The Census Environmental Impacts Frame[*]

John Voorheis[†], Jonathan Colmer, Kendall Houghton, Eva Lyubich,
Mary Munro, Cameron Scalera, Jennifer Withrow

August 27, 2024

Abstract

Research in environmental economics often relies on aggregated place-based data, limiting our understanding of the interplay between economic activity and environmental conditions due to within-location heterogeneity. To mitigate this we introduce the Census Environmental Impacts Frame (EIF), a new microdata infrastructure linking individual-level socioeconomic data and residential histories with high-resolution estimates of environmental conditions for nearly all U.S. residents over the past two decades. The EIF enables researchers to analyze the distribution of environmental amenities and hazards across individuals rather than places, evaluate when, how, and why environmental exposures have evolved over time, and investigate the economic consequences of exposure to various environmental conditions at unprecedented levels of detail. This paper describes the EIF, provides an illustration of aggregation bias in the context of air pollution, and summarizes opportunities for researchers.

JEL Codes: C55, C81, Q5.

[*] Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. We thank Nick Muller, Corbett Grainger and the participants at the AERE Annual Meeting, CRIW Conference on Natural Capital Accounting, and Census Bureau research seminars for helpful thoughts, comments, and discussions. We thank Surya Menon and Yolande Tra for research assistance. Colmer is grateful to the National Science Foundation (CAREER Award SES:2338220), Sloan Foundation, Russell Sage Foundation, Washington Center for Equitable Growth, Upjohn Institute, and the University of Virginia's Bankard Fund for Political Economy, Center for Real Estate and the Built Environment, and Environmental Institute for generous funding and support. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System (DMS) number: P-7505723, Disclosure Review Board (DRB) approval numbers: CBDRB-FY2022-CES010-017, CBDRB-FY23-CES014-003, CBDRB-FY23-CES014-005, CBDRB-FY2023-CES010-004 and CBDRB-FY23-CES019-009)

[†] Affiliations: Voorheis, Houghton, Lyubich and Withrow: Center for Economic Studies, US Census Bureau; Scalera: Columbia University; Colmer: University of Virginia; Munro: MITRE Corp. Corresponding Author: John Voorheis, john.l.voorheis@census.gov

1 Introduction

The natural environment is central to economic activity and human well-being. It provides essential inputs for production and directly influences various dimensions of well-being, including health, productivity, and recreation. While there has been tremendous progress in reducing exposure to some environmental hazards, other aspects of the environment are deteriorating, and threaten to undermine the progress that has been made, e.g., wildfires and air pollution (Shapiro, 2022; USGCRP, 2023; Burke et al., 2023b; Lang et al., 2024). Data limitations have constrained efforts to quantify the complex interplay between economic activity and the environment. Frequently, researchers are limited to using aggregate data about *places*, which severely constrains our ability to understand how environmental conditions are distributed, why differences in exposures exist, and their consequences for *people*.

To improve data access and help researchers move from a focus on places towards people, we introduce the Census Environmental Impacts Frame (EIF), a new microdata infrastructure that facilitates individual-level analysis of environmental exposures, their determinants, and their consequences across the United States. The EIF leverages confidential Census Bureau data to provide detailed demographic, economic, and location information for nearly all U.S. residents from the late 1990s onward. This comprehensive microdata unlocks several key research directions that are not feasible with place-based aggregate data: (1) a precise characterization of environmental exposures that accounts for the fact that people and environmental conditions are not uniformly distributed across space at a given point in time; (2) quantification of environmental exposures over time across and within-individuals accounting for migration decisions; and (3) detailed analyses of heterogeneity across intersectional individual-level identities, such as differences by race and sex, within a given part of the income distribution.

The EIF represents a significant advance in the data landscape. As new data become available at Census, the EIF will be updated to extend the economic panel and incorporate advances in the measurement of environmental conditions. When finalized, it will be accessible to approved researchers through secure Federal Statistical Research Data Centers (FSRDCs), opening up new opportunities for groundbreaking research. To bridge the gap in the interim until FSRDC-based research can commence, the Census Bureau will release data products derived from the EIF, including interactive visualizations and privacy-protected gridded population estimates. These releases will allow researchers to begin exploring the rich potential of this novel data infrastructure, while maintaining essential confidentiality safeguards.

