Water Contamination Sources Around Hennepin County MN

Mary Heise May 13, 2022

---

1. Introduction:

A high standard of drinking water quality is critical for the health of the public as well as the economy. It's often assumed that municipal drinking water is clean and safe as it's been regulated under the Safe Drinking Water Act since 1974. The truth is, however, that many chemicals aren't regulated or even tested for in drinking water. Meanwhile, industrial pollution continues to grow both in quantity and variety of chemicals released to the environment. The recent news of the health and environmental hazards of Per- and Polyfluoroalkyl Substances (PFAS), also referred to as "forever chemicals", illustrates perfectly that the challenge of protecting water is as persistent as these chemicals are. PFAS are not the only pollutants that remain in the environment for decades; other chemicals and some metals continue to accumulate and create health problems for both people and ecosystems.

Understanding past and current pollution is necessary to manage toxic waste and protect natural resources. Even if these polluting discharges don't directly impact waterways, airborne discharges eventually land and contaminate soils and water which eventually connect to a well or water intake pipe. For this reason, it's important to look at a large geographical area and all sources of pollution to assess the damage. This study aims to assess threats to water quality around the city of Minneapolis MN by analyzing a variety of pollution sources. Since pollution disperses widely in the environment, the larger area around Hennepin County was considered to factor in chemical transport pathways. Because it's difficult to quantify the effects of any contamination event by simply looking at quantities and types released, additional data was incorporated that models actual human health hazards.

2. Database:

Figure 1 shows the conceptual model of the database. It's broken down into two parts; one consisting of the tabular data and the second pertaining to the spatial aspect allowing for various proximity analyses on the tabular data.

