Documentation for RSEI Geographic Microdata (RSEI-GM)

Contents

Documentation for RSEI Geographic Microdata (RSEI-GM)

The RSEI Geographic Microdata are a unique dataset that provides detailed air model results from the RSEI model at a level of 810 meter square grid cells. The Microdata are very large, especially in their disaggregated form, which is more than 100GB in size per year for the entire United States.

RSEI also produces water Microdata, which provides toxicity-weighted concentrations for each stream segment downstream of a water discharge.

Ways to Get RSEI Results provides information on downloading the Microdata.

RSEI Microdata are summarized differently than the RSEI facility-level data

RSEI results are distributed in a compact form in tools like EasyRSEI and in applications like EPA's Envirofacts data warehouse. This facility-level dataset uses the same modeling results as the Microdata, but all of the grid-cell results for each chemical release from each facility are summed. The release[1] is the basic unit of analysis; when you rank states or counties in EasyRSEI, you are actually ranking the sum of all modeled releases (and transfers) from all reporting facilities in each state or county. Some of the releases may impact other states or counties, and some transfers to off-site incinerators may result in releases across the country from the reporting facility. The facility-level dataset is appropriate for ranking facilities, industries, chemicals, and screening for pollution prevention opportunities. Ranking or looking at trends for geographic areas can be useful as long as users understand the facility-based nature of the data.

With the Microdata, geographic analyses are more intuitive; when you rank states, you are ranking the impacts that occur within the geographic confines of each state. Impacts from off-site transfers are attributed to the grid-cells that correspond to the area around the facility that receives the transfer and releases the chemical into the environment. Because the Microdata are provided at the grid-cell level, you can look at small geographic areas, and examine the impacts that releases from multiple facilities may have on the same area.

Microdata users should familiarize themselves with RSEI methods and data

As with any model, RSEI is subject to the limitations of the underlying data sources and models that it incorporates. You should carefully consider the impact that the RSEI method may have on the results for any analysis. RSEI relies on TRI data, which only captures releases from certain industrial facilities; RSEI does not provide any information regarding many sources of environmental risk, including mobile sources, small industrial facilities, and hazardous waste (Superfund) sites.

The RSEI website contains important information about RSEI, and the RSEI documentation and help page contains links to the RSEI methodology and associated documentation.

Outliers should be checked

EPA's TRI program conducts an extensive data quality analysis after TRI reporting forms are received. However, facilities often make revisions to their TRI reporting, especially to the most recent annual report. RSEI data uses the same dataset as the TRI National Analysis, which represents a point in time in the autumn following the July deadline for reports. EPA's Envirofacts TRI search has more current data that should be checked for any revisions to anomalously high release amounts.

[1] More specifically, it is the release pathway; for instance, water releases can be split into different pathways for fish consumption and drinking water exposure.

In addition, some high toxicity-weighted concentrations are driven by reporting of chemical groups, like glycol ethers or various groups of metal compounds like lead compounds or nickel compounds. Because specific data on the types of these compounds are not reported to TRI, RSEI models a worst-case scenario and assumes the most toxic form of the chemical is being released in most of these cases. Exceptions include PACs and chromium and chromium compounds, for which RSEI assumes a more likely release profile based on National Emissions Inventory (NEI) data[2]. In these cases, additional investigation needs to be done to determine the exact form of the chemicals being released before any conclusions about potential risk can be drawn.

Adjustments must be made when doing time-series analysis

When using the Microdata over more than one year, users should take care to ensure that the same set of TRI reporting requirements are in force for the whole period. The Chemical table (Chemical.csv, available in the "data tables" or "Public Release" datasets) has a series of fields to help in picking the appropriate chemical set for the time series; the year in the field name is the beginning of the period for which reporting requirements are constant. For instance, the '98 Core Chemical Flag' field is "1" (or "True") for any chemical whose reporting requirements have been constant over the period 1998-present. Users should limit their time series to those chemicals whose requirements have not changed over the period being considered; otherwise increases in reporting may erroneously imply increases in potential impacts.

