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). With changes made to the model for Version 2.1, including more detailed air modeling close to the facility, and full-model results for surface water media, accurate locational data takes on additional importance.

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 (lat/long) on Form R ever. As facility reports of of lat/longs were subject to common reporting errors: 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 database of locational information, the Locational Reference Tables. Beginning with Version 2.1.5, RSEI adopts these coordinates for most reporting facilities as well.

For off-site facilities, the data quality is much worse, 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, and EPA has no central database for them. Little or no standardization is performed by the TRI program, therefore minor differences in an off-site facility record, such as a slight misspelling of the name, or "St." instead of "Street", can make two records look like two different facilities, when they are really the same.

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 lat/longs 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, Version 2.1.5, RSEI uses the best pick coordinates from EPA's Locational Reference Tables (LRT) for reporting facilities, and uses data from RSEI Version 2.1.3 only in those cases where the LRT data is not available. For off-site facilities, RSEI has created a master database of off-site facilities based on RSEI Version 2.1.3 that is updated with each new reporting year.

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 Agency-wide repository of locational information, the Locational Reference Tables (LRT). The LRT is a result of an EPA effort to obtain latitude/longitude coordinate information of documented origin for all of EPA's regulated facilities and sites. The LRT act as a storehouse for the actual locational data as well as the business rules that are applied to them in order to provide the most accurate information available for depicting the locations of federally regulated entities. While the LRT may contain several sets of coordinates for each facility, it designates one set as the "best pick:" according to a set of decision rules, the set of coordinates most like to accurately represent the location of the facility.

The database of all TRI reporting facilities for 1988-2005 includes 51,561 facilities[1]. Of the total number of facilities, 51,186 facilities were found in EPA's LRT system. EPA performed a county bounding check (to assure that the facility's best-pick LRT coordinates would plot within the reported county). The coordinates for 49,963 facilities passed the boundary test. These coordinates were applied to the TRI reporting facilities in the RSEI model. The MAD codes (Method, Accuracy and Description) codes that describe the quality of the coordinates for these facilities, are shown below in Table D-1. In the facility table, facilities assigned these coordinates are given the code "LRT" for LatLongSource and "Current" for the field LatLongTable.

[1] The Version 2.1.5 Public Release Installation CD contains TRI data from 1996-2005 only. However, the full data set is available for interested users. Throughout this appendix, statistics refer to the full data set, 1988-2005.

Table D-1
Method, Accuracy, Description (MAD) Codes for LRT Coordinates

MAD Code(s)	Description of MAD Code	Number of Records	% of All Facilities Using LRT Coordinates
Method*			
001, 002, 003, 004, 005, 006, 007	Address Matching	34,053	68%
008, 009, 010, 011	Census Block, Block Group, or Tract Centroid (1990)	102	0.02%
025	Classical Surveying Techniques	71	0.14%
012, 013, 014, 015, 016, 017, 028, 029	GPS	3,390	7%
018, 019, 020, 021, 030, 031, 032, 033,	Interpolation	5,335	11%
023, 024, 034, 035, 036	Public Land Survey	19	0.04%
027	Unknown	2,518	5%
Blank	No Code Given	4,475	9%
Accuracy (meters)*			
	0	2,748	6%
	1-10	2,626	5%
	11-50	3,815	8%
	51-150	32,814	66%
	151-500	1,138	2%
	501-1,000	562	1%
	1,001-10,000	209	0.42%
	10,001-25,000	1,676	3%
	>25,000	122	0.24%
	Blank	4,253	
Horizontal Datum			
002	NAD83	48,807	98%
Blank	No Value Available	1,156	2%

*The codes have been grouped for presentation purposes only. In the RSEI facility table they are listed individually.

There were 1,598 facilities without useable LRT coordinates. In these cases, additional coordinates were located and ranked in order of reliability (based on source, year and MAD codes). The same kind of county bounding test was performed on the highest-ranked

coordinates. If the coordinates passed the test, they were accepted; if they did not, the coordinates next in rank were tested. The ranked sources, with the number of facilities ultimately assigned coordinates from each source, are shown in Table D-2 below.

