ARTICLE

Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment

Magali N. Blanco[1]✉, Annie Doubleday[1], Elena Austin[1], Julian D. Marshall[2], Edmund Seto[1], Timothy V. Larson[1,2] and Lianne Sheppard[1,3]✉

BACKGROUND: Short-term mobile monitoring campaigns to estimate long-term air pollution levels are becoming increasingly common. Still, many campaigns have not conducted temporally-balanced sampling, and few have looked at the implications of such study designs for epidemiologic exposure assessment.
OBJECTIVE: We carried out a simulation study using fixed-site air quality monitors to better understand how different short-term monitoring designs impact the resulting exposure surfaces.
METHODS: We used Monte Carlo resampling to simulate three archetypal short-term monitoring sampling designs using oxides of nitrogen (NOx) monitoring data from 69 regulatory sites in California: a year-around Balanced Design that sampled during all seasons of the year, days of the week, and all or various hours of the day; a temporally reduced Rush Hours Design; and a temporally reduced Business Hours Design. We evaluated the performance of each design's land use regression prediction model.
RESULTS: The Balanced Design consistently yielded the most accurate annual averages; while the reduced Rush Hours and Business Hours Designs generally produced more biased results.
SIGNIFICANCE: A temporally-balanced sampling design is crucial for short-term campaigns such as mobile monitoring aiming to assess long-term exposure in epidemiologic cohorts.
IMPACT STATEMENT: Short-term monitoring campaigns to assess long-term air pollution trends are increasingly common, though they rarely conduct temporally balanced sampling. We show that this approach produces biased annual average exposure estimates that can be improved by collecting temporally-balanced samples.

Keywords: Air Pollution; Exposure Modeling; Environmental Monitoring; New Approach Methodologies (NAMs)

Journal of Exposure Science & Environmental Epidemiology (2023) 33:465–473; https://doi.org/10.1038/s41370-022-00470-5

INTRODUCTION

An increasing number of studies are using short-term monitoring campaigns to assess long-term air pollution levels [1–7]. Short-term mobile monitoring campaigns typically equip a vehicle with air monitors and collect samples while in motion (non-stationary sampling) and/or while stopped (stationary sampling). The focus of this analysis is on the latter mobile monitoring design. A single monitoring platform can be used to collect samples at many specified locations within a relatively short period of time, making it a time and cost-efficient sampling approach. Mobile campaigns are particularly well-suited for multi-pollutant monitoring of less frequently monitored traffic-related air pollutants that require expensive instruments or instruments that need frequent attention during the sampling period. And while a few studies have investigated the number of sampling locations and repeat samples needed to improve the resulting exposure surfaces from mobile monitoring campaigns [8, 9], to the best of our knowledge, none have considered the importance of conducting temporally-balanced sampling when the goal is estimation of an unbiased long-term average. This is particularly relevant for traffic-related pollutants since many experience strong diurnal and seasonal concentration trends [10, 11]. In general, many mobile monitoring campaigns have been short, lasting from a few weeks to months and with few repeat visits to each location spanning one to three seasons [1, 2, 6, 8, 11–34]. Most of these campaigns have conducted sampling during weekday business or rush hours, ignoring the surrounding hours, when air pollution concentrations can be drastically different.

The goal of this paper is to shed light on the temporal design of a short-term monitoring campaign for application to mobile monitoring for epidemiologic cohort studies. We carry out a set of simulation studies to better understand the role of monitoring design on the prediction of annual average surfaces. We use existing monitoring data from California to compare the primary, annual site averages when all the data are included to subsequent analyses utilizing subsets of the data. These data provide a unique

[1]Department of Environmental and Occupational Health Sciences, School of Public Health, University of Washington, Hans Rosling Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195, USA. [2]Department of Civil & Environmental Engineering, College of Engineering, University of Washington, 201 More HallBox 352700, Seattle, WA 98195, USA. [3]Department of Biostatistics, School of Public Health, University of Washington, Hans Rosling Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195, USA.
✉email: magali@uw.edu; sheppard@uw.edu

Received: 8 December 2021 Revised: 12 August 2022 Accepted: 15 August 2022
Published online: 31 August 2022

Table 1. Simulated sampling designs used to estimate site annual averages[a].

Design	Sampling Seasons	Sampling Days	Sampling Hours
Year-Around "Balanced" Design	Winter, spring, summer, fall	Mon–Sun	V1 (All Hours)[b] V2 (Most Hours): 5 AM–12 AM V3 (Truncated Hours): 6–9 AM, 1–5 PM, 8–10 PM
		5/7 weekday; 2/7 weekend samples	Random hours according to V1, V2, or V3
Two-Season Weekday "Rush Hours" Design	V4-5: winter & summer (2 wks/ season) V6-7: spring & fall (2-wks/ season)	Mon–Fri	7–10 AM, 3–6 PM
			Random Rush Hours according to V4-5 or V6-7
Two-Season Weekday "Business Hours" Design	V4-5: winter & summer (2 wks/season) V6-7: spring & fall (2 wks/ season)	Mon–Fri	9 AM–5 PM
			Random Business Hours according to V4-5 or V6-7

[a] V = version. See SI Table S1 for the exact sampling periods of the Business Hours and Rush Hours designs.
[b] The long-term sampling approach for this version produces gold standard (true) estimates.

opportunity to explore how short-term stationary sampling strategies can influence the resulting estimated annual-average concentration. Our analysis requires having a long-term, comprehensive set of measurement data, which therefore necessitates using fixed-site measurements rather than mobile measurements, to shed light on an aspect of study design for short-term stationary mobile monitoring.

METHODS

Data

We simulate three sampling designs (see below) using hourly observations for oxides of nitrogen (NOx) collected during 2016 from regulatory Air Quality System (AQS) sites in California. NOx was selected since it is a spatially and temporally variable traffic pollutant with a strong diurnal pattern [2, 35, 36], and it is measured at many regulatory monitoring sites in California, providing a large enough dataset for this analysis [37]. In sensitivity analyses we also consider NO and NO_2.