2 Data Construction

The Environmental Impacts Frame (EIF) is a modular infrastructure that combines confidential Census Bureau data with spatial data on environmental amenities and hazards. It consists of two core modules: a demographic "spine" containing basic information for the universe of individuals observed in the microdata, and an annual residential history file (RHF) for individuals in the spine. This modular set up maximizes flexibility by allowing researchers to combine the modules into a single panel or link any of the modules with external data sources. It also provides a foundation for the creation of additional linkable modules on, for example, housing characteristics, employment histories, and family structure. The EIF enables researchers to develop a systematic understanding of individual-level exposures to environmental hazards and amenities over several decades.

In the remainder of this section, we provide a more detailed overview of the two core modules, as well as the environmental data that has been incorporated into the EIF so far.

The Demographic Spine

The foundation of the EIF is the Census Bureau Numident – an administrative records dataset of individuals who have applied for Social Security Numbers (SSNs). We use the Numident to construct the core list of individuals in the EIF, which we refer to as the ”spine.”

The spine serves as the backbone upon which we merge all other information. Individuals are identified using Protected Identification Keys (PIKs), a unique identifier assigned to each individual via the Census data linkage infrastructure (Wagner and Layne, 2014).

The spine provides key demographic variables for each individual – date of birth, place of birth (state and county), an indicator for whether the individual was born outside the United States, date of death (if applicable), sex, race and ethnicity. This information is sourced from the Numident and an internal Census Bureau file called the Title 13 best race and ethnicity file, which aggregates information from survey and administrative records to provide a harmonized race and ethnicity classification that is consistent over time.

To ensure data quality and consistency, we process and clean the place of birth information in the spine. Using the July 2023 version of the place of birth geographic crosswalk, created as part of the Decennial Census Digitization and Linkage Project, we assign state and county FIPS codes to individuals born in one of the 50 states or Washington, D.C. For those born in the U.S., we successfully add state and county FIPS codes to over 99% of PIKs. Individuals born outside the 50 states or Washington, D.C., such as those born overseas or in U.S. territories, do not have detailed place of birth information in the spine and are maintained without cleaned place of birth information. Date of birth is used directly from the Numident, while date of death is constructed as a composite variable combining information from theNumident, Medicare data and from commercial data from the Veterans' Service Group of Illinois (VSGI).

The Residential History File

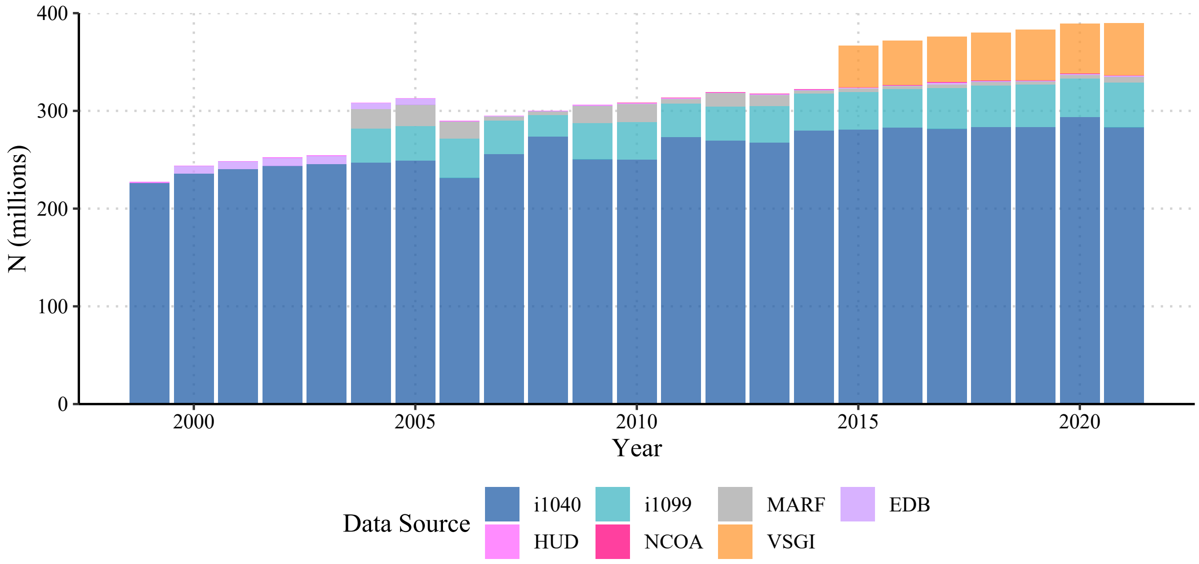
We use administrative data sources to construct residential histories for each individual PIK. The residential history file (RHF) identifies the most likely primary residential address for each person in a given year.

