Annals of Work Exposures and Health, 2022, Vol. 66, No. 5, 580–590
https://doi.org/10.1093/annweh/wxab105
Advance Access publication 26 November 2021
Original Article

Original Article

Probabilistic Machine Learning with Low-Cost Sensor Networks for Occupational Exposure Assessment and Industrial Hygiene Decision Making

Andrew N. Patton[1,*], Konstantin Medvedovsky[2], Christopher Zuidema[3], Thomas M. Peters[4] and Kirsten Koehler[1]

[1]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, USA; [2]QR Analytics, 3021 Cambridge Pl NW., Washington DC 20007, USA; [3]Department of Environmental and Occupational Health Sciences, University of Washington Hans Rosling Center for Population Health, 3980 15th Avenue NE, Seattle, WA 98195, USA; [4]Department of Occupational and Environmental Health, University of Iowa, 145 N. Riverside Drive, Iowa City, IA 52242, USA

*Author to whom correspondence should be addressed: Tel: +410-955-7706; e-mail: anpatt7@gmail.com

Submitted 10 March 2021; revised 18 October 2021; editorial decision 22 October 2021; revised version accepted 2 November 2021.

Abstract

Occupational exposure assessments are 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 exposure uncertainty and exposure variability for decision-making; however, the ratings can still fail to capture realistic exposure with small sample sizes. Therefore, low-cost sensor networks consisting of numerous lower-quality sensors have been used 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 readings to reference data highly accurately, entirely removing the need for laboratory calibration. Further, we indicate how the machine learning 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 for each measurement enables the use of the AIHA exposure ratings, which provide an enhanced industrial

What's Important About This Paper?

This study advances our knowledge about the performance and utility of low-cost sensor networks for exposure assessment. Specifically, this study demonstrates a method by which to efficiently and effectively manage sensor drift and calibration, and that exposure estimates based on sensor networks improve upon traditional compliance sampling methods and statistical analysis. Further, this study demonstrates that hazard maps that adhere to the AIHA exposure ratings framework can be used to optimize industrial hygiene resources.

decision-making framework as opposed to simply determining if a small number of measurements were above or below a pertinent occupational exposure limit. Lastly, we show how a probabilistic modeling exposure assessment with high spatiotemporal resolution data can prevent exposure misclassifications associated with traditional models that rely exclusively on mean or point predictions.

Keywords: carbon monoxide; exposure assessment; machine learning; occupational exposure assessment; sensor networks

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 indicates that samples should be taken as personal exposures for the highest risk workers, which are then compared against the PEL, although there is no requirement as to the number of samples taken. If the upper confidence limit of any samples (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). With small sample sizes, this type of compliance monitoring does little to represent true exposure concentrations in a workplace and usually results in underestimated exposures in facilities where concentrations are poorly managed and controlled, although overestimates are also possible (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 size estimation by acknowledging the likely lognormal distribution of the true contaminant exposures, and is, therefore, less likely to underestimate exposures (Rappaport, 1984; Tornero-Velez *et al.*, 1997; Tuggle, 1981). 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).

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). 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 may avoid the interpolation errors associated with low-resolution data (Koehler and Volckens, 2011; Koehler and Peters, 2013). These maps can provide industrial hygienists with exposure data to prioritize areas for intervention, within the hierarchy of controls. Personal exposures, collected within the breathing zone of workers, are the standard for occupational exposure assessment. However, such networks could also be used to provide estimates of personal exposures, while recognizing some potential for exposure misclassification due to strong gradients near sources and lower accuracy and precision of low-cost sensors compared to traditional industrial hygiene sampling equipment.

A major challenge associated with a sensor network is the need to calibrate each sensor to a reference standard, which ensures that the sensors provide reliable measurements across the range of possible environmental conditions and exposure settings (Afshar-Mohajer *et al.*, 2018; Borrego *et al.*, 2018; Datta *et al.*, 2020; Zamora *et al.*, 2019). Laboratory calibration is one method, but is highly time and equipment-intensive

(Zamora *et al.*, 2019; Zuidema *et al.*, 2019b). However, calibration models have been developed for use with environmental low-cost sensor networks that reduce or eliminate the need for pre-deployment laboratory calibrations or complex intra-deployment calibrations such as creating a range of models for each sensor and season (Borrego *et al.*, 2018; Datta *et al.*, 2020; Morawska *et al.*, 2018). In particular, non-linear/machine learning approaches (random forest, gradient boosted decision trees, neural networks, etc.) have had success modeling the non-linearity of the relationship between environmental conditions data and the contaminant of concern (Casey and Hannigan, 2018; Malings *et al.*, 2019; Zimmerman *et al.*, 2018).