We included 69 of 105 California AQS sites that met various criteria (SI Fig. S3). First, sites needed to have readings at least 66% of the time (5,797/ 8,784 hourly samples; 2016 was a leap year). Second, sites needed to have sampling throughout the year, such that data collection gaps were a maximum of 45 days long. These two criteria are similar to other air quality studies [38–40]. Third, sites were required to have sampled for at least 40% of the time during various two-week periods that were used in two of our "common" designs (described below). This sample size ensured that we could sample during these periods without replacement. Fourth, sites were required to have positive readings (> 0 ppb) at least 60% of the time, thus ensuring that sites had sufficient variability in their concentrations and allowing us to model annual averages on the natural log scale. Finally, sites in rural and industrial settings (as determined by the US EPA) [41] were excluded since these do not represent where the majority of people reside. The resulting sites were in both urban and suburban settings, in residential and commercial areas.

Sampling designs

We conducted simulation studies to characterize the properties of three sampling designs (Table 1, Supplementary Information [SI] Fig. S1). Each design has a long- and a short-term sampling approach. Long-term approaches use all of the data that meet each design's definition to estimate site annual averages and are analogous to traditional, fixed-site sampling approaches where sampling at a given location occurs over an extended period of time. Short-term approaches only collect 28 samples per site (distributed evenly during each sampling season) and are analogous to mobile monitoring campaigns that collect a few repeat samples per site. (The cut-off of 28 samples reflects our preliminary analyses showing that 28 hourly NOx samples are sufficient to estimate a site's annual average within about 25% error or less [SI Fig. S2].) Each design has multiple versions where samples are collected at slightly different times. The various design versions are intended to reflect the bias produced if only certain times are included in the measurements. We simulated each short-term sampling approach 30 times (Monte Carlo resampling), and hereafter refer to each of these simulations as a "campaign" since each represents a potential mobile monitoring study.

The Year-Around "Balanced" Design represents an "ideal" sampling scheme: sampling is conducted during all seasons, days of the week, and all or most hours of the day. Version 1 collects samples during all hours of the day. Versions 2–3 reduce the sampling hours to reflect the logistical constraints of executing an extensive campaign: samples occur during most hours of the day (5 AM–12 AM only; "Version 2") or during 6–9 AM, 1–5 PM and 8–10 PM ("Version 3"). Estimates from the long-term Balanced Design Version 1 are analogous to what might be collected from a traditional, year-around, fixed-site sampling scheme. For simplicity, we interchangeably refer to these as the "true" estimates or the "gold standard" hereafter, though we acknowledge that some error exists (e.g., due to missing hours or instrument accuracy).

The Two-Season Weekday "Rush Hours" and "Business Hours" Designs reflect common designs in the literature [2, 3, 11, 23, 33, 34, 42, 43]. Samples are collected either during summer and winter (Versions 4–5) or spring and fall (Versions 6–7). Sampling for each version occurs on weekdays during the same two-week period for all sites during each relevant season (See SI Table S1 for each version's exact sampling periods). Sampling is restricted to the hours of 7–10 AM and 3–6 PM (Rush Hours Design) or 9 AM–5 PM (Business Hours Design). The short-term approach collects 14 random samples during each season.

Prediction models

We estimated unweighted site annual averages based on the data collected during each campaign. We log-transformed these before using them as the outcome variable in partial least squares (PLS) regression models, which summarized hundreds of geographic covariate predictors (e.g., land use, road proximity, and population density; see SI Table S2 for the covariates considered) into two PLS components (using the `plsr` function in the `pls` package in R). We evaluated the performance of each campaign using ten-fold cross-validated (CV) predictions on the native scale, incorporating re-estimation of the PLS components in each fold. The cross-validation groups were randomly selected and, importantly, fixed across all campaigns to allow for consistent model performance comparisons across designs.

To best understand the role of design, we present results for annual average estimates, predictions, and model performance statistics. In descriptive analyses, we compare design-specific annual average estimates and predictions to the gold standard. We compare predicted site concentrations against predictions from the gold standard since epidemiologic air pollution studies often rely on predicted exposure, and the gold standard prediction represents the best possible prediction of annual-average concentrations that a study could hope to achieve. We complement this approach with model assessment evaluations of design-specific site predictions against two different references: an assessment against the true averages, and a traditional model assessment evaluation against the respective design-specific annual average estimates.

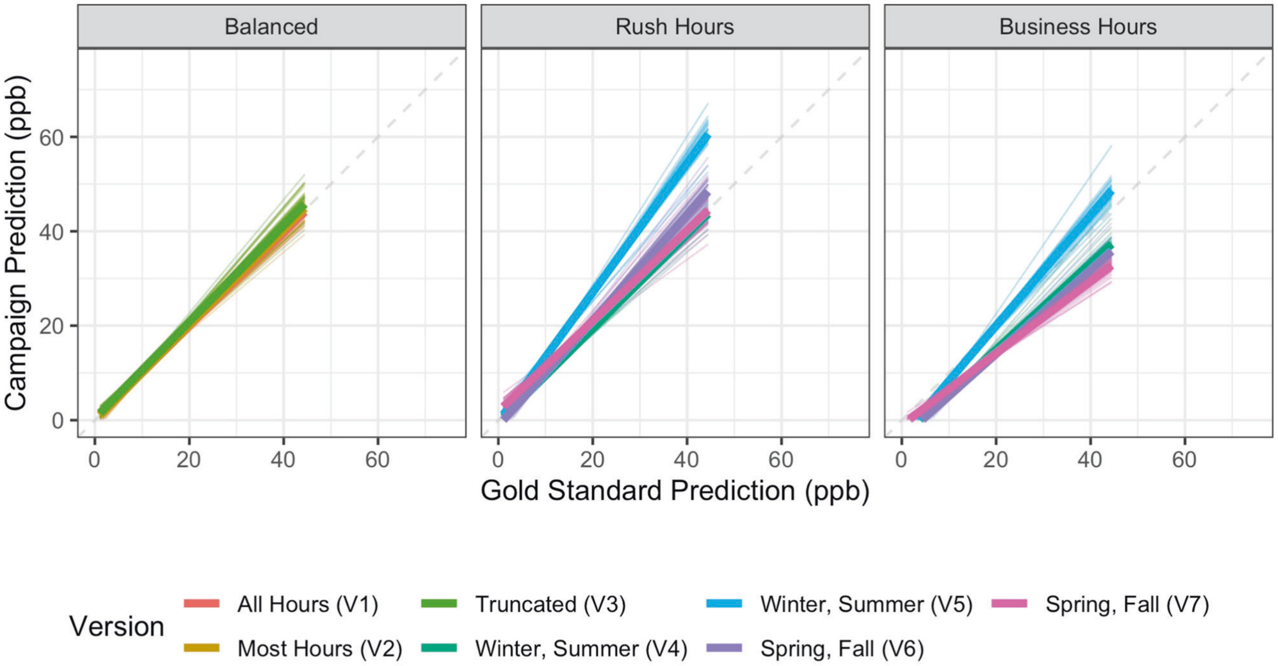

Fig. 1 Best fit lines of cross-validated short-term predictions for 30 campaigns vs the gold standard predictions for NOx. Thin transparent lines are individual campaigns, colored by design version; thicker lines are the overall version trend.

