

Data Documentation for Census Tract Environmental Quality Index

Disclaimer: This research **dataset** has been reviewed in accordance with U.S. Environmental Protection Agency (U.S. EPA), Office of Research and Development, and approved for release. Mention of brand names or vendors does not constitute an endorsement of products or services by the U.S. EPA.

Background

The Environmental Quality Index (EQI) accounts for the multiple domains of the environment with which humans interact. These domains include chemical, natural, built, and sociodemographic environments that have both positive and negative influences on health. An overall EQI was created for census tracts within the contiguous United States for 2006-2010 and 2011-2015.

Census tract-level EQI

Data Source Identification and Review

Data Selection

From within each domain — air, water, land, sociodemographic, and built environments — specific constructs or major areas were identified (Table 1). Constructs were initially identified based on the literature and frequently used datasets; constructs used in the 2006-2010 County EQI served as a starting point for the tract-level EQI [1]. The constructs included in census tract EQIs 2006-2010 and 2011-2015 were those constructs where data were available at the census tract-level.

Table 1. Constructs for each environmental domain at the census tract-level.

Domain	Constructs Considered	Constructs included
Air	1.) Hazardous air pollutants 2.) Criteria air pollutants	1.) Hazardous air pollutants 2.) Criteria air pollutants
Built Environment	1.) Roads 2.) Commute time 3.) Business environment 4.) Walkability 5.) Green space 6.) Highway safety 7.) Housing environment	1.) Commute time 2.) Business environment 3.) Green Space
Land	1.) Facilities 2.) Pesticides 3.) Agriculture 4.) Radon 5.) Mining Activity	1.) Facilities 2.) Pesticides
Sociodemographic	1.) Socioeconomic 2.) Crime 3.) Creative class representation 4.) Political character	1.) Socioeconomic 2.) Crime
Water	1.) Overall Water Quality	1.) Drought monitoring

	2.) Drought monitoring 3.) Drinking water supply 4.) General water contamination 5.) Domestic use 6.) Atmospheric deposition 7.) Chemical contamination 8.) Drinking water quality	2.) Drinking water supply
--	--	---------------------------

Data Source Search

Once the desired constructs were identified, potential sources for data were selected to represent those constructs. In general, a broad approach to data source identification was undertaken to:

- Identify EPA and non-EPA domain-specific environmental data sources for all census tracts in the contiguous United States;
- Summarize environmental data source availability, quality, spatial and temporal coverage, storage requirements, and acquisition steps; and
- Obtain the identified data.

Possible data sources were identified using Web-based search engines (e.g., Google), site-specific search engines (e.g., Federal and State data sites), literature-reported data sources (e.g., PubMed, ScienceDirect, TOXNET), and personal communications from data owners. Data that were available at—or had the potential to be aggregated to—the United States census tract-level were sought. Data were restricted to represent the years 2006-2010 and 2011-2015.

Table 2 identifies the data sources that were acquired and used for the construction of the EQI and includes a description of the data source and variables.

Table 2. Sources of data for air, water, land, built-environment, and sociodemographic domains for use in the tract-level Environmental Quality Index 2006-2010 and 2011-2015

AIR DOMAIN		
Source of Data	Description	Variables
CMAQ Fused Air Quality Surface Using Downscaling (FAQSD) [2]	Tract-level criteria air pollutant 5-year average	Particulate Matter _{2.5} (PM _{2.5}) Ozone
National-Scale Air Toxics Assessment (NATA 2011, 2014) [3]	HAPs Summation Variables based upon 140 (137 in 2011-2015) NATA pollutants using emissions information from the National Emissions Inventory and meteorological data input into the Assessment System for Population Exposure Nationwide model	*Sum of selected NATA contaminants

BUILT-ENVIRONMENT DOMAIN

Source of Data	Description	Variables
Dun and Bradstreet North American Industry Classification System codes (2008 [4], 2013 [5])	Description of food environment (fast food restaurants, groceries, convenience stores) and education environment (schools, daycares, universities) per census tract	†Education-related Environment – count of education-related businesses per square mile of census tract; Food Environment - ratio of negative food resources to total food resources per tract (count of negative food businesses /total food businesses)
National Land Cover Database [6]	Maintained by the U.S. Geological Survey. The National Land Cover Database provides information on land cover and land cover change for the Contiguous U.S. at different time points.	Census tract land area classified as natural cover and open space - percentage of open space in tract
United States Census (2006-2010 [2010], 2011-2015 [2015]) [7-9]	Census tract-level population commuting characteristics	Commute Time – time it takes to travel from home to work (minutes)