Table D-2
Sources for Coordinates for Facilities Without Valid LRT Coordinates

LatLongSource	Description	Number of Facilities	% of Facilities Without LRT Coordinates
2003 RSEI*	Coordinates taken from RY 2003 RSEI, Model Version 2.1.3. Field 'LatLongTable' describes data source of coordinates (see Table D-3).	811	51%
GDT	Coordinates taken from a geocoded data set used for RY 2003, address-match coordinates only.	9	1%
TRI	TRI preferred or submitted coordinates.	104	7%
ZIP	5-digit ZIP code centroid, either from an Arcview GIS program, or from a commercial web site.	657	42%

*Some of the coordinates in the RY 2003 RSEI model did not pas the county bounding test because a different series of quality checks was used in the that year.

The coordinates in 2003 were the result of a multi-step quality-assurance process using five different data sets. First, high-quality, well-documented coordinates were selected from the LRT system and from a dataset previously quality-assured by EPA (PREF94D). For facilities without coordinates from either of these sources, commercially geocoded coordinates were compared with the TRI reported data (submitted and preferred coordinates). In a series of steps, the best coordinates were selected from each facility. Facilities with no coordinates of acceptable quality were assigned ZIP code centroids based on their TRI-reported ZIP codes. Table D-3 below describes which data sources from RY 2003 were used in RY 2005.

Table D-3
Data Sources for Facilities with Coordinates Taken from 2003 RSEI*

LatLongTable	Description	Number of Facilities	% of Facilities with "2003 RSEI" Coordinates
QA_TRI	Ultimate source is TRI, as reported either in Pref94D or in the 'Facility' table in the current year TRI data freeze. Compared to geocoded coordinates and determined to be more accurate.	303	37%
QA_GDT	Ultimate source is geocoded data using GDT software, performed either by EPA in 1998 or by TCS in the current year for EPA. Compared to TRI lat/longs and determined to be more accurate.	199	25%
HQ Prefer	Ultimate source is 'PREF_LAT' and 'PREF_LONG' in Pref94D. Originally selected as High Quality Preferred Lat/longs; no comparison to geocoded results was performed.	142	18%
HQ LRT	Ultimate source is the high-quality LRT coordinates. Originally selected as High Quality Lat/longs; no comparison to geocoded results was performed.	95	12%
TRI_NOGDT	Ultimate source is TRI, as reported either in Pref94D or in the 'Facility' table in the current year TRI data freeze. One of approximately 400 facilities (in RY 2003) with TRI coordinates that did not have geocoding results to compare against.	34	4%
QA_LRT	Ultimate source is address-matched records from the LRT, which was combined with the geocoded and compared to TRI lat/longs and determined to be more accurate. or the	17	2%
ZIP	Zip code centroid used, found on internet look-up table.	7	1%
TRI	TRI data adopted during final QA process-facility's initial coordinates failed plotting.	5	1%
LRT_NOTRI	Ultimate source is LRT data. Adopted without comparison because no TRI coordinates were available.	4	0.49%
GDT_NOTRI	Ultimate source is geocoded data using GDT software, either by EPA in 1998 or by TCS in the current year for EPA. Adopted without comparison because no TRI coordinates were available.	3	0.37%

Table D-3
Data Sources for Facilities with Coordinates Taken from 2003 RSEI*

LatLongTable	Description	Number of Facilities	% of Facilities with "2003 RSEI" Coordinates
LRT	LRT coordinates adopted during final QA process- facility's initial coordinates failed plotting.	1	0.12%
Manual	Coordinates manually revised after plotting coordinates	1	0.12%

For more information on the method used in RY 2003, see Appendix D from RY 2003, Version 2.1.3.

3 Off-site Facilities

Data quality issues with the set of TRI off-site facilities are longstanding and serious: most notably that unique IDs are not used by TRI, and the addresses are not reported by the facilities themselves, but by those facilities that transfer waste to them. Given this, the accuracy of the reported addresses is questionable. In addition, because many different reporting facilities may be transferring their waste to the same facilities, there are many instances of the same facility being reported with many different permutations of name and address.

In RSEI Version 1.x, 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. 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.

