SAS Statistics

Carmen Azorín Marti
Sara Cerro Torres

Toxic chemicals

20th April 2024

Toxic Release Inventory

Laws have been in effect for more than 30 years that make us aware of what toxic chemicals we may be exposed to in our daily lives. Industrial facilities are required to report to the EPA's Toxics Release Inventory on a yearly basis .

The TRI can let you know which industrial facilities near you are emitting toxic substances into the air, water, and soil, which substances they're emitting, and how much. It can also tell you what pollution prevention methods are in place and which facilities are reducing their emissions.

There is a search tool that provides information on database data elements (columns), tables and subject areas available in the TRI database. This tool can be found at the link https://www.epa.gov/enviro/tri-customized-search, which has helped us to find different research areas of interest to us, as well as downloading CSV files. Some of the research areas are:

- Release Information for Water, Air, Surface and Underground Injection
- Source Reduction Quantity

- Source Reduction Activities: energy recovery from burned chemical waste, recycling of toxic chemicals, treated (destruction) of toxic substances

Research proposal

This proposal aims to investigate relationships within TRI data, focusing on release information, source reduction quantities and activities while incorporating spatial analysis based on facility locations.

Research Questions:

1. How do release estimation amounts vary across different environmental media (water, air, surface, underground injection) within TRI data?
2. Are there specific types of quantity reduction activities (energy production, recycling, toxic chemicals reduction) that show stronger associations with lower release estimation amounts?

Subject Area Data Files and Variables:

1. Release Information for Water, Air, Surface, and Underground Injection:
    - Data File: `release_info`
    - Variables of Interest:
        - `facility_id`: Identifier for reporting facility
        - `release_estimate_amount`: The release estimate (in pounds) reported by the facility
        - `environmental_medium`: Code indicating the environmental medium to which the toxic chemical is released from the facility (water, air, surface, underground injection).

Fourteen different environmental mediums are covered. You can find them in the following link https://enviro.epa.gov/enviro/ef_metadata_html.tri_page?p_column_name=ENVIRONMENTAL_MEDIUM

2. Source Reduction Information:
    - Data File: `reduction_quantity`
    - Variables of Interest:
        - `facility_id`: Identifier for reporting facility
        - `recyc_onsite_qty`: The total amount (in pounds) of the toxic chemical recycled onsite during the calendar year (January 1 - December 31) for which the report was submitted
        - `treated_onsite_qty`: The total amount (in pounds) of the toxic chemical treated onsite during the calendar year (January 1 - December 31) for which the report was submitted
        - `energy_onsite_qty`: The total amount (in pounds) of the toxic chemical in waste burned for energy recovery onsite during the calendar year (January 1 - December 31) for which the report was submitted

Let's get to work: SAS Programming

In order to start programming in SAS, it is necessary to download the CSV files from the official EPA website. To do this we can go to the following link: https://www.epa.gov/enviro/tri-customized-search, where they explain the necessary steps to download the CSV files of the thematic areas we are most interested in.

Once we have downloaded the files, we upload them to our SAS project folder. Afterwards, we will use the following code to convert the CSV files into tables to work with in SAS.

```
LIBNAME lib "/home/u63792450/sasuser.v94/SAS statistics/Project";

PROC IMPORT OUT=lib.quantity_reduction
            DATAFILE="/home/u63792450/sasuser.v94/SAS
statistics/Project/reduct_qty.CSV"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
RUN;

PROC IMPORT OUT=lib.release_info DATAFILE="/home/u63792450/sasuser.v94/SAS
statistics/Project/estimateRelease.CSV"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
RUN;
```

To analyse our data we must combine our two datasets using the industrial facility identifier. The datasets have to be sorted by this column.

```
PROC SORT DATA=lib.quantity_reduction;
BY tri_facility_id;
RUN;

PROC SORT DATA=lib.release_info;
      BY tri_facility_id;
RUN;

DATA lib.combined_data;
      MERGE lib.quantity_reduction (keep=tri_facility_id energy_onsite_qty
            recyc_onsite_qty treated_onsite_qty) lib.release_info
(keep=tri_facility_id
            release_estimation_amount environmental_medium);
      BY tri_facility_id;
RUN;
```

