

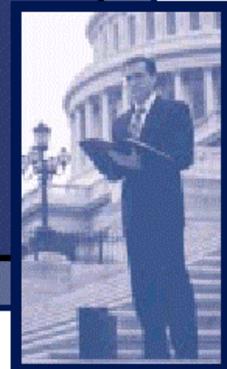


COPS Publication

Community Oriented Policing Services

www.usdoj.gov/cops/

Geocoding in Law Enforcement Final Report



Geocoding in Law Enforcement

Final Report

Prepared by:

**The Crime Mapping Laboratory
Police Foundation**

August 2000

**Report to the Office of Community Oriented Policing Services
Cooperative Agreement #97-CK-WXK-004**



Table of Contents

Introduction	4
The Geocoding Process	5
Step One: Prepare the geographic and tabular files for geocoding	5
Step Two: Specify geocoding parameters	10
Step Three: Geocode	11
Step Four: Review results	14
Step Five: If necessary, respecify parameters and geocode again	16
Conclusion	17



Introduction

Geocoding is the process of bringing tabular and geographic data together...

Geocoding is the process of bringing tabular and geographic data together based on a common geographic unit of analysis. A geographic unit of analysis refers to a spatial characteristic within the data that is necessary to locate it on a map such as address, zip code, beat, grid, etc. Tabular data are contained in a table and are a list of records that, along with information about the record, contain addresses or some other type of geographic variable. Examples of law enforcement tabular data are calls for service, crime, accidents, citations, sex offenders, and arrests. Along with information about the type of call or crime, date, time, etc., these data contain the location of the incident (address, grid, and/or beat), which will allow it to be geocoded. Geographic data are data that can be displayed on a map such as streets, census tracts, parcels, buildings, etc. Examples of geographic data specific to law enforcement are police districts, beats, or grid boundaries.

...geographic data used for geocoding are called "reference data"...

The purpose of geocoding is to assign tabular data to a location on the earth's surface to visualize the spatial characteristics of the data. It is analogous to placing a pin on a map in the appropriate location. Unlike the paper map, when geocoding in a geographic information system, data associated with the pin, or data point, are available. In order to geocode either electronically or manually, there must be a common geographic unit of analysis. In law enforcement, address is primarily used as the geographic unit of analysis to which tabular data are geocoded. However, other types of geographic units are also used such as parcels, zip codes, census tracts, census blocks, or beats. In the case of geocoding, geographic data used for geocoding are called "reference data" since the geographic data are used to reference the tabular data.

The following report is a discussion of the geocoding process, regardless of the Geographic Information Systems (GIS) software used, and some common problems faced when geocoding law enforcement data.



The Geocoding Process

In general, there are five basic steps in the geocoding process. They are:

1. Prepare the geographic and tabular files for geocoding.
2. Specify the geocoding parameters.
3. Geocode.
4. Review results.
5. If necessary, respecify parameters and geocode again.

Step One: Prepare the geographic and tabular files for geocoding

The first step is preparing the tabular and geographic files, which is the most important and can be the most time consuming of the five steps.

Tabular data. Tabular data used in law enforcement primarily consist of calls for service, crime, arrests, and accident data. Other types of tabular law enforcement data can include field information, sex registrant information, intelligence information, etc. In the past, much of this information was not captured electronically. However, currently, many departments either already have in place or plan to acquire an automated computer aided dispatch (CAD) system and a record management system (RMS) which have the capacity to store digital information as well as standardized tables of addresses and location names. The most important aspects of these systems as they relate to geocoding are that they reduce the need for manual data entry and provide clear and consistent rules in order to ensure that data are reliable and valid. Because individuals are the data entry mechanism, there is still the possibility of human error and inconsistencies, which is why training should coincide with the technology.