The method described above was last used in RSEI Version 2.1.3 (RY 2003) to create a master database of relatively unique facilities numbering approximately 47,000 (reduced from an original set of over 3 million). Approximately 35% of these facilities had an address match more accurate than a ZIP+4 match.

For RSEI Version 2.1.5 (RY 2005), the master database created for Version 2.1.3 was used as a starting point. The new RY 2005 TRI data (for all years) was matched back to the master database, and any unmatched facilities were 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 any previous year. Exact matching was performed first, and then a simplified version of the fuzzy matching used to create the master database was used to match the remaining facilities. The sections below

describe the process in detail: Section 3.1 describes how the master database was created- the collapse, geocoding and fuzzy matching. Section 3.2 describes how the current year data was matched to the master database.

3.1 Creating the Master Database

There are several data processing steps in determining unique facilities and their coordinates. First, in order to best determine unique facilities, the facility records were collapsed from approximately 3 million to almost 300,000 by removing the exact duplicates. Second, in order to expedite and improve the off-site facility locating process, TCS[2] geocoded the data and reported match rates. Finally, the geocoded off-site facility data was further collapsed in order to remove non-exact duplicates and determine truly unique off-site facilities and their addresses.

3.1.1 Collapsing Reported Off-site Facilities

In Version 2.1.3 (RY 2003), there were approximately 3 million off-site facility records in TRI. However, many of these facility records actually represent the same facility; they were just reported in slightly different ways by the facilities transferring chemicals to them. In addition, approximately 1 million records were blank or not viable records. In order to make the geocoding process more efficient, it was necessary to first collapse the list of all reported off-site facilities into possible unique facilities. The first collapsing procedure removed all records that were not viable along with all of the records that are exact duplicates. This first stage collapsed the off-site facility records from approximately 3 million to approximately 300,000.

Further collapsing, using algorithms in SAS to match addresses where the content is the same but the form is different (i.e., St. instead of Street), can bring the count down to approximately fifty thousand. However, the risk with this second collapse is in matching records that are not exactly the same, and also in picking one address form to represent that facility, where another form might be better for geocoding purposes. Therefore, to decrease potential error in geocoding unique facilities, the almost 300,000 facility address records were sent to the geocoding service.

3.1.2 Geocoding the Off-site Facilities

TCS evaluated the 300,000 off-site facility address records. Their geocoding efforts resulted in a 50% street address match; 0.18% ZIP+4 centroid match; 0.16% ZIP+2 centroid match; nearly 47% ZIP code centroid match; and nearly 3% unmatched records. At this point in the process, this numbers may be misleading, since many of the 300,000 facilities were duplicates. Presumably, some portion of the ZIP code matches and unmatched facilities have problematic street addresses that may be "corrected" by accepting the better data of some other record of the same facility.

[2] Thomas Computing Services, Lantana, FL.

3.1.3 Collapsing Off-site Facilities After Geocoding

A "fuzzy" matching SAS program (FIND_UNIQUE.SAS) was used to identify additional duplicate records that belong to a single unique facility. The term "fuzzy" refers to logical systems that do not require exact equality of two values in order to classify the two values as equal. In the name matching application, FIND_UNIQUE.SAS assigns two records to the same unique facility even if some identifying fields do not match exactly. This approach accommodates misspelled words and inconsistencies in how a facility might report its identifying information over time. For example, "DuPont," "Du Pont" and "E.I. DuDont" might all refer to the same facility. FIND_UNIQUE.SAS identifies a possible match based on similarity rather than exact equality in the name field and then decides whether to match the various spellings by examining the address fields.

Fuzzy matching always introduces the possibility of error. Two records may be matched that do not in fact belong to the same unique facility. Therefore, some discretion was applied in varying the program parameters and performing manual checks to balance two competing outcomes: a greater number of good/high confidence matches versus a greater number of erroneous matches.

The major parts of FIND_UNIQUE.SAS are:

1. Cleaning and conditioning the data;
2. Identifying a set of best names and addresses;
3. Matching records within the set of best names and addresses;
4. Finding indirect matches, where two records are matched not to each other but to a common third record.

The following sections describe in detail the SAS program and its application.

