US EPA ARCHIVE DOCUMENT

US EPA ARCHIVE DOCUMENT

Appendix A

Study Design and Survey Data Collection and Processing

Table of Contents

US EPA ARCHIVE DOCUMENT

Appendix A

Study Design and Survey Data Collection and Processing

This appendix describes the overall study design, implementation of the survey data collection, and the preparation of these data for the analyses described in Chapters 2 and 3 of this document. Section A.1 explains the statistical study design and the development of the original sampling frame, or list of facilities from which EPA selected facilities for the study. It also provides details on the design and implementation of the screener and long surveys developed to collect study data.

The remainder of this appendix provides details on the creation of electronic long survey databases and their use in providing data for data exploration (Chapter 2) and the risk assessment (Chapter 3). This includes how the long survey data were entered and archived in electronic formats (A.2); how facility-specific data supplemental to the survey were collected (A.3); how the data were processed for consistency and to provide inputs for modeling and data analysis (A.4); and a description of the statistical methodology used to weight up survey data and risk assessment results to estimates applicable to the entire population of surface impoundments with constituents or pH of concern (A.5).

A.1 Statistical Study Design and Survey Implementation

As described in Chapter 1, the Surface Impoundment Study is directed towards identifying and characterizing certain nonhazardous surface impoundments. An eligible impoundment is one that meets the criteria in the legislation or consent decree, regarding the wastes managed, and meets additional scope criteria described in Chapter 1, notably extreme pH conditions (i.e., less than 3 or greater than 11) or that one or more of 256 chemicals are present. In order to identify a representative sample of facilities with impoundments meeting the study criteria, EPA developed a two-phase or double-sampling design. In the two-phase design EPA collected some information on a relatively large sample of facilities through a screener survey and then used this information to select a second-phase subsample of facilities for which detailed facility and impoundment data were collected using a longer survey questionnaire.

EPA decided to collect data in the long survey for all eligible impoundments at the facilities in the sample. This decision meant that EPA would obtain an approximately equal probability sample of impoundments within primary sampling strata because facilities were selected with approximately equal probabilities within primary sampling strata (direct discharge facilities with high priority SICs, direct discharge facilities with low priority SICs, and zero discharge facilities). In addition, by collecting data for all eligible impoundments at sample facilities, EPA could overlay risk estimates for the separate impoundments to produce an integrated assessment of risk at the facility level. Facility-level risk estimates are important if a facility's nearby residents can be exposed to emissions from multiple sources (impoundments).

US EPA ARCHIVE DOCUMENT

A.1.1 Sampling Frame and Stratification

The sampling frame for nonhazardous industrial surface impoundments was based on available data identifying and listing facilities with surface impoundments that might meet the study criteria. Three primary sampling strata were defined for selection of facilities for the screener survey based on the facility's regulatory status under the Clean Water Act:

- **Direct discharge (Section 402) impoundments** treat waste in systems that ultimately discharge directly into surface waters. This subpopulation is regulated under CWA Section 402, which requires National Pollution Discharge Elimination System (NPDES) permits for all facilities that discharge to "waters of the United States."

- **"Zero discharge" impoundments** are not designed to discharge waste into the environment except through infiltration into soil or evaporation. Facilities that use infiltration or evaporation ponds for waste treatment or disposal may be regulated under a variety of state laws addressing both waste handling and groundwater protection. Specific regulations regarding these impoundments vary by State.

- **Indirect discharge (Section 307) impoundments** treat or hold waste prior to discharging to a publicly owned treatment works (POTW). Facilities that discharge significant waste flows to POTWs must comply with federal and local standards for pretreatment of waste in order to prevent adverse impacts on the public treatment plants. Local POTWs are the principal permitting authorities for CWA Section 307 facilities.

There are major differences in the sources and availability of data for defining the sampling frame for each of these subpopulations, and this affected the sampling frame and stratification for each. For the direct and zero discharger subpopulations, sampling frame data were adequate to use a stratified simple random sampling design, in which facilities were randomly selected from strata without replacement, and data were collected for all eligible impoundments at the facilities in the sample. For the indirect discharger subpopulation, limited sampling frame data led to a purposive (non-random) sample of facilities identified using anecdotal information. The chosen designs mean that the direct discharger and zero discharger samples are representative (although the sample is less representative for zero dischargers because their sampling frame was incomplete for some states). However, the non-random indirect discharger sample may not be representative.

A.1.1.1 Direct Discharge Facilities and Impoundments. The Permit Compliance System database (PCS) contains all facilities releasing waste to surface water, including those operating surface impoundments. EPA used this database as the sampling frame for the direct discharger subpopulation. EPA took the records in this database, as of late 1997, for facilities having SIC codes that were defined as the study's scope.

US EPA ARCHIVE DOCUMENT

Each PCS record related to a given discharge point, so a facility with multiple discharge points had multiple records. In addition, facilities with multiple permits were listed more than once. EPA combined multiple records for a given facility into one record only when it was quite clear that the records were for the same facility. EPA merged up to three different permits into a single facility-level record. The final count of records for facilities with SIC codes in the study's scope was 43,050.

EPA partitioned the sampling frame into three primary sampling strata, defined as:

1. Facilities in high-priority SICs (26, 2819, 2824, 2834, 2869, 2897, 2911, 30, 33, or 36)
2. All other facilities with in-scope SICs
3. Six pilot study facilities

Stratum 1, the high-priority SICs, were expected to contain a higher proportion of facilities that use surface impoundments to manage decharacterized wastewaters. Hence, this stratum was sampled at a much higher rate than Stratum 2, the remainder of the in-scope SICs, to ensure that the screener survey would include an adequate number of facilities using surface impoundments to manage decharacterized wastewaters. Each of these strata was then partitioned into substrata based on SIC codes, and the substrata were all sampled at the same rate within each primary sampling stratum. Hence, a stratified simple random sample of 2,000 facilities was selected from 15 sampling strata plus all six pilot study facilities.

A.1.1.2 Zero Discharge Impoundments. In this study, EPA defined zero discharge impoundments as those that are neither permitted under Section 402 of the Clean Water Act to release to surface water, nor permitted under Section 307 to pretreat waste before releasing it to a publicly owned treatment works (POTW). Because states are the primary regulators of zero discharge impoundments, state databases were the principal source of information on these impoundments. In addition, EPA identified some zero discharge impoundments in the Toxics Release Inventory (TRI) and the Aerometric Informational Retrieval System (AIRS) Facility Subsystem (AFS) databases. By assembling information from TRI, AFS, and available state data, EPA developed a list of 5,807 zero discharger facilities. EPA stratified the sampling frame according to general categories of completeness for the different state and federal data sources, and according to high and low priority SIC codes. A stratified random sample of 250 facilities was selected in the first stage using the same sampling rate for all strata except for the Oklahoma database of private sewage treatment facilities. EPA expected this group of facilities to be mostly out-of-scope, and if in-scope, to be relatively homogeneous. Hence, EPA sampled them at one-half the rate used for the other strata.

A.1.1.3 Indirect Discharge Impoundments. Section 307 of the Clean Water Act regulates indirect discharger facilities, which "pretreat" or hold waste prior to discharging it to a POTW. The total population of facilities required to pretreat their waste prior to discharge to a POTW is over 30,000; they are regulated and tracked by the approximately 2,000 POTWs that receive this pretreated waste. However, the POTWs do not routinely collect data on surface impoundment use by their pretreating customers, so there is no consistent data source from which to identify indirect dischargers that use surface impoundments. In addition to the 30,000 pretreaters, there

US EPA ARCHIVE DOCUMENT

are an unknown number of other indirect dischargers who are not required to pretreat their waste (and who discharge to POTWs outside the national pretreatment programs). Theoretically, any of these indirect dischargers could potentially use surface impoundments to store wastewater before discharging it. Based on information from EPA Regional pretreatment coordinators, it appears that only a very small proportion of these indirect discharger facilities are likely to use surface impoundments. From this information, EPA assembled a group of 35 facilities likely to operate indirect discharge impoundments and used this as a purposive sample to characterize the indirect discharger subpopulation.

A.1.2 Screener Survey Implementation

The sampling frame and stratification scheme led to a total of 2,285 facilities being selected for the screener survey, a short questionnaire designed to identify facilities and impoundments that meet the study criteria and thereby provide the sampling frame data for the long survey (see Attachment A1). These facilities included 2000 direct dischargers, 250 zero dischargers, and 35 indirect dischargers. Implementing the screener survey involved identifying and removing ineligible facilities from this sample, identifying and locating survey respondents to obtain the highest possible response rate, adjusting facility weights to account for survey nonresponse, and data entry, quality control, and processing.

A.1.2.1 Removal of Ineligible Facilities from Sample. The facilities chosen for the direct and zero discharger samples included a number of facilities that were outside the scope of the study. In many cases, the facilities selected in the sample were private residences or retail businesses that did not have activities in the SIC code range defined for the study, even though they were listed on the sample frame as having eligible SIC codes. EPA confirmed these sample members' status as "ineligible" using other data sources, and removed them from the sample. For the direct discharger sample, EPA determined that 138 facilities among the 2,000 direct dischargers were ineligible, and 74 facilities among the 250 zero dischargers were ineligible, resulting in 2,038 direct and zero discharger facilities in the sample.

A.1.2.2 Identifying Screener Survey Respondents. Once eligible facilities were identified, EPA needed to identify and locate the survey respondents. EPA found that the PCS data and the zero discharger frame data were frequently missing mailing address, location, and contact information. Of the 2,038 direct and zero discharger facilities, EPA found mailing addresses for 1,982. EPA found mailing addresses for all 35 indirect dischargers. Thus, the screener survey was mailed to 2,017 facilities.

The screener survey was mailed in February 1999. A large proportion of the surveys went to the appropriate individuals and were returned within the requested 45-day time frame with adequate information. EPA found that a significant proportion of the sample facilities had either changed ownership or names, or had ceased to exist during the period between 1990 and 1999, and required further tracing to locate individuals who were knowledgeable about those facilities' impoundments. Thus the screener survey data collection extended over a six-month period.

US EPA ARCHIVE DOCUMENT

EPA also needed to address the sampling frame multiplicity problem described in Section A.1.1.1. Any facilities with multiple permits that did not get merged into a single facility-level record on the sampling frame had multiple chances to be selected into the sample. Because being listed on the sampling frame more than once increases a facility's probability of selection, EPA needed to correct for this multiplicity, or being present on the sample frame more than once. EPA listed on the screener survey all wastewater permits that had been used to define the facility on the sampling frame, and asked each facility (on the screener survey) to list any additional permits that had been active for the facility at any time since June 1, 1990. In addition, EPA set up a computer-assisted telephone interviewing (CATI) application to call the screener survey respondents and probe for any additional permits that had not been listed on their screener survey responses. EPA then used both the responses to the original screener survey question and the responses to the supplemental CATI interviews to make weight adjustments for frame multiplicity (described in Section A.5).

EPA also used a CATI version of the mail survey to increase the response rate for approximately 100 of the mail screening survey recipients who did not provide their responses in a timely manner.

A.1.2.3 Screener Survey Weight Adjustments. For each of the 1,982 direct and zero discharge facilities mailed a screener, an initial sampling weight was computed by dividing the total number of facilities in the stratum (frame count) by the number of facilities selected into the sample from the stratum. Frame counts, sample sizes, and initial sampling weights for each stratum are provided in Section A.5, along with the detailed statistical methodologies. Sampling weights were not computed for the sample of 35 indirect discharger facilities because the sample was purposively selected and the survey results cannot be statistically extrapolated to any larger population.

Next, EPA needed to adjust these initial sampling weights for the sampling frame multiplicity described in Section A.1.1.1. After considerable data cleaning, multiplicity (number of linkages to the sampling frame) was determined for each facility that responded to the screening questionnaire. Because frame multiplicity must be known for every sample facility, not just the responding facilities, EPA computed, for each direct discharger sampling stratum, the average multiplicity among the respondents and used this value to impute multiplicity for each nonresponding facility. These multiplicity estimates were then used to adjust weights as described in Section A.5.

Weight adjustments to minimize bias due to survey nonresponse are based on models for the probability of not responding, using data that are available for both the respondents and the nonrespondents. For nonresponding facilities, EPA knew only the sampling stratum, and thus, EPA used sample-based ratio adjustments based on the sampling strata (Kalton and Maligalig, 1991). The nonresponse adjustments were defined only for the direct and zero discharge facilities because the indirect discharger sample was not a probability-based sample. Statistical details on facility weights and weight adjustments, including item-specific adjustments made during data analysis, can be found in Section A.5.

US EPA ARCHIVE DOCUMENT

A.1.2.4 Screener Survey Data Processing. In the screener survey (U.S. EPA, 1999b), EPA collected data on the facility's use of surface impoundments, and on the activities that were the source of the waste in the impoundment(s). For those facilities that reported using impoundments that met the criteria for being in the study, EPA also collected data on the facility's status as a hazardous waste generator, whether any impoundments contained decharacterized waste, whether the impoundments were used to treat waste biologically, and whether the impoundments had permanently stopped receiving waste.

When the screener surveys were returned, a coding clerk assigned codes for the closed-ended questions, according to a predetermined code list for the various response options. The surveys were then grouped into batches for tracking the hard copy survey forms and to subdivide the overall data entry task into more manageable segments. Double-extraction/double-entry was used to minimize data entry errors. Each coded response was entered into the data file twice, by different data entry staff, the files were electronically compared, and any differences were resolved by referring to the hard-copy forms.

EPA also performed a check on the responses indicating that there was no impoundment at the facility that met the study criteria. To perform the check, EPA drew a systematic random sample of every tenth response that indicated an absence of impoundments meeting the study criteria. For these responses, EPA obtained independent data (generally, state environmental agency files such as inspection reports) to verify these respondents' answers that no impoundments meeting the study criteria existed at these facilities. This check did not turn up any false negative responses.

Some facilities claimed their screening survey responses as Confidential Business Information (CBI), and EPA handled those facilities' screening survey responses data separately, in accordance with RCRA CBI procedures, but challenged all CBI claims. One screener survey response remains CBI.

EPA conducted a final edit of the screening survey data for all 1,787 completed screening surveys. This edit cleaned the data and ensured consistent formatting of responses and coded standardized responses for subsequent analyses. The cleaned data includes all screening survey data items, plus additional data needed for statistical analyses, and are available in electronic format (U.S. EPA, 1999b).

A.1.3 Long Survey (Second-Phase Sample)

For all facilities in the second phase sample, EPA prepared a long survey questionnaire requesting detailed information on the impoundments' design, operation, and closure practices as well as data on the wastewater and sludge composition and quantity. This three-part survey (U.S. EPA, 1999d) was developed by EPA to characterize the sample facilities with in-scope nonhazardous industrial surface impoundments and is the primary source of data for the Surface Impoundment Study (SIS), including the risk assessment, regulatory coverage, and other analyses presented in this report. EPA developed the sampling frame for this long survey from the screener survey data, as described in the following section.

US EPA ARCHIVE DOCUMENT

A.1.3.1 Long Survey Sampling Frame Development. While screener survey data collection was continuing through the summer of 1999, EPA needed to proceed with developing the sampling frame for the second phase sample of facilities that were to receive the long survey. The study's schedule required the long surveys to be mailed in the fall of 1999 so that the long survey data could be processed and analyzed for both the risk assessment and the regulatory coverage analysis. EPA chose to draw the second phase sample in two parts: a June 1999 sample, using the screener survey responses that had been received and processed by June 14, 1999, and a September 1999 supplementary sample to complete the sample with the facilities whose screener survey responses were processed after June 14, 1999, along with those that had claimed CBI status for all or part of their screener survey responses. The reason for this timing was so that EPA could collect publicly available data for most of the second phase sample facilities from state environmental agencies, along with the publicly available data being used to perform the false negative quality assurance check on the systematic random sample of screener survey responses.

After developing the complete set of non-CBI screeners, and reducing them to one record per facility, EPA determined which facilities were eligible for the second phase sample (long survey). The June sampling frame was developed from 1,597 completed screeners. Some facilities had more than one record in the combined hard-copy and CATI, non-CBI database. If there were screener surveys from both former and current owners, for the same facility, EPA kept the record for the current owner and deleted the record for the former owner. The resulting file contained 1,684 unique facilities with completed screeners.

The next step was to identify the facilities that were eligible for the second phase sample, according to their screener survey responses for the questions about the existence of an impoundment at the facility, meeting the criteria necessary for being in the study. Not all facilities answered the question about their facility's SIC code. In these cases, EPA obtained SIC codes from EPA databases or from descriptions of the facility's products or processes.

The file of facilities with a completed screener survey that were determined to be eligible for the second phase was the sampling frame for the second phase sample. The June 1999 sampling frame for the second phase sample consisted of 380 facilities; the non-CBI September 1999 sampling frame consisted of 43 facilities, and the CBI September 1999 sampling frame consisted of 9 facilities. EPA's objective was to obtain an overall sample of approximately 200 facilities, with approximately half of the facilities having at least one impoundment with decharacterized waste (to satisfy the requirements in the LDPFA), and approximately half of the facilities having never characteristic waste (to satisfy the requirements of the consent decree). In addition, EPA needed to balance the study resources so that direct and zero dischargers, and a few indirect dischargers, were included in the sample. With these general criteria, EPA selected sampling rates from the various strata that achieved the overall objectives, and resulted in the sample drawn as shown in Table A-1.