We propose to build a calibration model that improves upon existing non-linear/machine learning approaches by using probabilistic gradient boosted decision trees that model both a unique mean and standard deviation for each prediction. This model will be built 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 this modeling approach 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 for laboratory calibrations and multiple linear calibration models. Further, we show how these models can be used to create probabilistic exposure hazard maps as well as hazard maps with the AIHA exposure rating framework. Lastly, we show how a probabilistic model highlights the weakness of exposure assessments that rely strictly on a mean or point prediction to determine occupational exposures.

Methods

Sensor network data

A network 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 have been previously described in detail (Thomas *et al.*, 2018). Each of the 40 monitors in the network contained a sensor for PM (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 were averaged to five-minute intervals and wirelessly transmitted and stored in a database from 4 August 2017 to 27 March 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 were used for the remainder of the study to demonstrate the application of the method. Data from Thanksgiving and the following Friday, Christmas Eve, Christmas Day, New Year's Eve, and New Year's Day were excluded as operations were drastically altered from normal during these times.

Reference data

The reference data were collected with CO instruments (EL-USB-CO, Lascar Electronics, Erie, PA, USA) at three periods throughout the year, the first on 17 August 2017 (period 1), the second on 20 December 2017 and 21 December 2017 (period 2), and the third on 23 March 2018 and 26 March 2018 (period 3), resulting in a total of five reference sampling days for the entire deployment (Zuidema *et al.*, 2019a). For each sampling period, the reference instruments were placed at a total of ten unique locations within the bounds of the monitoring network. Some instruments were collocated with sensor monitors (approximately 85% of reference measurements), whereas some were intentionally noncollocated (approximately 15% of reference measurements). Reference data were collected throughout the workday on each measurement day. The distribution of all reference data is provided in Supplementary Fig. S1.

Training and testing data

In order to determine the quality of the calibration model, two types of training and test splits were used. The first, or the 'Full' split, is all of the data split randomly such that 80% of the data is used to train the model and the remaining fully separate 20% is used to evaluate the model (Pedregosa *et al.*, 2011). The second split is trained on one of the three sampling periods of data and tested on the other two sampling periods. 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.

Spatiotemporal weighting

The total set of exactly time and space collocated reference and sensor data pairs were 1785 unique data points—a fairly small amount of data. Therefore, a weighting system was used to expand the sample size by pairing all sensor measurements with all reference measurements with the goal of utilizing this weighting scheme to adjust each pair's contribution to the model tuning. Following the pairing of all measurements, the sample size for training increased from the number of exactly collocated in time/space measurements (n = 1785 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). Supplementary Fig. S2 provides a visual representation of this process. 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 as $C^{\Delta t}$, with C, a numerical constant, taking possible values of 0.70, 0.75, 0.80, 0.85, 0.90, and 0.99, representing between 2% weight (i.e., $0.70^{10} = 0.02$) and 90% weight at a Δt of 10 min. 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. The time weight and the spatial weight were multiplied together to form the total weight for the sensor-reference pair. 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 non-collocated measurements. Measurements with a total weight of less than 1% were eliminated from the dataset prior to modeling to reduce computational costs. 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. The sizes of the training and test sets based on maximum and minimum weighting parameters are shown in Supplementary Table S2.

Gradient boosted decision trees

Gradient boosted decision trees (GBDT) are a machine learning methodology that uses numerous weak decision tree learners (i.e. a machine learning algorithm) 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). This study will use NGBoost, an open source python library developed by the Stanford Machine Learning Group that boosts a base learner, in this case, a decision tree (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 of a Normal distribution with mean x and standard deviation $\sigma(N(x,\ \sigma))$, are modeled for each prediction, as opposed to a model like linear regression that provides a modeled mean, but constant standard deviation for all predictions (Duan *et al.*, 2019). Additional details on the model tuning process, as well as the packages used in the models, are provided in the Supplementary Material.

