The Privacy-Protected Gridded Environmental Impacts Frame*

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

July 16, 2025

Abstract

This paper introduces the Gridded Environmental Impacts Frame (Gridded EIF), a novel privacy-protected dataset derived from the U.S. Census Bureau's confidential Environmental Impacts Frame (EIF) microdata infrastructure. The EIF combines comprehensive administrative records and survey data on the U.S. population with high-resolution geospatial information on environmental conditions. While access to the EIF is restricted due to the confidential nature of the underlying data, the Gridded EIF offers a broader research community the opportunity to glean insights from the data while preserving confidentiality. We describe the data and privacy protection methods, and offer guidance on appropriate usage, presenting practical applications.

*Corresponding Author: John Voorheis, john.l.voorheis@census.gov. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Data Management System (DMS) number: P-7505723, Disclosure Review Board (DRB) approval numbers: CBDRB-FY24-0394 and CBDRB-FY24-CES023-019). 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 and Center for Real Estate and the Built Environment for generous funding and support, which has contributed to the development of the underlying Environmental Impacts Frame microdata.

†Affiliations: Voorheis, Houghton, Lyubich and Withrow: Center for Economic Studies, U.S. Census Bureau; Scalera: Columbia University; Colmer: University of Virginia; Munro: MITRE Corp.

1 Introduction

Advances in the availability of new microdata infrastructures, such as the Census Environmental Impacts Frame (EIF), provide unprecedented opportunities for researchers to expand our understanding of how environmental conditions are distributed across individuals with diverse demographic and economic characteristics, and to study the causes and consequences of environmental exposures (Voorheis et al., 2025). However, due to the highly confidential nature of the EIF, access is only available to researchers who pass necessary background checks, and whose projects have undergone a thorough approval processes. These barriers to entry unavoidably slow the research process and, consequently, delay the generation of new insights.

To address this challenge, we introduce the Gridded Environmental Impacts Frame—a publicly accessible data product which aggregates the EIF microdata in ways that preserve privacy while maintaining analytical consistency. The Gridded EIF expands access to institutions and researchers who would otherwise be unable to obtain approval to access the EIF . The data is organized on a fixed 0.01-degree grid (approximately 1km^2 in North America) and provides counts of the population by age, sex, and race and ethnicity as well as counts of the population by race and ethnicity and household income decile.

Providing the data in this form allows researchers to take advantage of certain features of the EIF, offering substantial benefits over traditional analyses of aggregated place-based data. First, the data enables distributional analysis by intersectional characteristics, such as by race *and* income. Second, the underlying administrative records are updated more frequently than most aggregate demographic data, making the gridded EIF more timely. For instance, as of May 2024, roughly two thirds of the U.S. population has appeared on a 1040 for tax year 2023 (filed in early 2024), and close to the entire population is covered in the 2023 vintage of the EIF's residential history file. This is in contrast to the most recently available ACS 5-year tables, which draw from survey responses as far back as 2018. Third, most individuals in the EIF can be geocoded to precise latitude and longitude, allowing for flexible aggregation to any geographic unit, not limited to just those that are defined by the Census Bureau (such as Census tracts). This is particularly important when analyzing hazards that do not align with administrative boundaries—in these cases, aggregating to units of fixed size (such as geographic grid) as opposed to fixed population (such as Census tracts) may enhance analysis. The Gridded EIF thus extends the analytical advantages afforded by intersectional information, fixed geographic structure, and high-frequency temporal coverage to a broader set of researchers and stakeholders.

This paper lays out the process for building the Gridded EIF, discusses the privacy-protection processes that we use, illustrates the consequences of these privacy-protection processes for inference in the data by drawing comparisons with the confidential microdata, and provides an illustrations of how the data could be used.

2 Building the Gridded Environmental Impacts Frame

The Gridded EIF is an aggregation of the EIF microdata Voorheis et al. (2025). The EIF microdata is built as a modular infrastructure, allowing researchers to construct custom datasets suited to each specific type of analysis. The three key components of the EIF are: 1) a residential history file (RHF) tracking individual residential locations for nearly the entire U.S. population[1] from 1999 to 2023, 2) a demographic spine containing basic demographic information (race, ethnicity, sex, place of birth, date of birth and mortality) for all individuals with social security numbers, sourced from the Social Security Administration's Numident file, and 3) an environmental hazards module, which harmonizes geospatial information on environmental hazards such as air pollution and wildfires, allowing for an easy link to the residential locations in the RHF. The EIF's modular design facilitates the integration of additional data sources, such as the Opportunity Databank, consisting of harmonized administrative tax data, or commercially provided housing data available in the Census

[1]Because the EIF relies on administrative records, it may not always fully capture all subpopulations, meaning that population totals in both the EIF microdata and the gridded EIF may differ from population totals in the Decennial Census or Census Bureau population estimates.