The final result was a sample of 216 facilities, plus the six pilot study facilities. However, one of the 222 facilities was included in both the June and September sample frames. Thus, the second phase sample consisted of 221 facilities, six of which were pilot study facilities.

US EPA ARCHIVE DOCUMENT

A.1.3.2 *Weight Adjustments for Ineligible Facilities and Nonresponses*. Theoretically, all facilities selected into the sample to receive the long survey should have been eligible for this phase of the study. That is, they should all have had at least one surface impoundment that satisfied the eligibility conditions in the screener survey. However, after they received the long survey, 21 facilities reported no eligible impoundments. By using extensive followup contacts, EPA determined the eligibility status of all facilities selected into the sample for the long survey. Hence, nonresponse adjustments were confined to adjustment for nonresponse among the sample facilities that were determined to be eligible for the survey.

For the full sample, there were only four eligible facilities that did not respond to the long survey, and one of those was an indirect discharge facility. Hence, for the weight adjustments for direct and zero discharge facilities, there were only three nonresponding facilities. Moreover, all three were direct discharge facilities whose screener data indicated that they did not handle any formerly characteristic waste.

The statistical analysis weights for the remaining 195 long survey respondents then were computed by adjusting the calibrated sampling weights for nonresponse among the eligible sample facilities. The weight adjustment process and results is described in detail in Section A.5. Because data were collected for all eligible impoundments at each responding facility (i.e., there was no subsampling of impoundments), these facility-level analysis weights also are appropriate for analysis of the impoundment-level data collected for the responding facilities.

A.1.3.3 *Long Survey Implementation*. The long survey questionnaire (U.S. EPA, 1999d) is a three-part form designed to collect the detailed information necessary for the risk assessment and regulatory gaps analysis as well as general characteristics of the study population. This information includes each facility's environmental setting (including receptor locations) and details on the design, operation, and history of each eligible surface impoundment, including the chemical composition of wastewater and sludge managed within these impoundments. The three parts include: Part A, basic facility identification information; Part B, an overview of the wastewater treatment system and environmental setting at the facility; and Part C, details about the design and operation of each in-scope impoundment. Part C also requested, for a list of 256 chemicals, chemical concentration data for wastewater, sludge, air, and leachate. Attachment A.1 includes electronic copies of the Part A, Part B, and Part C long survey forms.

The detailed information in the long survey required considerable effort to enter into an electronic format, standardize to consistent units and format, clean to correct skip pattern errors and other inconsistent responses, and process for data exploration and risk analyses. This was accomplished by creating and populating a series of relational databases, described in the subsequent sections, that hold the raw and processed survey data. Statistical methods were then applied (as described in Section A.5) to weight and analyze variables derived from the screener and long surveys (including risk assessment results) to characterize the population of nonhazardous industrial surface impoundments that meet the study criteria.

US EPA ARCHIVE DOCUMENT

Table A-1. Second Phase (Long Survey) Strata and Sample Sizes

Stage 2 Stratum	Type of Facility	Decharacterized Waste	SIC Priority	Frame Count	Sample Size
Non-CBI Stage 2 Strata and June Sample Sizes					
1	Direct Dischargers (DISCHARG=1)	Yes (Q16=1)	High (SIC_STR=1)	69	69
2			Low (SIC_STR=2)	7	4
3		Other (Q16=2 or missing)	High (SIC_STR=1)	183	61
4			Low (SIC_STR=2)	72	12
5	Zero Dischargers (DISCHARG=2)	Yes (Q16=1)	High (SIC_STR=1)	2	2
6			Low (SIC_STR=2)	4	4
7		Other (Q16=2 or missing)	High (SIC_STR=1)	13	13
8			Low (SIC_STR=2)	20	20
9	Preselected Indirect Dischargers (DISCHARG=3 and PREINDIR=1)	Yes (Q16=1)	High (SIC_STR=1)	2	2
10			Low (SIC_STR=2)	0	0
11		Other (Q16=2 or missing)	High (SIC_STR=1)	4	4
12			Low (SIC_STR=2)	4	4
13	Other Indirect Dischargers (DISCHARG=3 and PREINDIR=2)	Yes (Q16=1)	High (SIC_STR=1)	0	0
14			Low (SIC_STR=2)	0	0
15		Other (Q16=2 or missing)	High (SIC_STR=1)	0	0
16			Low (SIC_STR=2)	0	0
Total				380	195

(continued)

US EPA ARCHIVE DOCUMENT

Table A-1. (continued)

Stage 2 Stratum	Type of Facility	Decharacterized Waste	SIC Priority	Frame Count	Sample Size
Non-CBI Stage 2 Strata and September Sample Sizes					
1	Direct Dischargers (DISCHARG=1)	Yes (Q16=1)	High (SIC_STR=1)	4	4
2			Low (SIC_STR=2)	0	0
3		Other (Q16=2 or missing)	High (SIC_STR=1)	17	6
4			Low (SIC_STR=2)	4	1
5	Zero Dischargers (DISCHARG=2)	Yes (Q16=1)	High (SIC_STR=1)	0	0
6			Low (SIC_STR=2)	0	0
7		Other (Q16=2 or missing)	High (SIC_STR=1)	1	1
8			Low (SIC_STR=2)	0	0
9	Preselected Indirect Dischargers (DISCHARG=3 and PREINDIR=1)	Yes (Q16=1)	High (SIC_STR=1)	0	0
10			Low (SIC_STR=2)	0	0
11		Other (Q16=2 or missing)	High (SIC_STR=1)	2	2
12			Low (SIC_STR=2)	1	1
13	Other Indirect Dischargers (DISCHARG=3 and PREINDIR=2)	Yes (Q16=1)	High (SIC_STR=1)	2	0
14			Low (SIC_STR=2)	1	0
15		Other (Q16=2 or missing)	High (SIC_STR=1)	3	1
16			Low (SIC_STR=2)	8	0
Total				43	16

(continued)

US EPA ARCHIVE DOCUMENT

Table A-1. (continued)

Stage 2 Stratum	Type of Facility	Decharacterized Waste	SIC Priority	Frame Count	Sample Size
CBI Stage 2 Strata and Sample Sizes					
1	Direct Dischargers (DISCHARG=1)	Yes (Q16=1)	High (SIC_STR=1)	3	3
2			Low (SIC_STR=2)	0	0
3		Other (Q16=2 or missing)	High (SIC_STR=1)	4	1
4			Low (SIC_STR=2)	2	1
Total				9	5

A.2 Long Survey Data Entry

The goals of the data entry effort for the long survey were 1) to archive as complete a dataset as possible, in order to increase statistical confidence and 2) to maintain the integrity of the dataset through entry, processing, and analysis. This required rigorous quality assurance/quality control (QA/QC) procedures at every step of the process. The general QA/QC plan was to check all manually entered data 100 percent and to manually confirm that each data processing or analysis program was functioning correctly. Details on the data entry methodology and associated QC measures follow. This required entry of almost 200 Part A and Part B forms and, because many facilities had multiple eligible impoundments, over 500 Part C forms.

A.2.1 Data Entry Objectives

The overall objective of long survey data entry was to record and preserve, in an electronic format, exactly what the survey respondents reported on their returned forms. Although obvious typographical errors were corrected, entry staff were instructed not to judge how reasonable or consistent responses were, but to record them exactly as written. Database fields for margin notes from the long survey were included in for practically every question; this also enabled for typographic or other corrections to be recorded in the data entry database.

A.2.2 Data Entry Database

The data entry database for the long survey mirrors the design of the survey forms shown in Attachment A1. Data tables were indexed at the facility level (questions in Parts A and B), facility and impoundment level (Part C), and at a third level, by chemical for chemical data and by layer for liner and subsurface layer data. To help ensure consistent entry, coding tables were used for units and other repeated data elements. Duplicate tables were included in the entry database to allow for double extraction and double entry. Once double extraction/double entry

US EPA ARCHIVE DOCUMENT

comparisons were complete, these tables were removed from the database, resulting in the design described in this section. Although created and maintained in Microsoft Access, data design conventions include compatibility with *.dbf format, and programs are available to automatically export the database tables as .dbf or ASCII text files.

Data entry forms were developed that replicate the survey's appearance as closely as possible. This provided almost immediate familiarity with the entry screen for the data entry staff. Buttons were used to open text fields to record margin notes and comments. Drop-down boxes included standardized selections for units and other repeated data responses. EPA designed the survey to allow respondents to choose units for numeric values, resulting in a number of units being used for each numeric variable. As new units were encountered during data entry, the standard list of units was expanded to help ensure consistent and correct entry of each response.

Attachment A2 includes data entry database design documentation which describes the data table structure, linkages, codes, and the content of the various data fields. The database design is fully documented three parts, described briefly below.

A.2.2.1 Entity Relationship Diagram. Attachment A2-1 contains entity relationship diagrams that picture how the various tables that make up the data entry database are linked together using key fields. Links in this diagram are shown as one-to-many (where a table is related to several tables of the same structure) or one-to-one (where a table is linked to a single table). Tables are linked using one, two, or three key fields, depending on the number of tables linked and the position of the tables in the database. For example, because there can be multiple surface impoundments at a facility, there can be many surface impoundment data tables for each facility, with these multiple tables linked to the Surf_Imps table by the key fields FAC_ID and IMP_ID.

The first figure in Attachment A2-1 shows the overall database structure along with table structures for Part A and Part B of the long survey questionnaire. The remaining figures show the table structures and relationships for the Form C tables connected to the SURF_IMPS table. Survey questions corresponding to the data tables are listed at the top of each diagram.

A.2.2.2 Data Dictionary. Attachment A2-2 contains the data dictionary for the database tables shown in the entity relationship diagram (Attachment A2-1). This dictionary provides data type, size, and description (including long survey question number) for each field (column) in each database table, which are listed in the order of the survey questions and as they appear in the entity relationship diagrams. Data dictionaries for the coding tables are provided in alphabetical order at the end of this attachment.

A.2.2.3 Coding Tables. Attachment A2-3 contains the coding tables from the data entry database. In the SI survey database, coding tables serve the same function as a data entry code book: to ensure consistent responses for questions with answers that can be standardized, such as units or chemical names, or for questions with multiple choice responses (e.g., yes, no, don't know, or other). These tables were adapted from coding tables developed during survey design. Standardization (i.e., use of a table for multiple questions) was used wherever practical to minimize the number of tables and increase consistency within the database. During data entry,

US EPA ARCHIVE DOCUMENT

codes and their definitions are presented as drop-down boxes in the data entry forms to ensure correct and consistent data entry. The coding tables appearing in this document supercede those in the previous version in that they include additional rows for new values encountered during data entry. For example, the codes for concentration units expanded from 20 to over 40 possible entries during the course of data entry.

A.2.3 Data Entry Protocols

Data entry protocols were developed for and followed by data entry staff, and serve as a record of how data were entered. As new situations were encountered during data entry, the protocols were modified. The final protocol is included in Attachment A3.

Data entry protocols were developed to ensure consistent treatment of potentially inconsistent or incomplete data, and thereby minimize the double-entry comparison task and ensure a higher quality dataset. Perhaps the most important protocol was to record exactly, word-for-word, what was recorded in the survey, including the margin notes entered by the survey respondents. Another was to record a comment for every change made to correct obvious errors, resulting in a note wherever the database differs from the original survey.

Chemical data conventions were needed to ensure consistent treatment of nonstandard responses. Examples include: enter "cyanide" and "reactive cyanide" as total cyanide and "amenable" cyanide as free cyanide; sum individual alachlor values and enter total under "PCBs," including individual values in margin note; enter "chromium" values as total chromium. In each of these cases, notes were included in the database describing what was done. These and other data entry conventions are detailed in the data entry protocol in Attachment A3.

A.2.4 Digitizing Map Data

A geographic information system (GIS) was used to digitize residence and well locations from the marked topographic maps returned as question B3 of the Part B of the survey. Question B3 asked the survey respondents to mark wells, residences, and schools within a 2-kilometer radius of their surface impoundments on a U.S. Geological Survey (USGS) topographic map that was included with the survey form (see Attachment A1).

Survey response data for question B3 maps were used to develop a series of GIS map layers. The goals of these procedures were (1) to develop a series of GIS map and data layers that could be used to analyze spatial relationships among surface impoundment ponds, receptors, schools, and wells; and (2) to process and extract data to serve as inputs to risk assessment models. The coordinate locations of impoundment boundaries, individual residences, residential areas, schools and wells were entered into a GIS through "heads-up digitizing," a process whereby a GIS technician uses a mouse to enter the locations of features by pointing to them on a digitized image displayed on screen. A series of programs were written in Arc Macro Language (AML) to automate the data preparation and digitizing processes.

A.2.4.1 Map Preparation and Registration. Map preparation and registration consisted of three main steps:

US EPA ARCHIVE DOCUMENT

1. Obtain the necessary documents, including the map, image files of the map, and any additional annotation.
2. Assess the overall quality of the scanned image.
3. Create a registered image from the scanned image.

Obtain documents and images. Spatial data were acquired by physically searching the file of documents returned by each survey respondent in response to question B3. For most sites, these documents consisted of one or more hardcopy maps, which were usually annotated by the survey respondent to show the features to be digitized. In some most cases, these maps were the USGS topographic maps originally supplied to the respondent. In many cases, however, the respondent provided an alternate map or maps. These included other USGS topographic maps, photocopies of USGS maps, and a variety of non-USGS maps including site plan drawings and as-built diagrams.

Question 3B maps were labeled with preprinted labels containing a text ID and barcode. These maps were then scanned and converted to TIFF multiband ("composite") images.

Assess image quality. GIS technicians assessed the usability of each scanned image by displaying the map on the screen and viewing it to confirm that:

- all features shown on the map could be seen clearly on the image;
- registration marks and site ID label were clearly visible;
- there was no apparent distortion of the image;
- the image covered all of the area within 2km of the impoundments; and
- features and annotation added by the respondent were clearly visible.

The AML program epa_scanmap.aml prompted the user with a checklist and ensured consistency during this procedure. If the map was not usable and/or areas within the 2km buffer were missing, USGS Digital Raster Graphic (DRG) images of 1:24,000 Quads were downloaded via the internet and stored in the respective site directory as TIFF files.

Register image. The original maps provided to survey respondents were standard USGS 7.5-foot topographic quadrangles (1:24,000 scale). These maps contain registration marks for NAD 83 geographic coordinates near the four corners of the map area. Some of the maps returned by survey respondents were not standard USGS 7.5-foot topographic quadrangles. Although some of these maps contained registration marks labeled with geographic coordinates, others contained grid lines or registration marks based on arbitrary or unidentified coordinate systems. In some cases, no coordinate system or grid was shown on the map.

Prior to digitizing, each image was registered to a real world coordinate system so that subsequent measurements of distance and area could be expresses in real world units (as opposed to scanner inches). In most cases, the appropriate State Plane coordinate system was used. In this case, "appropriate" means the State Plane coordinate system zone specified on the map. Although the standard units of the State Plane coordinate system are generally feet, meters were used throughout this project.

US EPA ARCHIVE DOCUMENT

For maps with registration marks for NAD 83 geographic coordinates (primarily standard USGS topographic quadrangles), the program epa_box.aml was used to create a file containing the geographic coordinates of the four registration marks. The program then used this file to generate a map layer whose corners were coincident with the tic marks at the corners of the 7.5-foot topographic quadrangle and project this to the user-specified State Plane coordinate system zone.

A series of other programs, links.aml, register_image.aml, register_grayscale.aml and register_pseudocolor.aml, utilized Arc/Info's GRIDWARP command to identify registration marks and transform images to the appropriate State Plane coordinate system.

A.2.4.2 Digitizing Procedures. Features from all maps were digitized using the menu-driven digitize.aml program. Scanned images were displayed in the background and features were captured from these images using the cursor as the input device. Each feature type was stored as a separate map layer, or coverage, and a set of digitizing guidelines was developed (see Attachment A4-1). The coverage names, their contents and associated map symbols are shown in Table A-2.

All coverages contained fields for feature-specific margin notes, i.e. information that was noted on the map by the respondent, and digitizer's comments. Feature-specific margin notes and comments were added to individual features as they were digitized. Attachment A4-2 contains a list of standard digitizer's comments. Margin notes and comments that were not specific to one or more features were inserted into a text file specific to that site and image, i.e., 1234a.txt.

Because all coverages and all of their contained data items were created with the digitize.aml program, all coverages containing the same feature types have identically defined attribute tables. This ensures that coverages can be appended at some point in the future after they are projected to a common coordinate system.

A.2.4.3 QA/QC of Digitized Coverages. QA procedures were incorporated into the digitizing process and QC checks were carried out throughout the data development process through the use of computer programs that ensured standardization of data development.

