Exposome Data for the Health and Retirement Study International Network of Studies

User Guide, v1.0

Environmental Exposome

Jennifer D'Souza, Adam Taggart, Sara Adar

19-January-2024

Funding: R01AG030153

Table of Contents

EXECUTIVE SUMMARY AND OVERVIEW

The Gateway to Global Aging platform has provided a wealth of data across several countries that is harmonized such that cross-national comparisons can be made. This project aims to add *global harmonized environmental exposome data* to those Health and Retirement Study's International Network of Studies (HRS-INS) included in the Gateway. These data are intended to become accessible to researchers around the world to inform important scientific questions within and between countries.

This document provides user guidance and documentation for:

The environmental data included for harmonization. This includes the scientific motivation for its inclusion in the project as well as a summary of how the data was collected, processed, and links to the original source information. We also include meta data describing the coordinate reference system, variable names, and important notes for use. Finally, we include an example methods section and associated references for each parameter.

Assignment of environmental exposures to survey respondents. We provide guidance for exposure assessment to harmonize spatial and temporal averaging times for each environmental parameter across surveys. Code is also provided to assist in the assessment of exposures for each respondent using the cleaned data files provided by the Environmental Exposome Core at the University of Michigan School of Public Health.

Environmental exposome datasets that are ready for linkage with the surveys are located on a Google Drive hosted by the University of Michigan School of Public Health. Please contact gatewayexposome@umich.edu for access.

1.0 ENVIRONMENTAL EXPOSOME MEASURES

1.1 Overview of Environmental Measures

Measures of the environmental exposome currently available for HRS-INS surveys in the Gateway to Global Aging are outlined in Table 1 and described in detail in the following sections. Briefly, we have available the following data:

Outdoor air pollution

- Total fine particulate matter less than 2.5 microns in diameter ($PM_{2.5}$)
- $PM_{2.5}$ from specific emission sources such as energy generation from coal, traffic, indoor biomass burning, and agriculture
- Nitrogen dioxide (NO_2)
- Ozone (O_3)

Natural spaces

- Greenspace
- Bluespace

Physical stressors

- Nighttime light

Table 1. Summary of Spatial and Temporal Resolutions of Environmental Exposome Data

Exposure	Temporal Resolution	Approximate Spatial Resolution	Years
Total $PM_{2.5}$ Mass	annual average	0.0083º (1km^2)	2010-2019
Total $PM_{2.5}$ Mass	monthly	0.1º (~1km^2)	1998-2022
$PM_{2.5}$ Source Fraction	annual average	0.5º x 0.625º (~55km x ~70km) in North America, Europe, and East Asia; 2º x 2.5º (~225km x ~275km) elsewhere	2017
NO_2	annual average	0.0083º (~1km^2)	1990, 1995, 2000, 2005-2020
O_3	annual average	0.1º (~11km^2)	1990-2017
Greenspace	annual maximum annual minimum annual mean	250 meters (0.06 km^2)	2001-2021
Bluespace	static	150 meters (0.02 km^2)	2000-12
Nighttime Light	annual median	0.0042º (~500m^2)	2012-2021

1.2 Total $PM_{2.5}$ Concentrations in Outdoor Air

Description: Particulate matter of aerodynamic diameter less than 2.5 microns ($PM_{2.5}$) are extremely small particles that can be found in air. They are generated from a wide variety of sources ranging from natural emissions of windblown dust to anthropogenic emissions from fossil fuel combustion, traffic, industry, and fires. $PM_{2.5}$ is of interest for human health as their inhalation can initiate inflammation, oxidative stress, vascular changes, and autonomic imbalance that can ultimately impact health.

Spatiotemporal estimates of $PM_{2.5}$ concentrations ($\mu g/m^3$) have been generated for all locations at a monthly resolution on the globe between 1998 and 2022. These data, which are available at a 0.01^o x 0.01^o resolution, were generated using a fusion of satellite data (MODIS, VIIRS, MISR, and SeaWiFS), chemical transport model (GEOS-Chem), ground-based sun photometer observation (AERONET), ground-based monitoring data, and local characteristics of place. These models produced by van Donkelaar and colleagues combine these different data sources using a Geographically Weighted Regression (GWR).

These monthly estimates supplement the spatiotemporal annual mean concentrations of total $PM_{2.5}$ that are available at a $1km^2$ resolution from 2010-2019. These data were created using the Data Integration Model for Air Quality (DIMAQ) originally for the World Health Organization's Global Burden of Disease project and later refined for the Gateway to Global Aging project. These estimates were developed using a Bayesian hierarchical prediction model that leveraged satellite measurements of aerosol optical depth in the atmosphere, ground-level measures of pollution, chemical transport models, and correlations over space.

Citations:

van Donkelaar A, Hammer MS, Bindle L, Brauer M, Brook JR, Garay MJ, Hsu NC, Kalashnikova OV, Kahn RA, Lee C, Levy RC, Lyapustin A, Sayer AM, Martin RV. Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. Environmental Science & Technology, 2021, 55, 22, 15287-15300. doi:10.1021/acs.est.1c05309.

