How to Use the Combined TRI and Health Outcome Datasets

May 12, 2014

Contents

1 Introduction

This archive has been created as a tool for examining relationships between environmental toxicants and mortality for counties in the United States. The three dataset files in this archive contain data for toxicant releases, mortality, and other county characteristics, respectively. Brief descriptions of the three datasets will be provided in the next three sections. Sample codes to load and merge the datasets are then provided in the following section. The datasets are available in Stata, SAS and Excel.

The archive was developed and provided through the support of the Environmental Protection Agency Toxics Release Inventory University Challenge program for 2013-2014. It was developed by Michael Hendryx, Juhua Luo and Bo-chiuan Chen at the School of Public Health, Indiana University. We gratefully acknowledge the assistance of Jocelyn Hospital and other EPA staff in the development of this resource.

2 Toxics Release Inventory dataset

The Toxics Release Inventory dataset is stored as "*toxin*". The source for these data is the EPA's TRI.Net site: (`http://www2.epa.gov/toxics-release-inventory-tri-program/trinet`). Each row of data contains various measurements of chemical releases in a county for a specific year. There are 3 identification variables (`county, fips, year`) and 108 data variables for various chemical release measures. The 108 data variables can be characterized by 6 groups of chemicals, 9 types of releases, and 2 units of measurement.

The 6 groups of chemicals can be recognized by the first three or four characters of the variable name:

Variable Prefix	Chemical Group
airp	Hazardous Air Pollutants
carc	Carcinogens
mtl	Metals and Metal Compounds
pbt	Persistent Bioaccumulative Toxic
prio	Priority Pollutants
crcl	Comprehensive Environmental Response, Compensation, and Liability Act

The 9 types of releases can be identified by the next part of the variable name following an underscore (_):

Variable Part	Type of Releases
_total	Total On and Off-Site Releases
_onst	Total On-Site Releases
_air	Total Air Releases
_airf	Fugitive Air Emissions
_airp	Point Source Air Emissions
_water	Surface Water Discharges
_undg	Underground Injection
_land	Total On-Site Land Releases
_offst	Total Off-Site Releases

The 2 units of measurement can be identified by the suffix of the variable name:

Variable Suffix	Unit of Measurement
(none)	Pounds
x	Toxicity Weighted Pounds

For example, the variable `airp_air` is the total air releases for Hazardous Air Pollutants in pounds; the variable `carc_onstx` is the total on-site releases for Carcinogens in toxicity-weighted pounds.

3 Mortality dataset

The mortality dataset is stored as "*mortality*". The source for these data is the CDC Compressed Mortality File (`http://wonder.cdc.gov/mortsql.html`). Each row contains gender specific measurements on death from various causes in a county in a specific year. There are 5 identification variables (`county, fips, year, gender, male`) and 15 data variables for related measurements. The data variables can be characterized by 5 groups of mortality causes and 3 measurements of each (1 count of death and 2 mortality rates). The population size (`population`) is also provided.

The mortality dataset aggregates measurements over a five year period to obtain a more reliable estimate from small populations. The two time periods available are 2001-2005 and 2006-2010. The time periods can be identified using the '`year`' variable.

Value	Time Period
2005	2001 - 2005
2010	2006 - 2010

Gender can be identified using either the character variable '`gender`' or the numeric variable '`male`'.

Value of `gender`	Value of `male`	Gender of the Record
F	0	Female
M	1	Male

The 5 group of mortality causes can be recognized by the prefix of the variable name.

Variable Prefix	Group of Causes
total	All Causes
cancer	All Types of Cancer
breast	Breast Cancers
lung	Lung Cancers
cvd	Cardiovascular Diseases

The 3 measurements can be identified with the suffix of the variable name. The Age-Adjusted Mortality Rate is based on the 2000 U.S. Standard Population. All of the mortality rates are calculated per 100,000 persons.