3.1.4 Cleaning and Conditioning the Data

The first part of FIND_UNIQUE.SAS corrects common spelling errors and inconsistencies and prepares the data for the matching algorithms. Data cleaning begins with the removal of extraneous characters, regularization of spaces and conversion of all letters to upper case. Then, words that occur frequently but do not aid in matching are deleted. These words include "COMPANY," "LIMITED," "POST OFFICE BOX," "NOT AVAILABLE," and numerous other words and their associated abbreviations and variations. If such words remain in the match fields, then a name such as "COB CORPORATION, P.O. BOX 2" would appear very similar to "AC CORPORATION, P.O. BOX 10." The conditioning process converts the two names and addresses to "COB, 2" and "AC, 10," respectively.

Where relevant words commonly appear in various forms, the conditioning process substitutes a single form. For example, "NAT'L" and "NATIONAL" are both converted to "NATL." A frequency analysis and visual review of words in the database led to some regularizations of facility names, such as "ADM," "A.D.M." and "ARCHER DANIELS MIDLAND," or "EMPAC," "EMPACK" and "EMPAK."

For computational purposes, FIND_UNIQUE.SAS adds a leading blank and a trailing blank to each name and street address.

One example of how the conditioning might change a name field follows. If the reported name of a company is (the misspelling of "environmental" is intentional):

E	N	V	I	R	O	M	E	N	T	A	L		B	A	N	T	E	R			C	O	R	P	.

then the cleaned and conditioned version of the name would be:

	E	N	V	I	R		B	A	N	T	E	R	

The conditioning process concludes by correcting the state field when possible, based on the ZIP code field. FIND_UNIQUE.SAS does not assume that the ZIP code is correct whenever the reported state and ZIP code conflict. However, it does identify certain values in the state field as particularly susceptible to error. These suspect values were identified by checking reported state codes against reported city names and ZIP codes, using the *1996 World Almanac*. The conditioning process uses the state that corresponds to the reported ZIP code when the reported state is particularly susceptible to error or is not a valid state abbreviation. Table D-4 lists states that TRI reporters frequently misreport.

Table D-4
Suspect State Abbreviations

Reported State	Possible Actual State	Reported State	Possible Actual State
AR	AZ	MA	ME
AK	AR	MI	MS
AS	AR	MI	MO
CA	GA	MI	MN
IA	IN	MS	MO
IA	ID	NE	NV
II	IL	ON	OH
KT	KY	OP	OH
KU	KY	RH	RI
LA	AL		

State codes were discarded in favor of the state corresponding to the reported ZIP code if and only if:

1. The reported state is listed in the "Reported State" column, and
2. The state corresponding to the reported ZIP code is the state listed in the "Possible Actual State" column of the same row.

For example, if the reported ZIP code is "85607" and the reported state is "AR," then the program corrects the state to "AZ." However, if the reported ZIP code does not begin with "85," then this section of the program makes no change to the state code.

Another section of the conditioning process corrects state codes in certain city name and state code combinations. For example, where the reported city name is "BALTIMORE" and the reported state is "MA," the SAS program changes the state to "MD." The program also changes Canadian province codes to "CN."

3.1.5 Identifying a Set of Best Names and Addresses

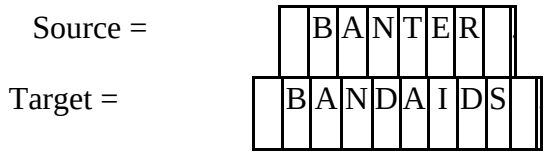
The purpose of the second part of FIND_UNIQUE.SAS is to reduce the number of records to be matched as quickly as possible. Since the time required to match all records in a dataset to each other increases exponentially as the number of records increases, it is important to perform preliminary matching using a simpler method where possible. FIND_UNIQUE.SAS does this by sorting records by facility name and comparing adjacent records. Thus, this early round of matching compares each record only to the preceding record and finds a match only in cases where the similarity is quite strong.

Specifically, the program sorts the data by the first ten non-blank characters in the facility name. If a reported facility name begins with the same ten characters as in the preceding record, the program compares the street addresses and ZIP codes and assigns three scores that measure the closeness of match in these location fields. If the scores exceed specified thresholds, then the program matches the two records to a single facility. Similarly the program then sorts the data by the first ten non-blank characters in the facility street address and compares the names in adjacent records.