The *digitize.aml* program was initiated at the command line and required a single parameter, the image ID of the image to be used for the current digitizing session. The menu interface to this program, displayed in Figure A-1, contained a large number of buttons which allowed the user to select the coverage to be edited (the "edit coverage"), add or delete features, assign impoundment ids, margin notes, or digitizer comments to feature databases ("feature attribute tables"), and perform all of the other normally-required processes. Also included was a button to allow the user to temporarily suspend menu input so that commands could be entered directly on the ArcEdit command line.

Most attributes were assigned to the feature attribute tables automatically, including the facility ID and map letter, type of source map, feature "origin" (preprinted or handdrawn), and attributes that controlled the symbolization of features in the graphic display.

US EPA ARCHIVE DOCUMENT

Table A-2. GIS Coverage Name, Type, and Content

Coverage	Type	Contents	Map Symbol
BOX_siteid	Line	Topographic map limits	
BUFF_2KM	Line	A system-generated 2-km buffer around impoundments	Thick red line
PONDS_PNT	Point	Impoundments represented by points	Blue dot
PONDS_POLY	Polygon	Impoundment boundaries of ponds with areas	Blue line
PROPERTY	Line	Site property boundary	Dashed red line
RECP_PNT	Point	Receptor locations – Individual buildings known or believed to be residences	Green dot
RECP_POLY	Polygon	Receptor locations – Urban or residential areas	Green line
RESP_2KM	Line	The 2-km radius as drawn by the survey respondent	Thick red line
SCHL_PNT	Point	Schools represented by point symbols and individual school buildings	Red dot
WELLS	Point	Wells (generally groundwater supply wells)	Hollow blue triangle with cross

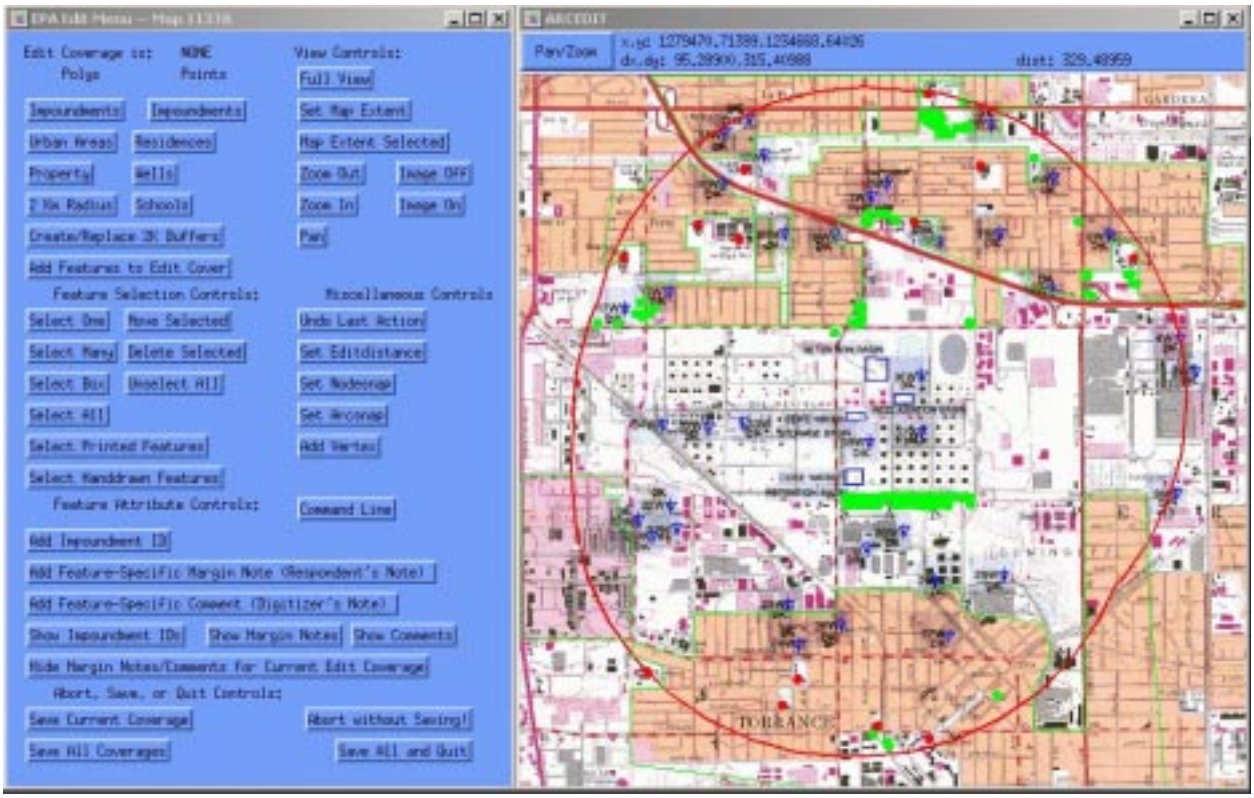

Figure A-1. Digitizing menu used in *digitize.aml* program.

US EPA ARCHIVE DOCUMENT

This interface was flexible enough to be used by both experienced GIS personnel and others without significant prior GIS experience. The program behind the interface was also designed to prevent or at least limit the inadvertent assignment of incorrect attributes. In other words, QA was integrated into the program wherever practical.

When all features had been digitized using the digitize.aml program, a program, qc_site.aml, was run to perform a series of automated QC checks on all coverages created for that site. These checks ensured that

- All the required coverages had been created
- All items in each feature attribute tables were present and correctly defined
- All standardized items (e.g. site, map, symbol, etc.) had correct values
- All lines in coverages containing polygon features formed closed polygons
- Lines did not intersect except at nodes (a requirement of lines that would be used to build polygon features).

The results of the program were written to a text file containing two types of messages. A warning message was issued if a coverage lacked features (e.g., the coverage containing residences lacked any points). Error messages were issued if any of the situations described in the above list dictated (e.g., the program found an incorrectly-defined item or an unclosed polygon). Corrections were made to the respective coverages, when necessary.

A second QC process involved the generation of large (24" x 36") checkplots of each image for examination by a quality control reviewer. The reviewer was a GIS analyst who had not been involved in that site's digitizing process. The checkplot displayed the map image and each of the digitized features drawn with its corresponding symbol (see Table A-2). The review process consisted of comparing the original map with both the checkplot and the digital coverages and carrying out the following steps:

- Determine whether all features had been digitized
- Determine whether all margin notes had been entered with feature data
- Determine whether appropriate digitizer's comments had been entered
- Determine whether non-feature-specific margin notes had been inserted into a text file.

The reviewer also examined receptor features to determine whether questionable residences should remain in the coverage. Ancillary data, such as Digital Ortho Quarter Quads, viewable via a web browser, were used in this determination. In the event that additional digitizing or revisions were needed, the map original was returned to the digitizer. Corrections were made and the review process was repeated.

A.2.4.4 Additional Data Modifications. Prior to final analyses of in-scope surface impoundments (described below), modifications were made to some features to improve the accuracy of analyses. Three types of modifications were made:

US EPA ARCHIVE DOCUMENT

- Many wells of WELLTYPE 7 or 14 (unknown or unspecified) were reclassified, based on ancillary data, such as documentation included with the survey.

- An examination of aerial photographs for specific ponds during the analysis of sites showing air risks revealed residences that had not been digitized. These were located in subdivisions that were developed after the USGS topographic quad was produced. In these cases, the additional residences were digitized.

- The assumption was made that all private drinking water wells should have residences associated with them. Residences were added to many sites to correspond with these wells.

A.2.5 Diagram Data, Elevation Data

Survey respondents were asked to supply diagrams containing information for three sets of questions in the survey. These diagrams contained information on facility wastewater treatment information (survey question B1), plan and elevation diagrams (question C10), and liner diagrams (question C11). Respondents often combined some or all of this information into a single diagram and or sent diagrams that combined different impoundments on a single diagram. To avoid making multiple scanned image files of these large format diagrams for each question, it was decided that each diagram should be scanned only once and then linked to the appropriate questions.

A.2.5.1 Database for Diagram Tracking and Linkages. A database system was developed to link each diagram to one or more facilities, impoundments, and uses. Tables A-3 and A-4 provide dictionaries for the two data tables in this database. Adhesive stickers were printed that contained a unique number printed in both Code 39 barcode and text. One sticker was placed on each diagram and its number was used as a "diagram number" to track and link the diagrams. The diagram number contained a checkdigit, which was used to detect and prevent data entry errors. Database tables were created in an Microsoft Access database to store linkage information. Simple data entry forms were created to permit linking the diagrams to their use(s) with simultaneous entry of plan and elevation data extracted from the diagrams.

Diagrams could be linked either to a facility (survey question B1, wastewater treatment diagrams) or to an impoundment (survey question C10, plan and elevation diagrams; survey question C11, liner diagrams). Wastewater treatment diagrams were linked to a facility by entering a record containing the diagram number and the facility ID in the table DIAG_WWT. Other diagrams were linked to an impoundment by entering a record containing the diagram number, the facility ID, and the impoundment ID in the table DIAG_IMP. In this way, a single diagram could be linked to a facility and one or more impoundments.

After linkage, the diagrams were scanned into TIFF format, which was converted to the more highly compressed (i.e., smaller files) GIF format for archiving. The resulting diagram files were titled with their diagram number (and .gif) as their file name. A simple report program was written in Microsoft Access that produced listings of documents by facility, impoundment, and use. This report was printed to an Adobe Acrobat (pdf) file for reference and use in retrieving the

US EPA ARCHIVE DOCUMENT

Table A-3. Structure of DIAG_WWT Database Table for Wastewater Treatment Diagrams (Question B1)

Field Name	Type	Size	Description
FAC_ID	Text	5	Facility ID - Linked to Table FAC_INFO
DIAG_ID	Text	15	Unique ID for diagram (from diagram sticker)

Table A-4. Structure of DIAG_IMP Database Table for Impoundment Diagrams (Question C10, plan and elevation views; Question C11, liner cross sections)

Field Name	Type	Size	Description
FAC_ID	Text	5	Facility ID - Linked with IMP_ID to table SURF_IMP
IMP_ID	Text	50	Impoundment ID - unique ID for impoundment at Facility
DIAG_ID	Text	15	Diagram ID - unique ID for diagram
C10_PLN	Boolean	1	True if diagram is a plan view of impoundment
C10_XST	Boolean	1	True if diagram is a elevation (cross section) view
C11_LNR	Boolean	1	True if diagram is a liner cross section

desired files for review. A copy of this report is included in Attachment A-5. The GIF format survey diagrams are archived and available on CD-ROM.

A.2.5.2 Processing of Elevation Data from Diagrams. In survey question C-10, Respondents were asked to supply plan and elevation diagrams for each surface impoundment. These diagrams were used to obtain the following elevation data:

- Ground elevation,
- Water table elevation,
- Base (bottom surface of the impoundment) elevation,
- Elevation of liquid level in the impoundment.

Maximum, minimum, and typical values (if supplied) were recorded for all elevation data, except for ground elevation, where an average value was recorded (if supplied). From this data, the following information was calculated:

- Distance of base from the water table,
- Distance of liquid level from the water table, and
- Height of liquid in the impoundment (i.e., distance of liquid level from base).

US EPA ARCHIVE DOCUMENT

Table A-5. Structure of IMP_ELEV Database Table for Impoundment Elevation Data (Question C-10)

Field Name	Type	Size	Description
FAC_ID	Text	5	Unique ID for each facility
IMP_ID	Text	50	Unique ID for impoundment at that facility
GR_EL	Double	8	Ground (reference) elevation - 0 when referenced to ground
GRELUTS	Long Integer	4	Units code for ground elevation
WT_MIN	Double	8	Minimum water table distance from ground
WT_MAX	Double	8	Maximum water table distance from ground
WT_TYP	Double	8	Typical water table distance from ground
WT_UTS	Long Integer	4	Units code for water table distances
B_MIN	Double	8	Minimum distance from ground to base of impoundment
B_MAX	Double	8	Maximum distance from ground to base of impoundment
B_TYP	Double	8	Typical distance from ground to base of impoundment
B_UTS	Long Integer	4	Units code for base distances
LH_MIN	Double	8	Minimum distance from ground to top of liquid surface
LH_MAX	Double	8	Maximum distance from ground to top of liquid surface
LH_TYP	Double	8	Typical distance from ground to top of liquid surface
LH_UTS	Long Integer	4	Units code for liquid distances
Comment	Text	250	Comment

A database table (IMP_ELEV) was created to store impoundment plan and elevation data extracted from diagrams. The structure of the database table (IMP_ELEV) is shown in Table A-5. The data entry form for the plan and elevation data was combined with the impoundment linkage form (mentioned above).

Because of the wide variety of diagrams supplied by participants, extraction of elevation data required some interpretation. In some instances, the needed elevation data was clearly noted on the diagram. In other cases, the needed data could be measured from scale drawings. For quality control, a second person compared all diagrams to their extracted values (i.e., 100 percent of all data was checked).

Upon completion of data entry and comparison, a senior review was conducted that focused on extreme data points including:

US EPA ARCHIVE DOCUMENT

- Facilities with the greatest differences between high and low water table values
- Impoundments with the greatest depth to the water table
- Impoundments with the water table at or above the base of the impoundment
- Impoundments with water table aboveground elevation
- Impoundments with base aboveground elevation
- Impoundments with the water table above the impoundment liquid level
- Facilities with the greatest distance between impoundment liquid level and water table
- Inconsistencies between elevation data and depth to saturated zone in survey question B-10.

This review considered approximately 15 percent of the facilities. The facility diagrams, surveys, and published data, such as USGS maps, were used in the review. Changes were made for 10 facilities based on review of elevation data. Corrections were made for three additional facilities based on the comparison of elevation data with question B-10. Additional corrections were made for three impoundments with unusually large distances between the water table and their impoundment liquid levels.

A.2.6 Quality Assurance/Quality Control

Extensive and rigorous QA/QC procedures were developed and followed throughout the data entry process. QA/QC procedures for map and diagram data have been described in the sections above. To achieve a 100-percent check for data entry, all survey data that were manually entered into the survey entry database from the hard-copy surveys were double-extracted and entered independently by two different staff members. To accommodate double-extraction/ double-entry, the data entry database contained duplicate tables for every data element as well as duplicate entry forms. Once both entries were complete, the two files were electronically compared, and, using the hard copy survey, a third staff member reconciled any differences. Other manually entered data were checked 100 percent.

For automated data processing, the data extraction/processing system was thoroughly validated before use. This involved manually checking enough of the data (usually 5 percent to 10 percent) to ensure that the system functioned properly. When conducting such checks, the QC procedures required that each unique calculation or data combination be checked at least once. In addition, a version control system was employed to ensure data integrity and that each analysis conducted with the most recent dataset. Detailed records were kept of every QC check, and these were reviewed during a final QA audit of the data entry process.

A.3 Collection of Supplementary Data

Secondary data sources included U.S. Census GIS data (used to supplement survey information on the number and location of people living around the site), GIS coverages of soils and aquifer data, USGS topographic maps, and river flow data from EPA's Basins database. These data were collected and used to provide more consistency and completeness for key data elements, or to provide data not directly available from the survey (e.g., population data).

US EPA ARCHIVE DOCUMENT

A.3.1 Development of Supplementary Spatial Data

A geographic information system was used to digitize residence and well locations from the marked topographic maps requested in question B3 of the Part B of the long survey. Question B3 asked the survey respondents to mark wells, residences, and schools within a 2-kilometer radius of their surface impoundments on a USGS topographic map that was included with the survey form (see Attachment B1). Because these maps were returned unmarked (or not returned) by a significant number of respondents and because the survey did not ask for population data, the GIS was used to supplement these data with U.S. Census data. In addition, the GIS was used to collect spatial data on the presence of waterbodies, wetlands, and managed areas with 2 km for the ecological risk assessment.

A.3.1.2 Data Processing and Spatial Analysis. Out of the total 157 facility sites with impoundments in-scope for the long survey (i.e., those with chemicals or pH of concern), a total of 153 returned maps with the survey, including 150 sites in the continental U.S., 2 sites in Alaska, and 1 site in Puerto Rico (four sites were determined to have missing geographic data). The geographic analysis was carried out for these sites to develop the sample data necessary to develop the following statistics about the distribution of wells, residences, population, and schools for impoundments with chemicals or pH of concern:

- Estimated number of groundwater supply wells, broken out by distance (0-150, 151-500, 501-1000, and 1001-2000 meters) from the impoundments, and cross tabbed by use (public, private drinking water, irrigation, livestock watering, don't know, other).
- Estimated number of residences, broken out by distance (0-150, 151-500, 501-1000, and 1001-2000 meters) from the impoundments.
- Estimated number of schools, broken out by distance (0-150, 151-500, 501-1000, and 1001-2000 meters) from the impoundments.
- Estimated number of people, broken out by distance (0-150, 151-500, 501-1000, and 1001-2000 meters) from the impoundments.

A simple Arc/Info distance function was used to process the school data, but the remaining questions required pre-processing of the digitized survey data and the 1990 U.S. Census data.