The traditional assessment compares the predicted exposures to the observed site measurements from which they were derived. This allows us to document the quantities that would normally be available from modeling the data measured from any specific campaign. We summarize the model performance in terms of cross-validated mean squared error (MSE)-based R^2 (R^2_{MSE}), regression-based R^2 (R^2_{reg}), and root mean squared error (RMSE). R^2_{MSE} assesses whether two sets of measurements such as estimates and predictions are the same (along the 1-1 line), and thus reflects both bias and variation around the one-to-one line (see SI Equations 1–3 for definitions). R^2_{reg}, on the other hand, assesses whether observations are linearly associated (based on the best fit line though not necessarily the 1-1 line) and thus adjusts for bias and slopes different than one. R^2_{reg} is defined as the squared correlation between two sets of measurements.

In sensitivity analyses, we repeated these simulations for nitrogen dioxide (NO_2) and nitrogen monoxide (NO), adding a two ppb constant to all of the hourly NO readings before log-transforming to eliminate negative and zero concentration readings. Furthermore, we conducted NOx simulations for a subset of sites ($N = 17$) within the Los Angeles (LA) and San Diego Counties, refitting PLS models to these sites alone. This region was meant to represent a potential area of interest for epidemiologic exposure assessment and one that could be more feasibly covered by a mobile monitoring campaign, though it had a reduced sample size.

Computing details are in SI Note S1.

RESULTS

Hourly readings

Sites ($N = 69$) had on average (SD) of 8090 (361) hourly readings, the equivalent of 337 [15] days of full sampling (See SI Table S3; note that this and many of the subsequent SI figures and tables also include results for NO and NO_2). Average (SD) hourly NOx concentrations were 16 [21] ppb (See SI Table S4). Sites had seasonal, daily, and hourly concentration patterns, with trends being more pronounced at some sites than others (See SI Figs. S4–S6).

Annual average estimates

Across the 69 monitor locations, gold standard annual average NOx concentrations had a median (IQR) of 14 [10–21] ppb and ranged from 3 to 56 ppb. Overall, the long-term and short-term sampling approach for each design had very similar distributions. The Balanced Design generally resulted in similar estimates as the true average; while the Rush Hours Design resulted in slightly higher annual averages; and the Business Hours Design resulted lower annual averages. See SI Table S5 and Fig. S7 for details.

Model predictions

The PLS model of the true annual average had a root mean square error (RMSE) of 7.2 ppb and a mean square error-based coefficient of determination (R^2_{MSE}) of 0.46.

We compared PLS model predictions from each short-term design to the gold standard model predictions. SI Fig. S9 shows the relative standard deviations of predictions by design, with 1 indicating that design predictions have the same standard deviation as the gold standard model predictions. Overall, the Balanced Design predictions have similar variability to those of the gold standard (range: 0.87–1.28), the Rush Hours Design predictions are more variable (range: 0.90-1.74), and the Business Hours Design predictions are mixed: some less and some more variable (range: 0.73–1.54). Figure 1 displays these comparisons as best fit lines. The scatterplots show that there are a few sites that have variable predictions in all designs. From the best fit lines, we observe that the short-term Balanced Design resulted in the most accurate predictions on average, as indicated by their overlapping general trends along the one-to-one line. The Rush Hours Design was more likely to have a positive general trend, while the Business Hours Design was more likely to have a negative general trend, indicating, for example, that higher concentrations were more likely to be over- or under-estimated, respectively. However, there was heterogeneity in this overall pattern across the various Rush and Business Hours Design versions. Furthermore, there was additional heterogeneity across individual campaigns. The SI contains comparable figures comparing design predictions to the gold standard and additional figures for NO and NO_2 (SI Figs. S10–S13).

Figure 2 shows site-specific comparisons of predictions across 30 short-term campaigns relative to the gold standard predictions for a stratified random sample of 12 sites in order to characterize relative bias (see SI Fig. S14 for all sites). Overall, the short-term Balanced Design predictions had a median (IQR) bias of 0.2 (−1 to