To ensure dataset integrity and handle missing values in the combined TRI dataset (`combined_data`) generated from the previous SAS program, we can exclude observations with missing values from the analysis.

```
DATA lib.combined_data;
      SET lib.combined_data;

      IF NOT MISSING(release_estimation_amount);

      IF NOT MISSING(treated_onsite_qty);

      IF NOT MISSING(energy_onsite_qty);

      IF NOT MISSING(recyc_onsite_qty);
RUN;
```

Analysis of relationships within the Toxic Release Inventory

1. *How do release estimation amounts vary across different environmental media (water, air, surface, underground injection) within TRI data?*

Look at some plots and basic descriptive statistics to investigate the data. The output suggests a big difference between the release estimation amount for the fourteen environmental mediums. Next fit a one-way analysis of variance model using proc glm. First we must tell SAS which variable is the classification variable, environmental medium.

```
PROC MEANS DATA=lib.combined_data NOPRINT;
      CLASS environmental_medium;
      VAR release_estimation_amount;
      OUTPUT out=release_stats MEAN=mean_amount MEDIAN=median_amount
```

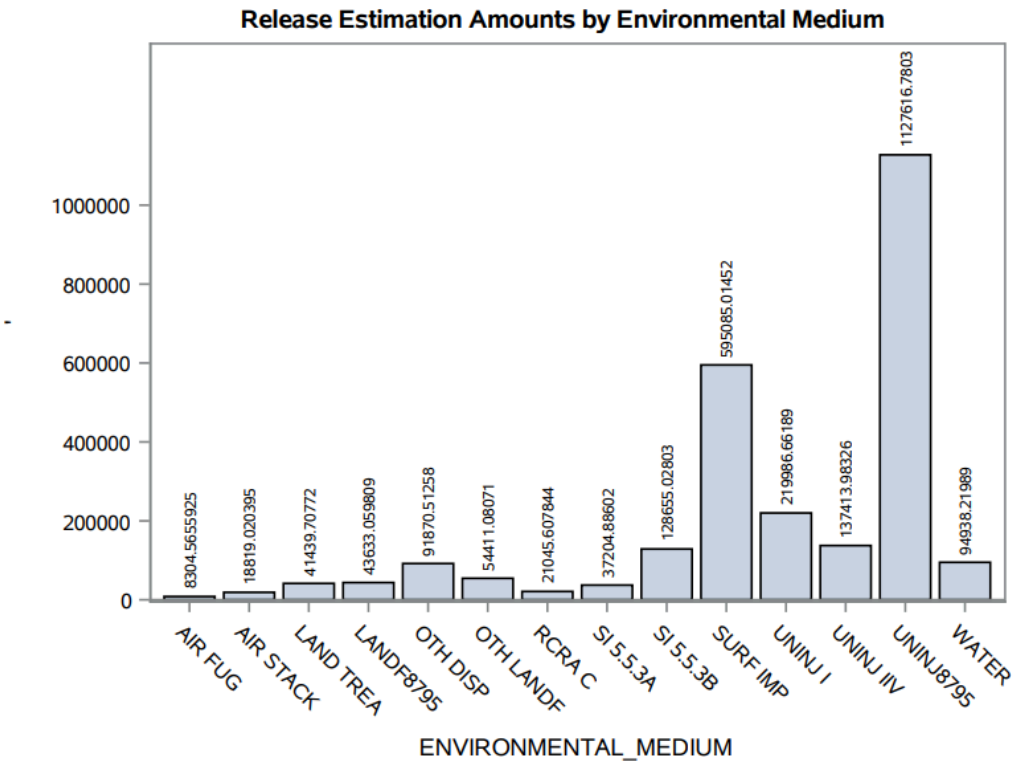
```
n=NOBS;
RUN;

PROC PRINT DATA=release_stats;
      TITLE
            'Descriptive Statistics for Release Estimation Amounts by
Environmental Medium';
      VAR environmental_medium mean_amount median_amount NOBS;
RUN;
```