Overlay Processing of In-Scope Impoundments. To develop the best estimate of wells, residences (households), and population surrounding the impoundments with constituents used census coverages and data were used to: (1) provide an indicator of average household size; (2) estimate the number of private drinking water wells, and (3) provide population data for population estimates. Census coverages and corresponding data were obtained via ftp download from the EPA server in Research Triangle Park. Additional processing was carried out to link block and block group variables with block coverages. Census data were not available for Puerto Rico, so the wells and residence analyses utilized only feature data on the map supplied by the survey respondent and no population data could be estimated.

US EPA ARCHIVE DOCUMENT

The most critical data processing steps for census/feature data analyses for each of the in-scope surface impoundment (excluding Puerto Rico) were as follows:

- Step 1. Create a set of buffers at distances of 150, 500, 1000 and 2000 meters, respectively, from the impoundment boundary.

- Step 2. Overlay buffers on census block group coverages to create new coverage of census blocks split by distance buffers, retaining the value of the original area of the census block for later analysis. Steps 1 and 2 were carried out using procbloc.aml. The resulting coverages were named BLR<Site ID><Impoundment Index>, e.g. BLR12341.

- Step 3. Overlay BLR coverage with RECP_PTS coverage and summarize number of receptor points per polygon, using rcp_over.aml.

- Step 4. Overlay BLR coverage with WELLS coverage and summarize number of wells per polygon, by welltype, using well_over.aml.

- Step 5. Populate new overlay coverages with census, receptor and well data, using linkwells.aml and blrprep.aml.

Figure A-2 shows an example of a BLR coverage, with surface impoundments, receptors and wells.

A.3.1.2 Dasymetric Mapping and Analysis Procedures for Human Receptor Data. As previously noted, the computation of distances for schools was straightforward because no census data were required. A GIS distance function (Arc/Info's NEAR command) was utilized to compute the distance of each school from each surface impoundment at the respective site. Distance were then categorized as belonging to Ring 1 (0 – 150m), Ring 2 (150.1 – 500m), Ring 3 (500.1 – 1000m) or Ring 4 (1000.1 – 2000m). Data were compiled in a file with the data structure shown in Table A-6.

A similar procedure was used to compute distances to marked wells, except that wells were broken out by well type. An additional analysis of wells was conducted, utilizing census data to provide supplemental data on the number of drinking water wells in the vicinity of a surface impoundment. The initial well distance file that was generated contained information about the distance of each marked well to each surface impoundment. The structure of this file is shown in Table A-7.

US EPA ARCHIVE DOCUMENT

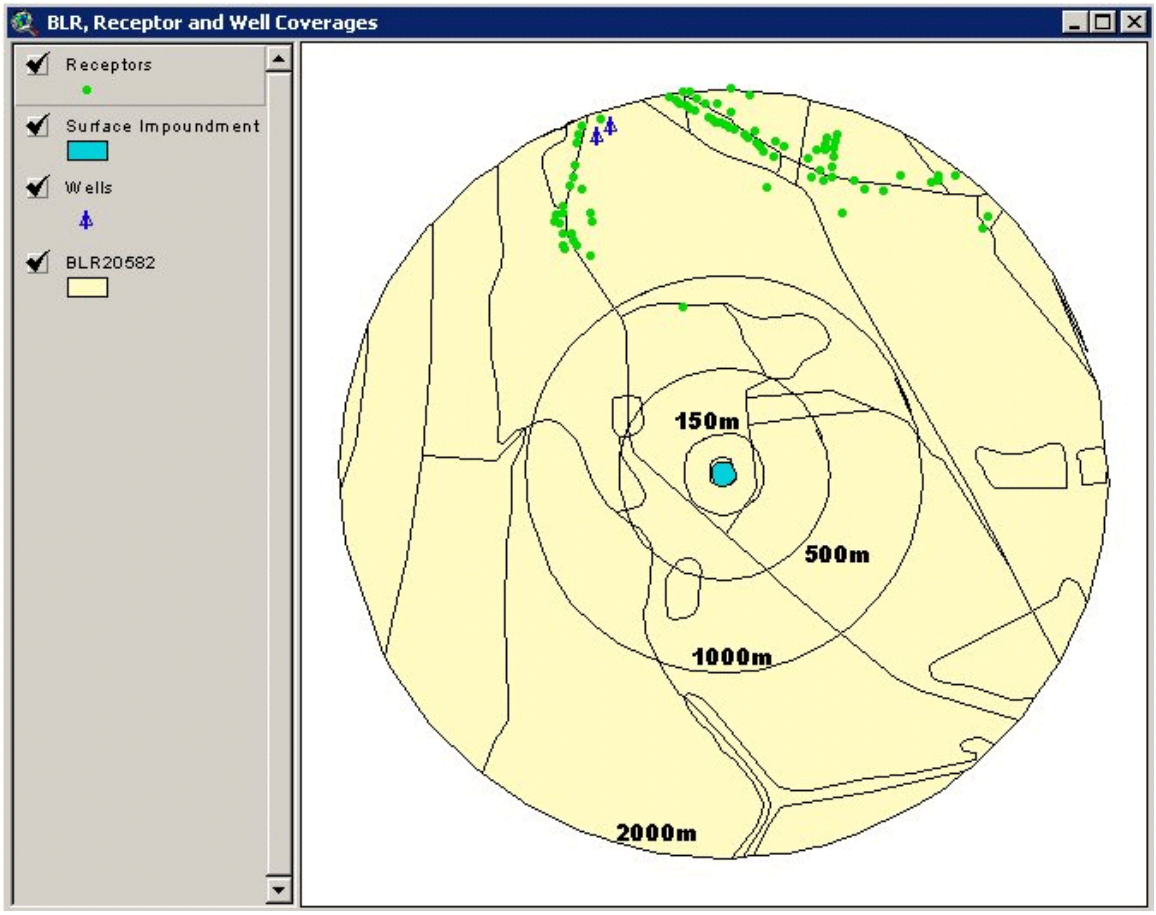

Figure A-2. Overlay of census blocks and distance rings with wells and receptors.

Table A-6. Table Structure for School Data

Variable Name	Description	Data Type
FAC_ID	Unique facility ID	Text
IMP_ID	Unique impoundment ID	Integer
RING_ID	Ring 1, 2, 3 or 4	Integer
RINGDIST	Ring distance of 150, 500, 1000 or 2000	Integer
AREAUNIT	Meters	Text
NMSCHOOL	Number of schools within specified ring	Integer

US EPA ARCHIVE DOCUMENT

Table A-7. Table Structure for Well Distance Data

Variable Name	Description	Data Type
FAC_ID	Unique facility ID	Text
IMP_ID	Unique impoundment ID	Integer
WELL_ID	Unique well ID: FAC_ID plus coverage ID	Integer
DIST	Distance from surface impoundment boundary	Float
XCOORD	X-coordinate in jState Plane meters	Float
YCOORD	Y-coordinate in State Plane meters	Float
WELLTYPE	Type of well	Integer

Census data were used to develop estimates of population and number of residences for each geographic unit that fell within the 2-km range of each eligible surface impoundment. The geographic unit of analysis was the result of a geographic overlay of census blocks and distance rings (at distances of 150 m, 500 m, 1,000 m, and 2,000 m from the surface impoundment, respectively). In some cases, the unit of analysis was an entire census block; in other cases, where a distance ring bisected it, the unit of analysis was a partial census block.

Dasymetric Mapping. Dasymetric mapping techniques were used to obtain a more accurate estimate of population and residence numbers than are possible by more traditional methods. Although the census block is the smallest geographic unit used by the U.S. Census, it is sufficiently large enough that variations in the numbers of people and residences within the block are obscured. This does not pose a problem when the entire block is the unit of analysis. With partial blocks, however, population and residence numbers for the entire census block must be reassigned to the partial block, keeping the block totals constant. Normally, this is done by prorating the block variables (such as population) by the area of the new, or partial block unit, that is, if the partial block was 75 percent of the size of the original, then 75 percent of the population would be assigned to that unit. The problem with this method is that, especially in rural areas, residences may be widely scattered and there may be large areas that are assigned a population when, in fact, they have none. The reverse is also possible, population undercounts in densely populated areas.

Dasymetric mapping uses supporting information about the distribution of a phenomenon to provide a more accurate representation of a (map) surface than that provided by standard data collection units, such as census blocks or block groups. In this case, the supporting information comes from the residences that were digitized from maps returned by the survey respondents. This information can be used to provide a better characterization of high and low density areas within the census block and to develop more accurate counts for enumeration units split by the 150, 500, 1,000 and 2,000 m rings. In other words, the presence of digitized points, and decisions made about their accuracy and currency, were used to weight the population and residence number estimations.

US EPA ARCHIVE DOCUMENT

Assumptions. Three different methods of geoprocessing and computation were developed, using assumptions based on the date of the map and the accuracy of the maps provided by the survey respondents. Assumptions:

- If map predates 1990 **and** no residences were marked on the map by respondent, then the 1990 census is the most accurate source of population and residence data.
- If the map predates 1990 and residences **were** marked on the map by respondent, the digitized map data represents the most accurate source of population and residence data in non-urban areas.
- If the map date is later than 1990, the digitized map data represents the most accurate source of population and residence data in non-urban areas, whether or not residences were marked on the map by respondent.
- Since individual receptor points are not present in urban areas, 1990 census data provide the most accurate source of population and residence data in those areas.

Decision Rules. Using the assumptions stated above, a decision tree, based on (1) presence of urban areas on map, (2) date of source map, and (3) whether respondent had marked residences on the map was used to determine which one of three processing and analysis routines would be used to most accurately estimate the population and number of residences within the 2-km surface impoundment buffer area. This decision tree is reflected in the Figure A-3.

Before implementing the decision tree, the map for each surface impoundment (n=517), was checked for: (1) presence of urban polygons, (2) map date, and (3) marked residences on map. Based on this check, one of the following three routines was implemented.

Routine A

This routine was used when no urban areas were contained in the geographic data and the map data were assumed to be more accurate than the census data. It is the simplest of the three routines. The number of receptor points in each geographic unit were counted. This provided the value for estimated number of residences. This value was then multiplied by the average number of people per housing unit, at the block group level (hereafter referred to as the block group housing unit size), to obtain the estimated number of persons.

Routine B

This routine was used when the map from which receptor points were digitized predated the 1990 census and there were no residences marked on the map by the respondent. Census data were assumed to be the most accurate source of information and census population totals for each block were held constant. However, the distribution of digitized points was used to weight population and residence numbers for partial blocks.

For census blocks that were not split by a distance ring (hereafter referred to as whole blocks), the estimated number of persons was simply the census population value for that block.

US EPA ARCHIVE DOCUMENT

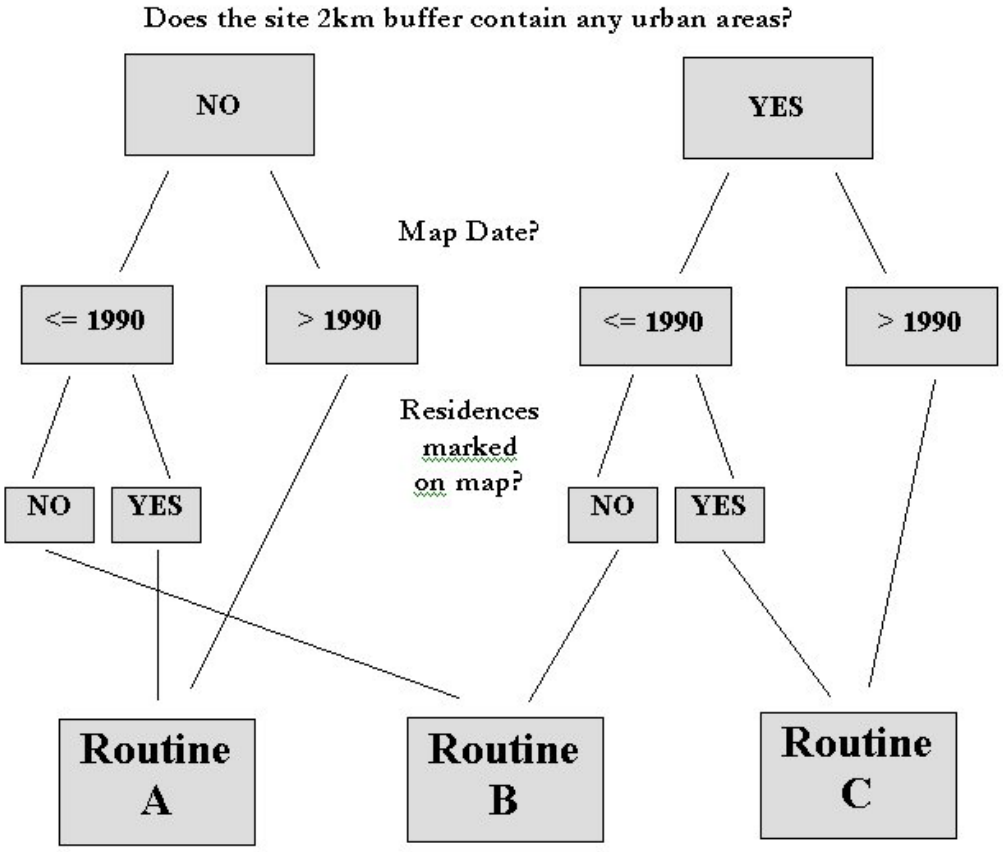

Figure A-3. Dasymetric procedure for integrating survey and U.S. Census data.

The estimated number of residences was computed by dividing the estimated number of persons by the block group housing unit size.

For census blocks that were split by a distance ring (hereafter referred to as partial blocks), the number of digitized receptor points was counted. If the partial block contained no receptor points, the block population was multiplied by the area proportion, and a percentage value was obtained by dividing the area of the partial census block by the area of the whole census block. The resulting value was the estimated number of persons for that partial block. The estimated number of residences was computed by dividing the estimated number of persons by the block group housing unit size.

For partial blocks that contained digitized receptor points, a revised block population value was obtained by multiplying the block population by the area proportion. Then, the total number of receptor points in the whole block was multiplied by the block group housing unit size. If that value exceeded the revised block population, the revised block population was divided by the number of receptor points in the block to come up with an estimated block housing unit size. This value was then multiplied by the number of receptor points in the partial

US EPA ARCHIVE DOCUMENT

block to determine the estimated number of persons. The estimated number of residences was equal to the number of digitized receptor points.

If the product of the number of receptor points in the whole block and block group housing unit size was less than the revised block population, the ratio of partial block receptor points to whole block receptor points was multiplied by block population to get the estimated number of persons. This value was divided by block group housing unit size to obtain the estimated number of residences.

Routine C

Routine C was used when the geographic data contained urban areas (with the exception of pre-1990 map dates with no residences marked on the map). The digitized receptor data were assumed to be accurate. However, census data were used in polygons with no digitized receptor points because areas delineated urban on topographic maps do not show individual residences.

For partial and whole blocks with digitized points, the number of receptor points were counted to estimate the number of residences. This value was then multiplied by the block group housing unit size to estimate population.

For whole blocks with no digitized receptor points, the estimated number of persons was the census population value for that block. The estimated number of residences was computed by dividing the estimated number of persons by the block group housing unit size. For partial blocks, the population value was the product of the area proportion and census block population. This value was divided by the census block group housing unit size to obtain estimated number of residences.

Routines A, B and C were incorporated into the program dasyprog.aml. After this program was run, the estimated number of residences was used to obtain an estimate of the number of drinking water wells, based on census data (census wells). Where the ratio of drinking water wells to housing units at the census block-group level was greater than 0.5, this ratio was multiplied by the estimated number of residences to obtain this value. Because those data were more complete (many respondents did not mark drinking water wells), the census wells were used in subsequent analyses in all cases except where marked private wells drinking-water wells were greater than the census well count.

The data obtained from the overlay analysis and dasymetric mapping procedures was compiled in a table with the structure shown in Table A-8.

A.3.1.3 Screening for Ecological Risk Modeling. GIS screening of sites with in-scope surface impoundments was conducted to determine the level and type of ecological risk assessment modeling. A series of GIS overlay procedures was developed and employed to examine spatial relationships between each surface impoundment site and (1) managed areas (such as parks and wildlife preserves), (2) land use categories, (3) permanently flooded woodlands, (4) Bailey's ecoregions, (5) fishable water bodies, (6) soils, and (7) groundwater geology. Attachment B4-3 contains a more detailed description of ecological screening overlay procedures.