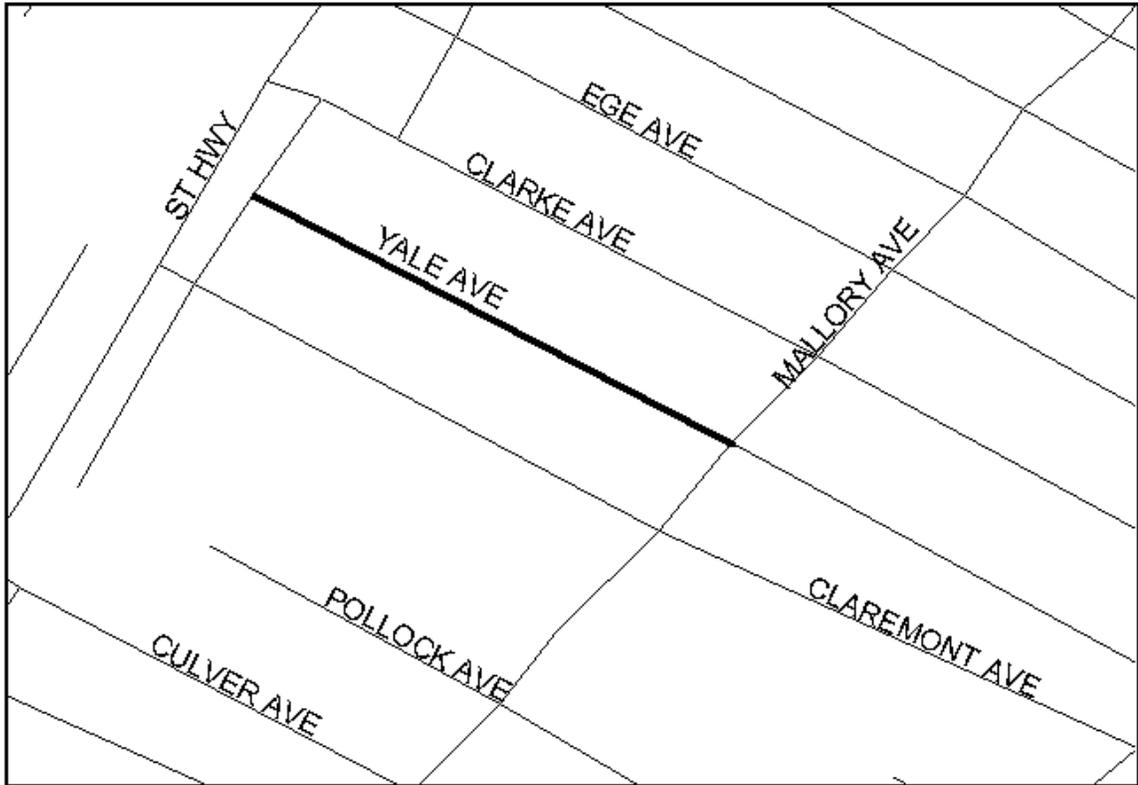
Geographic data. A variety of geographic data types may be used as a reference layer, though street files such as the Census Bureau's TIGER/Line ® files are the most commonly used. Street files are often called centerline files in that they depict the center of the streets by line segments. Details of curbs, alleys, cul de sacs, etc. are not depicted in centerline files. The following is an example of a street centerline file most commonly used.

Street files are often called centerline files in that they depict the center of the streets by line segments.



...street files are geographic databases containing records that describe interconnected street segments.

In this context, street files are geographic databases containing records that describe interconnected street segments. More specifically, each street is broken down into sections with address ranges associated with them. These sections usually span between cross streets. The following darkened line is an example of a street segment. The table below is the accompanying table.



<i>Street</i>	<i>Frontleft</i>	<i>Toleft</i>	<i>Frontright</i>	<i>To right</i>
YORK ST	473	479	478	480
YALE AVE	83	173	84	174
YALE AVE	1	81	2	82
WRIGHT AVE	1	91	2	96
WOODWARD ST	283	395	304	380

Street centerline databases generally have fields that describe the physical address along each street segment in the layer. Conventionally, addresses are recorded as 'from' and 'to' ranges for both the left and right sides of each street segment. In the above example, the centerline street segment of Yale Ave that is thicker than the others is represented by the address ranges of 83 to 173 on one side of the street and 84 to 174 on the other side, as highlighted in the table.