LAND DOMAIN

Source of Data	Description	Variables
EPA Facility Registry Service (2006-2010, 2011-2015) [10]	Maintained by EPA and provides locations of and information on facilities throughout the United States; different datasets within this database are updated at different intervals, but most are updated monthly; no set spatial scale across datasets. Some provide addresses, some geocoded addresses, etc.	†Facilities – natural log of the count of facilities per square mile of census tract
(1) United States Department of Agriculture Cropscape [11] (2) United States Geological Survey “Estimated Annual Agricultural Pesticide	(1) Percent land covered by specific crops tabulated to census tracts (2) Amount of total agricultural chemicals applied per square kilometer	†Pesticide Application- natural log of sum of pesticides in kg per square km

Use for Counties of the Conterminous United States” data for pesticide use [12, 13]

SOCIODEMOGRAPHIC DOMAIN

Source of Data	Description	Variables
United States Census (2006-1010 [2010], 2011-2015 [2015]) [7-9]	Census tract-level population and housing characteristics, including density, race, spatial distribution, education, socioeconomics, home and neighborhood features, and land use	<p>† Poverty – natural log of percent of families living below federal poverty level (percent);</p> <p>† Unemployment - natural log of percent of persons who are unemployed (percent);</p> <p>Bachelor’s Degrees – percentage of people 25 or older with a Bachelor’s degree;</p> <p>Income Spent on Rent – percentage of income spent on rent by tract residents</p>
ESRI Crime Index (2006-2010, 2011-2015) [14, 15]	Tract-level models based on reports of crime	†Crime - natural log transformed crime rate (count of crimes / tract population)

WATER DOMAIN

Source of Data	Description	Variables
United States Geological Survey Estimates of Water Use in the United States [16]	Percentage of tract that is self-service for water supply (example: a privately owned well), log transformed	†Self-Service of Water Supply - percent of population that is self-service for water source
Drought Monitor Data [17]	Percentage of tract in extreme drought from weekly data across 5 year average	Drought - percent of tract in extreme drought during the 5 year period

*Air domain: This symbol indicates the variable was summed then natural log transformed.

†Built-Environment, Land, Sociodemographic, Water domains: This symbol indicates variables that were natural log transformed.

Variable Construction

Approach

The overall summarized variable development process for tract-level EQI 2006-2010 and 2011-2015 was as follows:

- identify and develop relevant variables within each domain for each available year (2006-2010, 2011-2015),
- assess missing data and variability of each variable,
- assess collinearity among the variables,
- assess normality of variables and transform as necessary,
- remove any census tracts with population less than 10 and water census tracts.

Identification and Construction of Variables from Data Sources

EPA's 2006-2010 County-level EQI domains and variables served as a starting point for the 2006-2010 and 2011-2015 Tract-level EQI variable identification and construction process. Variables were created from selected data sources to represent the constructs. Ideally, developed variables would have measured or estimated values for each census tract in the United States. When this criterion was not met, or when a majority (>50%) of values were zero, the proportion of missing data and zero values were evaluated for variable inclusion. If a particular variable had information missing for many tracts, the nature of the missing data was evaluated. When it was determined that the missing data could be interpreted as meaningful zeros (i.e., no measures were taken because that condition did not occur in that census tract), the missing values were set to zero. If the missing data could not be determined to be legitimate zeros, and the data could not be reasonably averaged over geography, and the number of census tracts with missing data was too high (more than 50% of census tracts), the variable was not used in the tract-level EQI. The data reduction method Principal Component Analysis (PCA) is based on the variability between variables [18]; therefore, collinearity of variables was assessed.

We assessed if each of the variables was normally distributed. A key assumption for the PCA data reduction approach is that variables are distributed normally [18]. If data were non-normal, data were natural log transformed to increase normality. For those variables with zero values, half of the non-zero minimum value was added to all observations before log-transformation.