Descriptive Statistics for Release Estimation Amounts by Environmental Medium

Obs	ENVIRONMENTAL_MEDIUM	mean_amount	median_amount	nobs
1		39303.696091	610.78510351	204196
2	AIR FUG	8304.5655925	889.32644657	78401
3	AIR STACK	18819.020395	2037.2234781	76961
4	LAND TREA	41439.70772	133.7494714	2415
5	LANDF8795	43633.059809	0	4175
6	OTH DISP	91870.51258	1.8946618669	11165
7	OTH LANDF	54411.08071	154.60834146	3757
8	RCRA C	21045.607844	0	1789
9	SI 5.5.3A	37204.88602	0	252
10	SI 5.5.3B	128655.02803	90	1643
11	SURF IMP	595085.01452	111.72238809	2137
12	UNINJ I	219986.66189	2041.4054883	277
13	UNINJ IIV	137413.98326	0	201
14	UNINJ8795	1127616.7803	0	774
15	WATER	94938.21989	58.402942513	20249

```
PROC SGPLOT DATA=release_stats;
      TITLE 'Release Estimation Amounts by Environmental Medium';
      VBAR environmental_medium / RESPONSE=mean_amount
DATALABEL=mean_amount;
RUN;
```

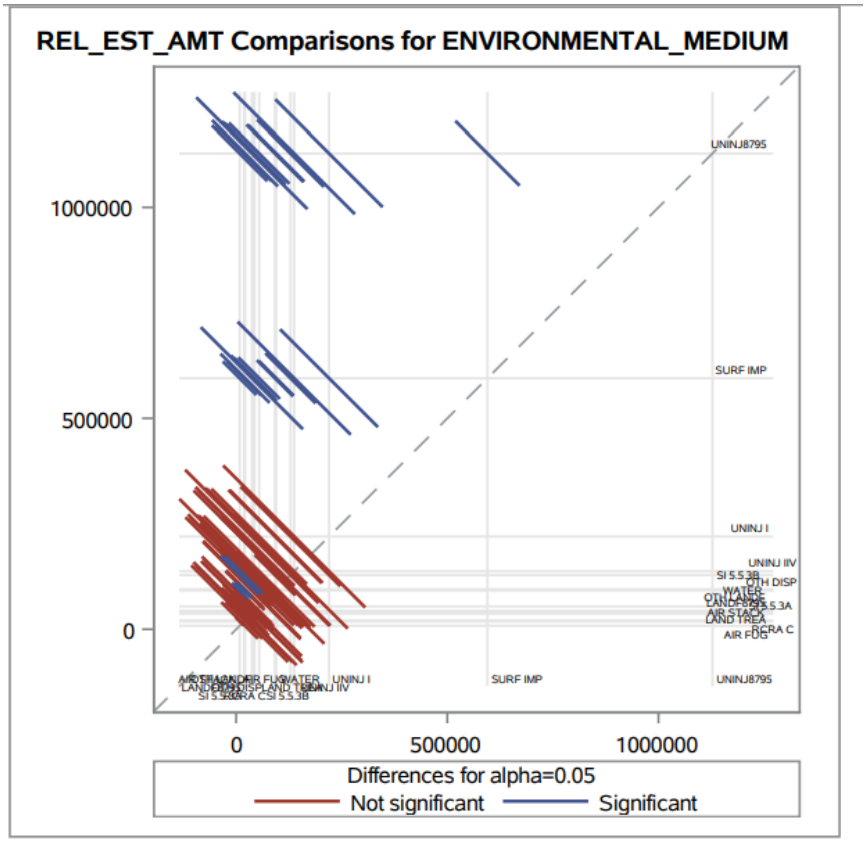


The majority of the bars representing different environmental media categories appear relatively low on the chart. This suggests that most environmental media types have relatively lower average release estimation amounts compared to a few specific categories.