Model features

Given that the sensor network was deployed in one contiguous climate-controlled indoor manufacturing floor, extreme outliers of environmental conditions (e.g. very high relative humidity or very cold temperatures) were interpreted as measurement errors. 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. Features for the model included temperature, relative humidity, time of the day, location, and the sensor reading. Additional information on model features is provided in Supplementary Material.

Model evaluation

The models were evaluated on two forms of predictive 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 Equation 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. The units of RMSE are those of the prediction, or in this case, ppm of CO.

Equation 1: Root Mean Square Error (RMSE):

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

While RMSE is used to evaluate a point prediction, CRPS is used to evaluate both the point prediction (mean) and the spread/uncertainty (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).

The results of the model evaluation were 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. This regression 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 was optimized for concentrations below 12 ppm where there was a linear response, whereas at concentrations greater than 12 ppm there was a non-linear response (Afshar-Mohajer *et al.*, 2018).

Spatial interpolation and hazard mapping

The predictions from the GBDT were used to create hazard maps via spatial interpolation of CO predictions across a rectangular 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 was conducted to interpolate over a range of possible values from the probabilistic output. 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 IDW error was determined using leave-one out cross validation (LOOCV) of IDW predictions (see online Supplementary Material for additional details). The IDW error was incorporated into a Monte Carlo simulation using all the$N_{NGBoost}(x,\ \sigma)$.

AIHA exposure ratings

In addition to visualizing percentiles of CO concentrations, the prediction distributions can be directly applied to the AIHA Exposure Rating framework (Ramachandran, 2005). Additionally, the flexibility of a custom OEL allows for selection of action thresholds that can serve as signal events before reaching a regulatory or health-relevant overexposure.

In order to generate exposure rating hazard maps, the 95th percentile (X_{95}) of$N_{IDW}(x,\ \sigma)$, the interpolated exposure distribution over all 223 days, at each location was calculated and assigned an exposure rating as shown in Table 1.

AIHA ratings are based on personal exposures, whereas the sensor network by definition measures area exposures. However, following interpolation, the concentration maps allow for an individual location to be assumed to be the exposure to a stationary worker. The estimation of personal exposures from area exposures has been conducted at this facility in previous work by Zuidema *et al.* (2019c) and is a valid method for personal exposure assessment when the exposure data and worker data are at a high enough spatiotemporal resolution.

Mean vs. 95th percentile exposures

Following IDW interpolation so that there are 233 daily exposure hazard maps, all mean exposure estimates across all grid locations and days were binned into quintiles (Quintile$_{\text{Mean}}$), and all 95th percentile exposure estimates across all locations and days were binned into separate quintiles (Quintile$_{95}$). Therefore, each location on each day had a 1–5 value for the mean exposure and a separate 1–5 value for the 95th percentile exposure. The Quintile$_{\text{Mean}}$ and Quintile$_{95}$ values for each location on each day were then compared to each other to determine if a location's Quintile$_{\text{Mean}}$ can properly estimate relative (compared across the facility) upper bound exposure levels.

Compliance sampling comparison

In the absence of large quantities of exposure data, AIHA exposure ratings are generally determined using Equation 2, where the upper confidence limit (UCL$_{95}$) of the 95th percentile of exposure is calculated (instead of the explicit 95th percentile (X_{95})) due to typically small sample sizes of less than ten. In Equation 2, γ is the confidence level, p is the proportion, n is the sample size, $\bar{y}$ is the natural log of the geometric mean of the sampling measurements, and s_y is the natural log of the geometric standard deviation of the sampling measurements, and *KK* is the value from a *K*-value table based on the corresponding parameters.

