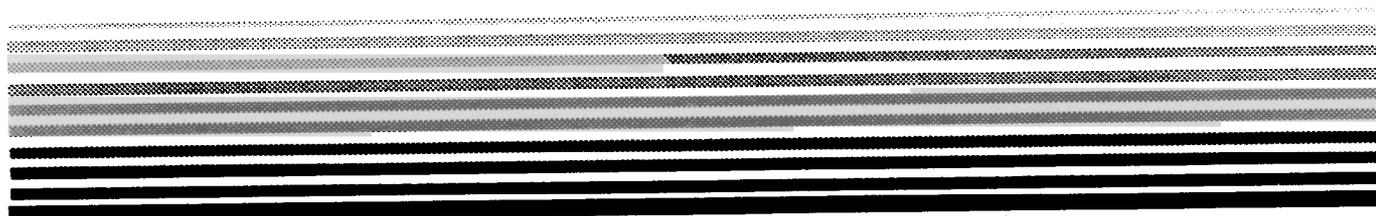
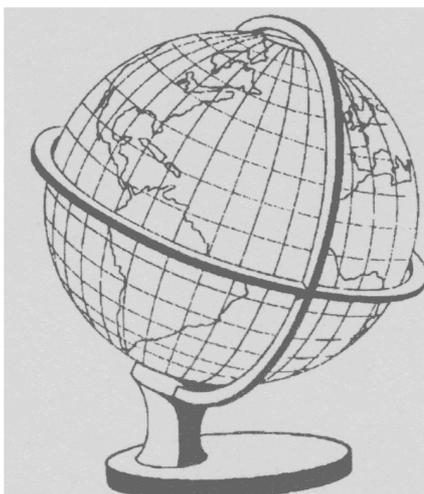




Locational Data Policy Implementation Guidance

Guide To Selecting Latitude/ Longitude Collection Methods



Contents

PREFACE	i-1
EXECUTIVE SUMMARY	ii-1
Chapter 1 BACKGROUND	1-1
1.1 Geocoding Study Methodology	1-1
Chapter 2 AUDIENCE AND PURPOSE	2-1
Chapter 3 BASIC TERMS, CONCEPTS AND DEFINITIONS	3-1
Chapter 4 IMPORTANT THEMES: PROGRAMMATIC DIVERSITY AND TWO ORGANIZATIONAL APPROACHES TO GEOCODING	4-1
Chapter 5 THE GEOCODING LIFECYCLE	5-1
5.1 Incremental field-Based Geocoding	5-1
5.2 Centralized Geocoding	5-2
5.3 The Geocoding Lifecycle	5-3
Chapter 6 GEOCODING METHODS, CAPABILITIES, REALISTIC COSTS AND ACCURACIES	6-1
Chapter 7 CROSS CUTTING IMPLEMENTATION ISSUES	7-1
7.1 The Need For A Geographic Reference Standard	7-1
7.2 Urban vs. Rural Geocoding: Productivity, Accuracy, And Cost Differences	7-2
7.3 Performing In-House Geocoding Or Contracting For Specialized Geocoding Services	7-3
7.4 EPA's Regulated Universe; Who Carries The Geocoding Burden ?	7-3
7.5 Accuracy Checking of Locational Data	7-5
7.6 Secondary Data Users, Multimedia Data Integration, Public Data Access, and Enforcement Data Requirements	7-8

Contents (continued)

Chapter 8	A FRAMEWORK FOR ESTIMATING GEOCODING.....8-1	
	LIFECYCLE COSTS	
8.1	<i>Characterize Existing Records.....8-1</i>	8-1
8.2	<i>Define Geocoding Requirements.....8-2</i>	8-2
8.3	<i>Select/Define Geocoding Methodology.....8-2</i>	8-2
8.4	<i>Estimate Geocoding Lifecycle Fixed and Variable Costs.....8-2</i>	8-2
Chapter 9	SUMMARY.....9-1	9-1
Appendix A	GEOCODING METHODS FACT SHEETS.....A-1	A-1
Appendix B	REFERENCESB-1.	B-1.

List of Figures and Tables

Figure 1	The Geocoding Lifecycle	5-3
Figure 2	Comparative Geocoding Costs And Accuracies	6-1
Figure 3	Bounding Boxes.....	7-7
Figure 4	Generic Geocoding Cost Estimation Model	8-3
Figure 5	Spreadsheet Version of Geocoding Cost Estimation Model	8-6
Table 1	EPA Data System's Locational Data Inventory Summary	7-5

Preface

This version of the Guide to Selecting Latitude/Longitude Collection Methods (the “Geocoding Study”) reflects the many comments and suggestions received in response to a prior draft version dated January 29, 1992. The major change to this version is the incorporation of review comments from EPA’s Office of Information Resources Management (OIRM) and Office of Pollution Prevention, Pesticides, and Toxic Substances (OPPPTS). Jeff Booth was the EPA project manager for this document and is part of the Program Systems Division of OIRM.

Support for the development of this document was provided by American Management Systems, Inc. under EPA contract #68-W9-0039, Delivery Order #025, and Booz, Allen & Hamilton, under EPA contract #68-W9-0037, Delivery Order #094.

Executive Summary

In May, 1990 EPA's Office of Administration and Resources Management (OARM) added a new chapter to the agency's Information Resources Management (IRM) Policy Manual. This chapter established the principles and methods underlying the collection and documentation of locational data. Following the addition of, and in response to that chapter, EPA initiated a geocoding study to analyze the realistic capabilities of various geocoding technologies and methods. "Geocoding" is the general term applied to different procedures, techniques, and technologies that are used to identify, quantify, and document locational coordinates.

The geocoding study was based upon literature reviews and vendor solicitations that were performed to establish a complete array of applicable geocoding methodologies and their associated advantages and disadvantages. Five major geocoding methodologies were chosen for review: Photo Interpretation, Map Interpolation, Global Positioning Systems (GPS), Address Matching, and ZIP Code Centroids. *The agency encourages the use of GPS technology specifically because the cost over time is expected to decrease while the ability to bolster skills and accuracy will increase.*

This guide describes each of these methods according to description, accuracy, cost, benefits and limitations. It also provides technical information about geocoding to organizations responsible for implementing EPA's Locational Data Policy (LDP). Both managers and technical personnel will find this information useful in selecting cost-effective geocoding methods.

In order to understand specific geocoding technologies, managers and technical specialists must understand that environmental programs are responsible for a diverse set of regulated entities and that the locational data collected for these entities has different organizational uses in response to regulatory and legislative requirements. Therefore, environmental organizations and government agencies have different locational data accuracy and quality needs which may influence their choice of geocoding technology(ies) or approach to implementing that technology.

The objective of this guide is to:

- Inform EPA program offices, States, and other parties affected by the Locational Data Policy, about the geocoding life cycle (i.e., the steps required to produce locational data independent of specific geocoding technologies or methods).
- Describe the *practically achievable capabilities*, realistic costs and accuracies, of different geocoding methods/technologies.
- Identify cross cutting geocoding issues.

- **Provide a framework for designing and estimating the cost of a geocoding projector program consistent with EPA IRM policy.**
- **Review some locational data accuracy checking methods.**

In addition, the guide provides basic terms, concepts and definitions regarding geocoding.

Geocoding has its own well-defined life cycle, a generic process independent of specific technologies or methods, that provides a realistic framework for estimating the true cost of implementing one or more of the available technologies. It also provides a fair basis of comparing geocoding methods. The way in which the life cycle will be employed, however, depends on the type of approach a manager or technical specialist is using. There are two fundamental organizational approaches to geocoding: 1) Incremental Field-based Geocoding, and 2) Centralized Geocoding. These approaches affect the way data is collected as well as the cost and time required to collect locational data.

In comparing different geocoding methods, managers and technical personnel should be aware of the differences between urban and rural geocoding, and understand issues involving the ability to perform in-house geocoding versus contracting specialized geocoding services.

It is important not only to understand the specifics of different geocoding methods, but also to realize the lack of locational data contained in existing databases covering regulated entities. An OIRM-sponsored regulatory review of EPA spatial data requirements concluded that States provide the majority of locational data, followed by the regulated community, EPA headquarters, and Regional offices. The Locational Data Policy encourages the integration of data based on location, thereby promoting use of EPA's extensive data resources for cross-media environmental analyses and management decisions. The cross-media benefit that the policy encourages will allow EPA program offices to share data collection responsibilities.

Because different geocoding methods can vary widely in applicability, accuracy, and cost, managers and technical specialists may need to characterize existing records, define geocoding requirements, and select/define geocoding methodology in order to estimate geocoding life cycle costs. Ultimately, estimating these costs depends upon defining and estimating fixed and variable costs.

Chapter 1
BACKGROUND

1. BACKGROUND

On May 17, 1990, after formal agency-wide review, the Locational Data Policy (LDP) became effective as an official directive under the 2100 series and Chapter 13 of the EPA Information Resources Management (IRM) Policy Manual. Its stated purpose is to establish "... the principles for collecting and documenting latitude/longitude coordinates for facilities, sites and monitoring and observation points regulated or tracked under Federal environmental programs within the jurisdiction of the Environmental Protection agency (EPA)."

Pursuant to this policy initiative, a Locational Accuracy Task Force (LATF) was formed in June of 1990 to develop a minimum locational data accuracy goal for EPA and to make recommendations to the EPA IRM Steering Committee about techniques for collecting locational data (i.e., geocoding). "Geocoding" is the general term applied to different procedures, techniques and technologies that are used to identify, quantify, and document locational coordinates. The findings of the LATF are incorporated into EPA's Locational Data Policy Implementation Guidance.

In the same time period, EPA initiated a study to analyze the realistic capabilities and costs of alternative geocoding technologies and methods. This document reports on the findings of that geocoding study by providing objective information and data about currently available geocoding technologies. This document is designed to assist organizations better understand technical, cost, and organizational issues that will directly impact their ability to comply with the LDP.

Two "themes" should be kept in mind while reading this document:

- Programs within the agency are responsible for regulating a diverse set of entities and, therefore, have a diverse set of locational accuracy needs.
- There are two fundamental approaches or combinations of approaches for obtaining locational coordinates: incremental and centralized. Programs must understand their own organizational requirements before deciding which approach to take.

These themes are presented in detail in Chapter 4.

1.1 Geocoding Study Methodology

The geocoding study is the basis for most of the recommendations in this report. It was conducted in six basic steps:

- Literature review.
- Study site selection.
- Existing EPA locational data in EPA data compilation.

- **Geocoding database creation.**
- **Geocoding test data analysis.**
- **Error/accuracy checking methods assessment.**

These steps are described in detail below.

