Posted on 1 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/essoar.168565375.52420155/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Uncertainty reduction and environmental justice in air pollution epidemiology: the importance of minority representation

Mariana Alifa[1], Stefano Castruccio[1], Diogo Bolster[1], Mercedes A Bravo[2], and Paola Crippa[1]

[1]University of Notre Dame
[2]Duke University

June 1, 2023

Abstract

Ambient air pollution is an increasing threat to society, with rising numbers of adverse outcomes and exposure inequalities across the globe. Reducing uncertainty in health outcomes models and exposure disparity studies is therefore essential to develop policies effective in protecting the most affected places and populations. This study uses the concept of information entropy to study tradeoffs in mortality uncertainty reduction from increasing input data of air pollution versus health outcomes. We study a case scenario for short-term mortality from fine particulate matter (PM2.5) in North Carolina for 2001-2016, employing a case-crossover design with inputs from an individual-level mortality dataset and high-resolution gridded datasets of PM2.5 and weather covariates. We find a significant association between mortality and PM2.5, and the information tradeoffs indicate that in this case increasing information from mortality may reduce model uncertainty at a faster rate than increasing information from air pollution. We also find that Non-Hispanic Black (NHB) residents tend to live in relatively more polluted census tracts, and that the mean PM2.5 for NHB cases in the mortality model is significantly higher than that of Non-Hispanic White (NHW) cases. The distinct distribution of PM2.5 for NHB cases results in a relatively higher information value, and therefore faster uncertainty reduction, for new NHB cases introduced into the mortality model. This newfound influence of exposure disparities in the rate of uncertainty reduction highlights the importance of minority representation in environmental research as a quantitative advantage to produce more confident estimates of the true effects of environmental pollution.

Hosted file

`963027_0_art_file_11004548_rpvxvy.docx` available at https://authorea.com/users/620448/articles/644451-uncertainty-reduction-and-environmental-justice-in-air-pollution-epidemiology-the-importance-of-minority-representation

Hosted file

`963027_0_supp_11004549_rgvxvy.docx` available at https://authorea.com/users/620448/articles/644451-uncertainty-reduction-and-environmental-justice-in-air-pollution-epidemiology-the-importance-of-minority-representation

Uncertainty reduction and environmental justice in air pollution epidemiology: the importance of minority representation

Mariana Alifa[1*], Stefano Castruccio[2], Diogo Bolster[1], Mercedes A. Bravo[3,4], Paola Crippa[1]

[1]Department of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, Notre Dame, IN, USA

[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA

[3]Global Health Institute, Duke University, Durham, NC, USA

[4]Children's Environmental Health Initiative, University of Notre Dame, South Bend, IN, USA

*Corresponding author: Mariana Alifa, Department of Civil and Environmental Engineering and Earth Sciences, 156 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556, USA, email: malifa@nd.edu

Key points

- We used information entropy to study efficient pathways for uncertainty reduction in an air pollution-mortality model for $PM_{2.5}$.
- We compared the uncertainty reduction effect of adding new data for Non-Hispanic Black (NHB) versus Non-Hispanic White (NHW) cases
- Introducing new NHB cases results in faster uncertainty reduction because of the differential $PM_{2.5}$ exposure in the NHB population.

Keywords

Air Pollution; Exposure Disparities; Information Entropy; Uncertainty Reduction; Environmental Justice; Risk Assessment

Abstract

Ambient air pollution is an increasing threat to society, with rising numbers of adverse outcomes and exposure inequalities across the globe. Reducing uncertainty in health outcomes models and exposure disparity studies is therefore essential to develop policies effective in protecting the most affected places and populations. This study uses the concept of information entropy to study tradeoffs in mortality uncertainty reduction from increasing input data of air pollution versus health outcomes. We study a case scenario for short-term mortality from fine particulate matter ($PM_{2.5}$) in North Carolina for 2001-2016, employing a case-crossover design with inputs from an individual-level mortality dataset and high-resolution gridded datasets of $PM_{2.5}$ and weather covariates. We find a significant association between mortality and $PM_{2.5}$, and the information tradeoffs indicate that in this case increasing information from mortality may reduce model uncertainty at a faster rate than increasing information from air pollution. We also find that Non-Hispanic Black (NHB) residents tend to live in relatively more polluted census tracts, and that the mean $PM_{2.5}$ for NHB cases in the mortality model is significantly higher than that of Non-Hispanic White (NHW) cases. The distinct distribution of $PM_{2.5}$ for NHB cases results in a relatively higher information value, and therefore faster uncertainty reduction, for new NHB cases introduced into the mortality model. This newfound influence of exposure disparities in the rate of uncertainty reduction highlights the importance of minority representation in environmental research as a quantitative advantage to produce more confident estimates of the true effects of environmental pollution.

1. Introduction

Air pollution is an increasing threat to today's society. Data from the Global Burden of Disease study ranked ambient pollution from $PM_{2.5}$ as the 5th leading global mortality risk factor in 2015, causing 4.2 million deaths and 103.1 million disability-adjusted life years due to health impacts such as lung cancer, lower respiratory infection, chronic obstructive pulmonary disease, cerebrovascular disease, and ischemic heart disease (Cohen et al., 2017). A recent update for this study (Fuller et al., 2022) reports a rise in ambient pollution attributable deaths to 4.5 million in 2019, a 7% increase since 2015 and a 66% increase since 2000, revealing that, despite increased awareness and attemtps at remediation of this problem, our efforts have so far been insufficient in protecting society from the harms of ambient pollution.

The United States stands out as a successful case of continued efforts to curb air pollutant emissions. The Clean Air Act required in 1970 that the Environmental Protection Agency (EPA) set National Ambient Air Quality Standard (NAAQS) for "criteria pollutants" and establish a network of ambient pollution monitoring stations to assess compliance to these standards. The first NAAQS specifically for $PM_{2.5}$ was issued in 1997 (once monitors were advanced enough to measure particles of this size), setting the standard for annual mean concentration at 15 $\mu g/m^3$ (EPA, 1997). However, subsequent findings of harmful health effects at air pollution concentrations that blend into background levels have prompted the continual lowering of NAAQS (McClellan, 2002). The standard for $PM_{2.5}$ was lowered to 12 $\mu g/m^3$ in 2012 (EPA, 2013), and a proposal issued in January of 2023 is now currently underway to further lower the NAAQS to 9-10 $\mu g/m^3$ (EPA, 2023). Although these nationwide measures have been effective in reducing overall levels of air pollution, they have not been as successful in curbing demographic and socioeconomic inequalities in relative exposure (Colmer et al., 2020; Liu et al., 2021).

Extensive research has found demographic and/or socioeconomic disparities in exposure to $PM_{2.5}$ and other air pollutants across different regions of the world (Hajat et al., 2015). In the United States, multiple studies have found that people of color have been systematically exposed to higher levels of air pollution (Colmer et al., 2020; Liu et al., 2021; Tessum et al., 2021). These racial disparities are not only found across different income levels, urbanicity levels, and emission types (Liu et al., 2021; Tessum et al., 2021), but they have also persisted despite the nationwide decreasing trend in air pollution seen in the last four decades, with studies identifying that the relatively most polluted census tracts in present day are largely the same census tracts that were most polluted in the 80s and the 90s (Colmer et al., 2020; Liu et al., 2021).

In light of this lack of progress in addressing both air pollution-related health outcomes at the global level and pollution exposure disparities at the national level, it is essential to develop policies that will effectively target the places and populations most affected by ambient air pollution. However, one of the multiple challenges to effective policy is the uncertainty affecting ambient pollution health impact assessments (HIAs) used to guide AQ standards from local and national (EPA, 2019; EU, 2008) to global (WHO, 2006) levels. These studies integrate multiple sources of information such as, among others, air pollution concentrations and related population exposure, physiological responses to pollution exposure, and their variation by individual-level factors (such as gender, age, body mass, race, etc.) as well as residential factors (such as proximity to water bodies or green spaces). Each of these sources of information involved in the air pollution HIA may introduce several different kinds of uncertainty into the final assessment model (Nethery & Dominici, 2019).

Among the many possible sources of uncertainty in HIAs, this study focuses on uncertainty stemming from incomplete knowledge of the pollution and/or health impact scenarios, caused by data scarcity in the input information. When there is a recognized scarcity in observational data precluding the full characterization of the pollution-exposure-effects scenario, action can be taken to augment the available input datasets to increase our knowledge of the problem and gain confidence in the results of the final assessment. Solutions to the problem of data scarcity have been indeed addressed extensively in both the air pollution and the epidemiology fields.

Air pollution research has proposed different approaches to data assimilation for better risk characterization, mainly by supplementing ground observations from official monitoring stations (for example, those from the United States' Environmental Protection Agency, EPA) with other sources of data, such as citizen-science observations (Bonas & Castruccio, 2021; Shen et al., 2021), satellite observations of atmospheric and aerosol properties (Van Donkelaar et al., 2021; Van Donkelaar et al., 2015; Zani et al., 2020), chemical transport models, or CTMs (Giani, Anav, et al., 2020; Giani, Castruccio, et al., 2020), and/or dispersion models (Bates et al., 2018). In cases where ground-based pollution data is sparse, CTMs able to reproduce monitored pollutant concentrations have also been used to make robust assessments of the region's pollution risks (Mead et al., 2018). Therefore, several studies have focused on localized downscaling of existing CTMs to achieve finer resolution in areas of interest (Tessum et al., 2017) or in the implementation of higher-resolution CTMs for a more accurate representation of meteorological, chemical and aerosol properties (Crippa et al., 2019).