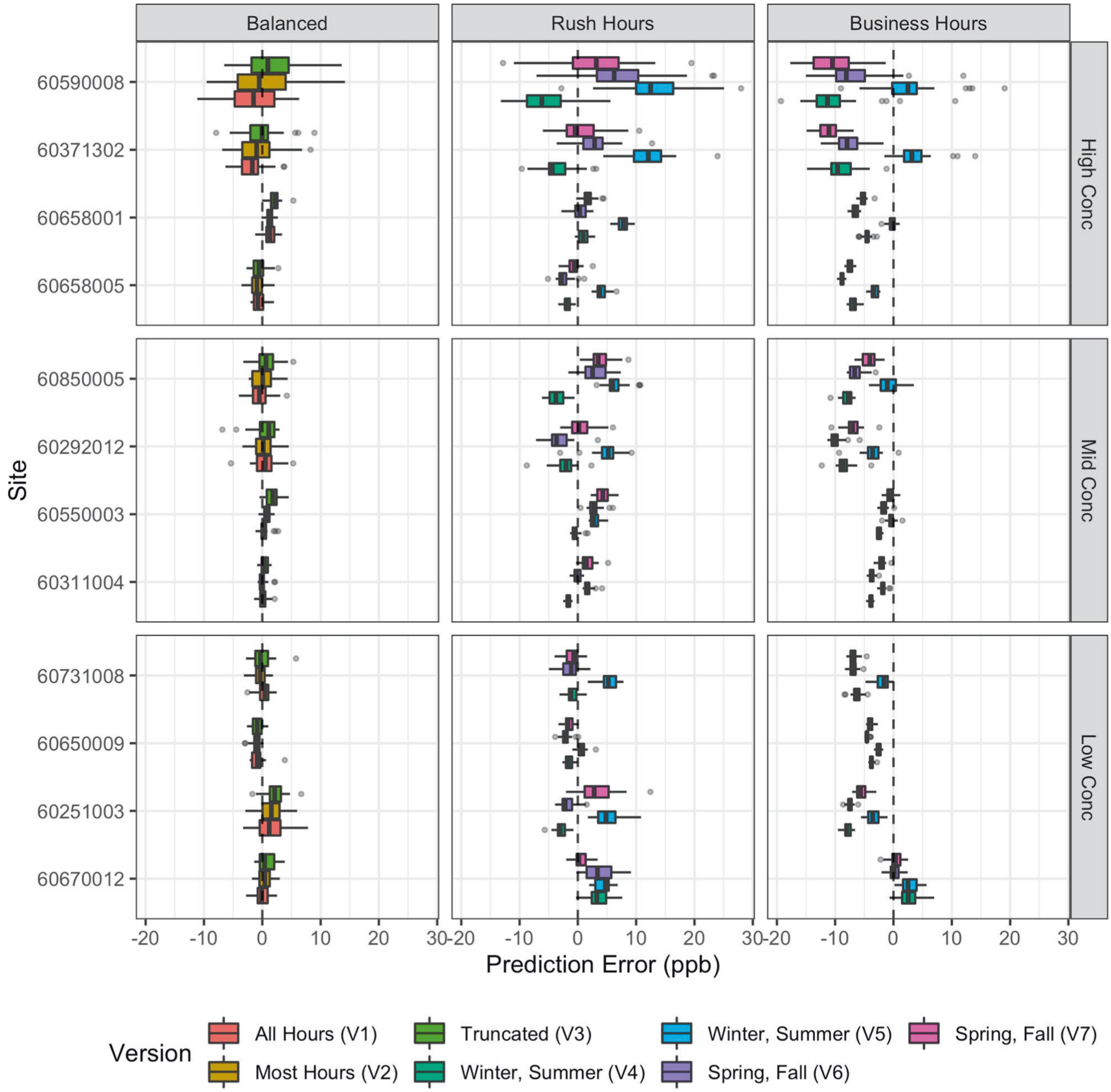

Fig. 2 Site-specific NOx prediction errors for short-term designs ($N = 30$ campaigns) as compared to the gold standard predictions (long-term Balanced Design Version 1). Showing a stratified random sample of 12 sites, stratified by whether true concentrations were in the low (Conc < 0.25), middle (0.25 ≤ Conc ≤ 0.75) or high (Conc > 0.75) concentration quantile and arranged within each stratum with lower concentration sites closer to the bottom.

1.4) ppb relative to the gold standard predictions (see SI Table S7 for details). All Balanced Design predictions were very similar to the gold standard predictions, though some sites frequently had larger biases. The Rush Hours and Business Hours Designs were more likely to consistently produce biased site predictions, with a median (IQR) bias of 1.2 (−1.2 to 4) ppb and −3.8 (−6.6 to −1.4) ppb, respectively. While the Rush Hours Designs generally resulted in higher predictions across sites (with some inconsistency across versions and sites), the Business Hours Design generally resulted in predictions that were lower than the gold standard predictions. There were also a few sites that tended to have more biased and/or more variable predictions relative to the gold standard across all designs. We observed similar patterns when looking at estimate (rather than prediction) biases (See SI Fig. S8).

Model assessment

Figure 3 shows the out-of-sample prediction performances relative to the observations from the true averages (left column) and the specific design (right column), for both the long-term and short-term approaches. The boxplots quantify the distribution of performance statistics across all 30 short-term campaigns while the squares show the performance of the long-term approach of the same design. When assessed against the true averages, the Balanced Design generally performs better than either the Rush Hours or Business Hours Design with higher CV R^2_{MSE} and CV R^2_{reg}, and lower CV RMSE estimates. This is particularly apparent for the long-term approach. Furthermore, within design, the long-term approach generally performs better than the majority of the short-term campaigns. There is considerable heterogeneity in