We select a single location for each individual in each year using data from, in order of preference, IRS income tax returns (Form 1040); Medicare Enrollment Database (EDB);

IRS Information Returns (Form 1099); Master Address File Auxiliary Reference File (MAF-ARF); US Department of Housing and Urban Development (HUD) Program Files; US Postal Service National Change of Address (NCOA) database; and VSGI data. Our preference ranking across administrative records sources is driven by three principles: we prefer data sources with broader coverage over data sources with narrower coverage; we prefer more precise location information over less precise location information; and we prefer data sources with less linkage uncertainty over sources with more uncertainty. Location information in the underlying data consists of a MAFID (a numeric ID associated with the Census Bureau's Master Address File, which lists housing units in the United States) and a ZIP code, which comes at two levels of detail – a five-digit code (zip5) and an additional four digits (zip9 or zip+4). We independently assign MAFID and ZIP codes (prioritizing zip+4 over zip5) based on their availability in the source data using the preference ranking described above. As a result, the ZIP codes and MAFIDs for a given PIK will not always correspond to the same source data.

Finally, we attach the longitude and latitude to every individual's residential address in each year. When a MAFID is available, we use its associated longitude and latitude from the MAF. When a MAFID is not available, we use the longitude and latitude of the ZIP code "centroid", which is defined as the population-weighted mean longitude and latitude of all residence within that ZIP code. Figure 1 presents the data source from which we construct the geographic coordinate locations of PIKs in the EIF. The majority of PIKs are assigned an address using 1040s.

Figure 1: Address Data Source



Source: Environmental Impacts Frame Spine and Residential History File, 1999-2021. **Notes:** This Figure reports the source of address data for individuals in the EIF over time.

Environmental Data

The EIF allows researchers to develop a systematic and comprehensive understanding of environmental exposures at a high spatial resolution, shifting the data constraint on spatial granularity from socioeconomic data to environmental data. Environmental modules that are currently being constructed are derived from data products on ambient exposures, modelled risk, or facility-level characteristics, including: air pollution concentrations; wildfire burn perimeters, wildfire risk metrics, and wildfire smoke plume data; historical flood inundations, rainfall, and flood risk metrics; hurricane wind field exposures; surface temperature, air temperature, and heat wave measures; proximity to Superfund sites, polluting facilities, brownfields, and other sites of interest; projected sea level rise inundation; and airborne toxic releases. Multiple ongoing research projects use this data to provide comprehensive evidence on the distribution of exposures, their causes, and their consequences (Colmer et al., 2023a; Chakma et al., 2023b,a; Burke et al., 2023a; Colmer et al., 2023c,b; Colmer and

Voorheis, 2024).