Previous work has also focused on assessing epidemiological uncertainty. For example, meta-analyses of epidemiological studies combine multiple previous studies' results for robustness (Atkinson et al., 2014; Pope et al., 2020). Another approach (Burnett et al., 2014) developed an integrated exposure-response model by combining epidemiological data from multiple $PM_{2.5}$ sources, such as ambient air pollution, active and second hand tobacco smoke, and household solid cooking fuel. A recent study (Coffman et al., 2020) derived distributions from existing epidemiological data to model uncertainty in the exposure-response curve at low levels of $PM_{2.5}$, for which data is usually sparse. Other studies have performed disaggregation of exposure data with the goal of improving health effect estimation in future epidemiological studies (Beckx et al., 2009; Breen et al., 2020).

Data scarcity in air pollution epidemiology studies also has environmental justice implications. Studies of air pollution epidemiology have been traditionally based on ambient air pollution monitoring data from the US Environmental Protection Agency (EPA), resulting in an urban bias in the assessment (Bell et al., 2004; Dominici et al., 2006) since the EPA prioritizes monitor placements in population-dense areas (Bravo et al., 2012; Miranda et al., 2011). Even within relatively-urbanized counties, minority populations have been found to live closer to sources of air pollution but further away from monitoring stations (Stuart et al., 2009). Recent research has therefore leveraged the use of satellite data, land use regression, and air quality models to expand and diversify the spatial area and thus, population, for which $PM_{2.5}$ exposures and health effects can be estimated (Ha et al., 2014; Hyder et al., 2014; Kloog et al., 2012; Qian et al., 2019).

Although the problem of data scarcity has been extensively studied as it relates to air pollution, epidemiology, and environmental justice, there remains a need for more interdisciplinary research linking the findings from all these fields under a single framework. We began addressing this need in a previous study (Alifa et al., 2022) where we adapted a methodology proposed in the hydrology field (De Barros & Rubin, 2008; De Barros et al., 2009) to create a novel framework that identifies the most efficient pathway to reduce uncertainty in estimates of air pollution-associated health risks. The studies in hydrology (De Barros & Rubin, 2008; De Barros et al., 2009) had explored the concept of uncertainty tradeoffs in the modeling of the health effects of groundwater contaminants combining the concept of information entropy with Bayesian inference methods; Our subsequent study (Alifa et al., 2022) adapted this framework for frequentist inference to study the effect of data increase on the reduction of air pollution mortality uncertainty, measured through the metric of information entropy, and visualize the tradeoffs in the resulting uncertainty of the mortality model depending on the kind of input data gained. The two cases presented in that study (Alifa et al., 2022), one with artificial data for $PM_{2.5}$ and mortality data used in a long-term exposure model, and one with real time-series data used in a short-term exposure model, demonstrated the applicability of the method for aiding stakeholders in choosing the most efficient pathway for HIA uncertainty reduction when limited resources (e.g. time, money, computational power) prevent them from investing in improvements for both pollution and health outcomes data.

We now seek to explore this framework further by applying it to a more complex case scenario involving spatio-temporal data. We use a case-crossover model design (Jaakkola, 2003) to investigate the association of short-term $PM_{2.5}$ exposure with mortality in North Carolina for the years 2001-2016, through the use of individual-level mortality data and high-resolution gridded datasets of $PM_{2.5}$ and weather covariates. This study aims to not only illustrate the usefulness of our information entropy tradeoff methodology to generate more robust impact assessments, but also to gain new knowledge of the influence of socio-demographic inequalities in the dynamics of uncertainty reduction.

The rest of the study is structured as follows: section 2 describes the datasets and methods used to study exposure disparities, pollution-mortality associations, and uncertainty tradeoffs from changes in input information. Section 3 presents the study results, and section 4 concludes with a discussion of the results' implications and dialogue with recent literature.

2. Methods

2.1 Data

Mortality data

We use individual-level mortality data for North Carolina from 2001 to 2016. The data was obtained from the North Carolina State Center for Health Statistics, Vital statistics department. Our analysis utilizes each participant's date of death, residential location, and race/ethnicity. We studied total mortality (all causes of death except external causes, International Classification of Diseases, ICD10, A00-R99). Other individual characteristics not analyzed in this work are also included in the mortality dataset, such as sex, age at death, education, and marital status. Additional analysis of the correlation of air pollution mortality with these individual-level variables, as well as that of residential and environmental variables, has been performed elsewhere (Son et al., 2020).

Air pollution data

We use daily gridded data from a 1km model of $PM_{2.5}$ concentration (Di et al., 2021). This ensemble-based model utilizes machine learning algorithms and multiple variables from monitoring stations from the Environmental Protection Agency (EPA), satellite measurements, land use terms, chemical transport model output, and others, to predict daily $PM_{2.5}$ for the entire United States. More details about model development and evaluation are available elsewhere (Di et al., 2019). The exposure assigned to each participant is based on the 1km gridcell that contains their residential location.

Weather data

We include daily gridded data on mean temperature and dewpoint temperature as covariates in our mortality modeling. Inclusion of these covariates is common practice in air pollution-epidemiology studies (e.g., (Nhung et al., 2017; Son et al., 2020)) to control for weather-related mortality. These data are obtained on a 4×4km grid from the Parameter-elevation Regressions on Independent Slopes Model (PRISM), which combines ground-based measurement station data with a digital elevation model to create gridded climate products for the U.S. Additional details are available elsewhere (Daly et al., 2008; PRISM Climate Group, 2004). Similarly to the air pollution data, each participant is assigned the weather data of the grid cell containing their residence.

Census data

We utilize US census data on race for the analysis of disparities in air pollution exposure. We chose the data for 2010 since this census year falls around the middle of the range of our analysis (2001-2016). A comparison with 2020 census data determined that although North Carolina's population is increasing, the changes in racial composition and spatial distribution of the population are small enough for the results of our study to not be affected by the choice of census year.

2.2 Exposure disparities

The 2010 US census reports 21.2% of the population of North Carolina was NHB, making them the largest racial minority in the state. Therefore, we focus our study of $PM_{2.5}$ exposure disparities on the NHB population.

We derive the average $PM_{2.5}$ concentration between 2001 and 2016 for each census tract in the state and compare these to the tract's %NHB using quantile regression (Koenker & Bassett

Jr, 1978; Koenker & Hallock, 2001). Quantile regression estimates the conditional quantile(s) of interest of the response variable (in this case, $PM_{2.5}$) as a linear combination of the predictor variable (in this case, %NHB). We model the 10th, 25th, 50th, 75th, and 90th percentile $PM_{2.5}$ using data from the 1405 census tracts in the state with NHB residents. Ordinary linear regression, in contrast, estimates the conditional mean of the response variable, only giving information about the relationship between air pollution levels and the percentage of NHB residents for the "average" census tract. Using quantile regression provides more comprehensive results, allowing us to study this relationship for the more and least polluted census tracts, as well as the median census tracts, thus exploring racial inequalities in exposure at different relative exposure levels.

In addition to state-wide results, we also investigate exposure disparities for the two most populated counties in the state: Mecklenburg County (population 923,427 in the 2010 census, 50.5% Non-Hispanic White (NHW) and 30.2% NHB) and Wake County (population 906,969 in the 2010 census, 62.2% NHW and 20.4% NHB). We report quantile regression results for each county, and we also compare the density function of the %NHB population in the least polluted census tracts in each county, determined as those with average $PM_{2.5}$ in the 1st quartile, to density function of %NHB in the most polluted census tracts (those with average $PM_{2.5}$ in the 4th quartile). This comparison of density functions provides an assessment of the differences in the racial distribution of the population between the most polluted and least polluted census tracts in the county.

2.3 Mortality modeling

We model the association between $PM_{2.5}$ and short-term mortality with a case-crossover design. This model uses each individual as their own control, eliminating the need to control for individual-level characteristics and thus greatly reducing the number of necessary covariates for good model specification. This low number of covariates presents an advantage for our goal of isolating the influence of increasing input data for a specific variable (in this study, either for $PM_{2.5}$ or mortality) on the uncertainty reduction of the epidemiology model. For a different type of model requiring more individual-level controls, the epistemic uncertainty introduced by a high number of covariates could obscure the uncertainty reduction achieved by any single variable's information gain. We select control days based on the same day of the week of the same month of the individual's death. Each case day therefore has more than one control, and we allow for bi-directional sampling of controls (selection of control days both before and after the individual's death) to control for bias from temporal trends in the pollution data (Navidi, 1998). Temperature and dewpoint temperature are also incorporated as covariates in the model.