FIND_UNIQUE.SAS calculates three scores that measure how well two names or two street addresses match. The definitions below use two new terms: *source* and *target*. The *source* is the set of words – i.e., name or street address – for which a match is sought. The *target* is the set of words that is being compared to the source. In the current part of the program, which compares adjacent records only, it does not matter which comparison value is designated the target and which is designated the source.

1. **Match Score**: The match score is the weighted proportion of letter pairs in the source also found in the target. A score of 0 means that no letter pairs in the source occur anywhere in the target. A score of 1 means that 100 percent of the letter pairs in the source also occur at least once in the target.

Example:

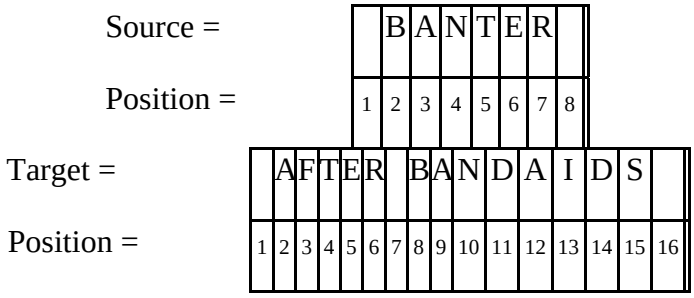


The eight letter pairs in the source are: _B, BA, AN, NT, TE, ER and R_, where "_" represents a blank. Of these, _B, BA and AN also appear in the target. Therefore, the unweighted match score is 3/8 or 37.5 percent.

FIND_UNIQUE.SAS introduces variable weights to allow the user to apply expectations about where errors are most important. In the current application, weights for letter pair matches decline exponentially so that matches near the beginning of the target are more valuable than later matches. The use of this model was based on an informal examination of the data.

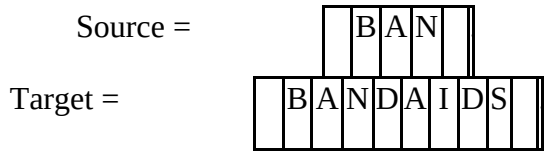
2. **Position Score**: The position score measures similarity in the sequencing of letter pairs. The reason this is important is that the match score gives credit for a letter pair match regardless of where the letter pair occurs. In the above example, if the target had been "AFTER BANDAIDS," the match score would have increased to 7/8 or 87.5 percent because the letter pairs TE, ER and R_ occur in "AFTER."

   The position score depends on where a matched pair is with respect to the first matched pair. In the following example, the first pair matched is _B, which occurs in the target at position 7.



The position score is similar to the root mean square (RMS) algorithm commonly used to measure error in diverse situations. A position score of 0 indicates that the matched letter pairs occur exactly in order and at the same relative positions in both the target and source. Higher scores indicate poorer matches.

3. **Leftover Score**: The leftover score measures the percent of the target that is not matched to any letter pairs in the source. The leftover score helps compensate for the tendency of the previous two scores to overmatch short sources. To illustrate, in the following example, the match score is 100 percent and the position score is 0 – both optimum values.



The leftover score measures the percentage of the target that is not matched to any letter pairs in the source. As in the match score, the leftover score uses a weighting system to give more weight to letter pairs that are most useful in discriminating between spelling variations and non-matching names. The best value is a leftover score of 0, and the worst value is 100 percent.

The comparison of adjacent records ends with one more iteration: by five digit ZIP code. The first two iterations examine records sorted by ten characters of the name and then by ten characters of the street address. The ZIP code iteration sorts all the records by five digit ZIP code and then compares adjacent records within each ZIP code for goodness of fit in both the name and street address fields.

FIND_UNIQUE.SAS allows the user to specify separate threshold values for each score and for each match field. In the iteration where names begin with the same ten non-blank characters, the thresholds for street address matches are relaxed slightly when five digit ZIP codes match exactly.

3.1.6 Matching Records Within the Set of Best Names and Addresses

The most powerful part of FIND_UNIQUE.SAS compares each record within a dataset to every other record, but it is also the slowest. For this reason, it is important to use Part II first to match closely-related records through comparisons of adjacent records.