Bureau, creating a breadth of research opportunities.

The Gridded EIF is constructed directly from the demographic spine and residential history modules of the EIF, which we combine with income information from the Opportunity Databank. Adjusted Gross Income (AGI) from IRS 1040s is aggregated to the household (Master Address File ID, MAFID) level using similar rules as those used in Kondo et al. (2024).[2] The Gridded EIF starts from this merged dataset in each year and aggregates the data in two ways. First, all geocoded individuals are assigned to a grid point on a fixed (unprojected) 0.01 degree grid, or about 1 km^2 in North America. The merged microdata is then collapsed by race, sex, and age groups (under 18, 18-64, and 65+); and by race/ethnicity and income deciles, within these grid points. Consistent with the EIF, we define harmonized mutually exclusive categories for race and ethnicity: Hispanic, non-Hispanic White, Black, Asian, and American Indian and Alaskan Native, and a residual group of individuals without race information or who are some other race or multiracial. Individuals who cannot be matched to the demographic spine (e.g., individuals without social security numbers) will not have age and sex information; the race-by-age-by-sex aggregations retain counts of these individuals as a separate group in each grid point. The grid system used is essentially identical in the North American domain to the widely used unprojected grid systems in atmospheric sciences, for instance the satellite-derived PM2.5 data from van Donkelaar et al. (2021), facilitating the harmonization of external environmental data sources at the level of variation in socioeconomic characteristics.

2.1 Noise Injection Process

The grid points used represent very small geographies, which means it is not possible to publish these raw tabulations under current Census Bureau disclosure avoidance guidelines

[2] For MAFIDs with fewer than 10 unique tax units, the household income is calculated as the sum of all unique tax-unit level AGI values in that MAFID. For MAFIDs with 10 or more unique tax units, we assume, consistent with the evidence in Kondo et al. (2024), that this is a data quality issue involving improper MAFID assignment in multiunit structures and assign the average tax unit level AGI as the household income for that MAFID.

without additional privacy protection. To protect the privacy of individuals in the underlying microdata, we use a noise infusion strategy inspired by the differential privacy methods used in the 2020 Decennial Census(Abowd et al., 2022). The noise injection method proceeds as follows. First, we inject mean zero, discrete Gaussian noise with standard deviation $\sigma = \frac{\gamma}{2\times\rho}$, where γ can be interpreted as the squared sensitivity of the statistics to be released and ρ is a parameter capturing the privacy loss budget, a quantification of the level of protection against reidentification provided by the noise injection (Dwork and Roth, 2014).[3] In a differential privacy framework, these two parameters together describe the privacy guarantee of the noise infusion method (Canonne et al., 2020).

We select values for γ and ρ in consultation with the Census Bureau's Disclosure Review Board. The resulting discrete Gaussian distribution leads most drawn values to fall between -3 and 3.[4]. This noise can be thought of as "on the order of rounding" as it represents a similar amount of coarsening as the rounding schemes used for official Census tabulations of the American Community Survey and other collections. This noise is only infused into demographic counts that actually exist in the underlying microdata for each grid point. If a particular race-by-income group or race-by-sex-by-age group does not exist in the underlying microdata, the count of these individuals is treated as structurally zero. This approach to noise injection and the selection of ρ and γ parameters differs slightly from strict differential privacy requirements, which would tune the parameters based on a measure of global sensitivity. Thus, our method is inspired by differential privacy methods, as in the methods used in Chetty and Friedman (2019), but lacks a formal privacy guarantee. The Census Bureau has judged that this lack of a formal guarantee is outweighed by the ability to release data products (such as the gridded EIF) which otherwise might not be feasible to release with a formal guarantee, and has judged the amount of privacy protection produced by the noise injection to meet the statutory guidelines of confidentiality under Title 13 of the U.S. Code.

[3]In a formal privacy setting, "sensitivity" represents the maximum amount that a single observation can affect a given statistic or set of statistics calculated from a given set of microdata.

[4]The parameters selected are $\rho \approx 0.33$ and $\gamma = 2$

The resulting datasets from this process consist of noisy integer counts by race-age-sex and race-income categories for each grid point. However, the discrete Gaussian noise can generate negative values that sometimes exceed the original counts in magnitude, creating negative cell values in the final data set. While this noise typically nets out when aggregating grid points to larger geographies, yielding non-negative estimates consistent with the underlying microdata, researchers may prefer strictly non-negative demographic counts at the grid level. To address this need, we implement a post-processing algorithm that exploits the nested structure of grids and the tendency for demographics and environmental hazards to be positively correlated within small geographic areas.