In addition to address ranges, street segments can be depicted by several variables such as street name, type, prefix, and suffix. Street centerline databases vary in how the addresses are described. In the previous example, address is depicted with a variable called “street” along with the address ranges. The following are tables with highlighted examples of two additional ways in which addresses are found in a geographic database. The prefix and suffix fields are used to store information such as quadrant. The first table highlights a street segment with the prefix W for West Northern Ave. The second table highlights a street segment with the suffix NW for Chesterfield Place Northwest. Type refers to the type of roadway, e.g. street, road, lane, and drive. The highlighted examples in the following tables show Ave for Northern Avenue and Pl for Chesterfield Place.

<i>Shape</i>	<i>L_f_addr</i>	<i>L_t_addr</i>	<i>R_f_addr</i>	<i>R_t_addr</i>	<i>Prefix</i>	<i>Name</i>	<i>Type</i>	<i>Suffix</i>
PolyLine	4901	5099	4900	5098	W	Vista	Ave	
PolyLine	5051	5099	5050	5098	W	Kaler	Cir	
PolyLine	5947	5999	6000	6098	W	Northern	Ave	
PolyLine	5401	5499	5400	5498	W	Orangewood	Ave	
PolyLine	5301	5399	5300	5398	W	Orangewood	Ave	
PolyLine	7451	7301	7452	7300	N	51st	Ave	
PolyLine	7499	7453	7498	7454	N	51st	Ave	
PolyLine	7515	7501	7516	7500	N	51st	Ave	
PolyLine	7605	7517	7604	7518	N	51st	Ave	
PolyLine	7647	7607	7648	7606	N	51st	Ave	

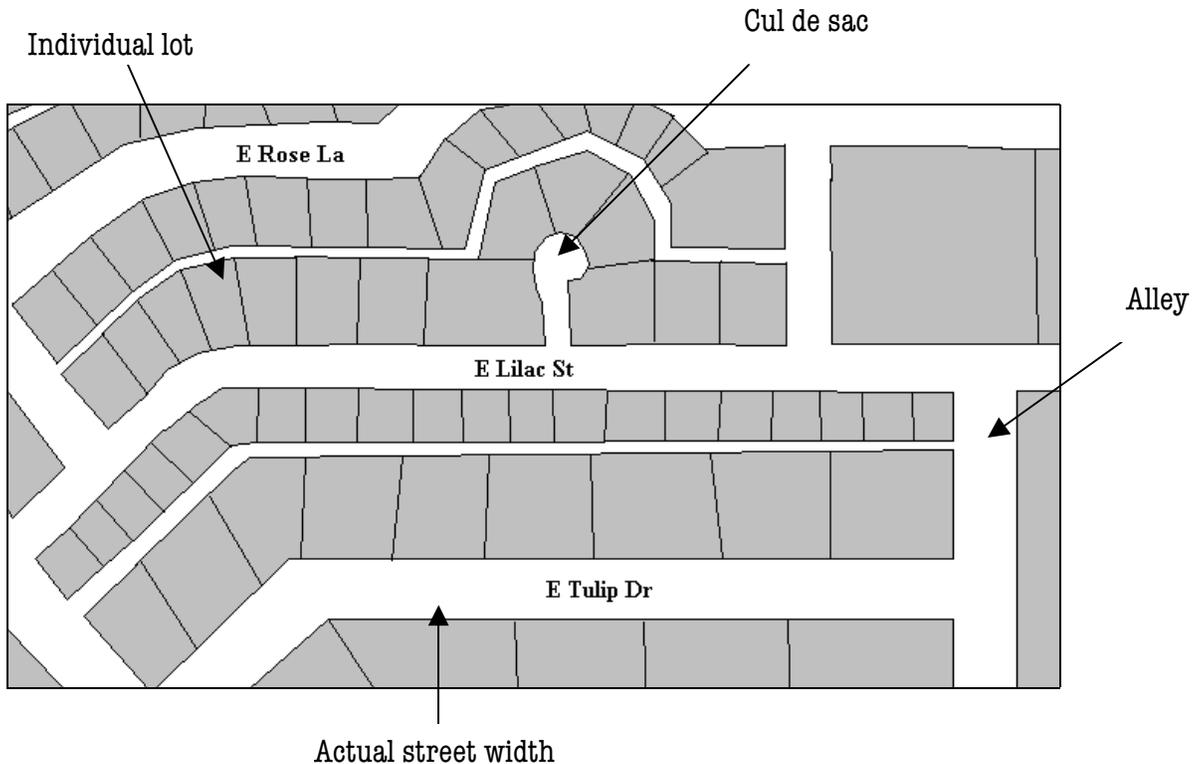
<i>Shape</i>	<i>L_f_addr</i>	<i>L_t_addr</i>	<i>R_f_addr</i>	<i>R_t_addr</i>	<i>Prefix</i>	<i>Name</i>	<i>Type</i>	<i>Suffix</i>
PolyLine	2799	2701	2798	2700		Albemarle	St	NW
PolyLine	3099	2901	3098	2950		Audubon	Ter	NW
PolyLine	2699	2601	2698	2600		Tilden	Pl	NW
PolyLine	2899	2801	2898	2800		Chesterfield	Pl	NW
PolyLine	2899	2801	2898	2800		Allendale	Pl	NW
PolyLine	4201	4249	4200	4298		Lenore	Ln	NW
PolyLine	4799	4601	4798	4600		Broad Branch	Rd	NW
PolyLine	4499	4423	4498	4400		Broad Branch	Rd	NW
PolyLine	4421	4401	0	0		Broad Branch	Rd	NW
PolyLine	4599	4583	4598	4546		Broad Branch	Rd	NW