Any census tracts with a population less than 10 and any water census tract, which is defined as any census tract where land area is equal to 0, were excluded. For the water census tracts, the sociodemographic variables in the index do not apply. In health outcome analyses, census tracts with $n < 10$ need to be suppressed, thus exclusion of census tract populations less than 10.

Air Domain

The air domain consists of two data sources, (1)CMAQ downscaler [2] representing criteria air pollutants and (2)NATA [3] representing HAPs.

Criteria air pollutants

Ozone and particulate matter are two common pollutants that are indicative of overall air quality with regards to non-point source pollutants. These two pollutants generally come from anthropogenic sources such as from car exhaust in the case of ozone or from fossil fuel power plants in the case of

PM_{2.5}. This data was provided through CMAQ, or Community Multi-Scale Air Quality model, which estimates ozone and particulate matter concentrations across time through models and simulations. For each census tract, the weekly data were summarized into a yearly average, which were then further summarized into a five-year average for the EQI's respective timeframe. This provided an estimate of typical values for these criteria pollutants at the census tract-level.

Hazardous Air Pollutants (HAPs)

Tract-level concentration estimates from National Air Toxics Assessment (NATA) were used for all HAPs included in the EQI. HAPs were selected for inclusion from the full NATA pollutant list. Using data from the 2011 NATA for 2006-2010 and the 2014 NATA for 2011-2015, variables were evaluated for collinearity and variability. Variables with any correlation coefficient >0.70 were examined, and representative variables were chosen for each pair or group of highly correlated variables. Of the remaining variables, all missing values were set to zero, with the assumption that lack of estimate for an area indicated low concern for contamination with a particular HAP, and the number of zero values was evaluated for each variable. Pollutants with more than 50% zero values were dropped. This process left 39 HAPs included in the 2006-2010 Census Tract EQI. Three of these 39 HAPs were not available in NATA 2014, thereby leaving 36 HAPs included in the 2011-2015 Census Tract EQI (Table 3). The HAPs did not generally follow a normal distribution; therefore, a summation variable was created, and natural log transformed. The summation reduced data dimensionality to reduce bias in the overall EQI from the air domain. Multiple HAPs existing individually in the analysis would heavily weight the air domain of the EQI. The summation additionally excluded outliers from the analysis as there were certain regions where the total amount of HAPs far surpassed the totals for the selected HAPs.

Table 3. NATA variables included in EQI 2006-2010 and 2011-2015

Pollutants in 2006-2010 EQI from NATA 2011	Pollutants in 2011-2015 EQI from NATA 2014
1_1_1-trichloroethane	
1-1-2-2-tetrachloroethane	1-1-2-2-tetrachloroethane
1,4-dichlorobenzene	1,4-dichlorobenzene
1,4-dioxane	
2_4-toluene diisocyanate	2_4-toluene diisocyanate
2-nitropropane	2-nitropropane
4,4'-methylenediphenyl diisocyanate (MDI)	4,4'-methylenediphenyl diisocyanate (MDI)
Acetophenone	Acetophenone
Acrylic Acid	Acrylic Acid
Benzidine	Benzidine
Biphenyl	Biphenyl
Carbonyl Sulfide	Carbonyl Sulfide
Chlorine	Chlorine
Chloroprene	Chloroprene
Cyanide Compounds	Cyanide Compounds
Dibenzofuran	Dibenzofuran
Dimethyl Formamide	

Dimethyl Phthalate	Dimethyl Phthalate
Dimethyl Sulfate	Dimethyl Sulfate
Epichlorohydrin	Epichlorohydrin
Ethylbenzene	Ethylbenzene
Ethylene Oxide	Ethylene Oxide
Ethylene Dichloride (1,2-dichloroethane)	Ethylene Dichloride (1,2-dichloroethane)
Formaldehyde	Formaldehyde
Hexachlorobenzene	Hexachlorobenzene
Hexamethylene Diisocyanate	Hexamethylene Diisocyanate
Hydrazine	Hydrazine
Hydrogen Fluoride	Hydrogen Fluoride
Hydroquinone	Hydroquinone
Isophorone	Isophorone
Maleic Anhydride	Maleic Anhydride
Manganese Compounds	Manganese Compounds
Mercury Compounds	Mercury Compounds
Methyl Tert-Butyl Ether	Methyl Tert-Butyl Ether
Phenol	Phenol
Phosphorus Compounds	Phosphorus Compounds
Propylene Oxide	Propylene Oxide
Quinoline	Quinoline
Tetrachloroethylene	Tetrachloroethylene
Vinyl Acetate	Vinyl Acetate

