**Harmonized E-PRTR and TRI Database**

With Orbis Identifiers and Toxicity-Weights.

Jan P. Hackmann1, Paul P. Momtaz2 \*

**Affiliations:**

1 TUM School of Management at the Technical University of Munich; Munich, 80333, Germany

2 Martin J. Whitman School of Management at the Syracuse University; Syracuse, NY 13244, USA

\*Corresponding author. Email: [ppmomtaz@syr.edu](mailto:ppmomtaz@syr.edu)

**Variables**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| PRTR\_FacilityID | A unique facility identifier derived from the emissions datasets (the Toxics Release Inventory (TRI) for U.S. facilities and the European Pollutant Release and Transfer Register (E-PRTR) for European facilities). |
| Orbis\_CompanyID | A unique Orbis identifier assigned to the company that owns the facility in the specified year. |
| Orbis\_CorporateID | A unique Orbis identifier assigned to the corporation that owns the company in the specified year. |
| Facility\_GeometryLat | The latitude coordinate of the facility’s geographic location. |
| Facility\_GeometryLon | The longitude coordinate of the facility’s geographic location. |
| NUTS3 | A five-digit regional identifier based on the Nomenclature of Territorial Units for Statistics (NUTS) classification system, representing in which NUTS-3 (smallest administrative units in the European Union used for statistical analysis) the facility is located. |
| County\_FIPS | A five-digit Federal Information Processing Standards (FIPS) code that uniquely identifies the U.S. county in which the facility is located. |
| Facility\_Industry | A unique facility-level industry identifier following the International Standard Industrial Classification (ISIC) Sections, transformed from NACE in E-PRTR and NAICS in TRI. A limited number of missing industry classifications is filled from the Orbis company data. |
| EU | A binary indicator equal to 1 if a facility is located in Europe (E-PRTR data), otherwise equal to 0 if the facility is located in the United States (TRI data). |
| Year | The calendar year in which the emissions are reported. |
| Emission\_Medium | The environmental medium to which the pollutant is released (air, water, or land). |
| Pollutant\_QuantityKg | The quantity (in kilograms) of a specific pollutant emitted to a particular medium from a unique facility in a given year. |
| Pollutant\_QuantityLbs | The quantity (in pounds) of a specific pollutant emitted to a particular medium from a unique facility in a given year. |
| EU\_PRTR\_Chemical\_Name | The pollutant name as reported in the E-PRTR dataset. |
| US\_TRI\_Chemical\_Name | The pollutant name as reported in the TRI dataset. |
| US\_TRI\_CAS\_Number | The Chemical Abstracts Service (CAS) Registry Number for the pollutant reported in the TRI dataset. |
| EU\_OECD\_Chemical\_Name | The pollutant name in the E-PRTR dataset according to the harmonized OECD PRTR pollutant nomenclature. |
| US\_OECD\_Chemical\_Name | The pollutant name in the TRI dataset according to the harmonized OECD PRTR pollutant nomenclature. |
| OECD\_Chemical\_Name | The standardized pollutant name from the harmonized OECD PRTR pollutant harmonizer. |
| OECD\_CAS\_Number | The CAS Registry Number or an OECD-assigned identifier for the pollutant, as provided in the harmonized OECD PRTR pollutant nomenclature. (For chemical groups identifiers are generated by OECD and are not standard CAS numbers.) |
| EU\_PRTR\_Threshold | The pollutant- and medium-specific E-PRTR reporting threshold (in kilograms.) |
| TACE | Toxicity Accounted Chemical Emissions (TACE) are non-greenhouse gas emissions, from the TRI and E-PRTR dataset, released into air, land, and water and normalized to reflect their relative toxicity. This normalization is achieved by dividing the quantity of each emitted chemical per medium (Pollutant\_QuantityKg) by its respective reporting threshold as defined in EU\_PRTR\_Threshold. |
| USEtox\_CAS\_Number | The CAS Registry Number for pollutants with established toxicity factors in the USEtox model. |
| EU\_UrbanRural | A binary indicator equal to 1 if an EU NUTS3 region is classified as “predominantly urban” according to Eurostat standards. This classification requires that over 80% of the region’s population resides in urban clusters, defined as contiguous 1 km² grid cells with at least 300 inhabitants per km² and a total of at least 5,000 residents; otherwise, the indicator is 0. |
| US\_UrbanRural | Binary indicator equal to 1 if an U.S. county is classified as "urban" following similar standards as Eurostat. Our classification requires that over 80% of the counties population resides in an "urban area" accoding to the US 2020 Census. According to the US 2020 Census an area is considered urban if it has at least 2,000 housing units or at least 5,000 residents; otherwise, the indicator is 0. |
| USEtox\_Factor\_UrbanAir | The USEtox characterization factors (CFs) for emissions to urban air. Emissions are classified as urban air if EU\_UrbanRural or US\_UrbanRural are euqal to 1. |
| USEtox\_Factor\_RuralAir | The USEtox CFs for emissions to rural air. Emissions that do not meet the criteria for “urban air” are considered “rural air.” |
| Nearest\_Coast\_Meter | The distance (in meters) from the emitting facility to the nearest coastline. |
| USEtox\_Factor\_SeaWater | The USEtox CFs for emissions to sea water. Emissions are classified as sea water discharges if the facility is within 500 meters of a coastline and the pollutant is discharged to water. |
| USEtox\_Factor\_FreshWater | The USEtox CFs for emissions to freshwater. Emissions to water sources that are not classified as sea water discharges are considered freshwater discharges. |
| USEtox\_Factor\_NaturalSoil | The USEtox CFs for emissions to natural soil. |
| USEtox\_Factor\_AgricSoil | The USEtox CFs for emissions to agricultural soil. |
| TACE\_USEtox | Toxicity-Accounted Chemical Emissions (TACE) using the USEtox model, calculated by multiplying the emission quantity (kg) by the corresponding USEtox CFs. |
| RSEI\_CAS\_Number | The CAS Registry Number for pollutants with toxicity weights in the Risk-Screening Environmental Indicators (RSEI) model. |
| RSEI\_RfC\_ToxWeight | The reference concentration (RfC)–based RSEI toxicity weight, representing an inhalation exposure (mg/m³) that is unlikely to pose appreciable risk of adverse health effects over a lifetime. |
| RSEI\_IUR\_ToxWeight | The inhalation unit risk (IUR)–based RSEI toxicity weight, denoting the upper-bound lifetime cancer risk from continuous inhalation exposure at 1 μg/m³. It is used to estimate cancer risk by converting exposure concentrations into dose-based risk assessments. |
| RSEI\_RfD\_ToxWeight | The reference dose (RfD)–based RSEI toxicity weight, indicating the daily oral intake (mg/kg-day) unlikely to produce adverse effects over a lifetime. |
| RSEI\_OSF\_ToxWeight | The oral slope factor (OSF)–based RSEI toxicity weight quantifying the incremental lifetime cancer risk per unit oral dose (mg/kg-day). Represents an upper-bound estimate of the low-dose slope of the dose-response curve for carcinogens. |
| TACE\_RSEI | Toxicity-Accounted Chemical Emissions (TACE) using RSEI toxicity weights, reflecting non-greenhouse gas emissions (in lbs) normalized by multiplying the emitted quantity of each pollutant by its respective RSEI toxicity weight. |