Table 1. 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	$X_{0.95} \le 0.10 * OEL$
2	Well controlled	Exposures infrequently exceed 50% of the limit and rarely exceed the limit	$0.10 * OEL < X_{0.95} \le 0.5 * OEL$
3	Nominally controlled	Exposures infrequently exceed the limit	$0.50 * OEL < X_{0.95} \le OEL$
4	Poorly controlled	Exposures frequently exceed the limit	$OEL \le X_{0.95}$

Equation 2: American Industrial Hygiene Association UCL_{95}

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

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, or 50 days) against the network-based assessments of the 95th percentile (X_{95}; n = 223 days) was conducted. To simulate a compliance sampling system, 50 randomly chosen locations were selected to represent stationary workers (do not move for entire eight-hour shift) on the factory floor. For each of the 50 locations, 3, 5, 10, or 50 random daily exposure values were selected from the 223 (days with sampling data) possible eight-hour daily time-weighted average (TWA) concentrations to simulate exposure assessments. The eight-hour TWAs were necessary since there are 12 h (6 am–6 pm) of exposure data available. 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 UCL_{95} was calculated based on ACGIH guidance using Equation 2 (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. Each location and sample size pairs were repeated 100 times to assess variability.

Software

All 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 the online Supplementary Material.

Results

Sensor modeling

The weighting parameter values that minimized model error via RMSE and CRPS were C= 0.75 and P= 2.0, resulting in a final dataset size of 5366 sensor/reference measurement pairs. The prediction results on the test sets for the four splits are presented in Table 2.

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 data from all three sampling periods, whereas period 1, period 2, and period 3 were tested on fully out of sample data. The relatively higher error in the period 2 split (this held true for all values of C and P that were investigated) is potentially due to atypical production schedules associated with reference sampling occurring shortly before the Christmas holiday. Additionally, using C = 0.75 and P = 2.0, all GBDT train and test splits (Table 2) 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 18 August 2017 and 23 March 2018 is shown in Fig. 1 with linear calibration, the GBDT calibration, and the reference measurement (Afshar-Mohajer *et al.*, 2018; Zuidema *et al.*, 2019a). The GBDT is both more accurate and captures peaks and valleys in reference measurements better than the linear model.

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 $N_{NGBoost}(x,\ \sigma)$ at each sensor location and time across the full eight-month deployment, which was then interpolated to produce probabilistic hazard maps. These probabilistic hazard maps can be created for any percentile of interest, instead of simply relying on a mean point prediction. The mean and 95th percentile hazard map for the entire dataset spanning 4 August 2017 to 27 March 2018 is shown in Fig. 2, overlaid with sensor locations denoted by "x".

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

Table 2. 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)	Lab calibration RMSE (ppm)
'Full' (80% of all data)	4293	20% of all data	1073	0.36	0.16	2.42
Period 1	682	Period 2 and Period 3	4684	1.91	1.11	
Period 2	2330	Period 1 and Period 3	3036	2.35	1.98	
Period 3	2354	Period 1 and Period 2	3012	1.39	0.73	

up by a constant, likely indicating that concentration variability is not uniform across the facility. However, given that these data are aggregated over 223 days, the 95th percentile for any daily (or smaller unit of time) exposure can be could be higher or lower than this range.

AIHA exposure rating hazard maps

The 95th percentile interpolated values from Fig. 2 can be directly used in the AIHA exposure ratings framework to assign an exposure rating. For the dataset used in this analysis, the CO exposures were significantly below the OSHA PEL of 50 ppm, and an artificial OEL of 8.25 ppm was used for illustrative purposes only (National Institute of Occupational Safety and Health, 2019). Using the full timeseries of network deployment over the dates in Fig. 2, the exposure category hazard map is shown in Fig. 3.

Areas in red are 'Poorly Controlled' meaning that for an illustrative 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.

Using mean concentrations to estimate 95th percentile concentrations

All $\text{Quintile}_{\text{Mean}}$ and Quintile_{95} pairs across each of the 223 days and grid locations were compared so that if a location's $\text{Quintile}_{\text{Mean}}$ and Quintile_{95} were in identical quintiles (both equal to 2 for example), then the mean exposure can be used to approximate relative upper bound exposure at that location (relative to other locations). However, if $\text{Quintile}_{\text{Mean}} < \text{Quintile}_{95}$, then the mean exposure underestimates the relative upper bound exposure, or if $\text{Quintile}_{\text{Mean}} > \text{Quintile}_{95}$, then the mean exposure overestimates the relative upper bound exposure – both are considered exposure misclassifications. The aggregated exposure misclassification counts are shown in Fig. 4 by season (based on the day they occurred), where each bar corresponds to the proportion of days in that season where more than 20% of the factory floor was misclassified.

