Technical Appendix F

Summary of Differences between RSEI Data and the TRI Public Data Release

Table of Contents

1 Introduction

The RSEI model currently uses a TRI data set that EPA uses in publishing the annual National Data Analysis. This data set is known as the data freeze, as it represents a static reference point for the TRI data, which is continually updated and revised. Currently, RSEI and the National Data Analysis both use the same data freeze. However, both the National Data Analysis and RSEI revise the data to increase its accuracy, and some of the revisions are performed for one dataset and not the other. The following sections discuss each of the types of the revisions made, and how they differ between RSEI and the National Data Analysis.

2 Offsite Facility Consolidation

In TRI, transfers to offsite facilities are reported by the facility transferring the chemical. Therefore, TRI contains multiple entries for the same offsite facility. Often, these entries are reported slightly differently by different facilities; for instance, one entry may be Columbia POTW at 1214 Main St., while the same offsite facility may have entry of Columbia Water Treatment Plant at 1214 Main Street. TRI lists entries like these separately. For modeling purposes, RSEI compares the text of all of the offsite facility entries to determine which entries actually represent the same facility, ultimately resulting in a set of unique offsite facilities. However, due to the large size of the original set, the comparison must be done automatically through a computer program, which is not as accurate as using human judgement. RSEI errs on the conservative side to avoid false matches, instead accepting a certain level of non-unique facilities to remain in the final set. For details on the method used, please see Technical Appendix D.

3 Offsite and Onsite Facility Latitude/Longitude Revisions

Facilities reporting chemical releases to TRI no longer report their latitude and longitudes. RSEI uses EPA's Locational Reference Tables (LRT) as the primary source of locational data. Additionally, RSEI geocodes the addresses of the consolidated set of offsite facilities to locate each offsite facility. TRI does not contain coordinates for offsite facilities. It should be noted, however, that in geocoding facilities the quality of the coordinates varies, from an exact match based on street address to a match based only on a 5-digit ZIP code. See Technical Appendix D for details on the deriving the locational data for both reporting facilities and offsite facilities.

4 Adjustments for Double-Counting

TRI facilities must report any chemicals that are transferred offsite to other facilities. These recipient facilities can dispose of the wastes in various ways, some of which are modeled by RSEI. Some of the offsite facilities that receive chemicals from TRI facilities are treatment, storage and disposal (TSD) facilities regulated under the federal Resources Conservation and Recovery Act (RCRA). These facilities were required to report their own releases to TRI for the first time in 1998. The new reporting requirement means that there is the potential for doublecounting releases that are transferred by a TRI facility to an offsite facility (and so reported to TRI), then released by a RCRA-regulated TSD (and reported again). The National Data Analysis uses facilities' RCRA identification numbers to match releases reported by TSD facilites to offsite transfers from other TRI reporters, and omits the matching offsite transfer from the summary. RSEI makes no such adjustments.

5 Selection of Industry Classification Codes

5.1 Introduction

The consideration of industrial sectors is an essential component of RSEI. The foremost reasons are that the 4-digit primary Standard Industrial Classification (SIC) code for a facility is used to estimate the stack air modeling parameters for those facilities for which facility-specific information is not available, and 4-digit SIC codes are used to estimate chromium speciation. The first two digits of a 4-digit SIC code define a major business sector, while the last two digits denote a facility's specialty within the major sector. Only facilities within certain SIC codes are subject to TRI reporting, based on the primary SIC code.

When submitting Form R reports by chemical, facilities are asked to always list the primary SIC code for the entire facility first, and then list up to five additional 4-digit SIC codes for other "establishments" (defined as "distinct and separate economic activities [that] are performed at a single physical location") which are associated with reportable releases and other waste management and source reduction activities. At least one Form R report is required for each chemical, but some facilities report multiple Form R reports for a single chemical to reflect the activities involving a TRI chemical at each establishment or group of establishments. This could lead to more than six 4-digit SIC codes being reported for a single facility.

While the instructions for reporting in the year 2000 have clarified the need to list the primary SIC code first in Section 4.5 of Form R, facilities commit errors in following these instructions and in providing valid SIC codes. When a facility reports more than one primary SIC code (based on all of its submitted Form R reports), RSEI performs a frequency analysis to assign a single primary SIC code to that facility. This analysis is performed on the "frozen" data set that is prepared by EPA prior to each year's Public Data Release. This data set represents not only the current year of reporting, but also any modifications or data corrections for all previous years.

Sector-based analyses are another important application of RSEI. The examination of the risk-related chronic human health impacts associated with various industrial sectors may be based on 2-, 3-, or 4-digit SIC codes. As noted above, RSEI lists up to six 4-digit SIC codes for each facility, with SIC Code 1 representing the primary industrial activity of the facility. When analyzing specific industrial sectors at the 4-digit level, users of the model can select: 1) all facilities which report that SIC code, and no other; 2) all facilities reporting that primary SIC code (although some may report other SIC codes as well); or 3) any facility that reported that SIC code (whether primary or not). Because some facilities engage in multiple activities at a single location and the aggregated chemical reporting by facility does not specify what portion of chemical releases are associated with each activity, analyses which are more inclusive in assessing a specific industry sector (e.g., option 3) will tend to overestimate the risk-related results by including releases from other activities at the selected facilities. This is an example of why it is important to perform additional analyses based on results from the rapid risk-screening afforded by RSEI. More highly aggregated results (such as "SIC Code Same 2 Digit") tend to reduce the uncertainty associated with some facility-specific analyses. This is because the definition of "SIC Code Same 2 Digit" requires that all 4-digit codes reported by a facility are covered by a single 2-digit code. This limits the range of activities those facilities are engaged in, but it also excludes facilities that operate several activities at a single location.