Variable Suffix	Unit of Measurement
_deaths	Number of Death
_cruderate	Crude Mortality Rate
_adjrate	Age-Adjusted Mortality Rate

4 County profile dataset

The county profile dataset is stored as "*county*". The sources of these data are the County Health Rankings (`http://www.countyhealthrankings.org/`) and the Area Health Resource File (`http://arf.hrsa.gov/`). Each row records various county population characteristics for a specific reporting year. There are 3 identification variables (`county, fips, year`) and 18 data variables.

The 18 data variables are listed below.

Variable Name	Description
smoking	Percent Adult Smoking
obese	Percent Adult Obese
inactive	Percent Physically Inactive
uninsure	Percent without Health Insurance
hischool	High School Average Freshman Graduation Rate (AFGR)
college	Percent with Post-Secondary Education
unemploy	Percent Unemployed
childpov	Percent Children in Poverty
asian	Percent Asain
black	Percent Black
hispanic	Percent Hispanic
nativeam	Percent Native American
otherace	Percent Other race
white	Percent White
tworace	Percent Multi-racial
povrate	Poverty Rate
sqmiles	Area in Square Miles
uic	Urbanization Index

5 Loading and merging datasets

5.1 To load the datasets into SAS

(1) Copy the `.sas7bdat` files into the desired directory.

(2) Submit LIBNAME statement to SAS, for example:

```
LIBNAME libry "/path/to/the/directory"
```

(3) Then the dataset will be accessible by using the SET statement in data steps or DATA= option in procedures. For example, to use the *toxin* dataset in a data step:

```
DATA newdata;
    SET libry.toxin;
    /* codes for data manipulation here */
RUN;
```

Or, to display simple statistics of smoking rate (`smoking`) using the *county* dataset in a MEAN procedure:

```
PROC MEANS DATA = libry.county;
    VAR smoking;
RUN;
```

5.2 To load the datasets into Stata

(1) Copy the `.dta` files into the desired directory.

(2) Submit the "use" command with a path to the dataset. For example, to use the *mortality* dataset in Stata:

```
use "/path/to/the/directory/mortality.dta"
```

5.3 General considerations prior to merging

The three datasets are organized as county-year data for *toxin*, as county-year-gender data for *mortality*, and as county-year data for *county*. The datasets can be merged using the common variable '`fips`' (Federal Information Processing Standards county code, a standard five digit code that represents every county in the United States). It is generally recommended to perform one-to-one merging. Thus, users are encouraged to manipulate the datasets prior to merging to ensure the single occurrence of values in key variables. It is not recommended that users use `county` as a merge variable since the different representations of county names among the data sources are preserved.

5.4 To merge the datasets using SAS

(1) Perform data manipulation tasks.

(2) Sort the resulting datasets by `fips`.

(3) Then the dataset can be merged by using MERGE statement in data steps. For example, to merge the edited *toxin* and *mortality* datasets into a *new* dataset:

```
DATA new;
    MERGE edited.toxin edited.mortality;
    BY fips;
    /* codes for post-merging manipulation here */
RUN;
```

5.5 To merge the datasets using Stata

(1) Perform data manipulation tasks.

(2) Load the base (master) dataset into Stata.

(3) Submit "merge" command with supplement (using) dataset. For example, to merge the edited *county* dataset to the dataset in memory:

```
merge 1:1 fips using "/path/to/the/directory/county.dta"
```