Built Domain

Data were identified at the census tract-level for three constructs: (1) Commute time [7, 8], (2) Business environment [4, 5], and (3) Green space [6, 19, 20] .

Commute Time

The average number of minutes employed persons spent commuting was obtained from the American Community Survey (ACS); note this variable was provided in both the 2010 five-year and 2015 five-year ACS and did not need further processing [7, 8].

Education-Related and Food Environments

Businesses represent an important component of the built environment and can contribute to the risk and amenity landscape. Variables representing various built-environmental features were constructed using 2008 (for 2006-2010 EQI) and 2013 (for 2011-2015 EQI) Dun and Bradstreet proprietary data [4, 5], which include more than 195 million records on businesses identified through the North American Industry Classification System (NAICS) codes. The variables that were constructed to represent the business environment construct included the (1) food environment and (2) education-related environment. Positive food environments included those that sold healthier foods, like grocery stores, sit-down restaurants, and organic shops, whereas the negative food environment included businesses

like fast-food restaurants and convenience stores (Table 4). Although related, these two food environments comprise different businesses and are not 100% inversely correlated. The positive food environment and negative food environment variables were summed to obtain a value for the total food environment in a census tract; we then divided the negative food environment by the total food environment to create a ratio of negative food environment to total food environment. The education-related environment (Table 4) was the total number of education-related businesses in a tract divided by the census tract area in square miles then standardized and log transformed.

Table 4. NAICS Codes used for Education-related and Food Environments

Variable Construct	NAICS Code	Description
Positive Food	452910; 452311 in 2013	Warehouse Clubs and Superstores
Positive Food	445110	Supermarkets and Other Grocery (except Convenience) Stores
Positive Food	445210	Meat Markets
Positive Food	445220	Fish and Seafood Markets
Positive Food	445230	Fruit and Vegetable Markets
Positive Food	445299	All Other Specialty Food Stores
Positive Food	311811	Retail Bakeries
Positive Food	445291	Baked Good Stores
Positive Food	446191	Food (Health) Supplement Stores
Positive Food	722110; 722511 in 2013	Full-Service Restaurants
Positive Food	722212; 722514 in 2013	Cafeterias
Negative Food	445120	Convenience Stores
Negative Food	722211; 722513 in 2013	Limited-Service Restaurants
Negative Food	722213; 722515 in 2013	Snack and Nonalcoholic Beverage Bars
Negative Food	722330	Mobile Food Services
Education	611511	Cosmetology and Barber Schools
Education	611610	Fine Arts Schools
Education	611110	Elementary and Secondary Schools
Education	611310	Colleges, Universities, and Professional Schools
Education	611210	Junior Colleges
Education	519120	Libraries and Archives
Education	611420	Computer Training
Education	611519	Other Technical and Trade Schools
Education	611410	Business and Secretarial Schools
Education	611513	Apprenticeship Training

Education	611512	Flight Training
Education	611691	Exam Preparation and Tutoring
Education	611692	Automobile Driving Schools
Education	611630	Language Schools
Education	611710	Educational Support Services

Green space

Exposure to green space has been associated with positive health. The green space variable, which is the census tract land area classified as natural cover and open space, was created by the EPA's EnviroAtlas [19, 20] using National Land Cover Database (NLCD) [6]. To create a green space variable, five total land cover groups were combined, those classified as natural land cover (barren land (rock/sand/clay/tundra/perennial ice), forest, shrubland/scrub land, herbaceous, and wetlands) and those classified as developed, open space, where impervious surfaces make up less than 20% of total cover and includes recreational areas such as grassy lawns, parks, and golf courses. This combined variable of natural land cover and developed, open space gave a percentage of the tract that had green space and ranged from 0-100 percent.