Shaddick G, Thomas ML, Green A, Brauer M, von Donkelaar A, Burnett R, Chang HH, Cohen A, Van Dingenen R, Dora C, Gumy S, Liu Y, Martin R, Waller LA, West J, Zidek JV, Pruss-Ustun A. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. Applied Statistics, 2018, 67(1), 231-253. https://doi.org/10.1111/rssc.12227

Original Source (monthly 1998-2022): https://sites.wustl.edu/acag/datasets/surface-pm2-5/#V5.GL.04
File names: Country_PM2.5_year_month

Original Source (annual, 2010-2019): Not publicly available
File names: PM2.5_DIMAQ_year

Coordinate Reference System: WGS84
Data type: Raster

Data type: Numeric
Units: $\mu g/m^3$

User Notes: The primary exposure estimates for $PM_{2.5}$ for the Gateway projects are concentrations at participant addresses derived from the monthly data and averaged over the 10 years before each survey visit (incorporating residential history as possible). Secondary exposure estimates of interest are the 1 and 5-year averages, averages over decades of life at the residential addresses, and estimates derived from the annual data.

Example Methods Section
"We estimated *10-year* average outdoor fine particulate matter ($PM_{2.5}$) concentrations for each participant based on their residential histories preceding each of their survey visits using the V5.GL.04 version of a global spatiotemporal model developed by van Donkelaar and colleagues. (van Donkelaar et al. 2020)

Briefly, this model integrates data from ground-level measurement stations, satellites, chemical transport and meteorological models, and area-level characteristics to estimate outdoor concentrations of $PM_{2.5}$ at a 0.01^o x 0.01^o resolution. This geographically weighted regression has excellent agreement with ground-level monitors (cross-validated R^2=0.90) and allows for estimation of concentrations at participant addresses even where ground-level monitoring is sparse or even non-existent."

1.3 Source-Specific $PM_{2.5}$ Concentrations in Outdoor Air

Description: Particulate matter of aerodynamic diameter less than 2.5 microns ($PM_{2.5}$) are generated by a variety of sources ranging from natural emissions of windblown dust to anthropogenic emissions from fossil fuel combustion, traffic, industry, and fires. The originating source has importance both for interventions but also for health since the emission source can alter the physical and chemical properties of the particles.

Spatial estimates of the fraction of $PM_{2.5}$ concentrations attributable to different sources have been predicted for all locations on the globe at a resolution of 2^o x 2.5^o with finer resolutions of 0.5^o x 0.625^o over North America, Europe, and Asia. These models were derived by serially running the GEOS Chem atmospheric chemical-transport model with all emission sources but one to isolate the unique contribution of that source to the total mixture of $PM_{2.5}$. Emissions data was derived from 2017 but assumed to be representative of other years.

Citation: McDuffie EE, Martin RV, Spadaro JV, Burnett R, Smith, SJ, O'Rourke P, Hammer M, van Donkelaar A, Bindle L, Shah V, Jaegle L, Luo G, Yu F, Adeniran J, Lin J, Brauer M. Source Sector and Fuel Contributions to Ambient $PM_{2.5}$ Attributable Mortality Across Multiple Spatial Scales, Nature Communications, 2021, 12:3594. doi: 10.1038/s41467-021-23853-y

Original Source: https://zenodo.org/record/4739100#.Y1AXlXbMLcs
Coordinate Reference System: WGS84
Data type: Raster

File names: GBD_raster_country_source or fuel category (see lists below)
Data type: Numeric
Units: Fractional percentages (e.g., 0.0245 is 2.45%)

List of Sources with Descriptions of Primary Emissions:

AFCID	*Anthropogenic fugitive, combustion, and industrial dust*
AGR	*Agriculture* - manure management, soil fertilizer emissions, rice cultivation, enteric fermentation, and other agricultural activities
ENEcoal	*Energy Production (coal combustion only)* - electricity and heat production, fuel production and transformation, oil and gas fugitive/flaring, and fossil fuel fires
ENEother	*Energy Production (all non-coal combustion)* - electricity and heat production, fuel production and transformation, oil and gas fugitive/flaring, and fossil fuel fires
GFEDagburn	*Agricultural Waste Burning* - solid waste disposal, waste incineration, waste-water handling, and other waste handling (from the GFED fires inventory)
GFEDoburn	*Other Open Fires* - deforestation, boreal forest, peat, savannah, and temperate forest fires (from the GFED fires inventory)
INDcoal	*Industry (coal combustion only)* – industrial combustion (iron and steel, non-ferrous metals, chemicals, pulp and paper, food and tobacco, non-metallic minerals, construction, transportation equipment, machinery, mining and quarrying, wood products, textile and leather, and other industry combustion) and non-combustion industrial processes and product use (cement production, lime production, other minerals, chemical industry, metal production, food, beverage, wood, pulp, and paper, and other non-combustion industrial emissions)
INDother	*Industry (all non-coal combustion)* – industrial combustion (iron and steel, non-ferrous metals, chemicals, pulp and paper, food and tobacco, non-metallic minerals, construction, transportation equipment, machinery, mining and quarrying, wood products, textile and leather, and other industry combustion) and non-combustion industrial

	processes and product use (cement production, lime production, other minerals, chemical industry, metal production, food, beverage, wood, pulp, and paper, and other non-combustion industrial emissions)
NRTR	*Non-Road/Off-Road Transportation* – rail, domestic navigation, other transportation
Other	*All Remaining Sources* - Includes volcanic SO2, lightning NOx, biogenic soil NO, ocean emissions, biogenic emissions, very short-lived iodine and bromine species, decaying plants (misc. inventories)
RCOC	*Commercial Combustion* – commercial and institutional combustion
RCOO	*Other Combustion* – combustion from agriculture, forestry, and fishing
RCORbiofuel	*Residential combustion (solid biofuel combustion only)* – residential heating and cooking
RCORcoal	*Residential combustion (coal combustion only)* – residential heating and cooking
RCORother	*Residential Combustion (all non-coal and non-solid biofuel)* –residential heating and cooking
ROAD	*Road Transportation* – cars, motorcycles, heavy and light duty trucks and buses
SHP	*International Shipping – international shipping and tanker loading*
SLV	Solvents - solvents production and application (degreasing and cleaning, paint application, chemical products manufacturing and processing, and other product use)
WDUST	*Windblown Dust*
WST	*Waste* – solid waste disposal, waste incineration, waste-water handling, and other waste handling