US EPA ARCHIVE DOCUMENT

Table A-8. Table Structure for Population, Residence, and Well Data

Variable Name	Description	Data Type
FAC_ID	Unique facility ID	Text
IMP_ID	Unique impoundment ID	Integer
PUBW_1	No. public GW wells 0 - 150m	Float
PUBW_2	No. public GW wells 151 - 500m	Float
PUBW_3	No. public GW wells 501 - 1000m	Float
PUBW_4	No. public GW wells 1001 - 2000m	Float
PRIDW_1	No. private DW GW wells 0 - 150m	Float
PRIDW_2	No. private DW GW wells 151 - 500m	Float
PRIDW_3	No. private DW GW wells 501 - 1000m	Float
PRIDW_4	No. private DW GW wells 1001 - 2000m	Float
IRRW_1	NO. irrigation GW wells 0 - 150m	Float
IRRW_2	No. irrigation GW wells 151 - 500m	Float
IRRW_3	No. irrigation GW wells 501 - 1000m	Float
IRRW_4	No. irrigation GW wells 1001 - 2000m	Float
COWW_1	No. livestock GW wells 0 - 150m	Float
COWW_2	No. livestock GW wells 151 - 500m	Float
COWW_3	No. livestock GW wells 501 - 1000m	Float
COWW_4	No. livestock GW wells 1001 - 2000m	Float
DKW_1	No. DK GW wells 0 - 150m	Float
KW_2	No. DK GW wells 151 - 500m	Float
DKW_3	No. DK GW wells 501 - 1000m	Float
DKW_4	No. DK GW wells 1001 - 2000m	Float
OTHERW_1	No. other GW wells 0 - 150m	Float
OTHERW_2	No. other GW wells 151 - 500m	Float
OTHERW_3	No. other GW wells 501 - 1000m	Float
OTHERW_4	No. other GW wells 1001 - 2000m	Float
CENPRW_1	No. 1990 census private GW wells 0 - 150m	Float
CENPRW_2	No. 1990 census private GW wells 151-500m	Float
CENPRW_3	No. 1990 census private GW wells 501 - 1000m	Float
CENPRW_4	No. 1990 census private GW wells 1001 - 2000m	Float
RES_1	Estimated residences 0 - 150m	Float
RES_2	Estimated residences 151 - 500m	Float
RES_3	Estimated residences 501 – 1000m	Float
RES_4	Estimated residences 1001 – 2000m	Float
POP_1	Estimated population 0 – 150 m	Float
POP_2	Estimated population 151 – 500m	Float
POP_3	Estimated population 501 – 1000m	Float
POP_4	Estimated population 1001 – 2000m	Float

US EPA ARCHIVE DOCUMENT

A.3.2 Surface Water Distances and Flow Data

Because the distance to surface water responses (survey question B12) were incomplete and did not include surface water flow data needed for the risk assessment, it was necessary to supplement these data with data collected from other sources. This data collection effort included largely a manual review of topographic maps and gathering of flow data from EPA's BASINS database.

A.3.2.1 Distance to Nearest/Nearest Waterbody and Ground Water Flow. The nearest fishable waterbody (FWB) in any direction and the nearest, downslope FWB (stream, lake, or pond) were identified using survey responses, site maps, atlases, topographical maps, and aerial photographs. FWBs were selected based on the following criteria:

- Lakes beyond the facility boundary but within 2 kilometers of the SI
- Streams that extended beyond the property boundary
- Streams that were order 3 or larger (The order of the stream was determined by tracing the convergence of tributaries with order 1 assigned to the furthest upstream segment indicated on the 1:24,000 topographic map (both ephemeral and perennial steams were assigned order 1). The streams were traced also using state atlases, hydrologic unit maps, and basin maps on the EPA "Know Your Watershed" web pages
- Waterbodies that did not meet the above criteria, but were closer to the SI than other waterbodies and were specifically mentioned by the respondent in Part B of the survey.

To determine the potential for a groundwater migration pathway from the SI to the FWB the following criteria were used:

- Respondent's geology summary from the B-form of the survey
- Regional geology information
- Topography as indicated on 1:24,000 or other available topographic maps.

In most cases the topography and stream flow were used as an indication of shallow groundwater flow to evaluate the potential contamination pathway to the FWB. In areas where there was the potential for fracture flow in shallow, hard-rock aquifers or in karst formations, it was automatically assumed that transport to the FWB was possible. Regional geology also was considered. Additional information was obtained from the following sources if supplemental information was required:

- DRASTIC: *A Standard System for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings*, Kerr ERL, Table 8, p.9 (Aller et al. 1987).

US EPA ARCHIVE DOCUMENT

- Hunt, 1974, *Natural Regions of the United States and Canada*, map of surface deposits, p. 122.
- USGS, 1984. Geologic Map of the United States
- USDA, state soil surveys as available
- Professional judgment.

Distances from the SI to the FWBs were typically measured using USGS 1:24,000 topographic maps. A 1:24,000 scaled rule was used for measurements, eliminating the need for conversions. Some facility packages did not include USGS maps, and a calculated scale was used.

Later analysis using generally newer aerial photography data (1977, 1997, 1998) than the USGS topographical maps indicated that some streams and lakes had been missed or were no longer present. The table was corrected based on this information. Independent, duplicate analysis was performed using the original assessment results for each facility as a 100 percent quality control check.

A.3.2.2 Collection of Surface Water Flow Data. Flow data were collected for all identified non-quiescent water-bodies (i.e., streams and rivers). Surface area was obtained for quiescent water-bodies (i.e., ponds and lakes). Stream attribute data included mean flow, 7Q10 flow, and stream width. The 7Q10 flow is representative of drought/low-flow conditions. No data were obtained for bay or ocean areas.

Three data sources were used to obtain stream data:

- EPA Office of Water. May 1996a. Database for *Better Assessment Science Integrating Point and Nonpoint Sources* (BASINS). EPA-823-R-96-001.
- Web pages: USGS, January 26 through 28, 2001. *United States NWIS-W Water Data Retrieval* Internet Site: http://waterdata.usgs.gov/nwis-w/us/
- van der Leeden et al., 1990. *The Water Encyclopedia - Second Edition*: Table 3-6 Flowing Water Resources of the United States, Data Source: Keup, L.E., 1985. Lewis Publishers, Inc., pp. 176.

EPA's BASINS model was the primary data source for 7Q10 and mean stream flow data for approximately 219 streams. The streams were found by searching the data tables and using GIS techniques to compare the site's latitude and longitude with the gaging station's coordinates. This search yielded all gages within 10 miles of the facility and a manual search was performed to narrow the list to modeled streams. The distance between the gage and the facility was calculated from the coordinate data using a spreadsheet. This calculation was validated for two facilities (in the Northwest and Southeast) using hand calculations and map comparisons.

US EPA ARCHIVE DOCUMENT

The USGS's NWIS-W water data and associated state geological survey web pages were used to obtain flow data for 56 streams that did not have appropriate data in the BASINS database. Typically, annual mean flow data were available from this source; however, only daily mean values were available for several streams. A relative annual mean was calculated from the daily mean values based on the data collected over the last 5 years or the available period of record. Generally, there was not enough data to obtain a 7Q10 value, and most of the data were not available in a format that could be downloaded into a digital file. As a result, a representative annual 7Q10 flow for streams was not calculated using the USGS gage data.

The distance along lines of latitude and longitude from the facility to gage stations was also provided in the tabulated results. Flow data webpages for the streams were printed and used to check the inputs of the original data table. A quality control check was performed on approximately six of the mean flows calculated from daily averages.

Table 3-6 Flowing Water Resources of the United States by Keup (van der Leeden et al., 1990) was used to estimate flow data for approximately 115 streams that were not listed in BASINS or on the USGS websites. The table correlates the stream's measured width and its estimated mean flow. Estimates based on the Keup's data are from end-of-stream locations. If actual data existed even many miles away (usually for large rivers) from the facility, the mean for the gage data was also presented along with the estimated Keup flow. Interpolations from the Keup data were independently verified.

The width of every stream was measured. At the end of the data collection activities, all data were queried to compare measured stream widths and flows to the estimated flows from the Keup's table. The query results showed that interpolations from the Keup data for mean annual flow compared well with actual gage information obtained from BASINS and the USGS sources. The estimates of the mean flow data appeared congruous for the set of streams to be modeled.

The surface areas for most of the lakes, ponds, and river inlets were measured on USGS 1:24,000 topographic maps using a planimeter. Some maps were of a different scale, and the planimeter was calibrated accordingly. Some waterbodies had areas below the limit of the planimeter. The areas of these waterbodies were estimated by multiplying the measuring length and width to find the square area. Some inlet areas were considered lake-like and areas were also determined for these waterbodies. The areas of approximately six, randomly selected waterbodies were independently checked for accuracy.

A.4 Data Processing

Data processing includes the calculations, conversions, and transformations necessary to prepare the basic survey data in the data entry database (described in Section A.2), for additional exploration and analysis. Data processing activities produced three primary products:

- Consolidated database, which is similar in basic structure and content as the data entry database except that units have been standardized and initial data cleaning (e.g., correction of skip pattern errors) has been conducted.

US EPA ARCHIVE DOCUMENT

- Risk assessment input database, which contains the chemical concentrations and associated surface impoundment characteristics necessary for modeling risks from surface impoundments.

- Derived variables, which were developed from survey data, risk assessment results, or combinations of these for estimating population characteristics using the statistical methodologies described in Section A.5.

Each of these processing activities is described in greater detail below.

The automated programs developed to create each of these data products were subjected to rigorous and complete QA/QC protocols. For all automated data processing, the data extraction/processing system was thoroughly validated before use. This involved manually checking enough of the data (usually 5 percent to 10 percent) to ensure that the system functioned properly. When conducting such checks, the QC protocol required that each unique calculation or data combination be checked at least once. In addition, a version control system was employed to ensure data integrity and that each analysis conducted with the most recent dataset.

A.4.1 Consolidated Database

The consolidated database is intended to serve as the final archive of survey data and contains the data in a form that makes it makes it useable for future analyses. To achieve these objectives, the consolidated database was designed to be consistent with the following criteria.

- Accurate reflection of survey responses, including margin notes as possible.

- Cleaning of conflicting responses and correction of skip pattern errors by respondents completing the survey.

- Conversion of quantitative data to standard units to enable meaningful analyses to be conducted.

- Collapse of chemical data from 8 tables in the data entry database into a single table in the consolidated database to allow easy comparison between the different sampling points (influent, effluent, and within the impoundment) and media (wastewater, sludge, leachate, air) requested by the survey.

Processing of the survey data in accordance with the last two criteria presented the biggest challenge, both from a programming and QA perspective. The chemical data from survey questions C23 and C24, required the most complex processing to convert units, calculate concentrations and mass per unit time values when possible, average different sampling periods, and, for wastewater influent only, combine data across multiple influent points. Over 40 different units used by survey respondents had to be converted to standard units for wastewater, leachate, sludge, and air. To document this process, Attachment A6 provides the data processing algorithms and unit conversions used for processing the chemical data. Each of these algorithms

US EPA ARCHIVE DOCUMENT

was checked manually during QC for correct program functioning. Records of these checks are available and archived and were subjected to a final GA audit.

To find conflicting data and skip pattern errors, SQL queries were performed based on the structure of the original survey form. For example survey question C2a asks if the impoundment has ceased receiving wastes since June 1, 1990, based on the response to this question the respondent should have or should not have responded to questions C2b, C2c, and C27a. By searching for instances that the response to question C2a was "no" or "don't know", but any one of C2b, C2c, or C27a is answered, or conversely if the response to question C2a was "yes", but there were no responses for C2b, C2c and C27a, problems were identified and rectified by looking at the responses to the questions as well as any margin notes made by the respondent. If the respondent clearly indicated that the impoundment was closed, but failed to answer the follow up questions, then non response codes could be entered as appropriate. In a couple of cases, the response to C2a was "yes", but the margin note for C27a clearly indicated that the impoundment was not closed. In this case the C2a response was rectified with the C27a margin note, and the change was noted in the C2a margin note. There also were instances when respondents answered all questions in some manner even if a response was not required. These spurious responses remain in the original data entry database, but were not transferred to the consolidated database.

Values for any survey response that could result in values reported with varying units were converted to a standard set of units. The only exception to this is the liner thickness response to question C12. Because liner thicknesses can vary greatly in magnitude by liner type, the values were transferred as provided with the actual survey units listed in the field provided. Similarly the chemical concentration data required conversion to a standard set of units as well as calculating concentrations and mass per unit time values, and combining data across multiple influent points for survey question C24a. Processing of the chemical data would also vary based on the type of data provided. Processing for concentration values, mass per unit time values, chemicals present with quantity unknown, non-detects, etc all required slightly different processing. A6-1 is an algorithm for the processing of each type of chemical data by survey question. This algorithm was used to document the processing as well as being used for quality control. The conversion functions written to convert survey data to a standard set of units are provided in A6-2.

Attachment A7 provides the basic design documents for the consolidated database. As described in section A.2 for the data entry database, these include entity relationship diagrams, a data dictionary, and copies of each coding table.

A.4.2 Risk Assessment Input Data

The risk assessment input database includes all of the chemical concentration data needed to run the risk assessment models. Design documents for this database, including a data dictionary and coding tables, are provided in Attachment A8. The risk assessment input database was populated in accordance with the risk assessment Technical Plan (see Appendix C), and included the following conventions developed to help reduce missing data and to ensure that the screening analysis was adequately protective.

US EPA ARCHIVE DOCUMENT

- Values below detection limits were entered at the detection limit given in the survey. Where the detection limit was not specified, a lookup table of default detection limits was used to fill a detection limit value in the risk assessment database.

- A nearest neighbor imputation methodology was applied to develop surrogate concentration data were chemicals are expected to be present, but quantities are unknown.

- Where sludge data were not available, partition coefficients were used to estimate sludge concentrations from wastewater concentrations

Each of these procedures is discussed in more detail in the following sections.

A.4.2.1 Detection Limits. Where the survey respondent entered a concentration value as less than a detection limit, for example "< 0.05 mg/L", a value at the detection limit (i.e., 0.05 mg/L) was placed in the risk assessment input database. This protective convention was adopted for the screening risk assessment to ensure broad coverage in cases where a chemical could be just below the detection limit. To ensure that this assumption is not overly conservative, chemical concentrations still exceeding risk criteria after the final phase of the risk analysis were examined as to their source (i.e., detection limit, surrogate) (see Section 3 and Appendix C).

For cases where the survey did not provide a detection limit (e.g., specified "not detected" or "ND"), a lookup table of default detection limits was developed considering analytical methods likely to be used for wastewater samples. The primary sources are summarizes as follows.

- **Wastewater.** EPA method 1624 and 1625 were selected because the methods are designed to meet the requirements of NPDES under 40 CFR parts 136.1 and 136.5. For inorganics (metals) standard methods for inductively coupled plasma (ICP) and cold-vapor atomic adsorption (CVAA) analyses were used. When detection limits for organic constituents were not available from these methods, the EPA 600 series for municipal and industrial wastewater was used. For any remaining constituents without detection limits, SW-846 EPA 8000 series was used. Finally, if no method was available, then a detection limit was pulled from the available detection limits in the survey database.

- **Sludge.** For the organics SW-846 EPA 8000 series was used. For method 8021, the method provided an estimated quantitation limit (EQL) of 0.1 mg/kg, and this value was used for applicable constituents. For 8081, a factor for sludge was calculated into the detection limit as noted in the spreadsheet. Finally, if no method is referenced, then the detection limit was pulled from the available detection limits in the survey database.

- **Air.** Detection limits in air were taken from the EPA report Ambient Measurement Methods and Properties of the 189 Clean Air Act Hazardous Air

US EPA ARCHIVE DOCUMENT

Pollutants. When a method is not referenced, then the detection limit was based on best professional judgment.

All detection limits were multiplied by a factor of ten to account for interferences. The final 1x and 10x values for wastewater, sludge, and air are provided in Attachment A-9 (Tables A-9-1, A-9-2, and A-9-3).

A.4.2.2 Surrogate Values. The Surface Impoundment Study Technical Plan for Human Health and Ecological Risk Assessment (Attachment C) specifies that in cases where the presence of a chemical in an impoundment can be inferred, but it is not possible to quantify a value, a value from a similar impoundment will be used to represent a likely concentration. These surrogate values were developed using a nearest neighbor imputation method which made it possible to maximize the use of presence information in the survey. Presence was inferred, and surrogate concentrations were sought, in three cases: (1) where the respondent had checked the "present but quantity unknown" (PQU) flag, (2) where the respondent had entered a chemical but provided no value (and did not check PQU), and (3) where chemicals were reported in wastewater effluent (to infer presence within the impoundment).

The imputation methodology is picture in Figure A-4 and described in the following text. Note that because detection limits were decided to be valid representations of concentrations in the impoundments for this risk analysis, the detection limit values described in Section A.4.2.1 were available and used for surrogates. All surrogate data processing was done on the constituent level and the maximum of the surrogate data gets filled into the CHEM_CONC Table in the risk assessment database with "Surrogate" marked as true.