Users should also be aware that seven high-scoring industry sectors were added for reporting year 1998, and so if a time series bridges that year, the new industries should be removed. The RSEI Facility table (Facility.csv) also has a True/False field called NewIndustryFlag that is "1" (or "True") for facilities in industries added in 1998. More information on the TRI industry expansion.

Water Microdata

RSEI water microdata are produced at a different level than the air microdata. Instead of the grid cell, the water microdata use the stream segment as the unit of analysis. Before using the water microdata, users should familiarize themselves with the RSEI water modeling methodology and be aware that TRI does not include all chemical discharges to water (see the RSEI website for more information on the TRI universe and where to go for information on other releases).

The RSEI water modeling is indexed by stream segment Common ID (COMID), which is a unique identifier assigned by the NHDPlus data set, Version 1. NHDPlus is based on the USGS's medium resolution National Hydrography Dataset (NHD), and adds important information like stream flow and velocity estimates.

The file "NHDMicroResults_conc_agg.csv" sums up the concentrations by release and COMID. In some cases, a stream path for a single release may split in two and then come together later downstream. In RSEI, those paths are calculated separately, but in this water Microdata file, the separate concentrations and associated metrics are summed (the disaggregated data are available upon request).

Getting Started

There are different kinds of files that contain Microdata results, and over time additional files will likely be produced to help users get the information they need in the easiest way possible.

- **Disaggregated Microdata files**- These are the raw Microdata files that contain the most disaggregated data possible. For each 810m grid cell, the file contains scores, concentrations, and tox-weighted concentrations for each chemical release. There may be multiple records for any one grid cell.Note that if two releases for the same chemical (either from different facilities or one from a stack release and one from a fugitive release from the same facility) affect the same grid cell, there will be separate records for each grid release. **Naming:** These annual files have historically been named MicroXXXX_YYYY, where

[2] The RSEI toxicity topic page has more information.

XXXX is the reporting year for the data freeze, and YYYY is the year of the data contained in the file. So Micro 2014_2010 is from the RY2014 RSEI update, and contains data for chemicals released in 2010. The **new naming convention** substitutes the version number for the version year, as in vXXX_micro_YYYY, where XXX is the version number and YYYY is the year of the data contained in the file; for example v234_micro_2014.csv.There is one annual file for the entire country, which is over 100 GB in size. See the Disaggregated Microdata table below for field names.