The post-processing algorithm proceeds as follows. First, within each 1-degree grid point (approximately 100km^2, or the size of Orange County, CA), we identify small noisy cells – those with raw noisy counts smaller than the absolute value of the most negative count in that grid point. For example, if the most negative value is -5, all cells with counts below 5 are classified as small noisy cells. Second, we calculate the sum of all small noisy cells by race group within each 1-degree grid, and replace small noisy cell values with this sum divided by the number of small noisy cells in that race-grid combination. This redistribution preserves ratio population distributions while forcing most cells to be non-negative. Third, we set any remaining negative values to zero. By redistributing small populations within larger areas, this algorithm better captures racial differences in environmental exposures for small racial groups than the raw noisy measures, provided that environmental hazards exhibit similar spatial correlation. We provide both raw noisy counts and post-processed counts, allowing users to apply alternative post-processing algorithms if they wish.

2.2 Demographic Data

The gridded EIF population data alone offers opportunities to explore the changing demographics of the United States. By providing harmonized and consistent measures of population by race and income, it enables researchers to explore demographic shifts at temporal and geographic scales previously unavailable in public use data products. Figure 1 illustrates the power of this approach, by mapping the Black population share within grid points in the Detroit Combined Statistical Area in 1999 and 2022. The 1999 map reproduces the well-documented segregation of the metro area, with the predominantly Black city of Detroit sharply abutting majority white suburbs, especially to the north on the Eight Mile Rd. border. The gridded EIF's annual data allow for nuanced analysis of how this demographic pattern has evolved over two decades. By 2022, the map shows substantial Black suburbanization in the Detroit metro area, with a marked reduction in the stark segregation between Detroit and its suburbs that characterized the region in 1999.

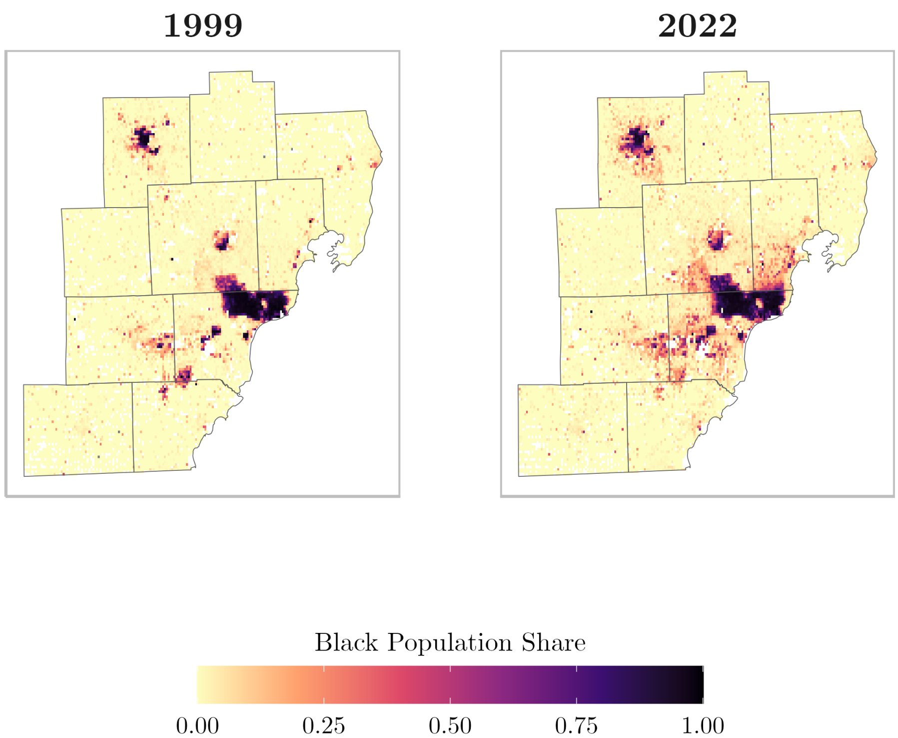
2.3 Incorporating Spatially-Resolved Environmental Data

To facilitate research on the distribution of exposure to environmental hazards, we have harmonized three widely used geospatial air pollution datasets to the same grid as the privacy-protected demographic data described above. These data are: satellite derived PM2.5 from van Donkelaar et al. (2021) and Shen et al. (2024) provided by the Atmospheric Composition Analysis Group (ACAG), spanning 1999-2022; satellite-derived PM2.5, NO2 and Ozone data from Requia et al. (2020), Di et al. (2020) and Di et al. (2019), provided by NASA's Socioeconomic Data and Applications Center (SEDAC), spanning 2000-2016; and Land-use regression based estimates of PM2.5, PM10, SO2, NO2, CO and Ozone from Kim et al. (2020), made available by the Center for Air, Climate, and Energy Solutions (CACES), spanning 1999-2020.