The imputation methodology employed a decision framework that was programmed into a data processing system to implement the methodology. The theme throughout the process is to find the most similar impoundment possible within the survey database that had data for the chemicals without values. Steps in the process include answering the following questions:

1. Are there any other impoundments at the same facility with data for the constituent?
   Yes
   1a. Are there any impoundments with the exact same treatment processes?
      Yes - fill surrogate data - finished
   1b. If the impoundment requiring surrogate data is aerated, are there any other impoundments which are aerated?
      Yes - fill surrogate data - finished
   1c. Are there any impoundments which perform the same function (treatment or non-treatment only)?
      Yes - fill surrogate data - finished

2. Are there any other impoundments with the same 6 digit SIC code with data for the constituent?
   Yes
   2a. Are there any impoundments with the exact same treatment processes?
      Yes - fill surrogate data - finished

US EPA ARCHIVE DOCUMENT

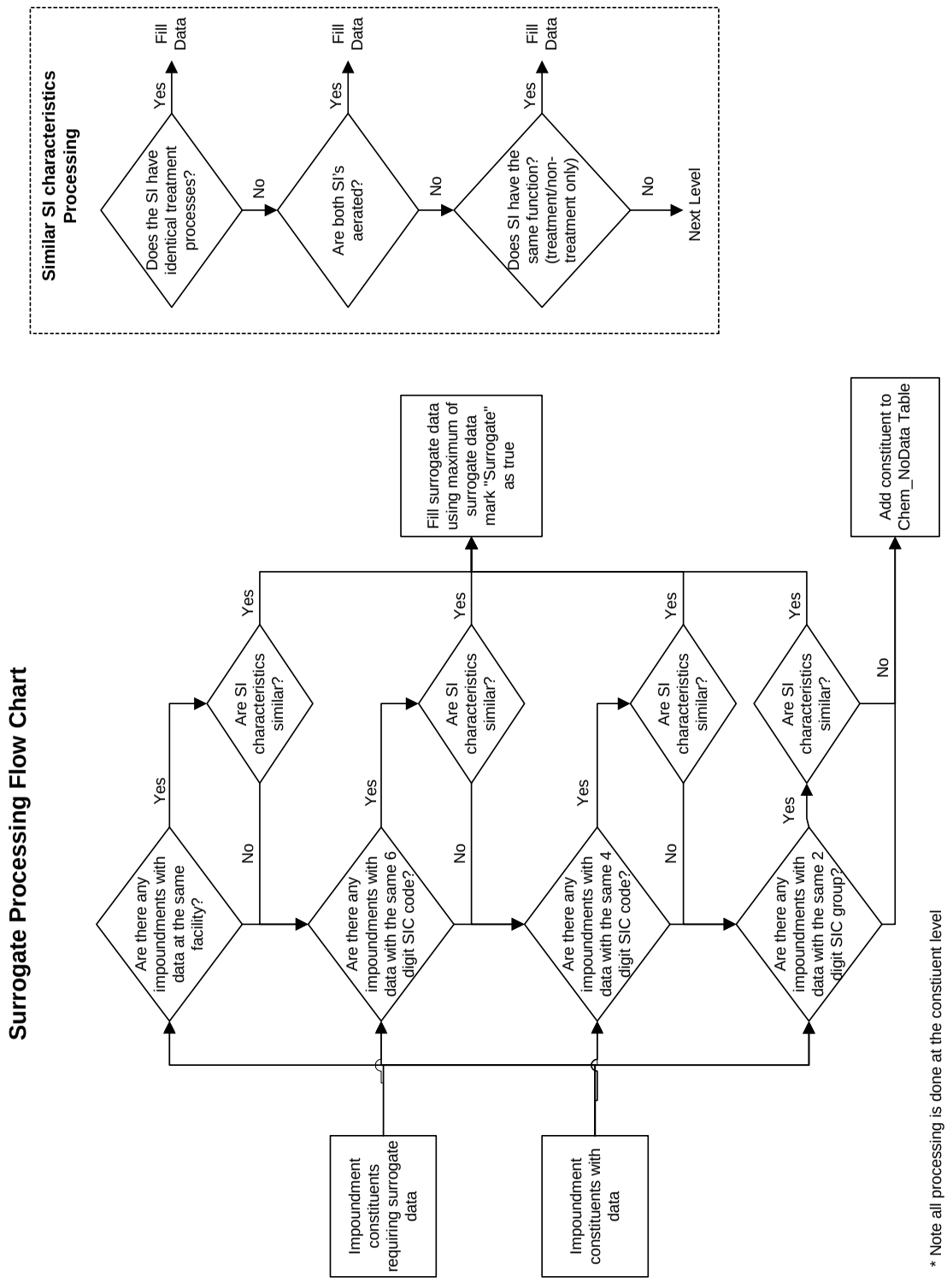

Figure A-4. Decision tree for identifying surrogate data for risk assessment

US EPA ARCHIVE DOCUMENT

2b. If the impoundment requiring surrogate data is aerated, are there any other impoundments which are aerated?
Yes - fill surrogate data - finished
2c. Are there any impoundments which perform the same function (treatment or non-treatment only)?
Yes - fill surrogate data - finished

3. Are there any other impoundments with the same 4 digit SIC code with data for the constituent?
Yes
3a. Are there any impoundments with the exact same treatment processes?
Yes - fill surrogate data - finished
3b. If the impoundment requiring surrogate data is aerated, are there any other impoundments which are aerated?
Yes - fill surrogate data - finished
3c. Are there any impoundments which perform the same function (treatment or non-treatment only)?
Yes - fill surrogate data - finished

4. Are there any other impoundments with the same 2 digit industry group with data for the constituent?
Yes
4a. Are there any impoundments with the exact same treatment processes?
Yes - fill surrogate data - finished
4b. If the impoundment requiring surrogate data is aerated, are there any other impoundments which are aerated?
Yes - fill surrogate data - finished
4c. Are there any impoundments which perform the same function (treatment or non-treatment only)?
Yes - fill surrogate data - finished

5. If there are still constituents requiring surrogates which can not be matched in steps 1 to 4, then add the constituents to the Chem_NoData table.

A.4.2.3 Estimating Sludge Concentrations from Wastewater Concentrations. When there is not a sludge concentration provided in the survey, but there is sludge within the impoundment, a sludge concentration was estimated from using waste-water partition coefficients (Kdw) for metals and a soil organic carbon-water partition coefficient (Koc) for organic constituents, along with total suspended solids (TSS) data pulled from the study survey. This approach accounts for contaminants sorbed to TSS, which is necessary when using total wastewater concentrations (versus dissolved).

The equations to be used are.

Kd (metals):

US EPA ARCHIVE DOCUMENT

Sludge_Conc = WW_Conc*([Kdw_L/kg]/(1+([Kd_L/kg]*([TSS_WW]/1000000))))

where:

WW_Conc is the wastewater concentration within the SI in mg/L
Kdw_L/kg is the 50th percentile waste-water Kdw value in L/kg
TSS_WW is the TSS value in mg/L.

Koc (organics):

Sludge_Conc = WW_Conc*(([Koc]*foc)/(1+(([Koc]*foc)*([TSS_WW]/1000000))))

where:

WW_Conc = wastewater concentration within the SI in mg/L
Koc = soil organic carbon / water partition coefficient in L/kg
foc = fraction organic carbon (waste solids)
TSS_WW = TSS value for wastewater in mg/L.

These two equations were derived from the following equation:

C_sol (mg/kg) = C_ww (mg/L) * { Kdw/ (1+Kdw [TSS]) }

where:

C_sol = solids concentration (sorbed, mg/kg)
C_ww = measured wastewater sample contaminant concentration (total, mg/L)
Kdw = waste-water partitioning coefficient = Koc*foc for organics (L/kg)
TSS = total suspended solids concentration (kg/L = g/cm3).

Total Suspended Solids (TSS). TSS values were obtained from the SI survey database (question C15) using the following hierarchy:

1. Use wastewater within the impoundment TSS value (WW_TSSV)
2. Use wastewater within the impoundment Total Solids value (WW_TSOLV)
3. Use wastewater within the impoundment MLSS value (WW_MLSSV)
4. Use wastewater within the impoundment MLVSS value (WW_MLVSS)
5. Use wastewater within the impoundment Biomass Concentration value (WW_BIOV)
6. Use wastewater influent TSS value (INF_TSSV)
7. Use wastewater influent Total Solids value (INF_TSOLV)
8. Use wastewater influent MLSS value (INF_MLSSV)
9. If still no data, use IWAIR default of 0.2 g/L (200 mg/L)

Fraction Organic Carbon (foc). With respect to foc, some correlations developed for biomass sludge in activated sludge systems suggest that an foc around 0.7 would be reasonable. Given that MLVSS is a measure of solids that volatilize at about 550 degrees centigrade, it is also reasonable to estimate fraction organic carbon (foc) in wastewater solids using the following equation:

US EPA ARCHIVE DOCUMENT

foc = MLVSS / (TSS or MLSS)

SI Survey data (MLVSS/TSS) that can be used to estimate foc for wastewater solids are limited to 99 pairs of MLVSS/TSS or MLVSS/MLSS data for influent, effluent, and wastewater in the impoundment (Table A-9). Review of these data shows little difference between the different sampling points and limited variability (overall coefficient of variation = 0.3). Based on these data the median fraction organic carbon (foc) value of 0.7 (70 percent organic carbon) was used as a typical value.

Table A-9. Summary Statistics: MLVSS/TSS

Medium	n	mean	StdDev	CV	min	10th %ile	median	90th %ile	max
wastewater	37	0.68	0.16	0.24	0.20	0.46	0.71	0.82	0.88
influent	30	0.61	0.28	0.46	0.03	0.12	0.71	0.89	0.93
effluent	32	0.71	0.15	0.20	0.32	0.51	0.70	0.88	1.00
all	99	0.67	0.20	0.31	0.03	0.32	0.70	0.86	1.00

StdDev = Standard deviation.
CV = Coefficient of variation (StdDev/mean).

Partition Coefficients for Organics (Koc). Soil organic carbon-water partition coefficients (Koc values) were extracted from the following readily available sources (listed in order or preference):

- IWEM model datafiles (178 values). This will ensure consistent Koc values with the groundwater modeling results described in Appendix C. IWEM Koc values are reported to be collected from Kollig et al. (1993).
- Kollig et al. (1993; 2 values), the EPA ORD reference containing peer-reviewed Koc values used in EPACMTP and IWEM.
- Superfund Chemical Data Matrix (SCDM; 29 values). Well-referenced EPA Superfund values used in Hazard Ranking System (HRS) (U.S. EPA, 1996b). Available online.

Values were found from these sources for most of the study organic chemicals. Koc values for nine study constituents were developed as follows:

- Extract log octanol / water partition coefficient (log Kow) from Hansch et al. (1995).
- Use following equation to calculate log Koc values from log Kow :

  log Koc = log Kow + 0.32.

US EPA ARCHIVE DOCUMENT

This equation was used in HWIR and in Kollig et al. (1993) to calculate Koc values.

The final Koc values used for sludge estimatation methodology as well as for the groundwater pathway exposure modeling are provided in Attachment A9. As shown above, Koc x foc was used to estimate waste-water partition coefficients (Kdw values) for organic constituents in the sludge estimation methodology. The two most significant uncertainties in this assumption are:

- the accuracy of applying soil Koc values to wastes
- limited data on waste organic carbon content.

Depending on waste streams, organic carbon content could contribute up to about a two-order of magnitude uncertainty factor to the Kdw value. The magnitude and impact of uncertainty introduced by the the applicability question is unknown.

Wastewater Partition Coefficients for Metals (Kdw). Waste solids / water Kdw values for metals were obtained from the HWIR modeling effort (U.S. EPA, 1999c). HWIR developed distributions from collected literature values for soil, sediment, suspended matter, and wastes. For wastes managed in surface impoundments, HWIR uses metal partition coefficients (Kd values) collected for suspended matter in surface water bodies. The distributions were based on collected data or, for metals where data were inadequate, using a regression equation relating soil and suspended matter log Kd values collected for other metals. These data show that suspended matter tends to have 2 to 3 times the affinity for metals than soil. This has been attributed to the higher surface area and organic carbon content of suspended particulate matter, which are also characteristics of solids in many industrial surface impoundments. The distributions for metals are provided in Attachment A9.

Significant uncertainties associated with the wastewater partition coefficients for metals include:

- Literature values are not a random, nonbiased sample and thus may not adequately represent the true distribution of partition coefficients.

- The accuracy of applying soil data to suspended solids; r^2 for the HWIR soil / supended matter regression equation is 0.37. However, the calculated values appear to be roughly in line with the measurements collected from published literature for other metals.

- The accuracy of applying surface water suspended solids data to waste solids.

The magnitude and impact of these uncertainties are uncertain in themselves, but, given the variability in partition coefficients, could be several orders of magnitude for a particular metal in a particular impoundment.

US EPA ARCHIVE DOCUMENT

A.4.3 Derived Variables for Exploration and Analysis

Both the survey findings presented in Chapter 2 and section A.5.2 of this attachment, and the risk results provided in Chapter 3 required weighting up to the entire population of facilities represented by the survey so that national level observations and conclusions could be made about nonhazardous industrial surface impoundments. This required development of derived variables and populating them from the surface impoundment database and the risk results. As with the other database discussed above, this was accomplished using automated data processing programs that were subjected to rigorous, complete QA/QC protocols to ensure that the programs are functioning as designed (i.e., all algorithms and calculations were hand-checked for each unique data situation).

Attachment A10 includes detailed specifications, by report question and variable, used to develop and check these derived variables. In each case, the source and destination of each variable is included in these tables, which are organized by section Chapter 2, Appendix B, and Appendix C) and variable level (facility or impoundment).

A.5 Data Analysis Methods

This section describes the statistical methodology underlying the population estimates computed using screener and long survey data. It is divided into two sections. The first section discusses how the statistical analysis weights were computed to account for the sampling design and to reduce the bias due to nonresponse. The second section discusses how these weights and features of the sampling design were used to compute robust, design-consistent estimates of sampling variances, standard errors, and confidence interval estimates of population parameters.

A.5.1 Statistical Analysis Weights

The statistical analysis weights for the observational units in any probability-based sample survey are the initial sampling weights adjusted to reduce the potential for bias due to survey nonresponse. The initial sampling weight for each unit is the reciprocal of the probability that the unit was selected into the sample. If each unit could have more than one linkage to the sampling frame (or list) from which the sample was selected, the initial sampling weights must be adjusted to compensate for this multiplicity. Finally, a model-based estimate of the probability of responding is usually used to reduce the bias due to nonresponse. The following sections discuss each of these steps for computing the statistical analysis weights for the Surface Impoundment Study.

A.5.1.1 Initial Sampling Weights. As described in Section A.1.1, major differences in the sources and availability of sampling frame data led to the definition of three primary sampling strata based on the facility's regulatory status under the Clean Water Act:

For direct discharge facilities, EPA constructed an essentially complete sampling frame of 43,050 facilities from the NPDES permits in the EPA's Permit Compliance System (PCS) database. EPA partitioned the sampling frame into three primary sampling strata, defined as follows:

US EPA ARCHIVE DOCUMENT

1. Facilities in high-priority SICs (26, 2819, 2824, 2834, 2869, 2897, 2911, 30, 33, or 36)
2. All other facilities with in-scope SICs
3. The six pilot study facilities.

Substrata were defined based on SIC codes resulting in a total of 15 sampling strata. A stratified simple random sample of 2,000 facilities was selected from the 15 sampling strata, and the six pilot study facilities were retained with certainty.

For zero discharge facilities, a sampling frame of 5,807 facilities was constructed from available state data and two federal databases: EPA's Toxics Release Inventory (TRI) and the Aerometric Information Retrieval System, Facility Subsystem (AFS). The sampling frame was stratified into 15 sampling strata based on general categories of completeness for the different state and federal data sources, and according to high and low priority SIC codes. A stratified random sample of 250 facilities was selected using the same sampling rate for all but one stratum.

Because local POTWs are the principal permitting authorities for indirect discharge facilities, anecdotal information collected from EPA, state and local personnel, and database information from EPA Region 7 was used to construct a sampling frame from which 35 facilities were purposively selected.

Subsequent to selection of this sample for the screener survey, EPA determined that some of the sample facilities were ineligible for Phase 2 of the study, and those facilities were removed from the sample before mailing the screener surveys. For each of the 1,984 direct and zero discharge facilities mailed a screener survey, the initial sampling weight was computed for the j-th facility in stratum r as follows:

$$w1(j) = N1(r) / n1(r) ,$$

where

N1(r) = Total number of facilities in stratum r, and
n1(r) = Number of facilities selected into the sample from stratum r.

The frame count, N1(r), sample size, n1(r), and initial sampling weight, w1(j), are shown for each stratum in Table A-10. Sampling weights were not computed for the sample of 35 indirect discharger facilities because the sample was purposively selected and the survey results cannot be statistically extrapolated to any larger population.