3 Coverage and Data Quality

The core modules of the EIF leverage the Census Bureau's data linkage infrastructure to combine multiple administrative and survey data sources, providing the most comprehensive information available on residential histories and basic demographics for United States residents. Despite the high quality of these data, some sub-populations may be underrepresented, or not covered at all. For example, because of its reliance on the Numident, the EIF spine excludes individuals who do not have a SSN, such as undocumented migrants. The EIF is also likely to systematically miss individuals who lack a connection to the formal economy in a given year, as its input datasets cover activities related to employment, asset ownership, and other aspects of the formal economy. Finally, data sources such as the EDB and HUD administrative data only cover eligible individuals who take up these programs, while the USPS NCOA data contain only self-reported moves. Although the EIF combines these datasets to provide robustness to selection issues, it is not possible to rule out the possibility that some individuals will be missing for non-random reasons.

To assess potential coverage issues in the EIF, we examine basic descriptive facts, using our best estimates of the US population as a benchmark. We do this two ways. First, we benchmark the RHF against the spine. The spine contains all non-deceased individuals who applied for SSNs, and may differ from the actual US population for three reasons: it excludes undocumented migrants and other residents who do not have a social security number, it can include SSN holders who are no longer residents of the US, and the Numident contains incomplete death information in earlier years, resulting in many individuals who died before the 1990s lacking a valid date of death(Finlay and Genadek, 2021).[1] Our second benchmark is the American Community Survey, which is a nationally representative annual sample, but

[1] Finlay and Genadek (2021) presents evidence that this issue is particularly acute for individuals who died before 1970.

in practice may have unequal PIK assignment rates or survey coverage across demographic groups when linked to other data (Wagner and Layne, 2014). The detailed results of this analysis are reported in Appendix B.

Our analysis shows that coverage rates for the EIF are encouragingly high, and have been improving over time. Overall we conclude that use of this data infrastructure should capture a set of individuals close to the actual US population, especially in recent years. Nevertheless, there remain small differences in coverage across groups which researchers should be aware of. In particular, coverage rates are slightly lower for men, for lower income individuals, and, depending on the year, for Hispanic and most non-Hispanic non-White race groups. Differences in coverage by race and ethnicity have declined over time.

4 Aggregation Bias and the Distribution of $PM_{2.5}$ Air Pollution

To underscore the advantages of individual level data over place-based aggregate data, we present a case study using $PM_{2.5}$ concentrations. When relying on public-use demographic data, researchers are typically constrained to administrative units like census tracts or counties. These large, non-uniform geographic areas can exhibit meaningful intra-area heterogeneity in both pollution concentrations and population distributions. Recent advancements in remote-sensing data products and computational capabilities have enabled finer-grained pollution measures on a $0.01° \times 0.01°$ grid (Van Donkelaar et al., 2016; Di et al., 2016; van Donkelaar et al., 2021). However, the necessity of aggregating these granular pollution data to match coarser demographic data has precluded fully leveraging this increased spatial resolution, potentially obscuring exposure differentials stemming from intra-area variation in populations.

Using the EIF, we document substantial variation in $PM_{2.5}$ concentrations across individuals within counties and census tracts in the contiguous United States. A variance

Table 1: The share of variance in $PM_{2.5}$ concentrations within different administrative regions, by race and ethnicity (2022)

	(1) All	(2) White	(3) Black	(4) Hispanic	(5) Asian	(6) AIAN
Within-County Variance Share	18.7%	19.2%	18.7%	19.4%	21.0%	14.6%
Within-Tract Variance Share	3.1%	4.0%	3.2%	2.1%	3.2%	3.4%

Source: Environmental Impacts Frame Residential History File, 1999-2022, and van Donkelaar et al. (2021). **Notes:** This Table reports $PM_{2.5}$ concentrations by race using cell, census tract, or county-level aggregations of pollution.

