Technical Appendix D

Locational Data for TRI Reporting Facilities and Off-site Facilities

Table of Contents

1 Introduction

The RSEI model uses latitude and longitude coordinates for each TRI reporting and off-site facility to locate each facility on the grid that underpins the model. The facility's location determines many of the modeling inputs, including the population exposed to air releases and the discharge stream reach (if any).

There are two types of facilities included in the model, TRI reporting facilities and off-site facilities. The quality of data varies significantly between the two types. TRI reporters submit their own addresses and, prior to Reporting Year (RY) 2005, submitted estimates of their latitude and longitude on Form R. As facility reports of coordinatess were subject to common reporting errors such as transposition of digits, confusion of latitude with longitude, lack of precision, and nonreporting, TRI no longer requires them to be reported, and instead relies on EPA's centralized databases of locational information. RSEI adopts EPA's coordinates for most reporting facilities as well.

For off-site facilities, the data quality is much lower, as the name and address of these off-site facilities are reported by the TRI reporters transferring the waste, not the receiving facility itself. The name and address tend to be reported in slightly different ways by different reporters, and often misspelled or misreported. Latitude and longitude are not reported. Little or no standardization is performed by the TRI program, so minor differences in an off-site facility record, such as a slight misspelling of the name, or "St." instead of "Street", can make it difficult to automatically match records.

In RSEI Version 1.x, reporting facilities were located on the grid using their reported latitudes and longitudes, and off-site facilities were located using the coordinates of the centroid of their 5-digit ZIP code. For Versions 2.1 through 2.1.3, the coordinates for both reporters and off-site facilities were improved using a commercial geocoding service. Geocoding is a process where a computer program uses street address, city, state, and ZIP code to match addresses to geographic points in Census TIGER files, and then determines the latitude and longitude of the address. For the current version, RSEI uses coordinates from EPA's Facility Registry Service (FRS) for most reporting facilities, and uses coordinates from earlier versions or satellite data only in those cases where the FRS data is not available or is inaccurate. For off-site facilities, RSEI uses a database of off-site facilities based on RSEI Version 2.1.3 that is updated with each new reporting year that contains both FRS coordinates and geocoded coordinates.

Section 2 describes the method and data sources used for reporting facilities. Section 3 describes how coordinates are determined for off-site facilities, including creating the master database, and matching the current year off-site facilities to that database.

2 TRI Reporting Facilities

The primary source for locational information for TRI reporting facilities is EPA's Facility Registry Service (FRS), a centrally managed database developed by EPA's Office of Environmental Information (OEI). The FRS contains accurate and authoritative facility identification records which are subjected to rigorous verification and data management quality

assurance procedures. FRS records are continuously reviewed and enhanced by a Regional Data Steward network and active State partners. RSEI accesses FRS data through EPA's Geospatial Data Access Project, which provides downloadable files of regulated facilities or sites in various formats. Within the file is key facility information, along with associated environmental interests for use in mapping and reporting applications.[1]

The database of all TRI reporting facilities for 1988-2016 includes 59,818 facilities. Of the total number of facilities, 58,553 facilities were assigned coordinates from EPA's FRS system. These coordinates were applied to the TRI reporting facilities in the RSEI model, and the facilities were assigned the codes "FRS" for LatLongSource. In some cases, manual quality assurance led to revision of existing coordinates; these 1,265 records show a source code of "Manual." Generally, these coordinates were generated through geocoding the address and visually checking placement on a satellite image.

The number of facilities with each type of coordinates (including FRS coordinates) assigned are shown in Table D-1 below.

Table D-1
Sources for Coordinates for Reporting Facilities

LatLongSource	Description	Number of Facilities	% of Facilities
FRS	Coordinates taken from EPA's FRS system	58,553	98%
Manual	Coordinates determined manually, either because coordinates were transposed, or based on visual inspection of map.	1,265	2%

3 Off-site Facilities

Accurate locational data for off-site facilities is more difficult to determine, because unique identifiers are not assigned by TRI, and the name and addresses are not reported by the off-site facilities themselves, but by the TRI-reporting facilities that transfer waste to them. Two main problems result from the manner in which off-site facilities are reported: 1) names and/or addresses of the off-site facilities may be reported incorrectly or incompletely; and 2) the same facility name and address may be reported in slightly different forms by different facilities, making it hard to determine unique facilities. Determining unique facilities is important for allowing comparisons of waste volume, hazard, and score across facilities, and can improve overall locational quality by matching records with correct and complete addresses to records that reference the same facility but with incomplete or inaccurate addresses.

[1] Data can be downloaded from this site: http://www.epa.gov/enviro/geo_data.html.