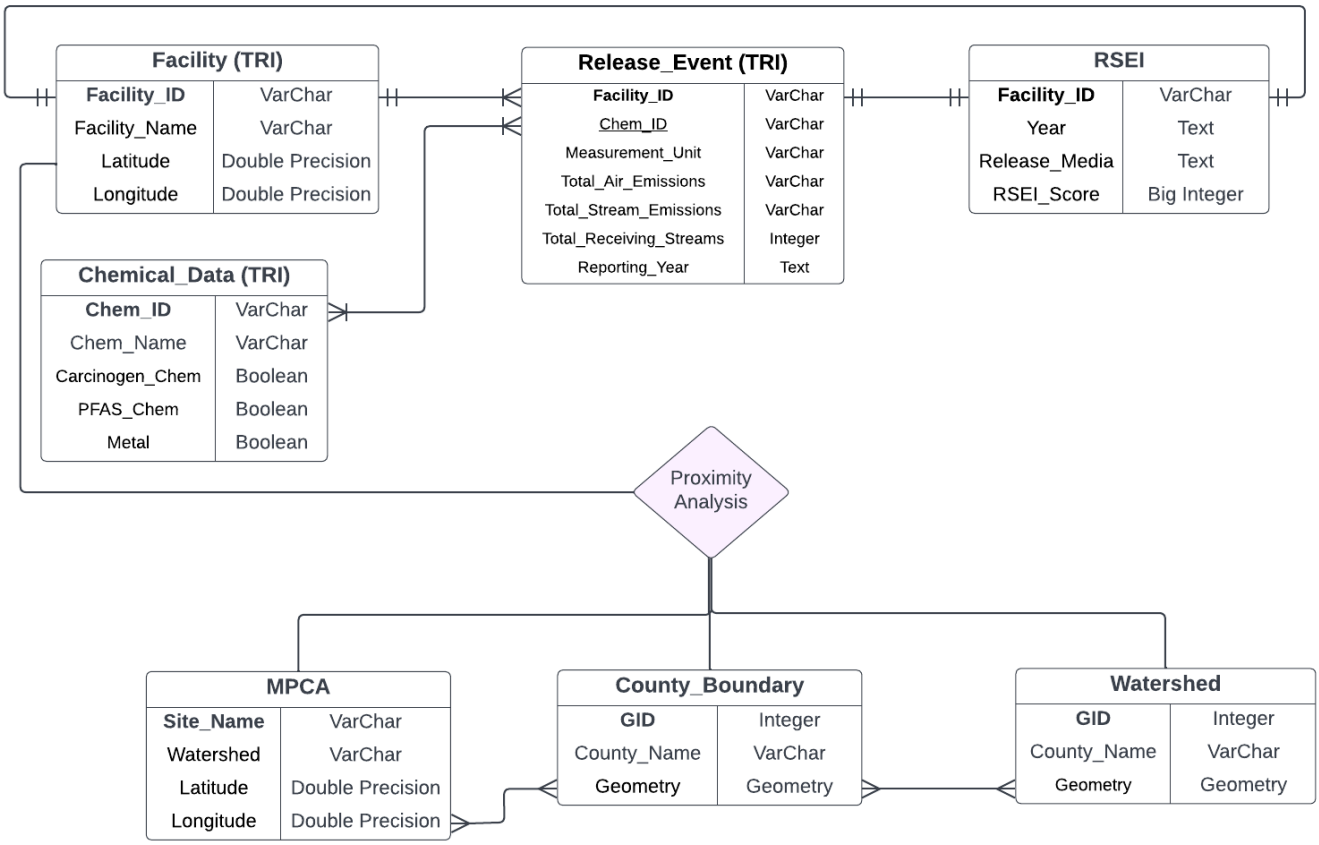


Figure 1. Conceptual model of the database.

3. Data:

Table 1. Data, sources and links used in the study analysis.

Data Source	Dataset	Purpose	File Format
Environmental Protection Agency	Toxic Release Inventory	Show location, quantity & type of toxic releases events annually.	txt
Environmental Protection Agency	Risk-Screening Environmental Indicators model	Modeled health risks of every TRI reported event, reported as a relative score.	txt
MN Pollution Control Agency	What's in My Neighborhood	Show accumulative point data of contaminated sites & release events in MN.	shp
MN Dept of Transportation	County Boundaries in MN	Create the boundaries of the Area of Interest (AOI) for the study.	shp
MN Dept of Natural Resources	Watershed Boundaries	Calculate pollution impacts within watershed boundaries.	shp

The study area boundary was derived from the MN Department of Transportation County Boundaries shapefile which contains geometries for all MN counties.

Pollution data originated from the Environmental Protection Agency (EPA) in two useful datasets that track point sources of pollution and the health risks associated with each release event. The Toxic Release Inventory (TRI) dataset reports on all pollution release events on an annual basis going back to 1986. Along with the facility and release location of each emission, listed in latitude/longitude coordinates, the data reports chemical name, type and quantity as well as discharge media – either air or water. The types are classified as PFAS, carcinogens or metals. The data was downloaded from the website as a zipped file with a batch of ten separate txt files which required further processing described in the Methods section.

The Risk-Screening Environmental Indicators model (RSEI) estimates potential health impacts of every TRI release event. The model considers the known chemical toxicity, quantity of release, transport through the environment and number of people exposed to the plume and quantifies the effects in a relative score with higher numbers indicating more detrimental impacts on the population. The score also differentiates between cancer and non-cancer health risks. This is useful data for showing cumulative impacts on populations that may be exposed to various types of exposure hazards. RSEI was downloaded as a single txt file for each year with no specific geometry attribute as it's associated with the TRI data facility location.

Additional point data on pollution sources is reported by the Minnesota Pollution Control Agency (MPCA). Their 'What's In My Neighborhood' database tracks pollution sources within communities going back to the early 1980's. The data records location in latitude/longitude, affected watersheds, site activity, industrial classification, active status of site and presence of site controls that reduce human exposure to contamination. This is a long term assessment of sites that need clean-up or are actively emitting pollution and range from animal feedlots to construction permits, hazardous waste facilities to brownfields. The MPCA data reveals the long term accumulation of contamination sites whereas the EPA data looks at single event releases. Together, both types of data more fully illustrate the whole scale of the problem.