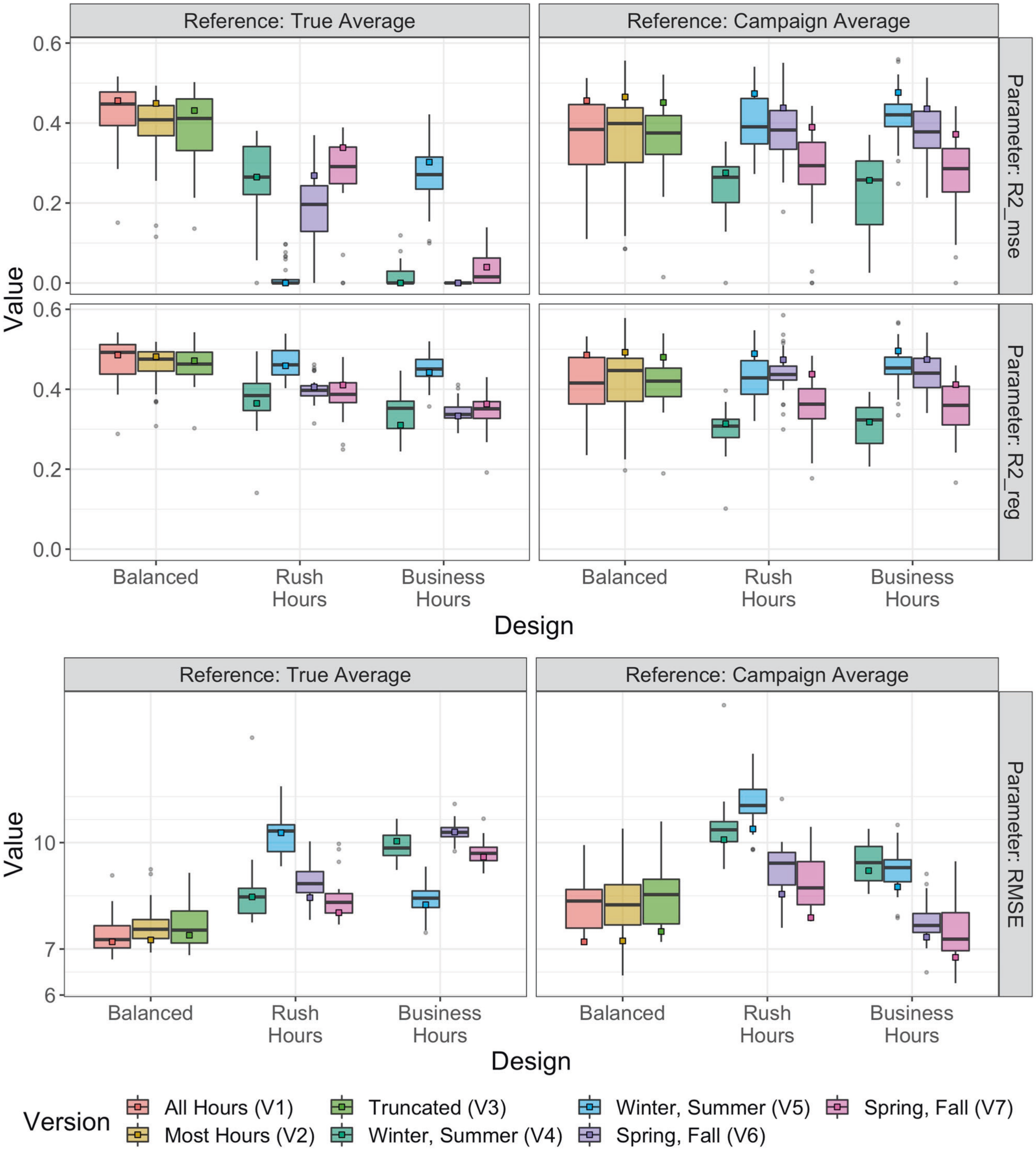

Fig. 3 Model performances. Performances (MSE-based R^2, regression-based R^2, and RMSE) are determined by each campaign's cross-validated predictions relative to: **a** the true averages (long-term Balanced Version 1), and **b** Its respective campaign averages. Boxplots are for short-term approaches (30 campaigns), while squares are for long-term approaches (1 campaign).

performance across the Rush Hours and Business Hours Design versions. In contrast, when assessed against observations from the same design, as would typically be done in practice, the role of sampling design on prediction performance is not as evident. The superior performance of the Balanced Design is not as apparent, and some of the Rush Hours and Business Hours Design versions appear to perform better. There are also a few campaigns that show poor performance, even under the Balanced Design. SI Fig. S15, S16 show similar results for NO_2 and NO, with NO showing more variability and some lower performing statistics. Stratifying by whether sites were considered to have high or low variability (based on hourly standard deviation estimates) showed similar R^2 and RMSE patterns (data not shown).

Sensitivity analyses

Findings were similar for sensitivity analyses for two other pollutants (see the SI for NO and NO_2 results presented alongside the NOx results) as well as within the smaller Los Angeles-San Diego region (see SI Fig. S20 and Table S8).

DISCUSSION

In this paper we have used existing regulatory monitoring data to deepen our understanding of the importance of short-term monitoring study design for application to stationary mobile monitoring for epidemiologic cohort studies. Others have shown that short-term data can be used to estimate long-term averages [2, 3]. What has been missing from the literature until now, however, is the impact of short-term monitoring study design on the accuracy and precision of long-term exposure estimates and model predictions. Our results indicate that for designs with a sufficient number of short-term samples at each location (about 28 or more), the design rather than the sampling approach (i.e., long- vs short- term) has the largest impact on the estimated annual averages. We focus the rest of this discussion on the short-term approaches for each design, which resemble mobile monitoring, though the long-term approaches produced similar results.

In terms of specific design, we found that the Balanced Design resulted in similar annual average predictions as those from the gold standard campaign. The similarity in annual averages and predictions across all of the Balanced Design versions suggests that campaigns with slightly reduced sampling hours (for example, due to logistical constraints) should to a large degree still produce unbiased annual averages at most sites. On the other hand, the Rush Hours Design was more likely to overpredict, while the Business Hours Design was more likely to underpredict site averages. These differences in results were likely because the Balanced Design captured much of NOx's temporal variability by allowing for samples to be collected during each season, day of the week, and all or most times of the day, all periods during which meteorology and traffic activity patterns impact air pollution concentrations (SI Figs. S4–S6). On the other hand, the Rush Hours Design was more likely to sample during high concentration times, while the Business Hours Design was more likely to sample during low concentration times (i.e., miss the rush hour times). Furthermore, we observed some prediction variability across the Rush Hours and Business Hours version, suggesting that the degree and direction of error is heavily impacted by the sampling window that happens to be selected. These conclusions were the same in the Los Angeles-San Diego sensitivity analysis, a geographic area that could more realistically be sampled by a mobile campaign.

At the site level, we saw that while any individual study campaign had the potential to produce biased estimates and predictions, the Rush Hours and Business Hours Designs were more likely to do so than the Balanced Design. The direction and magnitude of bias for each sampling design varied by site. This suggests a simple correction factor to adjust short-term measurements based on long-term observations at a small number of reference sites (for example using regulatory fixed sites), is unlikely to fully adjust for bias at the site level [44]. While many past campaigns have taken this approach to account for the fact that short-term stationary mobile sampling inherently misses some observations, this approach makes a strong assumption that all sites have the same temporal trends. SI Figs. S17–S19 illustrate the temporal trends for sites included in the Los Angeles-San Diego analysis and clearly shows how lower concentration "background" sites are also more likely to have less temporal variation when compared to other sites. Using these "background" sites (or any other site for that matter) to adjust readings at other sites would not substantially reduce the bias from an unbalanced sampling design. This may be especially pertinent for mobile monitoring campaigns since their increased spatial coverage is more likely to capture localized pollution hotspots that may have even more temporal variation. Sampling design should be prioritized, while temporal adjustment factors should be deemphasized or at least further investigated to establish their true value given their strong assumptions.

Furthermore, non-balanced designs may misrepresent some sites more than others and lead to differential exposure misclassification in epidemiologic studies since higher concentration sites were more likely to have greater degrees of bias and variation (Figs. 1, 2). While non-balanced designs may be appropriate for non-epidemiologic purposes including characterizing the spatial impact of traffic-related air pollutants during peak hours for urban planning and policy purposes, these could be misleading in epidemiologic applications.

In this study we were able to evaluate prediction model performance against the true annual average NOx exposure as well as against the observations typically available for model performance assessment. Performance assessment against the true averages indicates that the Balanced Design is clearly the best, and that there is little degradation in performance across versions of this design. This means it is possible to design high quality short-term stationary mobile monitoring studies that accommodate some measure of logistical feasibility, for example, by not requiring sampling in the middle of the night. In contrast, the performance of the Rush Hours and Business Hours Designs is comparatively worse, indicating that logistically appealing temporally limited sampling campaigns are inadequate for providing high quality annual averages. Further, the performance of these designs varies considerably and unpredictably depending upon the specific pair of two-week periods that are selected for sampling. Additionally, comparison of the two R^2 estimates (R^2_{MSE} and R^2_{reg}) indicates that not all of their poor performance is due to the inability to predict the same value as the truth (R^2_{MSE}), but due to systematic bias in the design. As noted earlier, R^2_{MSE} assesses whether two measurements are the same - along the 1-1 line, whereas R^2_{reg} simply assesses whether they are linearly associated.