The choice to investigate the pollution-mortality association in the short-term is motivated by the type of health data available for this study. We use a dataset where cases have been selected based on health outcome (in this case, mortality), making the data suitable for a short-term study using a case-control design and further, for a case-crossover design since we do not have data on other individuals who did not experience the outcome of interest (Belbasis & Bellou, 2018; Jaakkola, 2003). Since air pollution has been widely recognized to have both short-term and long-term effects, the same information tradeoffs methodology presented here could be applied to a different epidemiology model in the presence of health data suitable for a long-term study. For example, a long-term study could be performed using a cohort design, where participants are selected based on their degree of exposure to air pollution and placed into the "exposed" or "unexposed" group, and then health outcomes for these groups are observed and compared over a specified period of time (Belbasis & Bellou, 2018).

The coefficients of the case-crossover model are fit using conditional logistic regression (Pampel, 2020). If we describe mortality Y_i as following a Bernoulli distribution (equation (1a)), where Y_i can be equal to 1 for the day of death or 0 for the control day(s), and the probability that $Y_i = 1$ is P, then we can model the logged-odds of P as a linear relationship between our predictors of interest (equation(1b)):

$$Y_i \sim \text{Bernoulli (P)}; \tag{1a}$$

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta PM_{2.5} + \gamma T_t + \delta D_t, \tag{1b}$$

where α is the intercept and β is the fitted coefficient describing the association of $PM_{2.5}$ with mortality, also called exposure coefficient. We will focus on β for the study of uncertainty reduction in the case-crossover model (additional details are provided in section 2.4). The coefficients γ and δ describe the association of temperature (T) and dewpoint temperature (D), respectively. Solving for the odds by exponentiating equation (1b) gives us the expression:

$$\frac{P}{1-P} = e^{\alpha} \times e^{\beta PM_{2.5}} \times e^{\gamma T} \times e^{\delta D}, \tag{1}$$

where each exponent term can be interpreted as the odds ratio (OR) for the association of each covariate with mortality. Our main interest lies in the second exponent on the right-hand side, $e^{\beta PM_{2.5}}$. This term represents the OR for a $PM_{2.5}$ increment of 1 μg/m^3, which we will refer to as OR_1. For consistency with common practice in reporting of epidemiology results, we will report the OR for a $PM_{2.5}$ increment of 10 μg/m^3 (OR_{10}) which can be derived from OR_1 as:

$$OR_{10} = e^{\beta \times 10} = (e^{\beta})^{10} = (OR_1)^{10}. \tag{2}$$

We initially examine the association of mortality with $PM_{2.5}$ at multiple lags: lag0, lag1, and lag 2 (meaning the $PM_{2.5}$ on the day of death, 1 day before death, and 2 days before death, respectively). We also analyze two cumulative lags: lag01 (the cumulative effect of lags 0 and 1) and lag02 (cumulative effect of lags 0, 1, and 2), by fitting mortality against the average of the

$PM_{2.5}$ levels at the lags of interest. Then we perform stratified analysis to investigate differences in effects between the NHB and NHW populations at the aforementioned $PM_{2.5}$ lags. Since this stratified analysis performs multiple tests on subsets of the same dataset, we adjust its results for multiplicity by using the Bonferroni correction (Chen et al., 2017; Hochberg & Tamhane, 1987). Based on the results of the full model and the stratified analysis, we will select a single lag of $PM_{2.5}$ for further investigation of uncertainty tradeoffs. The temperature and dewpoint temperature covariates have the same lag as the $PM_{2.5}$ in each model fit.

2.4 Information change and uncertainty tradeoffs

This study adopts the uncertainty tradeoffs methodology developed in (Alifa et al., 2022) for the study of a realistic case scenario through the use of spatio-temporal data on pollution, mortality, and demographics. We will study how fitting the case-crossover model described in 2.3 with changing input information on mortality and air pollution (Y_i and $PM_{2.5}$ in equation (1)), respectively) affects the uncertainty of the pollution-mortality coefficient, β, in the model fit. We will also take advantage of the demographic information included in the mortality dataset to investigate racial differences in uncertainty reduction from improved health data.

<u>Uncertainty quantification of the mortality model</u>

We use the metric of information entropy to characterize the uncertainty of our estimate for the exposure coefficient, $\hat{\beta}$. Since we can assume $\hat{\beta}$ is a continuous random variable, its entropy can be defined as (Christakos, 2012):

$$\mathrm{H}(\hat{\beta}) = -\int_{-\infty}^{\infty} \mathrm{f}(\hat{\beta}) \ln(\mathrm{f}(\hat{\beta}))\, \mathrm{d}\hat{\beta}, \qquad (3)$$

where $\mathrm{f}(\hat{\beta})$ is the probability density function (PDF) of the estimate. As more input information is acquired for the model in equation (1)), the inference becomes more accurate such that $\hat{\beta} \rightarrow \beta$

in probability, which results in a reduction of $H(\hat{\beta})$. Our previous publication (Alifa et al., 2022) demonstrated several methods for deriving entropy both parametrically and non-parametrically. For this study, we derive $H(\hat{\beta})$ parametrically from the standard error of the exposure coefficient, $\hat{\sigma}^2_{\hat{\beta}}$, output from the conditional logistic regression fit. Assuming $\hat{\beta}$ to be asymptotically normal, we use the closed form equation for the entropy of a normal distribution,

$$H(\hat{\beta}) = \frac{1}{2}\log(2\pi e\hat{\sigma}^2_{\hat{\beta}}). \quad (4)$$

Additionally, the relative entropy $\Delta\mathbf{H}_{\hat{\beta}}$ is a useful metric to compare the uncertainty of different information stages. We can define the vector $\Delta\mathbf{H}_{\hat{\beta}}$ as:

$$\Delta\mathbf{H}_{\hat{\beta}} = \mathbf{H}_{\hat{\beta}} - H_{\hat{\beta},\mathrm{ref}}, \quad (5)$$

where $\mathbf{H}_{\hat{\beta}}$ is a vector containing $H(\hat{\beta})$ for different stages of information, and $H_{\beta,\mathrm{ref}}$ is the entropy for the information stage selected as reference. For this study we order the elements of $\mathbf{H}_{\hat{\beta}}$ from those computed with least to most information, and select the stage with most information as our reference, resulting in a $\Delta\mathbf{H}_{\hat{\beta}}$ that decreases towards 0.

Change in air pollution information

We generate different stages of air pollution information by upscaling the original 1km $PM_{2.5}$ model to two coarser resolutions, 6km and 12km. We then fit the model in equation (1) with the three different resolutions and compare $H(\hat{\beta})$ for the three cases. These different stages of information simulate a situation where stakeholders are currently operating with coarse-resolution output such as that from the EPA's Community Multiscale Air Quality Model (CMAQ, 12km resolution) or other similar gridded products, and want to explore the information benefits of downscaling their data to higher resolutions.

Change in mortality information

To change the amount of input mortality information, we fit equation (1) with varying number of mortality records. This simulates a case where stakeholders are interested in investigating the benefit of augmenting the health outcomes dataset used for their assessment, due to known or suspected missing cases in said dataset. We will investigate the effect of racial bias in the missing data by comparing the uncertainty reduction when cases are missing only from the NHW population versus cases missing only from the NHB population. We choose these two subpopulations for comparison since in the 2010 US census the racial majority in North Carolina was NHW with 65.2% of the population, while the largest racial minority was NHB, conforming 21.2% of the population. Since NHB cases represented about 20% of the study population, this is the maximum number of missing cases we explore for both races. Therefore, we initially fit the model with ~80% of the total mortality data, where the ~20% of missing cases are either all NHW or NHB patients. Then we increase the number of patients and repeat the fit again with ~90% of data, and lastly with 100% data coverage. We select missing cases at random from the pool of participants of the race of interest, and repeat each model fit 100 times to obtain ensemble results from which we compute the mean and 95% CI of $H(\hat{\beta})$ at each information stage.

Information yield curves

Information yield curves (Alifa et al., 2022; De Barros & Rubin, 2008; De Barros et al., 2009) are a graphical device designed to display the tradeoffs in uncertainty reduction between information gain in air pollution and health data. This tool plots together, in mirror image, the separate effects of information increase for each of these datasets on the uncertainty reduction of $\hat{\beta}$, enabling decision-makers to visualize the most efficient pathway to improve their assessment in their particular case scenario. In our previous study (Alifa et al., 2022) the changes in input data were first associated with changes in uncertainty for separate pollution and health models which when brought together would propagate to the final mortality uncertainty. Therefore, the information yield curve compared the changes in entropy for the separate pollution and health models (in the x axis) to the final change in entropy of the pollution-mortality assessment (in the y axis). The nature of the datasets in this current study requires a modification of the previous method by associating the changes in information for the input datasets directly with the changes in the final uncertainty of the case-crossover model fit. This results in an x-axis of qualitative nature, since there is no common unit to compare increased number of mortality records to increased resolution of the $PM_{2.5}$ grid. However, decision-makers taking advantage of this method in the future would be able to find a common metric for information increase from each dataset given their particular case scenario, such as cost of added data or time for data computation/procurement.