Literature reviews and vendor solicitations were performed to establish the universe of applicable geocoding methodologies and their inherent advantages and disadvantages. Five major geocoding methodologies were chosen for review: Photo Interpretation, Map Interpolation, Global Positioning Systems (GPS), Address Matching and ZIP Code Centroids. In addition, a description of photogrammetry has been added to this report. The reviews and solicitations were integrated and are contained in the *Geocoding Methods, Capabilities, Realistic Costs and Accuracies* section of this report. Loran-C was excluded from this study because the system is:

- **Not accurate enough to reliably provide locational data within the agency's 25 meter goal.**
- **Not widely used within the agency.**
- **Being supplanted by other technologies.**

To identify study sites with previously collected and well-documented accurate locational data, EPA Geographic Information System (GIS) Teams were solicited for candidate study sites. As a result of this solicitation, three study sites were chosen: Chattanooga, TN, Nashua River Basin, CT, and San Gabriel, CA. The Chattanooga study area was selected because over 1000 business entities in the study area already had airphoto-interpreted locational data coordinates. Nashua and San Gabriel were selected because both were slated for Global Positioning System (GPS) surveys within the time-frame of this project.

Data from various EPA systems were collected and integrated. Appropriate subsets were made from EPA data systems, including AIRS, Biennial Report, CERCLIS, FINDS, IFD, PCS, RCRIS, TRIS, and, in the case of Chattanooga, an existing integrated Chattanooga database created in 1985. These data were chosen because they are most characteristic of regulated entities in EPA's data holdings. Since many of the data systems contain different facility identifiers, matching facility records across the data bases was a semi-automated process. Matches were automatically made when possible, based on ID numbers. Once all possible automated matches were processed, further manual matching was accomplished based on the name and address of each remaining record.

To supplement the existing locational data in each study site, prioritized lists of regulated entities in each area were prepared and sent to EPA GPS survey teams. These teams collected GPS data for Chattanooga, Nashua and San Gabriel. In addition to GPS data, address-matched locational data were solicited from three different vendors for all facilities with accurate address records in the study sites.

The accuracy of each methodology was assessed by assuming one set of coordinates was “truth” and then comparing all other coordinate pairs to that set. From literature reviews and professional expertise, GPS was assumed to be the most accurate geocoding methodology studied because it represents the highest order of accuracy technology available. The second most accurate method was assumed to be Photo Interpretation/Map Interpolation based upon the scale and resolution of images used in the study. Locational discrepancies under 50 meters could not be resolved due to differences in measuring place. Based on these assumptions, GPS coordinates were used as “truth” if they were available for an entity. Otherwise, the photo-interpreted coordinates were used as “truth.”

A cost estimation model was developed to compare the fixed (i.e., equipment) and variable (i.e., labor) costs associated with each geocoding method. For a centralized geocoding process, a screening process was described that classified entities by the expected difficulty to geocode them. Based on the classification, the most cost-effective geocoding method could be applied to each class of entities.

Accuracy assessment methods were reviewed and tested using existing EPA locational data. Methods investigated included distance from ZIP Code centroid, point-in-polygon tests, and bounding box tests. All these tests essentially predict the “reasonability” of locational data coordinates, although the accuracy of these measures is highly variable.

A questionnaire was distributed to EPA data system managers to collect information about the status of locational data in EPA data systems. Responses from 23 data systems, representing 2.9 million data records were summarized (Table 1, page 7-6) to characterize the nature and sources of locational data in EPA data systems.

Chapter 2
AUDIENCE AND PURPOSE

2. AUDIENCE AND PURPOSE

This Guide to Selecting Latitude/Longitude Collection Methods is written for those organizations which are responsible for implementing EPA's Locational Data Policy, including:

- **EPA National System Administrators.**
- **EPA Program Management Branch Chiefs.**
- **EPA Regions and States.**
- **EPA contractors.**
- **The regulated community.**

The purpose of this Guide is to provide the reader with technical information about geocoding technology. This information will be presented in terms that are understandable both to management and technical personnel in order to assist in selecting cost-effective geocoding methods. There are five specific objectives:

- **To inform EPA program offices, States, and other parties affected by the LDP, about the geocoding life cycle (i.e., the process required to produce locational data independent of specific geocoding technologies or methods).**
- **To describe the *practically achievable capabilities*, realistic costs, and accuracies of different geocoding methods/technologies.**
- **To identify cross-cutting geocoding issues.**
- **To provide a framework for designing and estimating the cost of a geocoding project or program consistent with EPA IRM policy.**
- **To review selected locational data accuracy checking methods.**

Chapter 3
BASIC TERMS, CONCEPTS AND DEFINITIONS

3. BASIC TERMS, CONCEPTS AND DEFINITIONS

Geocoding, similar to other “high technologies,” has its own unique language. There are many ill-defined terms and phrases which can be confusing or misleading. Below, definitions of key terms and concepts commonly used throughout this document are provided to limit some of the potential confusion resulting from indiscriminate use of terminology:

- ***Geocoding*** -- The application of procedures, techniques, and technologies for the purpose of identifying, quantifying, and documenting geographic location or boundaries of a physical entity (e.g., facility, outfall pipe, Superfund site).
- ***Latitude/Longitude*** -- Latitude and longitude refers to the global reference system used to locate objects on the surface of the earth. Positions are referenced by the number of degrees north or south of the equator (latitude) and east or west of the prime meridian (longitude). Other commonly used ‘reference systems include Universal Trans-Mercator (UTM) and State Plane. Software exists to aid conversion of existing data to latitude/longitude coordinates.
- ***Global Positioning Systems (GPS)*** -- Global Positioning Systems are systems which derive location based on ground position relative to earth-orbiting satellites.
- ***Selective Availability*** -- Selective availability is the intentional degradation of the performance capabilities of satellite systems, such as GPS, for civilian users by the U.S. military, accomplished by artificially creating a significant clock error in the satellites.
- ***Address Matching*** -- Address matching is a semi-automated process for deriving latitude/longitude coordinates from street addresses.
- ***Photogrammetry*** -- Photogrammetry is a process for deriving reliable measurements by locating and measuring the position of an object on aerial photographs.
- ***Map Interpolation/Photo Interpretation*** -- Photo Interpretation/Map Interpolation is an integrated technique by which the user transfers information from an aerial photograph on to a map base and then extracts coordinate information via standard manual or digital map interpolation techniques.

- ***Land Surveying*** -- Land Surveying is a field-based process for determining location from direct measurements from a known baseline (typically geodetic survey monuments).
- ***Geocoding Accuracy*** -- Geocoding accuracy is a measure of the degree to which a geocoding process yields the true location of an object. If a process has a 25 meter accuracy, then the true location of the object must be within 25 meters of the reported location. For EPA, accuracy with a 95% level of confidence is espoused for locational data.
- ***Geocoding Precision*** -- Geocoding precision is a measure of the degree to which repeated measurements of an object's location yield the same result.
- ***Secondary Use*** -- The re-use of existing data in an application other than that for which it was primarily collected. Examples of typical secondary use applications include comparative risk studies and enforcement support.

Chapter 4

**IMPORTANT THEMES: PROGRAMMATIC DIVERSITY AND
TWO ORGANIZATIONAL APPROACHES TO
GEOCODING**

4. IMPORTANT THEMES: PROGRAMMATIC DIVERSITY AND TWO ORGANIZATIONAL APPROACHES TO GEOCODING

Two important themes need to be understood before considering specific geocoding technologies. First, environmental programs are responsible for a diverse set of regulated entities (e.g., facilities, air stacks, outfall pipes, monitoring wells, underground storage tanks, landfills, etc.). Organizational uses for locational data associated with these entities differ in response to legislative and regulatory requirements, as well as overall mission (i.e., Federal, state, regulated community, etc.). *As a result, environmental organizations and government agencies have different locational data accuracy and quality needs which may influence their choice of geocoding technology or approach to implementing that technology.* It also should be noted that the LDP guidance recommends that environmental organizations consider secondary uses for “their” locational data when selecting a geocoding method.

The following comparison is provided to illustrate differences in programmatic requirements for locational data. The Toxics Release Inventory System (TRIS) in EPA’s Office of Pollution Prevention, Pesticides, and Toxic Substances (OPPPTS) currently requires a single pair of latitude/longitude coordinates to geographically define the location of a reporting facility, regardless of its size. In comparison, Superfund site managers and environmental engineers require numerous locational coordinates, including depth and elevation data, to characterize a wide variety of natural and man-made phenomena within or around the boundaries of each site.

TRIS and Superfund program managers must adhere to the LDP. TRIS data are used primarily to ensure a *completenational inventory* and to support a variety of national, regional, and state-level planning analyses. Superfund data are used for *site-specific* engineering studies to determine the location and extent of environmental contamination in order to provide a scientific basis for removal or remedial actions.

Although both programs require high-quality locational data, specifications for amount, type (i.e., points, linear, polygonal, 3-dimensional elevation, or depth data), and accuracies/ precision of data differ dramatically. For example, many Superfund site analyses require more accurate locational data than EPA’s 25 meter goal. In contrast, primary use requirements of OPPPTS for TRIS locational data may lack justification for 25 meter accuracy. Other offices and systems have a mix of orientation (facility and non-facility level). For example, locational data in the Permit Compliance System (PCS), supporting EPA/Office of Water’s National Pollutant Discharge Elimination System (NPDES) Program, are of a mix of both facilities and outfall pipes. The extent to which secondary use requirements and data integration efforts substantiate the need for 25 meter accuracy in

latitude/longitude data (when the primary use does not) remains to be redetermined. Twenty-five meter accuracy is an agency-wide goal; however, it should be met wherever possible. Procedures for addressing these issues, planning for compliance, and/or obtaining waivers (if necessary) are presented more fully in the Locational Data Policy Implementation Guidance -- Guide to the Policy.¹

The second theme that needs to be understood is that there are two fundamental organizational approaches to geocoding: incremental field-based geocoding, and centralized geocoding. Incremental field-based geocoding is defined as the collection of locational data by EPA officials or their representatives during regularly scheduled field work (e.g., site inspections, compliance monitoring, site characterization, remedial or removal actions, etc.). In contrast, centralized geocoding is defined as a dedicated locational data collection effort used to populate a data base in one, unified effort. This *Guide* focuses primarily on centralized geocoding. During its deliberations, the LATF stated its preference for the incremental approach, although both have strengths and weaknesses. Chapter 6 presents more information about these approaches. This document highlights the impact of these geocoding approaches on various technical and management issues.

¹ Please note that the study which underlies much of the information provided in this document focused almost exclusively upon “point” data, or at least the notion of representing regulated entities, regardless of size or geographic extent, with one latitude/longitude coordinate pair. Therefore, those organizations seeking guidance on developing geographic base files containing polygonal data, also subject to the LDP, may have requirements that fall somewhat outside the scope of this Guide.

Chapter 5
THE GEOCODING LIFECYCLE

5. THE GEOCODING LIFE CYCLE

Geocoding has its own well-defined life cycle, a generic process independent of specific technologies or methods. The geocoding life cycle provides a realistic approach for estimating the true cost of implementing any geocoding technology. It also provides a fair basis of comparing geocoding methods. Before describing the six phases of the geocoding life cycle, however, it is important to understand the two organizational approaches to geocoding: incremental field-based and centralized geocoding. This document is more easily focused on centralized approaches although an incremental approach has its own benefits and limitations. A description of each is given below.

