Danielle for the last seven years.

All the participants in the consumer backpack gas station sampling.

NIOSH and CARTEEH for supporting my work.

NBA Slack (Nathan, Kostya, Ryan, Magnus, etc.) for python/modeling support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Exposure Assessment Strategies

In 2012, the National Academies published a resource detailing the needs for exposure assessments and exposure science in the 21st century (National Research Council 2012). Within the document, they outline short, medium, and long-term goals for research needs regarding modernizing exposure assessments into three categories. The three aims of this dissertation each correspond to one primary research need group, and all three together seek to address the criteria set out by the National Academies to push exposure assessment and exposure science into the 21st century.

Dissertation Aims and Structure

The three aims of this dissertation are as follows:

1) Determine benzene exposure and cancer risk from commercial gasoline station filling operations among non-occupational and occupational groups
2) Provide a probabilistic modeling framework for $PM_{2.5}$ exposure assessments with low-cost sensor networks that also reduces the need for lab calibrations
3) Within the context of the American Industrial Hygiene Associations' exposure ratings, develop a probabilistic calibration model for CO low-cost sensor networks that assists with regulatory decision making, while also demonstrating the improved utility of probabilistic models over point predictions for occupational exposure estimation.

Three manuscripts, Chapters II, III, and IV, make up the main body of the dissertation. Chapter I is the introductory materials and Chapter V is the conclusion.

Aim 1: Commercial Gasoline Station Filling

National Academies Criteria: "*Providing effective responses to immediate or short-term public-health or ecologic risks requires research on observational methods, data management, and models".*

In the United States as of 2013, nearly 40 million consumers purchase automotive gasoline per day (National Association of Convenience Stores 2013). Approximately 80% of that gasoline is purchased at convenience stores/gasoline stations and 16% is purchased at grocery or large chain retail stores (National Association of Convenience Stores 2019). The gasoline available for purchase in the United States is a non-unform hydrocarbon mixture that can vary by refinery, brand, time of year, etc. (ATSDR 1996; IARC 1989). Within the constituents of gasoline, benzene is one component that is federally regulated, with any refinery or importer average not to exceed 0.62% benzene by volume (Bruckner et al. 2008; Environmental Protection Agency 2012). The regulation is in place due to the established carcinogenicity of benzene as established by the Environmental Protection Agency, U.S. Department of Health and Human Services, and the International Agency for Research on Cancer (ATSDR 1996; IARC 1989; IRIS 2003). Benzene is primarily causative of acute myeloid leukemia (AML), but has also been linked to acute nonlymphocytic leukemia (ANLL) and myelodysplastic syndrome (MDS) as well (ATSDR 2006; Keenan et al. 2013). Notwithstanding the established carcinogenicity of benzene and the millions of exposed consumers per day, virtually no exposure assessments have been conducted in the United States to determine exposure and/or risk in this population in approximately the last 30 years, despite substantial studies on other downstream populations (Verma et al. 2001). In addition to exposures from gasoline station filling, benzene is an ambient pollutant from fossil fuel combustion and is considered by

the National Air Toxics Assessment (NATA) to be a 'national cancer risk contributor' from ambient exposure (ATSDR 2006; EPA 2015, 2018; Galbraith et al. 2010). Using a novel whole air self-sampling protocol, we conducted gasoline station filling operations exposure assessments on consumers, and corresponding probabilistic risk assessments for consumer and occupational scenarios. The risk assessment results indicated that filling operations for both consumers and occupational groups likely carry no excess risk of leukemia. Furthermore, the whole air self-sampling protocol developed for Aim 1 allows for short- or long-term whole air sampling of individuals based on tasks of interest or simply ambient exposures that can be then linked to the exposome, fulfilling the National Academies criteria for increased personal monitoring capabilities.

Aim 2: Low-Cost Sensor Networks for Ambient $PM_{2.5}$

National Academies Criteria: *"Supporting research on health and ecologic effects that addresses past, current, and emerging outcomes"*

Particulate matter ($PM_{2.5}$) air pollution is well established as a serious human health hazard, with the World Health Organization (WHO) estimating that $PM_{2.5}$ is causative of nearly 7 million deaths per year (World Health Organization 2018). $PM_{2.5}$ has both carcinogenic and non-carcinogenic health effects, and has been linked to lung cancer, cardiovascular disease, and stroke (International Agency for Research on Cancer 2016). Urban areas often experience the highest burden of $PM_{2.5}$ exposures due to the fact that $PM_{2.5}$ is produced via hydrocarbon combustion from vehicle traffic, electricity generation, cooking, and other sources. (International Agency for Research on Cancer 2016; Saha et al. 2020). In order to ensure that $PM_{2.5}$ levels meet federal regulatory standards established by the Clean Air Act, the Environmental

Protection Agency mandates that states operate high quality reference monitors in counties around the country (Environmental Protection Agency 2010; Maryland Department of the Environment 2018). However, given that these federally mandated monitors are stationary monitors required to prove compliance data at the county-level, there are only 275 monitors covering the over 110 million people in the 25 most populous urban areas – resulting in limited spatial coverage on the intra-city scale (Apte et al. 2017, Environmental Protection Agency 2010). In order to increase the spatial coverage of $PM_{2.5}$ monitoring, low-cost sensor networks consisting of numerous spatially distributed sensors of lower accuracy, precision, and cost than federal reference monitors have been implemented (Buehler et al. 2020; Datta et al. 2020).

These low-cost networks can measure exposure gradients on much smaller spatial scale across the area under observation than the reference monitors can provide (Piedrahita et al. 2014; Snyder et al. 2013; Szpiro et al. 2009). Despite increased spatial coverage and resolution, direct utilization of the raw $PM_{2.5}$ sensor data is not encouraged due to the reduction in accuracy and precision associated with the low-cost sensors that have not been corrected for environmental biases (Borrego et al. 2018; Buehler et al. 2020; Morawska et al. 2018). The raw sensor data often needs a lab calibration or other method to ensure the sensors are returning measurements that are accurate and precise, especially as $PM_{2.5}$ measurements are impacted by both temperature and relative humidity (Borrego et al. 2018; Datta et al. 2020; Levy Zamora et al. 2019; Morawska et al. 2018). However, the laboratory calibration process is highly time intensive and requires specialized facilities and equipment to calibrate each sensor for climate conditions of their deployed environment and the theoretical range of exposure levels (Levy Zamora et al. 2019). Frequently, linear calibration models are created to turn the raw sensor readings into a lab-calibrated value that can be used for exposure estimates (Datta et al. 2020;

Levy Zamora et al. 2019). While non-linear terms are added to the linear regression models, the calibration equations can fail to capture peaks and valleys in the reference measurements.

Despite traditionally a linear regression-based approach, calibration models have been conducted with a variety of modeling approaches including polynomial regression, gain/offset linear regression, land use regression, and machine learning methods like gradient boosted decision trees (GBDT) and random forests (Johnson et al. 2018; Lim et al. 2019). Using a novel fully probabilistic gradient boosting library (NGBoost) we developed models that provided increased point prediction and distributional accuracy to linear models. Furthermore, these models used raw data, not the calibrated data required for the linear model, showing that increased predictive power is possible without the intermediary step of lab calibration. Lastly, the unique modeled mean and variance for each prediction was used to create a framework for probabilistic spatial exposure assessments. The National Academies highlighted specifically that exposure modeling methodologies need to be modernized to better address and account for uncertainty in data and results. The use of NGBoost and a spatial interpolation Monte Carlo to propagate uncertainty throughout the exposure assessment process combined with utilizing sensor network data explicitly fulfills the research goals set out by the National Academies.

Aim 3: Occupational Low-Cost Sensor Networks for Industrial Hygiene Decision Making

National Academies Criteria: *"Addressing demands for exposure information among communities, governments, and industries with research that is focused, solution-based, and responsive to a broad array of audiences"*

The Occupational Safety and Health Administration (OSHA) promulgates legally enforceable permissible exposure limits (PEL) for workplaces in the United States. Additionally, OSHA provides guidance on a sampling strategy for conducting the compliance exposure assessment based on a very small number of samples taken on the highest risk workers (Ramachandran 2005). However, this strategy is likely to both underestimate the true exposures to the sampled workers and also entirely fail to discern exposures to workers not sampled (OSHA 2001; Tuggle 1981). However, the American Industrial Hygiene Association (AIHA) provides a framework that is less likely to underestimate exposures by incorporating the uncertainty of the true underlaying concentration distribution as opposed to just the sampling and analytical uncertainty of the OSHA strategy (Ramachandran 2005). The AIHA framework is four sets of exposure ratings ranging from 'Highly Controlled' to 'Poorly Controlled' which provide industrial hygienists with guidance about where to focus control efforts (Ramachandran 2005). Despite the improvements on the OSHA compliance sampling strategy, the AIHA framework still generally relies on a small number of personal samples. However, by utilizing low-cost sensor networks, multiple low precision/accuracy sensors spatially distributed across a workplace, exposure data can be collected at a small enough spatiotemporal scale to allow for substantial utility over single higher precision measurements (Peters et al. 2006; Thomas et al. 2018; Vosburgh et al. 2011; Zuidema et al. 2019). The high resolution data can then be spatially interpolated to create hazard maps, a type of mapping that shows exposure gradients over time, space, or both to provide full worksite exposure estimates (Koehler and Peters 2013; Koehler and Volckens 2011). These high-resolution hazard maps can be then used to estimate personal exposures based on the location of an individual on the facility floor. However, with multiple lower quality sensors, calibration to a reference standard is a non-trivial amount of work that can require frequent site visits, specialized laboratory facilities, and person hours (Afshar-Mohajer et al. 2018; Datta et al. 2020; Levy Zamora et al. 2019; Zuidema et al. 2019).

Entirely modeling based approaches to calibrate low-cost sensor networks (following collection of reference data) have been utilized to great effect, although the models that perform best tend to be non-linear machine learned based approaches (Zimmerman et al. 2018, Chapter III). Thus far, this approach has not been used on an entirely indoor occupational low-cost sensor network. Using data from an approximately 40 unit low-cost sensor network deployed in a heavy equipment manufacturing facility, we developed probabilistic calibration models for carbon monoxide (CO) using gradient boosted decision trees that eliminated the need for laboratory calibration (Zuidema et al. 2019). The model's predictions were spatially interpolated to create probabilistic concentration hazard maps as well as AIHA exposure rating hazard maps to create a framework that allows industrial hygienists to be more informed about where to direct control resources. Additionally, the probabilistic results showed how utilizing a mean or point prediction alone creates the potential to incorrectly identify areas of highest exposure, even when using large sample sizes associated with network data. The results of this chapter fulfill the requirements from the National Academies for research that is solution based, as we provide a framework for occupational exposure assessments that vastly outperform prior methodologies and can be used to improve both worker health and regulatory compliance.

References

Afshar-Mohajer N, Zuidema C, Sousan S, Hallett L, Tatum M, Rule AM, et al. 2018. Evaluation of low-cost electro-chemical sensors for environmental monitoring of ozone, nitrogen dioxide, and carbon monoxide. J Occup Environ Hyg 15:87–98; doi:10.1080/15459624.2017.1388918.

Apte J, Messier K, Gani S, Brauer M, Kirchstetter T, Lunden M, et al. 2017. High-resolution air pollution mapping with Google Street View cars: exploiting big data (Supplemental Material). Environ Sci Technol 51: 6999–7008.

ATSDR. 1996. Automotive Gasoline.

ATSDR. 2006. Toxicological Profile for Benzene.

Borrego C, Ginja J, Coutinho M, Ribeiro C, Karatzas K, Sioumis T, et al. 2018. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise – Part II. Atmos Environ 193:127–142; doi:10.1016/j.atmosenv.2018.08.028.

Bruckner J, Anand S, Warren A. 2008. Toxic Effects of Solvents and Vapors. In: *Casarett & Doull's Toxicology: The Basic Science of Poisons* (C. Klaasen, ed). McGraw Hill. 981–1051.

Buehler C, Xiong F, Levy Zamora M, Skog K, Kohrman-Glaser J, Colton S, et al. 2020. Stationary and Portable Multipollutant Monitors for High Spatiotemporal Resolution Air Quality Studies including Online Calibration. Atmos Meas Tech in review.

Datta A, Saha A, Zamora ML, Buehler C, Hao L, Xiong F, et al. 2020. Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmos Environ 242:117761; doi:10.1016/j.atmosenv.2020.117761.

Environmental Protection Agency. 2012. Gasoline Mobile Source Air Toxics.

Environmental Protection Agency. 2010. NAAQS Table. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table.

EPA. 2018. 2014 NATA Summary of Results.

EPA. 2015. Technical Support Document EPA's 2011 National-scale Air Toxics Assessment. Available: https://www.epa.gov/sites/production/files/2015-12/documents/2011-nata-tsd.pdf.

Galbraith D, Gross SA, Paustenbach D. 2010. Benzene and human health: A historical review and appraisal of associations with various diseases. Crit Rev Toxicol 40:1–46; doi:10.3109/10408444.2010.508162.

IARC. 1989. Gasoline. Monographs 45.

International Agency for Research on Cancer. 2016. Outdoor Air Pollution (Vol. 109).

IRIS. 2003. Benzene. Chem Assess Summ.

Keenan JJ, Gaffney S, Gross SA, Ronk CJ, Paustenbach DJ, Galbraith D, et al. 2013. An evidence-based analysis of epidemiologic associations between lymphatic and hematopoietic cancers and occupational exposure to gasoline. Hum Exp Toxicol 32:1007–1027; doi:10.1177/0960327113476909.

Koehler K, Peters T. 2013. Influence of Analysis Methods on Interpretation of Hazard Maps. Ann Occup Hyg 57: 558–570.

Koehler K, Volckens J. 2011. Prospects and pitfalls of occupational hazard mapping: "between these lines there be dragons". Ann Occup Hyg 55: 829–840.

Levy Zamora M, Xiong F, Gentner D, Kerkez B, Kohrman-Glaser J, Koehler K. 2019. Field and Laboratory Evaluations of the Low-Cost Plantower Particulate Matter Sensor. Environ Sci Technol 53:838–849; doi:10.1021/acs.est.8b05174.

Maryland Department of the Environment. 2018. Ambient Air Monitoring Network Plan for Calendar Year 2019.

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; doi:10.1016/j.envint.2018.04.018.

National Association of Convenience Stores. 2013. Fueling America: A Snapshot of Key Facts and Figures.

National Association of Convenience Stores. 2019. NACS | Selling America's Fuel. Available: https://www.convenience.org/Topics/Fuels/Who-Sells-Americas-Fuel [accessed 5 January 2020].

National Research Council. 2012. *Exposure Science in the 21st Century: A Vision and a Strategy*. The National Academies Press:Washington, DC.

OSHA. 2001. Appendix B to the Formaldehyde Standard. Code of Federal Regulations 29, Part 1910.1048.

Peters TM, Heitbrink WA, Evans DE, Slavin TJ, Maynard AD. 2006. The mapping of fine and ultrafine particle concentrations in an engine machining and assembly facility. Ann Occup Hyg 50:249–257; doi:10.1093/annhyg/mei061.

Ramachandran G. 2005. *Occupational Exposure Assessment for Air Contaminants*. 1st Editio. CRC Press:Minneapolis.

Saha PK, Sengupta S, Adams P, Robinson AL, Presto AA. 2020. Spatial Correlation of Ultrafine Particle Number and Fine Particle Mass at Urban Scales: Implications for Health Assessment. Environ Sci Technol 54:9295–9304; doi:10.1021/acs.est.0c02763.

Thomas G, Sousan S, Tatum M, Liu X, Zuidema C, Fitzpatrick M, et al. 2018. Low-Cost, Distributed Environmental Monitors for Factory Worker Health. Sensors 18:1411; doi:10.3390/s18051411.

Tuggle RM. 1981. The NIOSH decision scheme. Am Ind Hyg Assoc J 42:493–498; doi:10.1080/15298668191420134.

Verma DK, Johnson DM, Shaw ML, des Tombe K. 2001. Benzene and Total Hydrocarbons Exposures in the Downstream Petroleum Industries. AIHAJ - Am Ind Hyg Assoc 62:176–

194; doi:10.1080/15298660108984621.

Vosburgh DJH, Boysen DA, Oleson JJ, Peters TM. 2011. Airborne nanoparticle concentrations in the manufacturing of polytetrafluoroethylene (Ptfe) apparel. J Occup Environ Hyg 8:139–146; doi:10.1080/15459624.2011.554317.

World Health Organization. 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action.

Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, et al. 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech 11:291–313; doi:10.5194/amt-11-291-2018.

Zuidema C, Sousan S, Stebounova L V., Gray A, Liu X, Tatum M, et al. 2019. Mapping Occupational Hazards with a Multi-sensor Network in a Heavy-Vehicle Manufacturing Facility. Ann Work Expo Heal 63:280–293; doi:10.1093/annweh/wxy111.

CHAPTER II: BENZENE EXPOSURE AND CANCER RISK FROM COMMERCIAL GASOLINE STATION FUELING EVENTS USING A NOVEL SELF-SAMPLING PROTOCOL

1. Abstract

Tens of millions of individuals go to gasoline stations to purchase gasoline on a daily basis in the United States. One of the constituents of gasoline is benzene, a Group 1 carcinogen that has been strongly linked to both occupational and non-occupational leukemias. While benzene content in gasoline is federally regulated, there is approximately a thirty-year data gap in benzene exposures specifically from pumping gasoline. Using a novel self-sampling protocol with whole air canisters, we conducted a gasoline pumping exposure assessment for benzene, toluene, ethylbenzene, and xylene (BTEX) on Baltimore, MD area consumers. Exposures averaged 5.7 ppb, 23.5 ppb, 3.9 ppb, and 16.7 ppb respectively on 32 samples. Using the benzene exposure results, we conducted consumer and occupational probabilistic risk assessment scenarios and then contextualized the gasoline pumping risk with ambient benzene exposure risk. We found that the consumer scenarios did not approach the 1:1,000,000 excess risk management threshold and that the occupational scenario approached but did not exceed the 1:10,000 excess risk management threshold. Further, we found that in all case the ambient risk from benzene exceed that of pumping risk for consumers, but that in approximately 30% of occupational trials the pumping risk exceeded the ambient risk.

2. Introduction

According to the National Association of Convenience Stores, in 2019 there were approximately 129,000 convenience store gasoline stations and mass merchandising gasoline stations in the United States, accounting for 96% of all commercial gasoline sold (National Association of Convenience Stores 2019; NACS | Convenience Stores Fuel America). There were approximately 40 million fill ups per day at these gasoline stations as of 2012 (National Association of Convenience Stores 2013). The main form of gasoline sold is automotive gasoline, the primary fuel for internal combustion engines found in non-diesel cars,

motorcycles, non-diesel trucks, and other small engines (ATSDR 1995). Gasoline is a complex non-uniform mixture comprised of a variety of alkanes, alkenes, isoalkanes, cycloalkanes, cycloalkenes, and aromatics; many blends also contain performance-enhancing additives (IARC 1989). The exact ratios of these compounds vary by manufacturer and location, and even from batch to batch, depending on factors such as the source of the crude oil, the refining process used in its production, and the product specifications (ATSDR 1995; IARC 1989). Since the Clean Air Act Amendments of 1990, gasoline frequently contains ethanol in addition to petroleum products. The two most common mixtures in the United States are 10% ethanol/90% gasoline (E10) and 15% ethanol/85% gasoline (E15) (Alternative Fuels Data Center 2017). Approximately 95% of gasoline sold in the United States is E10 (Alternative Fuels Data Center 2017). However, the amount of ethanol used in blending can vary substantially, with maximums nearing 85% ethanol, primarily used outside of the United States (Alternative Fuels Data Center 2017).

Gasoline is a known human and animal carcinogen based on the toxicity of its components (ATSDR 1995; IARC 1989). Amongst the constituents of gasoline, benzene has the strongest body of evidence supporting its carcinogenicity (leukemias) in occupational and non-occupational settings, and the Environmental Protection Agency (EPA), U.S. Department of Health and Human Services, and International Agency for Research on Cancer (IARC), all identify benzene as a human carcinogen (ATSDR 2006; IARC 1989; IRIS 2003). There is epidemiological and toxicological evidence that excess benzene exposure can result in the development of acute myeloid leukemia (AML) in humans (Bergsagel et al. 1999; Paxton et al. 1994a, 1994b; Wong 1995). Other leukemias, including acute nonlymphocytic leukemia (ANLL) and myelodysplastic syndrome (MDS), have also been found to be associated with elevated benzene exposure (ATSDR 2006; Keenan et al. 2013). Benzene content is federally regulated,

with any refineries or importers required to average less than or equal to 0.62% benzene by volume in their gasoline (Environmental Protection Agency 2012). Generally, gasoline in the United States is likely to contain 0.5-2.0% benzene by volume (Bruckner et al. 2008; Environmental Protection Agency 2012). Additionally, in terms of non-occupational exposures, the National Air Toxics Assessment (NATA) ambient air pollution monitoring includes benzene as a 'national cancer risk contributor' and provides excess cancer risk associated with that ambient benzene exposure (EPA 2018b, 2018a).

Benzene exposure has been extensively studied in both upstream (petroleum extraction and production) and downstream (refining and marketing) settings (Verma et al. 2000, 2001). However, there is little information regarding potential exposures to the gasoline station consumer, a population of millions of individuals per day in the United States. According to the Agency for Toxic Substances and Disease Registry (ATSDR), non-occupational exposures to gasoline occur as a result of customers using the gasoline pumps and inhaling any volatilized part of the gasoline mixture (ATSDR 1995). The bulk of the studies and samples associated with consumers filling their own vehicles occurred in the 1980s and 1990s and were conducted by consulting firms or industrial sources (Northeast States for Coordinated Air Use Management 1985; Page and Mehlman 1989; Verma et al. 2001). Additionally, of the studies that were conducted in other countries (e.g., Singapore, Italy, England), approximately five percent of the mean benzene concentrations were greater than 2.5 ppm for short-duration, consumer focused measurements, the short-term occupational exposure limit issued by the American Conference of Governmental Industrial Hygienists. However, studies conducted in Europe in the 2000s indicate significantly reduced exposures compared to the 1980s and 1990s (OSHA 2020; Page and Mehlman 1989; Periago and Prado 2005). In addition to consumers, there are approximately 21,000 gasoline service station attendants across the

country as of 2019 who may also pump gasoline as part of their job description (BLS 2019). Furthermore, in New Jersey (and to a much smaller extent in Oregon), there are nearly 5,000 pump attendants who are legally required to pump gas for customers (Nobile 2018; Weller 2018).

We conducted an exposure assessment for consumers to characterize benzene and associated volatile organic compounds exposures associated with filling their gas tank using modern sampling and analysis methodologies. In addition, an exposure assessment was used to inform a consumer risk assessment for gasoline station filling. The risk assessment was extended to an occupational setting by developing a worker exposure scenario to estimate excess risk values for gasoline service station attendants and pump attendants. Lastly, the risk assessment results for the consumer and occupational groups were compared to the risk values provided by NATA in order to contextualize the risk from gasoline station benzene exposures with the risk from ambient benzene exposures.

3. Methods

3.1. Consumer Sampling – Design

The study participants consisted of 34 Baltimore, Maryland area consumers who were aged 18 or over, English speaking, literate, had a valid driver's license, had access to a working gasoline powered vehicle, and were able to fill their vehicle with at least five gallons of gasoline. Active smokers and pregnant or nursing women were excluded. At the time of consent into the study, each consumer was provided with a backpack (Figure 1) containing sampling equipment, demographic surveys, and fill-up specific questionnaires including questions such as 'How many times per month do you pump gas?', 'What make and model of car do you drive?', and

questions on related topics. The study was approved by the Johns Hopkins Institutional Review Board.

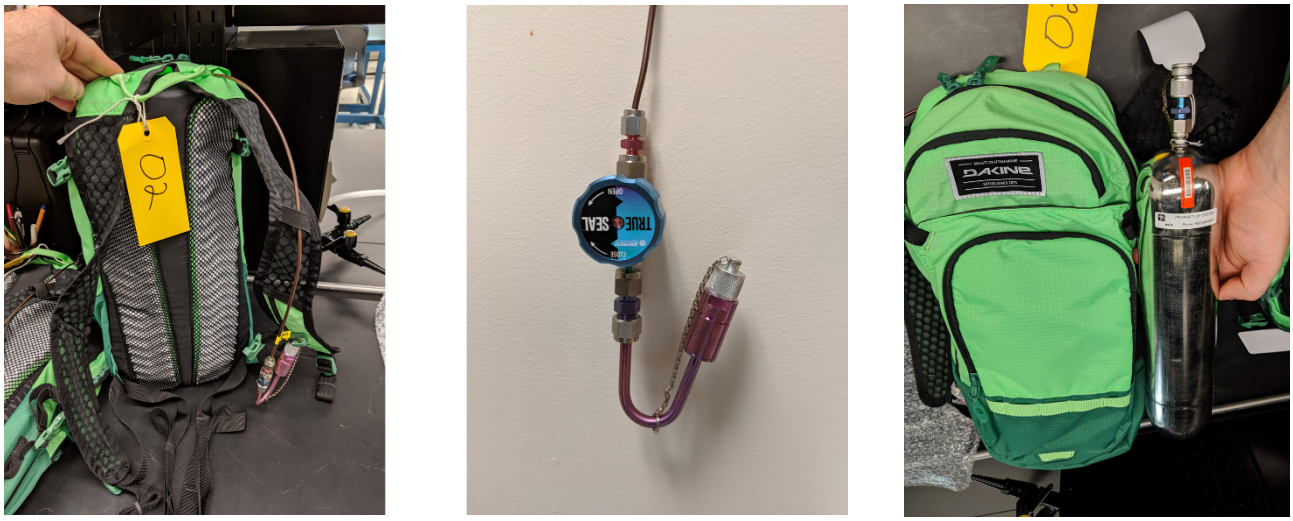


Figure 1: Consumer sampling backpack containing an evacuated steel canister, sampling line, flow regulator and start/stop knob, and MSR climate monitor

The consumer sampling was conducted using a backpack containing sampling equipment and electronics. The air sampling equipment was comprised of a 1.0 L MiniCan (Entech Instruments, Simi Valley, CA, USA) Silonite lined passivated steel canister evacuated to -30.00 mmHg and an attached Silonite lined steel sampling line with a flow regulator, open/close knob, and screw cap. Once opened, the canisters draw in whole air at 0.167 L/minute, for a six-minute operational limit. However, due to a loose gasket seal on the inlet valve with the first ten canisters used, a revised gasketless valve design was implemented for all other canisters. The flow rate, pressure, and canisters were otherwise identical. The backpack also contained an MSR 145 Data Logger (MSR Electronics GmbH, Seuzach, Switzerland) with temperature, relative humidity, and light sensors recording data on one-second intervals. Chubb Environmental Health Laboratory (Cromwell, CT, USA) provided the canisters and designed and

provided the custom sampling lines. Prior to providing each consumer with a backpack, each line was cleared with a clean vacuum canister and each sampling canister had its vacuum measured and recorded with an electronic pressure gauge.