Land Domain

The land domain consisted of two data sources, representing two constructs: (1) pesticide application [11-13] and (2) facilities [10].

Pesticide Application

Pesticide application for each tract was estimated using crop land coverage data from U. S. Department of Agriculture (USDA) CropScape and pesticide compound use in kilograms by crop type from U. S. Geological Survey's (USGS) Estimated Annual Agricultural Pesticide Use for Counties of the Conterminous United States data [11]. Crop-specific land coverage percentages were constructed by first downloading CropScape raster data for each state's cultivated crops for each year within the EQI's timeframe. A crop variable was created to map the Cropscape crop names to 10 standard crop names (Corn, Soybeans, Wheat, Cotton, Vegetable_and_Fruit, Rice, Orchards_and_Grapes, Alfalfa, Pasture_and_Hay, and Other Crops). The U.S. Census Bureau's 2010 TIGERLINES for the states, counties, and census tracts were then used to generate each crop's land coverage percentages for each geographic unit.

To estimate total pesticide use in kilograms for each census tract, USGS's "Estimated Annual Agricultural Pesticide Use for Counties of the Conterminous United States" [12, 13] data using the EPest high estimation method data were used to find pesticide use by crop group at the county-level. Data were then combined with the land coverage percentage data to estimate pesticide use per square kilometers of a census tract.

Using the combined and processed datasets, pesticidal chemical compounds for each year of the EQI's timeframe was summed by the chemical compound type. For each chemical compound in a census tract, only the maximum value over the five-year timeframe of the EQI was selected and was then summed with the other chemical compounds' maximum values within the five-year timeframe of the EQI to create the total pesticide use in kilograms for the census tract. The resulting value was then standardized by adding the non-zero minimum total pesticide use value across all census tract divided by 2. The standardized total pesticide use is then divided by the census tract's land area (km²) and log

transformed to create the final pesticide use construct.

Facilities

Large facilities have the capacity to affect land quality. The facilities included in the land domain are those represented on the EPA Facility Registry Service [10]. Because many census tracts had at least one of six facility types present within the Facility Registry Service (FRS), but no tracts had all types, a composite facilities data variable was constructed by summing the count of any one of the six facilities types (Brownfield sites, Superfund sites, Toxic Release Inventory sites, pesticide-producing-location sites large-quantity generator sites, and treatment, storage, and disposal sites) across the tracts divided by the census tract area in square miles then standardized and log transformed. Some facilities were denoted as having multiple types; these were not double counted. The same dataset downloaded in 2011 was used for both EQI time periods since dates in which a facility was first added to the FRS were inconsistent. Due to limitations with the FRS it is not possible to know with any accuracy exactly when a Facility was first added to the list. Some created dates would be overwritten automatically within the FRS database and would not match data available from other sources. In other words, the FRS data is accurate for when in time it was downloaded but due to this missing information with the date of creation, subsets by date of entry are not possible to construct. For example, the FRS only starts on Jan-1-2000 but the Superfund sites were largely given the designation decades prior to the year 2000. A limitation of the dataset is that it undercounts facilities for 2011-2015 because the data is only current through 2011 and facilities added after this are not included in the count.

Sociodemographic Domain

From the two identified data sources for the sociodemographic domain, five variables were created: (1) Bachelor's degrees, (2) Poverty, (3) Renter income, (4) Unemployment [7-9], and (5) Crime [14, 15].

Bachelor's Degrees, Poverty, Renter Income, and Unemployment

The four variables obtained from the U.S. Census were downloaded as constructed by the Census [7-9] and include: (1) percent earning a Bachelor's degree or higher among persons aged 25 years or older, (2) percent persons unemployed, (3) percent of families living below the Federal poverty line, and (4) gross rent as percentage of household income.

Crime

One variable was used to estimate the presence of crime at the census tract-level. This variable represents the summation of seven different crime types (murder, rape, robbery, assault, burglary, theft, and motor vehicle theft), was processed by ESRI, downloaded from ESRI [15], and used to represent a total crime variable for the tract level.

Water Domain

From the two identified data sources for the water domain, two variables were created: (1) Drought [17] and (2) Self-service water supply [16].