Because the Census Bureau’s mission to count and profile the nation’s people and institutions does not require very high levels of positional accuracy in its geographic products, its TIGER/Line ® files and centerline maps are designed to show only the relative position of geographic elements. Some level of correction is generally necessary when using these products.



A parcel file is a polygon layer used to keep track of lots, subdivisions, and ownership information primarily for planning and tax purposes.

As previously mentioned, another common layer used as a reference for geocoding is the parcel file. A parcel file is a polygon layer used to keep track of lots, subdivisions, and ownership information primarily for planning and tax purposes. The following is an example of how a parcel map looks in a GIS. Note that it is more specific than a centerline map in that alleys, cul de sacs, lot size, and street widths are discernable.



Parcel files provide a good base layer to examine relationships between land use, zoning, demographic data, and economic data as all of this information is contained in parcel files. Also, parcel reference layers can be very helpful in mapping law enforcement information because they can help to indicate the exact location of an incident, e.g. corner lot, and provide ownership information which may prove valuable, e.g. in the case of rental property.

While parcel files have the ability to be highly accurate, like street files, this accuracy is largely dependent on how often and how well they are maintained. Because of the nature of real estate and other planning mechanisms, keeping parcel files updated and accurate can be a full-time job for several technicians depending on the area.



For any crime mapping application to be effective, basic reporting procedures need to be uniform and consistent.

Successful geocoding depends not only on the quality of the geographic reference layer, but also the tabular data to be mapped. For any crime mapping application to be effective, basic reporting procedures need to be uniform and consistent. While most high-end geocoding software packages incorporate complicated algorithms making them somewhat forgiving, a well established reporting procedure is often the best way to ensure a high rate of accuracy.

Step Two: Specify geocoding parameters

Step two consists of specifying the parameters of both the reference data and the tabular data to determine how the geographic units from each source will be matched successfully.

Selection of which reference layer would be most effectively used for geocoding must consider the needs of the individual end user. Considerations of accuracy, cost of product, maintenance costs, utility, current reporting procedures, and compatibility with other existing reference layers must be dealt with on a case-by-case basis. In the case of law enforcement, most information is geocoded to the street reference layer. However, sometimes it is necessary to be more specific, in which case a parcel reference layer may be used. For example, one might use the street centerline file to pin map residential burglaries in an entire city since the exact location of the burglary would not be as important as getting a general idea of where the incidents are occurring throughout the city. On the other hand, one might use a parcel file to pin map residential burglaries in a one or two block area to determine exactly which houses were targeted and whether they were along alleys, wooded areas, on the end of the block, etc.

Parameters set to determine how tabular and geographic data are matched are set within the GIS software...

Parameters set to determine how tabular and geographic data are matched are set within the GIS software and may include spelling sensitivity of the match, address style, e.g., whether to match on address only or address and zip code, or whether to accept partial match scores. A match score denotes the number of address records from the tabular file that were located in the reference layer. Another parameter is offset distance, which is placing a data point a predetermined distance away from the original geocoded location. This may be done to place points on one side of the street segment or another to make the map more readable and/or realistic.