6 Datasets summary

6.1 Summary of Toxics Release Inventory dataset

Variable	Obs.	Unique	Mean	Min	Max	Label
county	58864	2762	(character variable)			County
fips	58878	2764	(character variable)			FIPS
year	58878	25	1999.2	1987	2011	Year
airp_total	55186	48906	1023997	0	4.73e+08	AIRP: Total On and Off-Site Releases
airp_onst	56401	47414	917807.2	0	4.73e+08	AIRP: Total On-site Releases
airp_air	55186	46452	546512.2	0	1.19e+08	AIRP: Total Air Releases
airp_airf	55186	36768	104691.5	0	1.43e+07	AIRP: Fugitive Air Emissions
airp_airp	55186	42748	441820.7	0	1.19e+08	AIRP: Point Source Air Emissions
airp_water	55116	12522	9466.8	0	1.11e+07	AIRP: Surface Water Discharges
airp_undg	55186	1364	83223.5	0	1.87e+08	AIRP: Underground Injection
airp_land	55186	14538	298823.6	0	4.73e+08	AIRP: Total On-site Land Releases
airp_offst	54559	27807	86971.0	0	4.55e+07	AIRP: Total Off-site Releases
airp_totalx	56401	51631	2.33e+12	0	4.42e+15	AIRP: Total On and Off-Site Releases (Toxicity x Pounds)
airp_onstx	56401	50354	1.88e+12	0	4.42e+15	AIRP: Total On-site Releases (Toxicity x Pounds)
airp_airx	56401	49515	1.09e+10	0	5.64e+12	AIRP: Total Air Releases (Toxicity x Pounds)
airp_airfx	56401	43402	4.22e+09	0	5.64e+12	AIRP: Fugitive Air Emissions (Toxicity x Pounds)
airp_airpx	56401	46111	6.71e+09	0	3.85e+12	AIRP: Point Source Air Emissions (Toxicity x Pounds)
airp_waterx	56401	18425	3.75e+08	0	1.57e+13	AIRP: Surface Water Discharges (Toxicity x Pounds)
airp_undgx	56401	1532	1.60e+10	0	4.16e+13	AIRP: Underground Injection (Toxicity x Pounds)
airp_landx	56401	17810	1.85e+12	0	4.42e+15	AIRP: Total On-site Land Releases (Toxicity x Pounds)
airp_offstx	56401	33728	4.53e+11	0	5.63e+14	AIRP: Total Off-site Releases (Toxicity x Pounds)
carc_total	49115	40796	401722.3	0	4.72e+08	CARC: Total On and Off-Site Releases
carc_onst	50049	38169	350563.8	0	4.72e+08	CARC: Total On-site Releases
carc_air	49115	36747	94849	0	1.15e+07	CARC: Total Air Releases
carc_airf	49115	26203	32492.8	0	5530824	CARC: Fugitive Air Emissions
carc_airp	49115	32420	62356.2	0	1.06e+07	CARC: Point Source Air Emissions
carc_water	49068	9122	1404.1	0	1.05e+07	CARC: Surface Water Discharges

Summary of Toxics Release Inventory dataset (continued)

Variable	Obs.	Unique	Mean	Min	Max	Label
carc_undg	49115	1033	16980.9	0	1.15e+07	CARC: Underground Injection
carc_land	49115	11884	243997.7	0	4.72e+08	CARC: Total On-site Land Releases
carc_offst	48620	23700	44944.9	0	4.55e+07	CARC: Total Off-site Releases
carc_totalx	49909	44379	2.56e+12	0	4.42e+15	CARC: Total On and Off-Site Releases (Toxicity x Pounds)
carc_onstx	49909	41979	2.09e+12	0	4.42e+15	CARC: Total On-site Releases (Toxicity x Pounds)
carc_airx	49909	40605	9.06e+09	0	4.52e+12	CARC: Total Air Releases (Toxicity x Pounds)
carc_airfx	49909	33129	2.90e+09	0	4.52e+12	CARC: Fugitive Air Emissions (Toxicity x Pounds)
carc_airpx	49909	36516	6.15e+09	0	1.80e+12	CARC: Point Source Air Emissions (Toxicity x Pounds)
carc_waterx	49909	13936	4.42e+08	0	1.57e+13	CARC: Surface Water Discharges (Toxicity x Pounds)
carc_undgx	49909	1137	1.83e+10	0	4.16e+13	CARC: Underground Injection (Toxicity x Pounds)
carc_landx	49909	14257	2.07e+12	0	4.42e+15	CARC: Total On-site Land Releases (Toxicity x Pounds)
carc_offstx	49909	28691	4.70e+11	0	5.63e+14	CARC: Total Off-site Releases (Toxicity x Pounds)
mtl_total	46766	35729	1028733	0	9.89e+08	MTL: Total On and Off-Site Releases
mtl_onst	48526	27553	786771.2	0	9.89e+08	MTL: Total On-site Releases
mtl_air	46765	23727	10958.0	0	5420000	MTL: Total Air Releases
mtl_airf	46765	15174	4039.4	0	4487777	MTL: Fugitive Air Emissions
mtl_airp	46765	20444	6918.6	0	2220000	MTL: Point Source Air Emissions
mtl_water	46692	11084	4487.9	0	1.05e+07	MTL: Surface Water Discharges
mtl_undg	46765	901	10985.0	0	3.68e+07	MTL: Underground Injection
mtl_land	46765	13760	789974.4	0	9.89e+08	MTL: Total On-site Land Releases
mtl_offst	46604	28617	213090.6	0	1.66e+08	MTL: Total Off-site Releases
mtl_totalx	48526	40665	1.72e+12	0	4.42e+15	MTL: Total On and Off-Site Releases (Toxicity x Pounds)
mtl_onstx	48526	35555	1.55e+12	0	4.42e+15	MTL: Total On-site Releases (Toxicity x Pounds)
mtl_airx	48526	32565	7.74e+09	0	5.64e+12	MTL: Total Air Releases (Toxicity x Pounds)
mtl_airfx	48526	23632	3.56e+09	0	5.64e+12	MTL: Fugitive Air Emissions (Toxicity x Pounds)
mtl_airpx	48526	28091	4.18e+09	0	1.50e+12	MTL: Point Source Air Emissions (Toxicity x Pounds)
mtl_waterx	48526	16670	3.89e+08	0	1.57e+13	MTL: Surface Water Discharges (Toxicity x Pounds)