A.5.1.2 Multiplicity Adjustments. The PCS data used to construct the sampling frame for the direct discharger sample were outfall- or pipe-level records. The first step was to collapsed the pipe-level records to the permit level by permit ID (NPID). Permits were then combined to the facility level. Because there was no unique facility ID to guide this process, permits were

US EPA ARCHIVE DOCUMENT

Table A-10. Initial Sampling Weights for the Screener Survey

Sampling Stratum	Frame Count	Sample Size	Initial Weight
Direct Discharge Facilities			
High-priority SICs:			
■ SIC 26	927	142	6.528
■ SIC 28	1019	156	6.532
■ SIC 29	440	67	6.567
■ SIC 30	1478	226	6.540
■ SIC 33	1752	268	6.537
■ SIC 36	919	141	6.518
Low-priority SICs:			
■ SIC 20-23	5169	141	36.660
■ SIC 24-27	3442	95	36.232
■ SIC 28-31	3000	82	36.585
■ SIC 32	3212	88	36.500
■ SIC 34	2680	73	36.712
■ SIC 35-39	3642	100	36.420
■ SIC 42-45[a]	2688	74	36.324
■ SIC 49[b]	9276	254	36.520
■ SIC 50-76[c]	3400	93	36.559
Pilot study facilities (certainty selections)	6	6	1.000
Direct discharger subtotal	43,050	2,006	NA
Zero Discharge Facilities			
States with complete databases:			
■ In TRI or AFS	228	13	17.539
■ High-priority SICs[d]	61	5	12.200
■ Low-priority SICs[d]	301	13	23.154
■ Unknown SIC[d]	1155	55	21.000

(continued)

US EPA ARCHIVE DOCUMENT

Table A-10. (continued)

Sampling Stratum	Frame Count	Sample Size	Initial Weight
■ SIC 4952[d]	891	22	40.500
States with general databases:			
■ In TRI or AFS	128	6	21.333
■ High-priority SICs	127	6	21.167
■ Low-priority SICs	543	25	21.720
■ Unknown SIC	1592	74	21.514
■ SIC 4952	95	3	31.667
States with partial databases:			
■ In TRI or AFS	116	4	29.000
■ With target SICs	121	6	20.167
■ Unknown SICs	117	4	29.250
■ SIC 4952	138	8	17.250
States with no relevant databases			
■ In TRI or AFS	194	6	32.333
Zero discharger subtotal	5,807	250	NA

TRI = EPA's Toxic Release Inventory.
AFS = EPA's Aerometric Information Retrieval System, Facility Subsystem.
[a]SICs 4212, 4213, 4231, and 4581
[b]SICs 4952 (excluding Publicly Owned Treatment Works), 4953, and 4959
[c]SICs 5085, 5093, 5169, 5171, and 7699 (transportation equipment cleaners only)
[d]Not in TRI or AFS.

merged to the facility level only when it was quite clear that there were multiple permits for the same facility. Up to 3 different permits were merged into a single facility-level record. Any facilities that had multiple permits that did not get merged into a single facility-level record on the sampling frame had multiple chances of being selected into the sample.

The screener survey listed all permits that had been used to define the facility on the sampling frame, and asked each facility to list any additional permits that had been active for the facility at any time since June 1, 1990. After considerable data cleaning, the multiplicity (number of linkages to the sampling frame) was determined for each facility that responded to the screener survey.

However, the frame multiplicity must be known for every sample facility, not just the responding facilities. Therefore, for each direct discharger sampling stratum, the average

US EPA ARCHIVE DOCUMENT

multiplicity was computed among the respondents and the multiplicity was imputed for each nonresponding facility within each sampling stratum to be the average multiplicity for that stratum. After having computed or imputed the multiplicity, m(j), for each direct discharge sample facility, the multiplicity-adjustment to the sampling weight was computed for the j-th facility as follows:

$w2(j) = 1 / m(j)$	for direct discharge facilities
$w2(j) = 1$	for zero discharge facilities.

Lessler and Kalsbeek (1992, Section 5.2.2) show how this using this multiplicity adjustment produces survey estimates that are design-unbiased.

A.5.1.3 Adjustment for Nonresponse to the Screener Survery. Weight adjustments to reduce the bias due to survey nonresponse are based on models for the probability of responding, using data that are available for both respondents and nonrespondents. Since the sampling stratum was the only thing we knew about the nonresponding facilities, we used sample-based ratio adjustments based on the sampling strata (see Brick and Kalton, 1996). The nonresponse adjustments were defined only for the direct and zero discharge facilities because the indirect discharger sample was not a probability-based sample.

The weight adjustment for nonresponse is simply the reciprocal of the weighted response rate in each weighting class. Therefore, strata for which the number of respondents was small (e.g., less than 20) were collapsed with similar strata to form weighting classes. However, assigning strata with dissimilar response rates to different weighting classes is necessary to reduce nonresponse bias.

Hence, after reviewing the pattern of study eligibility and survey response by sampling strata, it was decided that each of the 15 sampling strata for the direct discharger sample contained sufficient numbers of respondents to be a separate weighting class, and they are the first 15 weighting classes. However, because of the smaller sample size for the zero discharger sample, strata were combined to form weighting classes as follows:

- Weighting class 16 consists of zero discharger strata 1 through 4: the facilities from the TRI or AFS portion of the sampling frame;
- Weighting class 17 consists of zero discharger strata 5, 6, and 9: the facilities with high-priority SICs; and
- Weighting class 18 consists of the remainder of the zero discharger facilities.

Having defined the weighting classes for nonresponse adjustment, the weight adjustments were implemented for nonresponse in two stages. First, an adjustment was made for inability to determine whether or not a facility was eligible for the screener survey (i.e., was in operation at any time since June 1, 1990). The second stage of nonresponse adjustment was an adjustment for nonresponse among the facilities known to be eligible for the screener survey.

US EPA ARCHIVE DOCUMENT

The weight adjustment factor for inability to determine eligibility for the screener survey was computed for the *c*-th weighting class follows:

$$w_3(c) = \frac{\sum_{j \in c} w_1(j)\, w_2(j)}{\sum_{j \in c} w_1(j)\, w_2(j)\, I_k(j)} \quad ,$$

where $I_k(j)$ is an indicator that the eligibility status or the *j*-th facility is known, i.e.,

$I_k(j) = 1$ if the eligibility status of the *j*-th facility is known
$I_r(j) = 0$ otherwise.

This adjustment is equivalent to assuming that the proportion of sample facilities that are eligible for the screener survey (i.e., in operation at any time since June 1, 1990) is the same for facilities both with known and unknown eligibility status.

Similarly, the weight adjustment factor for survey nonresponse was defined for the *c*-th weighting class as follows:

$$w_4(c) = \frac{\sum_{j \in c} w_1(j)\, w_2(j)\, w_3(j)\, I_e(j)}{\sum_{j \in c} w_1(j)\, w_2(j)\, w_3\, I_r(j)} \quad ,$$

where I_r and I_e are indicators of response and eligibility status, respectively, i.e.,

$I_r(j) = 1$ if the *j*-th facility was a screener respondent
$I_r(j) = 0$ otherwise, and
$I_e(j) = 1$ if the *j*-th facility was eligible for Phase 1
$I_e(j) = 0$ otherwise.

These nonresponse adjustments are shown for each of the 18 weighting classes in Table A-11.

The final statistical analysis weight for the screener survey was defined for the *j*-th facility in the *c*-th weighting class as the product of the various weight components, as follows:

$w_5(j) = w_1(j)\, w_2(j)\, w_3(c)\, w_4(c)\, I_r(j)$.

A.5.1.4 <u>*Adjustment for Subsampling for the Long Survey*</u>. Respondents to the screener survey were eligible for selection into the subsample to receive the long survey if their screener survey data indicated that they satisfied the following conditions:

- Had an in-scope SIC[1]

[1] Major groups 20-39 and 97 plus codes 4212, 4213, 4231, 4581, 4952 (except Publicly Owned Treatment Works), 4953, 4959, 5085, 5093, 5169, 5171, and 7699 (transportation equipment cleaners only).

US EPA ARCHIVE DOCUMENT

- Were in operation at any time since June 1, 1990, and the time of the survey in the summer of 1999

- Used at least one direct- or zero-discharge surface impoundment to manage only nonhazardous waste.

A stratified random sample of 201 of the screener respondents, plus the six pilot study facilities, were selected to receive the long survey. The weight component for selection of this subsample was the reciprocal of the probability of selection. It was computed for the *j*-th facility in stratum *s* as

$$w_6(j) = N_2(s) / n_2(s) ,$$

where

$N_2(s)$ = Total number of facilities in stratum *s* eligible to be selected for the long survey sample, and
$n_2(s)$ = Number of facilities selected from stratum *s* to receive the long survey.

The frame count, $N_2(s)$, sample count, $n_2(s)$, and weight component, $w_6(j)$ are shown in Table A-12 for each sampling stratum.

A.5.1.5 <u>*Calibration to Screener Survey Weight Totals*</u>. The estimated number of facilities in the survey population using the long survey weights is not identical to the estimate based on the screener analysis weights (7,459 facilities) because the screener weights were not all the same within each stratum from which facilities were selected for the long survey. The screener sample weights provide a more reliable estimate of the size of the population because of the larger number of screener respondents, relative to the long survey subsample. Therefore, the long survey weights were calibrated to sum to the screener totals within each stratum used to select facilities for the long survey. In particular, the calibration weight factor was computed for the *j*-th facility in stratum *s* as follows:

$$w_7(j) = \frac{\sum_{i=1}^{N_2(s)} w_5(i)}{\sum_{i=1}^{N_2(s)} w_5(i)\, w_6(i)\, I_{S_2}(i)} ,$$

where I_{S_2} is a (0,1) indicator of inclusion in the long survey sample. These calibration adjustment factors also are shown in Table A-12.

US EPA ARCHIVE DOCUMENT

Table A-11. Weighting Class Adjustments for Screener Survey Nonresponse

Weighting Class	Adjustment for Inability to Determine Eligibility	Adjustment for Nonresponse Among Eligible Facilities
Direct Discharge Facilities		
High-priority SICs: ■ SIC 26	1.035	1.000
■ SIC 28	1.096	1.000
■ SIC 29	1.159	1.000
■ SIC 30	1.071	1.010
■ SIC 33	1.058	1.008
■ SIC 36	1.109	1.016
Low-priority SICs: ■ SIC 20-23	1.126	1.008
■ SIC 24-27	1.044	1.023
■ SIC 28-31	1.052	1.000
■ SIC 32	1.098	1.000
■ SIC 34	1.105	1.000
■ SIC 35-39	1.074	1.033
■ SIC 42-45[a]	1.119	1.000
■ SIC 49[b]	1.315	1.038
■ SIC 50-76[c]	1.105	1.012
Pilot study facilities (certainty selections)	1.000	1.000
Zero Discharge Facilities		
In TRI or AFS	1.105	1.000
High-priority SICs[d]	1.290	1.000
All other facilities[d]	1.154	1.012

TRI = EPA's Toxic Release Inventory.

AFS = EPA's Aerometric Information Retrieval System, Facility Subsystem.

[a]SICs 4212, 4213, 4231, and 4581

[b]SICs 4952 (excluding Publicly Owned Treatment Works), 4953, and 4959

[c]SICs 5085, 5093, 5169, 5171, and 7699 (transportation equipment cleaners only)

[d]Not in TRI or AFS.

US EPA ARCHIVE DOCUMENT

Table A-12. Subsampling and Calibration Weights for the Long Survey

Sampling Stratum[a]	Frame Count	Sample Size	Subsampling Weight	Calibration Weight
Direct Discharge Facilities				
Handles formerly characteristic waste and high-priority SIC	75	75	1.000	1.000
Handles formerly characteristic waste and low-priority SIC	7	4	1.750	0.894
Does not handle formerly characteristic waste and high-priority SIC	204	68	3.000	1.003
Does not handle formerly characteristic waste and low-priority SIC	78	14	5.571	0.938
Pilot study facilities (certainty selections)	6	6	1.000	1.000
Direct discharger subtotal	370	167	NA	NA
Zero Discharge Facilities				
Handles formerly characteristic waste and high-priority SIC	2	2	1.000	1.000
Handles formerly characteristic waste and low-priority SIC	4	4	1.000	1.000
Does not handle formerly characteristic waste and high-priority SIC	14	14	1.000	1.000
Does not handle formerly characteristic waste and low-priority SIC	20	20	1.000	1.000
Zero discharger subtotal	40	40	NA	NA

[a]Based on the screener survey data.

A.5.1.6 Adjustment for Nonresponse to the Long Survey. Theoretically, all facilities selected into the sample to receive the long survey should have been eligible for this phase of the study. That is, they should all have had at least one surface impoundment that satisfied the eligibility conditions in the screener survey. However, several facilities reported that they had no eligible impoundments and had completed the screener survey incorrectly. Using extensive follow-up contacts, the eligibility status was determined for all facilities selected into the sample for the long survey. Hence, nonresponse adjustments were confined to adjustment for nonresponse among the sample facilities that were determined to be eligible for the survey.

For the full sample, there were only four eligible facilities that did not respond to the long survey, and one of those was an indirect discharge facility. Hence, for the weight adjustments for direct and zero discharge facilities, there were only three nonresponding facilities. Moreover, all three were direct discharge facilities whose screening data indicated that they did not handle any formerly characteristic waste. Therefore, the weighting classes for nonresponse to the long survey will be defined as shown in Table A-13.

US EPA ARCHIVE DOCUMENT

Table A-13. Weighting Class Adjustments for Long Survey Nonresponse

Weighting Class[a]	Eligible Facilities	Responding Facilities	Nonresponse Adjustment
Direct Discharge Facilities			
Facility does not handle formerly characteristic waste	33	30	1.084
Facility and its impoundment(s) handle formerly characteristic waste	75	75	1.000
Facility handles formerly characteristic waste, but not its impoundment(s)	38	38	1.000
Direct discharger subtotal	146	143	NA
Zero Discharge Facilities			
All	35	35	1.000

[a]Based on the screener survey data.

The statistical analysis weights for the long survey respondents then were computed by adjusting the calibrated sampling weights, $w_5 * w_6 * w_7$, for nonresponse among the eligible sample facilities. Hence, the weight adjustment factor for nonresponse to the long survey was defined for the *k*-th weighting class as follows:

$$w_8(k) = \frac{\sum_{i \in k} w_5(i)\, w_6(i)\, w_7(i)\, I_e(i)}{\sum_{i \in k} w_5(i)\, w_6(i)\, w_7(i)\, I_r(i)} \quad ,$$

where I_r and I_e are indicators of long survey response and eligibility status, respectively, i.e.,

$I_r(i) = 1$ if the *i*-th facility was a long survey respondent
$I_r(i) = 0$ otherwise, and

$I_e(i) = 1$ if the *i*-th facility was eligible to receive the long survey
$I_e(i) = 0$ otherwise.

The final statistical analysis weight for the long survey then was defined for the *i*-th facility in the *c*-th weighting class as the product of the various weight components, as follows:

$$w_9(i) = w_5(i)\, w_6(i)\, w_7(i)\, w_8(k)\, I_r(i) \, .$$

Because data were collected for *all* eligible impoundments at each responding facility (i.e., there was no subsampling of impoundments), these facility-level analysis weights also are appropriate for analysis of the impoundment-level data collected for the responding facilities.

US EPA ARCHIVE DOCUMENT

A.5.1.7 Adjustment for Item Nonresponse. Using the final statistical analysis weights, w_5 and w_9, for the 1,774 screener survey and 195 long survey respondents, respectively, reduces the potential for bias due to nonresponse of eligible facilities selected for these surveys. However, some survey items have additional missing data among these survey respondents. Failure to adjust for nonresponse to individual data items again leads to nonresponse bias. In particular, all population totals will be underestimated if item nonresponse is ignored. Statistical imputation procedures are often used to replace missing data items because they result in simpler, more consistent, analyses. However, they also have the potential to distort relationships between variables (see Brick and Kalton, 1996).

Because of concern regarding the potential distortions that can result from using imputed data, weight adjustments were used to reduce the potential for bias due to item nonresponse, exactly as they were used to compensate for total survey nonresponse. In particular, if an analysis was based on *m* variables that were constructed from long survey data, the data used in the analysis were those belonging to the facilities (or impoundments) that had complete data for all *m* variables. The weight adjustment for item nonresponse was developed as a SAS macro so that it could easily be implemented for each individual data analysis for which complete data were not available for all long survey respondents. The adjustment was a standard weighting class adjustment. Because some analyses had high levels of missing data, the weighting classes used to adjust for long survey nonresponse were collapsed to the following three weighting clases:

- Direct discharge facilities that do not manage decharacterized waste (based on the screener survey data).
- Direct discharge facilities that do manage decharacterized waste (based on the screener survey data).
- Zero discharge facilities.

Hence, the weight adjustment factor for item nonresponse to the long survey was defined for the *l*-th weighting class as follows:

$$w_{10}(l) = \frac{\sum_{i \in l} w_9(i)\, I_e(i)}{\sum_{i \in l} w_9(i)\, I_r(i)} \quad ,$$

where I_r is an indicator of respondents with data for all *m* items used in a particular analysis and I_e is an indicator of the full set of long survey respondents, i.e.,

$I_r(i) = 1$	if the *i*-th facility or impoundment has data for all *m* variables used in the particular analysis
$I_r(i) = 0$	otherwise, and
$I_e(i) = 1$	if the *i*-th facility or impoundment was a long survey respondent
$I_e(i) = 0$	otherwise.