**Input files**

1. **OECD’s Harmonised List of Pollutants for Global PRTRs:**

The “OECD’s Harmonised List of Pollutants for Global Pollutant Release and Transfer Registers” helps standardize chemical reporting across various Pollutant Release and Transfer Register (PRTR) systems, enhancing data comparability at a global level. Originally introduced by the OECD’s Chemicals and Biotechnology Committee in 2014 and updated in 2022 (using data from the 2017 E-PRTR and 2019 TRI), this resource lists pollutants covered by different national PRTRs. We obtained the most recent version from the OECD in July 2024 (*1*) and used it to identify the pollutants shared by E-PRTR and TRI. Each record in this dataset corresponds to a single pollutant listed in at least one PRTR.

1. **European Pollutant Release and Transfer Register:**

The E-PRTR (version 9), managed by the European Environment Agency (EEA), contains annual industrial emissions to air, water, and land for facilities exceeding specific threshold values, as mandated by the Industrial Emissions Directive (2010/75/EU) and Regulation (EC) No 166/2006. The data encompass 65 industrial activities, including large installations and intensive livestock farms, to support regulatory compliance and public transparency. We focused on three core tables (2\_ProductionFacility, 2f\_PollutantRelease, 2c\_Function) from 2007–2021. Each observation represents a single reported emission of one pollutant to a specific medium at one facility. We obtained the data directly from the EEA website in January 2024 (*2*).