3. Results

3.1 Descriptive statistics

The mortality model had input of a total of 1,065,699 cases with 3,621,521 controls (3.40 controls per case). These cases contained more females than males (52.1% vs 47.9%), and the majority of deaths were from people older than 65 years old (75.4%). Most cases were Non-Hispanic White (77.4%), while the second most cases were Non-Hispanic Black (20.4%). Table S1 shows the full demographics of the mortality data used in the model.

The median of the $PM_{2.5}$ in the model was 9.5 μg/m^3, with lower bound (5th percentile) of 3.8 μg/m^3 and upper bound (95th percentile) of 21.5 μg/m^3. These quantiles varied by less than 0.1 μg/m^3 when recomputed separately for case days and control days. The median temperature was 15.7°C, with 5th and 95th percentiles of 0.7°C and 27.4°C, respectively. The median dewpoint temperature was 10.5°C and its 5th and 95th percentiles were -8.4°C and 21.9°C, respectively.

3.2 Exposure disparities

The quantile regression for the whole state shows a significant, positive correlation between average $PM_{2.5}$ and percent NHB population across all the quantiles modeled (Figure 1, panel a). This indicates that more polluted census tracts tend to have a higher percentage of NHB population across the entire state, regardless of the relative exposure level. Localized results from Mecklenburg and Wake counties (Figure 1, panels b and c) show the same significant, positive association for most quantiles studied. Figure 2 also shows that in both these counties, the majority of the least-polluted census tracts (those ranked in quartile 1 using average $PM_{2.5}$ as criteria) have a low percentage of NHB population, while the most polluted tracts (ranked in quartile 4) tend to have comparatively higher percentages of NHB residents.

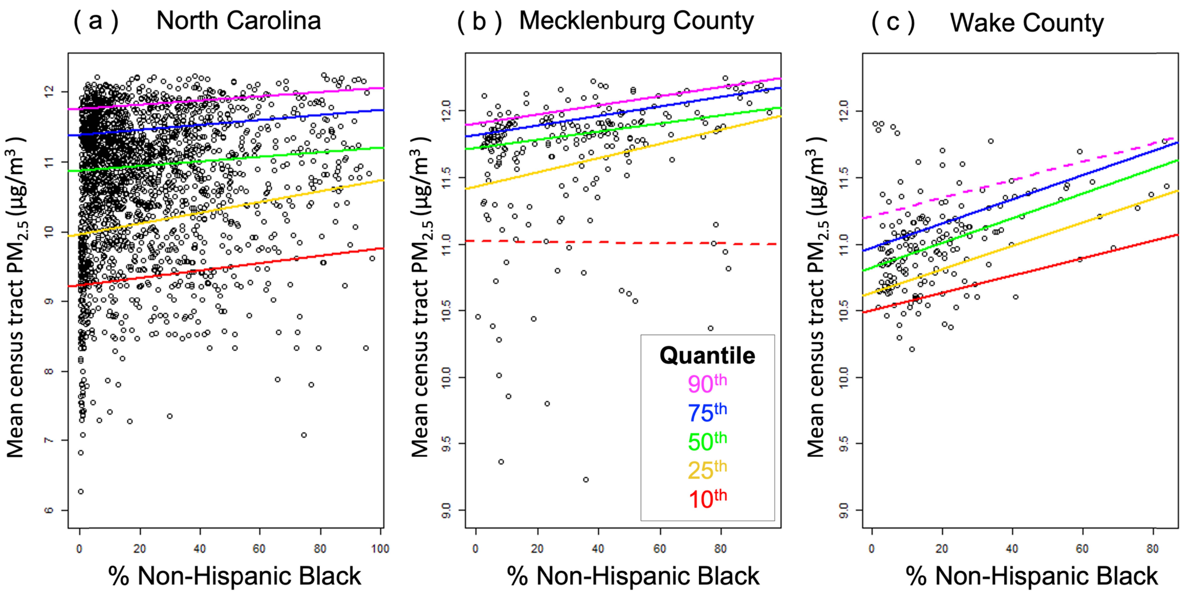

Figure 1. Quantile regression between census tract average $PM_{2.5}$ (years 2001-2016) and census tract percent of Non-Hispanic Black population for (a) all census tracts in North Carolina, (b) census tracts in Mecklenburg County, and (c) census tracts in Wake County. The inset in panel (b) provides a color reference for the quantiles plotted. Non-statistically significant results are represented with dashed lines. Note the y-axis scale in panel (a) is different from that in panels (b) and (c).

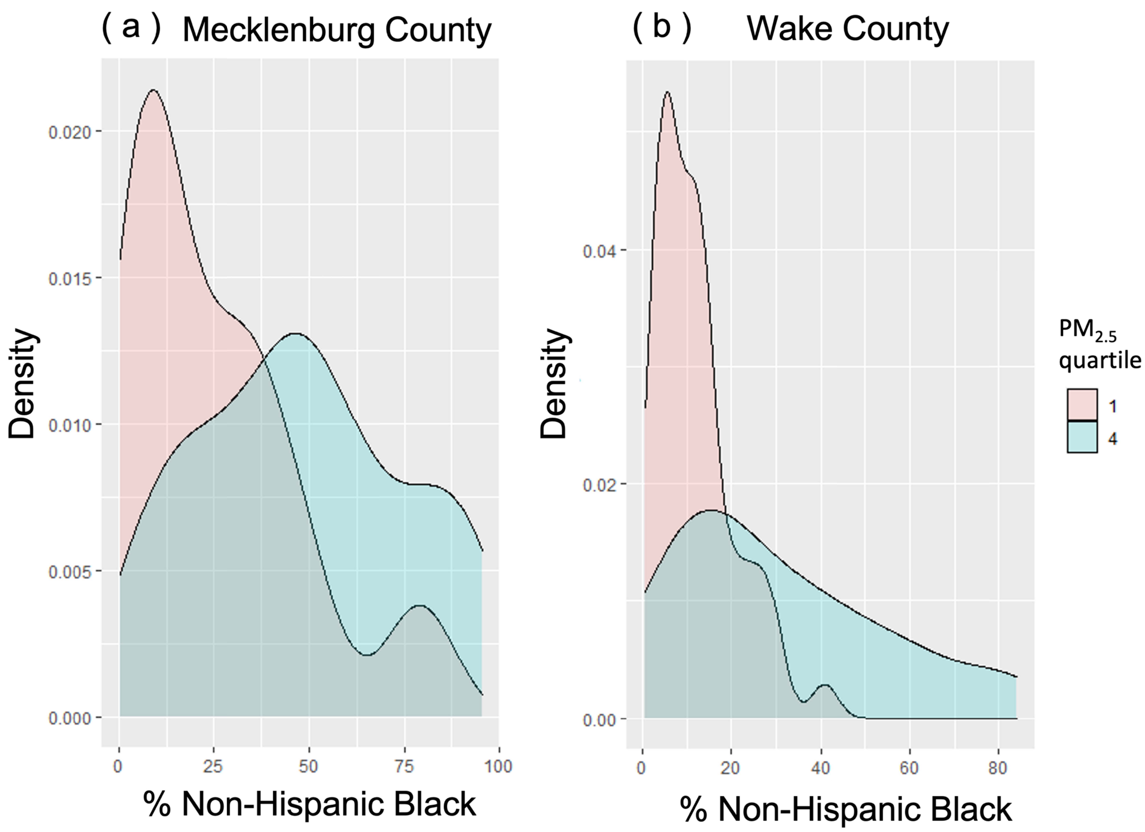

Figure 2. Density of percent Non-Hispanic Black for census tracts with average $PM_{2.5}$ in the first quartile (red) and in the fourth quartile (blue), for (a) Mecklenburg County and (b) Wake County.

3.3 Mortality model

We first present the results of the case-crossover model computed with the full record of mortality and using data from the highest resolution $PM_{2.5}$ gridded data (1km). We will later compare the changes in uncertainty for that model when fit with less data, by either reducing the number of mortality cases in the model or by using data from coarser $PM_{2.5}$ grids. All the model fits are performed with the same (4x4km) datasets for temperature and dewpoint temperature taken at the same temporal lags as the $PM_{2.5}$ data.

Table 1 reports the odds ratios for a 10 μg/m^3 increase in $PM_{2.5}$ (OR_{10}) and its 95% confidence intervals for the five different lags investigated. The significant associations observed were, in descending magnitude: for lag01, OR_{10} = 1.016 (95% CI 1.011–1.021); lag02, OR_{10} = 1.016 (95% CI 1.010–1.022); lag0, OR_{10} = 1.013 (95% CI 1.009–1.018), and lag1, OR_{10} =1.012 (95% CI 1.007–1.017). The association for lag2 was not statistically significant.

Our results were very similar to those of a previous study that used the same model design and mortality data (Son et al., 2020), with minor (and statistically non-significant) differences attributable to differences in sources and averaging techniques for the pollution and temperature data (comparison can be found in Figure S1).

Table 1. Odds Ratios and 95% confidence intervals for the association of PM2.5 with mortality at different lags. Non-significant results are colored in grey.