5.1 Incremental Field-Based Geocoding

Incremental field-based geocoding is a process of collecting locational data during planned field work. EPA personnel, state partners, or other delegated representatives already perform field work for various purposes such as facility inspections, compliance monitoring, and enforcement actions. Incremental field-based geocoding becomes an additional function or responsibility of field personnel and takes advantage of the opportunity presented by their already-established presence in the field. This approach to geocoding is incremental in the sense that field trips to regulated facilities and operating units occur intermittently on a selective basis.

There are two primary benefits of the incremental field-based approach to geocoding:

- The overhead costs associated with geocoding are reduced by limiting the significant administrative and travel costs associated with a more centralized approach (see below).
- Reliance on existing field personnel to perform geocoding produces certain efficiencies, particularly with their enhanced knowledge of the targeted facilities or regulated entities. One field crew can potentially geocode all regulated entities in a given location during a single site visit.

The two primary limitations of incremental field-based geocoding are:

- The time required for completion of a geocoding effort may be extensive because of the sample-based approach to site inspections typically used by EPA program offices.
- Most EPA field inspectors are not trained surveyors or geocoding technologists, a fact that may impact their productivity or the quality of geocoded data.

5.2 Centralized Geocoding

In contrast to incremental geocoding, the sole objective of centralized geocoding is to collect locational data for a specific geographic region and/or class of entities. Because it generally is not performed in conjunction with any other activities, centralized geocoding requires an independent organizational infrastructure and management process apart from existing programs. The purpose of centralized geocoding is to collect all of the targeted locational data in a concentrated effort.

The ongoing effort of the National Pollutant Discharge Elimination System (NPDES) program in EPA's Office of Water to geocode the location of all NPDES major industrial facility dischargers and their outfall pipes is a prime example of a centralized geocoding approach. This effort involved the mass mailing of several thousand U.S. Geological Survey (USGS) topographic maps (with detailed instructions) to NPDES facilities, the marking of facility and outfall pipe locations on these maps by the facility managers, and the digitizing of these data by EPA. The project has continued for more than a year with the geocoding of more than 7,000 facility and outfall pipes.

The primary benefits of a centralized approach to geocoding are:

- Geocoding projects can be accomplished relatively quickly (as compared to an incremental field-based approach) with limited impact on other programmatic functions.**
- A limited number of personnel needs to be dedicated to a geocoding effort, and provided the necessary technical training. Their productivity and efficiency will improve steadily as they gain experience throughout the geocoding life cycle.**
- A higher degree of management oversight and quality control can be exercised with a centralized geocoding project than with an incremental field-based approach.**
- On larger geocoding projects, the price-per-point drops over time until a "steady state" is reached.**

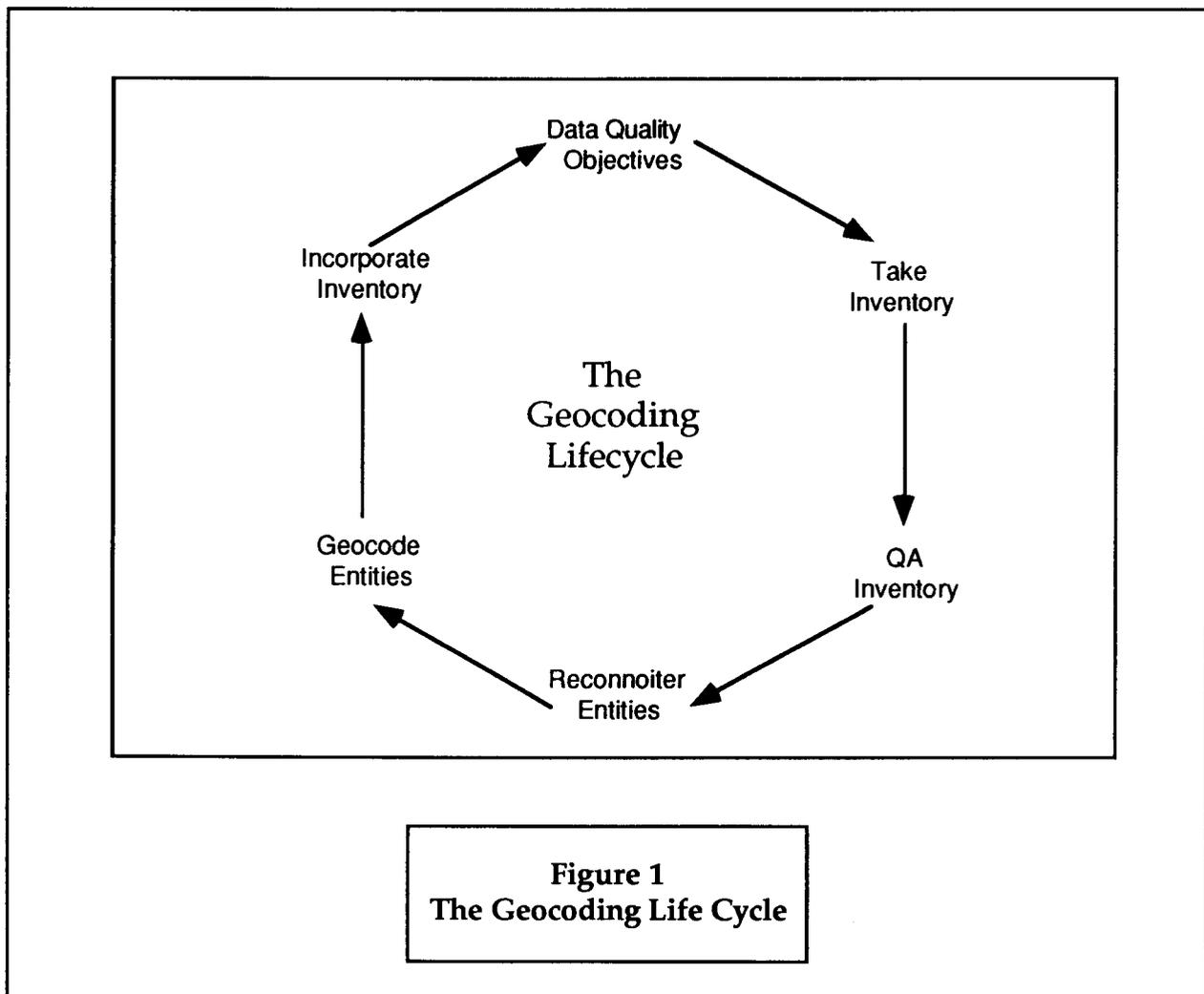
The primary limitations of a centralized approach to geocoding are:

- Centralization usually removes the component of on-site familiarity and thus may adversely impact data quality.**
- Centralized geocoding projects may require a disproportionate amount of resources for management overhead because they are not coupled with other activities that are being performed. This management overhead includes activities such as the creating an organizational**

work flow, and handling production logistics such as monitoring data and maps, quality assurance, and spread sheet design.

5.3 The Geocoding Life Cycle

As indicated in Figure 1, the geocoding life cycle has six major components. Each component “feeds” the subsequent one until the existing inventory is updated with new (or improved) locational data. This process continues until all data are collected or more rigorous data quality objectives (DQOs) are established. Higher expectations for data quality may be a function of improved technology that makes it cost effective to improve the quality of existing locational data. Organizations which initially comply with the 25 meter accuracy goal of the LDP may only be required to follow the geocoding life cycle once. After the locational data are incorporated into the data base, the initial work is complete. Updates and additions should be expected on a routine basis. The six components of the geocoding life cycle are defined below:



- ***Establish DQOs*** -- Determine the level of locational accuracy necessary for the intended primary and secondary use(s) of the data. The LDP has an accuracy goal of 25 meters. Each organization should assess its own programmatic requirements, consulting the full set of LDP Guidance materials, to determine if the agency's 25 meter accuracy goal helps or hinders their efforts to fulfill their mission. It also is crucial for them to give additional consideration to secondary users and data integration efforts. These issues should be addressed in each program's LDP Implementation Plan. A waiver from the LDP for organizations seeking a less rigorous accuracy requirement for their locational data may be possible.
- ***Take Inventory*** -- Determine the entities that need to be geocoded.
- ***Take a QA Inventory*** -- Determine whether existing address records are current and accurate. Physical site addresses are needed to perform address matching instead of mailing addresses.
- ***Reconnoiter Entities*** -- Determine, in a general way, where the target entities are located (e.g., on which maps, driving directions for field survey teams, etc.).
- ***Geocode Entities*** -- Assign latitude/longitude values to targeted entities by applying a specific geocoding technology (e.g., GPS, map interpolation, address matching, etc.).
- ***Incorporate Inventory*** -- Perform data QA/QC, check accuracy of geocoded data, and incorporate data into national or programmatic data systems.

The geocoding life cycle applies both to incremental field-based geocoding and centralized geocoding. The way in which the life cycle is implemented may vary, however. Some of the differences have been discussed above (e.g., time or personnel factors). For either approach, it is incumbent upon planners of geocoding projects to explicitly address how they will implement the geocoding life cycle cost-effectively to achieve their established DQOs.

Chapter 6

**GEOCODING METHODS, CAPABILITIES, REALISTIC
COSTS AND ACCURACIES**

6. GEOCODING METHODS, CAPABILITIES, REALISTIC COSTS AND ACCURACIES

Geocoding technology is changing rapidly. In recent years powerful, new techniques have become commercially available. Global Positioning Systems (GPS) are the most highly visible of these techniques. The LATF recommended GPS as the preferred geocoding technology for the agency, although they agreed that alternative methods, such as map interpolation or address matching, may be appropriate for certain activities (e.g., geocoding FINDS). A number of geocoding methods have been reviewed to estimate their cost and accuracy including: *GPS, Photogrammetry, Map Interpolation, Photo Interpretation/Map Interpolation, Address Matching and ZIP Code Centroid*. Appendix A presents descriptions of each geocoding method, summarizes their reasonably achievable accuracies, provides cost estimates, and highlights benefits and limitations.