These parameters are choices within the GIS software that will affect the geocode match rate. These preferences should be set and modified only after an examination of the tabular data. One should be completely familiar with the data in order to set these preferences. For example, if the spelling sensitivity parameter is lowered in a city that has two streets named Mill and Miller, inaccurate geocoding may result.



Step Three: Geocode

The common approach [in geocoding] is to place the address point along the street within the range of the street number.

Once the parameters are set, geocoding is just a push of a button that starts the GIS software application's geocoding process. Often, software packages use a fairly simple method to accomplish the actual placement of the tabular data in relation to the reference layer. The common approach is to place the address point along the street within the range of the street number. For example, this first table highlights one call for service at the address "292 Princeton Ave."

Shape	Eq	Str_name	Str_apf	Date	District	Grid	Call_cod	Priority
Point	37001	292 PRINCETON AVE		960622	S	1112	G1300	5
Point	21002	24 VIRGINIA AVE	F-BSMT	960813	S	1620	G1300	5
Point	09022	133 MARTIN LUTHER KING DR		960801	S	1603	G1300	0
Point	25002	227 MARTIN LUTHER KING DR		960604		1625	G1300	5

The following table depicts the data for the street reference layer and highlighted is the street segment in which "292 Princeton Ave" falls.

Shape	Street	Fromleft	Toleft	Fromright	Toight
PolyLine	PRINCETON AVE	289	315	290	316
PolyLine	E 53RD ST	33	65	34	66
PolyLine	ADLER ST	1	41	2	42
PolyLine	ALBERT PL	1	29	2	30
PolyLine	ARLINGTON AVE	23	71	24	70
PolyLine	ARLINGTON AVE	73	101	72	100

When an incident is geocoded, the point is placed relative to a proportioned length derived from the ratio of the difference between the maximum and minimum address values and the physical length of the line segment.



For example, a segment may represent a street that is 100 yards long with an address range of 101 to 151 for the left side. An address of 125 would be placed approximately halfway down the line on the left side. Although this method usually misses exact real world coordinates, it is often sufficient given the detail of the street file and necessary resolution of the final map. In this example, "292 Princeton Ave" is located near the beginning of the street segment.



As noted earlier, geocoding by street centerline may not be adequate to depict the exact location of an incident. In the case of mapping a few blocks of residential burglaries, one may want to geocode by parcel address. In some cases, such as recording traffic citations along routes that may not have readily discernable addresses, intersection address data can be recorded for later geocoding using street files. Intersection geocoding is often accurate to a higher resolution than street geocoding since there is no mathematical algorithm to define the physical location of the point generated; instead the point is placed at the center of the intersection of the appropriate line segments. However, even using intersection data to geocode is not completely accurate since, in reality, most incidents occurring at an intersection address do not occur exactly in the middle of where the two streets intersect.

...geocoding by street centerline may not be adequate to depict the exact location of an incident.



A similar geocoding strategy uses polygons as opposed to previously discussed street centerline files. This method is the same as geocoding specific address data in that a value from the tabular data must match that of a geographic reference layer. For example, calls for service data often have a grid, beat, or district field in addition to an address field. In this case, the incident data has a Census block number field.

<i>Shape</i>	<i>Str_name</i>	<i>Str_aprt</i>	<i>Call_cod</i>	<i>Priority</i>	<i>Census Block Number</i>
Point	237 MONTICELLO AVE		G13	7	340170006001
Point	95 STEGMAN ST		G1300	5	340170089000
Point	51 MERRITT ST	APT2C&E	G1320	3	340170056002
Point	576 COMMUNIPAW AVE		G1300	5	340170046002
Point	253 STEGMAN ST		G1300	5	340170043005