List of Fuel Categories with Primary Emissions (Note: These will not sum to 100% for each grid cell since they only include combustion sources of $PM_{2.5}$)

BIOFUEL	*Solid Biofuel (or Biomass) Combustion*- Solid biofuel
COAL	Total Coal Combustion – Hard coal, brown coal, coal coke
OILGAS	*Liquid Oil and Natural Gas Combustion* –light and heavy oil, diesel oil, and natural gas

User Notes: To calculate absolute contributions of $PM_{2.5}$ attributable to each source, location-specific total $PM_{2.5}$ (see above) must be multiplied by their location-specific fractional source contributions. These estimates of $PM_{2.5}$ attributable to each source averaged over the 10 years before each survey visit (incorporating residential history as possible) are the primary harmonized exposure estimates for the Gateway projects. Secondary exposure estimates of interest are 1 and 5-year averages and averages over decades of life at the residential addresses.

Example Methods Section:
"We estimated 1*0-year* average outdoor fine particulate matter ($PM_{2.5}$) concentrations for each participant based on their residential histories preceding each of their survey visits using the V5.GL.04 version of a global spatiotemporal model developed by van Donkelaar and colleagues. (van Donkelaar et al. 2020) Briefly, this model integrates data from ground-level measurement stations, satellites, chemical transport and meteorological models, and area-level characteristics to estimate outdoor concentrations of $PM_{2.5}$ at a 0.01^o x 0.01^o resolution. This geographically weighted regression has excellent agreement with ground-level monitors (cross-validated R^2=0.90) and allows for estimation of concentrations at participant addresses even where ground-level monitoring is sparse or even non-existent.

We further derived source-specific $PM_{2.5}$ concentrations by multiplying the total $PM_{2.5}$ at each address by the paired fraction of $PM_{2.5}$ attributable to *X* different emission sources. These fractions were derived by Mc Duffie and colleagues (McDuffie et al 2021) at a resolution of *X* (0.5^o x 0.625^o over North America, Europe, and Asia or 2^o x 2.5^o elsewhere) based on the serial exclusion of each source from the GEOS-Chem chemical transport model."

1.4 Nitrogen Dioxide NO_2 Concentrations in Outdoor Air

Nitrogen dioxide (NO_2) is a gas that can react to form $PM_{2.5}$ and ozone. It is commonly generated by the combustion of fuel, most notably in the transportation sector but also from power plants and industrial manufacturing. In urban areas, NO_2 is commonly used as an indicator of a mixture of pollutants generated by the transportation sector and traffic. It has been associated with adverse health effects in epidemiology studies.

Spatial estimates of NO_2 concentrations have been predicted for all locations on the globe at a resolution of 0.083^o every 5-years from 1990 to 2010 and then annually from 2010 to 2019 by Mohegh and Anenberg (Anenberg et al. 2022). These estimates were generated by extending an existing spatial prediction model (i.e., a land use regression model by Larkin and colleagues) to later years and refining this model for better performance in rural areas. The original spatial prediction model was derived from over 5000 monitors in nearly 60 countries (primarily in Europe, North America, and Asia) along with land use information and satellite data. Refinements to this model included further calibration using the Modern-Era Retrospective analysis for Research and Applications reanalysis product and additional satellite-based data.

Citation: Anenberg SC, Mohegh A, Goldberg DL, Kerr GH, Brauer M, Burkhart K, Hystad P, Larkin A, Wozniak S, Lamsal L. Long-term trends in urban NO_2 concentrations associated paediatric asthma incidence: estimates from global datasets. The Lancet Plantetary Health, 2022, 6(1): E49-E58. https://doi.org/10.1016/S2542-5196(21)00255-2

Related Citation: Larkin A, Geddes JA, Martin RV, Xiao Q, Liu Y, Marshall JD, Brauer M, Hystad P. Global land use regression model for nitrogen dioxide air pollution. Environmental Science & Technology, 2017, 51(12):6957-6964. doi: 10.1021/acs.est.7b01148.

Original Source: https://figshare.com/articles/dataset/Global_surface_NO2_concentration_1990-2020/12968114
Coordinate Reference System: WGS84
Data type: Raster

File names: NO2_GBD_year
Data type: Numeric
Units: ppb

User Notes: The primary exposure estimates for NO_2 for the Gateway projects are concentrations at participant addresses, averaged over the 10 years before each survey visit (incorporating residential history as possible). Secondary exposure estimates of interest are 1 and 5-year averages and averages over decades of life at the residential addresses. Areas of water are assigned a value of 0. For earlier years (1990-2000), the country boundaries are coarser, and some coastal areas may be missing.