3.2. Consumer Sampling – Procedure

Using the backpack from Figure 1, each consumer was directed to drive to the gas station of their choice (unknown to study staff), open the sampling line and cap, exit their vehicle, and pump gas as they normally would, enter their vehicle when finished, and then close the sampling line and cap. Participants were also instructed to remain near their vehicles while filling. The sampling backpacks were returned to study staff within 24 hours. Upon return of each backpack the canisters were checked to verify use; no other assessment of protocol adherence was performed. Personal sampling was conducted from August 2019 through March 2020, with sampling stopping due to COVID-19 related shutdowns.

3.3. Consumer Sampling – Analysis

Following collection, all canisters were measured for final pressure and returned to Chubb Environmental Health Laboratory (Cromwell, CT, USA). Canisters were analyzed for BTEX (benzene, toluene, ethylbenzene, xylene) via EPA TO-15 and TVOC (total volatile organic compounds) via NIOSH 1500. Any samples that were below limit of detection were assigned a value as the limit of detection divided by the square root of two (Hornung and Reed 1990).

3.4. Consumer Risk Assessment

Each consumer has a single benzene concentration for the recorded length of their fill-up based on the sampling results. Additionally, each consumer provided the number of times per month they typically fill up their vehicle from the questionnaires. Using the EPA's most conservative unit risk value for benzene inhalation carcinogenicity of 2.2×10^{-6}, the excess risk per million

people can be calculated following the standard EPA and NIOSH approach in Eq. 1a (EPA 1987; Integrated Risk Information System (IRIS) 2000; NIOSH 2017).

Eq. 1a: General non-occupational inhalation cancer risk assessment equation

$$Excess\ Risk_{per\ 1M} = 1{,}000{,}000 * UR * \left(\frac{CA * ET * EF * ED}{AT}\right)$$

From Eq. 1a, UR is the unit risk, CA is the benzene concentration, ET is the exposure time per day based on the length of time of fill-up, EF is the exposure frequency based on fill-ups per year, ED is the exposure duration of fifteen years, and AT is the averaging time of a lifetime of 70 years. Fifteen years was chosen for the exposure duration based on evidence in the literature that benzene exposures are causative of AML only at an approximate 10-20 year latency, and that exposures that occurred more than 20 years prior have no influence on the likelihood of developing leukemia (Finkelstein 2000; Galbraith et al. 2010; Glass et al. 2004; Richardson; Rinsky et al. 2002). Excess risk values that exceed 1:1,000,000 would be considered an unacceptable risk for consumers, a non-occupational group (EPA 2014; NIOSH 2017). However, in order to utilize all the collected data and expand the possible combinations of exposure and risk values, a probabilistic Monte Carlo risk approach will be utilized as recommended by NIOSH and the EPA for conducting risk assessments (Daniels et al. 2020; EPA 2014). Benzene concentrations were log-transformed and parameterized into $N(x, \sigma)_{Log(Benzene)}$. To determine the duration the canister was active, the flow regulator (constant rate) and canister had a maximum fill time of six minutes, corresponding to 5 ppm of vacuum decrease per minute of active sampling. Using the initial canister vacuum, subtracting the final canister vacuum, and then dividing by five produced an approximate sampling, or fill-up time. Fill times were parametrized into a truncated Normal distribution (Ntrunc) with a minimum

of 0.5 minutes, maximum of six minutes, and mean and standard deviation based on calculated fill times per consumer and then converted into $Ntrunc(x, \sigma, min = 0.5, max = 6)_{Minutes/Fill}$. Fill-ups per month were parameterized into a positive Poisson distribution of count data as $Pois(\lambda)_{Fill\ Days/Month}$, with λ as the mean of the fill-up counts. The full list of parameters for the consumer risk assessment are provided in Table 1.

Table 1: Parameters for probabilistic consumer risk assessment

Name	Variable	Value
Benzene Concentration	CA	$N(x, \sigma)_{Log(Benzene)}$
Exposure Time	ET	$Ntrunc(x, \sigma, min = 0.5, max = 6)_{Minutes/Fill}$
Exposure Frequency	EF	$Pois(\lambda)_{Fill\ Days/Month}$
Exposure Duration	ED	15 years
Averaging Time	AT	70 years
Unit Risk	UR	2.2×10^{-6}

In order to generate risk values for the consumer population, Eq. 1b, the probabilistic version of Eq. 1a, will be run 100,000 times with each iteration sampling from the distributions provided in Table 1.

Eq. 1b: Consumer inhalation cancer risk assessment equation with probabilistic components provided as their distributions

$$Excess\ Risk_{per\ 1M} = 1{,}000{,}000 * UR * \left(\frac{N(x,\sigma)_{Log(Benzene)} * (Ntrunc(x,\sigma,min=0.5,max=6)_{Min/Fill})/60 * (12 * Pois(\lambda)_{Fill\ Days/Month}) * ED}{AT}\right)$$

Following the resampling using Eq. 1b and the values from Table 2, percentiles of risk for the consumer population will be calculated from the resulting distribution of excess risk values.

3.5. Occupational Risk Assessment

While no direct occupational samples were collected, an occupational exposure scenario was considered using the near-pump concentrations from the consumer data and using a similar probabilistic methodology with appropriate exposure factors for an occupational setting. The consumer exposure concentrations of $N(x, \sigma)_{Log(Benzene)}$ will be reused directly for the occupational scenario whereas exposure time and exposure frequency will be new for the occupational exposure scenario. Exposure time (length of exposure per day) will be assumed to be normally distributed with a mean of seven hours and standard deviation of 0.5 hours, $N(7, 0.5)_{Hours/Day}$, with the expectation that this is a conservative estimate as it is possible that employees are not actively pumping gasoline an entire workday. Lastly, the exposure frequency (days exposed per year) will be drawn from a normal distribution with a mean of 260 days and a standard deviation of ten days, $N(260, 10)_{Days/Year}$, based on the 260-262 work days in a calendar year and the possibility an employee works more or less based on their personal situation (Office of Personnel Management 2011). Furthermore, the occupational risk assessment will be conducted with an excess risk management limit of 1:10,000 (NIOSH 2017). The parameters for the occupational risk assessment are shown in Table 2.

Table 2: Parameters for probabilistic occupational risk assessment

Name	Variable	Value
Benzene Concentration	CA	$N(x,\sigma)_{Log(Benzene)}$
Exposure Time	ET	$N(7,0.5)_{Work\ Hours/Day}$
Exposure Frequency	EF	$N(260,10)_{Work\ Days/Year}$
Exposure Duration	ED	15 years
Averaging Time	AT	70 years
Unit Risk	UR	$2.2x10^{-6}$

Again using 100,000 iterations, Eq. 1c and the values from Table 2 will be used to create the distribution of excess risk values for the occupational scenario.

Eq. 1c: Occupational inhalation cancer risk assessment equation with probabilistic components provided as their distributions

$$Excess\ Risk_{per\ 10K} = 10{,}000 * UR * \left(\frac{N(x,\sigma)_{Log(Benzene)} * N(8,1)_{Work\ Hours/Day} * N(260,10)_{Work\ Days/Year} * ED}{AT}\right)$$

3.6. National Air Toxics Assessment Risk Context

After the risk assessments have been completed for both the consumer and occupational groups, they will be contextualized with NATA provided excess cancer risk from ambient benzene concentrations that are provided on the census tract level (EPA 2018b). However, in order to expand the contextualization for consumers beyond just the specifics of the study population, a probabilistic Monte Carlo approach will be used that takes into account the

possibility of a consumer living and working in any area of Baltimore City or Baltimore County. Over the course of 100,000 iterations, two random census tract NATA excess risk values will be drawn from Baltimore City or Baltimore County, with one being assigned as the home tract with a weight of 0.8 and the other a work tract with a weight of 0.2, based on an approximate 40-hour work week. The census tracts will be weighted by population for the home tract, so a tract with a higher population is more likely to be selected than a less populated tract. The home and work values will be averaged according to their 0.8 and 0.2 weights. Each census tract pair's averaged NATA excess risk value will then be divided by a random draw from the consumer pumping risk distribution to create a distribution of ratios that compare consumer gasoline pumping risk to ambient risk. Additionally, the same process will be conducted for the occupational groups, where the average NATA excess risk value will be divided by a random draw from the occupational risk distribution. Because NATA excess cancer risk is reported as 1:1,000,000, the occupational pumping risk distribution will be converted to 1:1,000,000 to allow direct comparison.

3.7. Software

All data analysis and visualization was performed in R 4.0.2 "Taking Off Again" (R Core Team 2020).

4. Results

4.1. Consumer Sampling

From August 2019 through March 2020, 34 consumer samplers were collected. The temperatures during sampling ranged from -1.7°C to 33.9°C with a mean of 19.7°C and a median of 22.5°C. Two samples were not used in the exposure and risk assessment process. Consumer 29 had a canister leak in transport and was entirely discarded. Consumer 8's

canister was intact, but the reported sampling results were one to three orders of magnitude above the remaining samples. We believe this is the result of potential equipment malfunctions for Consumers 1-10 using inlet valves with loose gaskets that could require the consumer to touch the inlet orifice to move the gasket out of the way or even return the gasket after it fell out of the equipment entirely into the hand or onto the ground. Given the potential for gasoline contamination on hands while pumping or on the ground in the vicinity of a gasoline pump, Consumer 8's results will not be used for the remainder of the risk assessment. Additionally, the remainder of samples for Consumers 1-10 fell within plausible boundaries of exposure and were retained. Violin plots for the sampling distributions for BTEX and TVOC are provided in Figure 2, where plot width indicates density of samples and the height indicates concentration.

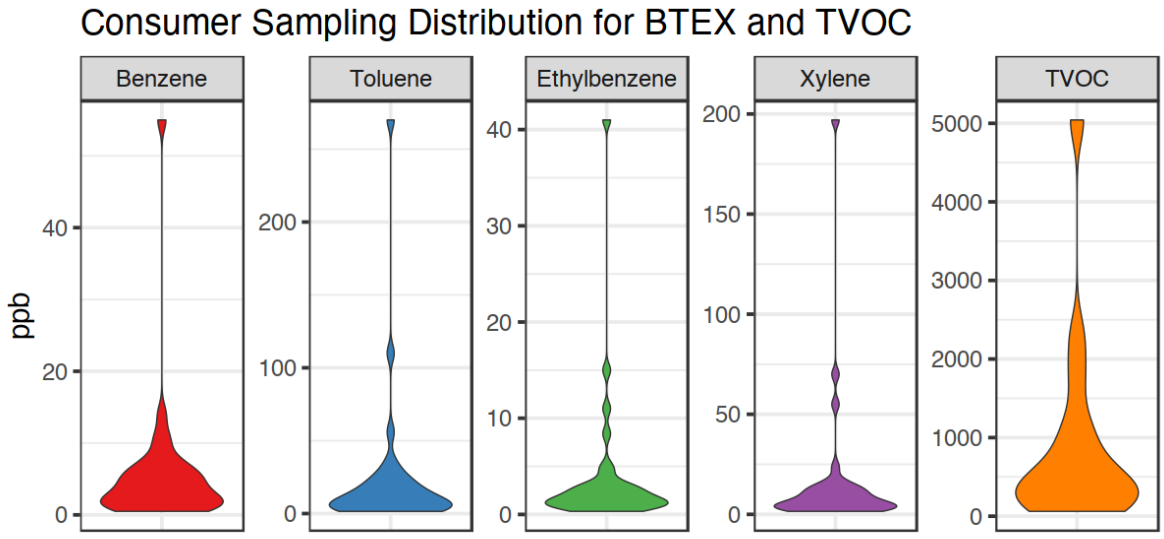


Figure 2: BTEX and TVOC concentration distributions from consumer sampling (n=32)

Despite sampling for benzene, toluene, ethylbenzene, and xylene, only the benzene concentrations were used in the risk assessment. IARC considers xylene and toluene to be Group 3, or not classifiable to human carcinogenicity, and neither xylene nor toluene have

inhalation unit risk values necessary to conduct an inhalation risk assessment (International Agency for Research on Cancer 1999a, 1999b; NIOSH 2019d, 2019c). While ethylbenzene is classified as Group 2B, or possibly carcinogenic to humans, the relevant exposure limits provided by OSHA (PEL 100 ppm) and NIOSH (REL 100 ppm) as well as epidemiological studies in ethylbenzene workers indicate that ethylbenzene's potential carcinogenic effects would require many orders of magnitude higher levels of exposure than seen in this chapter to produce excess risk (International Agency for Research on Cancer 2000; NIOSH 2019b).

Of the 32 valid samples, 31 were below the benzene NIOSH REL of 0.1 ppm (100 ppb) and 32 were below the OSHA PEL of 1.0 ppm (1000 ppb) (NIOSH 2019a). All samples were below the RELs and PELs for toluene, ethyl-benzene, and xylene (NIOSH 2019c, 2019d, 2019b). Both the REL and PEL for benzene are 8-hr time weighted averages (TWAs), whereas the samples are short term task lengths of less than six minutes. Therefore, the consumer samples are less than six-minute exposures at concentrations that NIOSH deems health protective for eight hours of exposure. The fully parameterized versions of the distributions introduced in Table 1 are presented below in Table 3 based on the results of the sampling. Plots of all distributions used in the risk assessments are presented in the Appendix.

Table 3: Fully parameterized distributions for probabilistic consumer risk assessment

Name	Variable	Value
Benzene Concentration	CA	$N(-5.73, 0.98)_{Log(Benzene)}$
Exposure Time	ET	$Ntrunc(3.08, 1.56\,, min = 0.5, max = 6)_{Minutes/Fill}$
Exposure Frequency	EF	$Pois(2)_{Fill\ Days/Month}$

4.2. Consumer Risk Assessment

Following the probabilistic risk assessment for consumers, zero percent of the 100,000 simulations exceeded the excess risk management level of 1:1,000,000. The full distribution is show in Figure 3a, where 1:1,000,000 is denoted by the vertical line at zero, or the base ten log of one. The 50th percentile of the consumer excess risk distribution was -2.8, the 75th was -2.5, and the 95th was -2.0. Therefore, the 95th percentile of risk was approximately 100 times lower than the excess risk management limit.

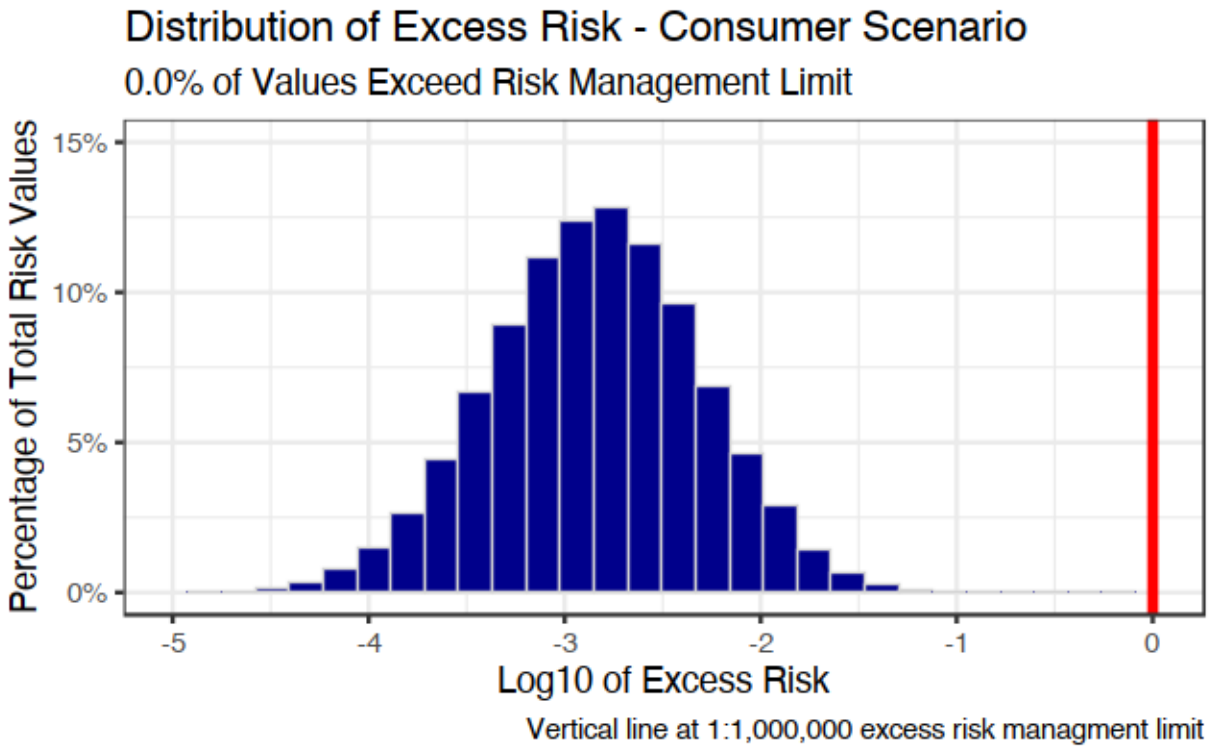


Figure 3a: Distribution of excess benzene related cancer risk from gasoline station pumping for consumers

4.3. Occupational Risk Assessment

The distribution of excess benzene related cancer risk from gasoline station pumping for occupational groups is shown in Figure 4b. The distribution exceeded the excess risk management limit of 1:10,000 on less than 0.01 percent of 100,000 trials, with a 50th percentile of -1.6, 75th percentile of -1.3, and a 95th percentile of -0.9. Therefore, the 95th percentile of occupational excess risk is approximately ten times less than the relevant 1:10,000 risk management limit. The entire excess risk distribution is presented in Figure 3b.

Figure 3b: Distribution of excess benzene related cancer risk from gasoline station pumping for occupational groups

4.4. National Air Toxics Assessment Risk Context

The results of the NATA ratio Monte Carlo are presented in Figures 4a and 4b, where both are on the scale of 1:1,000,000. Base ten log ratios greater than zero indicate that the pumping risk for consumer or occupational groups exceeds that of the NATA excess risk. For the consumers (Figure 4a), zero percent of the simulations exceeded zero. The 50th percentile of the ratio distribution was -3.44, the 75th percentile was -3.08, and the 95th percentile was -2.57, indicating that NATA excess risk was predominantly between two and three orders of magnitude larger than the excess risk from gasoline pumping alone. For the occupational group (Figure 4b), the 50th percentile was -0.24, the 75th percentile was 0.05, and the 95th percentile was 0.47. Based on the increased exposure duration and frequency, the log base ten ratio distribution for the occupational group exceeded zero on 28.9% of the simulations.

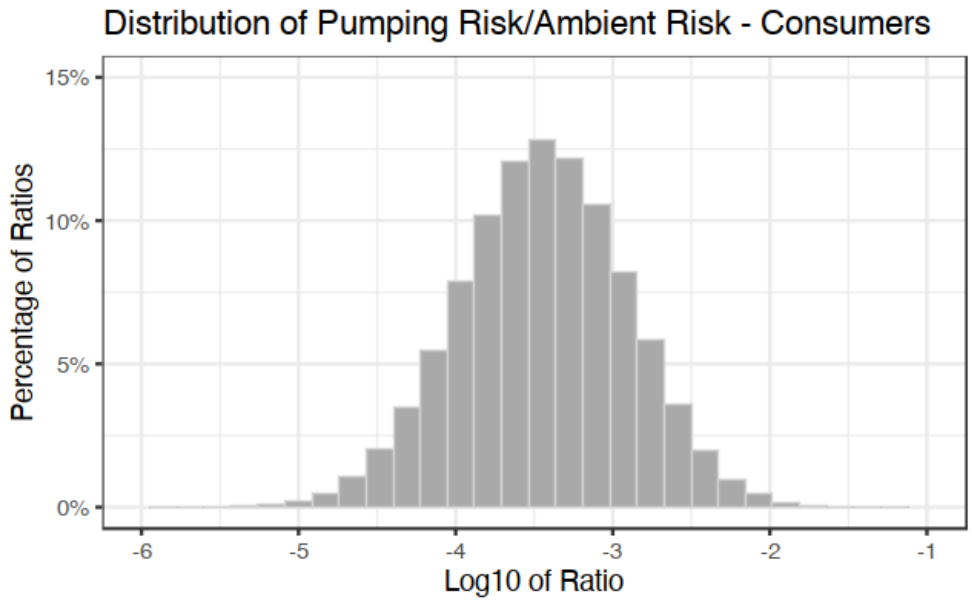


Figure 4a: Distribution of the ratio of gasoline pumping excess cancer risk from benzene exposure to excess cancer risk from ambient benzene exposure for consumers

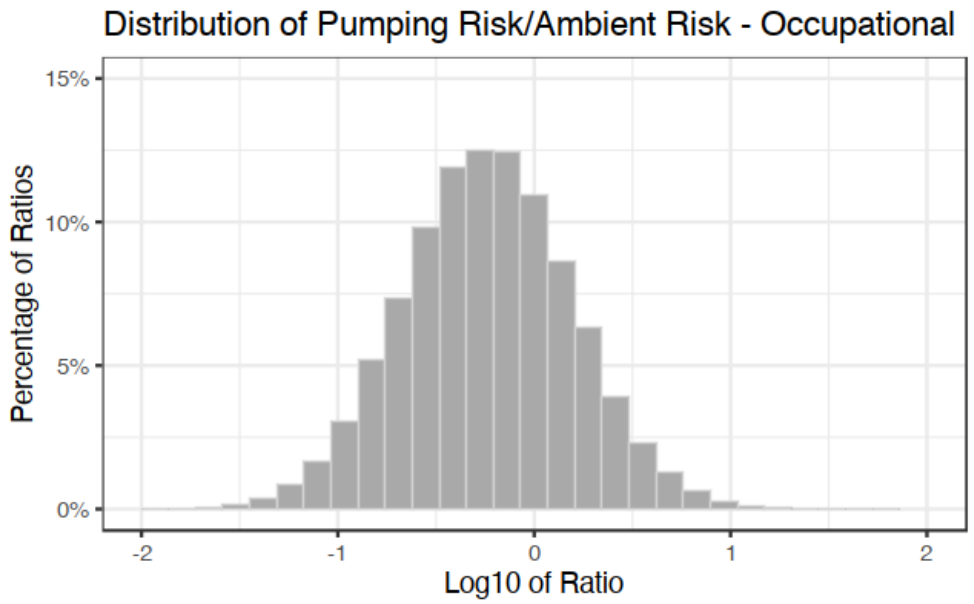


Figure 4b: Distribution of the ratio of gasoline pumping excess cancer risk from benzene exposure to excess cancer risk from ambient benzene exposure for occupational groups

5. Discussion

Millions of individuals per day are exposed to benzene, a known carcinogen, at commercial gasoline stations. In the United States, it has been nearly 30 years since a comprehensive evaluation of these exposures was conducted (Verma et al. 2001). Previous approaches used inconsistent or out of date sampling methodologies and were conducted via simulation studies. In addition, the results were often poorly documented and do not necessarily hold relevance based on the changes to fuel and fuel delivery technology (Alternative Fuels Data Center 2017; ATSDR 1996, 2006; IARC 1989). To address these data gaps and scientific challenges we implemented a novel self-sampling protocol that allowed consumers to perform their fill-ups as normally as possible, while collecting high-quality exposure data. The strengths of this protocol were that the consumers were likely to fill-up at a gas station they normally use, at a usual time, and in a more standard manner compared with a situation where the consumers were directed

to a set study site and observed. The intention was to capture the variability of possible exposure concentrations, and the self-sampling protocol was employed for that reason.

The exposure results for the consumers showed that 32 of the 33 viable samples for benzene were below the NIOSH Recommended Exposure Level of 100 ppb and 33 were below the OSHA Permissible Exposure Level of 1 ppm. While these are occupational standards, the REL is intended to be generally health protective (NIOSH 2018). In terms of the risk assessment, zero percent of the consumer risk distribution exceeded the 1:1,000,000 excess risk management limit. Using study participant specific data, 32 of 32 valid participants were between two and four orders of magnitude below the 1:1,000,000 excess risk limit. The occupational distribution had 0.006 percent of the risk values exceed the 1:10,000 excess risk management limit. Therefore, even with conservative assumptions for the occupational cohort, the excess risk management limit was not reached. These risk values for both the consumers and the occupational workers indicate that, in general when considering strictly pumping gasoline at commercial gasoline stations into automobiles, there is not an unacceptable cancer risk. It is important to note that these risk values do not take into account additional hazardous exposures that are possible at gasoline stations, particularly for an occupational cohort, such as sustained elevated $PM_{2.5}$ exposures from traffic or diesel exhaust fumes. Additionally, this risk assessment is explicitly only for exposures to benzene related to commercial gasoline station fill-ups and does not include other potential sources such as smoking cigarettes (Integrated Risk Information System (IRIS) 2000).

When contextualized with excess cancer risk from Baltimore City and Baltimore County ambient benzene concentrations (NATA), the consumer risk distribution did not exceed the NATA values, whereas approximately 29 percent of the occupational excess risk distribution did

exceed the NATA values, indicating that more cancer risk was due to occupational exposure than from ambient exposures. However, Baltimore City (mean = 4.4/1M, std = 0.31/1M) and Baltimore County (mean = 3.76/1M, std = 0.39/1M) are both in the 95th percentile of NATA excess risk from benzene nationwide. Therefore, in counties with less ambient benzene exposure, gasoline pumping could make up a larger percentage of an individual's total excess benzene risk than what was presented in Figure 4a and 4b. Additionally, the lack of spatial data associated with the gas stations (proximity to roadways, traffic patterns, etc.) does not allow for further characterization of station specific exposures.

The two main limitations of the exposure and risk assessment are the difficulty in verifying that the sampling protocol was followed by the participants and the likely low variability in sampling locations, population, and season. While COVID-19, reduced the intended sample size of 100 to 34 and the study duration to August 2019 through March 2020, there were still only plans to sample Baltimore area consumers. It is possible that additional exposure data outside of the existing distributions would be captured with additional multi-state and multi-season sampling.

6. Conclusion

Based on the results of the exposure assessment and the risk assessment, excess cancer risks from benzene exposures due to fuel pumping are low for both consumers and workers. In the context of Baltimore and other urban areas where excess cancer risk from benzene in ambient air is higher than the 1:1,000,000 excess risk management limit for the general population, consumer risks from re-fueling are very low. However, the upper 29 percent of the excess risk distribution for the worker scenario was equivalent or slightly higher than the ambient benzene risk but still below the NIOSH risk management limit. Additionally, the use of a novel self-sampling protocol for consumers allowed for a unique exposure assessment on an understudied

population that previously relied entirely on simulation studies. The use of whole air sampling means that the protocol can be reused for a range of chemical exposures of concern and could easily be extended to a longer duration task.

7. References

Alternative Fuels Data Center. 2017. Ethanol Blends. Available: https://www.afdc.energy.gov/fuels/ethanol_blends.html.

ATSDR. 1996. Automotive Gasoline.