Boundary data for both county and watershed levels originated from the MN Departments of Transportation and Natural Resources respectively. Both provide basic polygons of the entire state but were reduced to the Area of Interest (AOI) as described below.

4. Methods:

As mentioned above, the Tri dataset comes zipped in a batch of ten files. It was determined that one file provided more than adequate information. This particular file contained 269 columns of which only 22 were useful for this analysis. The header format was originally written with numbers, spaces, periods and all capitalized letters which generally pose problems for data extraction and querying. Additional preprocessing was required to avoid

these complications and was accomplished using a combination of Microsoft Excel, Jupyter Notebook and Python Pandas package. The notebook was also used to split the table into three separate First Normal Forms in order to align with good data management practices. An additional goal of automation was to streamline data acquisition for multiple years' worth of TRI reports in order to assess cumulative pollution impacts over time.

The RSEI and MPCA datasets were also preprocessed in Excel to eliminate extra columns and rename headers before saving as csv files for loading into the database. Windows Command Prompt was used to pull in all datasets with the loader script attached in the Appendix. Using Structured Query Language (SQL) in pgAdmin4, all datasets were clipped to the AOI (Figure 2) which was built by finding all counties surrounding Hennepin County. It was determined to exclude Sherburne county as an outlier as it barely touches Hennepin.

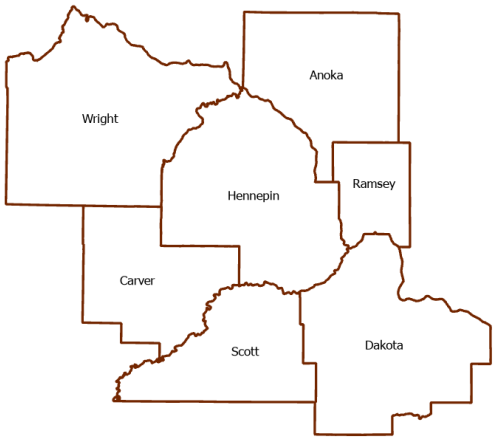


Figure 2. The study area consists of Hennepin County and six surrounding counties.

Once loaded, the TRI coordinates were converted into points with ST_MakePoint and the RSEI table was joined with the Facility table tri_fd to connect location information.

5. Data Analysis:

In the first phase of data exploration using Count, Sum and Group By statements, it was discovered that one facility was the major contributor of both air and water pollution in 2020, far surpassing all other facilities combined for water emissions and nearly matching them on air emissions. Further investigation is warranted to see what the long term historical trend is or if this is an anomaly for 2020 given reduced operations for many industrial facilities due to Covid-19. Calculating the average annual values for the entire AOI of TRI release events would reveal important information about the long term pollution burden on waterways as well as reveal more facility specific trends. Overlaying air and water emissions with watershed boundaries over time would provide more location specific impacts. Adding soil sample data to the database would be ideal for best understanding the complete cost of health and environmental emissions.

Breaking out the type of pollutant with more SQL aggregate functions revealed metals as the main toxin with carcinogens present in about half that quantity and zero pounds of PFAS in 2019. Since the public has just recently become aware of PFAS hazards and regulations are just beginning to be implemented, it's surprising that there were no reported releases of PFAS. This raises the need to dig farther back into the data to see how it's changed over time.

Calculating the density of combined MPCA contaminated sites along with 2020 TRI release events per county revealed Ramsey County as the most burdened with pollution. Some initial exploration on point density within watershed boundaries shows that the Minneapolis – Twin Cities watershed receives the most direct emissions and mainly by the one big polluting facility. Further analysis is needed to quantify pounds of emissions, split out air and water discharges and dig into the specifics of MPCA sites regarding active status, presence of site controls and area of impacted land.