Summary of Toxics Release Inventory dataset (continued)

Variable	Obs.	Unique	Mean	Min	Max	Label
mtl_undgx	48526	1042	1.36e+10	0	3.02e+13	MTL: Underground Injection (Toxicity x Pounds)
mtl_landx	48526	15897	1.53e+12	0	4.42e+15	MTL: Total On-site Land Releases (Toxicity x Pounds)
mtl_offstx	48526	33365	1.74e+11	0	2.44e+14	MTL: Total Off-site Releases (Toxicity x Pounds)
pbt_total	32960	24568	217930.2	0	4.71e+08	PBT: Total On and Off-Site Releases
pbt_onst	33266	20463	190732.5	0	4.71e+08	PBT: Total On-site Releases
pbt_air	32960	18566	1579.5	0	908900	PBT: Total Air Releases
pbt_airf	32960	9282	438.0	0	312287	PBT: Fugitive Air Emissions
pbt_airp	32960	16843	1141.5	0	890000	PBT: Point Source Air Emissions
pbt_water	32936	5705	90.1	0	71400	PBT: Surface Water Discharges
pbt_undg	32960	411	2155.8	0	8500395	PBT: Underground Injection
pbt_land	32960	9891	188677.9	0	4.71e+08	PBT: Total On-site Land Releases
pbt_offst	32839	16833	25520.6	0	2.90e+07	PBT: Total Off-site Releases
pbt_totalx	33113	24868	9.78e+09	0	3.26e+13	PBT: Total On and Off-Site Releases (Toxicity x Pounds)
pbt_onstx	33113	20869	7.95e+09	0	3.26e+13	PBT: Total On-site Releases (Toxicity x Pounds)
pbt_airx	33113	18805	4.60e+08	0	1.18e+12	PBT: Total Air Releases (Toxicity x Pounds)
pbt_airfx	33113	10492	9.72e+07	0	2.19e+11	PBT: Fugitive Air Emissions (Toxicity x Pounds)
pbt_airpx	33113	16871	3.63e+08	0	1.16e+12	PBT: Point Source Air Emissions (Toxicity x Pounds)
pbt_waterx	33113	6365	7075868	0	1.35e+10	PBT: Surface Water Discharges (Toxicity x Pounds)
pbt_undgx	33113	419	4.31e+07	0	1.53e+11	PBT: Underground Injection (Toxicity x Pounds)
pbt_landx	33113	9993	7.44e+09	0	3.26e+13	PBT: Total On-site Land Releases (Toxicity x Pounds)
pbt_offstx	33113	17351	1.83e+09	0	2.32e+12	PBT: Total Off-site Releases (Toxicity x Pounds)
prio_total	34812	25867	213115.1	0	4.71e+08	PRIO: Total On and Off-Site Releases
prio_onst	35242	22049	185029.6	0	4.71e+08	PRIO: Total On-site Releases
prio_air	34812	20308	3740.7	0	1184321	PRIO: Total Air Releases
prio_airf	34812	11054	1346.6	0	1179282	PRIO: Fugitive Air Emissions
prio_airp	34812	18121	2394.1	0	976400	PRIO: Point Source Air Emissions
prio_water	34787	5968	116.8	0	116278	PRIO: Surface Water Discharges
prio_undg	34812	477	2204.9	0	8500395	PRIO: Underground Injection