ATSDR. 2006. Toxicological Profile for Benzene.

ATSDR. 1995. Toxicological Profile for Gasoline.

Bergsagel DE, Wong O, Bergsagel PL, Alexanian R, Anderson K, Kyle RA, et al. 1999. Benzene and Multiple Myeloma: Appraisal of the Scientific Evidence. Blood 94.

BLS. 2019. Industries at a Glance: Gasoline Stations: NAICS 447. Ind a Glance. Available: https://www.bls.gov/iag/tgs/iag447.htm [accessed 6 October 2020].

Bruckner J, Anand S, Warren A. 2008. Toxic Effects of Solvents and Vapors. In: *Casarett & Doull's Toxicology: The Basic Science of Poisons* (C. Klaasen, ed). McGraw Hill. 981–1051.

Daniels R, Gilbert S, Kuppusamy S, Kuempel E, Park R, Pandalai S, et al. 2020. Current Intelligence Bulletin 69 - NIOSH Practices in Occupational Risk Assessment.

Environmental Protection Agency. 2012. Gasoline Mobile Source Air Toxics.

EPA. 2018a. 2014 NATA Summary of Results.

EPA. 2018b. 2014 National Air Toxics Assessment. Available: https://www.epa.gov/national-air-toxics-assessment/2014-national-air-toxics-assessment.

EPA. 2014. Risk Assessment Forum White Paper: Probabilistic Risk Assessment Methods and Case Studies. Available: https://www.epa.gov/sites/production/files/2014-12/documents/raf-

pra-white-paper-final.pdf.

EPA. 1987. The Risk Assessment Guidelines of 1986.

Finkelstein MM. 2000. Leukemia after exposure to benzene: temporal trends and implications for standards. Am J Ind Med 38:1–7; doi:10.1002/1097-0274(200007)38:1<1::AID-AJIM1>3.0.CO;2-9.

Galbraith D, Gross SA, Paustenbach D. 2010. Benzene and human health: A historical review and appraisal of associations with various diseases. Crit Rev Toxicol 40:1–46; doi:10.3109/10408444.2010.508162.

Glass DC, Sim MR, Fritschi L, Gray CN, Jolley DJ, Gibbons C. 2004. Leukemia risk and relevant benzene exposure period?Re: Follow-up time on risk estimates, Am J Ind Med 42:481-489, 2002. Am J Ind Med 45:222–223; doi:10.1002/ajim.10327.

Hornung RW, Reed LD. 1990. Estimation of Average Concentration in the Presence of Nondetectable Values. Appl Occup Environ Hyg 5:46–51; doi:10.1080/1047322X.1990.10389587.

IARC. 1989. Gasoline. Monographs 45.

Integrated Risk Information System (IRIS). 2000. Benzene - Carcinogenicity Assessment for Lifetime Exposure.

International Agency for Research on Cancer. 2000. Ethylbenzene. IARC Summ Eval 77.

International Agency for Research on Cancer. 1999a. Toluene. IARC Summ Eval 71.

International Agency for Research on Cancer. 1999b. Xylenes. IARC Summ Eval 71.

IRIS. 2003. Benzene. Chem Assess Summ.

Keenan JJ, Gaffney S, Gross SA, Ronk CJ, Paustenbach DJ, Galbraith D, et al. 2013. An evidence-based analysis of epidemiologic associations between lymphatic and hematopoietic cancers and occupational exposure to gasoline. Hum Exp Toxicol 32:1007–1027; doi:10.1177/0960327113476909.

NACS | Convenience Stores Fuel America. Available: https://www.convenience.org/Topics/Fuels/Fueling-America [accessed 5 January 2020].

National Association of Convenience Stores. 2013. Fueling America: A Snapshot of Key Facts and Figures.

National Association of Convenience Stores. 2019. NACS | Selling America's Fuel. Available: https://www.convenience.org/Topics/Fuels/Who-Sells-Americas-Fuel [accessed 5 January 2020].

NIOSH. 2019a. CDC - NIOSH Pocket Guide to Chemical Hazards - Benzene. NIOSH Pocket Guid to Chem Hazards. Available: https://www.cdc.gov/niosh/npg/npgd0049.html [accessed 1 March 2020].

NIOSH. 2019b. CDC - NIOSH Pocket Guide to Chemical Hazards - Ethyl benzene. NIOSH Pocket Guid to Chem Hazards. Available: https://www.cdc.gov/niosh/npg/npgd0264.html [accessed 14 September 2020].

NIOSH. 2019c. CDC - NIOSH Pocket Guide to Chemical Hazards - o-Xylene. NIOSH Pocket Guid to Chem Hazards. Available: https://www.cdc.gov/niosh/npg/npgd0668.html [accessed 14 September 2020].

NIOSH. 2019d. CDC - NIOSH Pocket Guide to Chemical Hazards - Toluene. NIOSH Pocket Guid to Chem Hazards. Available: https://www.cdc.gov/niosh/npg/npgd0619.html [accessed 14 September 2020].

NIOSH. 2017. How NIOSH Conducts Risk Assessments. Available: https://www.cdc.gov/niosh/topics/riskassessment/how.html.

NIOSH. 2018. NIOSH Potential Occupational Carcinogens | NIOSH | CDC. Available: https://www.cdc.gov/niosh/npg/nengapdxa.html [accessed 14 September 2020].

Nobile T. 2018. NJ coronavirus could have residents pumping their own gas. NorthJersey.com, April 17.

Northeast States for Coordinated Air Use Management. 1985. Evaluation of the health effects from exposure to gasoline and gasoline vapors. Unpubl Work.

Office of Personnel Management. 2011. Fact Sheet: Computing Hourly Rates of Pay Using the 2,087-Hour Divisor. Policy, Data, Overs Pay Leave.

OSHA. 2020. OSHA Annotated Table Z-1. Available: https://www.osha.gov/dsg/annotated-pels/tablez-1.html [accessed 20 October 2020].

Page NP, Mehlman M. 1989. Health Effects of Gasoline Refueling Vapors and Measured Exposures At Service Stations. Toxicol Ind Health 5:869–890; doi:10.1177/074823378900500521.

Paxton MB, Chinchilli VM, Brett SM, Rodricks J V. 1994a. Leukemia risk associated with benzene exposure in the pliofilm cohort: I. Mortality update and exposure distribution. Risk Anal 14:147–154; doi:10.1111/j.1539-6924.1994.tb00039.x.

Paxton MB, Chinchilli VM, Brett SM, Rodricks J V. 1994b. Leukemia risk associated with benzene exposure in the pliofilm cohort. II. Risk estimates. Risk Anal 14:155–161; doi:10.1111/j.1539-6924.1994.tb00040.x.

Periago JF, Prado C. 2005. Evolution of Occupational Exposure to Environmental Levels of Aromatic Hydrocarbons in Service Stations. Ann Occup Hyg 49:233–240; doi:10.1093/annhyg/meh083.

R Core Team. 2020. R: A Language and Environment for Statistical Computing.

Richardson DB. Temporal Variation in the Association between Benzene and Leukemia Mortality. Environ Health Perspect 116:370–374; doi:10.2307/40040154.

Rinsky RA, Hornung RW, Silver SR, Tseng CY. 2002. Benzene exposure and hematopoietic mortality: A long-term epidemiologic risk assessment. Am J Ind Med 42:474–480; doi:10.1002/ajim.10138.

Verma D, Johnson D, McLean J. 2000. Benzene and total hydrocarbon exposures in the

upstream petroleum oil and gas industry. AIHAJ - Am Ind Hyg Assoc 61: 255–263.

Verma DK, Johnson DM, Shaw ML, des Tombe K. 2001. Benzene and Total Hydrocarbons Exposures in the Downstream Petroleum Industries. AIHAJ - Am Ind Hyg Assoc 62:176–194; doi:10.1080/15298660108984621.

Weller C. 2018. Oregon now requires people to pump their own gas — and some New Jerseyans are freaking out. Business Insider, January 12.

Wong O. 1995. Risk of acute myeloid leukaemia and multiple myeloma in workers exposed to benzene. Occup Environ Med 52:380–384; doi:10.1136/oem.52.6.380.

8. Appendix

The following figures show the distributions used in the consumer and occupational risk assessments. These distributions were based on sampling information, professional judgement, and the literature (Office of Personnel Management 2011)

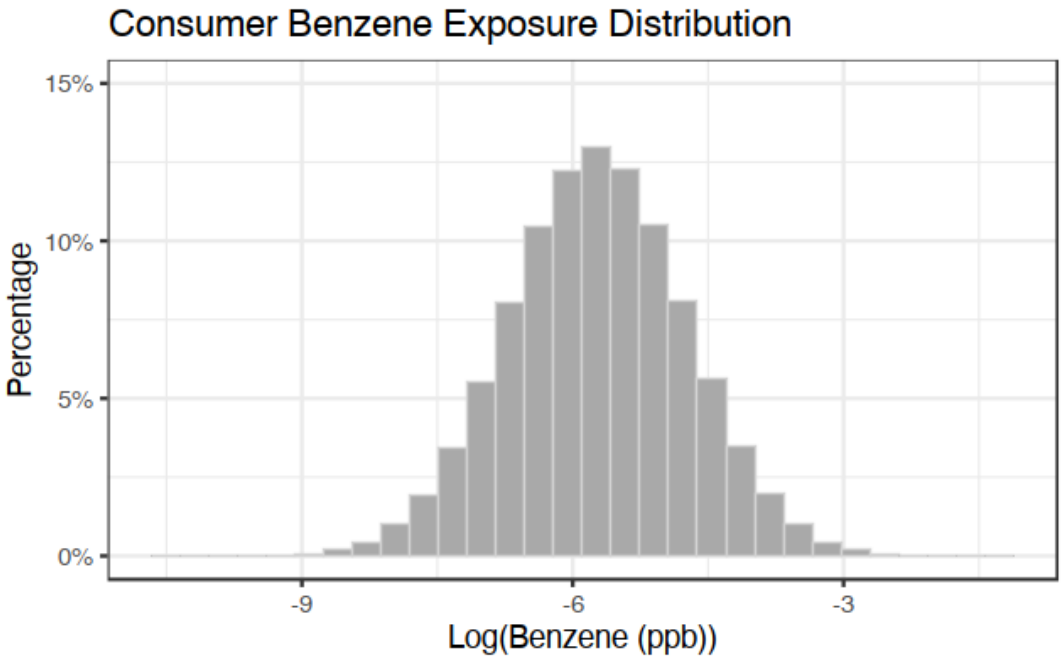


Figure A1: Distribution of consumer benzene exposures on log-ppb scale

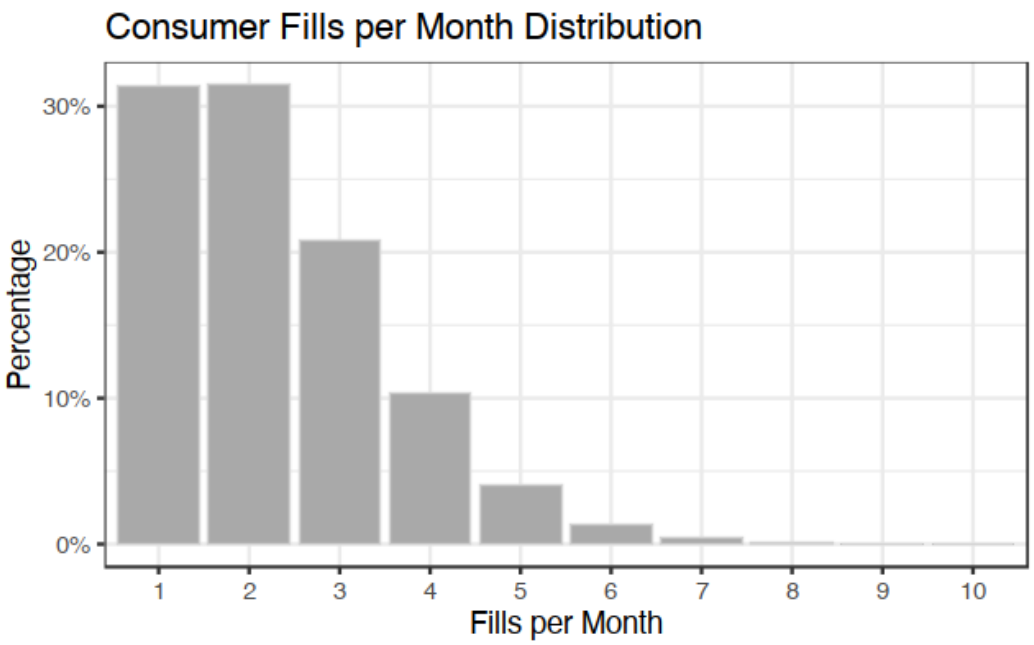


Figure A2: Distribution of consumer fill events per month

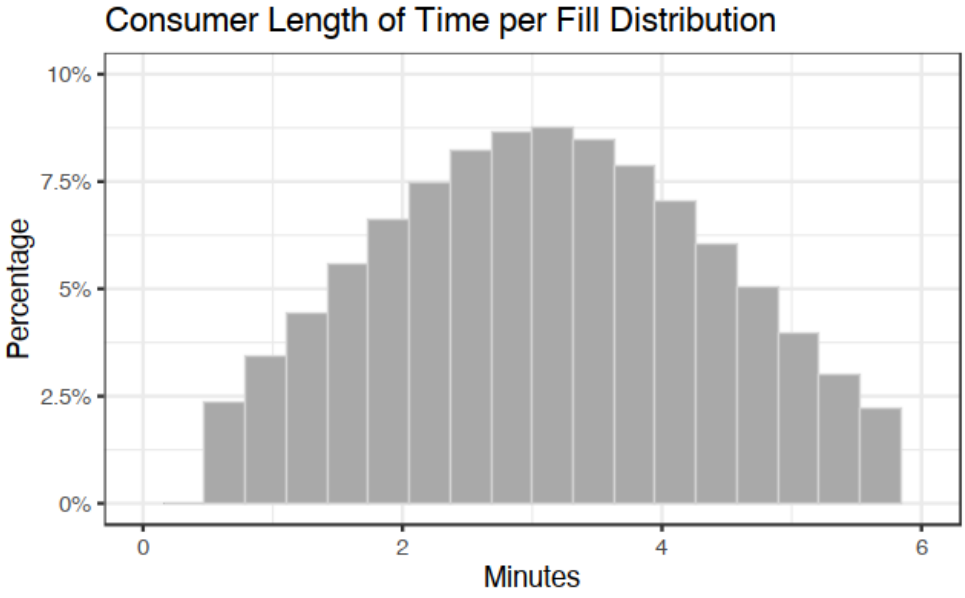


Figure A3: Distribution of consumer fill event durations

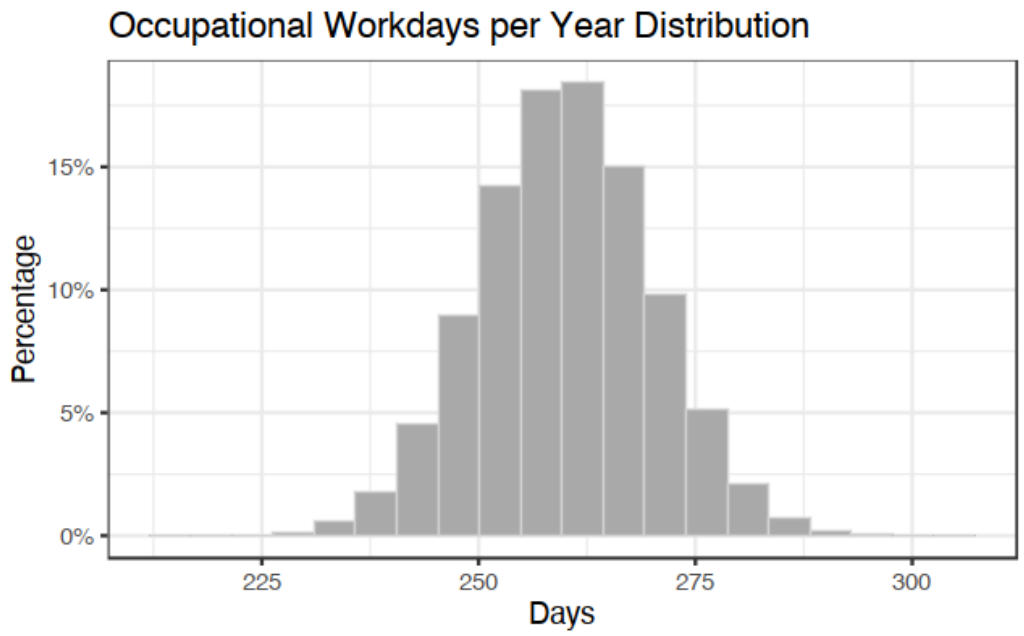


Figure A4: Distribution of occupational days exposed per year

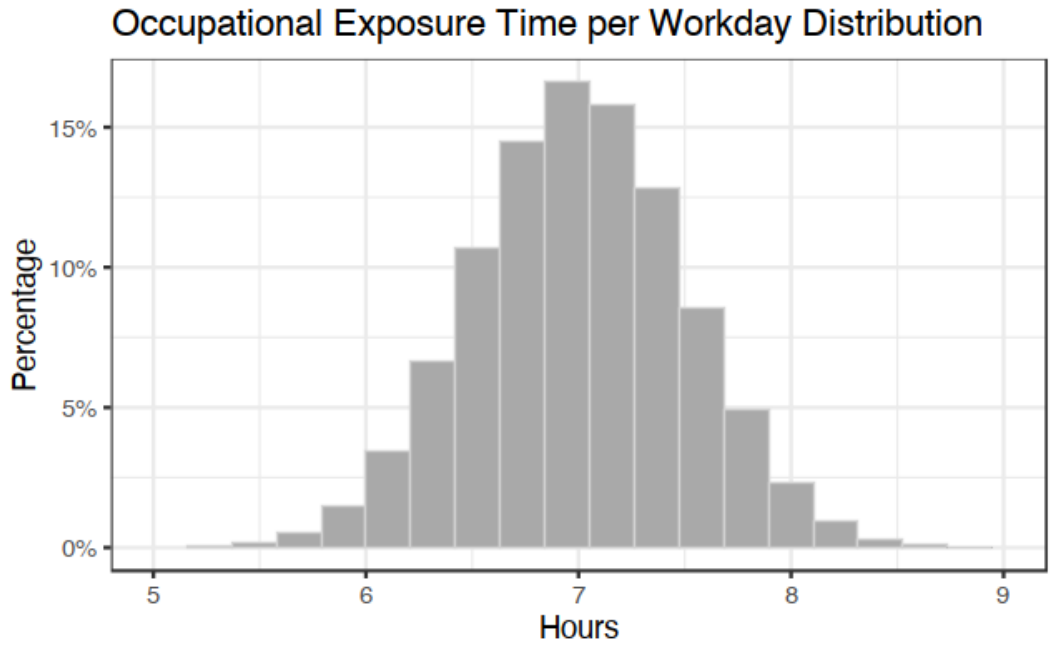


Figure A5: Distribution of occupational hours exposed per workday

CHAPTER III: MACHINE LEARNING FOR IMPROVING ACCURACY AND UTILITY OF LOW-COST AIR POLLUTION SENSOR NETWORKS FOR PROBABILISTIC SPATIAL EXPOSURE ASSESSMENT

1. Abstract

Low-cost sensor networks for monitoring air pollution are an effective tool for expanding spatial resolution beyond the capabilities of existing state and federal reference monitoring stations. Linear regression-based field calibration using co-located reference data is often used to improve low-cost monitor data quality. However, low-cost sensor data commonly exhibit highly non-linear biases with respect to climate conditions that cannot be captured by linear models. Hence, extensive lab calibrations are often carried out prior to field-deployment of these sensors, converting raw sensor readings to viable concentrations which are then subsequently input to the field calibration models. Using data from Plantower A003 $PM_{2.5}$ sensors deployed in Baltimore, MD from November 2018 through November 2019, we demonstrate that direct field-calibration of the raw low-cost sensor data using probabilistic gradient boosted decision trees (GBDT) can circumvent this two-step process and resource-intensive lab calibration, while also improving the point and distribution accuracy of the linear model. We provide a framework for utilizing the GBDT to conduct probabilistic spatial assessments of human exposure with inverse distance weighting (IDW) that predicts the probability of a given location exceeding an exposure threshold and provides percentiles of exposure. These probabilistic spatial exposure assessments can be scaled to time and space with minimal modifications. Here, we used the probabilistic exposure assessment methodology to create high quality spatial-temporal $PM_{2.5}$ maps on the neighborhood-scale in Baltimore, MD.

2. Introduction

According to The World Health Organization (WHO), fine particulate matter ($PM_{2.5}$) is responsible for approximately 7 million premature mortalities per year (World Health Organization 2018). Within the United States, 88,000 annual deaths are attributed to $PM_{2.5}$ exposure (Cohen et al. 2017). Further, $PM_{2.5}$ is considered a Group 1 carcinogen according to

the International Agency for Research on Cancer (IARC) (International Agency for Research on Cancer 2016). The WHO estimates that air pollution is responsible for approximately one quarter of all adult mortalities and leads to heart disease, stroke, and lung cancer (World Health Organization 2018). Given that one prominent source of $PM_{2.5}$ is combustion, concentrations are often highest in densely-populated urban areas with a higher levels of car/truck traffic and fuel combustion at power plants or on more localized scales leading to the potential for variability in $PM_{2.5}$ concentrations over small spatial scales. (International Agency for Research on Cancer 2016).

Within the United States, $PM_{2.5}$ concentrations are required to meet the primary and secondary National Ambient Air Quality Standards (NAAQS) established by the Environmental Protection Agency (EPA) via the Clean Air Act (Environmental Protection Agency 2010). For $PM_{2.5}$, the current primary, or health protective standard, is an annual mean of 12 $\mu g/m^3$ averaged over three years and the secondary, or infrastructure protective standard, is an annual mean of 15 $\mu g/m^3$ averaged over three years (Environmental Protection Agency 2010). In comparison, the WHO recommends that in order to reduce morbidity and mortality, $PM_{2.5}$ annual ambient concentrations should not exceed 10 $\mu g/m^3$ (World Health Organization 2018). Additionally, the EPA also provides a combined primary/secondary standard of a 24-hour 98^{th} percentile of 35 $\mu g/m^3$ averaged over three years (Environmental Protection Agency 2010).

In order to ensure that the air quality meets the NAAQS standards, the EPA requires that states operate monitoring sites with high quality sampling equipment that meets a Federal Reference Method (FRM) or Federal Equivalent Method (FEM) in major urban areas. However, there are only 935 $PM_{2.5}$ monitors to cover the entirety of the United States, and of the 25 most populous urbanized areas with a total population of 111 million people, there are only 282 $PM_{2.5}$ monitors

(Apte et al. 2017). Therefore, the spatial resolution on high quality $PM_{2.5}$ data can be severely lacking for major urban areas. For example, in Baltimore City, there is only a single FEM monitor administered by the Maryland Department of the Environment located near the geographic center of Baltimore City (Maryland Department of the Environment 2018). This is highly relevant as intra-city air pollution exposure ranges have been proposed to be as large or larger than exposure ranges between cities (Ye et al. 2020). Additionally, Saha et al. (2020) indicate that based on multi-site $PM_{2.5}$ monitoring and modeling in Pittsburgh, PA, there is substantial spatial variation in $PM_{2.5}$ concentrations within a city on a 1-4 km spatial scale (Saha et al. 2020). In addition to spatial resolution concerns, gravimetric methods in use at certain FRM stations requires 24-hr sampling, sacrificing the ability to measure $PM_{2.5}$ on shorter timescales (Maryland Department of the Environment 2018).

In order to fill the spatiotemporal data gaps in air pollution monitoring, low-cost sensor networks have been developed and deployed. These networks can consist of many types of sensors that, while less accurate than reference monitors, provide the ability to produce high resolution spatial and temporal measurements relevant at the urban level. (Piedrahita et al. 2014; Snyder et al. 2013; Szpiro et al. 2009). One example of a low-cost sensor network is the Solutions for Energy, Air, Climate, and Health (SEARCH) Center's investigation into neighborhood-level variations of air pollutant concentrations in Baltimore, MD which has been operational since December 2018 (Buehler et al. 2020; Datta et al. 2020). The SEARCH network encompasses low-cost monitors spread across the city measuring PM (mass and number concentrations), ozone, nitric oxide, nitrogen dioxide, carbon monoxide, carbon dioxide, methane, relative humidity, and temperature (Buehler et al. 2020). However, the reduction in precision compared with an FEM measurement adds complexity to the monitoring such that utilization of the raw sensor data is discouraged without accounting for environmental biases (Buehler et al. 2020;

Morawska et al. 2018). Therefore, in order to gather sensor data that is both accurate and precise, a combination of field and lab calibration is often recommended to ensure the sensor data is reliable (Borrego et al. 2018; Levy Zamora et al. 2019). Lab calibration is both labor intensive and requires laboratory facilities, which is not an option for all low-cost sensor network administrators. However, with the presence of an FEM monitor as gold standard, co-locating one or more network sensors with the FEM monitor allows for the creation of models that use raw sensor readings to accurately predict 'gold standard' or reference values.

Various strategies have been developed to optimize the efficacy of co-locations between low-cost sensor networks and reference monitors. Studies in California and China have co-located sensors with reference monitors at the start and end of the sampling period (Gao et al. 2015; Mukherjee et al. 2019). Others have conducted extensive lab calibration followed by rule based filtering or bias correction methods (Heimann et al. 2015; Mead et al. 2013). In addition, a variety of modeling approaches for training the calibration/modeling to reference data have been conducted, including polynomial regression, gain/offset linear regression, land use regression, and machine learning methods like gradient boosted decision trees (GBDT) and random forests (Johnson et al. 2018; Lim et al. 2019).

However, the existing approaches to model reference values from sensor measurements have not been evaluated at the spatial and temporal scales observed within the SEARCH network, and also only produce point estimates and/or confidence intervals for predictions. While point estimates and confidence intervals are useful, a full unique distribution for each prediction is a much more versatile model output. For example, the model uncertainty with the distribution can be used to answer questions such as what percentage of time the monitor reports values greater than a certain threshold. Linear regression only models the mean but not the shape or

spread of the data distribution. Many machine learning techniques allow for complex non-linear modeling of the mean. However, not all machine learning techniques provide any estimate of uncertainty around the prediction, let alone a full probability distribution. From a more applied sense, probabilistic modelling can be used in a probabilistic exposure assessment and risk assessment framework as suggested by EPA and NIOSH, both of which recommend simulations to more accurately characterize the full distribution of possible exposure and/or risk (Daniels et al. 2020; EPA 2014; NIOSH 2017).