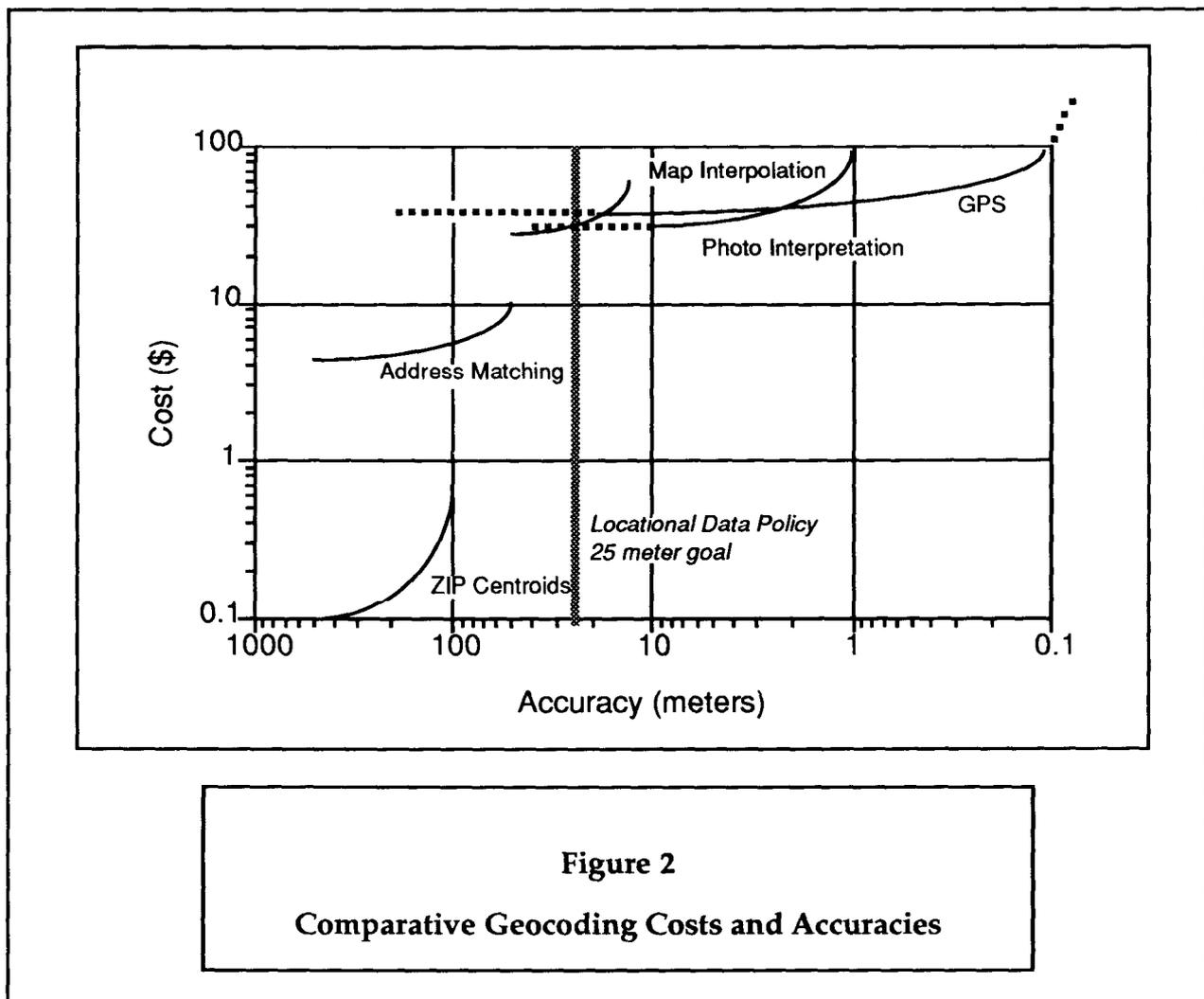


Figure 2, on the previous page, succinctly portrays the two primary evaluation parameters for selecting a geocoding technology: cost and accuracy. As the graph clearly illustrates, only Photo Interpretation and GPS consistently comply with or exceed EPA's 25 meter accuracy goal. Of some interest is their equally high costs per point relative to alternative, less accurate geocoding methods.

Chapter 7

CROSS CUTTING IMPLEMENTATION ISSUES

7. CROSS CUTTING IMPLEMENTATION ISSUES

A number of issues have been identified that apply to all geocoding methods. Each of these issues has a dramatic impact on programmatic geocoding strategies, procedures, and the selection/implementation of specific methods.

7.1 The Need for a Geographic Reference Standard

One of the fundamental problems encountered in documenting locational coordinates during the geocoding study was the lack of a geographic reference standard for EPA facilities or other entities to be geocoded. A geographic reference is the actual position where geocoding occurs. For example, the data analyzed in the geocoding study was derived from either a (theoretical) centroid, street entrance, or approximated block face of a given facility/entity. These differences are attributable to several different factors:

- **Some geocoding technologies have defacto geographic references; address matching, for example, at best approximates the location of an entity somewhere along a city block face.**
- **Entities vary greatly in size and dimension; an outfall pipe may be six inches in diameter while a large industrial plant may cover hundreds or thousands of acres.**
- **Access to regulated entities may be difficult and, therefore, may affect the location of a geographic reference; high fences, challenging terrain, or limits on access to a property may require geocoding personnel to compromise the ideal geocoding position.**

An inherent problem posed by a lack of geographic reference standard is the difficulty of assessing positional accuracy. It also handicaps the ability to subsequently compare locational data that are based on different geographic references. As a practical matter, however, a single geographic reference standard is inadequate. A workable geographic reference standard must recognize the real world limitations (as described above) and provide a hierarchy for implementing and documenting the reference standards employed. In lieu of a national standard, a consistent geographic reference, with well defined options, must be employed for each geocoding project. These references should be described explicitly in the LDP implementation plans prepared by each program.

7.2 Urban vs. Rural Geocoding: Productivity, Accuracy, and Cost Differences

The productivity, cost, and resulting accuracy of geocoding projects depend upon a number of factors including the technology employed, type and training of key personnel, nature of the regulated entities, and the geographic setting (urban versus rural). Experience has demonstrated that geocoding in rural and urban settings presents different problems, impacting productivity, accuracy, and cost, sometimes independent of the technology employed.

Address matching, GPS, and map interpretation are “sensitive” to urban versus rural geography for several reasons:

- **Urban environments are very dense and street address data bases are available that support geocoding accuracies to the nearest city block face or finer; rural environments often lack numerical addressing schemes (e.g., RD#4 vs. 123 Main Street) and street intersections can be “few and far between,” thus diminishing positional accuracy even where address geocoding is technically feasible (which, for many rural areas, it is not).**
- **Urban “canyons,” much like real world canyons, can interfere with radio transmissions and severely limit, or even preclude, GPS-based geocoding.**
- **Reconnoitering can be difficult in any setting; remote rural settings can be particularly difficult to locate thereby reducing overall productivity and raising costs.**
- **National “coverage” maps such as the USGS 1:24,000-scale topographic maps tend to be more current in urbanized areas. The presence of greater numbers of recognizable landmarks represented on these maps contribute to higher-order accuracies in map interpolation.**

These issues are only some examples of the differences encountered in urban versus rural geocoding. Lessons learned from urban and rural geocoding projects should be incorporated into future geocoding project plans. In some cases, different geocoding methods may be preferable in rural areas than those for urban areas. For example, where address matching may satisfy a geocoding requirement in an urban setting, GPS or photogrammetry may be preferable in rural settings. Using alternative geocoding technologies optimized for unique geographic settings may be more productive and may yield more accurate results.

7.3 Performing In-House Geocoding or Contracting for Specialized Geocoding Services

In recent years, an industry has developed to support geocoding requirements. Manufacturers of geocoding technology (e.g., GPS, analytical stereoplotters, GIS-based address matching) are selling state-of-the-art tools that enable performance of a full range of geocoding functions. In addition, consultants and specialized geocoding service bureaus have emerged so organizations can perform their own geocoding or contract for geocoding services. Two prerequisites to making a decision to perform geocoding in-house or through a professional contractor are described below.

- **Definition of DQOs -- A DQO statement will help determine the technologies of choice and, indirectly, decide whether service “bureau” or consultant options are preferable.**
- **Choice of an Incremental Field-based or a Centralized Geocoding Approach -- A decision to pursue an incremental approach probably implies that the responsible organization (or its agents, such as states) will conduct the actual geocoding. A centralized methodology opens up the possibility of employing specialized geocoding service providers.**

Other issues to consider are the amount of in-house expertise available to the organization, the impact on other programmatic activities of assigning personnel to geocoding tasks, and the productivity/cost-benefits of contracting for the service. Geocoding technology is quite sophisticated and evolving rapidly. Much of the productivity derives from allowing highly trained and experienced personnel employ the right tools in the most effective ways.

An in-house geocoding plan (as part of a LDP Implementation Plan) should be designed, and full life cycle cost estimates prepared. For comparison purposes, a Request for Information (RFI) can be sent to geocoding consultants and service bureaus to gather comparative cost data for the same work. Based on the results of an internal assessment and the responses to the RFI, a decision can be made on the most cost-effective way to proceed with a geocoding project (i.e., in-house or via contractor).

7.4 EPA’s Regulated Universe: Who Carries the Geocoding Burden?

A survey of EPA data systems was performed to determine the status of locational data in those systems. Responses were received for the following 23 systems:

- 305(b) Waterbody System (WBS).
- Aerometric Information Retrieval System (AIRS).
- Chemical Update System (CUS).
- Chemicals in Commerce Information System (CICIS).
- Comprehensive Assessment Information Rule Data System (CAIR).
- Comprehensive Environmental Response, Compensation, and Liability Information System (CERCLIS).
- Consolidated Docket System.
- Construction Grants GICS (CGGICS).
- Facility INdex System (FINDS).
- Federal Reporting Data System (FRDS).
- Hazardous Waste Data Management System (HWDMS).
- Management Information Tracking System (MITS).
- National Asbestos Registry System (NARS).
- Needs Survey (NEEDS).
- Permit Compliance System (PCS).
- Preliminary Assessment Information Rule (PAIR).
- Resource Conservation Recovery Information.
- Storage and Retrieval of Water Quality Information (STORET).
- Strategic Planning and Management System (SPMS).
- Superfund Enforcement Tracking System (SETS).
- Toxics Release Inventory System (TRIS).
- Underground Injection Control Tracking System (UIC).
- Water Quality Analysis System (WQAS).

Many data systems contained little or no locational data. The available data were most often generated from ZIP code centroids and map interpolation. Few systems have latitude/longitude accuracy standards or perform QA on locational data.

The findings of an OIRM-sponsored regulatory review for EPA spatial data requirements may be summarized as follows:

“Regulatory requirements for spatial data in EPA programs are limited. The requirements range from merely requiring that the location be provided in an unspecified manner to specifically requiring latitude and longitude to the nearest second [UIC inventory requirements in 40 CFR 144.26(b) (2)1. Seventeen regulations were identified that contained requirements for locational information. Of these seventeen, only seven specified latitude or longitude as a requirement. In most programs where latitude and longitude are required, no accuracy requirement was

specified; therefore, the accuracy of the data available from these programs is unknown. Only two of the seven regulations requiring latitude and longitude specified the level of accuracy (NPDES permits in 40 CFR 122.21 and WC inventory requirements noted above).²

Table 1 shows that, for the systems with locational data that responded, 71% of the locational data is supplied to EPA by the States. The next largest provider of locational data is the regulated community (12.7%). EPA Headquarters and regional offices provide only 8.5% of the locational data. FINDS survey results were omitted from the following table because it is a “secondary” source system (inclusion would have caused double-counting).