Lag	**OR_{10}**
Lag0	1.013 (1.009 - 1.018)
Lag1	1.012 (1.007 - 1.017)
Lag2	1.004 (0.999 - 1.008)
Lag01	1.016 (1.011 - 1.021)
Lag02	1.016 (1.010 - 1.022)

We also fit the case crossover models separately for the NHW and NHB cases to investigate effect differences between these population groups. Table 2 shows the OR_{10} and the (multiplicity adjusted) 95% confidence interval for each lag and race. The association between $PM_{2.5}$ and short-term mortality was significant in the NHW population for all lags except Lag2, the same lags where the association was also significant when the whole study population was represented (Table 1). This is a sensible result since the majority of the mortality cases studied come from the NHW population (77.4%). The results for the NHB population present wider confidence intervals, associated to the relatively lower number of cases that were used to fit the model since only 20.4% of the study population is NHB, making the multiplicity-adjusted results for NHB not statistically significant. We will use the Lag1 model for subsequent analysis since it was the lag with the closest to significant association for NHB.

Table 2. Odds Ratios and 95% confidence intervals for the association of $PM_{2.5}$ with mortality at different lags. Non-significant results are colored in grey.

Lag	OR_{10} NHW	OR_{10} NHB
Lag0	1.015 (1.010 - 1.020)	1.006 (0.992 - 1.021)
Lag1	1.013 (1.007 - 1.018)	1.010 (0.999 - 1.022)
Lag2	1.005 (0.998 - 1.011)	no effect
Lag01	1.018 (1.012 - 1.024)	1.011 (0.998 - 1.025)
Lag02	1.018 (1.011 - 1.025)	1.010 (0.994 - 1.026)

3.4 Uncertainty tradeoffs from information changes

To study uncertainty tradeoffs, we fit the model in equation (1) with varying input of either $PM_{2.5}$ data or mortality data (Y_i), in order to compare each of these datasets' influence in the final uncertainty of the case-crossover model, measured through the entropy of the exposure coefficient β, as explained in section 2.4.

First, we isolate the influence of changing air pollution data on the case-crossover model's uncertainty reduction. To achieve this, we fit the model with the full record of mortality data while varying $PM_{2.5}$ data, by fitting the model three times with $PM_{2.5}$ data of different resolutions (1km, 6km, and 12km). Figure 3 shows that fitting the model with finer resolution $PM_{2.5}$ data results in lower uncertainty of β. Since the $PM_{2.5}$ exposure is assigned based on each individual's gridcell of residence, a coarser grid may result in more deaths that happened the same day falling within the same gridcell, causing multiple cases to have identical $PM_{2.5}$ data. Although weather covariate data may still be different for each case (since these are always on the same 4km grid) making the cases sharing $PM_{2.5}$ data still likely distinct, the repeated sampling of the same $PM_{2.5}$ values does not provide new information to the model, therefore reducing the information value of the air pollution data input.

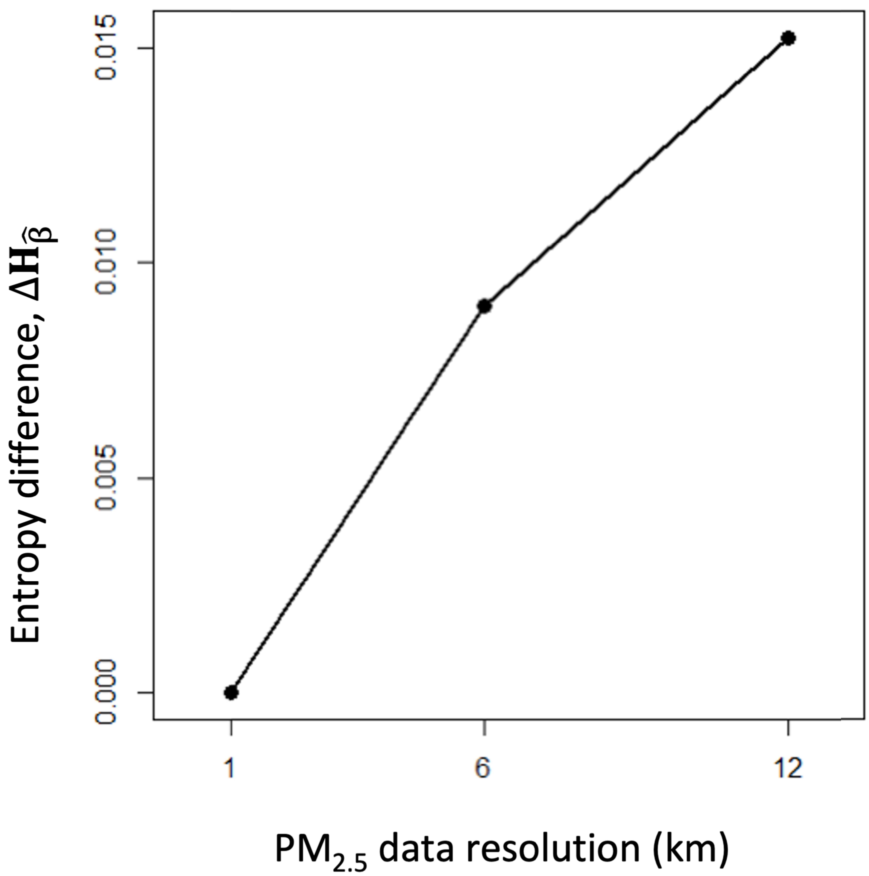

Figure 3. Entropy changes for the estimate of the exposure coefficient $\hat{\beta}$ for case-crossover model fit with $PM_{2.5}$ data of different spatial resolutions.

Then, we isolate the effect of changing mortality data in the uncertainty of the case-crossover model. To achieve this, we fit the model with the highest-resolution $PM_{2.5}$ data (1km) while varying the number of mortality cases input into the model. We do this analysis twice, selecting the missing cases to be either all from the NHW population or the NHB population, in order to investigate the effect of racial bias in the uncertainty reduction dynamics of health data. Since NHB cases represented approximately 20% of the study population, this is the maximum number of missing cases we explore for both races. Therefore, we initially fit the model with ~80% of data, and we then increase the number of cases to ~90% and finally to 100% data coverage. Figure 4 shows that while increasing the number of mortality cases reduces uncertainty in the model for both scenarios, the slope of uncertainty reduction is steeper when the new cases introduced are from the NHB population. The exposure disparities experienced by the NHB population shown in section 2.2 may be related to this difference, since differential exposure of a subpopulation may lead to a higher diversity of pollution data input in the model. This hypothesis is confirmed by the differences in the distribution of the mean of the Lag1 $PM_{2.5}$ data associated with cases and controls from the NHB population versus that one associated to the NHW population (Figure 5). The 95% confidence intervals between both distributions do not cross, making the mean $PM_{2.5}$ associated with NHB individuals statistically different from that of NHW individuals. At the lowest stage of information the model is fit with ~80% of the data, the majority of which comes from NHW individuals, so adding more data from NHW individuals will introduce samples from the $PM_{2.5}$ distribution that is already known the most. In contrast, new data from NHB individuals introduces information from a distribution of $PM_{2.5}$ that is different from the majority distribution, providing new information to the model and generating a faster uncertainty reduction. This result is not caused by the higher magnitude of the mean $PM_{2.5}$ for NHB shown in Figure 5, but by the fact that the NHB are a minority population with a statistically different $PM_{2.5}$ exposure distribution from that of the NHW population. Therefore, uncertainty reduction should have been steeper with new NHB data even if this subpopulation was exposed to less pollution than the NHW population, as long as the mean $PM_{2.5}$ between subpopulations remained statistically different.

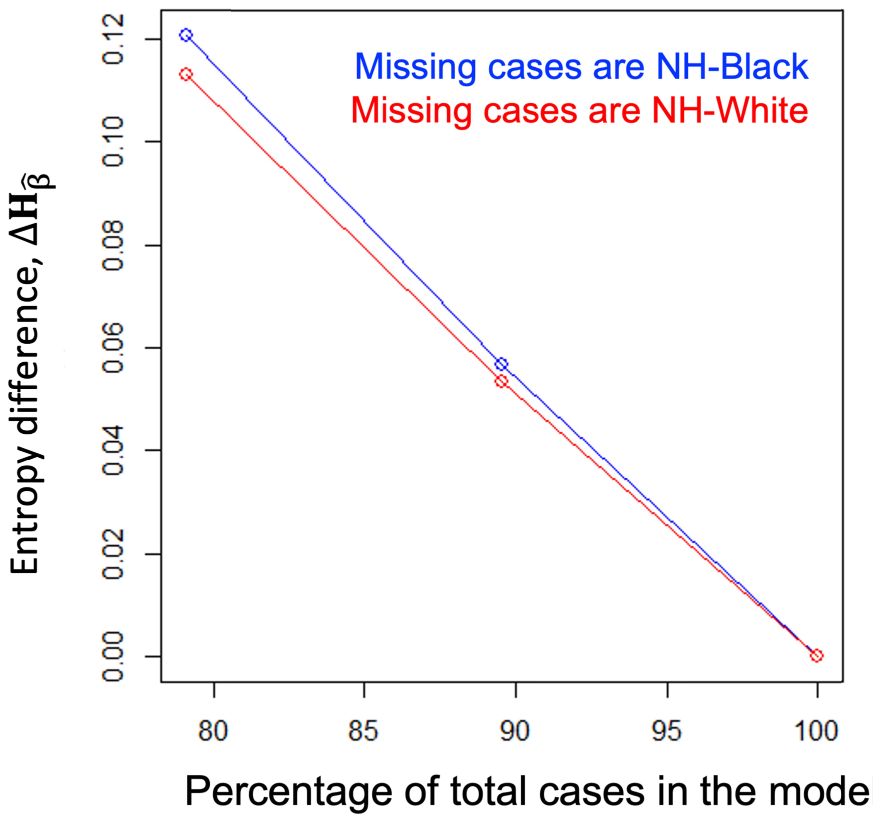

Figure 4. Entropy changes for the estimate of the exposure coefficient $\hat{\beta}$ for case-crossover model fit when more information is acquired for NHB cases only (blue series) or NHW cases only (red series).