We propose using probabilistic machine learning for calibration of low-cost $PM_{2.5}$ sensors, where both the spread and mean of the response is modeled with GBDT, resulting in several gains upon a traditional linear approach. First, like many other machine learning methods, GBDT can model non-linearity of the calibration equation with respect to temperature and relative humidity at a minimum. This is important as intermediate lab-corrections of the raw data using meteorological variables are often highly non-linear (Levy Zamora et al. 2019). Hence, linear field calibration models (like Datta et al. 2020) rely on these lab-corrections for the non-linear effects. Second, the GBDT proposed here directly used raw sensor data to predict reference concentrations. By utilizing raw data, this method removes the need for the lab calibration of the sensors. Additionally, we will show that the GBDT have more predictive accuracy than previously developed linear regression models for the same task. Further, we use the GBDT modeling results to conduct probabilistic spatial exposure assessments based on a Monte Carlo spatial interpolation. Finally, we aim to use the modeling and Monte Carlo approach to create highly customizable exposure assessments. For example, exposures can be aggregated by time and place in such a manner that the probabilistic nature of the exposure assessment is retained. These exposure assessments provide health relevant characterization of possible $PM_{2.5}$ exposures as opposed to a simple deterministic option.

3. Methods

3.1. Reference Data

There is one FEM monitoring site located in Oldtown in central Baltimore, and another in Essex on the eastern border of the city. Oldtown lies within the city limits of Baltimore in an area with high traffic density, whereas Essex is outside of the city limits and is within Baltimore County, the county surrounding Baltimore city. The Oldtown site measures $PM_{2.5}$ on an hourly basis using FEM Beta Attenuation, and the Essex site measures daily average $PM_{2.5}$ once every six days using a gravimetric FRM (Maryland Department of the Environment 2018). Both sites are operated by the Maryland Department of the Environment (MDE) (Maryland Department of the Environment 2018).

3.2. SEARCH Data

The SEARCH data in this chapter consists of hourly $PM_{2.5}$ measurements from November 2018 through November 2019 taken by 34 separate monitors. Each of the 34 monitors deployed in the network contains a Plantower A003 optical $PM_{2.5}$ sensor as well as a variety of other sensors for gaseous pollutants. Additionally, each monitor has a built-in temperature and relative humidity sensor. Each monitor contains both internal memory storage and a wireless cellular connection via a SIM card that uploads data to a remote server every ten seconds. The locations of the deployed monitors were chosen based on spatial and environmental factors as well as willingness of a property owner to host the monitor. The network has been online since October 2018. The locations of the SEARCH network monitors, Oldtown FEM Monitor (centrally located) and Essex FRM Monitor (eastern coast) are presented in Figure 1.

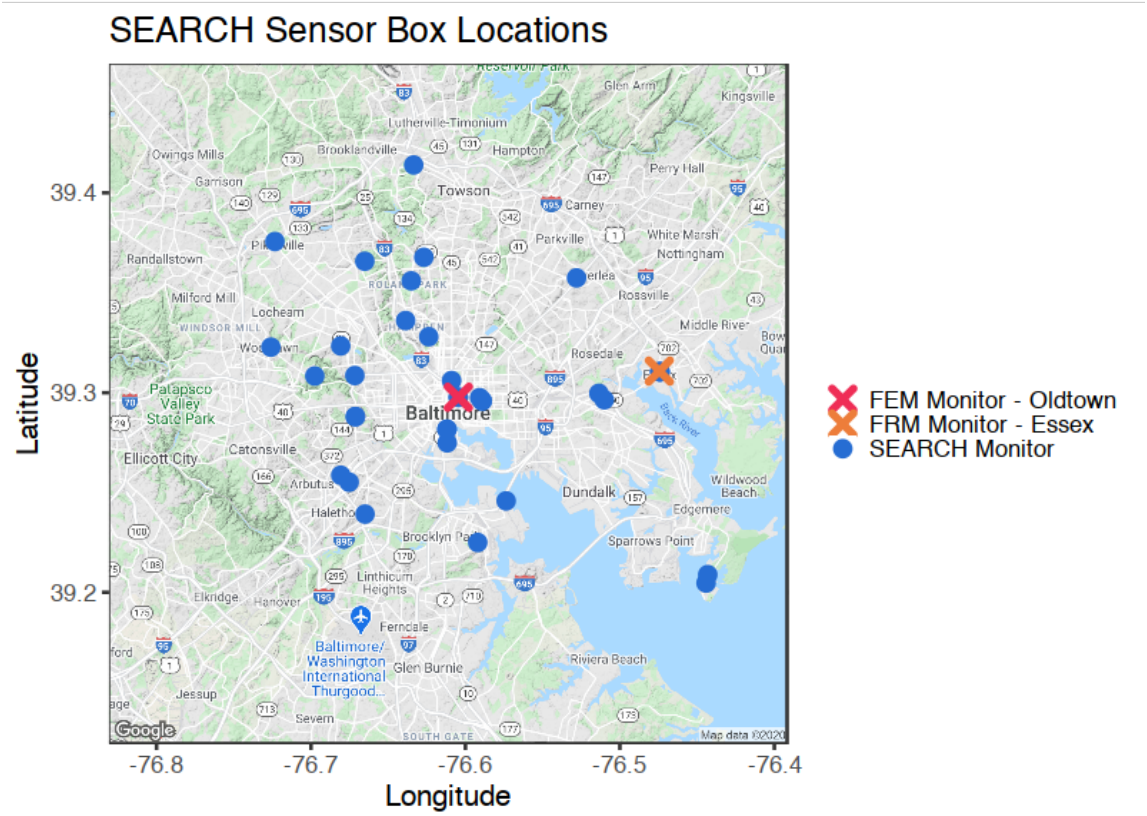


Figure 1: Map of Baltimore and Baltimore County SEARCH network monitors and FEM monitoring sites

There are two SEARCH monitors (B25 and B33) that were co-located with the Oldtown monitor and three monitors (B62, B21, and B8) that were co-located with the Essex monitor (monitor identification numbers are not indicative of the total number of monitors). However, only two of B62, B21, and B8 were ever active at one time. B25 and B33 were deployed from December 2018 through October 2019, and B61, B21, and B8 from February 2019 to August 2019. These monitors will serve as basis for the analysis for the remainder of the study.

3.3. Modeling

Three separate models were used to model sensor data to reference data. Two models are GBDT approaches that use raw sensor readings to model to reference data. In order to demonstrate the efficacy of GBDT, they were benchmarked against the linear regression approach conducted by Datta et al. (2020) that used lab-corrected sensor data.

3.3.1. Gradient Boosted Decision Trees

Gradient boosted decision trees (GBDT) are a powerful modeling tool for both regression and classification that excels on non-linear data (Friedman 2001). They dominate machine learning competitions for tabular data (as opposed to computer vison, language data, etc.) and work well on relatively small data sets out-of-the-box (Chen and Guestrin 2016; Duan et al. 2019). In contrast, deep learning techniques such as neural networks often require substantially more data and technical knowledge to fit and do not, as of yet, perform as well as GBDT on tabular data (Duan et al. 2019). GBDTs work by combining numerous weak learner decision trees additively in a forward stepwise pattern (Kuhn and Johnson 2013; Schapire 2003). Each weak learner is a slight improvement on a random guess, and as they are combined, the residual from each step is computed and used to generate the next weak learner (Kuhn and Johnson 2013). A variety of regularization approaches are used to avoid models that overfit to the data (Kuhn and Johnson 2013).

3.3.1.1. Light Gradient Boosting Machine (LightGBM)

One of the most widely used GBDT algorithms is LightGBM, created and developed as an open source project by Microsoft (Guolin et al. 2017). It can be considered one of the gold standard public libraries for GBDT alongside XGBoost and CatBoost (Chen and Guestrin 2016;

Dorogush et al. 2017). LightGBM models provide point predictions without estimates of uncertainty.

3.3.1.2. Natural Gradient Boosting (NGBoost)

NGBoost is a novel probabilistic gradient boosting framework – effectively serving as a wrapper that can boost a variety of base learning models such as decision trees, random forest, etc. (Duan et al. 2019). This chapter will use an NGBoost regression on a decision tree, resulting in a GBDT (all packages and libraries provided in Appendix A) (Pedregosa et al. 2011). NGBoost is a probabilistic modeling tool and returns a point prediction, (in this case a mean), and form of variance depending on the output distribution chosen. This variance is modeled for each individual prediction and is not a uniform value. While NGBoost does allow for a variety of custom distributions (Laplace, Log-Normal, etc.), a normal distribution was selected for this chapter based on the shape of the input data and the flexibility it provides. Therefore, the point prediction and mean of the normal distribution output are identical.

3.4. Linear Regression Baseline Model

Based off lab calibration work on the SEARCH monitors conducted by Levy Zamora et al. (2019), calibrations equations were developed to convert raw sensor readings to lab-corrected data. Using that lab-corrected data, a gain/offset linear regression was developed by Datta et al. (2020) to model to the reference data. The linear regression equation is shown in Eq. 1.

Eq. 1: Linear regression for $PM_{2.5}$ calibration to reference monitors using laboratory corrected sensor readings

$$PM_{2.5}Reference = \beta_0 + \beta_1 * RH + \beta_2 * T + \beta_3 * daytime + \beta_4 * weekend + \beta_5 * RH * PM_{2.5}Sensor + \beta_6 * T * PM_{2.5}Sensor + \beta_7 * daytime * PM_{2.5}Sensor + \beta_8 * weekend * PM_{2.5}Sensor$$

The covariates of this model are the lab-corrected low-cost sensor measurement ($PM_{2.5}$-Sensor), relative humidity (RH), temperature (T), a binary flag for between 7am and 5pm (daytime), and a binary flag for weekend (weekend). $PM_{2.5}$-Reference refers to the $PM_{2.5}$ as measured by a reference monitor. The model was trained on data where $PM_{2.5}$-Reference was the measurements from the Oldtown MDE reference monitor, and $PM_{2.5}$-Sensor was lab-corrected measurements from either monitors B25 or B33.

3.5. Model Features

In order to ensure a valid comparison between the existing linear regression and the GBDTs and demonstrate the efficacy of the GBDT with a small feature space, only five baseline features will be used in each model, the same ones present in Eq. 1: relative humidity (RH), temperature (T), daytime, weekend, and $PM_{2.5}$-Sensor. The primary difference between the linear regression and the GBDT is that linear regression used the lab-corrected $PM_{2.5}$ sensor data, whereas the GBDT used raw $PM_{2.5}$ sensor data. Additionally, whereas the linear regression explicitly includes interaction terms, the structures of decision trees include them implicitly.

3.6. Training and Testing Datasets

In order to compare and contrast results from the linear models using lab calibrated data presented by Datta et al. (2020) with the GBDTs, the time intervals for training and testing from the study will be duplicated with an additional 'monthly' interval added as well. The seven modeling splits are shown in Table 2. While the Full, Prospective, and three seasonal splits are intended to compare directly to Datta et al. (2020) and cover accuracy by season, the Essex split is intended to test the validity of the approach on fully out of sample data and monitors. The Monthly split is intended to test the model performance on small training set size as well as being able to capture some component of sensor drift.

Table 1: Training and testing splits for linear and GBDT models using Oldtown $PM_{2.5}$ data

Type	Training Set	Testing Set
Full	80% from 12/2018-11/2019	20% from 12/2018-11/2019
Prospective	all from 12/2018-7/2019	all from 8/2019-11/2019
Spring	80% from 3/2019-5/2019	20% from 3/2019-5/2019
Summer	80% from 6/2019-7/2019	20% from 6/2019-7/2019
Winter	80% from 12/2018-2/2019	20% from 12/2018-2/2019
Essex	100% Oldtown data from 12/2018-11/2019	reference MDE data from Essex (24hr averages) from 12/2018-11/2018
Monthly	each single month from 1/2019-10/2019	each subsequent month from 2/2019-11/2019

For example, in the full modeling split, the entire dataset of B25 and B33 data with corresponding Oldtown reference $PM_{2.5}$ (the target for the regressions) with be split 80% into a

training set and 20% into a test set.. For both NGBoost and LightGBM, 10% of the training data was set aside for each Full, Prospective, etc., for five-fold cross validation with early stopping during the tuning process to prevent overfitting. Hyperparameters, the rules by which the GBDTs learn and fit the data, were tuned using a random search over a grid of values (Pedregosa et al. 2011). The exact hyperparameter search space for each GBDT is provided in the Appendix A. It is important to note that the Essex split is trained on exclusively Oldtown data and tested on monitors co-located with the Essex MDE monitor. This entirely out of sample and separate monitor test is needed to ensure the validity of the methodology on monitors that were not used to build the model. However, given that the Essex monitor produces 24-hr average $PM_{2.5}$ concentrations, hourly predictions were made and then averaged up into a 24-hr prediction to allow for comparison. This averaging process will compress the values and, as such, the model evaluation metrics for the Essex split should not be directly compared to other train/test splits.

3.7. Model Evaluation

The primary method of evaluation of point predictions from the $PM_{2.5}$ models is root mean square error (RMSE) on the test set. RMSE is calculated as shown in Eq. 2. RMSE is defined as the square root of the mean of the squared residuals, and due to the squaring is always positive. Lower values of RMSE indicate more accurate predictions, and the values are on the scale of the predictors. The linear regressions, LightGBM, and NGBoost will all be evaluated on RMSE, shown in Eq. 2.

Eq. 2: Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Actual - Predicted)^2}{n}}$$

Actual is the measured $PM_{2.5}$ value by the MDE monitor, predicted is the model prediction, and n is the total number of samples under evaluation, in this application the number of hourly measurement and prediction pairs in the testing set. The second method of evaluation is the continuous ranked probability score (CRPS) and utilizes the probabilistic results of NGBoost and the confidence intervals of the linear regression. Probabilistic predictions provide more than just a point estimate, and therefore require evaluation of the spread around the point prediction as well as the point prediction itself. For linear regression, each prediction has the identical standard deviation around the mean, whereas NGBoost produces a unique standard deviation for each prediction based on the learned training data. CRPS will only be used to evaluate NGBoost and the linear regression as LightGBM does not provide intervals in its regression predictions. Similar to RMSE, CRPS is in the same units as the predicted variable, with smaller values indicating more accuracy. It can be considered a generalization of mean absolute error for probabilistic predictions (lower values indicate more accuracy) and takes into account the spread and mean of each prediction distribution (Gneiting and Raftery 2007). CRPS values were calculated using the properscoring package made available by The Climate Corporation (Barrett et al. 2015).

3.8. Spatial Interpolation

In order to use the results of the model to conduct exposure assessments, the calibrated data needs to be spatially interpolated across Baltimore. However, for NGBoost the model predictions are not simply hourly point predictions at each SEARCH monitor, but a mean and standard deviation of a normal distribution referred to as $N(x, \sigma)_{NGBoost}$. Therefore, a resampling process using the distributions as part of the interpolation process was conducted.

3.8.1. Inverse Distance Weighting

Inverse distance weighting (IDW) was used as the interpolation method. IDW operates under the assumption that locations in close proximity are more likely to have similar measurements than those further away, and that the weight of each known measurement in predicting at a location is inversely related to how far away the two are. The general formula for IDW is shown in Eq. 3 with d as distance between the interpolation location and the measured value, i is a sampled location, z_i the value at the sampled location i, and n as the total number of points used in the averaging.

Eq. 3: Inverse Distance Weighting (IDW)

$$z_{estimated} = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} z_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

The power parameter, p, is used to control the strength of the inverse distance relationship. For larger values of p, more distant measurements are devalued, whereas p = 0 corresponds to a straight average across all monitors. While the default selection for p is often 2.0, leave one out cross validation (LOOCV) on the SEARCH monitor predictions was conducted to ensure that the power parameter was selected properly and to ensure that interpolation error is propagated through the exposure assessment. IDW was conducted on a 256x256 grid using the spatstat package's IDW function in R (Baddeley et al. 2015).

The first step in the IDW LOOCV was to find 50 random hours where each monitor was active with at least eight other monitors active as well. Next, for one monitor at a time, that monitor was excluded (i.e. left-out) from one of the 50 hours and an IDW was conducted five times with power parameters equal to 1.0, 1.5, 2.0, 2.5, and 3.0. This was repeated so that each monitor resulted in 250 IDWs where it was not included in the interpolation. For each of the 250 IDWs, the prediction at the excluded monitor location was compared to the true value and residuals were calculated. After the LOOCV was complete, the residuals were grouped by power parameter and confirmed for normality via a Shapiro-Wilk normality test, and then parameterized into a single normal distribution by taking the mean and standard deviation as $N(0, \sigma)_{Error}$. The power parameter with the smallest standard deviation of residual was chosen as the optimal value.

3.8.2. IDW Monte Carlo

With the power parameter of the IDW selected, the next step is to conduct a Monte Carlo simulation using the IDW with the predicted distribution values from the NGBoost model. The Monte Carlo was chosen to run 250 simulations for each hour to balance runtime with accuracy. The IDW Monte Carlo was conducted using the following steps:

Step 1: Select a single one-hour slice of the data

Step 2: Produce a single draw from each SEARCH monitor's $N(x, \sigma)_{NGBoost}$ prediction from the hour selected

Step 3: Using the single draws from all available locations, conduct an IDW on a square 256x256 grid encompassing Baltimore city limits

Step 4: For each grid point sample from $N(0, \sigma)_{Error}$ and add to IDW predicted values

Step 5: Following the Monte Carlo, combine all 250 predictions per grid point to obtain the estimated concentration distributions

Following the Monte Carlo simulation, each grid cell's interpolated values were parameterized to a normal distribution following confirmation of normality via a Shapiro-Wilk normality test. Therefore, for each grid location results were recorded as $N(x, \sigma)_{IDW}$ based on the mean and standard deviation of the interpolated $PM_{2.5}$ values.

3.9. Exposure Assessment

The Monte Carlo results in each grid location by hour, having a predicted normal distribution of $PM_{2.5}$, is referred to as $N(x, \sigma)_{IDW}$. In order to aggregate an exposure assessment to administrative boundaries, the average of each $N(x, \sigma)_{IDW}$ within the borders of the administrative geometry was defined as the exposure for that zip code, Census Tract, neighborhood, etc. This exposure assessment can also be aggregated temporally, going from single hour bins to days, weeks, or months by averaging the $N(x, \sigma)_{IDW}$ for each hour bin up into the time units desired prior to spatial aggregation.

In order to demonstrate the probabilistic framework, three exposure metrics will be used in an example exposure assessment. The mean and 95th percentile prediction will be provided and are more conventional metrics. The third, is a threshold-based metric, offering the probability of exceeding a threshold for a given administrative region and time window. While there are monitors in many of the administrative regions that could theoretically provide single exposure values, the combination of multiple monitors will allow for a complete gradient across the area of interest that can be fit to any scale exposure assessment. Additionally, using multiple monitors for estimation increases the robustness of the estimate and includes information that

takes into account bordering regions. For $PM_{2.5}$, the example threshold was the primary EPA annual standard of 12 $\mu g/m^3$ (Environmental Protection Agency 2010). It is important to note that all three values are produced directly from the IDW results of one model. The exposure assessments were conducted on the Community Statistical Area (CSA) level, clusters of similar and known neighborhoods determined by the Baltimore City Planning Department (Baltimore City Department of Health 2017).

3.9.1. Exposure Variability

Exposure assessments were aggregated to daily, monthly, and three-month periods to evaluate the change in spatial variability with increasing length of aggregation time intervals. For each time period, starting on February 1, 2019, the standard deviation across all CSA predicted mean $PM_{2.5}$ exposures was calculated. The averages of the standard deviations were then compared across time aggregation intervals.

3.10. Software

All linear and GBDT modeling was conducted in Python 3.7.7 with spatial analysis and data visualization conducted in R 4.0.2 'Taking Off Again' (R Core Team 2020). The full list of modeling packages, libraries, and their version numbers is provided in Appendix B.

4. Results

4.1. Sensor Modeling

The predictions from the NGBoost model for the week of February 1, 2019 through February 7, 2019 are compared to the linear regression results and the corresponding MDE reference data in Figure 2a.

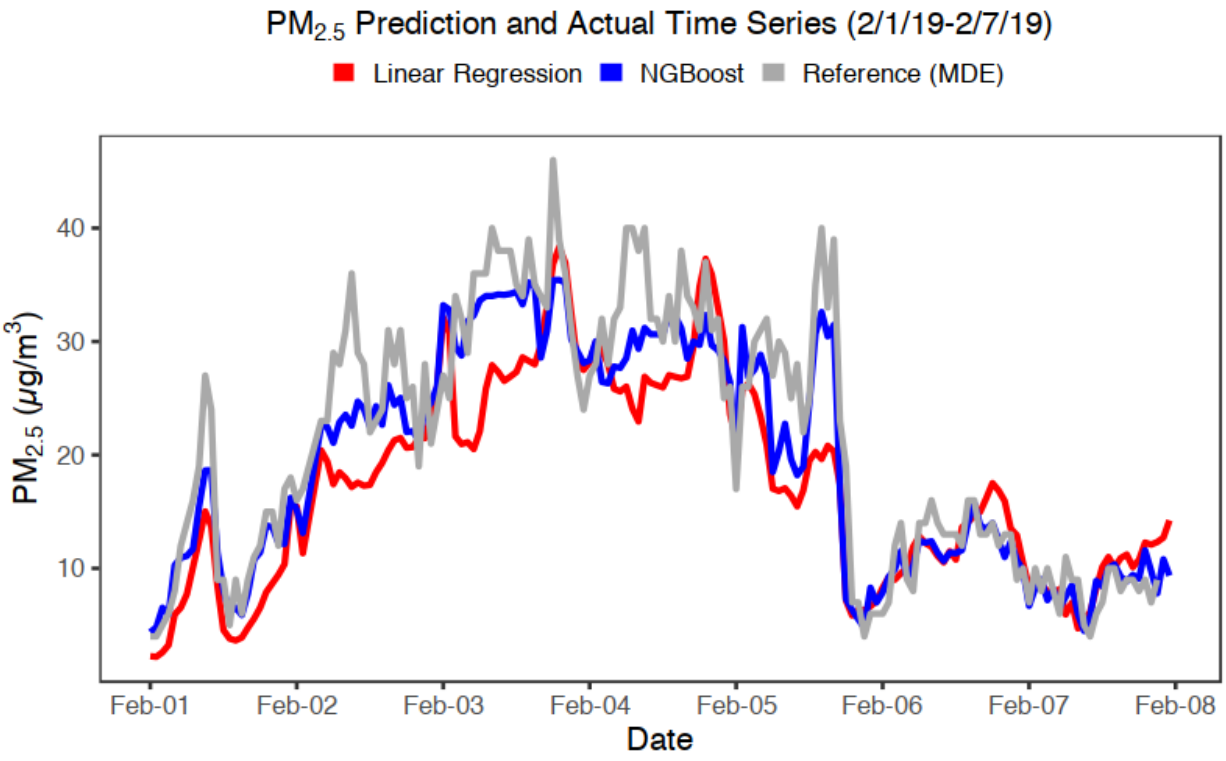


Figure 2a: Time series for the first week of February 2019 comparing linear regression and NGBoost predictions to the MDE reference standard

This week in February 2019 was an abnormally poor air quality event (mean MDE $PM_{2.5}$ during February of 8.6 $\mu g/m^3$). NGBoost generally did a better job of capturing high exposure, short duration events such as the highlighted week (e.g., Feb 05). Figure 2b shows a time series comparison for a week with more typical $PM_{2.5}$ concentrations in August 2019 (mean MDE $PM_{2.5}$ during August of 9.4 $\mu g/m^3$). Even at lower overall levels of $PM_{2.5}$, NGBoost was better able to capture short duration high and low concentrations better than the linear model. However, some peaks and troughs are missed by both models (e.g., Feb 02-03 and Aug 24-25).

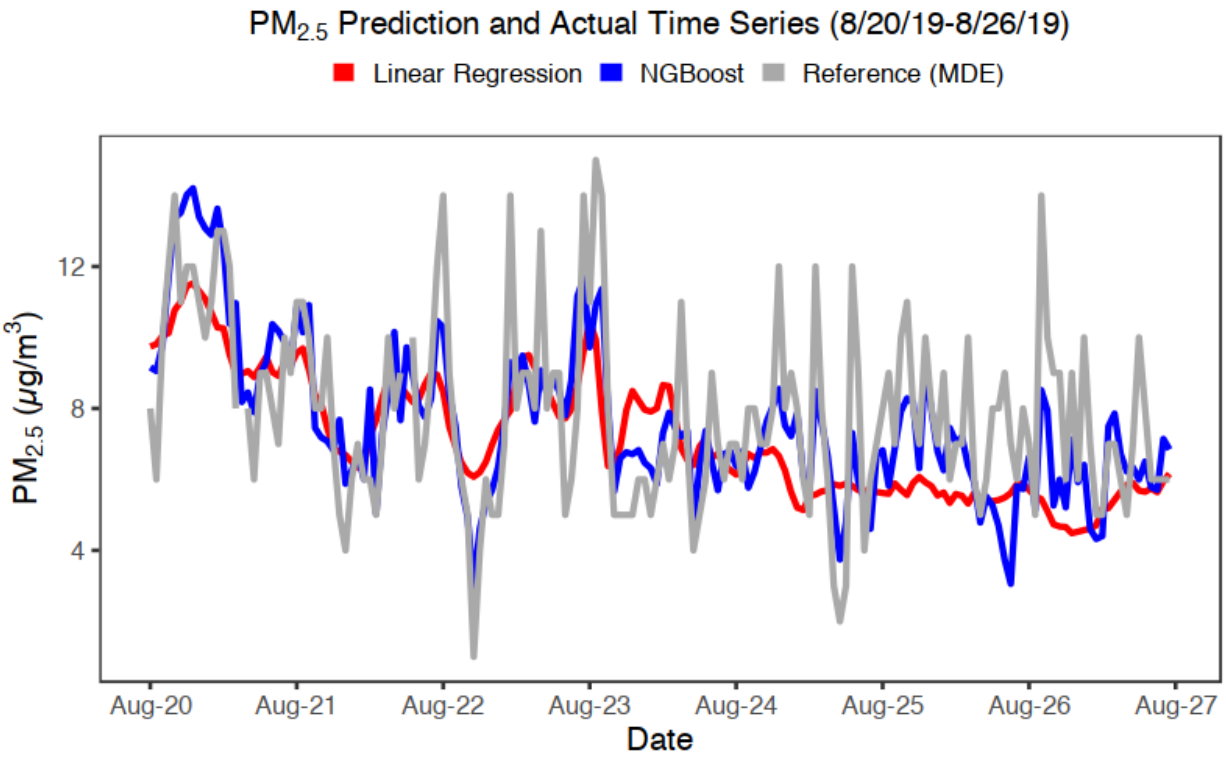


Figure 2b: Time series for the final week of August 2019 comparing linear regression and NGBoost predictions to the MDE reference standard

The model evaluation results are presented for RMSE in Table 2 and CRPS in Table 3. In general, both forms of GBDT and the linear regression are broadly comparable in terms of RMSE, although NGBoost performance is the best, with an average RMSE of 2.7 $\mu g/m^3$ compared to 2.9 $\mu g/m^3$ and 3.1 $\mu g/m^3$ for LightGBM and the linear regression, respectively, across all testing splits. NGBoost dramatically outperformed the linear regression in the winter split, with an RMSE of 2.9 $\mu g/m^3$ compared to 3.8 $\mu g/m^3$, a 30% increase in accuracy in that season. In terms of the probabilistic predictions, NGBoost also has at least a 30% decrease in average CRPS compared to the linear regressions, which takes into account both the spread and the mean of the prediction distribution. Therefore, the distribution spread for NGBoost is approximately one third more accurate than the spread for the linear regression model on a

per-prediction basis. These accuracy improvements are despite the fact that the GBDT (LightGBM and NGBoost) used the raw uncalibrated sensor data as opposed to the lab-corrected data used by the linear regression. Additionally, due to averaging across 24-hrs, the Essex evaluation results will have RMSEs that are far lower than those of the other evaluation splits. However, the improved accuracy of the GBDT compared with linear regression on the fully out of sample Essex data, demonstrates the transportability of the calibration approach to other monitors.