The bars labeled as "SURF IMP" (Surface Impoundments) and "UNINJ8795" stand out prominently due to their significantly higher heights compared to other bars. The taller height of these bars indicates that the release estimation amounts associated with "SURF IMP" and "UNINJ8795" environmental media categories are much higher on average compared to other categories.

```
PROC GLM DATA=lib.combined_data;
      CLASS environmental_medium;
      MODEL release_estimation_amount=environmental_medium;
      LSMEANS environmental_medium / PDIFF;
      RUN;
```

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
ENVIRONMENTAL_MEDIUM	13	1.8036488E15	1.3874221E14	40.43	<.0001



The statistically significant F-value (40.43 with p-value < 0.0001) indicates that there are significant differences in release estimation amounts across different environmental media categories.
The LS-Means graph reveals specific insights into the estimated means of release estimation amounts for each environmental media category. Most environmental media categories exhibit similar or relatively low release estimation amounts, as indicated by the low points in the graph. The exceptions are the SURF IMP and UNINJ8795 (Underground Injection) categories, which show significantly higher estimated means compared to other categories.

2. *Are there specific types of quantity reduction activities (energy recovery from burned chemical waste, recycling of toxic chemicals, treated (destruction) of toxic substances) that show stronger associations with lower release estimation amounts?*

First, we calculate summary statistics (e.g., mean, median, standard deviation) for `energy_onsite_qty`, `recyc_onsite_qty`, `treated_onsite_qty`, and `release_estimation_amount`.

```
PROC MEANS DATA=lib.combined_data MEAN MEDIAN STD;
      VAR energy_onsite_qty recyc_onsite_qty treated_onsite_qty
            release_estimation_amount;
      TITLE 'Descriptive Statistics for Quantity Reduction and Release
Information';
RUN;
```

Variable	Etiqueta	Media	Mediana	Desv. est
energy_onsite_qty		0.4072028	0	1.3915734
recyc_onsite_qty		0.9894617	0	2.0276725
treated_onsite_qty		2.0574618	1.0000000	2.5039131
RELEASE_ESTIMATION_AMOUNT	'	109525.06	2835.04	2443659.89

We create scatter plots to visualize relationships between quantity reduction variables (`energy_onsite_qty`, `recyc_onsite_qty`, `treated_onsite_qty`) and `release_estimation_amount`.

```
PROC SGPLOT DATA=lib.combined_data;
      TITLE
            'Relationship Between Energy Quantity Reduction and Release
Estimation Amount';
      scatter x=energy_onsite_qty y=release_estimation_amount /
            markerattrs=(symbol=circlefilled);
      xaxis label='Energy Onsite Quantity';
      yaxis label='Release Estimation Amount';
RUN;

PROC SGPLOT DATA=lib.combined_data;
      TITLE 'Relationship Between Teated Chemicals Quantity Reduction and
Release Estimation Amount';
      scatter x=treated_onsite_qty y=release_estimation_amount /
            markerattrs=(symbol=circlefilled);
      xaxis label='Treated Chemicals Onsite Quantity';
      yaxis label='Release Estimation Amount';
RUN;

PROC SGPLOT DATA=lib.combined_data;
      TITLE 'Relationship Between Recycling Quantity Reduction and Release
Estimation Amount';
      scatter x=recyc_onsite_qty y=release_estimation_amount /
            markerattrs=(symbol=circlefilled);
      xaxis label='Recycling Onsite Quantity';
      yaxis label='Release Estimation Amount';
RUN;
```

```
PROC ANOVA DATA=lib.combined_data;
      CLASS recyc_onsite_qty treated_onsite_qty energy_onsite_qty;
      MODEL release_estimation_amount=recyc_onsite_qty treated_onsite_qty
            energy_onsite_qty recyc_onsite_qty*treated_onsite_qty
            recyc_onsite_qty*energy_onsite_qty
energy_onsite_qty*treated_onsite_qty;
      MEANS recyc_onsite_qty treated_onsite_qty energy_onsite_qty/ tukey
cldiff;
      RUN;
```