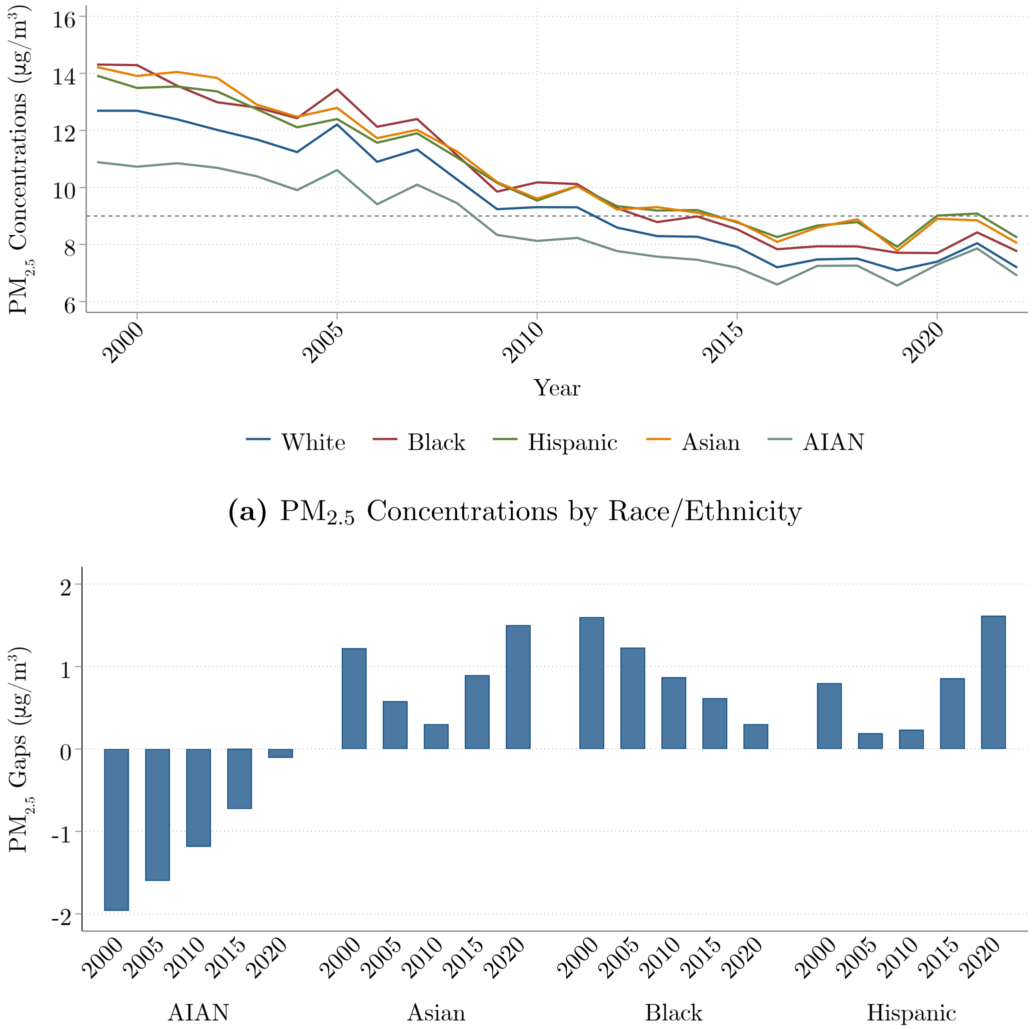
decomposition reveals that 49% of the total variation in $PM_{2.5}$ is within-county, with the relative importance of this intra-area component varying markedly across racial/ethnic groups – 32% for non-Hispanic White individuals, 21% for non-Hispanic Black individuals, 68% for Hispanic individuals, 58% for Asians, and 55% for American Indians and Alaska Natives (Table 2). As we shift to smaller geographic units, the within-region variation in $PM_{2.5}$ diminishes, with current remote sensing data constraints yielding negligible intra-census block group heterogeneity.

Motivated by these results, we explore how the geographic resolution of data affects inferences about the distribution of $PM_{2.5}$ exposures. We investigate trends in $PM_{2.5}$ exposure by demographic group, assigning pollution levels at the $1km^2$ grid cell, the most granular option available from satellite data. Figure 2 presents these trends from 1999 to 2022. Consistent with Currie et al. (2020), Panel A reveals substantial air quality improvements for all groups since 1999, and Panel B highlights the narrowing of the absolute Black-White gap. However, Panel B also highlights that reductions in absolute disparities have not been universal for all groups. While Hispanic-White and Asian-White $PM_{2.5}$ gaps narrowed between 1999 and 2010 they have since widened. In absolute terms, Hispanic and Asian individuals are now exposed to higher levels of $PM_{2.5}$ concentrations than Black individuals.[2] The advantageous AIAN-White gap has also shrunk substantially overtime.