Summary of Toxics Release Inventory dataset (continued)

Variable	Obs.	Unique	Mean	Min	Max	Label
prio_land	34812	10127	181252.8	0	4.71e+08	PRIO: Total On-site Land Releases
prio_offst	34566	17234	25983.6	0	2.95e+07	PRIO: Total Off-site Releases
prio_totalx	35116	27030	3.10e+10	0	3.78e+13	PRIO: Total On and Off-Site Releases (Toxicity x Pounds)
prio_onstx	35116	23414	2.26e+10	0	3.78e+13	PRIO: Total On-site Releases (Toxicity x Pounds)
prio_airx	35116	21385	7.20e+08	0	1.18e+12	PRIO: Total Air Releases (Toxicity x Pounds)
prio_airfx	35116	13141	1.57e+08	0	2.24e+11	PRIO: Fugitive Air Emissions (Toxicity x Pounds)
prio_airpx	35116	19026	5.62e+08	0	1.16e+12	PRIO: Point Source Air Emissions (Toxicity x Pounds)
prio_waterx	35116	7198	7147073	0	1.37e+10	PRIO: Surface Water Discharges (Toxicity x Pounds)
prio_undgx	35116	492	4.96e+08	0	1.15e+12	PRIO: Underground Injection (Toxicity x Pounds)
prio_landx	35116	10469	2.14e+10	0	3.78e+13	PRIO: Total On-site Land Releases (Toxicity x Pounds)
prio_offstx	35116	18380	8.45e+09	0	1.06e+13	PRIO: Total Off-site Releases (Toxicity x Pounds)
crcl_total	57681	51792	1746652	0	9.89e+08	CRCL: Total On and Off-Site Releases
crcl_onst	58744	50208	1518862	0	9.89e+08	CRCL: Total On-site Releases
crcl_air	57681	49030	659679.3	0	1.19e+08	CRCL: Total Air Releases
crcl_airf	57681	39607	126874.2	0	1.72e+07	CRCL: Fugitive Air Emissions
crcl_airp	57681	44654	532805.2	0	1.19e+08	CRCL: Point Source Air Emissions
crcl_water	57621	17975	83751.1	0	3.65e+07	CRCL: Surface Water Discharges
crcl_undg	57681	1647	150857.4	0	1.96e+08	CRCL: Underground Injection
crcl_land	57681	17205	652652.3	0	9.89e+08	CRCL: Total On-site Land Releases
crcl_offst	56970	31604	202292.7	0	1.66e+08	CRCL: Total Off-site Releases
crcl_totalx	58728	54284	2.24e+12	0	4.42e+15	CRCL: Total On and Off-Site Releases (Toxicity x Pounds)
crcl_onstx	58728	53018	1.80e+12	0	4.42e+15	CRCL: Total On-site Releases (Toxicity x Pounds)
crcl_airx	58728	51895	1.05e+10	0	4.53e+12	CRCL: Total Air Releases (Toxicity x Pounds)
crcl_airfx	58728	46223	3.92e+09	0	4.53e+12	CRCL: Fugitive Air Emissions (Toxicity x Pounds)
crcl_airpx	58728	47851	6.53e+09	0	3.85e+12	CRCL: Point Source Air Emissions (Toxicity x Pounds)
crcl_waterx	58728	22480	3.74e+08	0	1.57e+13	CRCL: Surface Water Discharges (Toxicity x Pounds)