1. **Toxics Release Inventory:**

The TRI Basic Plus Data Files (Type 1A) from the U.S. Environmental Protection Agency (EPA) provide details on facility-level chemical releases. Reporting requirements under EPCRA Section 313 apply to facilities with 10+ full-time employees that manufacture or process certain toxic chemicals above defined thresholds. We merged annual Type 1A files from 2007–2021 into a single dataset. Each observation represents a single reported emission of one pollutant to a specific medium at one facility. We obtained the data directly from the EPA TRI website in January 2024 (*3*).

1. **LinkTree OECD E-PRTR**

To align the OECD’s E-PRTR pollutant names with the actual pollutant names in the E-PRTR dataset, we created a manual “link tree” file. This resource standardizes naming conventions and ensures one-to-one matches for all E-PRTR pollutants referenced in the OECD file, which are (per the OECD list) also covered by TRI.

1. **LinkTree OECD TRI**

Similarly, we generated a manually created file to reconcile the OECD’s TRI pollutant names with those used in the actual TRI dataset. Only pollutants also present in the E-PRTR (per the OECD list) were included.

1. **RSEI Toxicity Weights**

The U.S. EPA developed the Risk-Screening Environmental Indicators (RSEI) model to generate continuous indices of potential human health risks from facility-level toxic chemicals. The model integrates data on chemical identities, emission quantities, fate and transport factors, and chronic human toxicity (cancer and noncancer) to estimate comparative risk-level contributions across facilities or regions. To facilitate cross-national comparisons of toxicity-weighted emissions, we used the EPA’s RSEI framework only the RSEI-published toxicity weights of the different pollutants in different media (air, surface water, and on-site land). These toxicity weights reflect standardized assumptions for metals, metal compounds, dioxins, and chemical categories. We obtained the RSEI toxicity weights data directly from the EPA website in May 2024 (*4*).

1. **USEtox Model**

We applied USEtox (version 2.13) characterization factors (CFs) for human health toxicity to convert reported emissions into toxicity potentials. USEtox CFs consider fate, exposure, and effect factors for each chemical, providing standardized toxicity metrics per kilogram emitted. We downloaded the USEtox CF data in February 2024 (*5*). Drawing on prior research (*6*), we refine CFs by distinguishing between urban and rural air, as well as seawater and freshwater, to improve accuracy:

* **Distance to Coast:** 
  + A 500 m threshold from the coastline (*7*) differentiates seawater from freshwater.
* **Urban/Rural Classification:**
  + Europe: NUTS 3 regions classified as “predominantly urban” if ≥80% of the population resides in contiguous high-density clusters. Data from Eurostat (*8*).
  + United States: Counties designated as urban if ≥80% of the population lives in a census-defined urban area. Data from the American Community Survey (*9*).

## **Data Set Generation**

## This document outlines the data processing workflow and toxicity analysis framework used to integrate emissions data from OECD, E-PRTR, and TRI with toxicity weighting from E-PRTR reporting thresholds, RSEI toxicity weights, and the USEtox model. The end goal is to produce standardized and comparable toxicity-adjusted emissions (TACE) for further analysis at the facility level. Both the E-PRTR and TRI datasets had already undergone preliminary processing prior to the data manipulation described below. Specifically, the datasets were restricted to facilities for which corresponding companies could be identified in Orbis and for which information on facility industry affiliation and other relevant criteria was available.

## **Data Processing Workflow**

## **Loading and Filtering the OECD Data**

## **Load the OECD Excel File**

## Import the OECD .xlsx file into a DataFrame (*OECD*).

## **Filter by Relevant Pollutants**

## Exclude any pollutants that are, according to the OECD, not covered, at least partly, by the E-PRTR and the TRI datasets. Retain only pollutants classified as:

## Covered

## Not individually reported but included in a group listing

## A category or group fully covered by multiple listings

## **Adding OECD Data to E-PRTR Data**

## **Load the E-PRTR Dataset** Import the E-PRTR emissions dataset into a data frame (*E-PRTR*).

## **Document Unique Pollutant Names** Extract the unique pollutant names in the *E-PRTR* dataset and save them to a .xlsx file to create a "link tree" matching E-PRTR pollutant names to the OECD naming conventions.

## **Load the E-PRTR–OECD Link Tree** Import the manually generated link tree that standardizes E-PRTR pollutant names to OECD naming.

## **Merge Link Tree with OECD** Combine the *OECD* DataFrame with the link tree DataFrame to propagate standardized names.