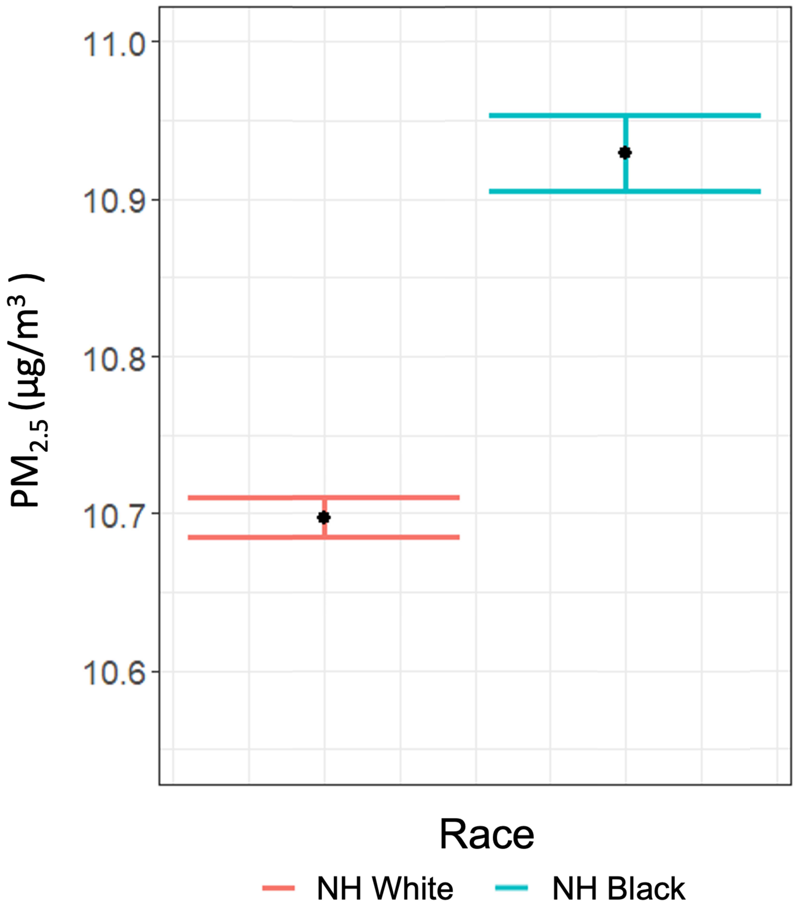

Figure 5. Mean of the lag-1 $PM_{2.5}$ associated with NHW cases (red) and NHB cases (blue) in the case-crossover model (equation (1)) computed with state-wide data, and its 95% confidence interval.

3.5 Information yield curve

While we showed in section 3.4 that increasing air pollution information and health effects information both reduce the uncertainty in the final mortality estimate, their contribution to uncertainty reduction is not equal. The information yield curve in Figure 6 compares the individual effects of information gain from each dataset in the model's uncertainty reduction. The dashed light-blue lines illustrate a graphical interpretation that can be used for decision-making purposes. If for a case scenario of interest, the target for mortality uncertainty reduction is $\Delta H_{\hat{\beta}}$ as indicated by the horizontal dashed lines, the change in the x axis required for the data in each side can be compared to find the most efficient pathway for uncertainty reduction. In the case below, increasing health data seems to reduce the uncertainty in the model more efficiently, since the same $\Delta H_{\hat{\beta}}$ can be achieved with a smaller change in x. However, the figure below presents a qualitative x-axis, as there is no common basis of comparison between increasing patient data and downscaling pollution model resolution. For a real-world scenario, stakeholders would be able to apply a common metric to these data improvements, such as cost or time, making the x-axis quantitative and potentially altering the decision-making outcomes presented here.

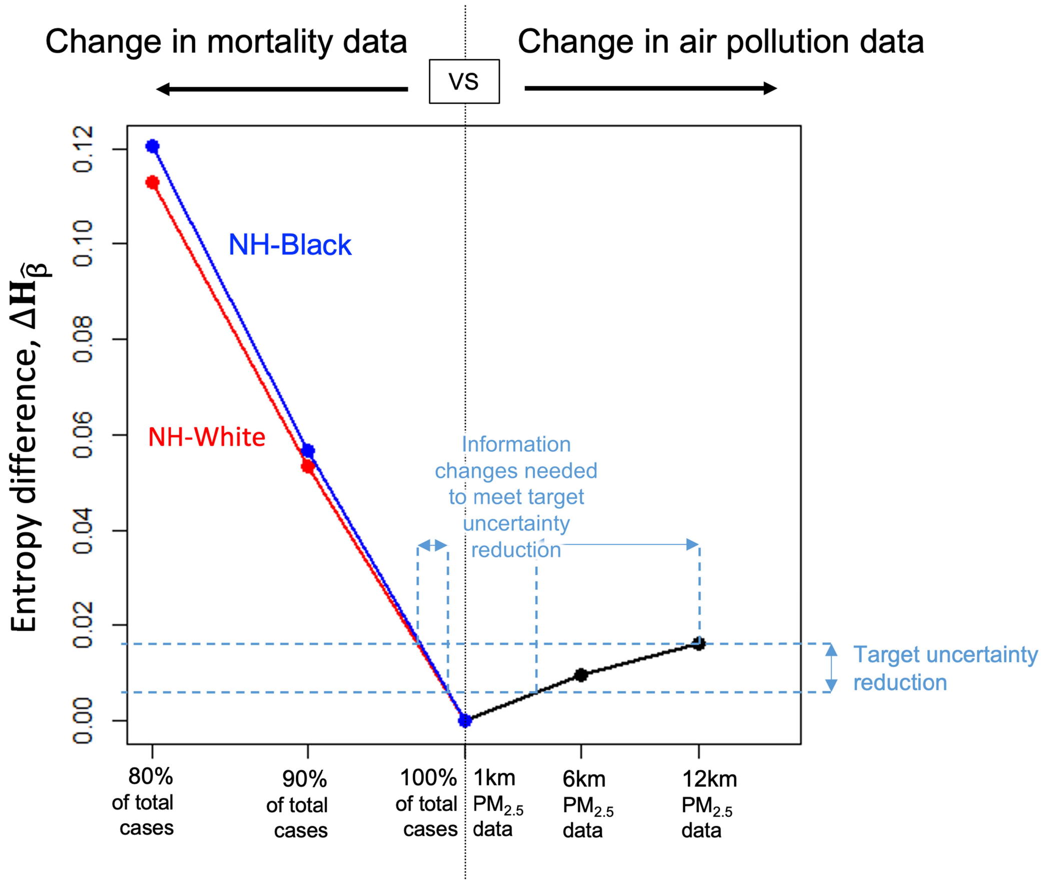

Figure 6. Information yield curve comparing the effect of information gain in mortality (left side) versus air pollution (right side) on the uncertainty reduction of the exposure coefficient in the case crossover model. The dashed light-blue lines provide graphical interpretation of the information yield curve by illustrating the different data increases necessary to achieve a fixed risk uncertainty reduction.

4. Discussion and conclusion

The results of this study illustrate the usefulness of our information entropy tradeoff methodology to not only generate more robust impact assessments, but also to gain new knowledge about the role of data from minority populations in the dynamics of uncertainty reduction.

We found associations between short-term $PM_{2.5}$ exposure and mortality for years 2001-2016 in North Carolina that were statistically significant and consistent with a previous study of the same mortality dataset (Son et al., 2020), despite the state's relatively low and decreasing air pollution levels. North Carolina had a state-wide average $PM_{2.5}$ concentration of 13.5 μg/m^3 in 2002, and state-wide decreases in concentrations resulted in the whole state presenting annual mean $PM_{2.5}$ below the EPA's standard of 12 μg/m^3 by 2016 (Bravo et al., 2022). Despite this improving trend in pollution concentrations, our findings add to the mounting evidence that particulate matter has detectable health effects even at pollution levels formerly seen as safe, motivating ongoing updates of air quality guidelines such as the EPA's proposal in January of 2023 to reduce the $PM_{2.5}$ standard to between 9 and 10 μg/m^3.