Origen	DF	Anova SS	Cuadrado de la media	Valor F	Pr > F
recyc_onsite_qty	9	4.7750991E14	5.3056656E13	8.91	<.0001
treated_onsite_qty	9	1.5143274E14	1.682586E13	2.83	0.0025
energy_onsite_qty	9	1.0712501E14	1.1902778E13	2.00	0.0353
recyc_ons*treated_on	81	3.1137648E15	3.8441541E13	6.45	<.0001
recyc_ons*energy_ons	81	2.7336479E14	3.3748739E12	0.57	0.9994
treated_o*energy_ons	81	2.8510659E14	3.5198344E12	0.59	0.9987

Thanks to the ANOVA test we can study the relationship between the release estimation amount and the amount of source reduction in each of the reduction forms.

In the test we can see that there are significant differences in the release estimation amount and the different amounts of source reduction in the three activities.

As for the recycling activity of toxic chemicals, we see that there is a big difference when the amount of recycled chemicals is 6 pounds compared to any other. In other

words, the facilities that recycle 6 pounds are the ones that have the highest release estimation amount.

As for the activity of destruction of toxic chemicals, we see that there is a big difference when the amount of chemicals destroyed is 3 or 4 pounds compared to when there are no chemicals destroyed at all.

As for the activity of burning toxic chemicals to recover energy, we see that there is no significant difference. In other words, there is no relationship between the amount of toxic chemicals burned for energy recovery and the emission of toxic chemicals from a facility.

Finally, we can see that there is an interaction between the activity of recycling toxic chemicals and the activity of burning them to produce energy.

Including distance to our work

One of our biggest dreams has always been to live and work in Los Angeles. We want to use the data we have on toxic emissions in the US to learn about the situation in Los Angeles. We know that the longitude and latitude in degrees are: 34.0523, -118.2437.

On the other hand, we have used the data on the locations of industrial facilities provided by the EPA to know the longitudes and latitudes in DMS.

In order to calculate the distances of each industrial facility and Los Angeles we have to convert the longitudes and latitudes to degrees and then calculate the difference using GEODIST.

First we will convert the CSV downloaded from the official EPA website to a SAS table.

```
PROC IMPORT OUT=lib.facility_information
DATAFILE="/home/u63792450/sasuser.v94/SAS
statistics/Project/facility_information.CSV"
```

```
            DBMS=CSV REPLACE;
      GETNAMES=YES;
RUN;
```

Facility Information:

- Data File: `facility_information`
- Variables of Interest:
  - `tri_facility_id`: Identifier for reporting facility
  - `region`: The EPA region in which the facility is located.
  - `fac_latitude`: The series of numbers that identifies the exact physical location of the facility as a measure of the angular distance north form the earth's equator to the center of the facility. The value is stored as degrees, minutes and seconds (0DDMMSS)
  - `fac_longitude`: The series of numbers which identifies the exact physical location of the facility. The right-justified value is stored as degrees, minutes and seconds (0DDDMMSS)

As shown in the summary, longitude and latitude are stored as a six or seven digit number indicating degrees, minutes and seconds. In order to convert this to degrees, we will first separate this long number into three separate numbers indicating degrees, minutes and seconds.

```
DATA lib.facility_information_decimal(DROP=fac_latitud fac_long);
      SET lib.facility_information;
      fac_latitud=fac_latitude;
      lat_sec=input(substr(fac_latitud, length(fac_latitud)-1,
length(fac_latitud)),
            best2.);
      fac_latitud=input(substr(fac_latitud, 1, length(fac_latitud)-2),
```

```
best12.);
      lat_min=input(substr(fac_latitud, length(fac_latitud)-1,
length(fac_latitud)),
            best2.);
      lat_deg=input(substr(fac_latitud, 1, length(fac_latitud)-2),
best12.);
      fac_long=fac_longitude;
      long_sec=input(substr(fac_long, length(fac_long)-1,
length(fac_long)), best2.);
      fac_long=input(substr(fac_long, 1, length(fac_long)-2), best12.);
      long_min=input(substr(fac_long, length(fac_long)-1,
length(fac_long)), best2.);
      long_deg=input(substr(fac_long, 1, length(fac_long)-2), best12.);
RUN;
```