Table 2: Model Evaluation Results Comparing Linear Regression with GBDT Across Identical Training and Test Splits - RMSE

Type	**RMSE ($\mu g/m^3$)**		
	Linear Regression	**NGBoost**	**LightGBM**
Full	3.2	2.8	2.9
Prospective	2.9	2.9	2.9
Spring	2.6	2.4	2.4
Summer	2.8	2.6	2.7
Winter	3.8	2.9	3.5
Essex	2.2 (24hr)	2.0 (24hr)	2.1 (24hr)

Table 3: Model Evaluation Results Comparing Linear Regression with GBDT Across Identical Training and Test Splits – CRPS

Type	CRPS ($\mu g/m^3$)		
	Linear Regression	**NGBoost**	**LightGBM**
Full	2.3	1.5	-
Prospective	2.2	1.7	-
Spring	1.9	1.4	-
Summer	2.1	1.4	-
Winter	2.6	1.7	-

It is worth noting that while average performance was somewhat comparable, NGBoost's RMSE spread across all train/test splits was much tighter (2.4-2.9 $\mu g/m^3$) than the linear regression (2.6-3.8 $\mu g/m^3$) or LightGBM (2.4-3.5 $\mu g/m^3$).

In addition to the full, prospective, and three seasonal splits, a monthly split was also performed. Both models were trained on one month of data and then tested on the sequential months. For example, training on May and testing on June-November. The results from this analysis are shown in Figure 3.

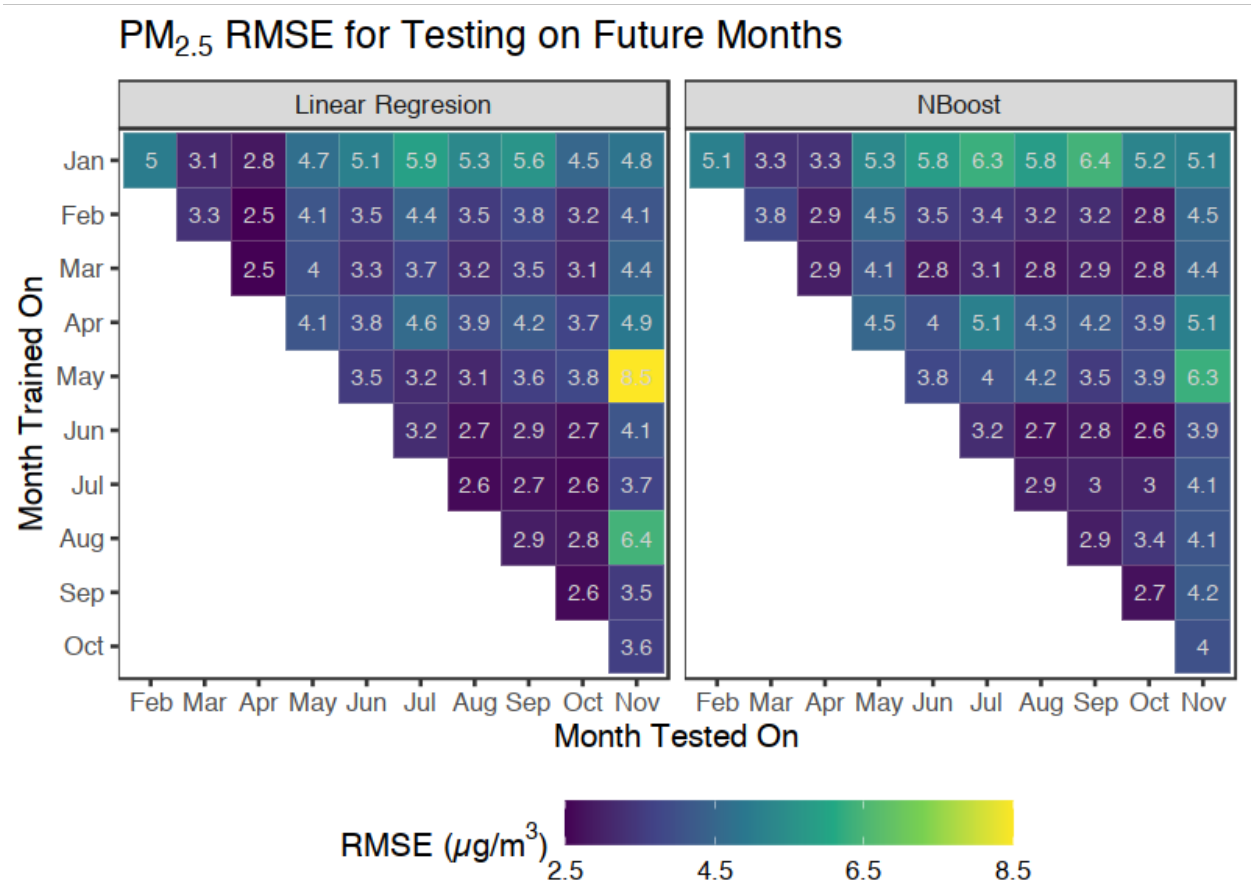


Figure 3: RMSE results from training on a single month and testing on subsequent months

In general, the performance capability for both models was acceptable for similar climate months in the near future (1-3 months in future from training month) with an NGBoost RMSE of 3.6 $\mu g/m^3$ and a linear regression RMSE of 3.4 $\mu g/m^3$. However, performance degraded sharply when predicting months 4-10 into the future, with an average NGBoost RMSE of 4.2 $\mu g/m^3$ and an average linear regression RMSE of 4.3 $\mu g/m^3$.

4.2. IDW Power Parameter Cross Validation

The LOOCV used to select the optimal power determined that the commonly used power value of 2.0 produced the residual distribution with the smallest standard deviation, although the

values across the search space of 1.0, 1.5, 2.0, 2.5, and 3.0 were within one percent of one another. The residual distribution was found to be $N(0, 1.8)_{IDW}$. Therefore, $N(0, 1.8)_{IDW}$ was added to each iteration of the IDW Monte Carlo to ensure that interpolation error is propagated through to the exposure assessment in addition to modeling error.

4.3. Exposure Assessment

IDW predictions were created for every hour from February 2019 through November 2019. An example of a single day exposure assessment on the CSA level within Baltimore city limits was conducted on June 5, 2019 to highlight the spatial variability observed using the network. June 5, 2019 had the largest difference in single day mean CSA predictions between any two monitors of a day with more than 20 monitors in operation of approximately 10.3 $\mu g/m^3$. Mean and 95th percentile $PM_{2.5}$ values are shown for June 5, 2019 in Figure 5a. For comparison, August 1, 2019, is shown in Figure 4b with a maximum CSA predicted difference of 3.5 $\mu g/m^3$. For both Figure 4a and 4b, although more pronounced in Figure 4a, there is a high concentration area in the center of the city, with additional high concentration areas in the northeast and northwest areas, likely corresponding to commuting traffic on major interstates I-83, which runs north-south through the center of the city and I-695 beltway that surrounds the city slightly outside the city limits (not shown).

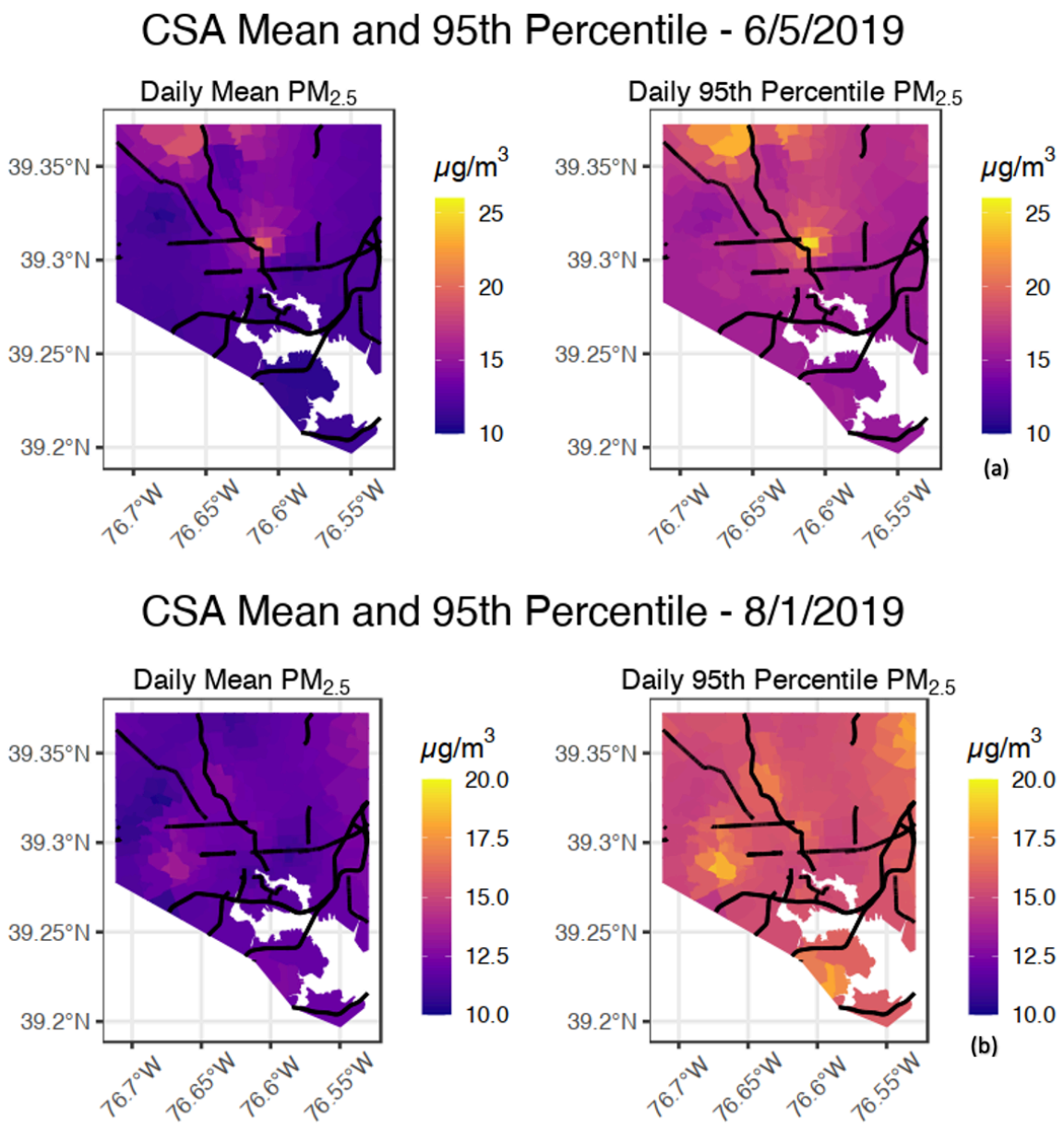


Figure 4a, 4b: Mean (left) and 95th Percentile (right) $PM_{2.5}$ exposures by Community Statistical Association (CSA) for June 5, 2019 and August 1, 2019 with major roads and highways in black.

While the mean and 95th percentile exposure are valuable metrics, they are easily obtainable from standard linear models. Therefore, the probabilistic nature of the NGBoost enables the

use of the model the mean and standard deviation for each prediction to be used to determine the probability of exceeding 12 $\mu g/m^3$ for the 24-hr period on June 5, 2019 as shown in Figure 5. The result is a powerful approach to assess spatially-resolved risk for exceeding threshold values in a complex urban landscape using low-cost distributed measurement networks.

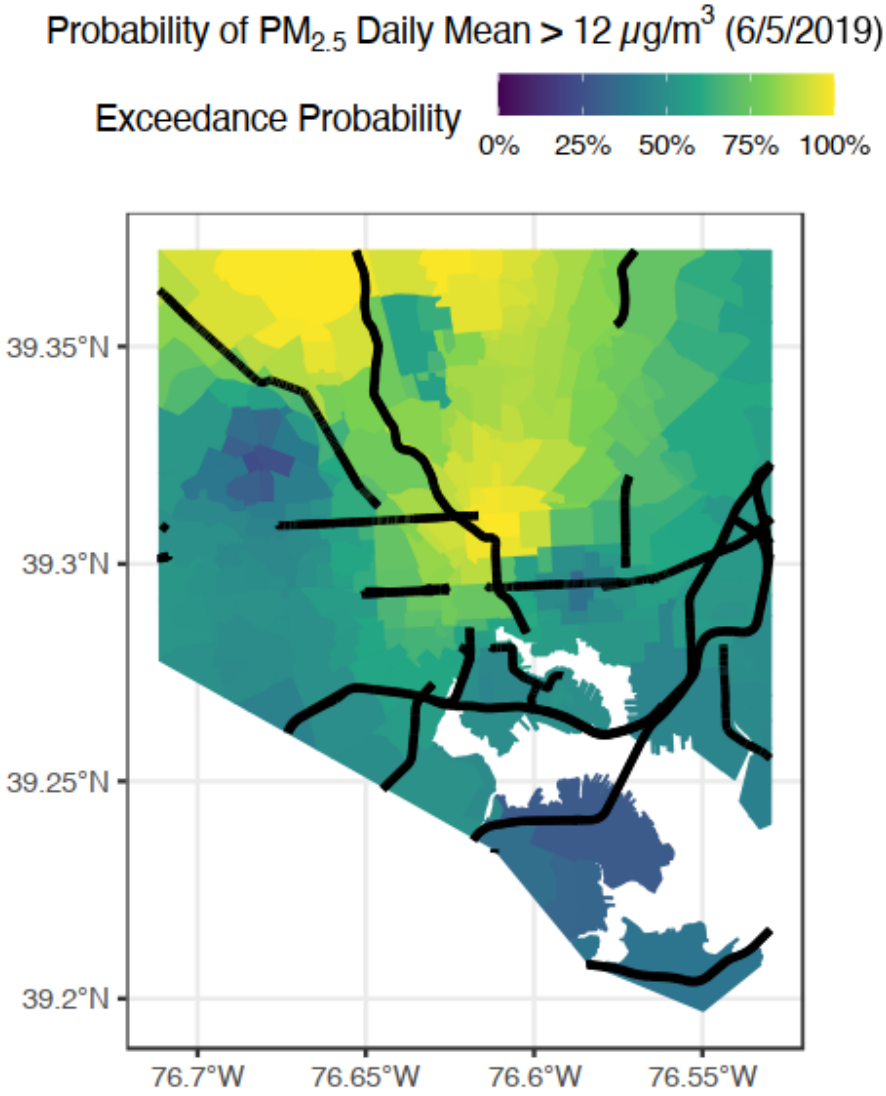


Figure 5: Probability of daily mean $PM_{2.5}$ exceeding 12 µg/m³ by CSA on June 5, 2019 with major highways and roads in black

Of the 278 CSAs in Baltimore on June 5, 2019, 158 had a greater than 50% chance of exceeding 12 $\mu g/m^3$ and 28 had a greater than 90% change of exceeding 12 $\mu g/m^3$. The CSAs with the highest exceedance probabilities are the center city areas near major commuting intersections (76.6°W, 39.3°N), and the northern areas that are adjacent to major interstate traffic (north and northwest borders), however, this is only one possible explanation for June 5th exposures, as traffic is not the only source or causative factor for $PM_{2.5}$ exceedances.

4.3.1. Exposure Variability

As the time period of aggregation increases (weeks, months, seasons, etc.) the spatial variability in $PM_{2.5}$ concentrations across the city decreases. The reduction in spatial variability averaging the monitor-specific $PM_{2.5}$ concentration from day, month, and three months are presented in Table 4. There is a 32% reduction in the standard deviation (spatial variability) across all sites comparing daily to monthly aggregations, and a 57% reduction comparing daily to three-month periods.

Table 4: Standard deviation of SEARCH monitor average measurements over three aggregation time periods

Time Period	**SD of Monitor Means ($\mu g/m^3$)**
2/1/2019	3.7
2/1/2019-2/28/2019	2.5
2/2019-4/2019	1.6

5. Discussion

The use of machine learning for predictive purposes in air pollution sensor data has seen substantial growth in the last several years. Large-scale approaches often utilize satellite data, country scale sensor networks, land use data, topography, etc. and have been built using random forests, GBDTs, and neural nets (Chang et al. 2020; Huang et al. 2018; Zhan et al. 2018; Zhao et al. 2020). On a smaller scale more analogous to SEARCH, personal monitoring device networks, mobile sampling networks, and city-scale sensor networks have also demonstrated the utility of machine learning regression techniques to optimize predictions and take into account environmental factors (Brokamp et al. 2017; Lim et al. 2019; Loh and Choi 2019; Zimmerman et al. 2018). However, while prediction of $PM_{2.5}$ using sensor measurements and additional data has been conducted by numerous studies, this chapter on SEARCH fills a unique position by providing a methodology for both increasing the utility of low-cost sensor networks by creating a probabilistic output useful for exposure assessments, a state-of-the-art model that improves on existing approaches, and also removes the need for lab-calibrated data, a time intensive process for mitigating environmental biases.

The optical sensors used in this work produce voltages that are automatically converted to a raw sensor concentration based on the manufacturer settings within the sensor. This raw sensor can be lab-calibrated (accounting for known environmental biases), a time and resource intensive process, in order to ensure accuracy and precision (Borrego et al. 2018; Levy Zamora et al. 2019). However, the SEARCH network is also deployed in the region of a Maryland Department of the Environment $PM_{2.5}$ monitoring station which measures $PM_{2.5}$ using a reference Federally Equivalent Method (FEM) $PM_{2.5}$ measurement (Maryland Department of the Environment 2018). Therefore, in previous studies, co-located SEARCH sensor monitors with the FEM monitor were used to develop a linear regression model that used lab-calibrated

sensor data, temperature, relative humidity, weekend (binary), and daytime (binary) to model gold standard $PM_{2.5}$ (Datta et al. 2020). Although, temperature and relative humidity are known parameters of concern when measuring $PM_{2.5}$, they have established non-linear relationships with the ultimate $PM_{2.5}$ measurement as well as each other, and these non-linear relationships are the reason why lab calibration is often necessary (Datta et al. 2020; Levy Zamora et al. 2019). Alternately, in order to capture the non-linear relationships without lab calibration, a non-linear modeling approach was used. Gradient boosted decision trees (GBDT), a popular tool for non-linear regression were used and showed that not only were they more accurate than traditional linear approaches without the additional lab calibration step. In addition, the GBDT utilized probabilistic methods, producing unique means (point predictions) and standard deviations for each prediction output, in contrast to linear regression which provides a uniform standard deviation across all predictions. Additionally, by evaluating the models on an entirely out of sample dataset (Essex split), the results show the GBDT calibration models can be transferable to other monitors in the network and improve upon existing linear models even when using raw, uncorrected data.

While creating models that produce accurate probabilistic predictions is interesting from an academic perspective, it is the application of the models that can result in actionable data products, as seen in Figure 5, that presents exceedance probabilities for relevant regulatory/health protective standards. Leveraging the probabilistic prediction optimized NGBoost allows for the use of the distribution for further analysis such as probabilistic risk assessments, more accurate best and worst case scenarios, and any other situation where a parameterized distribution would be more useful than a point prediction. Approximations such as using the 95^{th} percentile prediction from NGBoost would approach a worst-case scenario, but one that is modeled with a unique mean and standard deviation based on the input data.

This is in contrast to a linear regression where a 95th percentile is based on a uniform standard deviation across all data points. Therefore, GBDT allows for low probability occurrences such as the 95th percentile outcome to be modeled with precision, i.e., predictions for worst-case situations should still be accurate. Further, as demonstrated in Figure 5, any quantile of predictions could be used either in terms of exceedance probability or more simply, a straight quantile of exposure plot. Aside from strictly using predicted means, the use of the standard deviation of the predictions could be used to gauge the variation in exposures in a given location post-interpolation, identifying which locations experience the largest swings in $PM_{2.5}$ concentration. The GBDT used have the advantages over traditional linear models in terms of capturing the non-linear relationships between climate parameters as well as creating fully probabilistic outputs. Further, the exposure assessment process is highly flexible using this modeling approach. Any number of aggregations based on time and space could be conducted. Daily and seasonal values for CSAs were shown for illustrative purposes only. The model framework allows for consideration of any time and space aggregation of interest.

In terms of limitations, the specific tuned and fitted models covered in this paper are not universally applicable. The intent was to provide a framework for other investigators to use this approach on their own sensor networks and pollutants. Unique models should be tuned and fitted for every application, which is both a potential limitation but also lends itself to highly customizable solutions. Furthermore, the features for the GBDTs in this setting were not engineering or optimized but were simply the same as what was identified to be optimal for the linear regression by Datta et al. (2020). Therefore, it is likely that feature engineering specifically with the intent of improving the GBDT would yield increased accuracy.

Further research into these methods and data should consider the addition of more reference-sensor pairs that would allow for features that more completely characterize the local environment of each pair. For example, adding land use, topographic, or traffic features would be possible with ease with the GBDT approach. While adding FEMs would be impossible, a short-term high cost/accuracy instrument could be co-located with a number of monitors to provide reference data across the entire network. In addition to the potential for an expanded feature space, one of the primary adjustments to make in order to apply this framework is to determine the amount of training data needed. While this will vary by pollutant, features, model choice, and prediction quality desired, it is clear based on Figures 4a, 4b, and 5 that capturing climate variation across several months would be recommended. Further, it is possible that an ensemble of high bias low variance linear models (not likely to overfit, but likely overly generalized) and low bias high variance GBDT (possibility of overfit, but not overly generalized) would be useful in a setting where a limited amount of training data was available with no option to acquire more. This can be seen in Figure 3 where NGBoost performs equivalently to linear regression on single month training sets – likely slightly overfitting to that one month. Therefore, it is not surprising that with increased training data such as in the full and prospective splits, both NGBoost and LightGBM are more accurate than linear regression.

Lastly, It is possible that a temporal weighting feature that weights newer data more strongly in the model tuning process would additionally yield increased accuracy as a means to combat sensor drift (Levy Zamora et al. 2019). One method of temporal weighting would be to use a type of exponential decay on sample weight of a measurement. The parameters of the exponential weighting could be tuned using a variety of grid searchers or solvers to produce a weight that optimizes next day predictive power based on RMSE or any error metric of choice.

6. Conclusions

The framework for converting uncalibrated $PM_{2.5}$ sensor data into a probabilistic exposure assessment using probabilistic gradient boosted decision trees captures the non-linearity of the relationship between $PM_{2.5}$, relative humidity, and temperature, while providing more accurate and more useful probabilistic and deterministic output. The exposure assessments derived from the probabilistic modeling allows for small scale understanding of $PM_{2.5}$ exposure and variability that can be of use in acute and sub-chronic epidemiological studies. Additionally, with adjustments to the model and feature space this process could be applicable to other air pollutants of concern as well such as O_3, CO_2, or NO_2. The primary limitation of the study as presented is the lack of multiple sensor colocations with the reference site. Two sensors are permanently co-located with the reference, but additional information could be leveraged by rotating co-locations with other sensors. Further, GBDTs are likely to provide optimal utility in a situation with enough data to prevent overfitting, nearly a year of hourly data in this case. While the exact amount of data needed is not definable ahead of time, using linear regression or a blended approach would be suggested if there are concerns about a limited about of data.

7. References

Apte J, Messier K, Gani S, Brauer M, Kirchstetter T, Lunden M, et al. 2017. High-resolution air pollution mapping with Google Street View cars: exploiting big data (Supplemental Material). Environ Sci Technol 51: 6999–7008.

Baddeley A, Rubak E, Turner R. 2015. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press:London.

Baltimore City Department of Health. 2017. Neighborhood Health Profiles - Frequently Asked Questions | Baltimore City Health Department. Available: https://health.baltimorecity.gov/node/231 [accessed 30 September 2020].

Barrett L, Hoyer S, Kleeman A, O'Kane D. 2015. properscoring.

Borrego C, Ginja J, Coutinho M, Ribeiro C, Karatzas K, Sioumis T, et al. 2018. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise – Part II. Atmos Environ 193:127–142; doi:10.1016/j.atmosenv.2018.08.028.

Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmos Environ 151:1–11; doi:10.1016/j.atmosenv.2016.11.066.

Buehler C, Xiong F, Levy Zamora M, Skog K, Kohrman-Glaser J, Colton S, et al. 2020. Stationary and Portable Multipollutant Monitors for High Spatiotemporal Resolution Air Quality Studies including Online Calibration. Atmos Meas Tech in review.

Chang FJ, Chang LC, Kang CC, Wang YS, Huang A. 2020. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. Sci Total Environ 736:139656; doi:10.1016/j.scitotenv.2020.139656.

Chen T, Guestrin C. 2016. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 13-17-August-2016:785–794; doi:10.1145/2939672.2939785.

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, et al. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet 389:1907–1918; doi:10.1016/S0140-6736(17)30505-6.

Daniels R, Gilbert S, Kuppusamy S, Kuempel E, Park R, Pandalai S, et al. 2020. Current Intelligence Bulletin 69 - NIOSH Practices in Occupational Risk Assessment.

Datta A, Saha A, Zamora ML, Buehler C, Hao L, Xiong F, et al. 2020. Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmos Environ 242:117761; doi:10.1016/j.atmosenv.2020.117761.

Dorogush A, Ershov V, Gulin A. 2017. CatBoost: gradient boosting with categorical features support. Neural Inf Process Syst.

Duan T, Avati A, Ding DY, Thai KK, Basu S, Ng AY, et al. 2019. NGBoost: Natural Gradient Boosting for Probabilistic Prediction.

Environmental Protection Agency. 2010. NAAQS Table. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table.

EPA. 2014. Risk Assessment Forum White Paper: Probabilistic Risk Assessment Methods and Case Studies. Available: https://www.epa.gov/sites/production/files/2014-12/documents/raf-pra-white-paper-final.pdf.

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. Ann Stat 29:1189–1232; doi:10.1214/AOS/1013203451.

Gao M, Cao J, Seto E. 2015. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environ Pollut 199:56–65; doi:10.1016/j.envpol.2015.01.013.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378; doi:10.1198/016214506000001437.

Guolin K, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conf Neural Inf Process Syst.

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; doi:10.1016/j.atmosenv.2015.04.057.

Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, et al. 2018. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environ Pollut 242:675–683; doi:10.1016/j.envpol.2018.07.016.

International Agency for Research on Cancer. 2016. Outdoor Air Pollution (Vol. 109).

Johnson NE, Bonczak B, Kontokosta CE. 2018. Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. Atmos Environ 184:9–16; doi:10.1016/j.atmosenv.2018.04.019.

Kahle D, Wickham H. 2013. ggmap: Spatial Visualization with ggplot2. R J 5: 144–161.

Kuhn M, Johnson K. 2013. *Applied predictive modeling*. Springer New York.

Levy Zamora M, Xiong F, Gentner D, Kerkez B, Kohrman-Glaser J, Koehler K. 2019. Field and Laboratory Evaluations of the Low-Cost Plantower Particulate Matter Sensor. Environ Sci Technol 53:838–849; doi:10.1021/acs.est.8b05174.

Lim CC, Kim H, Vilcassim MJR, Thurston GD, Gordon T, Chen LC, et al. 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ Int 131:105022; doi:10.1016/j.envint.2019.105022.

Loh BG, Choi GH. 2019. Calibration of Portable Particulate Matter–Monitoring Device using Web Query and Machine Learning. Saf Health Work 10:452–460; doi:10.1016/j.shaw.2019.08.002.

Maryland Department of the Environment. 2018. Ambient Air Monitoring Network Plan for

Calendar Year 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; doi:10.1016/j.atmosenv.2012.11.060.

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; doi:10.1016/j.envint.2018.04.018.

Mukherjee A, Brown SG, McCarthy MC, Pavlovic NR, Stanton LG, Snyder JL, et al. 2019. Measuring spatial and temporal PM2.5 variations in Sacramento, California, communities using a network of low-cost sensors. Sensors (Switzerland) 19; doi:10.3390/s19214701.

NIOSH. 2017. How NIOSH Conducts Risk Assessments. Available: https://www.cdc.gov/niosh/topics/riskassessment/how.html.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825–2830.

Piedrahita R, Xiang Y, Masson N, Ortega J, Collier A, Jiang Y, et al. 2014. The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. Atmos Meas Tech 7:3325–3336; doi:10.5194/amt-7-3325-2014.

R Core Team. 2020. R: A Language and Environment for Statistical Computing.

Saha PK, Sengupta S, Adams P, Robinson AL, Presto AA. 2020. Spatial Correlation of Ultrafine Particle Number and Fine Particle Mass at Urban Scales: Implications for Health Assessment. Environ Sci Technol 54:9295–9304; doi:10.1021/acs.est.0c02763.

Schapire RE. 2003. The Boosting Approach to Machine Learning: An Overview. Springer, New York, NY. 149–171.

Snyder EG, Watkins TH, Solomon PA, Thoma ED, Williams RW, Hagler GSW, et al. 2013. The

changing paradigm of air pollution monitoring. Environ Sci Technol 47:11369–11377; doi:10.1021/es4022602.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. 2009. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21:n/a-n/a; doi:10.1002/env.1014.

World Health Organization. 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action.

Ye Q, Li HZ, Gu P, Robinson ES, Apte JS, Sullivan RC, et al. 2020. Moving beyond fine particle mass: High-spatial resolution exposure to source-resolved atmospheric particle number and chemical mixing state. Environ Health Perspect 128; doi:10.1289/EHP5311.

Zhan Y, Luo Y, Deng X, Grieneisen ML, Zhang M, Di B. 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ Pollut 233:464–473; doi:10.1016/j.envpol.2017.10.029.

Zhao Z, Qin J, He Z, Li H, Yang Y, Zhang R. 2020. Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. Environ Sci Pollut Res 1–18; doi:10.1007/s11356-020-08948-1.

Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, et al. 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech 11:291–313; doi:10.5194/amt-11-291-2018.

8. Appendix

8.1. (A) Modeling and Spatial Analysis Packages and Libraries

8.1.1. Python 3.7.3

lightgbm==2.3.1

ngboost==0.2.2.dev0

numpy=1.19.0

pandas=1.0.5

properscoring==0.1

scikit-learn==0.23.1

scipy==1.5.1

statsmodels==0.11.1

8.1.2. R 4.0.2 “Taking Off Again”

ggmap_3.0.0

nlme_3.1-149

rgdal_1.5-16

rpart_4.1-15

sf_0.9-5

sp_1.4-2

spatstat_1.64-1

tidyverse_1.3.0

8.2. (B) NGBoost Hyperparameter Search Grid

```
base_learner = DecisionTreeRegressor(criterion = 'mse')
param_grid = {
    'Base': [base_learner],
    'Base__max_depth': list(range(2, 200, 10)),
```

```
'Base__max_features': ['auto'],
'Base__min_samples_leaf': list(range(1, 200, 10)),
'Base__min_samples_split': list(range(2, 200, 10)),
'n_estimators': list(range(500, 3000, 500)),
'minibatch_frac': [1.0, 0.5] }
```

CHAPTER IV: PROBABILISTIC MACHINE LEARNING WITH LOW-COST SENSOR NETWORKS FOR OCCUPATIONAL EXPOSURE ASSESSMENT AND INDUSTRIAL HYGIENE DECISION MAKING

1. Abstract

Occupational exposure assessments have traditionally been dominated by small sample sizes and low spatial and temporal resolution with a focus on conducting Occupational Safety and Health Administration regulatory compliance sampling. However, this style of exposure assessment is likely to underestimate true exposures and their variability in sampled areas, and entirely fail to characterize exposures in unsampled areas. The American Industrial Hygiene Association (AIHA) has developed a more realistic system of exposure ratings based on estimating the 95th percentiles of the exposures that can be used to better represent capture some semblance of exposure uncertainty and variability for decision-making; however, the ratings can still fail to capture realistic exposure with few numbers of small sample sizes. Therefore, low-cost sensor networks consisting of numerous lower quality sensors have been implemented to measure occupational exposures at a high spatiotemporal scale. However, the sensors must be calibrated in the laboratory or field to a reference standard. Using data from carbon monoxide (CO) sensors deployed in a heavy equipment manufacturing facility for eight months from August 2017 to March 2018, we demonstrate that machine learning with probabilistic gradient boosted decision trees (GBDT) can model raw sensor reading to reference data highly accurately, entirely removing the need for laboratory calibration or developing multiple models by sensor, season, etc. Further we indicate how these models can produce probabilistic hazard maps of the manufacturing floor, creating a visual tool for assessing facility-wide exposures. Additionally, the ability to have a fully modeled prediction distribution enables the use of the AIHA exposure ratings, providing an enhanced industrial decision-making framework than simply whether a single measurement was above or below a pertinent occupational exposure limit. Lastly, we show how probabilistic modeling exposure assessment with high spatiotemporal resolution data can prevent inaccuracies associated with traditional models that rely exclusively on point predictions.

2. Introduction

Occupational exposure assessments in the United States are underpinned by the Occupational Safety and Health Administration (OSHA) permissible exposure levels (PEL), or the legally enforceable exposure threshold for a particular contaminant or hazard (Ramachandran 2005). However, the OSHA strategy for exposure assessment only requires a minimal number of samples taken from the highest risk workers, which are then compared against the PEL. If, for any of the samples, the upper confidence limit (exclusively based on sample and analytical error) exceeds the PEL, the workplace is determined to be in violation (Rappaport 1984; Tornero-Velez et al. 1997). This type of compliance monitoring does little to represent true exposure concentrations in a workplace due to the extremely limited spatial and temporal range of the sampling, and usually results in underestimated exposures in facilities where concentrations are poorly controlled, although overestimates are also possible (OSHA 2001; Rappaport 1984; Tornero-Velez et al. 1997; Tuggle 1981). However, the American Industrial Hygiene Association (AIHA) provides exposure assessment guidelines that more properly take into account the uncertainly of small sample estimation by acknowledging the likely lognormal distribution of the true contaminant exposures and is therefore less likely to underestimate exposures (Bullock et al. 2006; Hewett et al. 2006). Unlike compliance sampling which provides only a binary response, the AIHA guidelines provide a set of ratings, (highly controlled, well controlled, nominally controlled, poorly controlled), that assist with decision making processes post sample collection (Hewett et al. 2006; Ramachandran 2005). However, the AIHA ratings still often suffer from small available sample sizes, regardless of the appropriate uncertainty factors applied during exposure rating calculation (Hewett et al. 2006; Ramachandran 2005).

In an effort to fully characterize occupational exposures independent of compliance monitoring, low-cost sensor networks have been deployed, consisting of many lower accuracy, precision,

and sensitivity/specificity sensors spread throughout a facility that can collect measurements at high spatial and temporal resolution (Zuidema et al. 2019a). In addition to occupational settings, these low-cost networks have seen far more use in environmental applications (Buehler et al. 2020; Gao et al. 2015; Heimann et al. 2015; Lim et al. 2019). One of the primary benefits of the high spatiotemporal resolution from networks is that the facility's exposures can be visualized with hazard maps, a method that shows the gradient of exposures in a given area and avoids the interpolation errors associated with mapping with low resolution data (Koehler and Peters 2013; Koehler and Volckens 2011).

One of the primary challenges associated with multi-monitor low precision networks is calibration of each sensor to a reference standard (Afshar-Mohajer et al. 2018; Datta et al. 2020; Levy Zamora et al. 2019; Zuidema et al. 2019a). Calibration ensures that the sensors provide reliable measurements across the range of possible climate and exposure settings (Borrego et al. 2018; Datta et al. 2020; Morawska et al. 2018). However, laboratory calibration is a highly time-intensive process that can require specialized equipment and facilities (Levy Zamora et al. 2019; Zuidema et al. 2019c). However, calibration models have been developed for use with low-cost sensor networks that reduce or entirely eliminate the need for pre-deployment laboratory calibrations or complex intra-deployment calibrations such as unique models for each sensor or location or changing the model by season (Zimmerman et al. 2018, Chapter III). In particular, non-linear approaches (random forest, gradient boosted decision trees) have had success modeling the non-linearity of the relationship between climate data and the contaminant of concern (Zimmerman et al. 2018, Chapter III). In particular, in Chapter III we demonstrated the utility of using fully probabilistic non-linear calibration models as opposed to traditional linear approaches that fail to fully characterize the non-linearity of climate

and measurements, and also only provide point estimates with general measures of variance (e.g., confidence intervals).

In order to validate the probabilistic calibration model approach on occupational data collected with an entirely indoors low-cost sensor network, we will build a model on data described by Zuidema et al. (2019a) in a heavy equipment manufacturing facility in the United States. This network was composed of 40 monitors that were active from August 2017 to March 2018, collecting data on a range of physical and chemical hazards (Zuidema et al. 2019a). We will show that the use of probabilistic gradient boosted decision trees can leverage the full suite of data collected by the low-cost sensor network and produce accurate predictions of a reference dataset from each sensor measurement, eliminating the need laboratory calibrations and/or need multiple linear calibration models. Further, we show how these models can be used to create probabilistic hazard maps where exposure percentiles or exceedance probabilities can be visualized. Hazard maps using the AIHA exposure rating framework will also be created, providing a visual decision-making assistance tool for industrial hygienists. Lastly, we show how the use of the probabilistic models highlights the weakness of exposure assessments that rely strictly on a mean or point prediction to determine occupational exposures.

3. Methods

3.1. Sensor Network Data

An array of 40 monitors was deployed in a heavy equipment manufacturing facility (construction and forestry) in the United States from August 2017 through March 2018 (Zuidema et al. 2019a). Work tasks such as welding, machining, shot blasting, and laser cutting took place on the facility floor (Zuidema et al. 2019a). The development and specifics of the monitors has been previously described in detail (Thomas et al. 2018). Briefly, each of the 40 monitors in the

network contained a sensor for PM2.5 (GP2Y1010AU0F, SHARP Electronics, Osaka, Japan), CO (CO-B4, Alphasense Ltd., Essex, UK), noise (custom), and temperature and relative humidity (AM2302, Adafruit, New York, NY, USA) (Zuidema et al. 2019a). Each sensor recorded data on two second intervals that was averaged to five-minute intervals and wirelessly transmitted and stored in a database from August 4, 2017 to March 27, 2018 (Zuidema et al. 2019a). The layout of the sensors within the manufacturing facility was optimized for maximum spatial variability and to prevent duplicative locations (Berman et al. 2018; Zuidema et al. 2019a). Only temperature, humidity, and CO data will be used for the remainder of the study to demonstrate the application of the method. Data from Thanksgiving, Black Friday, Christmas Eve, Christmas Day, New Year's Eve, and New Year's Day was excluded as operations were drastically altered from normal during these times.

3.2. Reference Data

The reference data for CO was collected using an EL-USB-CO (Lascar Electronics, Erie, PA, USA) at three quarters throughout the year, the first on August 17, 2017 (Q1), the second on December 20, 2017 and December 21, 2017 (Q2), and the third on March 23, 2018 and March 26, 2018 (Q3) (Zuidema et al. 2019a). The reference instruments were placed at a total of ten unique locations within the bounds of the monitoring network on five days over the three quarters, some exactly spatially collocated with sensor monitors (approximately 85 percent of reference measurements), and some intentionally non-collocated (approximately 15 percent of reference measurements). Reference data was collected throughout the workday on each measurement day.

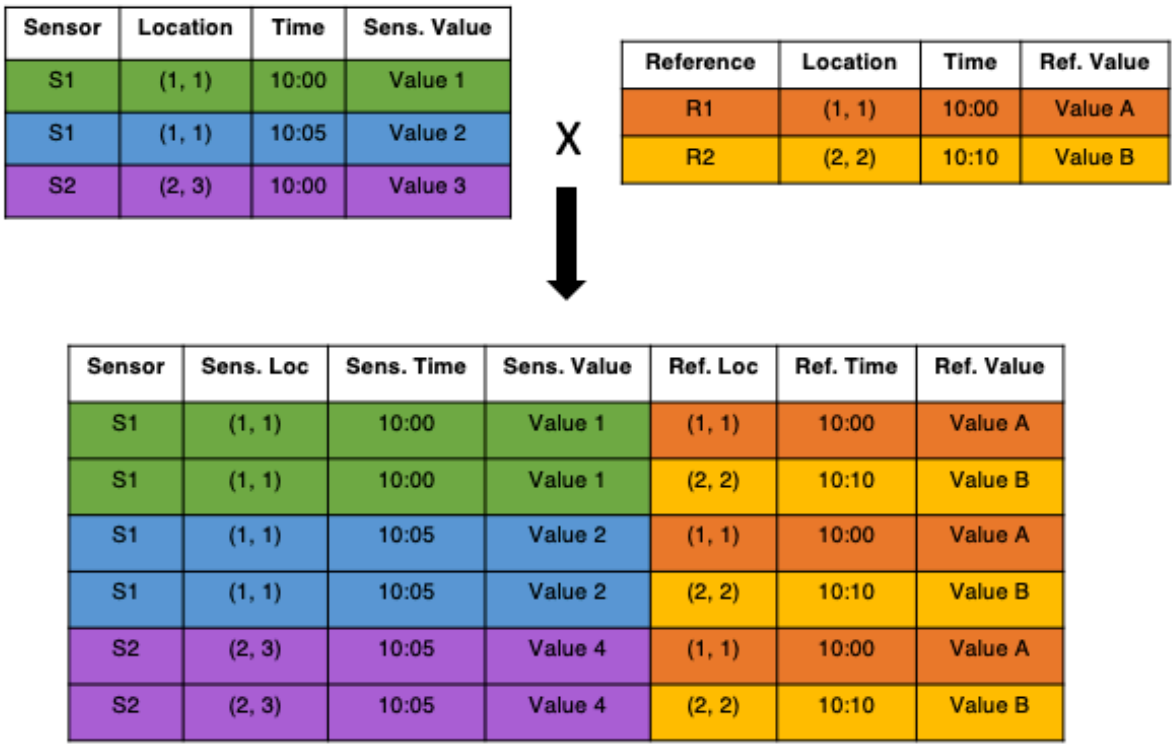
3.3. Spatiotemporal Weighting

Over the course of the five days of reference sampling, most monitors still did not have a collocated reference measurement. In order to expand the sample size from exclusively spatially and temporally collocated sensor and reference pairs, a weighting system was utilized. Within each of the five reference sample days, all sensor measurements were paired with all reference measurements as shown in an example version in Figure 1, where the blue, green, and purple unique network measurements are each paired with every orange and red unique reference measurement. Therefore, the sample size for training increased from the number of exactly collocated in time/space measurements (n = 1,785 unique sensor measurement-reference measurement pairs) to the count of sensor measurements multiplied by the count of all reference measurements (n = 1,182,290 unique sensor measurement-reference measurement pairs). Utilizing non-exactly collocated measurement pairs takes advantage of known spatial and temporal dependence in the data, i.e., measurements closer in space and time are more likely to be similar than those further apart.

The difference in minutes (Δt) between the two measurements was then used as an exponential weight in $C^{\Delta t}$, with C, a numerical constant, taking possible values of 0.70, 0.75, 0.80, 0.85, 0.90 and 0.99, representing between two percent weight (i.e., $0.70^{10} = 0.02$) and 90 percent weight at ten minutes. Additionally, the distance between the sensor monitor and the reference instrument in yards (D) for each sensor measurement was used to create a spatial weight of $1/(1 + D^P)$, where P takes possible values of 1.0, 1.5, or 2.0. Lastly, the time weight and the spatial weight were multiplied together to form the total weight for the observation. The result is a weighting factor with a maximum value of 1.0 (100%) for a sensor measurement that was spatially and temporally collocated with a reference measurement, and a smaller weight for sensor measurements that differed by time and/or space. The exact values for use in the final model

were determined by re-tuning the model with all possible pairs of C and P and finding the set with the least error. Measurements with a total weight of less than five percent were eliminated from the dataset prior to modeling to reduce computational costs (approximately 90% of the total sensor measurement-reference measurement pairs). For example, some of these eliminated measurements would include sensor and reference pairs from different days and/or hundreds of feet away.

Figure 1: Example of data expansion prior to spatiotemporal weighting where blue, green, and purple represent unique low-cost sensor network measurements and red and orange represent unique reference instrument measurements



3.4. Modeling

3.4.1. Gradient Boosted Decision Trees

Gradient boosted decision trees (GBDT) are a machine learning methodology that utilizes numerous weak decision tree learners combined in a forward stepwise additive manner (Kuhn and Johnson 2013; Schapire 2003). Each new weak learner is added based on the residual of the predictions from the previous step (Kuhn and Johnson 2013; Schapire 2003). Although there are a variety of GBDT implementations that could be used (XGBoost, LightGBM, CatBoost, etc.), this chapter will use NGBoost, an open source python library developed by the Stanford Machine Learning Group that boosts a base learner, in this case a scikit-learn DecisionTreeRegressor (Duan et al. 2019; Pedregosa et al. 2011). For regression, NGBoost supports fully probabilistic predictions where both the mean and standard deviation in the form $N(x, \sigma)$, are modeled for each prediction (Duan et al. 2019). Hyperparameters, or the values that control how the model learns, were tuned using a random search over a grid of parameter values with the specific grid details shown in Appendix A (Bergstra and Bengio 2012; Pedregosa et al. 2011).

3.5. Model Features

Given that the sensor network was deployed in one contiguous climate controlled indoor manufacturing floor, extreme outliers of climate conditions (*e.g.*, very high relative humidity or very cold temperatures) were interpreted as measurement errors. Therefore, to keep the measurements in the data set, but to minimize the effect of the climate measurement errors and preserve climactic trends within a day, a separate general additive model (GAM) was fitted to temperature and relative humidity with the minute of the day as the single covariate (Wood 2011). These smoothed temperature and relative humidity values for the facility as a whole (xTemperature and xRH respectively) were then assigned to each sensor measurement based

on minute and day to provide a more stable estimate of current climate conditions. The raw temperature and humidity measurements were also included as features the model to capture an estimate of any local events which occurred (*e.g.*, welding, heating, etc.), as well as to understand how a climate sensor misread could impact other measurements. The full set of features as well as the target and weights for the GBDT are presented in Table 1. All data, both reference and sensor, was restricted to the hours of 6:00 am-6:00 pm. Few if any workers were present outside this window.

Table 1: Description of target, features, and weighting for GBDT model

Data	Description (units)	Source	Model Use
CO Reference	Reference measurement (ppm)	Reference instrument	Target
CO Sensor	Raw sensor measurement (mV)	Sensor	Feature
Hour	Hour of day (6-18)	Sensor	Feature
Temperature	Raw sensor temperature (°C)	Sensor	Feature
RH	Raw sensor relative humidity in (%)	Sensor	Feature
xTemperature	Modeled facility temperature (°C)	Model	Feature
xRH	Modeled facility relative humidity (%)	Model	Feature
East	X-coordinates of sensor location on factory floor (feet)	Manual entry	Feature
North	Y-coordinates of sensor location on factory floor (feet)	Manual entry	Feature
Total Weight	Spatiotemporal weighting for sensor measurement (0-1)	Calculated	Weight

3.6. Training and Testing Datasets

In order to determine the quality of the GBDT model, two types of training and test splits were used. The first, or the 'Full' split, is all of the data split randomly 80/20 into training and test set

(Pedregosa et al. 2011). The second is training on one quarter of data (one of Q1, Q2, Q3) and testing on the other two quarters. These splits will determine not only the validity of the model on out of sample data but will also assist with the determination of how much reference data is needed to build an effective model. The sample sizes for each split with the most restrictive (smallest possible sample sizes) and least restrictive (largest possible sample sizes) and weighting are shown in Table 2.

Table 2: Training and testing set sizes for all train/test splits with most restrictive (smallest sample size) and least restrictive weighting (largest sample size)

Training Set & Split Name	**Training Size (Min/Max)**	**Testing Set**	**Testing Size (Min/Max)**
'Full' (80% of all data)	2,859/36,087	20% of all data	715/9,022
Q1	455/5,708	Q2, Q3	3,119/39,401
Q2	1,552/19,561	Q1, Q3	2,022/25,548
Q3	1,567/19,840	Q1, Q2	2,007/25,269

3.7. Sensitivity Test

In addition to various training and test splits, a sensitivity test on the number of datapoints included in the model was conducted using the optimized C and P parameters. Instead of using an 80/20 training split, the model will be trained and tested on 60/40, 40/60, and 20/80. These intentionally down-sampled training set models were then compared with the full dataset model baseline for accuracy.

3.8. Model Evaluation

The models were evaluated on two forms of accuracy, root mean square error (RMSE) and continuous ranked probability score (CRPS), both exclusively on test set data not used in model training. RMSE is the square root of the mean of the squared residuals and is always greater than or equal to zero. Smaller values of RMSE indicate a more accurate model. The equation for RMSE is shown in Eq. 1, where 'Actual' is the reference measurement, 'Predicted' is the mean of the $N(x,\sigma)$ prediction from the GBDT, and n is the total number of predictions under evaluation. Additionally, the units of RMSE are those of the prediction, or in this case, ppm of CO.

Eq. 1: Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Actual - Predicted)^2}{n}}$$

While RMSE is used to evaluate a point prediction, CRPS is used to evaluate both the point prediction (mean) and the spread (standard deviation) of the $N(x,\sigma)$ prediction for each measurement. Similar to RMSE, CRPS is in the units of the prediction and smaller values indicate a more accurate fit. CRPS is a generalized form of mean absolute error for use with probabilistic predictions (Gneiting and Raftery 2007). CRPS values were calculated using the python properscoring package provided by The Climate Corporation (Barrett et al. 2015; Pedregosa et al. 2011).

The results of the model evaluation will also be compared to the results obtained using the linear calibration equation developed and utilized by Afshar-Mohajer et al. (2018) and Zuidema et al. (2019a) respectively. These linear regressions modeled low-cost sensor response to known concentrations of CO in a laboratory setting, but did not include temperature or relative humidity as covariates (Afshar-Mohajer et al. 2018). Of note, the linear calibration model in laboratory settings failed when reference concentration measurements exceeded 12 ppm, as the low-cost sensor became saturated and produced a non-linear response at high concentrations (Afshar-Mohajer et al. 2018).

3.9. Spatial Interpolation and Hazard Mapping

The predictions from the GBDT are used to create hazard maps via spatial interpolation of CO predictions using a 256x256 grid to cover the approximately 800,000 square foot factory floor. However, in order to best utilize the fully probabilistic predictions, a Monte Carlo resampling will be conducted to interpolate over the full range of possible values. Interpolation was conducted using inverse distance weighting (IDW), a non-statistical process by which values closer to the interpolation point are given more weight than those further away. The equation for IDW is shown in Eq. 2 where d_i is the distance between the interpolation point and the measured value, i is a location with known data, z_i is the data value at the location i, and n is the total number of points used in the averaging. The power parameter, p, determines the intensity of the inverse distance relationship with larger values weighting further apart locations less. The value of P was set at 2.0 based on convention and prior research (Chapter III). IDW was conducted using the spatstat package in R (Baddeley et al. 2015).

Eq. 2: Inverse Distance Weighting (IDW)

$$z_{estimated} = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} z_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

Prior to performing the interpolation for the exposure assessment, the interpolation error was determined using a methodology from Chapter III for repeated leave-one out cross validation (LOOCV) of IDW predictions. 500 random five-minute measurement times were selected over the entire study period, and for each measurement time an IDW was performed with each active monitor left out once (approximately 40 IDWs per measurement time). Residuals were calculated from the difference of the IDW prediction at the left-out monitor location and time and the mean of the $N(x,\sigma)_{GBDT}$ for that location and time. The resulting residual distribution was parameterized into an appropriate symmetrical probability distribution based on negative loglikelihoods.

Following determination of the IDW error, the Monte Carlo simulation using all the $N(x,\sigma)_{GBDT}$ can be conducted using methods described in Chapter III, where the IDW is repeatedly conducted using values from $N(x,\sigma)_{GBDT}$ with each interpolation location also sampling from the parameterized error distribution and adding that sample to the IDW result. The product is an interpolated surface where each grid point is $N(x,\sigma)_{IDW}$, incorporating uncertainty from the modeling process and error from the interpolation. The interpolated surface can be aggregated to any timescale greater than the minimum sensor resolution by averaging the $N(x,\sigma)_{GBDT}$ at each monitor across the time period of interest prior to the interpolation. As the surfaces are

probabilistic; various percentiles can be mapped to show best case scenarios, worst case scenarios, etc.

3.10. Using Mean Concentrations to Estimate 95th Percentile Concentrations

In order to characterize the spatial relationship between mean exposures and 95th percentile exposures, daily (n = 223) hazard maps were generated. Each location on the 256x256 grid will have its mean and 95th percentile of exposure via $N(x,\sigma)_{IDW}$ separately binned into daily quintiles, resulting in $Quintile_{Mean}$ and $Quintile_{95}$ respectively. The $Quintile_{Mean}$, $Quintile_{95}$ location pairs will then be compared to determine if a location's $Quintile_{Mean}$ can properly estimate relative (compared across the facility) upper bound exposure levels. For example, if a location's $Quintile_{Mean}$ and $Quintile_{95}$ are both in the same quintile, then the mean exposure can properly estimate relative upper bound exposure at that location. However, if $Quintile_{Mean} < Quintile_{95}$, then the mean exposure underestimates the relative upper bound exposure, or if $Quintile_{Mean} > Quintile_{95}$, then the mean exposure overestimates the relative upper bound exposure.