Table 2, Panel A reports exposures shown in 2 in 2000 and 2022 as well as the change between these years. Panels B and C report the same outcomes using mean $PM_{2.5}$ calculated over larger geographic areas: Census tracts and counties. Relative to cell-level assignment of $PM_{2.5}$ concentrations, assigning individuals their census tract's average concentrations results in a small attenuation. This suggests that aggregation bias, at least along margins of race and ethnicity, are not a first-order concern at geographies smaller than or equal to Census tracts. By contrast, assigning individuals their county's average concentrations results in a significant attenuation. This attenuation varies substantially by race and ethnicity. We calculate that average concentrations in 2000 are 1.94, 1.61, 0.70, 0.63, and 0.53 $\mu g/m^3$ lower for Hispanic, Asian, AIAN, Black, and White individuals respectively. In 2022, these differences are 1.06, 0.83, 0.54, 0.4, and 0.36 $\mu g/m^3$. Importantly, aggregation bias affects not only the magnitude of exposures but also their relative ranking across groups. When individuals are assigned their county's average concentrations, the average Black individual is exposed to the highest $PM_{2.5}$ concentrations in 2022, rather than the third highest as

[2] One possible explanation for this pattern is the increasing severity of wildfire smoke, which disproportionately affects areas with large Hispanic and Asian populations.

Figure 2: Racial and Ethnic Trends and Gaps in $\text{PM}_{2.5}$ Concentrations



(a) $\text{PM}_{2.5}$ Concentrations by Race/Ethnicity

(b) $\text{PM}_{2.5}$ Gaps Relative to NH White Individuals

Source: Environmental Impacts Frame Residential History File, 1999-2022, and van Donkelaar et al. (2021).
Notes: See Section 2 for details on construction.

Table 2: Aggregation Bias in $PM_{2.5}$ Concentrations

	(1)	(2)	(3)	(4)	(5)
	White	Black	Hispanic	Asian	AIAN
Panel A: Cell					
2000	12.69	14.29	13.49	13.91	10.73
2022	7.19	7.76	8.25	8.05	6.91
Δ 2000–2022	-5.50	-6.53	-5.24	-5.86	-3.82
Panel B: Census Tract					
2000	12.66	14.28	13.48	13.91	10.69
2022	7.16	7.75	8.23	8.04	6.84
Δ 2000–2022	-5.50	-6.53	-5.25	-5.87	-3.85
Panel C: County					
2000	12.16	13.66	11.55	12.30	10.03
2022	6.83	7.36	7.19	7.22	6.37
Δ 2000–2022	-5.33	-6.30	-4.47	-5.08	- 3.66

Source: Environmental Impacts Frame Residential History File, 1999-2022, and van Donkelaar et al. (2021). **Notes:** This Table reports $PM_{2.5}$ concentrations by race using cell, census tract, or county-level aggregations of pollution.

indicated by the cell-level data.

These findings suggest that the use of county-level demographic data is likely to result in significant bias when analyzing $PM_{2.5}$ disparities. Aggregation bias is less of a concern with tract-level data, but its use limits scope for additional layers of heterogeneity analysis. Our results underscore the value of leveraging granular, individual-level data, such as that provided by the EIF, to accurately assess environmental exposures and disparities across demographic groups.

5 Conclusion

This paper introduces the Environmental Impacts Frame (EIF), a novel microdata infrastructure that allows researchers to combine individual-level data on socio-economic circumstances and granular information about where people live with spatial data on environmental conditions.