Example of Methods Section:
"We estimated 1*0-year* average outdoor nitrogen dioxide (NO_2) concentrations for each participant based on their residential histories preceding each of their survey visits using model predictions available at a resolution of approximately $1km^2$. These estimates were available every 5-years from 1990 to 2010 and then annually from 2010 to 2019 (Anenberg et al. 2022). NO_2 was predicted by extending an existing spatial prediction model (Larkin et al 2017) derived from over 5000 monitors in nearly 60 countries (primarily in Europe, North America, and Asia) to additional years. Additional refinements to this model were made for rural settings including further calibration using the Modern-Era Retrospective analysis for Research and Applications reanalysis product and additional satellite-based data.

1.5 Ground-Level Ozone (O_3) Concentrations in Outdoor Air

Ozone (O_3) is a gas that is formed when emissions from sources including cars and industry react together in the presence of sunlight and heat. Ozone is a highly reactive gas that can be harmful to human health.

Spatiotemporal estimates of ambient ozone were generated at a 0.1^0 x 0.1^0 resolution between 1990 and 2017 across the world by Becker and colleagues (2023). Here, ozone is defined as the highest six-month running average of eight-hour daily maximum ozone mixing ratios (i.e., fractional concentration in parts per billion [ppb]). Concentrations were estimated using Bayesian Maximum Entropy (BME) fusion model to combine 9 atmospheric chemistry models. The modeling also incorporates a Regionalized Air Quality

Model Performance (RAMP) framework to flexibly correct for bias that occurs due to different model performance around the globe.

Citation: Becker JS, DeLang MN, Chang KL, Serre ML, Cooper OR, Wang H, Schultz MG, Schroder S, Lu Xiao, Zhang L, Deushi M, Josse B, Keller CA, Lamarque JF, Lin M, Liu J, Marecal V, Strode SA, Sudo K, Tilmes S, Zhang Li, Brauer M, West JJ. Using regionalized air quality model performance and Bayesian Maximum Entropy data fusion to map global surface ozone concentrations. Elementa: Science of the Anthropocene. 2023. 11(1):0025. https://doi.org/10.1525/elementa.2022.00025

Related Citation: Malashock DA, DeLang MN, Becker JS, Serre ML, West JJ, Chang KL, Cooper OR, Anenberg SC. Estimates of ozone concentrations and attributable mortality in urban, peri-urban and rural areas worldwide in 2019. Environ. Res. Lett. 2022. 17:054023 doi: 10.1088/1748-9326/ac66f3

Original Source: Provided by Jason West at University of North Carolina
Coordinate Reference System: WGS84
Data type: Raster

File names: Country_O3_year
Data type: Numeric
Units: ppb

User Notes: The primary exposure estimates for O_3 for the Gateway projects are concentrations at participant addresses, averaged over the 10 years before each survey visit (incorporating residential history as possible). Secondary exposure estimates of interest are 1 and 5-year averages and averages over decades of life at the residential addresses.

Example of Methods Section
"Long-term averages of ozone concentrations were calculated at participant addresses over the previous X years before their survey. Ozone levels are reported in parts per billion (ppb) using a global model of ozone (Becker et al. 2023) that estimated levels at a 0.1^0 x 0.1^0 resolution between 1990-2017. Briefly, Becker et al combined a rich set of ground level ozone monitoring data with multiple global atmospheric models using a Bayesian Maximum Entropy model with a Regionalized Air Quality Model Performance (RAMP) bias correction for differential model performance around the globe. These estimates are refinements of models originally developed for use in the Institute for Health Metrics and Evaluation Global Burden of Disease project and subsequently used in worldwide estimates of attributable mortality (Malashock et al. 2022)."

1.6 Greenspace

Greenspace is a measure intended to reflect exposure to nature and natural settings. It is hypothesized that greenspace can have important impacts on health through pathways including increased exercise, social engagement, and reduced stress. It can also provide protection from environmental contaminants and shade from sunlight and heat.

The Normalized Difference Vegetation Index (NDVI) has been provided as a marker of greenspace. NDVI is measured by the Moderate Resolution Imaging Spectroradiometer aboard NASA's Terra satellite (MODIS-Terra) satellite and was pre-processed by the Google Earth Engine. The MOD13Q1 version 6 data span from 2000 to 2021 and are available at a 250m, 16-day resolution. We have calculated the maximum, minimum, and mean NDVI for each survey year.

NDVI ranges from -1 to 1, with densely vegetated areas like forests approaching 1 and water, ice, pavement, and bare soil exhibiting low positive or negative values. Generally:

-1 to 0: Inanimate Objects or Dead Plants
0 to 0.33: Unhealthy plant
0.33 to 0.66: Moderately healthy plant
0..66 to 1: Healthy plant

(source: https://eos.com/blog/ndvi-faq-all-you-need-to-know-about-ndvi/)

Citation: Didan K, Munoz AB, Solano R, Huete A. MODIS Vegetation Index User's Guide (MOD13 Series). 2015. Doi: 10.5067/MODIS/MOD13Q1.006

Original Source: https://lpdaac.usgs.gov/products/mod13q1v006/ (version 6, Accessed 2022-12-07)
We downloaded, multiplied by the Scale Factor and processed using Google Earth Engine. A sample script is available in the Appendix.
Coordinate Reference System: WGS84
*Data File Type: R*aster

File names: NDVI_MAX/MIN/MEAN_MODIS_year
Data type: Numeric
Units: unitless

User Notes: The primary exposure estimate for greenspace for the Gateway projects is the maximum NDVI averaged over a 1km buffer around each participant address. This is intended to capture vegetation in a person's neighborhood regardless of the amount of greenness currently present based on the season. Secondary exposure estimates of interest are at 0, 250, and 5000 meters as well as minimum and mean NDVI values. The primary exposure duration is averaged over the 10 years before each survey visit (incorporating residential history as possible) with 1 and 5-year averages and averages over decades of life at the residential addresses being secondary.