To demonstrate the utility and quality of inference of the Gridded EIF at various geographic scales, we reproduce the core descriptive results presented in Colmer et al. (2024). Figure 2 presents race-specific income–PM2.5 exposure gradients for 2022 using the gridded EIF and the ACAG PM2.5 data.[5] We calculate these profiles using both the raw and post-processed values from each race-income-grid point. For all race groups the two noisy

[5]We use the V5.GL.04 version of the ACAG data, which is referred to as *pm25_ACAG_gwr* in the gridded EIF data.

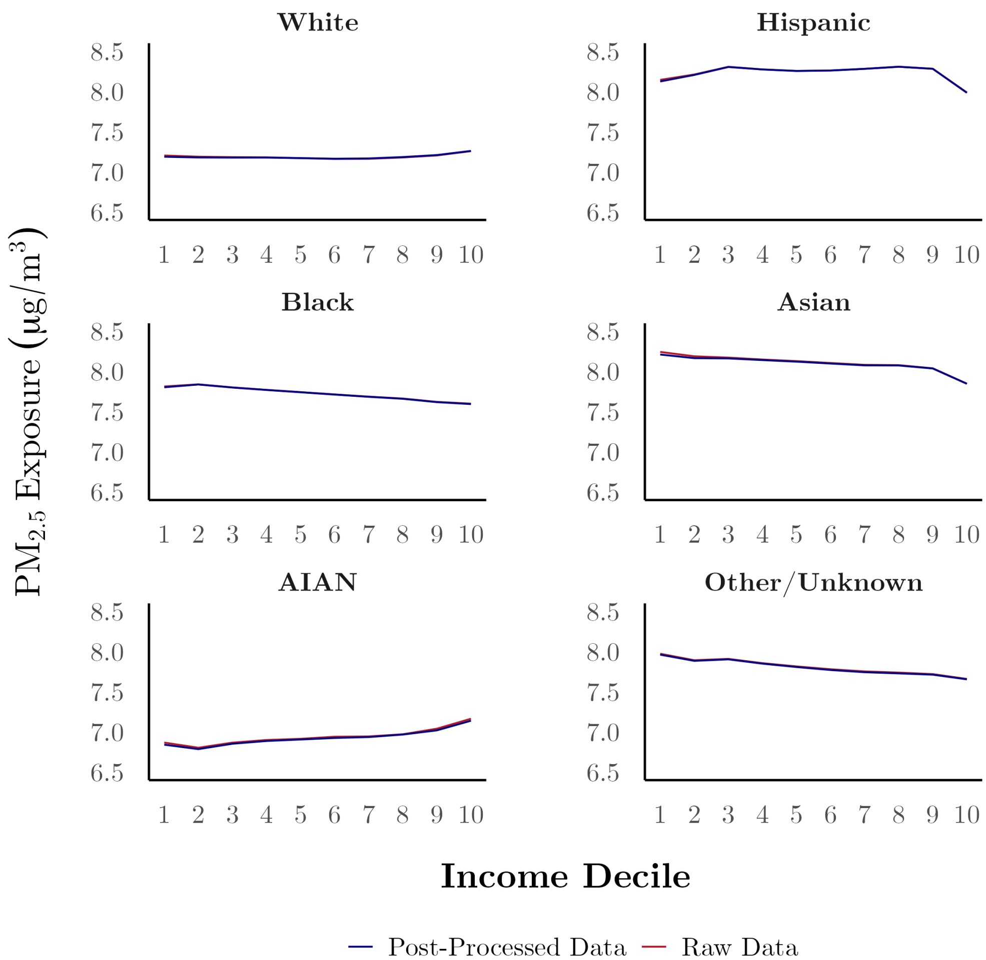
Figure 1: Black Population Share, Detroit Combined Statistical Area



Source: Gridded Environmental Impacts Frame. These maps show the Non-Hispanic Black population within each 0.01 degree grid point as a share of all individuals with non-missing race information, in 1999 and 2022. Opacity of each grid point is based on grid point total population, with lower population cells being more transparent.

measures produce very similar results. Figure 3 replicates this exercise for the Mankato, MN CZ, a typical small metro area with a population of around 100,000. At this smaller geographic scale, differences between the two noisy measures are more apparent. In particular, the raw noisy measure exhibits greater variability than the post-processed measure, though absolute differences remain small.