3.11. AIHA Exposure Ratings

In addition to visualizing percentiles of CO concentrations, the prediction distributions can be directly utilized with the American Industrial Hygiene Associations' Exposure Rating framework which is provided in Table 3 (Ramachandran 2005). These ratings are intended to allow industrial hygienists to make clear determinations as to where to focus resources for additional controls, if necessary. Additionally, the flexibility of a custom OEL allows for selection of action thresholds that can serve as signal events before reaching a regulatory over exposure or a health relevant over exposure. In order to generate exposure rating hazard maps, the 95th percentile ($X_{0.95}$) of $N(x,\sigma)_{IDW}$ at each point was calculated and classified based on AIHA instructions.

Table 3: AIHA Exposure Rating framework with ratings, descriptions, explanations, and numerical interpretations

Exposure Rating	Description	Explanation	Numerical Interpretation
1	Highly controlled	Exposures infrequently exceed 10% of the limit	$X_{0.95} \leq 0.10 * OEL$
2	Well controlled	Exposures infrequently exceed 50% of the limit and rarely exceed the limit	$0.10 * OEL < X_{0.95} \leq 0.5 * OEL$
3	Nominally controlled	Exposures infrequently exceed the limit	$0.50 * OEL < X_{0.95} \leq OEL$
4	Poorly controlled	Exposures frequently exceed the limit	$OEL \leq X_{0.95}$

However, the AIHA exposure ratings are generally determined using Eq. 3, where the upper confidence limit of the 95th percentile of exposure is calculated due to typically small sample sizes (less than ten). In Eq. 3, γ is the confidence level, p is the proportion, n is the sample size, $\bar{y}$ is the natural log of the geometric mean, and s_y is the natural log of the geometric standard deviation, and K is the value from a K-value table based on the corresponding parameters.

Eq. 3: American Industrial Hygiene Association UCL_{95}

$$UCL_{95} = e^{(\bar{y} + K_{\gamma,p,n} * s_y)}$$

Due to the extremely dense data as a result of the sensor network, the 95th percentile was directly determined without any additional uncertainty introduced by the use of the K-factor in Eq. 3. The AIHA ratings are assumed to be based on personal exposures, while the sensor network by definition measures area exposures. However, following interpolation, the concentration surfaces allow for an individual location to be assumed to be the exposure to a stationary worker.

3.12. Compliance Sampling Comparison

In order to demonstrate the accuracy of compliance sampling to assess occupational exposures, a comparison of UCL_{95} for small sample sizes (3, 5, 10, 50) against the network-based assessments was conducted. Given the large sample size of the network measurements following the construction of the daily (n = 223) interpolated hazard maps, the network exposure 95th percentile will be considered the reference the UCL_{95} is compared against. To simulate a compliance sampling system, at 50 randomly chosen locations on the factory floor (to approximate the exposures for 50 stationary workers), a series of random daily exposure values were chosen from the 223 possible daily time-weighted average concentrations; each of three, five, ten, and 50 daily exposure values were each randomly sampled 100 times. For example, one location using three days of exposures would be analogous to a stationary worker in that location being sampled on three days. Then, using the number of daily TWA exposures, the upper confidence limit of the 95th percentile of exposure (UCL_{95}) was calculated based on ACGIH guidance using Eq. 3 (Hewett et al. 2006; Ramachandran 2005). The exposure rating and estimated exposure for the worker based on the UCL_{95} for each of the number of days sampled was compared against the exposure rating and estimated X_{95} exposure based on the sensor network concentrations.

3.13. Software

All linear and GBDT modeling was conducted in Python 3.8.3 with spatial analysis conducted in R 4.0.2 'Taking Off Again' (R Core Team 2020). The full list of modeling packages, libraries, and their version numbers is provided in Appendix A.

4. Results

4.1. Sensor Modeling

The weighting parameter values that minimized model error via RMSE and CRPS were C = 0.75 and P = 2.0. Using C = 0.75 and P = 2.0, the models were refit and the prediction results on the test sets for the four splits are presented in Table 4.

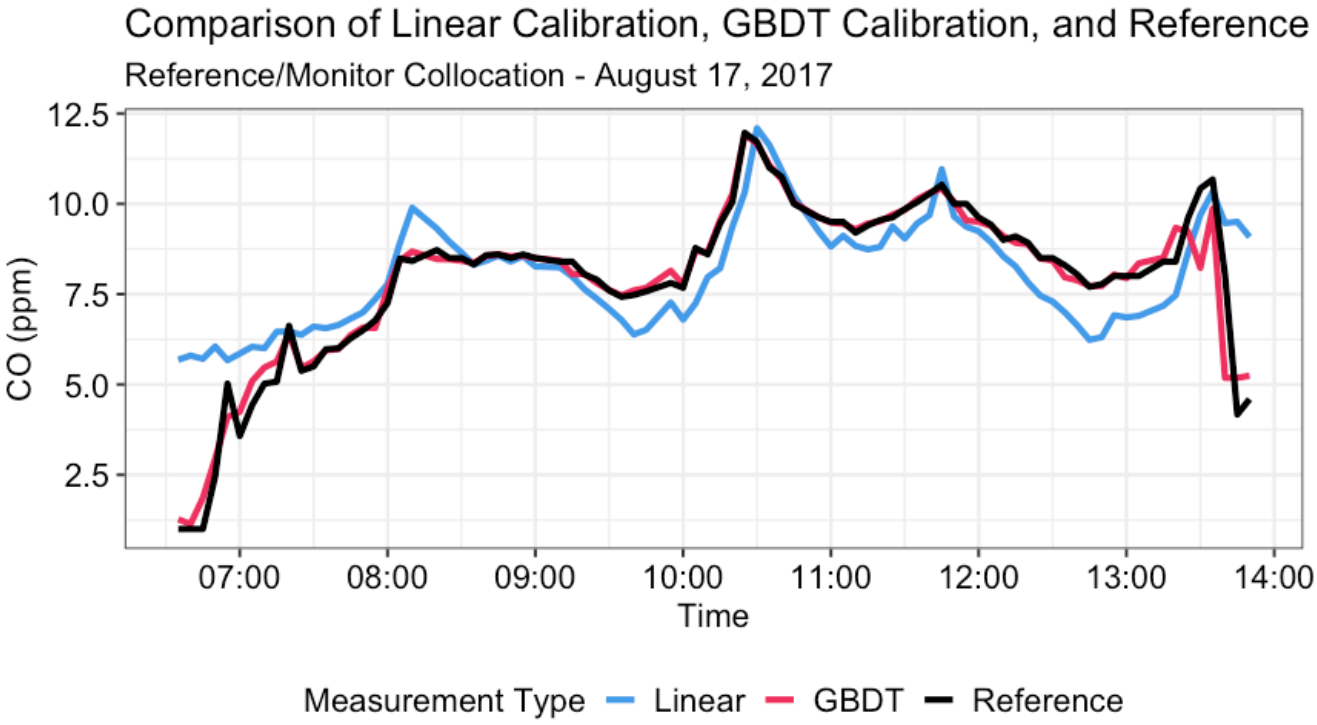
Table 4: Sample size, RMSE, and CRPS for all test/train splits with C = 0.75 and P = 2.0

Training Set & Split Name	Training Size	Testing Set	Testing Size	RMSE (ppm)	CRPS (ppm)
'Full' (80% of all data)	4,293	20% of all data	1,073	0.36	0.16
Q1	682	Q2, Q3	4,684	1.91	1.11
Q2	2,330	Q1, Q3	3,036	2.35	1.98
Q3	2,354	Q1, Q2	3,012	1.39	0.73

The full split, with an RMSE of 0.36 ppm on the test set, was almost four times lower than for the other splits. This accuracy is likely a result of the larger and more diverse training set and a test set that is composed of the same time mix of data as the training set whereas Q1, Q2, and

Q3 were tested on fully out of sample data. The relatively higher error in the Q2 split (this held true during the C and P tuning as well) is potentially due to atypical production schedules associated with reference sampling occurring shortly before the Christmas holiday (Thomas et al. 2018; Zuidema et al. 2019a). Additionally, using C = 0.75 and P = 2.0 all GBDT train and test splits performed better than the previously utilized linear calibration equation which produced RMSE values of 2.42 ppm on a full dataset test. The time series of a single collocated reference and monitor pair on August 18, 2017 is shown in Figure 2 with linear calibration, the GBDT calibration, and the reference measurement (Afshar-Mohajer et al. 2018; Zuidema et al. 2019a). As seen in previous work, the GBDT is both more accurate and able to capture more peaks and valleys in reference measurements than a linear model (Chapter II).

Figure 2: Time series comparison of linear calibration, GBDT calibration, and reference measurements at single reference-monitor collocation on August 18, 2017



The results for this sample size sensitivity test with C = 0.75 and P = 2.0 are shown in Table 5. The error increases with decreased sample size. However, the only approximately 52% increase in RMSE with an 80% decrease in training sample size could point to the possibility of reducing the number of reference sample days for future network deployments.

Table 5: RMSE and CRPS for baseline (80/20) split and additional splits with reductions in training size with C = 0.75 and P = 2.0

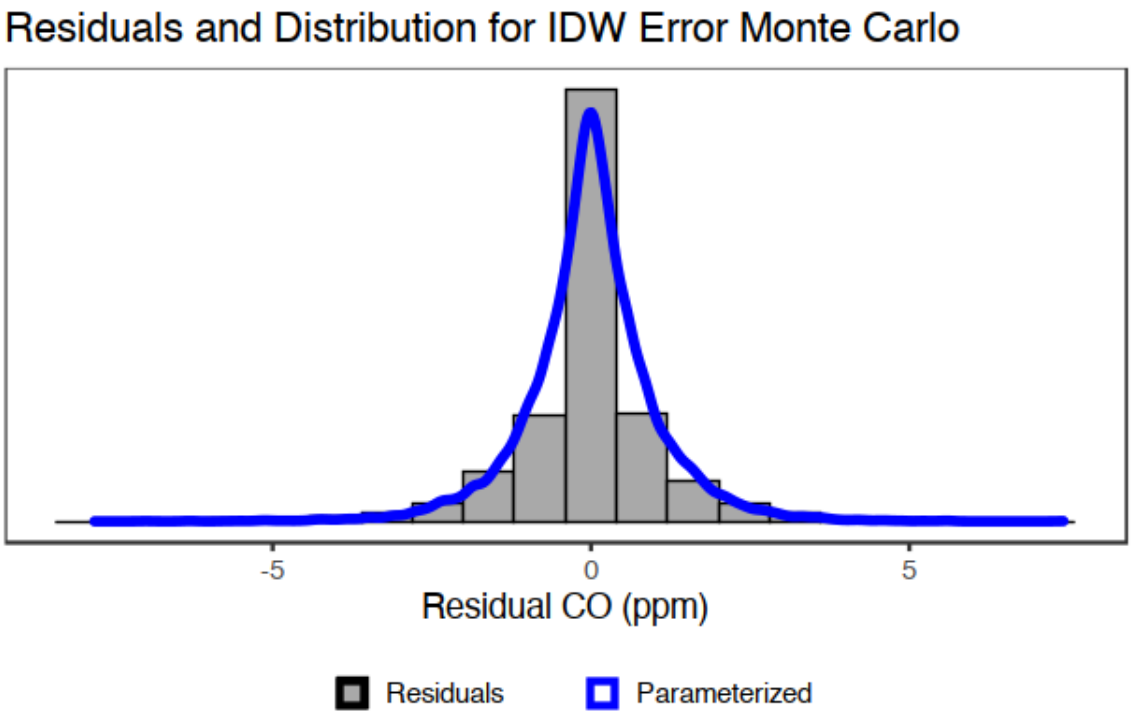
Training/Test Split	Training Size	Testing Size	RMSE (ppm)	CRPS (ppm)
80/20 (baseline model)	4,293	1,073	0.36	0.16
60/40	3,220	2,146	0.43	0.17
40/60	2,146	3,220	0.46	0.19
20/80	1,073	4,293	0.55	0.24

4.2. Spatial Interpolation and Hazard Mapping

The tuned full split model on the entire dataset with C = 0.75 and P = 2.0 was used to predict $N(x, \sigma)_{GBDT}$ at each sensor location and time across the full eight-month deployment. These predictions at sampled locations were then interpolated and used to determine the IDW error. Following the 500 measurement periods where LOOCV was conducted for each active monitor, the residual distribution was most accurately parametrized via likelihood estimation into a Laplace distribution of $\mu = 0.0, b = 0.71$, referred to as $Laplace(0.00, 0.71)_{Error}$. Likelihood estimation was conducted using the MASS and ExtDist packages in R (Venables and Ripley 2002; Wu et al. 2015). The residual distribution and parameterized distribution are shown in Figure 3. $Laplace(0.00, 0.71)_{Error}$ was sampled on each iteration of the IDW Monte Carlo and

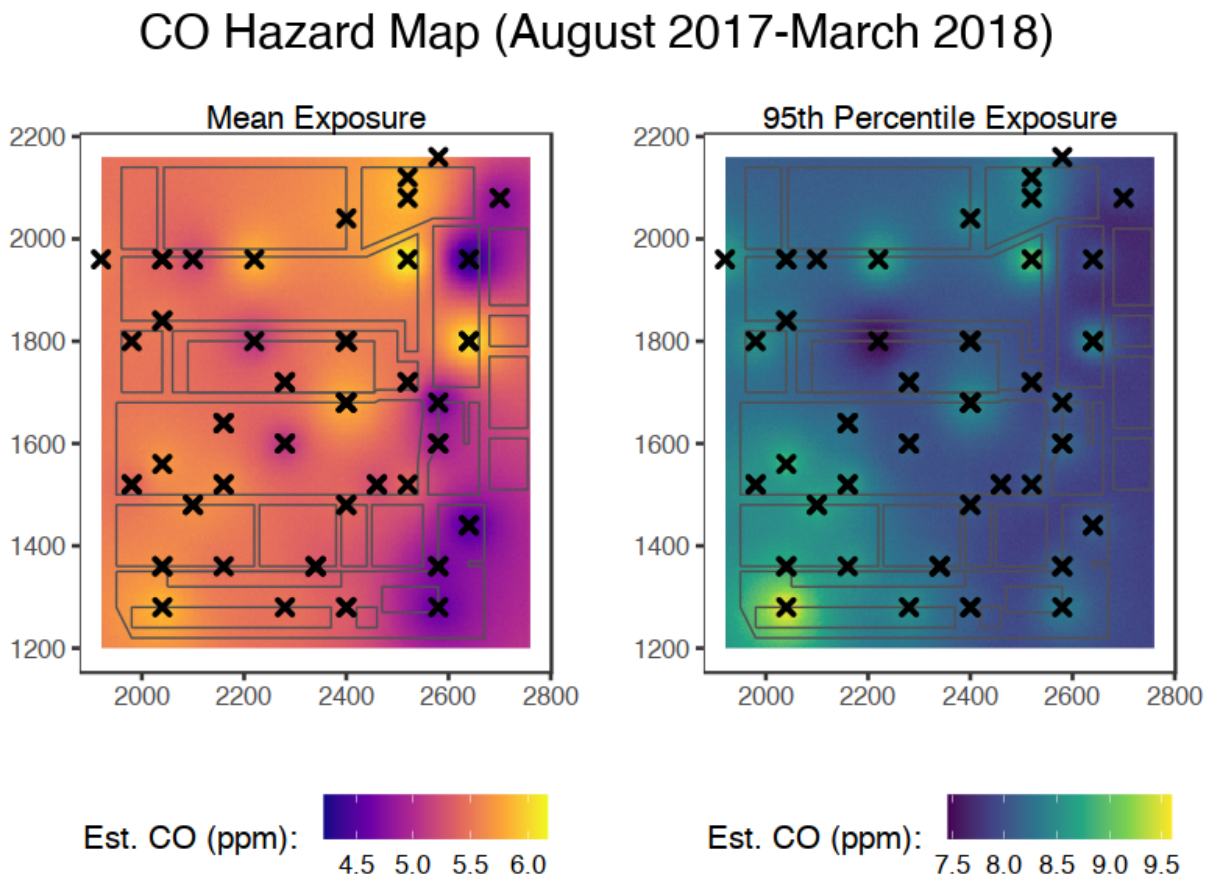
added to each location's IDW predicted value to propagate modeling and interpolation error throughout the interpolation, resulting in a probabilistic surface of CO concentration data.

Figure 3: Residuals from LOOCV IDW Monte Carlo and corresponding parameterized distribution of Laplace(0.00, 0.71)



With probabilistic interpolated surfaces, hazard maps for any percentile of interest can be created, instead of simply relying on a mean point prediction. The mean and 95th percentile hazard map for the entire dataset spanning August 4, 2017 to March 27, 2018 is shown in Figure 4, overlaid with locations of the sensors denoted by "x".

Figure 4: Hazard map of estimated mean and 95th percentile CO concentrations (ppm) for August 4, 2017 to March 27, 2018

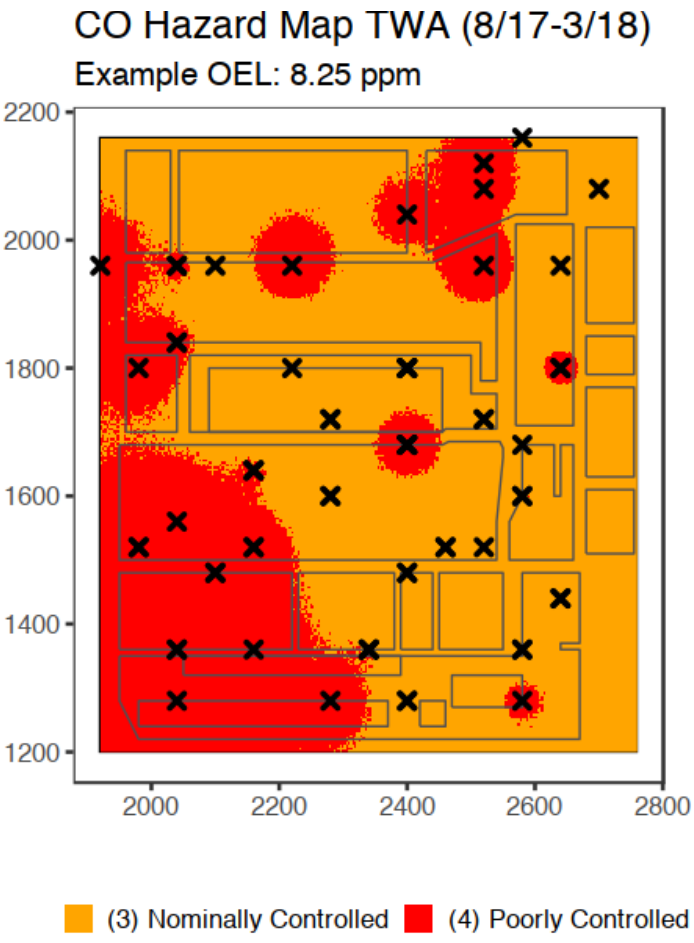


The mean exposure hazard map and 95th percentile exposure hazard map have concentrations ranging from approximately 4.5 ppm to 6.0 ppm and 7.5 ppm to 9.5 ppm respectively. The primary point of interest is that the 95th percentile concentration map is not simply a linear multiple of the mean concentration map scaled up by a constant, likely indicating that concentration variability is not uniform across the facility. However, given that this data is aggregated over 223 days, the 95th percentile for any daily exposure can be could be higher or lower than this range.

4.3. AIHA Exposure Rating Hazard Maps

Utilizing the 95th percentile interpolated values from Figure 4 allows for direct use in the AIHA exposure ratings framework. Each grid location can then be assigned to an exposure category based on a selected occupational exposure limit (OEL). For the dataset used in this analysis, the CO exposures were significantly below the OSHA PEL of 50 ppm, and an entirely artificial OEL of 8.25 ppm was used for illustrative purposes only (NIOSH 2019). Using the full timeseries of network deployment, as shown in Figure 4, the exposure category hazard map is shown in Figure 5.

Figure 5: AIHA exposure ratings based on full dataset prediction and interpolation for August 4, 2017 to March 27, 2018



Areas in red are 'Poorly Controlled' meaning that for an assigned OEL of 8.25 ppm, the 95th percentile of exposure is greater than the OEL. Orange areas are 'Nominally Controlled' with the 95th percentile of exposure between one half of the OEL and the OEL. Due to the simulation process and the probabilistic model, there are granularities, or non-contiguous exposure zones in the hazard map at the borders of well-defined exposure regions. However, the micro-scale exposure gradients should not be used by industrial hygienists as evidence of true exposure variation, but as a level of general variability and uncertainty in a transition region.

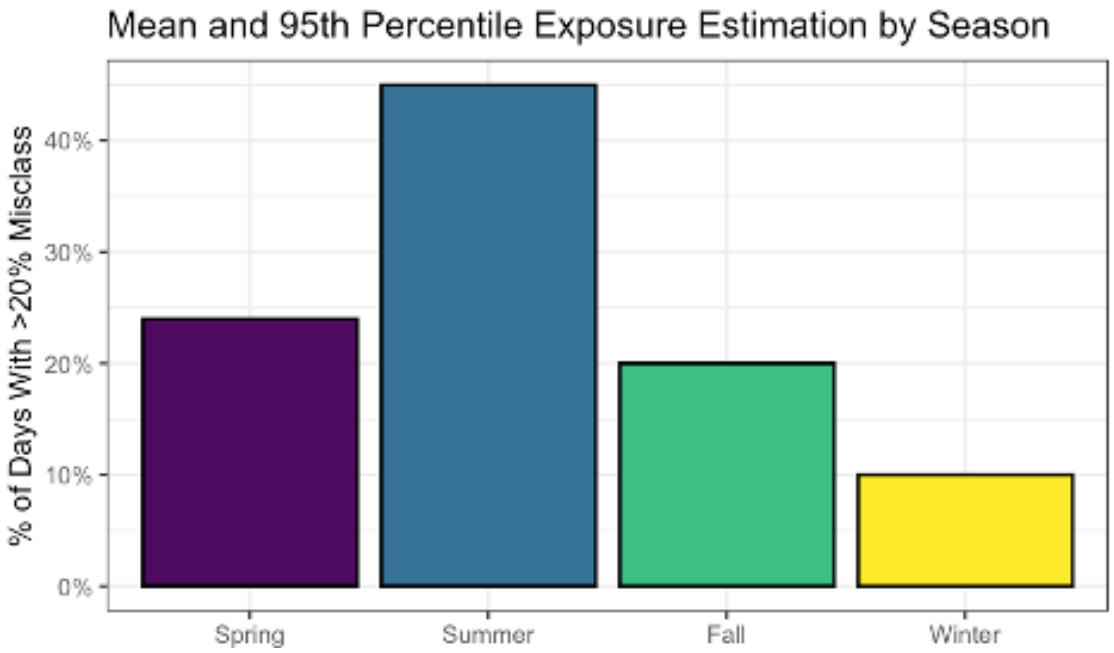
4.4. Using Mean Concentrations to Estimate 95th Percentile Concentrations

The 223 $Quintile_{Mean}$ and $Quintile_{95}$ pair counts were aggregated and grouped by season. If a location's $Quintile_{Mean}$ and $Quintile_{95}$ are in identical quintiles, then the mean exposure can be used to approximate relative upper bound exposure at that location. However, if $Quintile_{Mean} < Quintile_{95}$, then the mean exposure underestimates the relative upper bound exposure, or if $Quintile_{Mean} > Quintile_{95}$, then the mean exposure overestimates the relative upper bound exposure – both are considered exposure misclassifications. The exposure misclassification counts are shown in Figure 6 where each season is broken out by the percentage of days in that season where greater than 20 percent of the factory floor has an exposure misclassification.

Comparing across seasons, estimating relative upper bound exposure from the mean is least effective in the summer months (August and September), with approximately half the days having at least 20 percent of the facility inaccurately classified. The remaining three seasons were slightly more accurate using the mean to estimate the 95th percentile, with between 10

and 25 percent of the days in those seasons having at least 20 percent of the facility misclassified.

Figure 6: Percentage of days, by season, where greater than 20 percent of the factory floor would be misclassified using mean exposures ($Quintile_{Mean}$) to estimate relative upper bound exposures ($Quintile_{95}$) (August 4, 2017 to March 27, 2018)



4.5. Compliance Sampling Comparison

The results of the comparison between simulated compliance sampling and the sensor network for exposure assessment are shown in Table 6. For the majority of the locations (simulated workers), the compliance sampling significantly overestimated the exposures both in terms of the peak exposures (UCL_{95} vs. 95^{th} percentile) and the exposure rating categories. Approximately two thirds of locations were incorrectly classified into a higher exposure category (4 as opposed to 3) than they should have been. Geometric means and standard deviations were used for the mean comparisons due to the lognormal distribution of the UCL_{95} values

when using different sampling data. The geometric standard deviation is across 100 replicates and all locations. Even with 50 sampling days, the geometric mean UCL_{95} is more than 10 percent higher than the X_{95} from the network data.

Table 6: Comparison of simulated compliance sampling exposures with sensor network exposures for 50 workers/locations and range of 'sampling days', each repeated 100 times

Simulated Compliance Sampling			**Network Sampling**	**% of Locations with Overestimated Exposure Category**
Sampling Days	**Geo. Mean UCL_{95} (ppm)**	**Geo. SD UCL_{95}**	**Geo. Mean 95th Percentile (ppm)**	
3	30.55	2.60	8.14	68
4	18.56	1.75		68
5	15.11	1.48		68
10	10.96	1.19		68
50	9.06	1.06		65

5. Discussion

The primary utility of a low-cost sensor network is to capture measurements on a smaller spatial and temporal scale than a single reference monitor (Piedrahita et al. 2014; Snyder et al. 2013; Szpiro et al. 2009). In an occupational setting, these networks can monitor an entire facility floor, characterizing exposures across time and space in a way that would be impossible with traditional industrial hygiene sampling (Thomas et al. 2018; Zuidema et al. 2019b).

However, sensors in low-cost networks tend to be individually less accurate and precise than more expensive instruments that are used for either calibration or confirmatory sampling (Datta et al. 2020; Levy Zamora et al. 2019; Thomas et al. 2018; Zuidema et al. 2019a). Therefore, the low-cost sensors need to be calibrated, either in the lab or the field, to provide measurements that are usable from an exposure assessment context, especially given the non-linear relationship of certain pollutants of concern with climate data (Datta et al. 2020; Zuidema et al. 2019c). Instead of conducting any pre-deployment lab calibrations or intra-deployment sensor adjustments, this chapter utilized a fully probabilistic gradient boosted decision tree (GBDT) that modeled reference measurements directly from raw sensor readings in millivolts, as well as climate and location data for each low-cost sensor. The fully probabilistic GBDT (NGBoost) allowed for a unique mean and standard deviation to be modeled for each prediction. Additionally, by utilizing a spatiotemporal weighting system, the size of the dataset used for modeling increased substantially by not exclusively using reference and sensor measurement pairs that were perfectly collocated in time and space. For example, a sensor measurement that was ten feet and five minutes from a reference measurement was not discarded but weighted according to a tuning process that optimized for predictive accuracy.