Example Methods Section
"We estimated exposure to greenspace using Normalized Difference Vegetation Index (NDVI) using data available through the Google Earth Engine and derived from 250 meter resolution images captured by the Moderate Resolution Imaging Spectroradiometer aboard NASA's Terra satellite (MODIS-Terra). Specifically, we accessed version 6 of the MOD13Q1 data. NDVI serves as an indicator of vegetation within an area with values ranging from -1 to 1, with densely vegetated surfaces like forests approaching 1 and water, pavement, and bare soil exhibiting negative values.

For our analyses, we estimated the maximum annual NDVI in a 1km buffer around participant locations and generated a *10*-year average exposure based on their residential histories. We selected the maximum NDVI for the year to capture the presence of vegetation, irrespective of its greenness at a specific time or the duration of its greenness throughout the year."

1.7 Bluespace
Bluespace is a measure that reflects exposure to water (i.e., lakes, rivers, oceans) in natural settings. It is hypothesized that bluespace can have important impacts on health through pathways including increased exercise, social engagement, and reduced stress. It may also provide some cooling properties from extreme heat but may exacerbate flooding during extreme weather events.

Bluespace was estimated using a static map of open water bodies from the Land Cover CCI Climate Research Data Package (Water Bodies 4.0). This map is available at the 150m resolution and has been coded as a binary variable with 0 for land and 1 for water by leveraging multiple datasets from 2000 through 2012. The overall accuracy of the model was estimated between 98 and 100% with slightly lower accuracy (74-89%) when focusing on complex water body mapping such as coastlines and river banks.

Citation: Lamarche C, Santoro M, Bontemps S, D'Andrimont R, Radoux J, Giustarini L, Brockmann C, Wevers J, Defourny P, Arino O. Compilation and Validation of SAR and Optical Data Products for a Complete and Global Map of Inland/Ocean Water Tailored to the Climate Modeling Community. Remote Sensing. 2017; 9(1):36. https://www.mdpi.com/2072-4292/9/1/36

Original Source (website/link when applicable): http://maps.elie.ucl.ac.be/CCI/viewer/download.php
Coordinate Reference System: WGS84
Data File Type: Raster

File Names: nir_blue
Data type: Binary (0=land, 1=water)
Units: unitless

User Notes: The primary exposure estimate for bluespace for the Gateway projects is percent of bluespace within a 1km buffer around each participant address. Secondary exposure estimates of interest are at 250m and 5 km. The primary exposure duration is averaged over the 10 years before each survey visit (incorporating residential history as possible) with 1 and 5-year averages and averages over decades of life at the residential addresses being secondary.

Example Methods Section
"We estimated exposure to bluespace using data from the Land Cover CCI Climate Research Data Package (Water Bodies 4.0) (Lamarche et al 2017). These data were compiled using 9 different data sources for water between 2000 and 2012 to improve the spatial accuracy of the water boundaries. For our analyses, we estimated the percent bluespace in a 1km buffer around participant locations and generated a *10*-year average exposure based on their residential histories assuming that there is temporal stability in the water boundaries. In secondary analyses we looked at other distances from participants' homes including 250m and 5km."

1.8 Light at Night (LAN)
Nighttime light reflects human activities in neighborhoods. This parameter is of scientific interest as it may disrupt circadian rhythms and interfere with sleep. It is also used by some economists as a measure of wealth in low- and middle-income countries.

Annual global nighttime light has been produced from monthly cloud-free average radiance grids from the NASA/NOAA Visible Infrared Imaging Radiometer Suite (VIIRS) between 2012 and 2021. Elvidge and colleagues conducted pre-processing of the information to remove outlier features such as biomass burning and the aurora as well as moonlight. We utilized median masked global raster data.

Citation: Elvidge CD, Zhizhin M, Ghosh T, Hsu FC, Taneja J. Annual time series of global VIIRS nighttime lights derived from monthly averages:2012 to 2019. Remote Sensing, 2021, 13(5): 922, doi:10.3390/rs13050922

Original Source: https://eogdata.mines.edu/nighttime_light/annual/v21/
Coordinate Reference System: WGS84
Data type: Raster

File Names: Nightlight_VIIRS_year
Data type: numeric
Units: nanoWatts per square centimeter per steradian (nW/cm2/sr)

User Notes: The primary exposure estimate for light at night for the Gateway projects is the 10-year average estimated at the participant address before each survey visit (incorporating residential history as possible). Secondary analyses include the 1 and 5-year averages and averages over decades of life. LAN is extremely skewed. Log-transforming may aid in visualizing the data (maps, boxplots). Additional documentation: https://eogdata.mines.edu/products/vnl/

Example Methods Section
"Light at Night (LAN) exposures were obtained from the Joint Polar-orbiting Satellite System (JPSS) and the Visible and Infrared Imaging Suite (VIIRS) Day Night Band (DNB) technology and assigned to each participant address in the 10-years prior to each survey. These data were processed by the Earth Observation Group to remove pixels that were sunlit, moonlit or cloudy (Elvidge et al., 2012). We used annual median-masked rasters, version VNL2.1 which has an improved method of removing outliers. These data were at the 500m resolution from 2012-2021."