Part III simultaneously scores and evaluates four match fields: name, street address, state and ZIP code. The program compares each record (source records) to all other records (target records). If the source record matches multiple target records, then the source is assigned to the target with the most frequently reported identifying data.

For example, assume that all of the following records match each other and they are all in the same state:

Name	Street	ZIP	Frequency
BANTER	10 MAIN ST.	12345	10
BANTER	P.O. BOX 40	12345	2
BATNER	10 MIAN ST.	72345	1

The "Frequency" column indicates how many times each version of the identifying data occurs in the database. Ten times, the facility reported its name as "BANTER," its street address as "10 MAIN ST." and its ZIP code as "12345." Since this combination of identifying information occurs more frequently than the other two, FIND_UNIQUE.SAS assigns "BANTER," "10 MAIN ST." and "12345" to all thirteen records.

As part of this step of the program, the data are exported to an Excel spreadsheet, where some manual matches and corrections supplement the SAS matching. The data are then imported into SAS again, where processing continues.

3.1.7 Finding Indirect Matches

In the final part of FIND_UNIQUE.SAS, the program consolidates all the information about matching records and finds a set of unique facilities. In particular, Part IV finds indirect matches, where record A matches record B and record B matches record C but a comparison of A to C fails the goodness of fit thresholds. In this case, A and C should be matched even though they fail in the direct comparison.

In the following hypothetical example, the first record might match the third record by a four letter-pair match in the name field (_B, BA, ER and R_) with an optimal position score of 0, combined with an exact match in the street address field and an exact five-digit ZIP code match.

Name	Street	ZIP
BANTER	10 MAIN ST.	123450040
TRENTON PLANT	P.O. BOX 40	12345
BATNER TRENTON PLT	10 MAIN ST.	12345

The second record might also match the third record based on good match and position scores in the name field and an exact match in the ZIP code field. Therefore, all three records pertain to a single unique facility, even if the first and second records might fail to match using a direct comparison.

3.1.8 Identifying and Assigning the Best State Match

The fuzzy matching program results in two output files: (1) the original file of offsite facilities in which each observation is labeled with the identification number ("ID_MATCH") of a unique off-site address for which it matches (approximately 3 millions records), and (2) a file which represents the legend of unique off-site records based on the ID_MATCH identification number. The latter file contains the records used in the display of off-site facility information in the RSEI

model, such as the best name and address or locational coordinates determined from earlier routines of the fuzzy matching program. However, this unique addresses file does not output the best state associated with each facility as it does for name, street, city, and zip.

To develop a state value for each of the unique off-site addresses, the facilities were plotted to retrieve the state in which they mapped to. Similarly, the state corresponding to the best zip value was also retrieved (i.e. BEST_ZIP as determined by the fuzzy matching program). A separate analytical routine was then performed in SAS to determine the BEST_STATE value. This analysis required the following preparatory procedures:

1. The original file of approximately 3 million reported off-site facility records was sorted based on the unique off-site identification number it was assigned;
2. The frequency of the reported state within each ID_MATCH group of records was calculated;
3. The state most frequently reported for each ID_MATCH group was retained.

As a result of these procedures, three different fields containing various state values could be compared for each unique off-site facility: the plotted state, the state corresponding to the BEST_ZIP, and the state reported with the highest frequency. The following rules were applied in their comparison and in the determination of the final BEST_STATE value:

1. If the plotted state = BEST_ZIP state = reported state, then the state was considered valid;
2) Alternatively, if any two of the three fields matched, then that state value was used;
3) Finally for instances in which the latitude or longitude = 0 or was blank, no plotted state could be determined, so the reported state, if available, was used.

Of the unique addresses that resulted from running the collapse program, less than 200 were not resolved with a BEST_STATE based on this methodology. These remaining off-site facilities were exported into Excel and manually evaluated since the three state fields were in disagreement. The three state fields were used as a guide and provided context during this manual verification of BEST_STATE. Some of the reasons for how the BEST_STATE was determined for these records, included:

1) Some combination of city/zip/state was confirmed on www.usps.com;
2) The state was in the facility name;
3) Searched on some combination of the name/street address/city on www.google.com for an exact match.