While our chosen proof of concept relates to air pollution, the EIF enables comprehensive analysis of environmental amenities, hazards, regulations, and policies. For example, the EIF is an ideal framework for studying the distribution of exposure to extreme weather and natural disasters, such wildfires and hurricanes, extreme heat, flood events, and sea level rise. More broadly, the EIF's longitudinal nature makes it well-suited for analyzing dynamic responses, making it the only population-scale framework with the capacity to seriously engage with questions related to environmental migration, sorting, siting, and environmental gentrification.

The EIF will evolve over time, with planned extensions including incorporating additional income measures, worker-firm links, historical environmental data, family linkages, and housing characteristics. It will integrate with existing Census Bureau data infrastructure, enhancing existing data products (e.g. the Census Community Resilience Estimates) and environmental exposure statistics. It will also be available to researchers working on approved projects at Federal Statistical Research Data Centers (FSRDCs). By representing the rich interconnections between people, economic activity, and environmental conditions, the EIF can fundamentally advance our understanding of coupled human and natural systems in the United States.

References

Burke, M., J. Colmer, C. Scalera, and J. Voorheis, "Wildfire Smoke and $\text{PM}_{2.5}$ Disparities in the United States," 2023. Mimeo.

_ , M.L. Childs, B. de la Cuesta, M. Qiu, J. Li, C. Gould, S. Heft-Neal, and M. Wara, "The Contribution of Wildfire to PM2.5 trends in the USA.," *Nature*, 2023, *622*, 761–766.

Chakma, T., J. Colmer, and J. Voorheis, "The Causes and Consequences of Urban Heat Islands," 2023. Mimeo.

_ , _ , and _ , "Racial Heat Disparities in the United States Are Not Reconciled by Differences in Economic Circumstances," 2023. Mimeo.

Colmer, J. and J. Voorheis, "Microdata and Natural Capital Valuation," in Mary Bohman, Eli Fenichel, and Nicholas Muller, eds., *Measuring and Accounting for Environmental Public Goods: A National Accounts Perspective*, University of Chicago Press, 2024.

_ , _ , and B. Williams, "Air Pollution and Economic Opportunity in the United States," 2023. Mimeo.

_ , _ , and _ , "Who Weathers the Storm? The Unequal Effects of Hurricanes in the United States," 2023. Mimeo.

_ , S. Qin, J. Voorheis, and R. Walker, "Income, Wealth, and Environmental Inequality in the United States," 2023. Mimeo.

Currie, J., J. Voorheis, and R. Walker, "What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality," *NBER Working Paper 26659*, 2020.

Di, Q., I. Kloog, P. Koutrakis, A. Lyapustin, Y. Wang, and J. Schwartz, "Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States," *Environmental Science & Technology*, 2016, *50* (9), 4712–4721.

Donkelaar, A. Van, R.V. Martin, M. Brauer, N.C. Hsu, R.A. Kahn, C. Levy, A. Lyapustin, A.M. Sayer, and D.M. Winker, "Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors," *Environmental Science & Technology*, 2016, *50* (7).

Finlay, K. and K. Genadek, "Measuring All-Cause Mortality With the Census Numident File," *American journal of public health*, 2021, *111.*

Lang, M. W., J. C. Ingebritsen, and R. K. Griffin, *Status and Trends of Wetlands in the Conterminous United States 2009 to 2019*, Washington, D.C.: U.S. Department of the Interior; Fish and Wildlife Service, 2024.

Shapiro, J., "Pollution Trends and US Environmental Policy: Lessons from the Past Half Century," *Review of Environmental Economics and Policy*, 2022, *16* (42–61).

USGCRP, *Fifth National Climate Assessment*, Washington, DC, USA: U.S. Global Change Research Program, 2023.

van Donkelaar, A., M. Hammer, L. Bindle, M. Brauer, J. Brook, M. Garay, C. Hsu, O. Kalashnikova, R. Kahn, C. Lee, R. Levy, A. Lyapustin, A. Sayer, and R. Martin, "Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty," *Environmental Science & Technology*, 2021, *55* (22).

Wagner, D. and M. Layne, "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications (CARRA) Record Linkage Software," *Mimeo*, 2014.