## **Merge with E-PRTR** Merge the enriched *OECD* DataFrame with the *E-PRTR* dataset using:

## EU\_PRTR\_Chemical\_Name (from *E-PRTR*)

## EU\_OECD\_Chemical\_Name (from *OECD*) Apply an inner join to retain only matched chemical names in both datasets.

## **Adding OECD Data to TRI Data**

## **Load the TRI Dataset** Import the TRI dataset into a data frame (*TRI*).

## **Document Unique Pollutant Names** Extract unique TRI pollutant names and save them to a .xlsx file to create a "link tree" matching TRI pollutant names to the OECD naming conventions.

## **Load the TRI–OECD Link Tree** Import the manually generated link tree that standardizes TRI pollutant names to OECD naming.

## **Merge Link Tree with OECD** Combine the *OECD* DataFrame with the link tree DataFrame to propagate standardized names.

## **Merge with TRI** Merge the enriched *OECD* DataFrame with the *TRI* dataset using:

## US\_TRI\_Chemical\_Name (from *TRI*)

## US\_OECD\_Chemical\_Name (from *OECD*) Apply an inner join to retain only matched chemical names in both datasets.

## **TRI Unit Conversion**

## Emission quantities in TRI may be reported in various units (e.g., metric tons, grams). Apply a standard conversion to pounds to ensure consistent units across all datasets.

## **Toxicity Analysis: TACE**

## **Merging E-PRTR Reporting Threshold Data**

## E-PRTR reporting thresholds were added manually into the OECD PRTR chemical harmonization file.

## For TRI pollutants not explicitly mentioned but included in the same chemical group, the maximum E-PRTR reporting threshold in that group is used.

## Based on the chemical name and medium, the corresponding E-PRTR reporting threshold in kg (EU\_PRTR\_Threshold) is added to the respective emissions in the *E-PRTR* and *TRI* DataFrames.

## **Computing TACE**

## **Handling Zero or Missing Thresholds**

## Thresholds with a value of zero are replaced with a small constant (0.001) to prevent division by zero and to ensure the variable remains continuous (a zero threshold indicates all emissions must be reported, implying high toxicity.)

## **Threshold-Based Calculation** TACE is computed by dividing the emission quantity in kilograms (Pollutant\_QuantityKg) by the pollutant-specific threshold (E-PRTR Threshold).

## **Store the Result** Save the resulting toxicity-adjusted emission in TACE.

## **Toxicity Analysis: RSEI**

## **Integrating RSEI with E-PRTR**

## **Load RSEI Toxicity Scores** Import the RSEI toxicity scores from a .xlsx file into a DataFrame (*RSEI*).

## **Merge with E-PRTR** Merge *E-PRTR* with *RSEI* using:

## OECD\_CAS\_Number (from *E-PRTR*)

## RSEI\_CAS\_Number (from *RSEI*) *Note:* *E-PRTR* uses OECD\_CAS\_Number due to the lack of a dedicated CAS identifier in the original E-PRTR dataset.

## **Convert E-PRTR Quantities** Convert *E-PRTR* emission quantities from kilograms (Pollutant\_QuantityKg) to pounds (Pollutant\_QuantityLbs) for consistency.

## **Integrating RSEI with TRI**

## **Load RSEI Toxicity Scores** Again, use the *RSEI* dataset as above.

## **Merge with TRI** Merge *TRI* with *RSEI* using:

## US\_TRI\_CAS\_Number (from *TRI*)

## RSEI\_CAS\_Number (from *RSEI*)

## **Computing TACE-RSEI**

## **Apply the Toxicity Factor** Multiply the emission quantity (in pounds, Pollutant\_QuantityLbs) by the corresponding RSEI toxicity weights. The specific weight depends on the emission medium:

## *Air Emissions*: Use the maximum of RSEI\_RfC\_ToxWeight or RSEI\_IUR\_ToxWeight.

## *Water and Land Emissions:* Use the maximum of RSEI\_RfD\_ToxWeight or RSEI\_OSF\_ToxWeight.

## **Store the Result** Save the resulting RSEI toxicity-adjusted emission as TACE\_RSEI.

## **Toxicity Analysis: USEtox**

## **Integrating USEtox with E-PRTR**

## **Load USEtox Characterization Factors** Read inorganic and organic CFs for human toxicity from USEtox .xlsx files into a DataFrame (*USEtox*).