Summary of Toxics Release Inventory dataset (continued)

Variable	Obs.	Unique	Mean	Min	Max	Label
crcl_undgx	58728	1839	1.55e+10	0	4.16e+13	CRCL: Underground Injection (Toxicity x Pounds)
crcl_landx	58728	20851	1.78e+12	0	4.42e+15	CRCL: Total On-site Land Releases (Toxicity x Pounds)
crcl_offstx	58728	37003	4.33e+11	0	5.63e+14	CRCL: Total Off-site Releases (Toxicity x Pounds)

6.2 Summary of Mortality dataset

Variable	Obs.	Unique	Mean	Min	Max	Label
county	12568	3142	(character variable)			County
fips	12568	3142	(character variable)			FIPS
year	12568	2	2 007.5	2 005.0	2010.0	Year
gender	12568	2	(character variable)			Gender
male	12568	2	0.5	0.0	1.0	Male
population	12564	12206	236 409.3	157.0	2.47e+07	Population
total_deaths	12525	3856	1 946.6	10.0	151495.0	All Cause Deaths
total_cruderate	12441	12337	1 014.6	123.6	2286.3	All Cause Crude Mortality
total_adjrate	12441	12349	881.1	291.3	3573.4	All Cause Adjusted Mortality
cancer_deaths	12260	1882	469.0	10.0	35168.0	Cancer Deaths
cancer_cruderate	11852	11517	233.9	36.3	657.1	Cancer Crude Mortality
cancer_adjrate	11852	11448	204.8	62.7	587.7	Cancer Adjusted Mortality
breast_deaths	4493	493	89.4	10.0	5562.0	Breast Cancer Deaths
breast_cruderate	3095	2903	30.0	10.3	72.3	Breast Cancer Crude Mortality
breast_adjrate	3095	2815	24.8	11.2	57.1	Breast Cancer Adjusted Mortality
lung_deaths	10848	965	148.5	10.0	8938.0	Lung Cancer Deaths
lung_cruderate	9192	8759	72.5	5.6	241.2	Lung Cancer Crude Mortality
lung_adjrate	9192	8728	65.3	9.6	236.7	Lung Cancer Adjusted Mortality
cvd_deaths	12367	2238	687.2	10.0	64808.0	Cardiovascular Deaths
cvd_cruderate	12131	11919	361.0	35.5	1043.7	Cardiovascular Crude Mortality
cvd_adjrate	12131	11901	306.1	87.6	842.7	Cardiovascular Adjusted Mortality

6.3 Summary of County profile dataset

Variable	Obs.	Unique	Mean	Min	Max	Label
fips	12564	3141	(character variable)			FIPS
county	12564	3175	(character variable)			County
year	12564	4	2 011.5	2 010.0	2 013.0	Data year
smoking	9980	1703	21.5	0.0	49.1	% Adult smoking
obese	12564	293	29.5	11.7	47.6	% Adult obese
inactive	6282	274	27.9	10.1	43.9	% Physically inactive
uninsure	12560	377	18.8	3.3	54.3	% Uninsured
hischool	12313	4892	80.6	0.0	100.8	Highschool AFGR
college	9423	6802	53.1	0.0	100.0	% Some college
unemploy	12560	204	8.1	1.1	29.7	% Unemployment
childpov	12560	529	22.8	2.6	67.1	% Children in poverty
asian	12564	99	0.8	0.0	46.0	% Asain
black	12564	496	8.8	0.0	86.5	% Black
hispanic	12564	381	6.2	0.0	97.5	% Hispanic
nativeam	12564	184	1.9	0.0	94.2	% Native American
otherace	12564	227	2.6	0.0	39.1	% Other race
white	12564	587	84.4	0.0	99.7	% White
tworace	12564	81	1.4	0.0	28.4	% Two race
povrate	12564	300	13.7	0.0	49.1	Poverty rate 2000
sqmiles	12564	3102	1 126.2	0.0	145 899.7	Area in square miles
uic	12564	12	5.5	1.0	12.0	Urbanization index