Beginning with RY2006, TRI requires facilities to submit North American Industry Classification System (NAICS) codes instead of SIC codes. RSEI treats NAICS codes in the same way as SIC codes. Because modeling parameters are still collected by SIC code, the historically-reported SIC codes for facilities are retained in the model along with the NAICS codes, and facilities can be selected by either code. For facilities that did not report prior to RY 2006 and therefore have no historically-reported SIC codes, RSEI uses the official Census NAICS-SIC crosswalk to derive SIC codes based on the facilities' reported NAICS codes.

5.2 Processing TRI Data

For each new year of annual release data (as reported in the TRI National Data Analysis), the RSEI methodology examines the current Form R reports for each TRI reporting facility (or those reports in the last year of reporting for facilities no longer reporting to TRI). There are normally over a million latest-year submissions, with a smaller number of submissions with SIC code information on them. The set of latest-year submissions with SIC code information is examined for the purposes of assigning a primary SIC code to individual facilities (either current reporters or historical reporters). Some small number of these submissions (usually less than 1%) have invalid primary codes (shown by the code INVA in the TRI data). Most facilities report only one SIC code (86 percent in RY 2000, the first year this analysis was performed). The rest report multiple SIC codes, but designate one code as the facility's primary SIC code. A few facilities report more than one primary SIC code. In these cases, RSEI performs a frequency analysis and selects the code most often designated as primary to be that facility's primary SIC code. Similarly, if more than five additional SIC codes are submitted, RSEI selects the five most frequently submitted (this is very infrequent).

Therefore, the main difference between RSEI's treatment of SIC codes and that of the National Data Analysis is that in cases where more than one primary codes is reported, RSEI selects the most frequently reported. There are also differences between how RSEI and the National Data Analysis aggregate SIC data to the 2-digit level; RSEI uses different codes, and differentiates between facilities that have multiple codes in original industries, new industries, or both. The following section provides the decision rules that RSEI follows in processing the TRI data.

5.3 Methodology for Assigning Primary and Secondary SIC Codes to TRI Facilities

Data will be pulled from the TRI_SUBMISSION_SIC table, which contains the following fields:

TRI_FACILITY_ID
DOC_CTRL_NUM
SIC_CODE
SIC_SEQUENCE_NUM
PRIMARY_IND

The following decision rules will be utilized in pulling the revised set of SIC codes:

1. Retain data for the facility's most recent submission year only.

2. Identify primary SIC codes. Position of SIC reported (SIC1-SIC6) is indicated by a combination of SIC_SEQUENCE_NUM (values = 1 through 6) and PRIMARY_IND (1 or 0). Primary SIC is determined by PRIMARY_IND=1. [Note: SIC_SEQUENCE_NUM is always 1 when PRIMARY_IND=1.] For facilities without a PRIMARY_IND=1 flag (19 facilities for year=2000), assign the code NR (“Not Reported or Not Applicable”) as “SIC Code 1" in RSEI.

3. All other SICs are considered "non-primary.” Non-primary observations equal to NA or INVA will be deleted. Primary observations equal to NA or INVA will be retained, at least temporarily (see below).

4. Perform a frequency analysis on SIC by TRI_FACILITY_ID for facilities with multiple “primary” SICs. The most frequently occurring SIC, if available, is chosen. If there is a tie, then the default ascending order of SICs will prevail and the lower numerical SIC code will be chosen (e.g., 2911<2912). SIC codes which appear valid (i.e., 4-digit number), if present, will be chosen over NA or INVA (i.e., NA or INVA may be most frequent, but there is also a valid SIC code). In the case where all primary SICs contain either NA or INVA, then see Step 6 below. [A new Flag Field, “MultiplePrimarySIC,” will be created as a variable in RSEI that identifies those facilities that have reported multiple primary SIC codes.]

5. The same frequency analysis will be performed on the non-primary SICs, and the top five will be retained (sorted by most frequently occurring, tie goes to lowest SIC). [For year=2000, no facility had more than six different SIC codes reported.]

6. The primary and non-primary list will then be combined to create a maximum of six SICs per facility. All primary SICs that are NA will be deleted and replaced with the code NR (“Not Reported or Not Applicable”) - the code INVA will be retained. If primary SIC was NA or INVA, then facility stack information will be based upon median stack parameters for all stacks.

7. The TRI_FACILITY table contains more records that the number of facilities associated with the TRI_FACILITY_SIC table. The extra facilities will have a blank SIC record in RSEI (as will those facilities with a primary SIC code of NA) that will be labeled NR for easier searching. For these facilities, the facility stack information will be based upon median stack parameters for all stacks.