US EPA ARCHIVE DOCUMENT

The final statistical analysis weight, adjusted for item nonresponse, then was defined for the *i*-th facility in the *l*-th weighting class as follows:

$$w_{11}(i) = w_9(i)\, w_{10}(l)\, I_r(i) .$$

Hence, each analysis was based on complete data cases with a statistical adjustment for nonresponse to the set of data items used in each particular analysis. This ensures that the estimated numbers of facilities and impoundments in the survey population are consistent across all analyses.

Nevertheless, estimates of population totals for other population characteristics (e.g., the total number of impoundments with liners) may be somewhat inconsistent from one analysis to the next because of different missing data patterns. However, when the extent of missing data is low (e.g., 10 percent or less), the inconsistencies will be small.

A.5.1.8 *Analysis Domains*. Statistical analyses were performed primarily for two populations of facilities and the surface impoundments used to manage non-hazardous wastes at those facilities. This section describes and briefly characterizes each of these populations.

The first population of particular interest consists of those facilities in the screener survey population that had at least one eligible impoundment, as defined for that survey. The specific characteristics of that population of facilities are as follows: facilities with in-scope SICs[2] in the United States that were in operation at any time between June 1, 1990, and the summer of 1999 that used at least one direct- or zero-discharge surface impoundment to manage only nonhazardous wastes resulting from any one of the following processes:

- A manufacturing process other than heat transfer
- A direct-contact heat transfer process
- Equipment washing, product washing, or washing surfaces (e.g., buildings or floors)
- Spill cleanup
- Air pollution control
- Materials handling (e.g., valve/pump drips collected in a sump and mixed with rainwater)
- Boiler blowdown
- Laundering
- Leachate (liquid percolated through or drained from a waste management unit).

Because of many false positive responses to the screener survey, the best estimate of the size of this population is based on the long survey responses. The estimated number of such facilities is 7,459, based on 184 such facilities that responded to the long survey. The estimated number of surface impoundments at these facilities that meet these same eligibility conditions is 16,782, based on 562 such impoundments reported in the long survey.

[2] Major groups 20-39 and 97 plus codes 4212, 4213, 4231, 4581, 4952 (except Publicly Owned Treatment Works), 4953, 4959, 5085, 5093, 5169, 5171, and 7699 (transportation equipment cleaners only).

US EPA ARCHIVE DOCUMENT

The second population of particular interest consists of those facilities in the first population that used at least one eligible surface impoundment to manage at least one of the target chemicals identified in the long survey or had an extreme pH in an eligible impoundment . The specific characteristics of this population of facilities are as follows: facilities in the first population whose direct- and zero-discharge impoundments were used only to manage non-hazardous waste for which *any* of the following were true:

- 30-day average pH was less than 3
- 30-day average pH was greater than 11
- at least one target chemical was managed in the surface impoundment.

The estimated number of such facilities is 4,457, based on 157 such facilities that responded to the long survey. The estimated number of surface impoundments at these facilities that meet these same eligibility conditions is 11,863, based on 531 such impoundments reported in the long survey.

A.5.2 Estimation Procedures

This section discusses the statistical analysis procedures used to compute point estimates of population totals, means, and proportions for the populations of facilities and impoundments discussed above. In addition, it describes how standard errors were computed for these population estimates, how estimates with poor precision were identified, and how confidence interval estimates can be generated. All the standard errors were produced using RTI's SUDAAN software for analysis of data from complex sample surveys (Shah et al, 1997).

A.5.2.1 Point Estimates. If Y_i denotes a measured quantity for the *i*-th facility or impoundment (e.g., number of eligible impoundments or presence of a liner), then the population total for characteristic Y_i was estimated as

$$\hat{Y} = \Sigma\, w_i Y_i \quad ,$$

where w_i denotes the statistical analysis weight and Σ denotes summation over either all facilities in the sample or over all impoundments at these facilities (depending on whether the outcome, Y_i, is a facility-level or impoundment-level outcome). In the same manner, the population mean for characteristic Y_i was a ratio estimate computed as follows:

$$\bar{Y} = \Sigma\, w_i Y_i / \Sigma\, w_i \quad ,$$

Likewise, population proportions were ratio estimates, computed as follows:

$$\hat{P}_x = \Sigma\, w_i X_i / \Sigma w_i \quad ,$$

where X_i =1 for those facilities or impoundments with the characteristic of interest (e.g., ever managed RCRA characteristic hazardous waste) and X_i = 0 otherwise.

US EPA ARCHIVE DOCUMENT

In addition, estimates of population totals, means, and proportions were generated for various subpopulations, or analysis domains (e.g., states or SIC codes). In these cases, the estimators of the population totals, means, and proportions were generated by substituting the product $d_i w_i$ for w_i in the above formulas, where $d_i = 1$ if the facility or impoundment is a member of the analysis domain and $d_i = 0$ otherwise.

A.5.2.2 *Standard Errors*. The standard error of an estimate is a common statistical measure of its precision. It is the standard deviation of the sampling distribution of the estimate or, alternatively, is the square root of the variance of the estimate. That is, if one were to replicate the sample selection and data collection procedures many times in exactly the same way and with exactly the same population, the standard error of the estimate is the standard deviation of the values of that estimate that would be generated by those samples.

Estimates of variances and standard errors of survey statistics were computed using RTI's SUDAAN software. For nonlinear survey statistics, such as estimated means and proportions, SUDAAN uses the classical first-order Taylor Series linearization method (Wölter, 1985).

Because the number of facilities and impoundments in the target population is much greater than the number included in the sample, calculation of standard errors was simplified by treating the initial sample of facilities selected for the screener survey as having been selected with replacement. Hence, computation of standard errors only required identifying the analysis strata and primary sampling units (PSUs) used at the first stage of sample selection. Because facilities were selected directly in the initial sample for the screener survey, they are the PSUs. Because each analysis stratum must contain at least two responding facilities in order to calculate standard errors, some of the sampling strata shown in Table A-10 were collapsed to form analysis strata as shown in Table A-14. In addition, when the number of facilities with complete data for the set of items entering a particular analysis was low, adjacent analysis strata were sometimes collapsed, but strata representing direct discharge facilities were never collapsed with strata representing zero discharge facilities.

The procedures used by SUDAAN to estimate variances and standard errors can best be explained by introducing some mathematical notation to represent the statistical analysis strata, facilities, impoundments, and observations. Hence, let

$h = 1, 2, ..., 15$ denote the 15 statistical analysis strata shown in Table A-14
$i = 1, 2, ..., n_h$ denote the sample facilities in stratum h and
$j = 1, 2, ..., m_{hi}$ denote the impoundments at the i-th facility in stratum h.

If Y represents an impoundment-level characteristic (e.g., concentration of a target analyte), then let

Y_{hij} = the value of the outcome, Y, for the j-th impoundment at the i-th facility in stratum h

and

$$Y_{hi} = \sum_{j=1}^{m_{hi}} Y_{hij} \quad .$$

US EPA ARCHIVE DOCUMENT

However, if *Y* represents a facility-level characteristic (e.g., number of years of operation), then let

Y_{hi} = the value of the outcome, *Y*, for the *i*-th facility in stratum *h*.

Using this notation, whether *Y* represents an impoundment-level characteristic or a facility-level characteristic, Y_{hi} is a facility-level outcome, and it is helpful to further let

$$Z_{hi} = w_{hi} Y_{hi} \quad .$$

Then, the estimated population total for characteristic *Y* can be represented as

$$\hat{Y} = \sum_{h=1}^{15} \sum_{i=1}^{n_h} Z_{hi} \quad .$$

Sampling variances for estimated totals were then estimated as

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{15} n_h S_h^2 \quad ,$$

where

$$S_h^2 = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (Z_{hi} - \bar{Z}_h)^2 \quad ,$$

and

$$\bar{Z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi} \quad .$$

In order to illustrate how SUDAAN computed sampling variances for estimates means and totals, it is helpful to represent these ratio estimates as

$$\hat{R} = \frac{\sum_{h=1}^{15} \sum_{i=1}^{n_h} w_{hi} Y_{hi}}{\sum_{h=1}^{15} \sum_{i=1}^{n_h} w_{hi}} \quad .$$

The estimated variance of the ratio was then based on the following "linearized value:"

$$Z_{hi}^* = \frac{w_{hi} (Y_{hi} - \hat{R})}{\sum_{h=1}^{15} \sum_{i=1}^{n_h} w_{hi}} \quad .$$

The variance of the estimated mean or proportion was then computed as the estimated variance for the population total of the linearized values, Z_{hi}^*, i.e,

$$\hat{V}(\hat{R}) = \hat{V}(Z_{hi}^*) \quad ,$$

where the latter variance is computed using the formula presented above for estimating the variance of a population total.

US EPA ARCHIVE DOCUMENT

The standard errors computed in this manner account only for the uncertainty resulting from random errors, primarily those due to making inferences from a sample, rather than from a census of all facilities in the population. They do not account for potential sources of systematic error (or bias), such as the incomplete nature of the sampling frame for zero dischargers (see Table A-10), response errors, data entry errors, etc.

When a cell sample size (i.e., the number of observations upon which a total, mean, or the denominator of a proportion is based) is small (e.g., less than 30), the standard error calculated by SUDAAN often is underestimated. In that case, the survey design effect, which typically exceeds one (1), may be estimated to be less than one, suggesting that the survey achieved greater precision than a simple random sample. Hence, if $\hat{\theta}$ represents an estimated total, mean, or proportion and $SE(\hat{\theta})$ represents its standard error calculated by SUDAAN, the standard error

used for that estimate was calculated as

$$se(\hat{\theta}) = Max\left[SE(\hat{\theta}), \frac{SE(\hat{\theta})}{\sqrt{DEFF(\hat{\theta})}}\right] \quad \text{when } n < 30$$

$$se(\hat{\theta}) = SE(\hat{\theta}) \quad \text{when } n \geq 30,$$

where DEFF is the Type 1 survey design effect calculated by SUDAAN and n is the cell sample size. Hence, the standard error calculated by SUDAAN was inflated to compensate for underestimation when the cell sample size was small (<30) and the survey design effect was less than one (1).

Estimates with Poor Reliability

When cell sample sizes are small, weighted population estimates may not be reliable, and their standard errors may not be accurately estimated. Therefore, estimated totals and means are flagged in the report as being unreliable when the relative standard error (RSE) of the estimate is 50 percent or more. That is, if $\hat{\theta}$ represents an estimated total or mean, then that estimate is flagged as unreliable if

$$\frac{se(\hat{\theta})}{\hat{\theta}} > 0.50 \quad .$$

RSEs do not work as well as measures of precision for estimated proportions because an estimate, $\hat{P}$, and its complement, $(1 - \hat{P})$, have the same variance but quite different RSEs. Therefore, the statistic used to flag estimates of proportions as unreliable is the RSE of the natural logarithm of $\hat{P}$. In particular, the estimate, $\hat{P}$, of a population proportion is flagged as unreliable if

$$\frac{se(\hat{P}) / \hat{P}}{-\ln(\hat{P})} > 0.275 \quad \text{when } \hat{P} < 0.50$$

or

$$\frac{se(\hat{P}) / (1 - \hat{P})}{-\ln(1 - \hat{P})} > 0.275 \quad \text{when } \hat{P} \geq 0.50 \quad .$$

US EPA ARCHIVE DOCUMENT

The upper bound, 0.275, is an *ad hoc* bound that has been found to produce reasonable results.

Table A-14. Analysis Strata Used for Variance Estimation

Analysis Stratum	Number of Long Survey Respondents
Direct Discharge Facilities	
■ SIC 26	27
■ SIC 28	29
■ SIC 29	21
■ SIC 30	6
■ SIC 33	20
■ SIC 34-39	6
■ SIC 20-23	6
■ SIC 24-27	6
■ SIC 28-31	9
■ SIC 32	6
■ SIC 49-76[a]	7
Pilot study facilities (certainty selections)	6
Direct discharger subtotal	149
Zero Discharge Facilities	
In TRI or AFS	5
High priority SICs[b]	6
All other facilities[b]	24
Zero discharger subtotal	35

TRI = EPA's Toxic Release Inventory.
AFS = EPA's Aerometric Information Retrieval System, Facility Subsystem.
[a]SICs 4952 (excluding Publicly Owned Treatment Works), 4953, and 4959, 5085, 5093, 5169, 5171, and 7699 (transportation equipment cleaners only)
[b]Not in TRI or AFS.

A.5.2.3 <u>*Confidence Intervals*</u>. The reported standard errors also can be used to compute confidence interval estimates of population totals, means, and proportions. If $\hat{\theta}$ represents an estimated total, mean, or proportion, an approximate 100(1-α) percent confidence interval estimate of that parameter can be calculated as

$$\hat{\theta} \pm t_{df,1-\alpha/2}\, SE(\hat{\theta}) \quad ,$$

US EPA ARCHIVE DOCUMENT

where t is the 100(1-α/2) percentile of the Student's t distribution with df degrees of freedom and $SE(\hat{\theta})$ is the standard error of the estimate. The appropriate degrees of freedom is

$$df = \sum_{h=1}^{H} (r_h - 1) \quad ,$$

where h represents the analysis strata (see Table A-14) and r_h is the number of responding facilities in analysis stratum h.

These confidence intervals are valid so long as the number of facilities contributing to the estimated total, mean, or proportion is large enough that the sampling distribution of the sample total, mean, or proportion is approximately a Student's t distribution.

Because of the relatively large number of facilities in the sample for the surface impoundment study, the resulting degrees of freedom usually are greater than 30, and the appropriate value to use from the Student's t distribution is actually the 100(1-α/2) percentile of the standard normal distribution. In that case, the approximate 95 percent confidence interval estimate of a population parameter (total, mean, or proportion) becomes

$$\hat{\theta} \pm 1.96\, SE(\hat{\theta}) \quad .$$

Confidence interval estimates are reported only when the cell sample size is sufficiently large to support a reasonably precise estimate. Therefore, confidence interval estimates are not reported for those estimates that are flagged as unreliable based on the criteria discussed above.

A.6 References

Aller, L., T. Bennett, J.H. Lehr, R.J. Petty, and G. Hackett. 1987. *DRASTIC: A Standard System for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings*. EPA-600/2-87-035. Robert S. Kerr Environmental Research Laboratory, Ada, OK.

Brick, J.M., and G. Kalton. 1996. Handling missing data in survey research. *Statistical Methods in Medical Research* 5:215-238.

Hansch, C., A. Leo, and D. Hoekman. 1995. *Exploring QSAR: Hydrophobic. Electronic, and Steric Constants*. American Chemical Society, Washington, DC.

Hunt, C. B., 1974. Natural Regions of the United States and Canada. W. H. Freeman and Company, San Francisco, California.

Kalton, G. and D.S. Maligalig. 1991. "A Comparison of Methods of Weighting Adjustment for Nonresponse." *Bureau of the Census 1991 Annual Research Conference Proceedings,* pp. 105-110.

Kollig, H.P., J.J. Ellington, S.W. Karickhoff, B.E. Kitchens, J.M. Long, E.J. Weber, and N.L. Wolf. 1993. *Environmental Fate Constants for Organic Chemicals Under Consideration*

US EPA ARCHIVE DOCUMENT

for EPA's Hazardous Waste Identification Projects. U.S. Environmental Protection Agency. Office of Research and Development. Athens, GA.

Lessler, J.T. and W.D. Kalsbeek (1992). *Nonsampling Error in Surveys*. New York, NY: Wiley.

Shah, B.V., Barnwell, B.G., Bieler, G.S. (1997) SUDAAN User's Manual. Research Triangle Institute, RTP, NC.

U.S. EPA (Environmental Protection Agency). 1999b. *Screening Survey for Land DisposalRestrictions Surface Impoundment Study*. Washington, DC. February.

U.S. EPA (Environmental Protection Agency). 1999d. *Survey of Surface Impoundments*. Washington, DC. November.

U.S. EPA (Environmental Protection Agency). 1999a. Chemical Data Base for HWIR99. Office of Research and Deveopment. Athens, GA.

U.S. EPA (Environmental Protection Agency). 1999c. *Surface Water, Soil, and Waste Partition Coefficients for Metals*. National Exposure Research Laboratory. Athens, GA. June 22.

U.S. Geological Survey (USGS), 1984. Geologic Map of the United States. USGS Branch of Distribution at Federal Center, Denver, CO.

U.S. EPA (Environmental Protection Agency). 1996b. *Superfund Chemical Data Matrix*. Office of Emergency and Remedial Response, Washington, DC. http://www.epa.gov/oerrpage/superfnd/web/resources/scdm/index.htm

U.S. EPA (Environmental Protection Agency). 1996a. *Better Assessment Science Integrating Point and Nonpoint Sources* (BASINS), Version 1. EPA-823-R-96-001. Office of Water, Washington, DC.

USGS (United States Geological Survey). Updated daily. United States National Water Information System (NWIS) water data retrieval internet site accessed on January 31, 2001. http://waterdata.usgs.gov/nwis-w/us/.

van der Leeden, F., Troise, F. L., Todd, D. K., 1990. *The Water Encyclopedia - Second Edition*. Chelsea, Michigan; Lewis Publishers, Inc.

Wölter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

US EPA ARCHIVE DOCUMENT