Further, it is notable that the standard approach to model assessment, comparing model predictions to observations collected during the sampling campaign, doesn't clearly reveal the superior performance of the Balanced Design or the inherent flaws of the Rush Hours and Business Hours Designs. In fact, the Rush Hours and Business Hours Designs sometimes perform better than the Balanced Design when evaluated against the campaign's observations. This is because the evaluation doesn't take into account that the observations are biased because of the sampling design.

It is notable that occasionally there was an "unlucky" short-term campaign with meaningfully poorer performance than the other campaigns of the same design. This was more likely in the non-balanced designs. It may be possible that this result is driven by a few high-leverage outlier sites that impact the prediction model performance.

Our study focused on short-term stationary campaigns with 28 repeat samples per site. We did not consider campaigns with fewer or more visits. As evident in SI Fig. S2, the percent error in estimating the annual average from fewer than 25 visits is much higher, suggesting that site estimates will be considerably noisier in mobile campaigns with few repeat visits, regardless of the study design. Prediction model performance is thus likely to decrease as the number of visits per site decrease. Logistically, it is also difficult to achieve temporally balanced sampling with fewer than 28 site visits. Furthermore, we note that this study focused on a few generalizable, common designs in the literature, though other approaches have been taken.

In putting these results in context, it is important to recognize that in this simulation study we are using NOx hourly averages to approximate potentially shorter-term sampling durations that could be collected, for example, during a mobile monitoring campaign (e.g., a few minutes or less). Shorter duration sampling will affect the noise in the data. For comparison, however, our additional evaluations of minute-level data suggest that the decrease in percent error in going from two-minute to hour-long samples is at most a few percent because of serial correlation in the data. This thus gives us confidence that the findings from this work are still generalizable to more common, shorter-term stationary monitoring campaigns with sampling periods closer to a few minutes.

Further, our study took place throughout California, a large, geographically diverse area with varying climate profiles [45]. While such a large sampling domain would be challenging for a real-world monitoring campaign, the overall conclusions of this study – the importance of temporally-balanced sampling, are also supported in the Los Angeles-San Diego sensitivity analysis. In terms of the siting criteria for the regulatory monitoring sites where the data came from, locations are generally meant to capture representative population exposures, including near roadway, at various spatial scales ranging from microscales (< 100 m range) to regional scales in order to inform regulatory compliance [46, 47]. This should thus have provided us with decent spatial coverage and concentration variability. Many air pollution exposure studies intended for application to epidemiologic cohorts, in fact, rely on this network of regulatory monitors [48]. Still, when compared to most short-term monitoring campaigns, this study's larger domain and reduced exposure variability may have produced lower prediction model performances than would otherwise be expected.

Another distinction is that while we sampled measurements within sites at random, campaigns typically sample from sites along a fixed route or in a designated area. The actual sampling scheme will thus depend on the exact route developed and the number of platforms deployed, both of which are beyond the scope of this paper. In general, sampling along a route also induces some spatial correlation in the mobile monitoring data. This dependence is often overlooked in short-term monitoring campaigns and was not addressed in this study. Furthermore, we did not consider the importance of the distribution of sampling locations in this study, which is particularly relevant when the exposure assessment goal is an epidemiologic application. Selecting sites that are representative of the target cohort's residence locations will ensure the spatial compatibility assumption is met, which is an important way to reduce the role of exposure measurement error in epidemiologic inference [49].

Our evaluation focused on NOx, NO, and NO_2, which are quickly and moderately decaying air pollutants [35]. Campaigns that measure these pollutants may be more susceptible to sampling design than campaigns that measure less spatially- and/or temporally- variable pollutants such as $PM_{2.5}$. We selected NOx, NO, and NO_2 because these traffic-related pollutants are often measured in short-term campaigns, and data for these pollutants are more widely available. Non-criteria pollutants, for example ultrafine particles (UFP), however, have also received increasing attention in recent years given their emerging link to adverse health effects [50–53]. Still, high-quality information about their spatial distribution is essentially absent, and most studies have implemented short-term mobile sampling approaches that may not be temporally [54] balanced and potentially be misleading for application to epidemiologic inference.

An important next step in this work is to understand whether the differences in exposure estimates that we observed across study designs have a meaningful impact on epidemiologic inferences. This is of particular interest considering that year-around, balanced designs are resource-intensive and rare, while shorter and more convenient designs are more common in the literature.

CONCLUSIONS AND RECOMMENDATIONS FOR MOBILE MONITORING CAMPAIGNS

Short-term monitoring study design should be an important consideration for campaigns aiming to assess long-term exposure in an epidemiologic cohort. Given the temporal trends in air pollution, campaigns should implement balanced designs that sample during all seasons of the year, days of the week, and hours of the day in order to produce unbiased annual averages. Nonetheless, restricting the sampling hours in balanced designs will still generally produce unbiased estimates at most sites. On the other hand, unbalanced sampling designs like those often seen in the literature are more likely to produce biased annual averages, with some sites being more biased than others. And while predictions from these restricted designs may at times perform similarly to balanced designs (or, more problematically, may erroneously *appear* to perform similarly when evaluated against measurements which are themselves biased samples), this performance may strongly depend on the exact sampling period chosen and may thus be difficult or impossible to anticipate prior to conducting a new sampling campaign. Furthermore, the differential exposure misclassification that may result from these designs may be problematic in epidemiologic investigations. Finally, studies that implement unbalanced sampling designs are likely to have hidden exposure misclassification given that both the observations and model predictions may be systematically incorrect. By implementing a balanced sampling design, campaigns can increase their likelihood of capturing accurate annual averages.

DATA AVAILABILITY

Air pollution data are available through the EPA (https://www.epa.gov/outdoor-air-quality-data). The covariates used in this analysis for regulatory sites are freely available through various online sources and may be available from the authors upon request.