Among these records were some facilities for which the lat/long coordinates were deleted. Reasons for deleting lat/long coordinates included: (1) they were erroneous (e.g., the facility was actually located in the UK or Canada, or the search on name and address revealed a different state that was NOT adjacent – if the state was adjacent, the lat/long was not deleted), or (2) no supporting data to make a determination could be found using any of the above mentioned methods. Finally, only three records resulted with no state value at all; and those lat/longs were also deleted because two were in the UK and one was located in Canada. These records were then re-appended to the larger off-site address file resulting in the complete set of unique off-site addresses.

3.1.9 Results

The geocoding procedure and the SAS algorithms collapsed the number of off-site records from the initial 3 million to a final set of 46,768 records. As shown in Table D-5, approximately 34 percent of the unique facilities were matched to high-quality street addresses. Note that each unique address may represent multiple reports of off-site transfers from multiple Form R's.

Table D-5
Final Status of All Off-site Facilities in Master Database

Coordinate Matches	Number of Records	Match Percentage
Street address (including hand matches)	16,096	34%
ZIP+4 centroids	138	<1%
ZIP+2 centroids	74	<1%
5-digit ZIP code centroids	27,113	58%
Unable to geocode or match to geocoded	3,347	7%
Total Unique Facilities	**46,768**	

3.2 Matching Current Year Data to the Master Database

The RY2005 TRI data (for all years) that was processed for RSEI Version 2.1.5 includes more than 5 million offsite records. As described above, these include many exact duplicates, invalid records, blank records, and records that are not exactly duplicative but refer to the same facility. The last step in processing the off-sites involves 1) removing the invalid records, blank records, and exact duplicates; 2) matching the exact duplicates back to the master database; 3) fuzzy matching the records that do not match exactly; and 4) adding the unmatched remainder to the master database.

Exact and fuzzy matching are both necessary because there are no unique IDs assigned in TRI to the off-site facilities, and with every reporting year update, all of the unique keys that link the data tables in the TRI data change. So there is no way to link an off-site facility from the RY2005 TRI data, even for an older year, into the master database except by matching the text fields. And a link must be created between the RY2005 record and the master database, because the link to the reporting facility goes through the RY2005 off-site record. If this matching process were not performed, there would be no way to tell what reporting facilities were transferring chemicals to these off-site facilities, or what quantities they were transferring.

3.2.1 Creating the Collapsed Set

The entire set of off-site facilities for all years of TRI reporting (1988-2005) included in the RY2005 TRI data had more than 5 million records. Once the invalid or blank records were removed, and the exact duplicates removed, the collapsed set contained 372,713 unique records. These records are unique in the sense that there are no exact duplicates; however, many of the records in reality refer to the same off-site facility.

3.2.2 Exact Matching

The set of RY2005 off-site facilities with the blank/invalid/duplicate records removed was matched to the RSEI Version 2.1.3 master database. Approximately 350,000 records from the collapsed set were matched exactly to records in the master database.

3.2.3 Fuzzy Matching

The approximately 22,000 remaining off-site records from the collapsed RY2005 set were matched to the master database using a fuzzy matching program similar to the one described above. This process matched approximately 6,000 records from the collapsed set to records in the master database. This left an unmatched remainder of approximately 16,000 records. This remainder was examined by hand. Approximately 300 matches were made by hand. The remaining unmatched off-site records were added to the master database.

3.2.4 Results

The following table shows the results for the new master database, with the unmatched current-year off-site facilities added. Note that because no new geocoding has been done, the absolute number of everything but the last category has remained the same. Only the number of "unable to geocode or match to geocoded" facilities has changed.

Table D-6
Final Status of All Off-site Facilities in New Master Database

Coordinate Matches	Number of Records	Match Percentage
Street address (including hand matches)	16,096	26%
ZIP+4 centroids	138	<1%
ZIP+2 centroids	74	<1%
5-digit ZIP code centroids	27,113	43%
Unable to geocode or match to geocoded	19,359	31%
Total Unique Facilities	**62,780**	