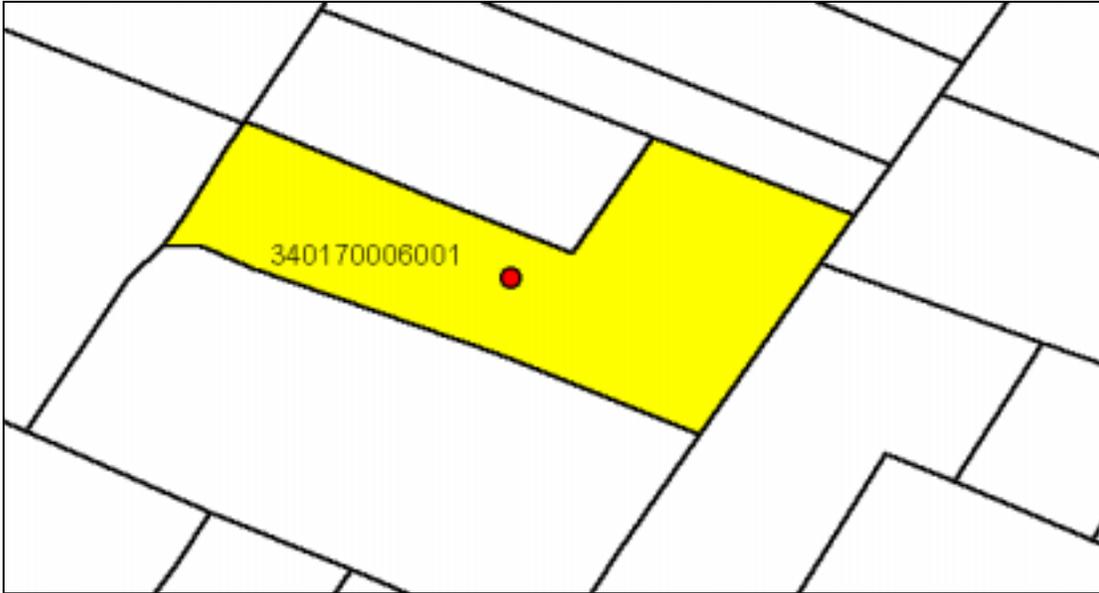
When geocoding polygons, the grid, beat, or district value is geocoded to the center of the polygon that represents the grid, beat, or district layer. In this example, the Census block group layer matches that from the incident data.

<i>Shape</i>	<i>Area</i>	<i>Census Block Group</i>	<i>Pop1990</i>	<i>Pop1999</i>
Polygon	0.12219	34017000600	5227	5521
Polygon	1.38054	34017000901	0	12
Polygon	0.17121	34017000100	5396	5633
Polygon	0.10119	34017000200	4715	4681
Polygon	0.11900	34017000300	4071	4187
Polygon	0.17429	34017000400	3467	3509

\



Once the incident is geocoded, the point is placed in the exact middle of the polygon, as are all the points geocoded to that polygon.



...many people geocode by address first and aggregate the data by polygon after the fact to avoid losing address data.

This type of geocoding is often used to create graduated color thematic maps by aggregating the data by polygon since data geocoded by polygon no longer can be located by address. However, many people geocode by address first and aggregate the data by polygon after the fact to avoid losing address data. Other types of geographic polygon data that can be used as reference layers for geocoding are zip codes, census blocks, census tracts, county, state, and country. Geocoding data at the polygon level is particularly useful when a dataset is very large or when address information is not as reliable as the grid, beat, etc. information.

Step Four: Review results

Step four, reviewing results, is a matter of determining whether the geocoding process was successful and how successful it was. Most GIS software will have geocoding outcome statistics that indicate how many of the cases were successfully geocoded, how many are a partial match, and how many were not geocoded at all. Obviously, the ideal successful geocoding rate is 100%. Depending on the number of cases and the purpose of the analysis, a rate of 95% can also be acceptable. The important factor in accepting a geocoding rate less than 100% is understanding why some of the incidents are not geocoded. In some cases, missing data or data outside



the mappable jurisdiction make up the incidents that are not geocoded. However, many times, there are other reasons that can and should be corrected.

In the case of a street centerline layer, geocoding problems are generally manifested in the form of missing or incorrect address ranges or missing street segments. Street layers often lack address ranges for streets because of recent construction of new roads or errors in the original creation process of the reference layer. The quality of these layers will depend both on those maintaining the layers and the growth rate of the jurisdiction. For example, in a small town that has not added new streets in fifty years, the street reference layer would be easy to maintain, not only because nothing is being added but also because the town has relatively few streets to maintain. On the other hand, in a large city that is growing rapidly, the street layer would be much more difficult to maintain.