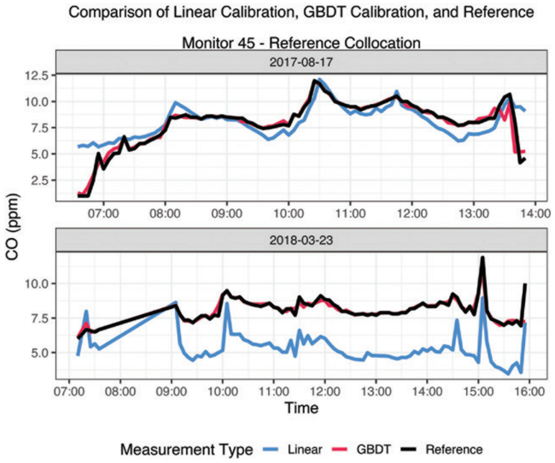

Figure 1. Time series comparison of linear calibration, GBDT calibration, and reference measurements at single reference-monitor collocation on 18 August 2017 and 23 March 2018.

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% of the facility inaccurately assigned. The remaining three seasons were slightly more effective, with between 25 and 10% of the days in those seasons having at least 20% of the facility as exposure misclassifications. The full distribution of floor percentage misclassification by season is provided in the online Supplementary Fig. S4.

Compliance sampling comparison

The results of the comparison between simulated compliance sampling on a small number of days and the sensor network approach in which personal exposures may be estimated over a long duration of time are shown in Table 3. For the majority of the locations (approximating the personal exposure for stationary workers without strong sources in their breathing zone that are unlikely to be captured by the network), the limited number of samples used in traditional compliance sampling significantly overestimated the exposures when estimating the UCL_{95} (which is compared against the X_{95} from the sensor network) to determine the exposure rating. Approximately two-thirds of locations were incorrectly classified into a higher exposure rating (four as opposed to three). Geometric means and standard deviations of the UCL_{95} were calculated for the 100 replicates of each condition (i.e., number of samples and location). Even with 50 sampling days, the geometric mean UCL_{95} is more than 10% higher than the X_{95} from the network data.

Discussion

This study utilized a NGBoost probabilistic gradient boosted decision tree that modeled reference measurements directly from raw sensor voltages and

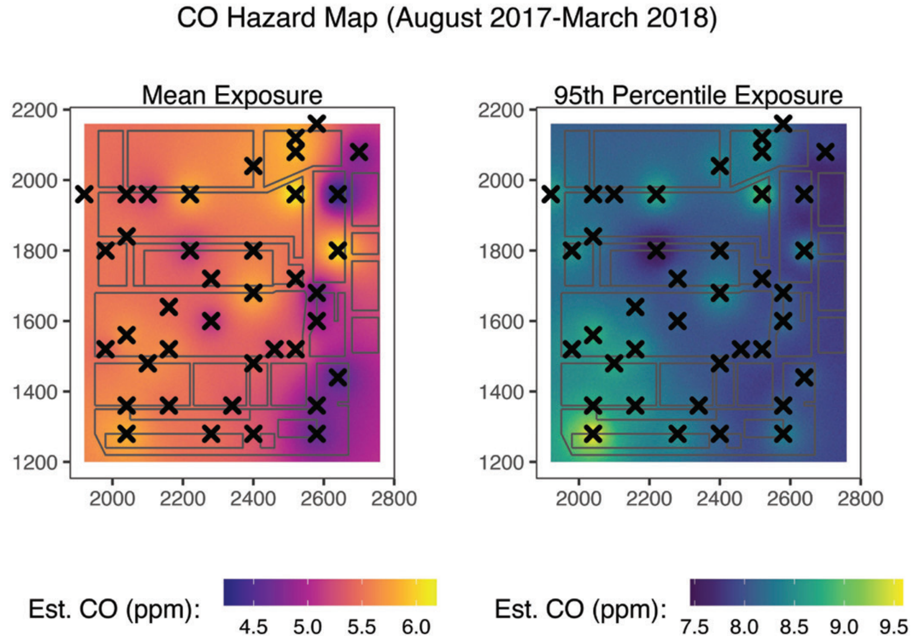

Figure 2. Hazard map of estimated mean and 95th percentile CO concentrations (ppm) for 4 August 2017 to 27 March 2018.