1.9 Spatial Splines

Environmental exposures are inherently spatial and thus there is a risk of these exposures being correlated with other features of place. While we attempt to adjust for these characteristics in our epidemiological analyses using personal and neighborhood features, there is the possibility of residual confounding by place. Therefore, we have developed code that can be used to estimate spatial splines to flexibly account for residual confounding. The code in the appendix will generate continuous place-based indicators to model place flexibly, allowing the user to specify the number of degrees of freedom desired.

Citations: Keller JP and Szpiro AA. Selecting a Scale for Spatial Confounding Adjustment. Journal of the Royal Statistical Society: Series A Statistical Society. 2020. 183(3): 1121–1143. doi:10.1111/rssa.12556.

Paciorek CJ. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Statistical Science. 2010. 25(1):107–125. doi: 10.1214/10-STS326

Original Source: NA, but R code is provided in the appendix
Coordinate Reference System: NA

Variable Names: spatialbasis1ofX
Data type: numeric
Units: unitless

User Notes: The code outputs *x* variables corresponding to the degrees of freedom specified. All *x* variables are entered into the model. For example for 5 degrees of freedom,Y=spatialbasis1of5 +spatialbasis2of5+ spatialbasis3of5 +spatialbasis4of5 +spatalbasis5of5. The degrees of freedom are not set. In HRS (United States of America), 10df was used.

Example Methods Section
"We adjusted for unmeasured confounding by location using spatial splines with *X* degrees of freedom. Spatial splines are a flexible approach to address spatial confounding by giving more degrees of freedom to areas more densely populated areas, as opposed to using categorical variables that do not account for population density- e.g. census divisions, state. These variables were created by applying thin plate regression splines to the coordinates of participant locations. Then using *x* degrees of freedom, *x* continuous variables are used to describe the participant's location."

2.0 USING ENVIRONMENTAL EXPOSOME DATA

2.1 Temporal Averaging

Recommended: 1-year, 5-year, 10-year

2.2. Spatial Buffers

Recommended: at location, 250m, 1km, 5km
Assuming that there are geocoordinates available for participant locations (e.g. latitude, longitude) resolution.

2.3 Spatial Adjustment

Environmental exposures and other unmeasured factors can be correlated with space. Therefore there is a need to ensure adjustment for space. If coordinates are available for participant locations, we recommend using spatial splines (code provided in the Appendix).

3.0 APPENDIX

3.1 Assigning Exposures to Participants (R Code)

This is sample code in R- which is free – but often has updates so please be advised that review and modifications may be required during implementation.

This links data to locations and buffers.

```
#Code Created for Gateway User Guide
#Note that this one transitions to the sf and stars R packages
#instead of using the rgdal packages
#VERY helpful sites for stars:
#https://r-spatial.github.io/stars/articles/stars1.html
#https://tmieno2.github.io/R-as-GIS-for-Economists/index.html
#Example using Ireland

#install.packages('stars')
#install.packages('terra')
#install.packages("exactextractr")
        #this is a MUCH faster way of calculating means over buffers
library(stars)
library(terra)
library(raster)
library(exactextractr)


#Load (x,y) coordinates for points
irl_sample<-read.csv('S:\\Adar\\Gateway\\Data\\User Guide\\Test\\irl_sample.csv')
class(irl_sample)
names(irl_sample)
        #set population data frame and convert to spatial object
pop_sample_sp<-st_as_sf(irl_sample, coords = c("x", "y"), crs = "WGS84")
        #plot points
plot(st_geometry(pop_sample_sp),cex=0.1)

############
#Load Rasters
############
setwd('S:\\Adar\\Gateway\\Data\\By Country\\Ireland\\')

list_files<-list.files()
        ####
        #PM2.5
        ####
        #subset to pm25
        list_pm25_dimaq<-subset(list_files, substr(list_files,1,10)=='PM25_DIMAQ')
        pm25_brick<-raster::brick(lapply(list_pm25_dimaq,raster))

        ####
        #NO2
        ####
        #subset to NO2
        list_no2_gbd<-subset(list_files, substr(list_files,1,7)=='NO2_GBD')
        no2_brick<-raster::brick(lapply(list_no2_gbd,raster))

        ####
        #Source Fraction
        ####
        #subset to Source fraction
        list_gbd_src<-subset(list_files, substr(list_files,1,10)=='GBD_raster')
        src_brick<-raster::brick(lapply(list_gbd_src,raster))

        ####
        #Source Fraction
        ####
        #subset to Source fraction
        list_gbd_src<-subset(list_files, substr(list_files,1,10)=='GBD_raster')
        src_brick<-raster::brick(lapply(list_gbd_src,raster))
```

```
#############
#Get Values at Different Buffers
############
#############
#Make Buffers around points
############
        #250m buffer
buffer250m <- st_buffer(pop_sample_sp, dist = 250)
        #1km buffer
buffer1km <- st_buffer(pop_sample_sp, dist = 1000)
        #5km buffer
buffer5km <- st_buffer(pop_sample_sp, dist = 5000)

############
#Buffer Size=0 (Extract to Point)
############

pm25_all<-terra::extract(pm25_brick, pop_sample_sp)
summary(pm25_all)

no2_all<-terra::extract(no2_brick, pop_sample_sp)
summary(no2_all)

src_all<-terra::extract(src_brick, pop_sample_sp)
summary(src_all)


############
#Buffer Size=250m
############
        #plot to check
plot(st_geometry(buffer250m[buffer250m$X %in% c(1:10),]), border = "red", lwd = 2, col = NA)
plot(st_geometry(pop_sample_sp[pop_sample_sp$X %in% c(1:10),]),
  add = TRUE,
  cex = 1,
)
pm25_250m<-exact_extract(pm25_brick,
                         buffer250m,
                        fun='mean')
        #better looking names
oldnames<-names(pm25_250m)
names(pm25_250m)<-lapply(oldnames,function(x){
  paste0("PM25_DIMAQ_",  substr(x, nchar(x)-4+1, nchar(x)), "_Mean_250m")
        })

no2_250m<-exact_extract(no2_brick,
                         buffer250m,
                        fun='mean')
        #better looking names
oldnames<-names(no2_250m)
names(no2_250m)<-lapply(oldnames,function(x){
  paste0("NO2_GBD_",  substr(x, nchar(x)-4+1, nchar(x)), "_Mean_250m")
        })

pm25_no2_250m<-cbind(pm25_250m, no2_250m, pop_sample_sp)
head(pm25_no2_250m)
summary(pm25_no2_250m)

############
#Buffer Size= 5km
############
```

```
        #plot to check
plot(st_geometry(buffer5km[buffer5km$X %in% c(1:10),]), border = "red", lwd = 2, col = NA)
plot(st_geometry(pop_sample_sp[pop_sample_sp$X %in% c(1:10),]),
  add = TRUE,
  cex = 1,
)

#Blue Space within Buffer Distance
irl_blue<-raster("S:\\Adar\\Gateway\\Data\\By Country\\Ireland\\irl_blue.tif")
        #get max value in 5km buffer
irl_blue_5km<-exact_extract(irl_blue,
                            buffer5km,
                           fun='mean')
summary(irl_blue_5km)
irl_blue_5km<-cbind(pop_sample_sp, irl_blue_5km)
```