The approach used in RSEI has evolved over the years, as TRI reporting has changed from paper submissions to electronic reporting (resulting in a general increase in data quality), EPA information systems like FRS have improved, and geocoding has become more readily available. The next section briefly describes the changes in approach. Interested users are referred to the Appendix D for any previous version of the model for details on that year's approach. Sections 3.2 and 3.3 describe the current approach.

3.1 Historical Method to Process Off-Site Records

In RSEI Version 1 (released in 1999), all off-site facilities had been located on the model grid using the centroid of the off-site facility's ZIP code. For Versions 2.1 through 2.1.3, the off-site locations were substantially improved using fuzzy matching to collapse the database of off-site facilities, and commercial geocoding to assign a location to each off-site facility in the collapsed database. Briefly, the entire set of off-site facilities was geocoded by TCS, and then the whole set was run through a series of matching programs in SAS, designed to match facilities to each other, on name first (based on the assumption that a third party is most likely to get a facility's name correct), providing leeway for non-exact matches, and then moving through the rest of the facility's address and determining if it is a plausible match. This method was last used in RSEI Version 2.1.3 (RY 2003).

Starting with RSEI Version 2.1.6 (RY 2006), a new method was used that preserves the matches made in previous years (rather than starting from scratch each year). The master table created for Version 2.1.3 was used as a starting point. This master table contains a key that identifies each unique off-site record, and contains every permutation of that record that has been submitted to TRI over the course of TRI reporting, which includes several hundreds of thousands of records. For instance, if a reporting facility listed an off-site transfer as going to "ABC Waste Management" in 2000 and another facility listed one as going to "ABC Waste Mgmt." at the same (or similar) address in 2005, each name would have a separate record in the master table, but be keyed to the same unique record in the collapsed off-site table. The table "offsite" that is described in the RSEI data dictionary contains a single record for each unique key, and consists of approximately 50,000 records (the size of the table varies by year).

Each year, the current year TRI data (for all years) is matched back to the master table, and any unmatched facilities are added to the master list. Because there are no unique IDs or keys, there is no easy way to match the current year TRI data to the master. Exact matching is performed first, and then a simplified version of the fuzzy matching used to create the master database is used to match the remaining facilities.

3.2 Revisions to the Off-Site Database for Version 2.3.6 (RY2016)

In previous years, most of the quality assurance and checking with satellite data was focused on two kinds of off-site facilities: incinerators and publicly-owned treatment works (POTWs), since these are the only off-site facilities that RSEI models. The rest of the off-site facilities were processed in the same way for consistency, but because the records were not modeled, inaccuracies had less practical importance. For Version 2.3.6, RSEI includes an off-site transfer map in the EasyRSEI dashboard, so, for the first time, all off-site facility locations can be

examined and off-site transfers can be analyzed and compared. Prior to this release, a new data quality approach was developed to increase the accuracy of non-modeled off-site records. The method consists of fuzzy matching reported records against records in the FRS system, geocoding non-matched facilities, and manually verifying facilities receiving large transfers. The starting point was the master and off-site tables from version 2.3.5, so previous work in determining locations and unique facilities is preserved.

3.2.1 Matching reported data against FRS data

Reported information (name, address and RCRA identifier where available) in the master table was compared to FRS entries in several iterations. Exact matches of all fields were accepted, as were exact matches of all fields except name (FRS and TRI identifiers remain with a physical location through ownership changes). For remaining facilities, all fields were fuzzy matched against all FRS records. Fuzzy matching uses established algorithms to standardize data fields (e.g., changing abbreviations like Rd. to Road) and evaluate the text inside a field against another field and provide a score from 0 to 100 describing how closely all included fields match. Note that any exact matches are assigned a score of 100.

Fuzzy matching was conducted in several iterations (varying the FRS data subset, including or not including the RCRA as a field to match on, etc.). In each iteration, the scored matches were sorted in descending order, and visually reviewed to determine the appropriate score cutoff (i.e., at what score do false matches begin to appear). Matches above that score cutoff were accepted.

For reported records that were not matched, off-site facilities were sorted in descending order by pounds transferred and manually matched where possible.

In past RSEI updates, reported off-site records were fuzzy matched against each other to determine unique off-site facilities. The master off-site table lists all reported records, and each one is assigned a unique key that identifies a relatively unique off-site record and corresponds to the collapsed list of off-sites in the off-site table. Once these new matches were established, the highest-scoring match in each group was selected for the collapsed off-site table and the FRS information (lat/long, FRSID) of that match was assigned to the off-site record in the collapsed table. The exception is if the existing off-site record in the collapsed table had been manually edited for accuracy, in which case that existing record was retained.