**Table 1
EPA Data System's Locational Data Inventory Summary**

	Data Entry	Provide Lat/Long	Perform QA	Perform Update	Total number * of records
EPA Central	38,690 3.7%	14,404 3.6%	186,683 25.3%	298,166 37.2%	537,943
EPA Region	341,938 32.7%	19,390 4.9%	224,315 30.4%	280,532 35.0%	866,175
State	657,734 62.9%	282,696 71.4%	306,958 41.6%	118,625 14.8%	1,366,013
Submitter	0 0.0%	50,255 12.7%	0 0.0%	0 0.0%	50,255
Other (Contractor)	7,320 0.7%	28,966 7.3%	19,923 2.7%	104,198 13.0%	160,407
Total number of records*	1,045,682	395,711	737,879	801,521	2,980,793

* The number of data records is not equal to the number of regulated entities. For example, if TRIS, PCS, and AIRS contain data records for a facility, then multiple records are listed for that facility.

he impact of these numbers on EPA and individual program directions for geocoding national data bases should be carefully considered because much of the burden for implementing EPA’s Locational Data Policy may fall to the States via grants and performance agreements.

² Battelle Contract No. 68-03-3534, Work Assignment No. H2-51, Task 1.a, pg. 10).

7.5 Accuracy Checking of Locational Data

There are several methods that can be reemployed to assess the accuracy of geocoded locational data. All of these methods assume that information in addition to lat/long (e.g., correct address) is known about the location in question. The additional information is used to generate a second estimate of the site location. Geocoding accuracy is then checked by comparing lat/longs from the geocoding effort with the estimated site location.

It should be noted that the accuracy checking methods described below are generally inappropriate for higher accuracy geocoding technologies, such as GPS or photogrammetry. The difficulty in checking the accuracy of GPS, for example, is the general lack of availability of higher accuracy "benchmark" data with which to compare the GPS readings. This issue will need further analysis and guidance as the agency programs begin to conduct GPS surveys.

The simplest accuracy assessment method is to compare the location in question to a known location such as a ZIP Code centroid or a Place-name location from the USGS Geographic Names Index System (GNIS). The distance between the two points is compared to a threshold distance. If the distance is less than the threshold, then the location is accepted, otherwise, the location is rejected. The difficulty then becomes setting that threshold distance. Given the wide range in ZIP code sizes, a single threshold is difficult to set. For example, in a nationwide analysis of address-matched TRIS facility locations and ZIP centroids, 98% of the locations were accepted using a 10 km threshold.

A better method of checking locational data accuracy is to perform a point-in-polygon analysis. This method compares the location in question (a point) to a small area polygon such as a ZIP code or county boundary. If the point falls inside the correct polygon, then the location is accepted. Otherwise, the point is rejected. However, this method is computationally intensive. Calculating whether or not a point falls within a complex polygon is a computationally difficult task (although straightforward through the use of GIS technology).

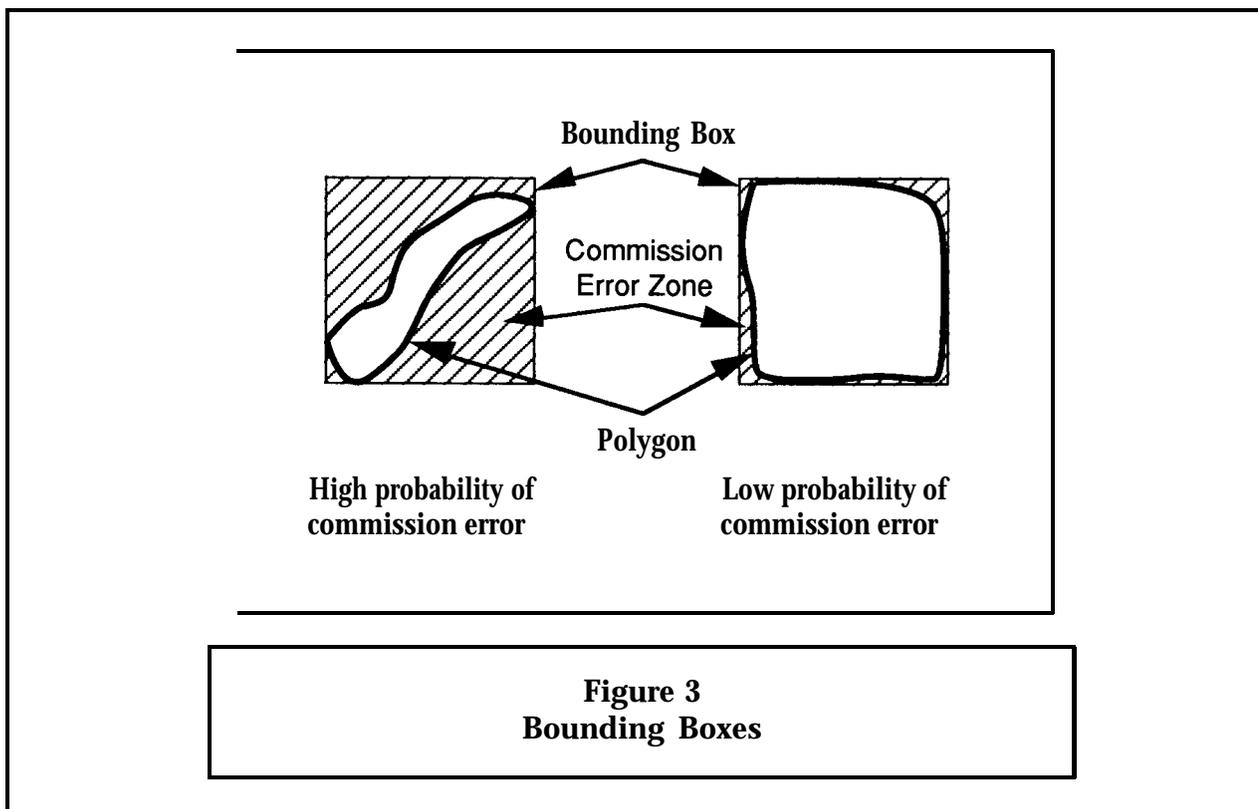
The accuracy of the point-in-polygon assessment method is dependent on the size of the polygon and the accuracy of the polygon boundary delineation. ZIP code polygon boundaries tend to vary among vendors. An approach to mitigate boundary uncertainty is to apply a buffer around the polygon and then to accept the additional points that fall outside a polygon, but inside the buffer. Similar to setting a distance threshold in the previous accuracy checking method, setting an appropriate buffer size is difficult. In an accuracy assessment exercise of the TRIS database, the reviewers found that a 1-2 km buffer around rural ZIP code polygons generally worked well.

A simplified approach that approximates the point-in-polygon accuracy assessment method, and is nearly as easily computed as the ZIP code centroid method, is the

bounding box test. To execute this test, one must assume that if a point falls inside a rectangle that bounds the small area polygon, then the point also falls inside the polygon.

Determining whether a point is inside a generic polygon is a complex geometric problem that is computationally intensive. In comparison, determining whether a point is contained inside a rectangle is computationally simple if the rectangle is oriented along the north-south axis. The point's latitude and longitude are simply compared to the latitude and longitude ranges defined by the rectangle. If both the latitude and longitude are within the range, the point lies inside the rectangle. While this test is not as accurate as a true point-in-polygon test, it can be implemented easily in almost any programming language. The bounding box test is susceptible to commission error but the likelihood of error reduces as the shape of the polygon becomes closer to the shape of the rectangle (Figure 3).

In analyzing the locational data from TRIS, submitted coordinates were usually either very accurate or very inaccurate. Based on this observation, the likelihood of commission error was reduced. The test could be improved by classifying the bounding boxes based on the area of the commission error zones. Based on this classification, statistics could be generated about the test's accuracy. OPPPTS has generated bounding boxes for ZIP codes and counties. They could be used by any data base management system to identify inaccurate lat/long coordinates.



7.6 Secondary Data Users, Multimedia Data Integration, Public Data Access, and Enforcement Data Requirements

The intent of the LDP is to “...extend environmental analyses and allow data to be integrated based upon location, thereby promoting the enhanced use of EPA’s extensive data resources for cross-media environmental analyses and management decisions.”³

The cross-media benefit that the policy encourages will allow sharing of data collection responsibilities. Instead of each program collecting its own locational data for an entity, program offices can “pool their resources” and collect all required data compliant with the 25 meter accuracy goal. This ability should make it more feasible for program offices to comply with the accuracy goal.

³ U.S EPA Locational Data Policy

Chapter 8

**A FRAMEWORK FOR ESTIMATING GEOCODING
LIFECYCLE COSTS**

8. A FRAMEWORK FOR ESTIMATING GEOCODING LIFE CYCLE COSTS

What is geocoding going to cost? This important question has no simple answer. One simplistic approach to estimating geocoding costs would be to multiply current industry cost-per-point estimates (for each technology) against the total number of regulated entities. Application of a more robust cost estimation methodology, like the one described below, will provide more reliable results.

8.1 Characterize Existing Records

All geocoding efforts should commence with a thorough review of existing databases of entities to be geocoded. A profile of these records should characterize their number, type, address quality, and urban/rural geographic distribution. Some of the reasons for performing this review are described below:

- The number of records will give an indication of the level of effort required.
- Knowledge of types of entities (e.g., outfall pipe, wellhead, industrial facility, underground storage tank, monitoring well, etc.) is important for selection of appropriate geocoding methods. It also will indicate potential geocoding problems including accessibility, standard geographic referencing, and field logistics.
- Address quality is particularly important. First, accurate addresses are required for efficient field logistics (i.e., an entity must be located on the ground before it can be geocoded). Addresses also must represent the actual location of the entity, not a mailing address. Addresses that do not commence with numerics (e.g., RD #3) are difficult or impossible to address-match. This factor can impede field logistics and may preclude the use of address matching technology as a geocoding solution.
- Knowledge of the geographic distribution of regulated entities, in particular the percentage that are “urban” or “rural” areas, is very useful. Depending upon the definition of urban/rural, organizations can predict geocoding productivity (in the sense of being able to assign a lat/long coordinate of acceptable accuracy to a given entity). For example, an urban area analysis of a data base can reveal the applicability of using address matching for the portion of the database that resides within known coverages of geocoded street addresses (available from the U.S. Census Bureau or from commercial data purveyors).

8.2 Define Geocoding Requirements

After a profile of the existing records inventory is completed, geocoding requirements and constraints should be identified. Various methodologies are available to guide organizations through the process of requirements analysis. One recommended approach is the DQOs methodology documented by EPA's Office of Research and Development (ORD).

A requirements analysis should answer the following fundamental questions:

- What is the intended primary use of these data?
- What are the possible secondary uses of these data?
- What are the programatically-defined minimum accuracy requirements?
- Can the 25 meter accuracy goal be achieved? If not, how can a waiver be substantiated? (a waiver process will be overseen by EPA's IRM Steering Committee).
- What are the economic and time constraints?