3.2 Calculating Exposure Averages (1-year, 5-year, 10-year, decadal)

```
###################
#Calculating Exposure Averages Using intervalaverage
###################
```

R package intervalaverage

Structure of data, Long

3.3 Generate Spatial Splines

```
#Creating Thin plate spline functions for lat/long

#desc
#Create thinplate splines for points
#[using unpenalized thin-plate regression splines (TPRS) in the MGCV package]
#(Wood 2003). TPRS are a flexible way of adjusting for spatial confounding.
#Using singular value decomposition, they decompose the distance matrix of all
#participant locations into a set of basis functions, the first k of which are
#included as adjustment covariates in the health models (Wood 2003).
#end desc

#Michael's code (transcribed)
#library(mgcv)
#library(splines)
#n<-50
#s1<-runif(n^2)
#s2<-runif(n^2)
#y<-s1+s2^3 +s1*s2 + s1*4*sqrt(s2)+sin(4*s1)
#y<-100*(y-min(y))/diff(range(y))
#z<-gam(y~s(s1,s2,fx=TRUE,k=10))
#H<-predict.gam(z,type="lpmatrix")
#library(data.table)
#setwd()
#x<-fread("unique_locs_all.txt")
#x[,y:=1L]
#
#ndf<-c(10L,15L,20L)
#H_list<-lapply(ndf,function(q){
#           m<-gam(y~s(lambert_x,lambert_y, fx=TRUE, k=q+1L),data=x)
#           H_temp<-predict.gam(m,type="lpmatrix")
#           out<-as.data.table(H_temp[,-1])
#           newnames<-paste0("spatialbasis",1:ncol(out),"of",q)
#           setnames(out,newnames)
#           out
#           })
#           cbind(H_list)
#v15<-gam(y~s(lambert_x,lambert_y,fx=TRUE,k=16),data=x)
#H15<-predict.gam(v15,type="lpmatrix")
#
#v20<-gam(y~s(lambert_x,lambert_y,fx=TRUE,k=21),data=x)
#H20<-predict.gam(v20,type="lpmatrix")
```

```
#what I actually ran
library(mgcv)
library(splines)
n<-50
s1<-runif(n^2)
s2<-runif(n^2)
y<-s1+s2^3 +s1*s2 + s1*4*sqrt(s2)+sin(4*s1)
y<-100*(y-min(y))/diff(range(y))
z<-gam(y~s(s1,s2,fx=TRUE,k=10))
H<-predict.gam(z,type="lpmatrix")


library(data.table)
jen_loc<-read.csv("C:\\Users\\changj\\starbucks_us_locations.csv")

View(jen_loc)
names(jen_loc)
jen_loc<-jen_loc[c("long","lat")]
jen_loc<-na.omit(jen_loc)
dim(jen_loc)
library(ggplot2)
library(usmap)
library(maptools)
jen_trans<-usmap_transform(jen_loc)

plot_usmap() +
  geom_point(data = jen_trans, aes(x = long.1, y = lat.1),
          color = "red", alpha = 0.25)

setwd("C:\\Users\\changj\\")
x<-fread("starbucks_us_locations.csv")
x[,y:=1L]
names(x)


# test on 10
ndf<-c(5L,7L,10L)
H_list<-lapply(ndf,function(q){
              m<-gam(y~s(lat,long, fx=TRUE, k=q+1L),data=x)
              H_temp<-predict.gam(m,type="lpmatrix")
                        #type="lpmatrix" then a matrix is returned which yields the values of the
                        #linear predictor (minus any offset) when postmultiplied by the parameter
                        #vector (in this case se.fit is ignored). The latter option is most useful
                        #for getting variance estimates for quantities derived from the model: for
                        #example integrated quantities, or derivatives of smooths. A linear predictor
                        #matrix can also be used to implement approximate prediction outside R
                        #(see example code, below)
              out<-as.data.table(H_temp[,-1])
              newnames<-paste0("spatialbasis",1:ncol(out),"of",q)
              setnames(out,newnames)
              out
              })
cbind(H_list)
head(H_list)
jen_sp<-cbind(H_list,jen_loc)
```