3.2.2 Geocoding off-site facilities

In order to locate facilities that could not be matched to FRS records, the unmatched off-site records were geocoded using ESRI geocoding services. Street address-level matches were accepted; lower-quality matches were evaluated individually. If the address could not be evaluated, the ZIP code was used for the facility location.

3.2.3 Manual verification and potential data quality issues

Matches for the off-site facilities with the largest volume of transfers received were individually verified, and transfers to cities with certain known large off-sites were also examined to make sure that any relevant transfers were matched to the correct record. However, given the data quality of reported transfers, especially in the early years of TRI reporting, some data quality issues remain, and care should be taken with the results. Some of the major issues are:

- A large portion of the pounds reported as being sent off-site do not have a valid reported receiving facility. That portion varies by year, but was 39% in 2016. This includes records where the reporting facility has left all fields blank, reported an incomplete name and address, or reported generic text like "NA," "Nonhazardous scrap metal buyer," or "Chicago" with no additional information.
- This revision process relied on previous fuzzy matching of raw off-site records into groups of records associated with each unique off-site. If erroneous matches were made in the past, they will remain in the current dataset. Future QA will involve verification of the groupings.
- Reporting inconsistencies in the data may lead to incorrect off-site identification. In some cases, reported RCRA identifiers are inconsistent with reported with name and address. Sometimes the inconsistency can be explained, for instance if the RCRA identifier corresponds to a destination facility and the reported name and address correspond to a hauling company (in which case the name and address associated with the RCRA identifier would be used), or vice versa. In other cases, the correct facility is not clear; generally the default is to use the reported name and address over the RCRA identifier.
- Wastewater treatment plants (POTWs) are often reported as "City Sewer Department" with the address being City Hall or the local water board office, rather than the physical plant. Corrections have been made for some of the larger cities/POTWs, where we manually assign POTW releases in a city to the largest POTW if the physical plant is not specified. However, some FRS records for POTWs erroneously locate them at offices, and some of those cases remain in these data.
- Because we have one set of off-sites for all years, the name of any given off-site may not match any off-site record. Off-sites can change ownership frequently, so there may be one off-site record to many reported records at the same site but with different names.
- False matches can occur if two reported records have a general address location like "Highway 34" and similar names like "Bob's Scrap Yard" and "Joe'' Scrap Yard." Off-sites often cluster in industrial areas and sometimes street address numbers are not reported.
- The algorithms employed have attempted to identify duplicate records, but not all duplication has been eliminated. Because FRS was used as a primary source, any duplication in that data system may get replicated in the off-site data. FRS strives to contain only unique facilities, but duplication exists, and is more prevalent in earlier years. Also, since state is used as a primary field in the RSEI matching algorithms, records that have the wrong state reported may not get matched to the correct record.

3.3 RY 2016 Results

For Version 2.3.6- RY 2016, the reported off-site records have been processed into 452,258 reported records (for all years), and collapsed into 40,326 relatively unique off-site facilities. Of these facilities, 14,542 were assigned FRS identifiers. The locational coordinates were obtained from the sources shown in Table D-2 below.

Table D-2
Sources for Coordinates for Off-Site Facilities

Source for location	Number of off-site facilities	Percent of total off-site facilities
Geocoding using ESRI (2017)	24,123	60%
FRS	12,387	31%
Manual (using satellite imagery)	2,550	6%
Geocoding using commercial company (TCS, 2005 and earlier)	977	2%
NA (no location available)	109	<1%
ZIP Code	104	<1%
Inactive FRS (coordinates taken from FRS entry that is no longer active)	64	<1%
DMR (EPA's Discharge Monitoring Report, using data from water discharge permits)	12	<1%

The general quality of reported TRI data has improved over the years, and EPA's FRS system is more complete and more accurate in later years as well. Table D-3 below shows that the percent of off-site facilities assigned FRS coordinates has increased from 31% for all years, to 55% for the most recent year (2016).

Table D-3
Sources for Coordinates for Off-Site Facilities for 2016 Only

Source for location	Number of off-site facilities	Percent of total off-site facilities
FRS	4,754	55%
Geocoding using ESRI (2017)	2,369	27%
Manual (using satellite imagery)	1,406	16%
Geocoding using commercial company (TCS, 2005 and earlier)	67	<1%
ZIP Code	51	<1%
Inactive FRS (coordinates taken from FRS entry that is no longer active)	17	<1%
DMR (EPA's Discharge Monitoring Report, using data from water discharge permits)	5	<1%
NA (no location available)	4	<1%

[revised 12/21/2017]