8.3 Select/Define Geocoding Methodology

Estimating the cost of geocoding depends upon a number of factors, including the methodology or technology employed. In some cases, multiple methodologies may be useful based on entity significance or location. Characterization of an existing records inventory, when placed in the context of an organization's geocoding requirements, will help to shed light on a preferred methodology (assuming there is broad familiarity with the available technologies). Once a technology is selected and a preliminary geocoding approach is defined, the geocoding life cycle fixed and variable costs may be estimated. It is critical that the analysis go beyond selecting a specific technology (e.g., GPS) and that a detailed geocoding implementation plan be prepared. This plan should include a decision about the overall geocoding approach: incremental field-based or centralized geocoding. A well-defined geocoding implementation plan not only will provide a sound a basis for a reasonably accurate cost estimate, but also is required by the LDP.

8.4 Estimate Geocoding Life Cycle Fixed and Variable Costs

Estimating geocoding life cycle costs ultimately depends upon defining and estimating fixed and variable costs. Whereas each geocoding technology has some specific fixed costs (e.g., GPS survey stations), these unique costs can be incorporated into a generic cost model. Figure 4 portrays a generic Geocoding Cost Estimation

GEOCODING COST MODEL	Take Inventory	Locate Address	Locate Facility	Geocode Feature	Update Inventory
Fixed Costs					
Set-up & Management Costs	The setup costs to get a proper list of "candidate entities". Cost of project design, initial staffing, client communications	The setup costs to get a valid mailing "Address" / phone	The setup costs to get an accurate street "Location"	The setup costs for a precise "Position"	The setup costs for updating the original source files
Other Fixed Costs					
Equipment	Computer hardware for database mgmt. PCs, IBM 3090	Business directories Map catalogues	Input/ Output Devices Business directories Map catalogue and storage facilities	GPS(s) Inertial Systems Laser theodolite	Communications system to source database
Software	Computer Systems Develop Method Timeline Program DBMS	Develop Method Map ID System	Design/Adapt System for Address matching Duplicate matching Map Coordinates	Document operational steps Write interface software	Define Requirements Design new data fields/ definitions
Other Infrs.	Download Database DB QC Procedures Storage Space for equipment, maps, digitizers, etc.	Set Up Database Design Address Cross-check and Duplication Systems	Secure or confirm Location	Set accuracy determination method Design data and exception-reporting forms	Verification of update Calculation of Changes and Costs
<div style="border: 1px solid black; padding: 10px; width: fit-content; margin: 0 auto;"> <p>Figure 4 Generic Geocoding Cost Estimation Model</p> </div>					

(continued on next page)

GEOCODING COST MODEL	Take Inventory	Locate Address	Locate Facility	Geocode Feature	Update Inventory
<i>Variable Costs</i>					
Profess- ional Costs					
Training	DB Administrator / Records Clerk	Address Editing	Valid Location criteria. Operator Training	Operator Training	
Ramp-up Costs					
Pilot Tests	Test database of 5% of entities in a pilot. Determine unusable record rates	Verify existence and phone (timeliness issue for pilot) Categorize/ cost Address confirmation steps	Secure confirmed street location. Confirm Location Identify Map (Quad) Identify Street Seg	Identify and geocode specific features. Determine production rates and accuracy's	Upload update fields. Acceptance test update software
Retooling Costs	Design Simultaneous extract process	Secure New Directories/ Maps	Rewrite Locational Procedures/ Systems	New Hardware Rework field procedures	Debug and rework programs
Staff Functions	Data extractor DB Administrator	Mail or Phone Address Confirmer	Create master Location coding file Confirm Location Identify Map (Quad) Identify Street Seg	Equipment operator(s)	Programmer
Quality Mgmt	Compare Source and database	Edit Final List Review rejects	Cross-check with ZIP, quad or county	Edit Positional Accuracy /Precision Cross-check	Compare Source and Update
Space & Facilities		Workrooms Project Storage Mail Facilities	Hard copy storage	Electronic and hard copy storage	

**Figure 4 (continued)
Generic Geocoding Cost Estimation Model**

Model that can be used as a framework for any program's geocoding requirements, independent of the technology employed⁴. Actual estimates of geocoding costs should account for existing capital equipment (e.g., GPS receivers) versus needed purchases.

All aspects of the model are not applicable in every case, Applicability depends in large measure on the technology to be employed and the underlying approach (i.e., incremental field-based or centralized geocoding). The model is flexible enough, however, to accommodate most geocoding methodologies.

The proposed Geocoding Cost Estimation Model is generic inasmuch as it is not dependent upon any one specific geocoding method or technology. It defines the overall processes required for any geocoding methodology. Although the details often differ from one method to another, the parameters in Geocoding Cost Estimation Model capture the full cost of any method employed.

Figure 5, a spreadsheet version of the proposed Geocoding Cost Estimation Model, is structured as a two dimensional matrix containing a Cost Axis (y) and a Process Axis (x). The Cost Axis contains two major cost categories: Fixed and Variable. Fixed Costs are divided into three subcategories: Set-up Costs, Other Fixed Costs, and Ramp-up Costs. The Process Axis contains 5 major categories: Take Inventory, Locate Address, Locate Facility, Locate Feature, and Update Inventory.

8.4.1 Fixed Costs

Investments in a geocoding infrastructure (without regard to the particular geocoding method) are referred to as "fixed costs." Geocoding large numbers of geographically distributed facilities and entities is logistically complex and information intensive. In order to achieve high orders of productivity, thereby reducing unit costs (i.e., the cost of geocoding a single entity), a geocoding infrastructure, or "system," is absolutely essential. This "system," while including a specific geocoding technology (e.g., GPS), encompasses a much wider range of activities and overall capabilities (discussed below).

As noted above, fixed costs fall into three major categories: set-up costs, other fixed costs, and ramp-up costs.

Set-up and other fixed costs are one-time investments in the necessary components of any geocoding process. Ramp-up costs include the costs of necessary trial and error (i.e., pilot testing) associated with implementing the geocoding "system" established during the set-up phase. Set-up costs include:

- Project planning and logistics.

⁴ Text contained in each of the matrices are examples of types of geocoding costs incurred and is not intended to be a comprehensive list.

	Take Inventory	Locate Address	Locate Facility	Geocode Feature	Update Inventory
FIXED COSTS					
<i>Set-up Costs (hours)</i>					
<i>Set-up Costs (\$)</i>					
<i>Training (hours)</i>					
<i>Training (\$)</i>					
<i>Other Fixed Costs (\$)</i>					
Equipment (\$)					
Software (\$)					
Other Infrastructure (\$)					
<i>Ramp-up Costs (hours)</i>					
Pilot Tests (hours)					
Retooling (hours)					
<i>Ramp-up Costs (\$)</i>					
Subtotal (hours)					
Subtotal (\$)					
VARIABLE COSTS					
<i>Professional Costs (hours)</i>					
Staff Functions (hours)					
Quality Management (hrs)					
<i>Professional Costs (\$)</i>					
<i>Other Variable Costs (\$)</i>					
Transportation (\$)					
Space & Facilities (\$)					
Materials (\$)					
Subtotal (hours)					
Subtotal (\$)					
TOTAL (HOURS)					
TOTAL (\$)					

Figure 5
Spreadsheet Version of Geocoding Cost Estimation Model

- **Staff supervision.**
- **Communications with participating organizations (i.e., contracts, regional offices, regulated facilities).**
- **Implementation of financial management and control systems.**

Other fixed costs include:

- **Equipment (i.e., computer hardware, map files, magnetic tapes or diskettes, furniture, etc.).**
- **Software (i.e., data management, spreadsheet, statistical analysis).**
- **Development of rules and procedures for geocoding personnel or regulated community.**
- **Training.**

8.4.2 Variable Costs

Variable costs are the costs of fully implementing the geocoding “system” over time. Major variable cost parameters include personnel/ functions, materials, quality management, transportation, space, and facilities. Although variable costs are expected to become consistent as the number of facilities or entities increases, there are many factors that can greatly affect the per-entity cost (i.e., communications, logistics, timing, public support).

Together, fixed and variable costs provide a well-rounded measure of the “true” costs of geocoding. The geocoding cost model encompasses systemic costs associated with geocoding that are much broader than costs associated just to applying a given method (e.g., digitizing a point on a paper map).

The process of identifying the location of a facility and, more importantly, an entity (e.g., smoke stack, outfall pipe), is more cumbersome and complex than it may appear upon first consideration. This fact is independent of the specific geocoding technology employed. The geocoding cost model accounts for five sequential and interdependent process steps, discussed below.

8.4.3 Process Parameters

To geocode accurately, one needs to know where the object to be geocoded is located before one can define where it is (to whatever degree of precision required or possible). To illustrate the point, take the example of sending a team from EMSL-LV into the field in the San Gabriel Basin to conduct a GPS survey of regulated entities. First, a list of features to be geocoded was compiled “take inventory.” Second, addresses for the candidate features were

identified “locate address.” Third, candidate features were mapped to support field logistics “locate facility.” Fourth, the GPS survey team located the facility and, subsequently, located the candidate features on the ground “locate feature.” The final step, “update inventory,” or successfully loading locational data into the host system, has yet to be completed.

As the Geocoding Cost Estimation Model indicates, the five step geocoding process incurs fixed and variable costs. By comparing the cost parameters with the process parameters, a fairly comprehensive understanding of cost emerges. In addition, the cost model provides a framework for comparing the estimated (and actual) costs of alternative geocoding technologies.

Chapter 9
SUMMARY

9. SUMMARY

This Guide To Selecting Latitude/Longitude Collection Methods is useful in informing EPA program offices, states, and other parties affected by the LDP about the geocoding lifecycle. In order to understand what geocoding is and the different methods available, a manager or technical specialist must be able to understand the capabilities, realistic costs, and accuracies of different geocoding methods/technologies. In order to implement the effective use of geocoding, cross-cutting geocoding issues must be identified, a framework for designing and estimating the cost of a geocoding project or program consistent with EPA IRM policy must be provided, and locational data accuracy checking methods must be reviewed.

Different organizations operate differently and have varying requirements. Therefore, the extent to which any one geocoding method will be employed may vary. In conjunction with the LDP, however, this Guide will provide Federal agencies, states, and other parties with guidelines and techniques to implement useful latitude/longitude collection methods.

Appendix A
FACT SHEETS ON GEOCODING METHODS

Global Positioning Systems (GUS)	A-1
Photogrammetry	A-3
Map Interpolation	A-5
Photo Interpretation/Map Interpolation	A-7
Address Matching	A-9
ZIP Code Centroids	A-11

Global Positioning Systems (GPS)

Description

GPS is an earth-surface positioning system that utilizes a constellation of earth-orbiting satellites deployed and maintained by the Department of Defense (DOD). GPS produces latitude/longitude coordinates by relying on established trigonometric principles, timing, range measurements, and several statistical models. Analogous to traditional land surveying, GPS requires a minimum of four simultaneous satellite observations to precisely position a point on the earth. Typical GPS surveying steps include:

- Plan detailed field logistics.
- Assemble GPS survey team.
- Travel to the site of entities.
- Establish GPS base station that collects data simultaneously with, and from the same satellite constellation as the GPS units in the field.
- Take GPS “reading” and move to the next entity until all readings are completed.
- Perform post-processing/QA on raw GPS data in an office setting.
- Enter locational data into inventory or programmatic data system.