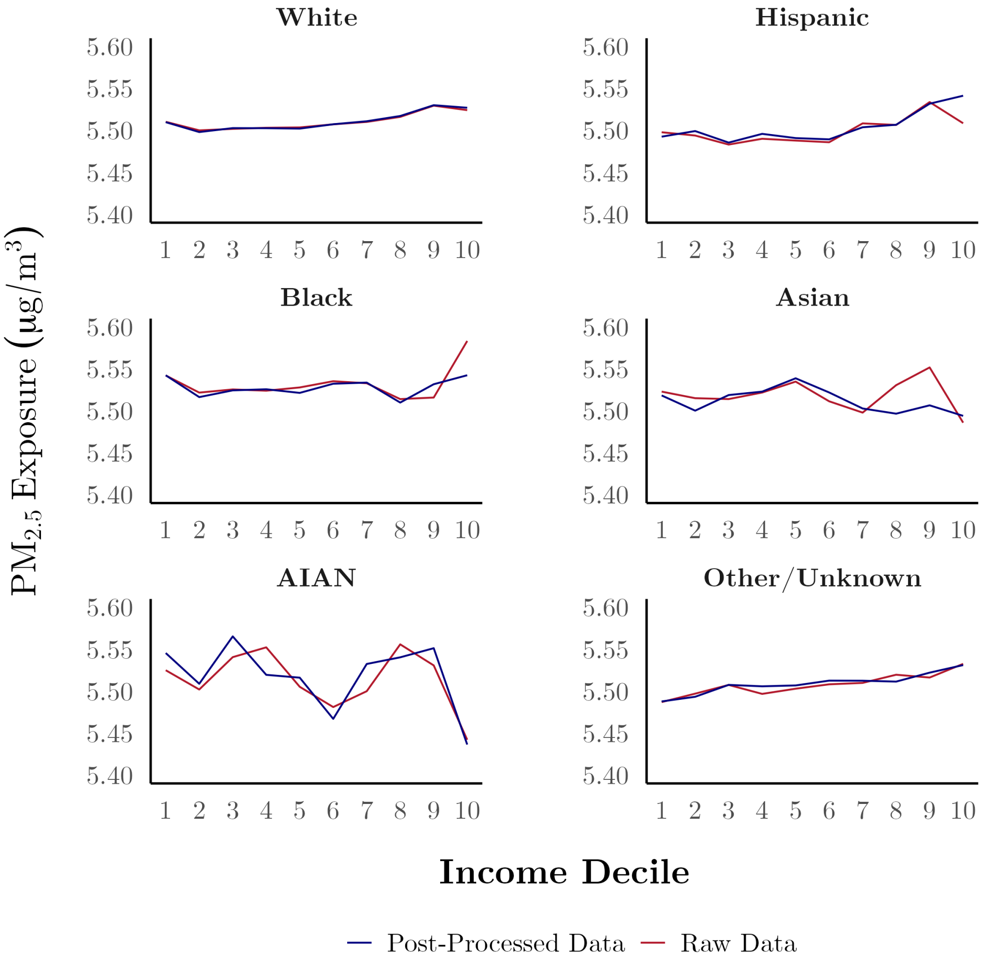
Figure 2: Race by Income PM2.5 Gradients, National



Source: Gridded Environmental Impacts Frame. This graph show the average PM2.5 exposure by race and income decile, for the nation as a whole, using the two noisy measures in the gridded EIF.

Extending this analysis across all CZs, we find that raw noisy measures produce mean absolute errors roughly 65% larger than post-processed measures when compared to the con-

Figure 3: Race by Income PM2.5 Gradients, Mankato, MN



Source: Gridded Environmental Impacts Frame. This graph show the average PM2.5 exposure by race and income decile, for Mankato, MN, using the two noisy measures in the gridded EIF.