- **Aggregated Microdata files (Grid Cell files)**- Aggregated Microdata files use the same data as the disaggregated files, but sum the chemical releases over each grid cell. Because the values are summed, unweighted concentrations are not available (the sum of the concentrations of different chemicals would be meaningless). **Naming:** These annual files have historically been named MicroXXXX_YYYY, where XXXX is the reporting year for the data freeze, and YYYY is the year of the data contained in the file. So Micro 2014_2010 is from the RY2014 RSEI update, and contains data for chemicals released in 2010. The **new naming convention** substitutes the version number for the version year, as in vXXX_micro_YYYY, where XXX is the version number and YYYY is the data year; for example v234_micro_2014.csv.These files have historically been named in the format AggMicroXXXX_YYYY_GCZZ, where XXXX is the reporting year for the data freeze, YYYY is the year of the data contained in the file, and ZZ is the 2-digit grid code (see Field 1 in the Table 1 below for grid codes). The **new naming convention** substitutes the version number for the version year, as in vXXX_aggregated_micro_gcZZ_YYYY; for example, v234_aggregated_micro_gc14_2014.csv.
- **Block Group Microdata files**- These files are the same as the aggregated Microdata files, but instead of being presented at the grid cell level, the values are averaged over Census block groups. The file BG_RSEI_XXXX_3yr is a csv file with the block group-level data averaged over 2012 through 2014. There are also shape files (tl_2010_bg_US_RSEI) with the same data; that is, the .dbf file and the .csv have the same fields.
- **Water Microdata files-** This file contains the toxicity-weighted concentrations downstream of TRI discharges by stream segment. All years of data are contained in the file, which is named NHDMicroResults_conc_agg_XXXX, where XXXX is the reporting year of the data freeze.
- **Census Crosswalk Files-** Each set of crosswalk files links the RSEI grid cell geography to a different US decennial census year. There is one crosswalk for each area and decennial Census year (1990, 2000, 2010). Crosswalk files are named by area (Alaska, Con(terminous) US, etc.). The last three fields in each file contain percent values that can be used to adjust the block or cell contents when performing the crosswalk. PCT_B_C and PCT_C_B are area-weighted and can be used for metrics that do not involve population, such as concentration and toxicity-weighted concentration. PCT_PC_B is population weighted, and can be used to crosswalk fields that involve population, like score and pop. Note that the "PCT_CP_B" field is not available for the territories (VI, PR, GU, AS, MP). The Northern Mariana Islands are in the Guam file and the Virgin Islands are in the Puerto Rico file. There are no crosswalks for Puerto Rico, the Virgin Islands, Mariana Islands, Guam, or American Samoa for 1990. For these areas, RSEI uses 2000 block boundaries and scales each cell's population by the overall ratio of 1990/2000 population for each area.
- **Public Data Release Files (data tables)-** These tables contain TRI reporting data and other data used in modeling, including facility locations and chemical toxicity. Figure 1 below shows the relationships between the main public data release tables. The RSEI Data Dictionary provides field descriptions.
- **Shape files-** Shape files of the RSEI grid geography are available by region, in two versions: polygon and center point. The RSEI Data Dictionary provides field descriptions.

Linking Tables

The Microdata must be used together with the other data tables provided with the RSEI public data release of the same vintage. If you are using the all-years RY2014, Version 2.3.4 Microdata, you must use the data tables from the Version 2.3.4 release, which are also provided on the same sites as the Microdata, under "data tables" or "public release."

If you are using the three-year dataset from RY2015 (V 2.3.5 Microdata), you must use the data tables from Version 2.3.5, which are also provided on the RSEI ftp site. Each table links into the Microdata using a key field (like ChemicalNumber, FacilityNumber, etc.), as indicated in Figure 1, below. Field descriptions can be found in the RSEI Data Dictionary. The public data release files allow you to extract parameters like chemical name, CAS number, facility name, TRI facility ID (TRIFID) and any other associated data contained in the RSEI databases.

Geographies

RSEI Microdata are presented at the grid-cell level, but can be transposed onto Census block geographies using the Census crosswalks. It is important to consider the nature of the metric being used when changing geographies.

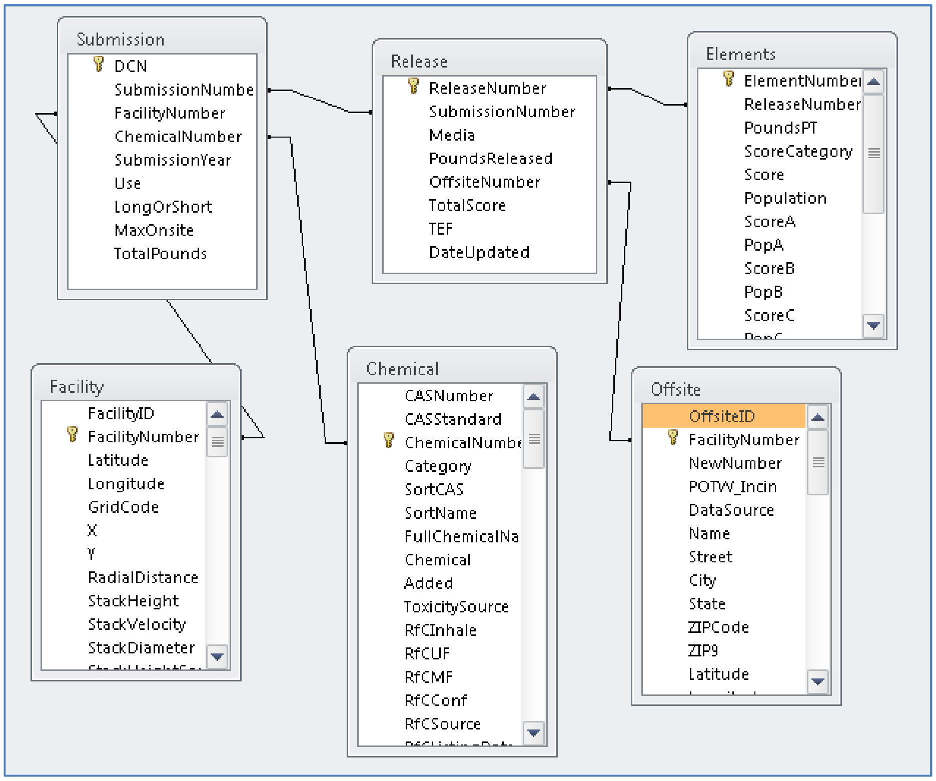
When using the crosswalk from grid cells to blocks:

- Concentration and toxicity-weighted concentration are like levels that describe the entire grid cell. When crosswalking from grid cell to block, the concentrations at the block level should be expressed as an area-weighted average, using the PCT_C_B field in the Crosswalk table as the weight. When performing the calculation, multiply each value (toxicity-weighted concentration or concentration) by the size of the grid cell (656,100m^2) and the PCT_C_B value. Sum the rows over each block ID and divide the resulting values by the area of each block (not provided, but available from the Census).
- Score and population are additive- the score of a block is the sum of the scores of the grid cells composing the block. But because the grid cell populations were originally transposed from the blocks that have varying population densities, to crosswalk back to the block, you need to use a population weight rather than an area weight in cases where a cell is not completely within a block. PCT_CP_B is provided in the crosswalk for this purpose.

When aggregating from census blocks to other census geographies:

- Population and scores can be summed across lower geographies and rolled up (i.e., from blocks to block groups or census tracts).
- Concentrations and toxicity-weighted concentrations should be expressed as an area-weighted average.

Figure 1. Relationships between public data release tables



Note: Aggregated microdata and crosswalks are in separate files for each grid; other files have a field describing the grid. Grid codes are as follows:

- **14=Conterminous US**
- **24=Alaska**
- **34=Hawaii**
- **44=Puerto Rico/Virgin Islands**
- **54=Guam/Marianas**
- **64=American Samoa**

RSEI recent version numbers and TRI reporting years (RY)

- V2.3.1- RY 2010
- V2.3.2- RY 2011
- V2.3.3- RY 2012
- V2.3.4- RY 2014
- V2.3.5- RY 2015

Appendix 1. Field Descriptions for Microdata Tables

These tables, along with the tables for the RSEI facility-level data, can also be found in the RSEI Data Dictionary.

Disaggregated Microdata Table

Field Number	Name	Description
1	GridCode	Identifies grid. . 14=Conterminous US 24=Alaska 34=Hawaii 44=Puerto Rico/Virgin Islands 54=Guam/Marianas 64=American Samoa
2	X	X-coordinate of grid cell
3	Y	Y Coordinate of grid cell
4	ReleaseNumber	Internal unique identifier for release (lookup in table “Release”)*
5	ChemicalNumber	Internal unique identifier of released chemical (lookup in table “Chemical”)*
6	FacilityNumber	Internal unique identifier of releasing facility (lookup in table “Facility” if media = 1 or 2; if media = 6 or 750 or 754, then lookup in table “Offsite”)*
7	Media	Code describing media into which chemical is released. (lookup in table “Media”)*
8	Conc	Concentration of chemical for release/media at grid cell (µg/m^3)
9	ToxConc	Concentration multiplied by inhalation toxicity weight.
10	Score	Risk-related score (surrogate dose * toxicity weight * population)
11	ScoreCancer	Risk-related score (surrogate dose * toxicity weight * population) using only toxicity values for cancer effects
12	ScoreNonCancer	Risk-related score (surrogate dose * toxicity weight * population) using only toxicity values for noncancer effects
13	Pop	Number of people in grid cell (may be interpolated)