The results from the model and sensitivity tests indicated that CO reference concentrations can be predicted with high accuracy (between two- and five-times reduction in error from traditional linear methods) when training on reference data that covers a range of climate conditions (August, December, and March), but training in a single season reduces the accuracy substantially. While this chapter is in an occupational setting measuring a gas, a similar probabilistic GBDT calibration model for environmental spatial exposure assessments on low-cost sensor networks measuring particulate matter was developed by in Chapter III with comparable predictive results. Therefore, in two very different exposure scenarios, the

probabilistic GBDT calibration method for spatial exposure assessments has been shown to improve upon existing approaches and provide utility (unique estimates of variance on a per-prediction level, capture non-linear relationships, etc.) not possible with traditional calibration or modeling approaches.

However, the accuracy of the model alone is not the primary advantage to this probabilistic calibration approach. The probabilistic predictions can be interpolated across the entire facility floor to produce hazard maps showing exposures for a user-selected percentile (Figure 3) or in terms of threshold exceedance probabilities (e.g., the likelihood that a given location exceeded 5 ppm) (see Chapter III). The ability to provide a full distribution of values based on the $N(x, \sigma)$ at each grid location adheres to the American Industrial Hygiene Association's (AIHA) exposure categories, which explicitly rely on the 95th percentile exposures, as opposed to mean or median exposures (Ramachandran 2005). Figure 5 presents an example of an exposure ratings assessment that provides a simple (easily understandable by non-technical employees), graphical display of an exposure scenario. Such maps could be used by an industrial hygienist to indicate that areas of a facility that need immediate attention or to prioritize controls or additional high-quality sampling (badges, canisters, etc.) at those locations and tasks. However, exposure rating variation between adjacent grid cells should not be used by industrial hygienists as evidence of micro-scale exposure gradients, but as a level of general variability and uncertainty. Furthermore, as seen in Table 7, the utilization of traditional compliance monitoring sampling protocols is non-competitive with network-based exposure assessments in terms of both exposure category accuracy and exposure accuracy.

An additional point of emphasis for using the 95th percentile is that using mean exposures alone can potentially result in exposure misclassifications. For example, as shown in Figure 6,

on approximately 50 percent of the days the network was active in the Summer, 22 percent of the factory floor would have had incorrect exposure assignments using exclusively the mean exposure as opposed to the 95th percentile exposures. Using the mean as a proxy for the upper bound exposures essentially prevents the accurate characterization of facility-wide exposures. These misclassifications could lead to prioritization of industrial hygiene control resources on areas that may not yield desired reductions in worker exposures.

The primary limitations for this chapter are the lack of additional exact co-locations between sensors and the reference instruments, which would obviate the need for the computationally costly spatiotemporal weighting and also provide a larger training set. Although the spatiotemporal weighting was able to provide an increase in sample size over requiring an exact time and space match, additional exact co-location data would allow for additional testing as well, measuring error by location of reference instrument or time of day for example. Further improvements on the model could be made in by incorporating recent measurements explicitly, where $time_0$ is predicted by $time_{-1}$, $time_{-2}$, etc.

The exposure assessment framework presented in this paper was explicitly not about the actual CO exposure levels compared to the true PEL, as all concentrations measured suggest the concentrations were very well controlled. The goal was to demonstrate the utility of a modeling process that is substantially more complex than traditional calibration methods. Despite the increased technical costs, implementation of this framework is expected to provide immense time savings on laboratory and air sampling data collection and analysis. These time savings are on top of the ability to create highly accurate probabilistic predictions with models that can handle enormous amounts of data, both in terms of potential features and number of measurements. Based on the information provided, it should be possible to implement this

framework on already existing systems or as part of the design of an entirely new network. Lastly, the models demonstrated here could be relatively easily implemented with a live or streaming data pipeline that could serve as a real-time exposure assessment tool or early warning system across an entire facility.

6. Conclusions

Using probabilistic gradient boosted decision trees is an effective way to calibrate an indoor occupational low-cost sensor network. In the occupational setting these probabilistic models prevent both the need for lab calibration and also the need to collocate reference instruments at all sampling points. They also create predictions that can be accurately used to conduct spatial probabilistic exposure assessments following the American Industrial Hygiene exposure ratings guidelines. Additionally, the model used in this chapter demonstrated how utilizing simple means or medians and small sample sizes are not sufficient for high quality occupational exposure assessments as exposure misclassification is likely. Finally, these models could be used to predict on live streaming sensor data, creating a real-time exposure assessment tool.

7. References

Afshar-Mohajer N, Zuidema C, Sousan S, Hallett L, Tatum M, Rule AM, et al. 2018. Evaluation of low-cost electro-chemical sensors for environmental monitoring of ozone, nitrogen dioxide, and carbon monoxide. J Occup Environ Hyg 15:87–98; doi:10.1080/15459624.2017.1388918.

Baddeley A, Rubak E, Turner R. 2015. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press:London.

Barrett L, Hoyer S, Kleeman A, O'Kane D. 2015. properscoring.

Bergstra J, Bengio Y. 2012. Random Search for Hyper-Parameter Optimization. J Mach Learn

Res 13: 281–305.

Berman JD, Peters TM, Koehler KA. 2018. Optimizing a sensor network with data from hazard mapping demonstrated in a heavy-vehicle manufacturing facility. Ann Work Expo Heal 62:547–558; doi:10.1093/annweh/wxy020.

Borrego C, Ginja J, Coutinho M, Ribeiro C, Karatzas K, Sioumis T, et al. 2018. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise – Part II. Atmos Environ 193:127–142; doi:10.1016/j.atmosenv.2018.08.028.

Buehler C, Xiong F, Levy Zamora M, Skog K, Kohrman-Glaser J, Colton S, et al. 2020. Stationary and Portable Multipollutant Monitors for High Spatiotemporal Resolution Air Quality Studies including Online Calibration. Atmos Meas Tech in review.

Bullock W, Ignacio J, Mulhausen J. 2006. *A Strategy for Assessing and Managing Occupational Exposures*. 2nd Editio. American Industrial Hygiene Association Press:Fairfax, VA.

Datta A, Saha A, Zamora ML, Buehler C, Hao L, Xiong F, et al. 2020. Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmos Environ 242:117761; doi:10.1016/j.atmosenv.2020.117761.

Duan T, Avati A, Ding DY, Thai KK, Basu S, Ng AY, et al. 2019. NGBoost: Natural Gradient Boosting for Probabilistic Prediction.

Gao M, Cao J, Seto E. 2015. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environ Pollut 199:56–65; doi:10.1016/j.envpol.2015.01.013.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378; doi:10.1198/016214506000001437.

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;

doi:10.1016/j.atmosenv.2015.04.057.

Hewett P, Logan P, Mulhausen J, Ramachandran G, Banerjee S. 2006. Rating exposure control using Bayesian decision analysis. J Occup Environ Hyg 3:568–581; doi:10.1080/15459620600914641.

Koehler K, Peters T. 2013. Influence of Analysis Methods on Interpretation of Hazard Maps. Ann Occup Hyg 57: 558–570.

Koehler K, Volckens J. 2011. Prospects and pitfalls of occupational hazard mapping: “between these lines there be dragons”. Ann Occup Hyg 55: 829–840.

Kuhn M, Johnson K. 2013. *Applied predictive modeling*. Springer New York.

Levy Zamora M, Xiong F, Gentner D, Kerkez B, Kohrman-Glaser J, Koehler K. 2019. Field and Laboratory Evaluations of the Low-Cost Plantower Particulate Matter Sensor. Environ Sci Technol 53:838–849; doi:10.1021/acs.est.8b05174.

Lim CC, Kim H, Vilcassim MJR, Thurston GD, Gordon T, Chen LC, et al. 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ Int 131:105022; doi:10.1016/j.envint.2019.105022.

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; doi:10.1016/j.envint.2018.04.018.

NIOSH. 2019. CDC - NIOSH Pocket Guide to Chemical Hazards - Carbon Monoxide. NIOSH Pocket Guid to Chem Hazards. Available: https://www.cdc.gov/niosh/npg/npgd0105.html.

OSHA. 2001. Appendix B to the Formaldehyde Standard. Code of Federal Regulations 29, Part 1910.1048.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825–2830.

Piedrahita R, Xiang Y, Masson N, Ortega J, Collier A, Jiang Y, et al. 2014. The next generation

of low-cost personal air quality sensors for quantitative exposure monitoring. Atmos Meas Tech 7:3325–3336; doi:10.5194/amt-7-3325-2014.

R Core Team. 2020. R: A Language and Environment for Statistical Computing.

Ramachandran G. 2005. *Occupational Exposure Assessment for Air Contaminants*. 1st Editio. CRC Press:Minneapolis.

Rappaport SM. 1984. The rules of the game: An analysis of Osha's enforcement strategy. Am J Ind Med 6:291–303; doi:10.1002/ajim.4700060407.

Schapire RE. 2003. The Boosting Approach to Machine Learning: An Overview. Springer, New York, NY. 149–171.

Snyder EG, Watkins TH, Solomon PA, Thoma ED, Williams RW, Hagler GSW, et al. 2013. The changing paradigm of air pollution monitoring. Environ Sci Technol 47:11369–11377; doi:10.1021/es4022602.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. 2009. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21:n/a-n/a; doi:10.1002/env.1014.

Thomas G, Sousan S, Tatum M, Liu X, Zuidema C, Fitzpatrick M, et al. 2018. Low-Cost, Distributed Environmental Monitors for Factory Worker Health. Sensors 18:1411; doi:10.3390/s18051411.

Tornero-Velez R, Symanski E, Kromhout H, Yu RC, Rappaport SM. 1997. Compliance Versus Risk in Assessing Occupational Exposures. Risk Anal 17:279–292; doi:10.1111/j.1539-6924.1997.tb00866.x.

Tuggle RM. 1981. The NIOSH decision scheme. Am Ind Hyg Assoc J 42:493–498; doi:10.1080/15298668191420134.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Fourth Edi. Springer, New York, NY.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc 73: 3–36.

Wu H, Godfrey J, Govindaraju K, Pirikahu S. 2015. ExtDist: Extending the Range of Functions for Probability Distributions. R package version 0.6-3.

Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, et al. 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech 11:291–313; doi:10.5194/amt-11-291-2018.

Zuidema C, Sousan S, Stebounova L V., Gray A, Liu X, Tatum M, et al. 2019a. Mapping Occupational Hazards with a Multi-sensor Network in a Heavy-Vehicle Manufacturing Facility. Ann Work Expo Heal 63:280–293; doi:10.1093/annweh/wxy111.

Zuidema C, Stebounova L V., Sousan S, Gray A, Stroh O, Thomas G, et al. 2019b. Estimating personal exposures from a multi-hazard sensor network. J Expo Sci Environ Epidemiol 30:1013–1022; doi:10.1038/s41370-019-0146-1.

Zuidema C, Stebounova L V., Sousan S, Thomas G, Koehler K, Peters TM. 2019c. Sources of error and variability in particulate matter sensor network measurements. J Occup Environ Hyg 16:564–574; doi:10.1080/15459624.2019.1628965.

8. Appendix

8.1. (A) Modeling and Spatial Analysis Packages and Libraries

8.1.1. Python 3.8.4

ngboost==0.2.2.dev0

numpy=1.19.0

pandas=1.0.5

properscoring==0.1

scikit-learn==0.23.1

scipy==1.5.1

statsmodels==0.11.1

8.1.2. R 4.0.2 “Taking Off Again”

ggmap_3.0.0

MASS_7.3-52

mgcv_1.8-32

nlme_3.1-149

numDeriv_2016.8-1.1

optimx_2020-4.2

rdist_0.0.5

rgdal_1.5-16

rpart_4.1-15

sf_0.9-5

sp_1.4-2

spatstat_1.64-1

tidyverse_1.3.0

zoo_1.8-8

8.2. (B) NGBoost Hyperparameter Search Grid

```
base_learner = DecisionTreeRegressor(criterion = 'mse')
param_grid = {
    'Base': [base_learner],
    'Base__max_depth': list(range(2, 200, 10)),
    'Base__max_features': ['auto'],
    'Base__min_samples_leaf': list(range(1, 200, 10)),
    'Base__min_samples_split': list(range(2, 200, 10)),
    'n_estimators': list(range(500, 3000, 500)),
    'minibatch_frac': [1.0, 0.5]
     }
```

CHAPTER V: CONCLUSION

Summary Findings

1) Determine benzene exposure and cancer risk from commercial gasoline station filling operations among non-occupational and occupational groups

The goal of Chapter II was to determine benzene exposures and cancer risks for consumer and occupational groups from commercial gasoline station filling activities. We sampled 34 individuals (32 valid samples) while they filled up their vehicles at commercial gas stations from August 2019 through March 2020. All were analyzed for benzene, toluene, ethylbenzene, xylene, and total volatile organic compounds (TVOC). All samples were below the OSHA PEL and NIOSH REL for ethylbenzene, toluene, and xylene. Thirty-two of the samples were below the OSHA PEL for benzene, and 31 were below the benzene NIOSH REL. When combined with demographic information captured in surveys, the probabilistic risk assessment for benzene related cancer for consumers averaged three orders of magnitude below the relevant 1:1,000,000 excess risk management level, with zero exceedances. For the occupational scenario, less than 0.01 percent of the probabilistic trials exceeded the 1:10,000 excess risk management limit, with an average of 65 times lower than the limit. These low excess risk values for the consumers and occupational scenarios indicate that there is minimal excess risk for pumping operations at commercial gasoline stations.

2) Provide a probabilistic modeling framework for $PM_{2.5}$ exposure assessments with low-cost sensor networks that also reduces the need for lab calibrations

The aim of Chapter III was to build a probabilistic machine learning calibration model for low-cost sensor networks that outperformed traditional linear regression approaches, reduced the need for laboratory calibrations, and could be used in novel probabilistic exposure assessments. We built a probabilistic gradient boosted decision tree (GBDT) using NGBoost that outperformed traditional linear regression as quantified by RMSE and CRPS for modeling a low-cost sensor to a reference standard. In terms of RMSE, the GBDT averaged 2.7 $\mu g/m^3$ across all test splits and the linear regression averaged 2.9 $\mu g/m^3$, and in terms of CRPS, the GBDT averaged 1.5 $\mu g/m^3$ and the linear regression averaged 2.2 $\mu g/m^3$. In addition, the GBDT utilized raw data, as opposed to the lab corrected data of the linear regression, completely eliminating the initial calibration step. The GBDT was also fully probabilistic such that each prediction has a modeled mean and standard deviation unlike the linear regression that provides a single variance value across all predictions. The prediction distributions were then leveraged with an inverse distance weighting Monte Carlo simulation to conduct spatial exposure assessments that contained modeled and interpolated uncertainty to produce a probabilistic output with a mean and standard deviation of exposure aggregated to the neighborhood level.

3) Within the context of the American Industrial Hygiene Associations' exposure rating categories, develop a probabilistic calibration model for CO low-cost sensor networks to assist with regulatory decision making, while also demonstrating the improved utility of probabilistic models over point predictions for occupational exposure estimation.

The goal of Chapter IV was to create a probabilistic machine learning calibration model for CO on an occupational low-cost sensor network that would be able to be integrated with the AIHA exposure ratings framework and at the same time show the failings of a traditional compliance

sampling (mean/median focused) approach. NGBoost was used to create a fully probabilistic calibration model from raw sensor readings in millivolts to a reference concentration. Spatiotemporal weighting on millions of measurements was tuned and optimized to reduce computational costs and increase training and test set sizes. The model had an RMSE of 0.36 ppm predicting on test data similar to the training set and approximately 1.60 ppm on fully out of sample test data. The predictions were interpolated using a Monte Carlo framework developed in Chapter III and were used to create both probabilistic hazard maps of CO concentrations and AIHA exposure rating hazard maps. Lastly, the probabilistic model demonstrated how using a simple point prediction for exposure estimation is likely to underestimate and mischaracterize peak spatial exposures.

Future Research & Public Health Implications

Future research is needed to collect individual exposures on a chronic level to fully characterize exposures across time and space. While single canisters/samples can inform about single task exposures, the chronic or day-to-day exposures remain poorly characterized beyond large scale ambient information like NATA. Combinations of in-home and personal networked monitors that push data to the cloud continuously should be used in conjunction with conventional whole air sampling to provide a full 24-hour picture of exposures. This data should be collected with the plan of using more advanced and computationally intensive modeling methods such as neural networks that can learn relationships in high dimensional data to predict adverse health outcomes. While BTEX and TVOC were studied in Chapter II, additional pollutants of concern could be analyzed using this framework.

In terms of probabilistic exposure assessments for low-cost sensor networks, additional research is needed in terms of understanding how sensor drift impacts measurement and

modeling error. Specifically, characterizing both the drift and accuracy/precision on a per-sensor level as part of the modeling process remains an area needing additional research. Lastly, how and when to ensemble high variance/low bias and low variance/high bias models when utilizing limited data sets would allow sensor networks to be deployed and used quickly without having to wait for many months for enough data to run machine learning models without the risk of overfitting. Using environmental low-cost networks with probabilistic/machine learning calibration models, while more computationally difficult to create than linear regression models, can produce more accurate sensor predictions that allow for more nuanced understandings of health effects resulting from environmental exposures. Additionally, while the development of these models is non-trivial from a time perspective, their development can be conducted without the need for specialized laboratory facilities, and their flexibility and portability would also contribute to significant time savings upon traditional approaches. Lastly, while only $PM_{2.5}$ and CO were used as example pollutants in Chapter III and Chapter IV, there is no reason why this approach would not work on a variety of pollutants measured by a networked sensor, as long as appropriate interfering parameters are available. In particular, compounds such as ozone are difficult to calibrate with traditional methods and machine learning approaches could learn the complex non-linear relationships, allowing previously unusable data to be modeled with accuracy.

CURRICULUM VITAE

ANDREW N. PATTON

PERSONAL DATA

Department of Environmental Health and Engineering
Johns Hopkins University
Bloomberg School of Public Health
615 North Wolfe Street
Baltimore, Maryland 21205
Phone: 610-620-5576
Email: andrew.patton@jhu.edu
URL : www.andrewpatton.org

EDUCATION AND TRAINING

PhD	2021	Johns Hopkins Bloomberg School of Public Health	Exposure Science and Environmental Epidemiology
MS	2014	University of San Francisco	Environmental Management
BS	2011	University of California, Berkeley	Molecular Toxicology

PROFESSIONAL EXPERIENCE

Consultant
AP Analytics
2016 – 2020

Epidemiology Research Associate
Epidemiology Division
Marin County Health and Human Services
2017 – 2019

Associate Health Scientist
Cardno ChemRisk
2011-2016

PUBLICATIONS

Submitted

1. Innes GK, Nachman KE, Abraham AG, Casey AJ, **Patton, AN**, Price AB, Tartof SY, Davis MF. *United States organic and conventional meat associations with multi-drug resistant organisms.* Environmental Health Perspectives. *In Review,* 2020.

2. **Patton, AN.**, Levy-Zamora, M., Fox, M., Koehler, K. *Benzene exposure and cancer risk from commercial gasoline station fueling operations using a novel self-sampling protocol.* International Journal of Environmental Research and Public Health. *In Review, 2021.*

Articles, Editorials, and Other Publications not Peer Reviewed

1. **Patton, A.**, Ereman, R., Willis, M., Hannah, H., Arambula, K. Development of Text-Based Algorithm for Opioid Overdose Identification in EMS Data. Online J Public Health Inform 11(1): e28. 2019.

Interviews and Articles for Print and Internet Media

1. Hoffman, R. and **Patton, A.** Sixers' new shooting gravity: Creating space for Joel Embiid and Ben Simmons. The Athletic. December 11, 2020. https://theathletic.com/2251603/2020/12/11/sixers-new-shooting-gravity-creating-space-for-joel-embiid-and-ben-simmons/

2. Patton, A. Bleggi, A. Nuckols, G. What Factors Influence Injury Risk in Powerlifters. Stronger by Science. February 17, 2020. https://www.strongerbyscience.com/powerlifting-injuries-factors/

3. Patton, A. What exactly is load management?. Nylon Calculus. November 8, 2019. https://fansided.com/2019/11/08/nylon-calculus-load-management/

4. ClinicalAthlete Podcast. Injuries in Powerlifting – Crunching the Numbers with **Andrew Patton**. Episode 44. July 2019.

5. Patton, A. How can we visualize a player's shooting gravity?. Nylon Calculus. July 22, 2019. https://fansided.com/2019/07/22/nylon-calculus-visualizing-nba-shooting-gravity/

6. No-Lift Powerlifting Podcast. **Andrew Patton** (Stronger by Science). Episode 70. June 2019.

7. **Patton, A**., Bleggi, A., Nuckols, G. Injuries in Powerlifting: Basic Results. Stronger by Science. June 2019. https://www.strongerbyscience.com/powerlifting-injuries-results/

8. **Patton, A**., Bleggi, A., Nuckols, G. Injuries in Powerlifting: Background and Overview. Stronger by Science. December 2018. https://www.strongerbyscience.com/powerlifting-injuries-background/

Presentations

1. "Machine Learning and EMS Data for Opioid Overdose Surveillance". University of Miami Department of Public Health Sciences. Distinguished Lecture Series. February 17, 2020.

CURRICULUM VITAE

ANDREW N. PATTON

PART II

TEACHING

Classroom Instruction

Teaching Assistant:

PH.182.615 “Airborne Particles”. Topics included: properties of aerosols, uniform particle motion, particle size statistics, acceleration and curvilinear motion, Brownian motion/diffusion, filtration, aerosol sampling, optical properties, electrical properties, respiratory deposition.

- 3rd term, 2018
- 1st term, 2019 (online)

PH.185.621 “Spatial Analysis III: Spatial Statistics”. Topics include: geostatistics, point pattern data analysis, area level data analysis, interpolation.

- 3rd term, 2018

AS.280.101 “Introduction to Public Health”. Topics included: major causes of morbidity and mortality, the socioeconomic, behavioral, and environmental factors that affect health, the analytical methods used in the field, the role of government in protecting the public’s health

- 1st Term, 2019

Guest Lecturer:

Other Universities:

“How to Explain Yourself Graphically”. Invited speaker for Data Science in Sports, Brigham Young University, presented electronically, October 2020.

“Practical Modeling”. Invited speaker for Advanced Data Analysis, University of San Francisco. April 2019.

“Geostatistics”. Invited speaker for Advanced Quantitative Methods. University of San Francisco. April 2018.

PRESENTATIONS

Invited Lectures:

May 2019	"Preventing the Next Overdose" Speaker at Opioid Data Task Force Meeting – California Department of Public Health.
March 2019	"DIY Opioid Surveillance and Outreach System" Speaker at Syndromic Surveillance and Public Health Emergency Preparedness, Response and Recovery Committee – International Society for Disease Surveillance.
November 2018	"Opioid Overdose Surveillance and Classification with R" Speaker at R Group for Biosurveillance – International Society for Disease Surveillance.

Scientific Meetings:
*Indicates Conference Presenter

May 2020	***Patton, A**., Zamora-Levy, M., Fox, M., Koehler, K. Development of Exposure Factors for Benzene, Toluene, Ethyl-Benzene, and Xylene (BTEX) from Total Volatile Organic Compounds (TVOC) in Automotive Gasoline. Transportation, Air Quality, and Health Symposium. Riverside, CA. (Cancelled due to COVID-19).
June 2019	Hannah, H., Arambula, K., **Patton, A.,** Hansen, R. Willis, M., Ereman, R. Cost-effectiveness of Offering Treatment to Non-Fatal Opioid Overdoses Encountered by Emergency Medical Services (EMS) in Marin County, California. Council of State and Territorial Epidemiologists Annual Meeting. June 2019. Breakout Presentation. Arambula, K., Hannah, H., **Patton, A.,** Willis, M., Ereman, R., Preventing the Next Overdose: An Emergency Medical Services-Based Non-Fatal Opioid Overdose Surveillance and Telephone Outreach Pilot Program. Council of State and Territorial Epidemiologists Annual Meeting 2019. Breakout Presentation.

February 2019 ***Patton, A**. Leukemia Risk Assessment Approximation from Commercial Gasoline Station Benzene Exposures. Abstract at the Center for Advancing Research in Transportation Emissions, Energy, and Health (CARTEEH) Symposium, Austin, TX.

***Patton, A.** Development of Text-Based Algorithm for Opioid Overdose Identification in EMS Data. International Society for Disease Surveillance Annual Meeting. Oral Presentation. Honorable Mention for Best Student Submission.

April 2018 ***Patton, A.** Dementia needs assessment and case projections for Marin County, California (2015-2045). Abstract 11th Annual Research on Aging Showcase. Johns Hopkins Bloomberg School of Public Health. Baltimore, MD.

May 2016 ***Patton, A.** Upstream Land Use and Surface Water Pesticide Concentrations in the Salinas River Watershed. Abstract at the Northern California Society of Environmental Toxicology and Chemistry Annual Meeting, Oakland, CA.

November 2015 ***Patton, A.**, Saah, D. Organophosphate Pesticides, Surface Water, and California Agriculture. Abstract at the Society of Environmental Toxicology and Chemistry Annual Meeting, Salt Lake City, UT.

March 2012 Donovan, E., Grespin, M., Cyrs, W., **Patton, A.,** and Finley, B. Airborne Asbestos Concentrations During Work Involving Asbestos-Containing Floor Tiles: A Review of the Published and Unpublished Literature. Abstract at International Society of Exposure Sciences Annual Meeting. San Francisco, CA

ADDITIONAL INFORMATION

Personal Statement of Research and Research Objectives

My research goals are to improve exposure assessments and risk assessments by utilizing novel machine learning techniques to improve low-cost sensor networks. analyze data. I have developed novel exposure assessment frameworks for occupational and environmental low-cost sensor networks. The frameworks utilize probabilistic machine learning (gradient boosted decision trees) calibration models and Monte Carlo simulation during spatial interpolation to produce probabilistic exposure assessment maps/data products that incorporate modeling and interpolation uncertainty. Furthermore, the probabilistic models significantly outperform traditional linear models from a predictive capacity and do so without any need for laboratory correcting or calibrating the data ahead of time. I am interested in future research that implements deep learning to assist with out-of-scope predictions on low-cost sensor networks, and then utilizing the trained models in a streaming data scenario to create a real-time exposure assessment tool. Finally, I would also like to investigate how best to ensemble low-variance

high-bias and high-variance low-bias calibration models to take advantage of the strengths of both – particularly in a small data scenario, or when initially deploying a network.

Keywords

Occupational exposure assessment
Hazard mapping
Machine learning
R
Environmental exposure assessment
Air pollution
Data science
Python