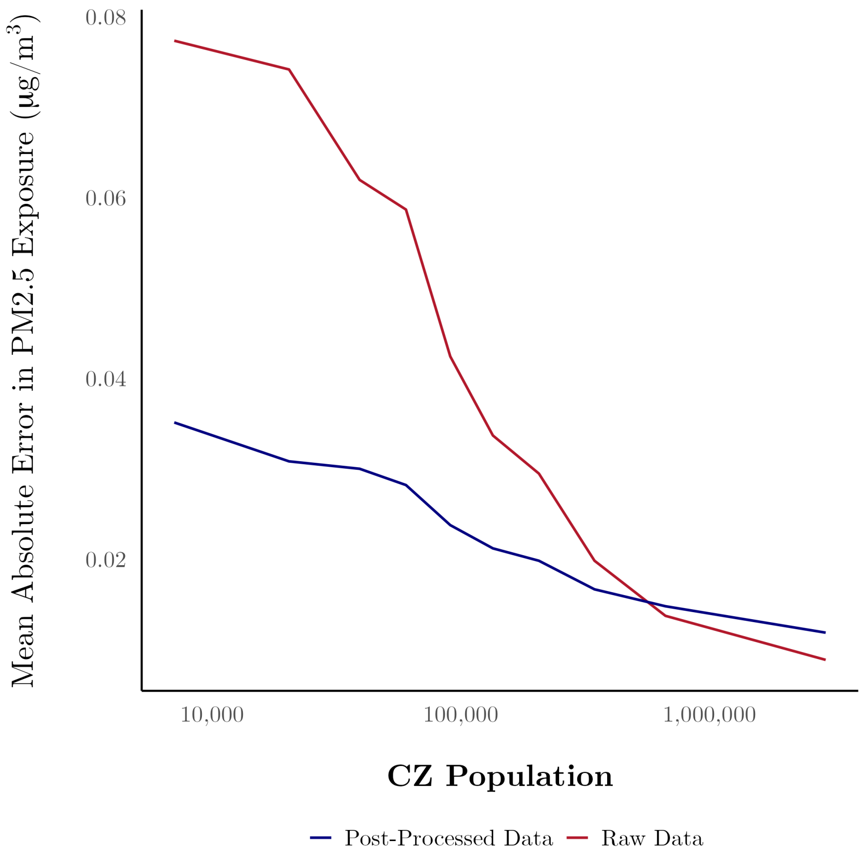
fidential microdata. This varies systematically by CZ size, as shown in Figure 4. For the smallest CZs (bottom two quintiles by population), the raw noisy measures exhibit mean absolute errors more than double those of the post-processing measures, while for larger CZs (the top 10% by population), the post-processing noisy measures produce errors about 25% larger than raw measures. However, mean absolute errors are small across the distribution. The crossover between the two error curves occurs near the population threshold where Census Bureau disclosure avoidance rules permit direct estimates from the confidential microdata (the population of the smallest state, Wyoming, which was around 600,000 in 2023).

3 "Real-Time" Analysis

In addition to the standard gridded EIF dataset produced using the 1999-2023 microdata files for the EIF, we also create a "real time" version of the EIF for 2024. This real time file starts from the most recent EIF residential histories, and then updates residential locations for all individuals who have filed their tax year 2023 1040 tax return through the spring of the current (2024 when this first version of the gridded EIF was created). This provides the most up-to-date information on individual's locations available in the administrative records underlying the EIF.

This version of the EIF enables near real-time analysis of populations exposed to environmental hazards. This could be particularly useful in the context of short term forecasting tools. The National Weather Service and other agencies have increasingly accurate forecasting capabilities for a number of weather and air quality related hazards. By combining these forecases with gridded demographic data, researchers and policymakers can characterize exposed populations by vulnerability indicators derived from age, income, and race, facilitating more targeted emergency preparedness and response strategies.

Figure 4: Mean Absolute Error in PM2.5 Exposure Estimates, Gridded EIF vs. Confidential Microdata



Source: Gridded Environmental Impacts Frame. This graph show the Mean Absolute Error between the actual CZ-level average PM2.5 exposure and the average PM2.5 exposure derived from the noisy measures in the gridded EIF, by CZ population.

Figure 5: The spatial distribution of hurricane-force wind speed probabilities, Hurricane Milton.