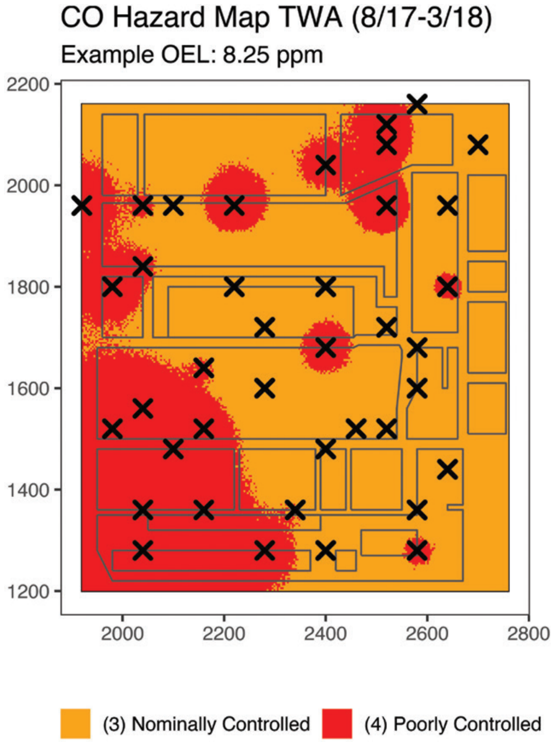

Figure 3. AIHA exposure ratings based on full dataset prediction and interpolation for 4 August 2017 to 27 March 2018.

environmental condition and location data in a low-cost occupational sensor network. Using NGBoost allowed for a unique mean and standard deviation to be modeled for each prediction, unlike a traditional linear regression approach which provides a mean and constant standard deviation. In addition, the NGBoost was between two and five times more accurate in terms of RMSE than an existing laboratory-based calibration model. As low-cost sensors are individually less accurate and precise than more expensive instruments, some form of calibration is needed for the success of the network (Datta *et al.*, 2020; Thomas *et al.*, 2018; Zamora *et al.*, 2019; Zuidema et al., 2019a, 2019b). Therefore, the success of the modeling approach in this paper highlights the potential for reductions or eliminations of the highly time and equipment intensive process of laboratory calibration if appropriate field reference data can be collected (e.g. it may be worthwhile to do span or zero checks for sensors to identify malfunctioning sensors before installation without necessitating a full calibration).

The results from the model and sensitivity tests (see online Supplementary Material) indicated that CO reference concentrations can be predicted with high accuracy when training on reference data that covers a range of environmental conditions (August, December, and March), but training in a single season reduces the accuracy substantially. However, the accuracy of the model