6. Challenges and Lessons Learned:

The format of the TRI and RSEI EPA tables created several problems with cleaning and loading data. In addition to the difficult header format, the first column kept defaulting as the index therefore shifting all the headers over one column. Pandas helped after much trial and error but eventually it just made sense to use Excel to remove extra columns, rename them and filter out by state. I finally resorted to this after several tables were downloaded locally but with the columns misaligned after I thought I solved the problem. Excel has many useful tools but it was quite a learning curve in itself. It's also a time consuming and tedious process when several tables need the same cleaning. Building stronger coding skills is a must to avoid this problem for future work.

Splitting the TRI data down to the barest First Normal Forms wasn't actually helpful as all queries relied on the joined tables. It would have been better to combine the Facility and Pollution Release tables and only split out the Chemical schema as the Chemical data was scarcely called upon. Given the scope of this analysis, working with a larger table would likely not create any issues. Also, I initially wanted to bring in as many columns that appeared helpful or interesting but this added undue difficulties for preprocessing and loading. It would have been better to use the bare minimum of data and bring in additional data if needed.

Data types posed issues in both the data loading and analysis processes. To facilitate data loading, Variable Character was selected as the default after other types failed to load. This required casting columns later when performing date and arithmetic queries and added unnecessary complications. Best practices are to troubleshoot during the data loading process in order to assign the most workable data type from the very beginning.

There were problems with differing or no projections when running spatial analysis queries. The first challenge was to get the separate latitude/longitude coordinates into a workable format. This was accomplished with subqueries calling upon ST_SetSRID and

ST_MakePoint functions. Then ST_Transform was called upon to convert to a local projection to accommodate distance measurements. It would be helpful in future work to better prepare and coordinate projections as a part of the data cleaning process rather than the analysis phase. This would allow the focus to remain on the question at hand rather than the lack of a proper UTM.

7. Solutions:

This database provides a good start to understanding what pollution is being emitted in the dense urban Minneapolis/Hennepin County area and possible impacts on water quality and human health. It catches a snapshot in time as well as the potential long term assessment of pollution trends. This information can be used to drive environmental and public health policies, mitigate contamination impacts and better regulate industries. By bringing in demographic data, it can also promote environmental social justice by raising awareness of the most hard hit areas of the city in relation to more vulnerable populations which tend to face greater exposure to toxic pollution.

Water quality for residents can be improved by targeting the worst affected areas and offenders for mitigation and regulation. Annual updates to the database incorporating the newest EPA reports would enable continual monitoring of the situation and relatively short response times to emissions spikes. Using the same automation process for data acquisition and SQL analysis will ensure consistency and ease for future ongoing monitoring efforts as well as historical pollution assessments.

Appendix:

1. **Create Table Statements**

```
-- Load in TRI Facility
DROP TABLE IF EXISTS tri_2020;

CREATE TABLE tri_2020
(
year text,
tri_fd varchar,
city text,
county text,
state text,
lat double precision,
long double precision,
chemical_name varchar,
measurement text,
total_receiving_streams smallint,
total_water_discharge double precision,
total_on_site_releases double precision,
primary key (tri_fd)
);

\copy tri_2020 from
'C:\Users\mmMary\Documents\GIS_Classes\pfas_project\DataBases\tri_2020.csv' with header CSV;

-- Load in TRI chem
DROP TABLE IF EXISTS tri_chem2020;

CREATE TABLE tri_chem2020
(
chem_id varchar
chemical_name varchar,
is_carcinogen boolean,
is_pfas boolean,
is_metal boolean,
primary key (chem_id)
);

\copy tri_chem2020 from
'C:\Users\mmMary\Documents\GIS_Classes\pfas_project\tri_chem2020.csv' with header CSV;

-- Load in MPCA Site data
DROP TABLE IF EXISTS mpca;

CREATE TABLE mpca
(
site_id varchar,
name varchar,
```

```
city text,
county text,
watershed varchar,
latitude double precision,
longitude double precision,
primary key (site_id)
);

\copy mpca from
'C:\Users\mmMary\Documents\GIS_Classes\pfas_project\DataBases\mpca_sites.csv' with header
CSV;

-- Load RSEI
DROP TABLE IF EXISTS rsei_2020;

CREATE TABLE rsei_2020
(
year smallint,
tri_fd varchar,
name varchar,
release_media text,
chemical_name varchar,
rsei_score double precision,
latitude double precision,
longitude double precision,
primary key (tri_fd)
);

\copy rsei_2020 from
'C:\Users\mmMary\Documents\GIS_Classes\pfas_project\DataBases\DATA\rsei_2020.csv' with
header CSV;
```