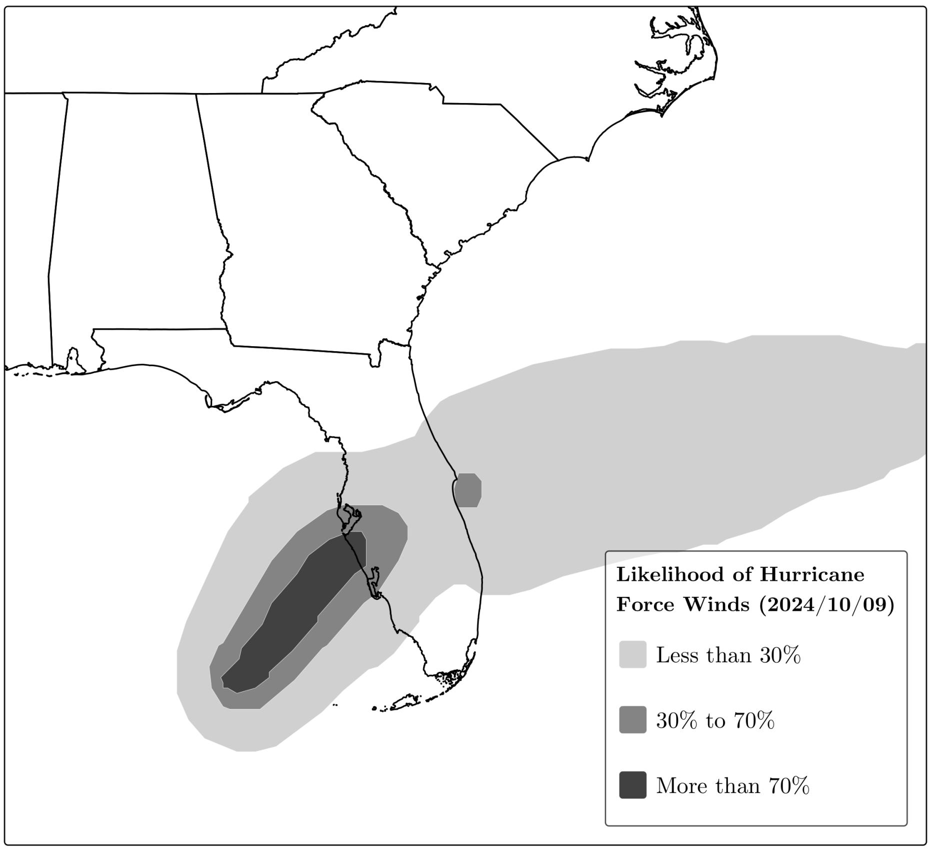


Figure 6: Hurricane Force Wind Probabilities

Notes: This figure presents hurricane-force wind speed probabilities for Hurricane Milton as provided by the National Hurricane Center (NHC) on October 9, 2024, at 11:00AM.

This capability was utilized in advance of Hurricane Milton (U.S. Census Bureau, 2024a). We combined the gridded EIF with the forecast products that the National Hurricane Center produced immediately before Hurricane Milton made landfall in Florida on October 9th. These forecasts include predicted probabilities of hurricane-force and tropical-storm-force winds, as well as the predicted probability of flash flooding. Figure 6 shows the extent of the forecasts for hurricane force winds as of 11:00am on the day Milton made landfall.

We intersected the version of these forecast products produced at 11:00am ET on the day Milton made landfall with the gridded EIF to explore the characteristics of the population that was forecast to be exposed to environmental hazards associated with Hurricane Milton. Table 1 summarizes the demographic characteristics of the population that was forecasted to be exposed to hurricane force winds, compared to the demographics of the overall real-time gridded EIF for 2024. The population forecasted to be exposed to hurricane force winds was substantially older and more likely to be non-Hispanic White relative to the overall U.S. population. The data's granular structure also enabled analysis of race-by-income decile exposure patterns, as detailed in U.S. Census Bureau (2024a). Following landfall, we intersected the Gridded EIF with observed hazard data to identify actually exposed populations, then compared pre- and post- landfall results to characterize populations exposed to forecast errors. The complete analysis can be found in U.S. Census Bureau (2024b).

Pre- and post-landfall analysis of hurricanes is only one such application of the gridded EIF for real time analysis purposes. There are several other applications under active development. Geospatial data and forecast products for a much broader set of discrete extreme weather events beyond exposure to hurricane wind fields can be combined with the gridded EIF to explore a much broader array of hazards. This includes extreme precipitation events, coastal flooding due to storm surge, riverine flooding, extreme heat events, tornadoes and other convective storms. Finally, by combining the gridded EIF with the National Weather Service's Air Quality forecast product and real-time pollution monitoring data, it is possible to track the characteristics of the population exposed to hourly or daily fluctuations in air quality, analysis that could be embedded in a continuous monitoring tool.

Table 1: Socioeconomic characteristics of the population forecasted to experience hurricane-force winds (70 percent or more probability) compared to the national population.

	Population	Percent	National percent	Difference
Non-Hispanic White	547,242	70.71	55.03	15.68
Non-Hispanic Black	44,713	5.78	11.60	-5.82
Hispanic	84,998	10.98	15.80	-4.82
Non-Hispanic Asian	12,814	1.66	4.13	-2.47
Non-Hispanic AIAN	1,678	0.22	0.83	-0.61
Other	82,515	10.66	12.61	-1.95
Over the Age of 65	244,158	31.55	18.26	13.29
Under the Age of 18	109,691	14.17	19.69	-5.52
Lower Income	154,601	19.95	20.02	-0.07
Total Population Exposed	773,960			

Notes: This table presents the sociodemographic characteristics of the populations forecasted to be exposed to hurricane-force winds with probability greater than 70 percent, using probabilities provided by the National Hurricane Center (NHC) on October 9, 2024, at 11:00AM. Column 1 provides population counts. Column 2 provides percentages of the total exposed population. Column 3 reports each group's share of the national population. Column 4 reports the difference in group shares. Source: Gridded Environmental Impacts Frame and NHC data.