Reasonably Achievable Accuracy

Under normal conditions, using differential GPS surveying techniques (i.e., using two receivers simultaneously, one from a known position and one from the entity being geocoded), 5 to 20 meter accuracies can be achieved. Accuracies ranging from 5 to 100 meters are possible depending upon a variety of technical and geographical factors (discussed below). GPS is one of the few geocoding techniques that can consistently produce locational data that complies with EPA’s 25 meter accuracy goal if properly executed.

Cost

The cost of GPS-produced locational data is influenced primarily by equipment, labor, and transportation. GPS receivers cost as little as \$2,500 and as much as \$100,000, although prices are dropping quickly due to commercial competition and technological breakthroughs. Base stations, computers, and other field equipment increase costs significantly. Labor costs are incurred for survey planning, training, field work, and post processing. Finally, transportation and related travel costs can be relatively high for extended field surveys. For centrally planned, dedicated GPS surveys, the average cost per point ranges from \$75 to \$125. at a 2 to 5 meter accuracy

The cost per point of incremental GPS surveys (i.e., performed by field staff already “on-site” conducting other duties) is expected to be less expensive (\$35 to \$70 within a 2 to 5 meter accuracy) than centrally planned and executed GPS surveys.

Benefits

- GPS is the LATF’s recommended technology because of its demonstrated ability to geocode data that comply with, or are more accurate than, the Agency’s 25 meter accuracy goal.
- GPS rapidly is becoming a standard geocoding technology in the surveying community; many of the early problems associated with this emerging technology are being resolved through extensive field work.
- GPS material costs are dropping rapidly due to high product demand and commercial competitiveness.
 - Many states are beginning to implement GPS base station networks.

Limitations

- Field access to some EPA regulated entities by GPS survey crews maybe impossible or difficult at best; this limitation is true for all field-based geocoding methods.
- DOD has authority to invoke Selective Availability (SA) of its satellites and, therefore, degrade the achievable accuracy of GPS surveys to approximately 100 meters.
- Reception problems in urban areas and areas of high physical relief can limit GPS accuracies to 200 meters and, in some cases, preclude readings entirely.
- Post processing of GPS survey readings is usually necessary to achieve appropriate orders of accuracy, and is usually time-consuming and complex.
- Data exchange between different brands of receivers may be difficult, time consuming, or, in some cases, impossible.
 - Standard procedures for collecting reliable GPS data are yet to be developed.

Photogrammetry

Description

Photogrammetry is defined as the “...art and science of obtaining reliable measurements from photographs.*” Photogrammetric sciences are a fundamental part of modern map making and most small and medium scale maps are made from aerial photographs. The aerial photographic holdings in EPA and other agencies of the Federal government are a wealth of spatial and temporal data about environmental conditions and processes. The Environmental Monitoring Systems Laboratory in Las Vegas (EMSL-LV) currently provides information that is interpreted from aerial photographs to characterize hazardous waste sites, analyze wetlands, identify ecological resources and meet a number of environmental monitoring needs. EMSL-LV has now acquired the capability to supply highly accurate metric, or measurement, information for similar applications.

Photogrammetric data is produced on very precise photo measurement devices called “analytical stereoplotters.” These devices are typically calibrated to the micron level and enable the scientist to create complex mathematical models that correct for known distortions in the photographs. From these three dimensional photo models, highly accurate measurements and positional data can be derived for mapping and analytical purposes. This data can be produced in digital format directly for input in a Geographic Information System (GIS).

Cartographic information can be produced from aerial photographs to the locational specifications of the U.S. National Map Accuracy Standards. These can be traditional map features such as roads or hydrology, or special map layers such as historical hazardous waste site activity or fractures in the bedrock. Any information that can be derived from an aerial photo can be accurately mapped in a digital format. Once the photo model is established, thematic information represented by points, lines, and polygons can be input directly into digital format without transfer to a hard-copy map and digitizing from the map base. This saves time and reduces spatial error propagation.

Exact measurements can be accomplished on an analytical stereoplotter to help characterize activity of environmental interest. For example, in studying hazardous waste sites, volumes of waste accumulations and changes in such volumes are needed to evaluate remedial options. Also, precise distance and area measurements can be utilized for risk assessment and other site characterization activities.

Cartographic information that depicts the elevation of the land surface, such as the contour map or the digital elevation model (DEM) can routinely be produced by photogrammetric techniques. The resolution of this data can be tailored to the specific needs of the project.

* (ASPRS, 1991)

Any feature that is observable on an aerial photograph can be accurately referenced to a coordinate system. Photogrammetry can be extremely useful for collecting and recording the coordinate data that is required by the LDP. Conversely, information that is not readily visible on photographs, such as property boundaries or pipeline locations, can be digitally superimposed onto the photo model for special mapping or interpretive purposes.

Reasonably Achievable Accuracy

Photogrammetric accuracies are dependent on film scale, the quality of ground control data, and a number of other factors, but sub-meter accuracies are routinely achievable from standard photo products.

Cost

The cost of photogrammetrically-derived locations can vary considerably depending on a number of factors, such as the scale of the photos and the quality of the ground control. However, price ranges from \$25 to 100 per point are reasonable expectations.

Benefits

- **Photos represent permanent records of environmental conditions.**
- **Photogrammetry can produce extremely high accuracies.**
- **Photogrammetry is time-tested and legally defensible in the most rigorous courtroom setting.**
- **Any feature on a photograph can be precisely geocoded with ease. The coordinate definition for linear and polygonal features can be determined with the same ease as point features.**
- **Full Photogrammetric capabilities now exist within EPA.**
- **QA/QC parameters are inherent in the mathematical models that are used in the photogrammetric process.**
- **Photogrammetry does not require a field visit to the actual entities being geocoded and can be utilized when site access is impeded.**

Limitations

- **Photogrammetry requires capital equipment and trained technicians.**
- **The process cannot be easily performed outside a laboratory.**

Map Interpolation

Description

Map interpolation involves direct measurement from existing paper maps. Typical map interpolation steps include:

- Develop a candidate list of regulated entities to be geocoded, including addresses and facility IDs.
- Sort the candidate list by map upon which they are expected to be located.
- Acquire the appropriate maps (e.g., USGS 1:24,000-scale topographic maps).
- Identify map symbols that represent the candidate entities or that provide sufficient contextual information to geocode required entities.
- Measure the location of candidate entities using various methods (i.e., bar scale, engineer's scale or graduated ruler, electronic digitizer).
- Enter locational data into inventory or programmatic data system.

Reasonably Achievable Accuracy

Achievable accuracies on USGS 1:24,000-scale quadrangles (available nationally) range from 12 to 50 meters for easily identifiable features. Larger scale (i.e., higher resolution) maps, such as those produced for hazardous waste site investigations, support achievable accuracies in the 5 to 25 meter range.

Cost

Dedicated map interpolation managed and performed in an office setting costs between \$40 to \$60 per point. The cost of manual geocoding using maps directly in the field (by field "inspectors") should be considerably less, approximately \$28 to \$40 per point.

Benefits

- Maps at a consistent scale (1:24,000) are available nationally from the U.S. Geological Survey.
- Maps are inexpensive when compared to other geocoding technologies.
- Map interpolation can occur in the field, in the office, or on the telephone talking to a site operator; costs and accuracies differ significantly.

- **Non-rectified air photographs can provide useful ancillary information to existing paper maps during manual map interpolation.**

Limitations

- **Map interpolation does not always yield coordinate data that achieve EPA's 25 meter accuracy goal for locational data.**
- **Not all entities are identifiable on maps; studies have shown that as many as 40% or more are not identifiable on a single map source.**
- **Accuracy of map interpolation depends upon map source currency and scale, the object being geocoded, identifiable landmarks, existence of street names and address ranges, and skill of the interpolator. These factors are highly variable and, therefore, produce inconsistent geocoding results.**
- **Misidentification of an entity on the map can result in inaccuracies of greater than 100 meters.**

Photo Interpretation/Map Interpolation

Description

Photo Interpretation/Map Interpolation (PI/MI) is actually an integrated technique by which the user transfers information from an aerial photograph on to a map base and then extracts coordinate information via standard manual or digital map interpolation techniques. Current and historical aerial photographs contain a wealth of environmental data that are not routinely placed on standard map products. However, a 'raw' aerial photo contains no inherent coordinate information. By 'interpreting' the information on an aerial photo, such as the location of a hazardous waste site, and then transferring that information to a standard map base, coordinate data can be extracted.

Photographic Interpretation is performed by viewing aerial photographs through microscopes or stereoscopes. Stereoscopic viewing creates a perceived three-dimensional effect which, when combined with viewing at various magnifications, enables the analyst to identify signatures associated with different features and environmental conditions. The term "signature" refers to a combination of visible characteristics (such as color, tone, shadow, texture, size, shape, pattern, and association) which permit a specific object or condition to be recognized on aerial photography.

By correlating observable features on the photograph, such as the road network, with the same set of features on a standard map, the analyst can transfer the location of other photographic elements on to the map base for gee-referencing. The typical PI/MI process would involve the following steps:

- Obtain aerial photograph(s) and maps of the area of interest.
- Interpret photographs for specific 'signatures' of environmentally significant activity or location.
- Transfer data to the map.
- Interpolate coordinate location(s) from map.
- Enter locational data into inventory or programmatic data system.

Reasonably Achievable Accuracy

Accuracies of PI/MI are dependent the spatial accuracy of the basic map product and the skill of the interpreter in transferring the information. However, from standard maps such as the USGS 1:24,000 series, accuracies of 10-15 meters are routine.

cost

Costs for this method generally will be from \$40 to \$100 per point.

Benefits

- **Field visits are not required.**
- **Aerial photos and maps are common resources in EPA projects.**
- **PI/MI amenable to office environment.**
- **Relatively current aerial photography routinely available.**
- **Permanent record of activities and other programmatic purposes served by photos.**

Limitations

- **Moderate to high level of manual effort is required.**
- **Final accuracy is dependent on photo and map scale.**
- **Formal training in photographic interpretation techniques may be required, depending on the specific information being extracted.**

Address Matching

Description

Address matching is a set of semi-automated operations for deriving latitude/longitude coordinates from street addresses. Address matching is achieved by “comparing” a tabular file of street addresses with a digital cartographic street network file (e.g., Bureau of Census’s GBF/DIME and TIGER) that contains street names, address ranges, and latitude /longitude coordinates of the street network. Street address geocodes (lat/long) are generated by matching the equivalent street name in the network file and interpolating along its associated address range. Numerous commercial Geographic Information System (GIS) software packages (e.g., ARC/INFO) provide address matching capabilities, and several commercial firms perform address matching services. Typical steps required for address matching include:

- Compile address file for target entities.
- Check for correct and consistently formatted addresses.
- Submit the “clean” address file to a service bureau.
- Enter locational data into inventory or programmatic data system.