We also explored tradeoffs between data increases in air pollution or health outcomes in the uncertainty reduction of the case-crossover model used to investigate the pollution-mortality relationship. The information yield curve presented in Figure 6 compared the different uncertainty reduction effects of augmenting information in air pollution and health data. While both data types reduce uncertainty in the case-crossover model when information is increased, the effect of new data for mortality resulted in a steeper rate of uncertainty reduction. One qualification of this outcome is that information increase was done by different methods for each dataset, making the comparison of information change merely qualitative as there is no common variable in the x-axis of the information yield curve. If this method were applied to a scenario where information increases are associated to costs, time, or, as done in our previous study (Alifa et al., 2022), pollution/health model uncertainties, the comparison could be done qualitatively and the decision-making outcomes of the information yield curve may change. The goal of this work is not to provide an absolute answer to the choice between investing in pollution versus health information, but to develop a framework applicable to any data set and environmental exposure scenario used in any epidemiological model.

The positive relationship between average $PM_{2.5}$ and %NHB population found at the census tract level through quantile regression is consistent with previous findings of disparities in exposure for the NHB population in both nationwide (Miranda et al., 2011; Tessum et al., 2021; Woo et al., 2019); and regional (Bravo et al., 2016; Servadio et al., 2019; Stuart et al., 2009) studies. Our study of Mecklenburg and Wake counties further illustrated the presence of this inequality for the most populated areas of the state, which experience relatively higher levels of air pollution. However, the state-wide positive association found with respect to all the concentration quantiles also reveals that exposure inequalities can be detected not only among counties such as Mecklenburg and Wake with high emissions (placed in the high $PM_{2.5}$ quartiles), but also among counties with lower emissions (those in the low $PM_{2.5}$ quartiles), indicating that these racial inequalities may be independent from the relative difference in pollution levels between counties that have different emission types or levels of urbanicity, agreeing with recent nationwide findings (Liu et al., 2021; Tessum et al., 2021). These findings of exposure disparities are not reflected in the results of the stratified case crossover model, possibly due to the relatively low $PM_{2.5}$ levels in the state that result in relatively small magnitude of exposure disparities.

A key finding of this paper is that disparities in $PM_{2.5}$ exposure can affect model uncertainty reduction. If exposure from a certain minority subpopulation (in this case, the NHB population) is significantly different than that of the majority population, as shown in Figure 5, then data from this minority have relatively higher information value resulting in a faster rate of uncertainty reduction in the mortality model (Figure 4). The authors hypothesize that this result is transferrable to the study of any minority subpopulation (by race, income, residential location, etc.) that experiences a different exposure from the majority, implying that minority representation in environmental research benefits not only the minorities in question, but also the researchers and stakeholders performing the research. In a situation where there is a known or suspected environmental exposure difference between sub-populations, ensuring the representation of all groups in the data used for the environmental impact assessment will result in a wider sampling of the problem's information space, providing the quantitative advantage of reduced uncertainty. Since minority groups have been found to be both over-exposed and at times under-monitored (Stuart et al., 2009), the application of this framework will also provide researchers with increased awareness of both exposure and information disparities by design, contributing to the ongoing work of environmental justice.

There still remain multiple interesting opportunities for future expansion of the uncertainty reduction framework proposed in our first study (Alifa et al., 2022) and further expanded in this present work. One possible next step in future work is considering a case scenario where the assessment goes from an initial baseline of comparatively scarce pollution, epidemiology, or demographic information to subsequent stages of more information, via data augmentation methods such as assimilation, disaggregation, and/or downscaling. This work would require the integration of multiple datasets (e.g., by combing air pollution monitoring station data, gridded CTM output, and area-based demographic and health outcomes data), introducing new kinds of epistemic uncertainties, such as those stemming from errors in pollution and exposure measurements, model specification, data aggregation, and extrapolation of exposure-response functions, among others (Nethery & Dominici, 2019). These uncertainties are different from the one addressed in our framework in that they increase monotonically with the increase of input data, having the potential to obscure any uncertainty reduction from information gain if the epistemic errors in the data are too high (Rao, 2005). For this reason, our work so far has taken advantage of full datasets and simulated information scarcity by modeling only subsets of this data, which has allowed us to explore the proposed framework without having to deal with the epistemic uncertainties introduced by data assimilation errors.

The choice of North Carolina for this case study was prompted by the unique availability of high-resolution mortality data, but the relatively low $PM_{2.5}$ levels in the state prevented us from incorporating true data assimilation into this project, since the noise introduced by multiple $PM_{2.5}$ data sources would have been greater than the signal of the $PM_{2.5}$ data itself. This limitation speaks to the wider issue of data scarcity in air pollution, health outcomes, and demographics for the regions of the world that are most in need of epidemiology and exposure disparities studies.

The framework developed here could still be useful, however, for a case of interest where there is availability of pollution data only. As mentioned in the introduction, multiple methods to augment air pollution observations through assimilation of other datasets such as CTMs, satellite data, citizen-science observational networks have been devised in recent years. In a scenario where stakeholders want to augment their observational network but are unsure of which method to choose for the task, studying the information entropy tradeoffs between different data assimilation methods may be an efficient way to inform a decision. Furthermore, if demographic data is also available (such as census data), stakeholders would be able to investigate how information increases from different air pollution sources have different effects in the uncertainty of the estimates of exposure inequalities between different subpopulations, and whether focusing on augmenting data in regions with high versus low concentrations of minority populations yields different effects in uncertainty reduction.

As the scientific community continues efforts to improve characterization of environmental exposure effects for overlooked areas and populations around the world, the framework presented here gives researchers a new opportunity to elevate minority representation from a qualitative afternote in a study's discussion section to a centerpiece of the study's design, aiding a quantitatively more accurate analysis and producing confident estimates of the true effects of environmental pollution.

Acknowledgements

This publication is based upon work supported by the Lucy Family Institute for Data & Society at the University of Notre Dame, grant number 22006.

Open Research

The detailed death records data were obtained from the Children's Environmental Health Initiative (CEHI) at Notre Dame (Children's Environmental Health Initiative, 2020). These data are governed by data use agreements with data providers and protocols reviewed and approved by the Institutional Review Board (IRB) at the University of Notre Dame. The data may be accessed through a collaboration request to CEHI: https://www.cehidatahub.org/collaborate. The

1km gridded air pollution data was obtained from NASA's SEDAC (Di et al., 2021) and can be downloaded here: https://sedac.ciesin.columbia.edu/data/set/aqdh-pm2-5-concentrations-contiguous-us-1-km-2000-2016/data-download. The 4km gridded temperature and dewpoint temperature was obtained from the PRISM Climate Group at Oregon State University (PRISM Climate Group, 2004) and can be downloaded here: https://prism.oregonstate.edu/downloads/. The 2010 census data can be downloaded from the Census Bureau, https://data.census.gov/. All analyses were performed using R Statistical Software (v 4.2.3, R Core Team, 2023).

5. References

Alifa, M., Castruccio, S., Bolster, D., Bravo, M., & Crippa, P. (2022). Information entropy tradeoffs for efficient uncertainty reduction in estimates of air pollution mortality. *Environmental Research*, *212*, 113587.

Atkinson, R., Kang, S., Anderson, H., Mills, I., & Walton, H. (2014). Epidemiological time series studies of PM2. 5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*, *69*(7), 660-665.

Bates, J. T., Pennington, A. F., Zhai, X., Friberg, M. D., Metcalf, F., Darrow, L., . . . Russell, A. (2018). Application and evaluation of two model fusion approaches to obtain ambient air pollutant concentrations at a fine spatial resolution (250m) in Atlanta. *Environmental Modelling & Software*, *109*, 182-190.

Beckx, C., Panis, L. I., Uljee, I., Arentze, T., Janssens, D., & Wets, G. (2009). Disaggregation of nation-wide dynamic population exposure estimates in The Netherlands: Applications of activity-based transport models. *Atmospheric Environment*, *43*(34), 5454-5462.

Belbasis, L., & Bellou, V. (2018). Introduction to epidemiological studies. *Genetic epidemiology: methods and protocols*, 1-6.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Jama*, *292*(19), 2372-2378.

Bonas, M., & Castruccio, S. (2021). Calibration of Spatial Forecasts from Citizen Science Urban Air Pollution Data with Sparse Recurrent Neural Networks. *arXiv preprint arXiv:2105.02971*.

Bravo, M. A., Anthopolos, R., Bell, M. L., & Miranda, M. L. (2016). Racial isolation and exposure to airborne particulate matter and ozone in understudied US populations: Environmental justice applications of downscaled numerical model output. *Environment international*, *92*, 247-255.

Bravo, M. A., Fuentes, M., Zhang, Y., Burr, M. J., & Bell, M. L. (2012). Comparison of exposure estimation methods for air pollutants: ambient monitoring data and regional air quality simulation. *Environmental research*, *116*, 1-10.

Bravo, M. A., Warren, J. L., Leong, M. C., Deziel, N. C., Kimbro, R. T., Bell, M. L., & Miranda, M. L. (2022). Where is air quality improving, and who benefits? A study of PM2. 5 and ozone over 15 years. *American Journal of Epidemiology*, *191*(7), 1258-1269.

Breen, M., Chang, S. Y., Breen, M., Xu, Y., Isakov, V., Arunachalam, S., . . . Devlin, R. (2020). Fine-scale modeling of individual exposures to ambient PM2. 5, EC, NOx, and CO for the coronary artery disease and environmental exposure (CADEE) study. *Atmosphere*, *11*(1), 65.