Drought

Drought Monitor at University of Nebraska Lincoln provides weekly GIS drought data for the contiguous U.S. [17] There are 5 levels to drought classification ranging from not drought impaired to exceptional drought. The two highest levels (extreme or exceptional drought) were used to indicate a sufficient level of drought for inclusion in the EQI. Using the sum of extreme or exceptional drought weeks within a census tract, the final drought variable was calculated for each census tract by dividing the sum by the total number of weeks in a year to get the average drought value.

Self-Service

Self-service of water supply was included because of the different regulations regarding municipal water sources and self-managed domestic systems such as wells. Using USGS's water use data [16], the data for the population using self-supplied water is provided at the county level. This value was then multiplied by the proportion of each census tract's population to the total county population and then multiplied by 100 to obtain the percent tract population on self-supplied water. The resulting percentage represents an estimate for self-service of water supply variable.

Data Reduction and Index Construction

Overall Approach

After variable development, all the variables were combined into an index representing the overall environmental quality. The specific tasks required for index construction were as follows:

- Include all the variables from each domain in a PCA to empirically summarize environmental context (retaining the first component as the index)
- Assess the valence of the variable loadings; if loadings were not in the correct direction to ensure a higher value on the index corresponded to worse environmental quality, corrected valence when necessary
- Repeat the previous steps for each of the five RUCA strata such that each RUCA had its own overall index

Principal components analysis (PCA)

PCA is a data reduction technique frequently used to create indices for inclusion in statistical models [18]. PCA analyzes total variance, and the loading represents the correlation between the variable and the component. PCA assumes no underlying latent variable structure but, rather, seeks to empirically summarize multiple possible domains.

PCA was chosen for data reduction for several reasons. An empirical summary of the constituent domains of the EQI was desired. The differing scales at which variables were measured needed to be accommodated. Because PCA standardizes measures prior to combining, the differing scales are less problematic. Finally, PCA enables variable loadings to differ by their relative importance to the total component. This feature enabled exploration of variable loading differences for interpretation and assessment of particularly influential or meaningless variables. However, PCA requires at least three variables [21] and since the water and land domains did not meet this requirement, domain specific indices could not be done at the census tract-level.

Following the PCA process, the indices were valence-corrected for use. "Valence correction" refers to reorientation of PCA output for (1) uniformity of interpretation of indices (2) uniformity in orientation of RUCA stratified indices. Therefore, the loading valence needed to be corrected prior to the construction of the indices to ensure that higher values on the overall EQI, signify worse environmental quality.

EQI indices are designed such that lower values represent "better" quality and higher values represent "worse" quality. Therefore, health beneficial variables should be inversely loaded on the overall component (with a "-" sign) and a potentially health harming variable will be represented with a "+" loading. Given that the first principal component was taken to represent environmental quality and that the orientation of these indices was designated as going from better to worse quality (negative to

positive index value), it was necessary to reverse the component variable loadings vector from a PCA output if the variables included in the index were incorrectly loaded. Determination of variables as beneficial or detrimental to human health across domains was done a-priori based on literature evidence and content matter judgement (Table 5). Reorientation of PCA derived indices through multiplication of the component variables loading vector by -1 preserves:

- the direction of the relationship among the variables for a given PCA (i.e. variables that loaded with same signs will retain same signs and variables that loaded opposite to each other will retain opposite signs after reversal and therefore the pattern of correlations among the variables will remain intact) and
- the magnitude of correlation among variables (reversal of loading signs does not impact the magnitude of the loading) [22].

Table 5. A Priori Loading Directions for Census Tract EQI Variables

Variable Name	Assumed Direction
PM _{2.5}	+
Ozone	+
Sum of selected NATA contaminants	+
Commute time	+
Education-related environment	-
Food environment	+
Census tract land area classified as natural cover and open space	-
Facilities	+
Pesticide Application	+
Unemployment	+
Poverty	+
Bachelor's Degrees	-
Crime	+
Income Spent on Rent	+
Self-Service of Water Supply	+
Drought	+

PCA analyzes the total variance of the variables included in the command. To construct the EQI, variables were entered into the PCA, which produced variable loadings, roughly equivalent to the “weight” or contribution that each variable made toward explaining the total variance. The loading associated with each variable then was multiplied by the mean value of that variable at the given geography (census tract for the EQI), and these weighted mean values were summed.