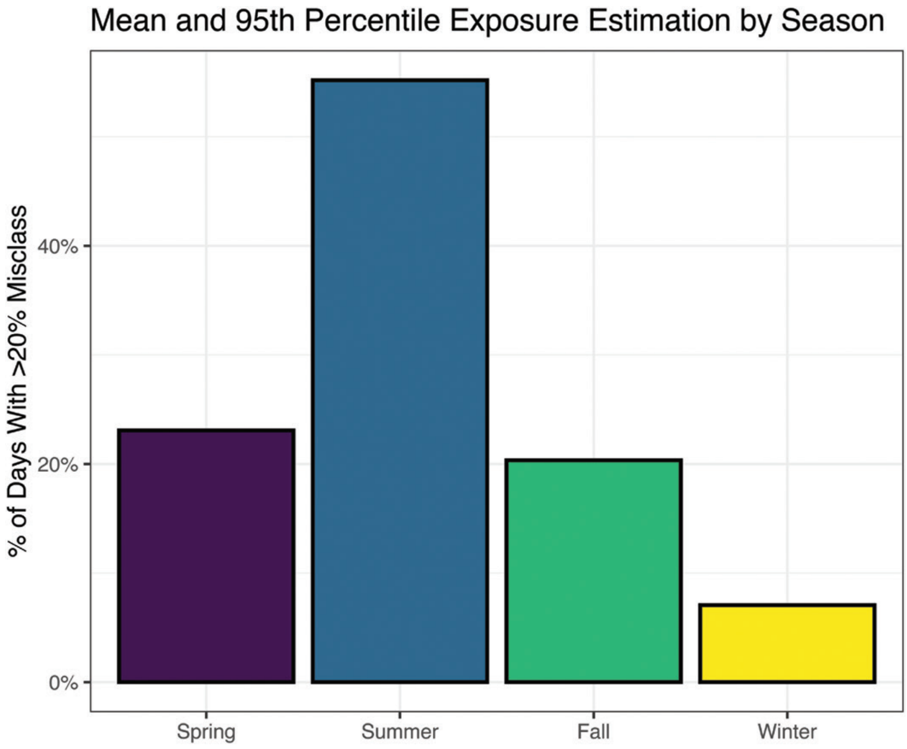

Figure 4. Percent of days by season (4 August 2017 to 27 March 2018) where greater than 20% of the factory floor has exposure misclassifications by using the mean exposure (QuintileMean) as a proxy for relative upper bound exposures (Quintile95).

Table 3. 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 AIHA exposure rating
Sampling days	Geo. mean UCL_{95} (ppm)	Geo. SD UCL_{95}	Geo. mean X_{95} (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

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 (Fig. 2) or in terms of threshold exceedance probabilities (e.g., the likelihood that a given location exceeded an OEL). 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 ratings, which explicitly rely on the 95th percentile personal exposures, as opposed to mean or median exposures (Ramachandran, 2005). Maps of exposure ratings could be used by an industrial hygienist to determine where to prioritize controls or additional high-quality sampling (badges, canisters, etc.). Furthermore, as seen in Table 3, if network-based monitoring can be validated against traditional personal exposures, this approach could provide better estimation of the exposure category, as well as capturing temporal variability in exposures for many workers.

An additional point of emphasis for using the 95th percentile is that using mean exposures alone can potentially result in exposure misclassifications of upper bound exposures. For example, a high predicted mean concentration in a location does not necessarily imply a high predicted upper bound concentration in that location. For example, as shown in Fig. 4, on approximately 45% of the days the network was active in the Summer, greater than 20% of the factory floor would have had incorrect exposure assignments using the mean exposure as opposed to the 95th percentile exposures. When small sample sizes are collected, estimates of the mean can be more reliable than estimates of the UCL_{95}. However, relying on the mean as a proxy for the upper bound exposures may lead to misclassifications over large portions of the facility. 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 study 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 colocation data would allow for additional testing as well, measuring error by location of reference instrument or time of day for example. Future improvements on the model could be made in by incorporating recent measurements explicitly, where $time_0$ is predicted by $time_{-1}$, $time_{-2}$, etc. Additionally, sensitivity tests to determine of the exact number of reference instruments neededsd to achieve acceptable model accuracy would be valuable future work. Furthermore, the AIHA framework relies on the use of personal exposure assessments. Future work should include the incorporation of the personal exposure methodology from Zuidema *et al.* (2019c), where an individual's personal exposure was estimated via their location over time throughout the course of a day. This methodology could be used to remove the artificial constraint on a 'stationary worker'. Additionally, exposures that are specifically worker-generated in extreme proximity to the worker's breathing zone could potentially be underestimated by a networked approach, although Zuidema *et al.* (2019c) discuss this possibility as well. Lastly, the models demonstrated here could be implemented with a live or streaming data pipeline that could utilize network data to serve as a real-time exposure assessment tool or early warning system across an entire facility.

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 AIHA exposure ratings guidelines. Further, 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, while substantially more complex than traditional calibration methods, 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. Additionally, the model used in this study demonstrated how utilizing mean or median exposures determined from small sample sizes are not sufficient for high-quality occupational exposure assessments as exposure misclassification is likely.

Supplementary Data

Supplementary data are available at *Annals of Work Exposures and Health* online.

Acknowledgments

Funding for this project was provided by U.S. Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health (NIOSH) grant R01OH010533. A.N.P. was supported by a grant from NIOSH to the Johns Hopkins Education and Research Center for Occupational Safety and Health (award number T42 OH0008428). The authors declare no conflict of interest relating to the material presented in this manuscript.

Data Availability

No new data were generated or analyzed in support of this research.

References

Afshar-Mohajer N, Zuidema C, Sousan S, *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.

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 Health*; **62**: 547–58.