3.4 Download NDVI from Google Earth Engine

Google Earth Engine is freely available and only requires creating an account.

```
//Example is 2000 AND INDIA
//Code for Downloading MODIS NDVI for Gateway Countries

//Loading MODIS data
var modis = ee.ImageCollection('MODIS/006/MOD13Q1');

//Creating an array of feature collections
```

```
    //extent of india
    // class      : Extent
    // xmin       : 68.2056
    // xmax       : 97.39556
    // ymin       : 6.755997
    // ymax       : 35.67455

    // var xMin =68.2056;
    // var yMin = 6.755997 ;
    // var xMax = 97.39556;
    // var yMax = 35.67455 ;
    //Region 1
    // class      : Extent
    var xmin1       = 68 ;
    var xmax1       = 98 ;
    var ymin1       = 6;
    var ymax1       = 20;

    //Region 2
    // class      : Extent
    var xmin2       = 68 ;
    var xmax2       = 98 ;
    var ymin2       = 19;
    var ymax2       = 36;
    // Construct a rectangle from a list of bounding coordinates.
    var rectangleBounds1 = ee.Geometry.Rectangle(
      [xmin1, ymin1, xmax1, ymax1]
    );
    Map.addLayer(rectangleBounds1, {}, 'rectangleBounds1');


    var rectangleBounds2 = ee.Geometry.Rectangle(
      [xmin2, ymin2, xmax2, ymax2]
    );
    Map.addLayer(rectangleBounds2, {}, 'rectangleBounds1');

//One for each region
var regions = [rectangleBounds1, rectangleBounds2
];
  regions.forEach(
  function(region){
     var modis = ee.ImageCollection('MODIS/006/MOD13Q1')
                 .filterDate('2000-01-01', '2021-12-31')
                 .filterBounds(region)
                 .select('NDVI');
     for (var year = 2001; year <= 2021; year=year+1) {
     var startDate = year + '-01-01'; //Start Jan 1st
     var endDate = year + '-12-31'; //Ends Dec 31st
     var annual = modis.filter(ee.Filter.date(startDate,endDate)) //Filter by dates
     .qualityMosaic('NDVI') //Take the MAX
     // .clipToCollection(region) ;//Clip to regions
     Export.image.toDrive({
       image: annual,
       description: 'region'+(regions.indexOf(region)+1).toString()+'_'+year+'_MAX',
       folder: 'Max NDVI (MODIS)- India', // should match your Google Drive folder
       scale: 250, // resolution
       region: region,
       crs: 'EPSG:4326', // can be whatever you want 4269 might be better for NA
       maxPixels: 1e10}); // may need to boost to meet size
     }
  }
  );
```

3.5 Download Land Cover Classification Data from Google Earth Engine

```
//Code to Download Land Cover Classifications Types 1 & 2
//Type 1: Land Cover Type 1: Annual International Geosphere-Biosphere Programme (IGBP) classification
//Type 2: Land Cover Type 2: Annual University of Maryland (UMD) classification

//FOR EACH COUNTRY NEED TO CHANGE: INDIA
// Line 12: cty_poly- make sure the polygon is right
// Line 33: make sure the folder name is updated
// Line 48: make sure the folder name is updated


//Import Country Boundary
var cty_poly = ee.FeatureCollection('users/changj/ind_poly');
print(cty_poly);

//Display the shapefile into the interactive map
Map.addLayer(cty_poly);

//Filter Land Cover by Country Polygon (masks the other values)
var collection = ee.ImageCollection('MODIS/006/MCD12Q1')
  .filterDate('2001-01-01', '2020-01-01')
  .filterBounds(cty_poly);  // Intersecting with country
  print(collection);

//Select the LC_Type1 Band: Land Cover Type 1: Annual International Geosphere-Biosphere Programme (IGBP) classification
var lc_type1=collection.select('LC_Type1');

//Export
/*call up batch function
https://gis.stackexchange.com/questions/248216/exporting-each-image-from-collection-in-google-earth-engine
*/
var batch = require('users/fitoprincipe/geetools:batch')

batch.Download.ImageCollection.toDrive(lc_type1, 'India/LC_Type1',
                {scale: 500,
                 region: cty_poly,
                 type: 'float'})


//Select the LC_Type2 Band: Land Cover Type 2: Annual University of Maryland (UMD) classification
var lc_type2=collection.select('LC_Type2');

//Export
/*call up batch function
https://gis.stackexchange.com/questions/248216/exporting-each-image-from-collection-in-google-earth-engine
*/
var batch = require('users/fitoprincipe/geetools:batch')

batch.Download.ImageCollection.toDrive(lc_type2, 'India/LC_Type2',
                {scale: 500,
                 region: cty_poly,
                 type: 'float'})
```