REFERENCES

1. Hankey S, Marshall JD. Land use regression models of on-road particulate air pollution (Particle Number, Black Carbon, PM2.5, Particle Size) using mobile monitoring. Environ Sci Technol. 2015;49:9194–202.
2. Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, et al. High-resolution air pollution mapping with google street view cars: Exploiting big data. Environ Sci Technol. 2017;51:6999–7008.
3. Hatzopoulou M, Valois MF, Levy I, Mihele C, Lu G, Bagg S, et al. Robustness of land-use regression models developed from mobile air pollutant measurements. Environ Sci Technol. 2017;51:3938–47.
4. Patton AP, Perkins J, Zamore W, Levy JI, Brugge D, Durant JL. Spatial and temporal differences in traffic-related air pollution in three urban neighborhoods near an interstate highway. Atmos Environ. 2014;99:309–21
5. Van den Bossche J, Peters J, Verwaeren J, Botteldooren D, Theunis J De, et al. Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset. Atmos Environ. 2015;105:148–61.
6. Kerckhoffs J, Hoek G, Messier KP, Brunekreef B, Meliefste K, Klompmaker JO, et al. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ Sci Technol. 2016;50:12894–902.
7. Xie X, Semanjski I, Gautama S, Tsiligianni E, Deligiannis N, Rajan TR, et al. A review of urban air pollution monitoring and exposure assessment methods. ISPRS Int J Geo-Inform. 2017;6:389.
8. Hatzopoulou M, Valois MF, Levy I, Mihele C, Lu G, Bagg S, et al. Robustness of land-use regression models developed from mobile air pollutant measurements. Environ Sci Technol. 2017;51:3938–47.
9. Messier KP, Chambliss SE, Gani S, Alvarez R, Brauer M, Choi JJ, et al. Mapping air pollution with google street view cars: Efficient approaches with mobile monitoring and land use regression. Environ Sci Technol. 2018;52:12563–72.
10. Yu CH, Fan Z, Lioy PJ, Baptista A, Greenberg M, Laumbach RJ. A novel mobile monitoring approach to characterize spatial and temporal variation in traffic-related air pollutants in an urban community. Atmos Environ. 2016;141:161–73.
11. Batterman S, Cook R, Justin T. Temporal variation of traffic on highways and the development of accurate temporal allocation factors for air pollution analyses. Atmos Environ. 2015;107:351–63.