Borrego C, Ginja J, Coutinho M, *et al*. (2018) Assessment of air quality microsensors versus reference methods: the EuNetAir Joint Exercise—part II. *Atmos Environt*; **193**:127–42.

Buehler C, Xiong F, Zamora ML, *et al*. (2020) Stationary and portable multipollutant monitors for high spatiotemporal resolution air quality studies including online calibration. *Atm Measur Tech*; **14**:995–1013.

Casey JG, Hannigan MP. (2018) Testing the performance of field calibration techniques for low-cost gas sensors in new deployment locations: across a county line and across Colorado. *Atm Measur Tech*; **11**:6351–78.

Datta A, Saha A, Zamora ML, *et al*. (2020) Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. *Atm Environ*; **242**:117761.

Duan T, Avati A, Ding DY, *et al*. (2019) *NGBoost: Natural Gradient Boosting for Probabilistic Prediction*. arXiv:1910.03225v4 [cs.LG].

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.

Gneiting T, Raftery AE. (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*; **102**:359–78.

Heimann I, Bright VB, McLeod MW, *et al*. (2015) Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. *Atm Environ*; **113**:10–9.

Hewett P, Logan P, Mulhausen J *et al*. (2006) Rating exposure control using Bayesian decision analysis. *J Occup Environ Hyg*; **3**: 568–81.

Koehler KA, Peters TM. (2013) Influence of analysis methods on interpretation of hazard maps. *Ann Occup Hyg*; **57**: 558–70.

Koehler K, Volckens J. . (2011) Prospects and pitfalls of occupational hazard mapping: "between these lines there be dragons". *Ann Occup Hyg*; **55**:829–40.

Kuhn M, Johnson K. (2013) *Applied predictive modeling*. New York, NY: Springer.

Lim CC, Kim H, Vilcassim MJR *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.

Malings C, Tanzer R, Hauryliuk A, *et al*. (2019) Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atm Measur Tech*; **12**:903–20.

Morawska L, Thai PK, Liu X *et al*. (2018) Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environ Int*; **116**: 286–99.

National Institute of Occupational Safety and Health. (2019) *Carbon Monoxide*. NIOSH Pocket Guide to Chemical Hazards. Washington DC: NIOSH.

Pedregosa F, Varoquaux G, Gramfort A, *et al*. (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res*; **12**:2825–30.

Ramachandran G. (2005) *Occupational Exposure assessment for air contaminants*. 1st ed. Boca Raton FL: CRC Press.

Rappaport SM. (1984) The rules of the game: an analysis of OSHA's enforcement strategy. *Am J Ind Med*; **6**:291–303.

R Core Team. (2020) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schapire RE. (2003) *The boosting approach to machine learning: an overview*. New York, NY: Springer. p. 149–71.

Thomas G, Sousan S, Tatum M, *et al*. (2018) Low-cost, distributed environmental monitors for factory worker health. *Sensors*; **18**:1411.

Tornero-Velez R, Symanski E, Kromhout H, *et al*. (1997) Compliance versus risk in assessing occupational exposures. *Risk Anal*; **17**:279–92.

Tuggle RM. (1981) The NIOSH decision scheme. *Am Ind Hyg Assoc J*; **42**:493–8.

Zamora ML, Xiong F, Gentner D, *et al*. (2019) Field and laboratory evaluations of the low-cost plantower particulate matter sensor. *Env Sci Technol*; **53**:838–49.

Zimmerman N, Presto AA, Kumar SPN, *et al*. (2018) A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atm Measur Tech*; **11**: 291–313.

Zuidema C, Sousan S, Stebounova LV *et al*. (2019a) Mapping occupational hazards with a multi-sensor network in a heavy-vehicle manufacturing facility. *Ann Work Expo Health*; **63**: 280–93.

Zuidema C, Stebounova LV, Sousan S *et al*. (2019b) Sources of error and variability in particulate matter sensor network measurements. *J Occup Environ Hyg*; **16**: 564–74.

Zuidema C, Stebounova LV, Sousan S, Gray, A, Stroh, O, Thomas G, Peters TM, Koehler K. (2019c) Estimating personal exposures from a multi-hazard sensor network. *J Exp Sci Environ Epidemiol*; **30**: 1013–1022.