Now we can convert the DMS data to degrees using the formula:

$$VALUE = (degrees) + (minutes/60) + (seconds/3600)$$

```
DATA lib.facility_information_decimal(DROP=lat_sec lat_min lat_deg long_sec
            long_min long_deg);
      SET lib.facility_information_decimal;
      facility_latitude_decimal=lat_deg+lat_min/60+lat_sec/3600;
      facility_longitude_decimal=long_deg+long_min/60+long_sec/3600;
RUN;
```

Finally, we can now calculate the distances from each industrial facility to our selected city.

```
DATA lib.facility_distance_to_los_angeles;
      SET lib.facility_information_decimal;

      /* Calculate distance to Los Angeles (34°03'08.1"N, 118°14'37.3"W) */
      distance_to_los_angeles=GEODIST(facility_latitude_decimal,
            -facility_longitude_decimal, 34.0523, -118.2437, 'M');

      /* Coordinates of Los Angeles in decimal degrees */
      DROP fac_latitude fac_longitude facility_latitude_decimal
            facility_longitude_decimal;
```

```
RUN;
```

Incorporate distance to the analysis

To incorporate distances between industrial facilities and Los Angeles into the analysis of quantity reduction and release information, we will leverage the calculated distances to explore potential spatial relationships and their impact on hazardous substance releases.

```
PROC SORT DATA=lib.facility_distance_to_los_angeles;
      BY tri_facility_id;
RUN;

DATA combined_data_with_distance;
      MERGE lib.combined_data lib.facility_distance_to_los_angeles;
      BY tri_facility_id;
RUN;
```

First, we categorize facilities into groups based on their distances to Los Angeles. This will allow us to compare release estimation amounts across different distance categories.

```
PROC RANK DATA=combined_data_with_distance OUT=combined_data_with_distance
            groups=10 /* Specify the number of distance groups */;
      var distance_to_los_angeles;
      ranks distance_group;
RUN;
```

We conduct one-way ANOVA to test for differences in release estimation amounts across multiple distance groups.

```
PROC GLM DATA=combined_data_with_distance;
      CLASS distance_group;
      MODEL release_estimation_amount=distance_group;
      MEANS distance_group / tukey cldiff;
      TITLE 'One-way ANOVA: Release Estimation Amount by Distance Group';
```

```
RUN;
```

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
distance_group	9	1.3560709E15	1.5067454E14	24.49	<.0001

The ANOVA test already warns us that there are significant differences between the groups, as the p-value < 0.001. Using Tukey's HSD test for release_estimation_amount we see significant differences between distance groups to Los Angeles. We can see that distance groups 1 and 2 have very significant differences with respect to the rest of the groups. This means that the facilities located in the two groups closest to Los Angeles emit considerably more toxic chemicals.

Conclusion

In conclusion, this project has underscored the critical importance of analyzing and understanding the dynamics of toxic emissions through the Toxic Release Inventory (TRI). The findings indicate significant variability in estimated release amounts based on different environmental media and toxin reduction activities implemented at industrial facilities. These results highlight the need for more stringent regulatory policies and tailored mitigation strategies that are adapted to the specific characteristics of each medium and type of industrial activity to effectively reduce toxic emissions. Additionally, the inclusion of spatial analysis in this study has revealed how geographic proximity to urban areas like Los Angeles can influence emission patterns, emphasizing the importance of considering geographic context in environmental strategy planning and execution. Ultimately, this work not only provides a comprehensive insight into the management and impact of toxic wastes but also lays a solid foundation for future research and developments in environmental regulation.