The quality of these [geographic] layers will depend both on those maintaining the layers and growth rate of the jurisdiction.

Because street files are generally manually updated, they are also prone to data entry errors. The combination of these problems can be difficult to solve. Many agencies produce and maintain their own street centerline files or rely on their city or town's engineering department to maintain them, and others have opted to purchase street centerline files that have been corrected and are bundled with contracts for regular updates. However, for layers particular to the police department, such as police district boundaries or beat boundaries, most likely someone at the police department will be tasked to maintain these. Fortunately, these layers are typically composed of many less geographic components than a street reference file. In any case, interagency cooperation can prove invaluable as a means for resolving geocoding problems.

Even if the geographic reference layers are complete, there may still be problems with the tabular data used by law enforcement agencies, which stem primarily from having incorrect information or inaccurate data entry. Address errors in calls for service, crime, or accident data are often the source of much of the woes of geocoding in law enforcement and can be categorized into five common types: incorrect street numbers, street name errors, direction errors, incorrect intersections (in the case of intersection geocoding), and discrepancy errors.

- Street number errors can be the result of entry mistakes and are frequently identified as range errors; the address identified is either less than or greater than the address range of the corresponding records in the source file.



- Street name errors can occur for a variety of reasons and are a result of the address record not matching the street name in the source database. These errors are among the most common, with reasons including misspellings, abbreviations, inconsistent street types, or lack of compliance with city addressing standards.
- Direction errors occur when the direction code of the address record does not match the direction code in the source database, either because they are incorrect or missing.
- Unmatched intersections can result from inaccurate intersection addresses or incorrect format. For instance, the intersection of Baltimore and Eutaw is referred to as “Baltimore & Eutaw” and will not be recognized by many GIS software packages if entered as “Baltimore/Eutaw” or “Eutaw & Baltimore.”
- Other errors can occur because there is a discrepancy between the common name of a location and the official address. For example, a police agency may have in its computer aided dispatch system a street called “Route 123,” but the official name in the street centerline file may be “US HWY 123;” thus, they will not match.

Step Five: If necessary, respecify parameters and geocode again.

Step five is only necessary when there is not a 100% geocoding rate, which seems to happen most of the time. Once the data have been geocoded and problems have been determined by the geocoding process, there are a few solutions. Some may choose to take care of the problems by changing the preferences, e.g. lowering the spelling sensitivity. However, as noted earlier, this should be done cautiously and with full knowledge of the tabular and geographic datasets, as changing preferences such as spelling sensitivity can lead to incorrect address locations.

Another solution is developing alias tables, which allow for multiple names for a single entity. For example, a location such as a park or a mall may be referred to in a computer aided dispatch system as “Central Park” or “Centertown Mall” instead of by address to make them easy for dispatchers to enter. An alias table would replace the common name with the appropriate address in the GIS system so it can be located on the map. However, even though alias tables are helpful, it is still recommended that data be cleaned at the source to reduce confusion and facilitate consistency in data. Data are not only used for mapping purposes, so cleaning the data in the GIS would not solve data integrity issues when the data are used elsewhere.



The best way to correct data is to fix it at the source; in other words, bring the errors to the attention of those administering the computer aided dispatch system, to those maintaining the street centerline file, to the officers writing the reports, or to the clerks entering the data. However, this may not always be a realistic or a fast enough solution. In small projects or where a large number of cases share the same few mistakes, manual correction may be possible. Large datasets can often be corrected more efficiently when imported into a database or spreadsheet program (for example, Excel or Access) which has a “search and replace” utility. In addition, specialized software programs, referred to as “scrubbers” can be useful for large projects, and in instances where data are regularly imported.

Conclusion

These five steps outline the general process of geocoding law enforcement data. Because the types of data, problems, and solutions are common among law enforcement agencies, there is much to learn and share with one another. This overview has not detailed the specific methodology of geocoding in a particular GIS software program since the software used varies from agency to agency. For a specific outline of how to geocode in particular GIS software applications, see the text manual or electronic help files accompanying the software.