2. SQL Queries

a. A SQL query that involves only 1 table.

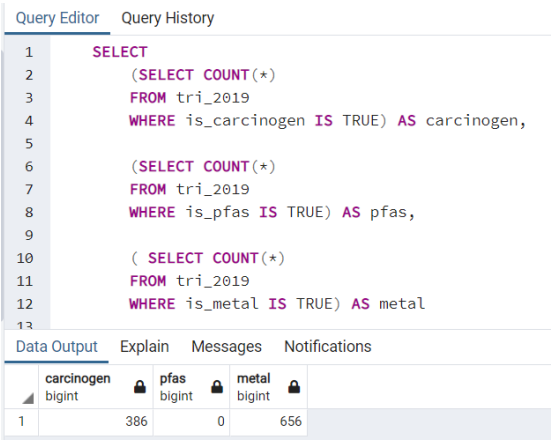
```
-- Find how many pollution releases of each type of toxin for 2019 study area
-- Show numbers for carcinogens, pfas & metals side by side in one table
-- * was used as it returns a very short concise output

  SELECT
    (SELECT COUNT(*)
    FROM tri_2019
    WHERE is_carcinogen IS TRUE) AS carcinogen,

    (SELECT COUNT(*)
```

```
FROM tri_2019
WHERE is_pfas IS TRUE) AS pfas,

( SELECT COUNT(*)
FROM tri_2019
WHERE is_metal IS TRUE) AS metal
```



b. A SQL query that involves 2 or more tables with a join.

```
-- Get highest rsei score per county

SELECT
  t20.county,
  SUM(rs.rsei_score) AS county_rsei
FROM
  rsei_2020 rs
  INNER JOIN tri_2020 t20 ON (t20.tri_fd = rs.tri_fd)
WHERE
  t20.is_carcinogen IS TRUE
GROUP BY
  t20.county
ORDER BY
  county_rsei desc
```

```
17  SELECT
18    t20.county,
19    SUM(rs.rsei_score) AS county_rsei
20  FROM
21    rsei_2020 rs
22    INNER JOIN tri_2020 t20 ON (t20.tri_fd = rs.tri_fd)
23  WHERE
24    t20.is_carcinogen IS TRUE
25  GROUP BY
26    t20.county
27  ORDER BY
28    county_rsei desc
29
```

Data Output | Explain | Messages | Notifications

	county text	county_rsei double precision
1	DAKOTA	20511839
2	RAMSEY	5519672
3	HENNEPIN	4426937
4	ANOKA	2578319
5	CARVER	2493052
6	SCOTT	796807
7	WRIGHT	27