4 Access to the Gridded EIF

The Gridded EIF is available to the public as an ongoing Census experimental data product, facilitating exploration of the distribution of exposure to environmental hazards. These files are available via the Census Bureau's website[6]. For each year 1999–2023, race-age-sex and race-income datasets are made available, indexed by grid point centroid. As subsequent years of the EIF and the gridded EIF aggregates are generated and approved for public release, these will be made available on this site. These files are distributed in Parquet format due to the large size of the gridded datasets. Additionally, the harmonized pollution data described above are made available for the years 1999-2023, again indexed by grid point centroid, and will be likewise updated as new data becomes available. To conduct geospatial analyses of other environmental hazards, or indeed any spatial variation, we will also distribute a grid topology file in R data format that provides a mapping to the gridded EIF. Example code to combine the gridded EIF files additional documentation are also available for potential users.

The Gridded EIF provides an easily accessible way for researchers to engage with many of the patterns and trends available in the restricted use microdata, without the access restrictions and disclosure constraints that come with the use of confidential Census Bureau data. As needs and uses for this gridded data evolve, the Census Bureau will continue to update and enhance the information available.

[6]The data and associated documentation are available here: https://www.census.gov/data/experimental-data-products/gridded-eif.html.

References

Abowd, John, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev, "The 2020 Census Disclosure Avoidance System TopDown Algorithm," *Census Working Paper Series*, 2022.

Canonne, Clément L., Gautam Kamath, and Thomas Steinke, "The Discrete Gaussian for Differential Privacy," *CoRR*, 2020, *abs/2004.00010.*

Chetty, Raj and John Friedman, "A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Samples," *NBER Working Paper Series*, 2019.

Colmer, Jonathan, Suvy Qin, John Voorheis, and Reed Walker, "Income, Wealth, and Environmental Inequality in the United States," *Working Paper*, 2024.

Di, Qian, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M Benjamin Sabath, Christine Choirat, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, Loretta J Mickley, and Joel Schwartz, "An Ensemble-based Model of PM2.5 Concentration Across the Contiguous United States with High Spatiotemporal Resolution," *Environment International*, 2019, *130.*

_ , _ , _ , _ , _ , _ , M. Benjamin Sabath, Christine Choirat, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, Loretta J. Mickley, and Joel Schwartz, "Assessing NO2 Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States," *Environmental Science Technology*, 2020, *54.*

Dwork, Cynthia and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, 2014, *9* (3–4), 211–407.

Kim, Sun-Young, Matthew Bechle, Steve Hankey, Lianne Sheppard, Adam A. Szpiro, and Julian D. Marshall, "Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression," *PLoS ONE*, 2020, *15.*

Kondo, Illenin, Kevin Rinz, Natalie Gubbay, Brandon Hawkins, John Voorheis, and Abigail Wozniak, "Granular Income Inequality and Mobility Using IDDA: Exploring Patterns across Race and Ethnicity," *NBER Working Paper*, 2024.

Requia, Weeberb, Qian Di, Rachel Silvern, James T. Kelly, Petros Koutrakis, Loretta J. Mickley, Melissa P. Sulprizio, Heresh Amini, Liuhua Shi, and Joel Schwartz, "An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-level Ozone in the Contiguous U.S.," *Environmental Science Technology*, 2020, *54.*

Shen, Siyuan, Chi Li, Aaron van Donkelaar, Nathan Jacobs, Chenguang Wang, and Randall V. Martin, "Enhancing Global Estimation of Fine Particulate Matter

Concentrations by Including Geophysical a Priori Information in Deep Learning," *ACS EST Air*, 2024, *1.*

U.S. Census Bureau, "Real-Time Analysis: Hurricane Milton," *US Census Bureau Technical Report*, October 9th 2024.

_ , "Real-Time Analysis: Hurricane Milton, Post-Landfall Report," *US Census Bureau Technical Report*, October 28th 2024.

van Donkelaar, Aaron, Melanie S. Hammer, Liam Bindle, Michael Brauer, Jeffery R. Brook, Michael J. Garay, N. Christina Hsu, Olga V. Kalashnikova, Ralph A. Kahn, Colin Lee, Robert C. Levy, Alexei Lyapustin, Andrew M. Sayer, and Randall V. Martin, "Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty," *Environmental Science Technology*, 2021, *55.*

Voorheis, John, Jonathan Colmer, Kendall Houghton, Eva Lyubich, Mary Munro, Cameron Scalera, and Jennifer Withrow, "The Census Environmental Impacts Frame," *Working Paper*, 2025.