Burnett, R., Pope III, C. A., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., . . . Brauer, M. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environmental health perspectives*, *122*(4), 397-403.

Chen, S.-Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease*, *9*(6), 1725.

Children's Environmental Health Initiative. (2020). *North Carolina Detailed Death Records during the period 2000 - 2017. [Data set]. CEHI.* https://doi.org/10.25614/ddrgeo_2000_2017.

Christakos, G. (2012). *Random field models in earth sciences*. Courier Corporation.

Coffman, E., Burnett, R. T., & Sacks, J. D. (2020). Quantitative Characterization of Uncertainty in the Concentration–Response Relationship between Long-Term PM2. 5 Exposure and Mortality at Low Concentrations. *Environmental Science & Technology*, *54*(16), 10191-10200.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., . . . Dandona, R. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907-1918.

Colmer, J., Hardman, I., Shimshack, J., & Voorheis, J. (2020). Disparities in PM2. 5 air pollution in the United States. *Science*, *369*(6503), 575-578.

Crippa, P., Sullivan, R., Thota, A., & Pryor, S. (2019). Sensitivity of simulated aerosol properties over eastern North America to WRF-Chem parameterizations. *Journal of Geophysical Research: Atmospheres*, *124*(6), 3365-3383.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., . . . Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology: a Journal of the Royal Meteorological Society*, *28*(15), 2031-2064.

De Barros, F., & Rubin, Y. (2008). A risk-driven approach for subsurface site characterization. *Water resources research*, *44*(1).

De Barros, F., Rubin, Y., & Maxwell, R. M. (2009). The concept of comparative information yield curves and its application to risk-based site characterization. *Water Resources Research*, *45*(6).

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., . . . Schwartz, J. (2019). An Ensemble-based Model of PM2.5 Concentration Across the Contiguous United States with High Spatiotemporal Resolution. *Environment International*, *130*, 104909.

Di, Q., Wei, Y., Shtein, A., Hultquist, C., Xing, X., Amini, H., . . . Mickley, L. J. (2021). *Daily and Annual PM2.5 Concentrations for the Contiguous United States, 1-km Grids, v1 (2000 - 2016)* NASA Socioeconomic Data and Applications Center (SEDAC).

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama*, *295*(10), 1127-1134.

EPA, U. S. (1997). National ambient air quality standards for particulate matter: Final rule. *Fed. Regist.*, *62*(138), 38,651-638,701.

EPA, U. S. (2013). National Ambient Air Quality Standards for Particulate Matter. *Fed. Regist.*, *78*(10), 3,086-083,287.

EPA, U. S. (2019). *Integrated Science Assessment (ISA) for Particulate Matter*

EPA, U. S. (2023). Reconsideration of the National Ambient Air Quality Standards for Particulate Matter. *Fed. Regist.*, *88*(18), 5,558-555,719.

EU. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*.

Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., . . . Corra, L. (2022). Pollution and health: a progress update. *The Lancet Planetary Health*.

Giani, P., Anav, A., De Marco, A., Feng, Z., & Crippa, P. (2020). Exploring sources of uncertainty in premature mortality estimates from fine particulate matter: the case of China. *Environmental Research Letters*, *15*(6), 064027.

Giani, P., Castruccio, S., Anav, A., Howard, D., Hu, W., & Crippa, P. (2020). Short-term and long-term health impacts of air pollution reductions from COVID-19 lockdowns in China and Europe: a modelling study. *The Lancet Planetary Health*, *4*(10), e474-e482.

Ha, S., Hui, H., Roussos-Ross, D., Haidong, K., Roth, J., & Xu, X. (2014). Th effects of air pollution on adverse birth outcomes. *Environmental research*, *134*, 198-204.

Hajat, A., Hsia, C., & O'Neill, M. S. (2015). Socioeconomic disparities and air pollution exposure: a global review. *Current environmental health reports*, *2*(4), 440-450.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. John Wiley & Sons, Inc.

Hyder, A., Lee, H. J., Ebisu, K., Koutrakis, P., Belanger, K., & Bell, M. L. (2014). $PM_{2.5}$ exposure and birth outcomes: Use of satellite- and monitor-based data. *Epidemiology*, *25*(1), 58-67.

Jaakkola, J. (2003). Case-crossover design in air pollution epidemiology. *European Respiratory Journal*, *21*(40 suppl), 81s-85s.

Kloog, I., Melly, S. J., Ridgway, W. L., Coull, B. A., & J., S. (2012). Using new satellite based exposure methods to study the association between pregnancy $PM_{2.5}$ exposure, premature birth and birth weight in Massachusetts. *Environmental Health*

18(11).

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, *15*(4), 143-156.

Liu, J., Clark, L. P., Bechle, M. J., Hajat, A., Kim, S.-Y., Robinson, A. L., . . . Marshall, J. D. (2021). Disparities in air pollution exposure in the United States by race/ethnicity and income, 1990–2010. *Environmental Health Perspectives*, *129*(12), 127005.

McClellan, R. O. (2002). Setting ambient air quality standards for particulate matter. *Toxicology*, *181*, 329-347.

Mead, M. I., Castruccio, S., Latif, M. T., Nadzir, M. S. M., Dominick, D., Thota, A., & Crippa, P. (2018). Impact of the 2015 wildfires on Malaysian air quality and exposure: a comparative study of observed and modeled data. *Environmental Research Letters*, *13*(4), 044023.

Miranda, M. L., Edwards, S. E., Keating, M. H., & Paul, C. J. (2011). Making the environmental justice grade: the relative burden of air pollution exposure in the United States. *International journal of environmental research and public health*, *8*(6), 1755-1771.

Navidi, W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics*, 596-605.

Nethery, R. C., & Dominici, F. (2019). Estimating pollution-attributable mortality at the regional and global scales: challenges in uncertainty estimation and causal inference. *European heart journal*, *40*(20), 1597-1599.

Nhung, N. T. T., Amini, H., Schindler, C., Joss, M. K., Dien, T. M., Probst-Hensch, N., . . . Künzli, N. (2017). Short-term association between ambient air pollution and pneumonia in children: A systematic review and meta-analysis of time-series and case-crossover studies. *Environmental Pollution*, *230*, 1000-1008.

Pampel, F. C. (2020). *Logistic regression: A primer*. Sage publications.

Pope, C. A., Coleman, N., Pond, Z. A., & Burnett, R. T. (2020). Fine particulate air pollution and human mortality: 25+ years of cohort studies. *Environmental research*, *183*, 108924.

PRISM Climate Group. (2004). Oregon State University. https://prism.oregonstate.edu/. Data accessed September 2021.

Qian, D., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., . . . Lyapustin, A. (2019). An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment international*, *130*, 104909.

R Core Team, R. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, K. S. (2005). Uncertainty analysis in atmospheric dispersion modeling. *Pure and applied geophysics*, *162*(10), 1893-1917.

Servadio, J. L., Lawal, A. S., Davis, T., Bates, J., Russell, A. G., Ramaswami, A., . . . Botchwey, N. (2019). Demographic inequities in health outcomes and air pollution exposure in the Atlanta area and its relationship to urban infrastructure. *Journal of Urban Health*, *96*, 219-234.

Shen, P., Crippa, P., & Castruccio, S. (2021). Assessing Urban Mortality from Wildfires with a Citizen Science Network. *Air Quality, Atmosphere & Health.*, *Under review*.

Son, J.-Y., Lane, K. J., Miranda, M. L., & Bell, M. L. (2020). Health disparities attributable to air pollutant exposure in North Carolina: Influence of residential environmental and social factors. *Health & place*, *62*, 102287.

Stuart, A. L., Mudhasakul, S., & Sriwatanapongse, W. (2009). The social distribution of neighborhood-scale air pollution and monitoring protection. *Journal of the Air & Waste Management Association*, *59*(5), 591-602.

Tessum, C. W., Hill, J. D., & Marshall, J. D. (2017). InMAP: A model for air pollution interventions. *PloS one*, *12*(4), e0176131.

Tessum, C. W., Paolella, D. A., Chambliss, S. E., Apte, J. S., Hill, J. D., & Marshall, J. D. (2021). PM2. 5 polluters disproportionately and systemically affect people of color in the United States. *Science Advances*, *7*(18), eabf4491.

Van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., . . . Lee, C. (2021). Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science & Technology*, *55*(22), 15287-15300.

Van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2015). Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental health perspectives*, *123*(2), 135-143.

WHO. (2006). *Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global update 2005 - Summary of risk assessment*. W. H. Organization.

Woo, B., Kravitz-Wirtz, N., Sass, V., Crowder, K., Teixeira, S., & Takeuchi, D. T. (2019). Residential segregation and racial/ethnic disparities in ambient air pollution. *Race and social problems*, *11*, 60-67.

Zani, N. B., Lonati, G., Mead, M., Latif, M., & Crippa, P. (2020). Long-term satellite-based estimates of air quality and premature mortality in Equatorial Asia through deep neural networks. *Environmental Research Letters*, *15*(10), 104088.