Alternatively:

- Compile address file for target entities.
- Check for correct and consistently formatted addresses.
- Load appropriately formatted digital cartographic data base into a GIS.
- Run addresses through address matching software in batch or interactive mode; recheck and correct addresses which cannot be processed.
- Enter locational data into inventory or programmatic data system.

Reasonably Achievable Accuracy

The accuracy of address-matching depends upon the accuracy of the street network file and the length of the street segment upon which interpolation is performed. Accuracies are higher in urban areas, and significantly lower in rural areas. Reasonably achievable accuracy for address matching ranges from 50 to 500 meters.

Cost

The cost of address matching differs if it is performed “in-house” versus through a service bureau. Average commercial address matching costs have been documented in the \$1.25 to \$4.00 per point range. Preprocessing costs for compiling and normalizing address files adds an incremental cost of anywhere from \$2.00 to \$6.00 a

point, resulting in an estimated per point cost of \$3.25 to \$10.00. The cost of in-house address matching is considerably higher per point if up-front investments are considered. For example, the cost of building the properly formatted street network file in a GIS is significant.

Benefits

- **Address matching is a relatively low cost, batch-oriented method of producing latitude/longitude data from address data bases.**
- **Address matching can be performed in the office or through a number of qualified address matching service providers.**

Limitations

- **Address matching may not yield measurements that comply with EPA's locational accuracy goal of 25 meters.**
- **At best, address matching produces locational coordinates that approximate the property parcel centerline of a given facility; pipes, stacks, underground storage tanks, and wellheads do not have addresses per se, which limits the utility of address matching for these types of entities.**
- **Address matching in rural areas remains a somewhat unreliable method, because of the distance between street intersections and the extensive use of ZIP code centroids as default locational data. Furthermore, rural addresses may not exist in some jurisdictions.**

ZIP Code Centroids

Description: ZIP code centroid geocoding is an automated operation for deriving latitude/longitude coordinates from ZIP codes contained in street addresses. This geocoding method assigns the same lat/long of the ZIP code centroid location to all entities within the same ZIP code. The term “ZIP code centroid” is often misleading. ZIP code centroid geocodes are procured from a vendor and not from the U.S. Postal Service (USPS). Each vendor’s geocodes are based on the vendor’s interpretation of where ZIP code boundaries lie. The USPS defines ZIP codes in terms of optimal carrier routes, as opposed to explicitly drawing boundaries on a map. The lack of fixed ZIP code boundaries creates some discrepancies between different vendor’s centroids. Typical steps required for ZIP centroid geocoding include:

- Acquire a ZIP code centroid file from a vendor.
- Check for correct ZIP codes in street address records.
- Run ZIP codes through matching software in batch mode.
- Submit alternatively “clean” address file to a service bureau.
- Enter locational data into inventory or programmatic data system.

Reasonably Achievable Accuracy

The accuracy of ZIP code centroids depends upon the size of the area served by the ZIP code. ZIP codes are extremely variable in size, being defined by mail volume and population or mail drop density. Therefore, the areal extent of urban ZIP codes is rather small (in some cases the 5-digit ZIP codes is merely a city block), while rural ZIP codes are either extremely large (150,000 square miles in one Alaska ZIP code) or are served by Post Office boxes. As a result, ZIP code accuracies are higher in urban areas and significantly less so in rural areas. Reasonably achievable accuracy for ZIP code centroid ranges from 50 to 500 meters in urban areas to many kilometers in rural areas.

Cost

The cost of ZIP centroid geocoding differs if it is performed “in-house” versus through a service bureau. Average per point ZIP centroid matching costs has been documented in the \$0.01 to \$0.60 range. Preprocessing costs for compiling and normalizing address files adds an incremental cost of anywhere from \$0.01 to \$0.10 a point, resulting in an estimated per point cost of \$0.02 to \$0.70. The major costs associated with in-house generation of lat/longs as ZIP code centroids is the cost of procuring a ZIP code centroid file from a vendor and the cost of correcting/ normalizing street address files. EPA’s Office of Information Resources Management (OIRM) currently maintains a license to a commercial ZIP code centroid file.

Benefits

- **ZIP code geocoding is a relatively low-cost, efficient centralized method of producing locational data from existing address files.**
- **ZIP code geocoding has been widely employed at EPA for many years and is a well-understood technique.**

Limitations

- **ZIP code geocoding fails to meet EPA's 25 meter accuracy goal for most locational data.**
- **At best, ZIP code centroids produce locational data coordinates that approximate the centroid of the block where the facility is located; pipes, stacks, underground storage tanks, and wellheads do not have addresses, which greatly limit the utility of ZIP centroids for these types of regulated entities.**

ZIP code geocoding in rural areas produces higher-order inaccuracies than in urbanized areas because of the typically larger spatial extent of rural ZIP codes.

ZIP code boundaries cannot be reliably produced from vendor to vendor because of the way ZIP code boundaries are defined by the U.S. Postal Service.

Appendix B
REFERENCES

References

Battelle, 9/1990. Final Letter Report on Regulatory Review for EPA Spatial Data Requirements to U.S. Environmental Protection Agency, Office of Information Resources Management. Contract No. 68-03-3534, Work Assignment No. H2-51, Task 1.a.

Bissex, D. A., C. J. Franks, and A. Heitkamp., 1990. Quality Assurance for Geographic Information Systems. Urban and Regional Information Systems Proceedings, Edmonton, Alberta, Vol. 2, pp.106-118.

Bolstad, P. V., P. Gessler, and T. M. Lillesand., 1990. Potential Uncertainty in Manually Digitized Map Data. International Journal of Geographic Information Systems, Vol. 4(4), pp.399-412.

Croswell, P. L., 1987. Map Accuracy: What is it, Who Needs it and How Much is Enough. Urban and Regional Information Systems Proceedings, Edmonton, Alberta, Vol. 2, pp.48-62.

Fitzsimmons, C. K. 1988. Evaluation of Selected Methods for Determining Geographic Coordinates. Unpublished Report for Exposure Assessment Division, Environmental Monitoring Systems Laboratory Las Vegas, NV. Prepared under EPA Contract CR 812189-02.

Hunter, G. J., and I. P. Williamson, 1990. The Need for a Better Understanding of the Accuracy of Spatial Databases. Urban and Regional Information Systems Proceedings, Edmonton, Alberta, Vol. 4, pp.120-128.

Hurt, J., 1989. GPS: A Guide to the Next Utility. Trimble Navigation Ltd., Sunnyvale, CA.

Kruczynski, L. R., 1990. An Introduction to the Global Positioning System and its use in Urban GIS Applications. Urban and Regional Information Systems Proceedings, Edmonton, Alberta, Vol. 3, pp.87-91.

Palmer, D., 1989. Designing a GIS/LIS: Some Accuracy and Cost Considerations. Urban and Regional Information Systems Proceedings, Edmonton, Alberta, Vol. 2, pp.52-56.

Slonecker, E. T., and J. A. Carter, 1990. GIS Applications of Global Positioning System Technology. GPS World, Vol. 1(3), pp.50-55.

Slonecker, E. T., and M. J. Hewitt III, 1991. Evaluating Locational Point Accuracy in a GIS Environment. Geo Info Systems, Vol. 1(6), pp.36-44.

Slonecker, E. T., and N. Tosta, 1992. National Map Accuracy Standards: Out of Sync, Out of Time. Geo Info Systems, Jan. 1992, pp. 20-26.

US EPA. 1991. TRI Location Data Quality Assurance for GIS Final Report (unpublished report for US EPA Office of Toxic Substances, Washington, D. C.)

From Photogrammetric Engineering and Remote Sensing:

Acceptance Tests

Testing Land-Use Map Accuracy: Another Look, Michael C. Ginevan, 10: Ott 79:1372

Accuracy

Map Accuracy, E.J. Schlatter, 10:206

map, Funk, 24:392

map evaluation at Army Map Service, Coulthart, 23:855

testing, USGS, 20:181

map, Weg, 27:148

Effects of Interpretation Techniques on Land-Use Mapping Accuracy, Floyd M. Henderson, 3: Mar 80:359

Accuracy Specifications

Map Accuracy Specifications, Adopted by ASP in 1940 50th Anniversary Highlights, 2: Feb 84:237

Accuracy Test

The Minimum Accuracy Value as an Index of Classification Accuracy, Stan Aronoff, 1: Jan 85:99

Analysis of Variance

Analysis of Variance of Thematic Mapping Data, George H. Rosenfield, 12: Dec 81:1685

Cartography

Spatial Accuracy Specifications for Large Scale Topographic Maps, Dean C. Merchant, 7: Jul 87:958-961

Category Variances

Spatial Correlation Effects Upon Accuracy of Supervised Classification of Land Cover, James B. Campbell, 3: Mar 81:355

Classification Accuracy

***Accuracy Assessment: A User's Perspective*, Michael Story and Russell G. Congalton, 3: Mar 86:397**

Interpolation

least squares, Kraus, 38:487,1016

methods, Schut, 40:1447

***Interpolation of a Function of Many Variables, II*, Arthur, 39:261, 31:348**

***Performance Evaluation of Two Bivariate Processes for DEM Data Using Transfer Functions*, Mohamed Shawki Elghazali and Mohsen Mostafa Hassan, 8: Aug 86:1213**

Interpolation, Height

Experience with Height Interpolation by Finite Elements, H. Ebner and F. Reiss, 2: Feb 84:177

Interpolation, Raster

A Comparative Analysis of Polygon to Raster Interpolation Methods, Keith C. Clark, 5: May 85:575

Map Accuracy

***The Map Accuracy Report: A User's View*, Stan Aronoff, 8: Aug 82:1309**

***Accuracy Specifications for Large-Scale Maps*, The Committee for Specification and Standards, American Society of Photogrammetry, 2: Feb 85:195**

***Aerial Verification of Polygonal Resource Maps: A Low-Cost Approach to Accuracy Assessment*, Thomas H. George, 6: Jun 86:839**

***A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data*, Russell G. Congalton, 5: May 88:593-600**

***Using Spatial Autocorrelation Analysis to Explore the Errors in Maps Generated from Remotely Sensed Data*, Russell G. Congalton, 5: May 88:587-592**

***ASPRS Interim Accuracy Standards for Large-Scale Maps*, The American Society for Photogrammetry & Remote Sensing, 7 Jul 89:1038-1040**

Accuracy Assessment of a Landsat-Assisted Vegetation Map of the Coastal Plain of the Arctic National Wildlife Refuge, Nancy A. Felix and Daryl L. Binney, 4: Apr 89:475-478