Aggregated Microdata Table

Field Number	*Name*	*Description*
1	X	X-coordinate of grid cell.
2	Y	Y Coordinate of grid cell
3	NumberOfFacilities	Number of facilities with releases affecting grid cell.
4	NumberOfReleases	Number of individual releases affecting grid cell.
5	NumberOfChemicals	Number of chemicals with nonzero concentrations for grid cell.
6	ToxConc	Concentration multiplied by inhalation toxicity weight, summed over all chemicals impacting cell.
7	Score	Risk-related score (surrogate dose * toxicity weight * population), summed over all chemicals impacting cell.
8	Pop	Population of grid cell.
9	ScoreCancer	Risk-related score (surrogate dose * toxicity weight * population) using only toxicity values for cancer effects.
10	ScoreNonCancer	Risk-related score (surrogate dose * toxicity weight * population) using only toxicity values for noncancer effects.

Averaged Block Group Microdata

Field Number	*Name*	*Description*
1	GEOID10	US Census Block Group ID.
2	ALAND10	Land area of the block group (m^2).
3	AWATER10	Water area of the block
4	TOXCONC	Average toxicity-weighted concentration of the cells in the block group, averaged over three years.
5	PTOXCONC	Percentile associated with field TOXCONC.
6	SCORE	Sum of the risk-related score (surrogate dose * toxicity weight * population) of the cells in the block group, averaged over three years.
7	PSCORE	Percentile associated with field SCORE.
8	NCSCORE	Sum of the risk-related scores (surrogate dose * toxicity weight * population) of the cells in the block group, averaged over three years. Score is calculated using only noncancer toxicity weights.
9	PNCSCORE	Percentile associated with field NCSCORE.
10		Sum of the risk-related scores (surrogate dose * toxicity weight * population) of the cells in the block group, averaged over three years. Score is calculated using only cancer toxicity weights.
11	PCSCORE	Percentile associated with field CSCORE.
12	POP	Sum of the population of the cells in the block group, averaged over three years.

Averaged Block Group Microdata

Field Number	Name	Description
13	PPOP	Percentile associated with field POP.
14	COVERED	Internal field.
15	FOUND	
16	GC	Grid code.

Water Microdata

Field Number	Name	Description
1	ReleaseNumber	Internal unique identifier for release (links to Release table).
2	Counter	Auto-increment count of COMIDs
3	ComID	"Common Identifier" of a flowline (sub-segment of a reach)- atomic unit of reach data that matches one-to-one to NHD.
4	ReachCode	Code for reach.
5	Conc	Concentration of chemical in flowline (mg/L).
6	Sequence	Number defining pathway of release (used to indicate branching).
7	TravelTime	Time(s) for release to go from top of flowline to bottom.
8	TravelLength	Distance (m) for release to go from top of flowline to bottom.
9	Paths	Number of branches in stream path.
10	FCode	Descriptor from NHD for type of flowline (e.g., pipeline, stream)
11	ResCode	Internal code.

Census Crosswalk Table

Field Number	Name	Description
1	GridID	Identifies grid. 14=Conterminous US 24=Alaska 34=Hawaii 44=Puerto Rico/Virgin Islands 54=Guam/Marianas 64=American Samoa
2	X	X coordinate of the cell address.
3	Y	Y coordinate of the cell address
4	Block_ID00	US Census Block ID.
5	UR	Internal
6	PCT_B_C	Percent of the Census block that is within the cell (Block to Cell).

Census Crosswalk Table

Field Number	*Name*	*Description*
7	PCT_C_B	Percent of the cell that is within the Census block (Cell to Block).
8	PCT_CP_B	Percent of the cell's population that is within the Census block (Population-Cell to Block).

[updated 1/14/2017]