12. Kerckhoffs J, Hoek G, Gehring U, Vermeulen R. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ Int. 2021;154:106569.
13. Minet L, Liu R, Valois MF, Xu J, Weichenthal S, Hatzopoulou M. Development and comparison of air pollution exposure surfaces derived from on-road mobile monitoring and short-term stationary sidewalk measurements. Environ Sci Technol. 2018;52:3512–9.
14. Abernethy RC, Allen RW, McKendry IG, Brauer M. A land use regression model for ultrafine particles in Vancouver, Canada. Environ Sci Technol. 2013;47:5217–25.
15. Farrell W, Weichenthal S, Goldberg M, Valois M-F, Shekarrizfard M, Hatzopoulou M. Near roadway air pollution across a spatially extensive road and cycling network. Environ Pollut. 2016;212:498–507.
16. Montagne DR, Hoek G, Klompmaker JO, Wang M, Meliefste K, Brunekreef B. Land use regression models for ultrafine particles and black carbon based on short-term monitoring predict past spatial variation. Environ Sci Technol. 2015;49:8712–20.
17. Patton AP, Zamore W, Naumova EN, Levy JI, Brugge D, Durant JL. Transferability and generalizability of regression models of ultrafine particles in urban neighborhoods in the boston area. Environ Sci Technol. 2015;49:6051–60.
18. Ragettli MS, Ducret-Stich RE, Foraster M, Morelli X, Aguilera I, Basagaña X, et al. Spatio-temporal variation of urban ultrafine particle number concentrations. Atmos Environ. 2014;96:275–83.
19. Rivera M, Basagaña X, Aguilera I, Agis D, Bouso L, Foraster M, et al. Spatial distribution of ultrafine particles in urban settings: A land use regression model. Atmos Environ. 2012;54:657–66.
20. Saraswat A, Apte JS, Kandlikar M, Brauer M, Henderson SB, Marshall JD. Spatiotemporal land use regression models of fine, ultrafine, and black carbon particulate matter in New Delhi, India. Environ Sci Technol. 2013;47:12903–11.
21. Simon MC, Patton AP, Naumova EN, Levy JI, Kumar P, Brugge D, et al. Combining measurements from mobile monitoring and a reference site to develop models of ambient ultrafine particle number concentration at residences. Environ Sci Technol. 2018;52:6985–95.
22. van Nunen E, Vermeulen R, Tsai M-Y, Probst-Hensch N, Ineichen A, Davey M, et al. Land use regression models for ultrafine particles in six European areas. Environ Sci Technol. 2017;51:3336–45.
23. Weichenthal S, Ryswyk KVan, Goldstein A, Bagg S, Shekkarizfard M, Hatzopoulou M. A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. Environ Res. 2016;146:65–72.
24. Weichenthal S, Van Ryswyk K, Goldstein A, Shekarrizfard M, Hatzopoulou M. Characterizing the spatial distribution of ambient ultrafine particles in Toronto, Canada: A land use regression model. Environ Pollut. 2016;208:241–8
25. Yu CH, Fan Z, Lioy PJ, Baptista A, Greenberg M, Laumbach RJ. A novel mobile monitoring approach to characterize spatial and temporal variation in traffic-related air pollutants in an urban community. Atmos Environ. 2016;141:161–73.
26. Saha PK, Li HZ, Apte JS, Robinson AL, Presto AA. Urban ultrafine particle exposure assessment with land-use regression: Influence of sampling strategy. Environ Sci Technol. 2019;53:7326–36.
27. Sabaliauskas K, Jeong CH, Yao X, Reali C, Sun T, Evans GJ. Development of a land-use regression model for ultrafine particles in Toronto, Canada. Atmos Environ. 2015;110:84–92.
28. Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in. Eur – ESCAPE Proj Atmos Environ 1994. 2013;72:10–23.
29. Wolf K, Cyrys J, Harciníková T, Gu J, Kusch T, Hampel R, et al. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Sci Total Environ. 2017;579:1531–40.
30. Cattani G, Gaeta A, Di Menno di Bucchianico A, De Santis A, Gaddi R, Cusano M, et al. Development of land-use regression models for exposure assessment to ultrafine particles in Rome, Italy. Atmos Environ. 2017;156:52–60.
31. Hoek G, Beelen R, Kos G, Dijkema M, Zee SC, van der, Fischer PH, et al. Land Use Regression Model for Ultrafine Particles in Amsterdam. Environ Sci Technol. 2011;45:622–8.
32. Saha PK, Hankey S, Marshall JD, Robinson AL, Presto AA. High-spatial-resolution estimates of ultrafine particle concentrations across the continental United States. Environ Sci Technol. 2021; Available from: https://doi.org/10.1021/acs.est.1c03237
33. Tessum MW, Sheppard L, Larson TV, Gould TR, Kaufman JD, Vedal S. Improving air pollution predictions of long-term exposure using short-term mobile and stationary monitoring in two US metropolitan regions. Environ Sci Technol. 2021;55:3530–8.
34. Saha PK, Zimmerman N, Malings C, Hauryliuk A, Li Z, Snell L, et al. Quantifying high-resolution spatial variations and local source impacts of urban ultrafine particle concentrations. Sci Total Environ. 2019;655:473–81.
35. Karner AA, Eisinger DS, Niemeier DA. Near-roadway air quality: Synthesizing the findings from real-world data. Environ Sci Technol. 2010;44:5334–44.
36. Riley EA, Banks L, Fintzi J, Gould TR, Hartin K, Schaal LN, et al. Multi-pollutant mobile platform measurements of air pollutants adjacent to a major roadway. Atmos Environ. 2014;98:492–9.
37. US EPA. Air Quality System (AQS). US Environmental Protection Agency. 2019 [cited 2019 Jun 4]. Available from: https://www.epa.gov/aqs
38. MESA Air. Data Organization and Operating Procedures (DOOP) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) and Associated Studies. MESA Air; 2019. Available from: https://www.uwchscc.org/MESAAP/Documents/MESAAirDOOP.pdf
39. Young MT, Bechle MJ, Sampson PD, Szpiro AA, Marshall JD, Sheppard L, et al. Satellite-Based NO2 and model validation in a national prediction model based on universal kriging and land-use regression. Environ Sci Technol. 2016;50:3686–94.
40. Keller JP, Olives C, Kim SY, Sheppard L, Sampson PD, Szpiro AA, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. Environ Health Perspect. 2015;123:301–9.
41. US EPA. AirData Pre-Generated Data Files. US Environmental Protection Agency. 2019 [cited 2019 Dec 7]. Available from: https://aqs.epa.gov/aqsweb/airdata/download_files.html
42. Minet L, Gehr R, Hatzopoulou M. Capturing the sensitivity of land-use regression models to short-term mobile monitoring campaigns using air pollution micro-sensors. Environ Pollut. 2017;230:280–90.
43. Saha PK, Li HZ, Apte JS, Robinson AL, Presto AA. Urban ultrafine particle exposure assessment with land-use regression: Influence of sampling strategy. Environ Sci Technol. 2019;53:7326–36.
44. Chastko K, Adams M. Assessing the accuracy of long-term air pollution estimates produced with temporally adjusted short-term observations from unstructured sampling. J Environ Manag. 2019;240:249–58.
45. Li L, Girguis M, Lurmann F, Pavlovic N, McClure C, Franklin M, et al. Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. Environ Int. 2020;145:106143.
46. US EPA. Electronic Code of Federal Regulations (e-CFR), Title 40, Chapter 1, Subpart C, Part 58, Appendix E to Part 58 - Probe and Monitoring Path Siting Criteria for Ambient Air Quality Monitoring. United States Environmental Protection Agency (US EPA); 2021 [cited 2021 Aug 4]. Available from: https://www.ecfr.gov/cgi-bin/retrieveECFR?n=40y6.0.1.1.6#ap40.6.58.0000_0nbspnbspnbsp.e
47. CARB. Annual Network Plan - Covering Monitoring Operations in 25 California Air Districts. California Air Resources Board (CARB); 2021. Available from: https://ww2.arb.ca.gov/sites/default/files/2021-07/2021%20Annual%20Network%20Plan_1.pdf
48. Hoek G. Methods for assessing long-term exposures to outdoor air pollutants. Curr Environ Health Rep. 2017;4:450–62.
49. Szpiro AA, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics. 2013;24:501–17.
50. Kilian J, Kitazawa M. The emerging risk of exposure to air pollution on cognitive decline and Alzheimer' s disease e Evidence from epidemiological and animal studies. Biomed J. 2018;41:141–62.
51. Lane KJ, Levy JI, Scammell MK, Peters JL, Patton AP, Reisner E, et al. Association of modeled long-term personal exposure to ultrafine particles with inflammatory and coagulation biomarkers. Environ Int. 2016;92–93:173–82.
52. Weichenthal S, Olaniyan T, Christidis T, Lavigne E, Hatzopoulou M, Van Ryswyk K, et al. Within-city Spatial Variations in Ambient Ultrafine Particle Concentrations and Incident Brain Tumors in Adults. Epidemiology. 2020;31:177.
53. US EPA. Integrated science assessment (ISA) for particulate matter (final report, Dec 2019). US Environ Prot Agency. 2019.
54. Riffault V, Arndt J, Marris H, Mbengue S, Setyan A, Alleman LY, et al. Fine and Ultrafine Particles in the Vicinity of Industrial Activities: A Review. Crit Rev Environ Sci Technol. 2015;45:2305–56.

AUTHOR CONTRIBUTIONS

MNB: data curation, formal analysis, software, investigation, methodology, software, visualization, writing – original draft and editing. AD: formal analysis, validation, visualization, writing – review, and editing. EA: writing – review, and editing. JDM: Conceptualization, writing – review, and editing. ES: writing – review, and editing. TL: Writing – review, and editing. LS: Conceptualization, methodology, funding acquisition, resources, supervision, writing – original draft, review, and editing.

FUNDING

This work was funded by the Adult Changes in Thought – Air Pollution (ACT-AP) Study (National Institute of Environmental Health Sciences [NIEHS], National Institute on Aging [NIA], R01ES026187), and BEBTEH: Biostatistics, Epidemiologic & Bioinformatic Training in Environmental Health (NIEHS, T32ES015459). Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. CR-83998101) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICAL APPROVAL

No humans or animals were involved in this work.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41370-022-00470-5.

Correspondence and requests for materials should be addressed to Magali N. Blanco or Lianne Sheppard.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.