c. A SQL query using a sub query or common table expression

```
-- Get a breakdown of rsei score & mpca site metrics per county
-- Join the rsei with tri data to bring in location info for rsei score
-- Get the total sum of rsei scores per county.

With rsei AS (
  SELECT
    t20.county,
    sum(rs.rsei_score) AS county_rsei
  FROM
    rsei_2020 rs
    INNER JOIN tri_2020 t20 ON t20.tri_fd = rs.tri_fd
  GROUP BY
    t20.county
  ORDER BY
    county_rsei
),
-- Join tri & mpca data to quantify number of mpca sites per county.
mpca AS (
  SELECT
    t20.county,
    COUNT(site_id) as total_mpca_sites
  FROM
    tri_2020 t20
    JOIN mpca mp ON t20.county = UPPER(mp.county)
  GROUP BY
    t20.county
  ORDER BY
    t20.county,
    total_mpca_sites
```

```
)
-- Get percentage for both rsei score & mpca site density for each county &
-- show all results in one table.
SELECT
  r.county,
  r.county_rsei,
  m.total_mpca_sites,
  cast (
    100 * r.county_rsei / SUM(r.county_rsei) OVER () AS int
  ) AS rsei_percent,
  cast (
    100 * m.total_mpca_sites / SUM(m.total_mpca_sites) OVER () AS int
  ) AS mpca_percent
FROM
  rsei r
  JOIN mpca m ON r.county = m.county
GROUP BY
  r.county,
  r.county_rsei,
  m.total_mpca_sites
ORDER BY
  r.county
```

```
31  -- Get a breakdown of rsei score & mpca site metrics per county
32  -- Join the rsei with tri data to bring in location info for rsei score
33  -- Get the total sum of rsei scores per county.
34
35  With rsei AS (
36    SELECT
37      t20.county,
38      sum(rs.rsei_score) AS county_rsei
39    FROM
40      rsei_2020 rs
41      INNER JOIN tri_2020 t20 ON t20.tri_fd = rs.tri_fd
42    GROUP BY
43      t20.county
44    ORDER BY
45      county_rsei
```

Data Output | Explain | Messages | Notifications

	county text	county_rsei double precision	total_mpca_sites bigint	rsei_percent integer	mpca_percent integer
1	ANOKA	6233509	435902	4	6
2	CARVER	7495396	86924	5	1
3	DAKOTA	95830130	752094	64	10
4	HENNEPIN	15034973	4843980	10	67
5	RAMSEY	24551525	986670	16	14
6	SCOTT	1730938	140976	1	2
7	WRIGHT	2250	33320	0	0

d. A spatial SQL query

```
-- Get combined TRI & mpca point density per county

-- Create geometry & set projection for tri points
-- Count each point
```

```
WITH tri_cte AS (
  SELECT
    mc.name,
    Count(
      ST_Contains(
        mc.geom,
        (
          ST_SetSRID(
            ST_MakePoint(long, lat),
            4326
          )
        )
      )
    )
            as total
  FROM
    metro_counties mc,
    tri_2020 t
  WHERE
    ST_Contains(
      mc.geom,
      (
        ST_SetSRID(
          ST_MakePoint(long, lat),
          4326
        )
      )
    ) is True
  GROUP BY
    mc.name
  union  -- Show results in same table
-- Create geometry & set projection for mpca points
-- Count each point
  SELECT
    mc.name,
    Count(
      ST_Contains(
        mc.geom,
        (
          ST_SetSRID(
            ST_MakePoint(longitude, latitude),
            4326
          )
        )
      )
    ) as total
  FROM
```

```
  metro_counties mc,
  mpca m
 WHERE
  ST_Contains(
   mc.geom,
   (
    ST_SetSRID(
     ST_MakePoint(longitude, latitude),
     4326
    )
   )
  ) is True
 GROUP BY
  mc.name
)
 -- Add together mpca & tri 2020 points for each county
SELECT
 name,
 SUM(total) AS point_total
FROM
 tri_cte
GROUP BY
 name
```

```
83  ---------------------------------
84  -- Get combined TRI & mpca point density per county
85
86  -- Create geometry & set projection for tri points
87  -- Count each point
88
89  WITH tri_cte AS (
90    SELECT
91      mc.name,
92      Count(
93        ST_Contains(
94          mc.geom,
95          (
96            ST_SetSRID(
97              ST_MakePoint(long, lat),
```

Data Output | Explain | Messages | Notifications

	name character varying (100)	point_total numeric
1	Anoka	6581
2	Carver	2835
3	Dakota	8099
4	Hennepin	23963
5	Ramsey	10483
6	Scott	3252
7	Wright	3341