Rural-Urban Commuting Area

The overall EQI was created for each census tract in the contiguous United States. Recognizing that environments differ dramatically across the rural-urban continuum [23], the decision was made that the EQI would be most useful if it accommodated rural-urban environmental differences. The census tract EQI was stratified by Rural-Urban Commuting Areas (RUCAs) [24]. RUCA represents an appropriate level of measurement because it is based on year 2010 census tract designations of primary and secondary traffic flow patterns. RUCA codes are similar to Rural Urban Continuum Codes (RUCCs) in their attention to metropolitan and micropolitan areas, but further divide tracts by the type of community to which

commuting traffic primarily flows [23, 24]. It includes 10 major primary traffic flow categories with 1 to 6 secondary flow subcodes with a focus on commuting patterns. For instance, RUCA code 8 is classified as a small town with high commuting (primary flow 30% or more to a small urban cluster) and has subcodes of 8.1 (secondary flow 30%–50% to an urban area), 8.2 (secondary flow of 30%–50% to a large urban cluster), 8.3 (secondary flow 10%–30% to an urban area), and 8.4 (secondary flow 10%–30% to a large urban cluster).

For the census-tract EQI, RUCA primary traffic flows were used to construct RUCA-stratified EQIs. The RUCA categories based on primary flows are as follows: 1) urban core area (RUCA code 1); 2) suburban area (RUCA code 2); 3) micropolitan area (RUCA codes 3, 4, 5, 6); 4) small town area (RUCA codes 7, 8, 9); and 5) rural area (RUCA code 10). Loadings on the stratified and non-stratified sets of indices were assessed to determine loading heterogeneity across tracts. Because these loadings differed meaningfully by RUCA level, RUCA-stratified EQIs were constructed for each census tract.