## **Merge with E-PRTR** Merge *E-PRTR* with *USEtox* using:

## OECD\_CAS\_Number (from *E-PRTR*)

## USEtox\_CAS\_Number (from *USEtox*)

## **Incorporate Urban/Rural Attributes**

## Convert facility coordinates (Facility\_GeometryLat and Facility\_GeometryLon) into a GeoDataFrame and perform a spatial join with the NUTS3 administrative boundaries.

## Merge the NUTS3-level EU\_UrbanRural variable (from a Eurostat land-use dataset) into the *E-PRTR* DataFrame.

## **Integrating USEtox with TRI**

## **Load USEtox Characterization Factors** As above, read inorganic and organic CFs for human toxicity from the USEtox dataset (*USEtox*).

## **Merge with TRI** Merge *TRI* with *USEtox* using:

## US\_TRI\_CAS\_Number (from *TRI*)

## USEtox\_CAS\_Number (from *USEtox*)

## **Incorporate Urban/Rural Attributes**

## Convert facility coordinates (Facility\_GeometryLat and Facility\_GeometryLon) into a GeoDataFrame and perform a spatial join with county (or other relevant) boundaries in the US.

## Merge the relevant US Census land-use dataset (which contains the US\_UrbanRural binary variable) into the *TRI* DataFrame for the corresponding county.

## **Computing TACE-USEtox**

## **Air Emissions** If EU\_UrbanRural or US\_UrbanRural equals 1 (urban), multiply the emission quantity (in kg, Pollutant\_QuantityKg) by USEtox\_Factor\_UrbanAir; otherwise, use USEtox\_Factor\_RuralAir.

## **Water Emissions** Check Nearest\_Coast\_Meter. If < 500, multiply Pollutant\_QuantityKg by USEtox\_Factor\_SeaWater; otherwise, use USEtox\_Factor\_FreshWater.

## **Land Emissions** Compare USEtox\_Factor\_NaturalSoil and USEtox\_Factor\_AgricSoil. Use the higher factor (worst-case) to multiply by Pollutant\_QuantityKg.

## **Store the Result** The final USEtox toxicity-adjusted emission is stored in TACE\_USEtox.

## **Further, Use**

## All three toxicity metrics—TACE, TACE\_USEtox, and TACE\_RSEI—are computed row-by-row across the dataset. These facility-specific emissions disaggregated by pollutant and release medium can be aggregated at the facility level:

## Sum TACE, TACE\_USEtox, and TACE\_RSEI scores per facility per year to get annual facility-level toxicity accounted chemical emissions scores.

## This aggregation yields facility-level toxicity indicators that support robust comparisons of emissions across facilities while accounting for differences in pollutant toxicity.

**References**

1. Organization for Economic Co-operation and Development (OECD). (2022). Harmonized List of Pollutants for Global Pollutant Release and Transfer Registers (PRTRs) (OECD Series on Prevention and Control of Pollutant Releases). OECD Publishing. <https://doi.org/10.1787/39657758-en>
2. European Environment Agency. (2023). Industrial reporting under the Industrial Emission Directive 2010/75/EU and European Pollutant Release and Transfer Register Regulation (EC) No 166/2006 – Database version 9 [Data set]. European Environment Agency. <https://sdi.eea.europa.eu/data/3da7d329-beea-4a7b-89bc-d45fc1c4b8ac>
3. U.S. Environmental Protection Agency. (2024). TRI Basic Plus data files: Calendar years 1987–present. U.S. Environmental Protection Agency. <https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-plus-data-files-calendar-years-1987-present>
4. U.S. Environmental Protection Agency. (2024, March). RSEI Toxicity Weighting Spreadsheet v.2.3.12. <https://www.epa.gov/rsei/rsei-toxicity-data-and-calculations>
5. USEtox Team. (2024). USEtox (corrective release 2.13) [Software]. UNEP/SETAC Life Cycle Initiative. <https://usetox.org/model/download/usetox2.13>
6. Erhart, S., & Erhart, K. (2023). Environmental ranking of European industrial facilities by toxicity and global warming potentials. Scientific Reports, 13(1), 1772. https://doi.org/10.1038/s41598-022-25750-w
7. OpenStreetMap contributors. (2024). Coastlines [Data set]. OpenStreetMap Foundation. <https://osmdata.openstreetmap.de/data/coastlines.html>
8. European Commission. (2024). *NUTS 2024 classification* [Data file]. Eurostat. <https://ec.europa.eu/eurostat/web/nuts/overview>
9. U.S. Census Bureau. (2021). County-level urban and rural information for the 2020 Census [Data file]. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html>