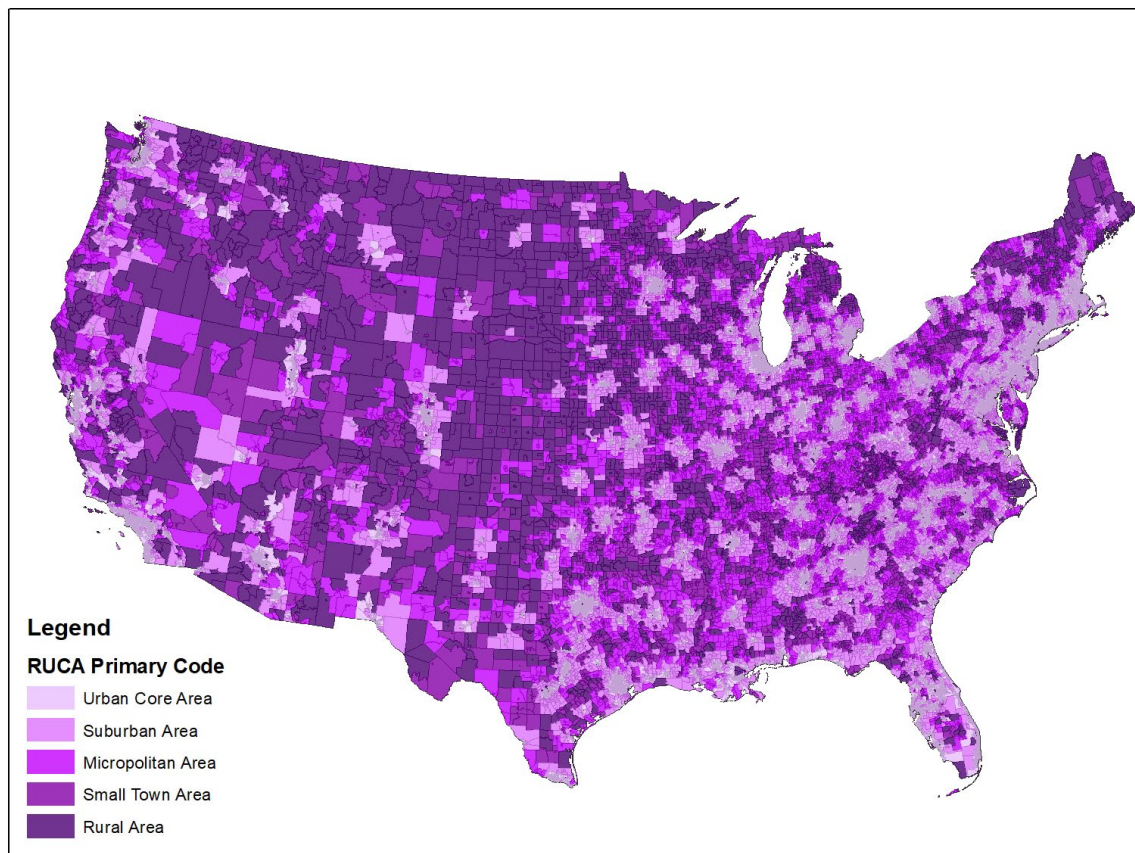


Figure 2. Rural-urban commuting area (RUCA) stratification for all tracts in the United States

The first principal component was the unique linear combination that accounted for the largest possible proportion of the total variability in the component measures. Therefore, the first component was retained as the EQI. This process was undertaken separately for each of the RUCA strata.

References

1. United States Environmental Protection Agency (EPA), *U.S. EPA. Environmental Quality Index - Technical Report (2006-2010) (Final, 2020)*. 2020.
2. United States Environmental Protection Agency (EPA), *CMAQ: The Community Multiscale Air Quality Modeling System*. 2022, United States Environmental Protection Agency (EPA).
3. United States Environmental Protection Agency (EPA), *National Air Toxics Assessment*. 2018, United States Environmental Protection Agency (EPA).
4. Dun & Bradstreet (D&B), *Dun and Bradstreet Products*. 2008.
5. Dun & Bradstreet (D&B), *Dun and Bradstreet Products*. 2013.
6. Homer, C., et al., *Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information*. Photogrammetric Engineering and Remote Sensing, 2015. **81**(5): p. 345-354.
7. United States Census Bureau, *2006-2010 5-year American Community Survey (ACS)*. 2010, United States Census Bureau.
8. United States Census Bureau, *2011-2015 5-year American Community Survey (ACS)*. 2015.
9. United States Census Bureau, *American Community Survey (ACS)*. 2024, United States Census Bureau.
10. United States Environmental Protection Agency (EPA), *Facility Registry Service (FRS)*. 2025.
11. United States Department of Agriculture (USDA), *Cropscape*. 2023, United States Department of Agriculture.
12. Baker, N.T. and W.W. Stone, *Estimated annual agricultural pesticide use for counties of the conterminous United States, 2008–12: U.S. Geological Survey Data Series 907*. 2015, United States Geological Survey (USGS).
13. United States Geological Survey (USGS), *Estimated Annual Agricultural Pesticide Use for Counties of the Conterminous United States, 2013-17 (ver. 2.0, May 2020)*. 2020: United States Geological Survey.
14. Federal Bureau of Investigation, *Uniform Crime Reporting: Crime in the U.S.* 2020.
15. ESRI, *Crime Indexes*. n.d., ESRI.
16. United States Geological Survey (USGS), *Estimated Use of Water in the United States*. 2010.
17. University Nebraska-Lincoln, *U.S. Drought Monitor - Data Download*. 2023.
18. Tabachnick, B.G. and L.S. Fidell, *Using Multivariate Statistics*. 5th ed. 2007: Pearson Allyn & Bacon.
19. United States Environmental Protection Agency (EPA), *EnviroAtlas Green space dataset*. 2017.
20. United States Environmental Protection Agency (EPA), *EnviroAtlas*. 2025.
21. Watkins, M.W., *Exploratory Factor Analysis: A Guide to Best Practice*. Journal of Black Psychology, 2018. **44**(3): p. 219-246.
22. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments*. Philos Trans A Math Phys Eng Sci, 2016. **374**(2065).
23. Hall, S.A., J.S. Kaufman, and T.C. Ricketts, *Defining urban and rural areas in U.S. epidemiologic studies*. J Urban Health, 2006. **83**(2): p. 162-175.
24. United States Department of Agriculture (USDA), *Rural-Urban Continuum Codes*. 2013.