Section 6 – Geocoding Report

STANDARD GEOCODING APPROACH

Geocoding is the process that converts addresses to locations on a map. The goal of legal services geocoding is to assign a geographic coordinate pair (Lat, Long), a zip code and a 2000 Census geography identifier to as many case records as possible while maintaining an acceptable level of accuracy and error. Geocoding also describes the success rate, the accuracy and confidence of the results. The project design calls for a maximum number of cases to be geocoded with minimal intervention, using 'off-the-shelf' commercially available software and data tools.

Geocoding Software

The geocoding software used for this project is MapInfo Corporation's commercial product MapMarker Plus. It is one of the most widely used professional-grade geocoders on the market. The MapMarker ESP configuration was used because it is optimized for Microsoft SQL Server, which is the database management system this project employs.

MapMarker Plus performs multiple tasks required in this project. First, it standardizes addresses to meet United States Postal Service Coding Accuracy Support Systems (USPS CASS) requirements. Second, it corrects and/or appends ZIP Codes and ZIP+4s. Third, it assigns map coordinates to street addresses, street intersections, ZIP Codes, ZIP+2s, and ZIP+4s. Fourth, it assigns street addresses and ZIP+4s to Census geographies.

Geocoders such as MapMarker Plus consist of two major components: the address dictionary, which contains all the data such as street addresses, ZIP Codes and map coordinates; and the geocoding engine, which is the software that matches an input address to the address dictionary. MapInfo issues quarterly releases to the MapMarker Plus engine and address dictionary, both to take advantage of fresh data and to meet USPS CASS requirements. For this project, the same version (7.0 dated July 2001) that was used for the original Georgia project was used. It features the July 2001 edition of the address dictionary, which MapInfo generates from the following three sources:

- Address data from the USPS ZIP+4 database, released July 2001
- Street geometry[1] from Geographic Data Technology's (GDT) *Dynamap-2000* product, released July 2001
- ZIP Code and ZIP+4 Centroids[2] from GDT, released July 2001

Geocoding Process

In geocoding, procedures are used to optimize the success rate of map coordinate assignment at each geographic level: ZIP Code, County, Census tract, Census block group and Census block.

1 The term "street geometry" refers to digital maps of streets, which is the source of the map coordinates used in MapMarker Plus.

2 Center point of a geography.

These procedures maximize the number of unbiased and useable case address geocodes. In addition, every geocoded case record is assigned a result code indicating the accuracy of the geocoder in resolving each case address to the geography.

Case Record Preparation

Geocoding systems like MapMarker parse an address into components: house number, street name, street type (e.g. ST), street prefix, suffix directionals like N or S, and unit numbers (e.g. apartment numbers). For this project, unit components like apartment numbers do not contribute to the assignment of a geocode or even its precision. Given two apartment units with the same street address, MapMarker would give both cases the same geocoded location because both are in the same building. Thus, unit number components add noise to the address that the geocoder must sort through when parsing and geocoding.

Consider the following examples:

Input Address	**Standardized Address and assigned Longitude/Latitude**
2100 Roswell Road, room 200c	2100 ROSWELL RD STE 200C -84.491330 / 33.966202
2100 Roswel Rd #200c (sic)	
2100 Roswell Road Suite 200	
2100 Roswell Rd Suite 500	2100 ROSWELL RD STE 500 -84.491330 / 33.966202

Despite the variations in the first three cases, MapMarker correctly interprets all three addresses the same way, and returns the same standardized address and geocoded location. The two distinct units, 200 and 500, get the same geocode. If the unit information were stripped off, MapMarker would have less data to parse, yet still return the same geocoding results. With less data to parse, MapMarker processes more records in less time. Moreover, with a simplified address, it is more likely to find a single close match in the address dictionary.

Prior to geocoding, the addresses are simplified by removing individual apartment unit information. Using standard SQL language, each address is scanned for a pattern and then appropriate adjustments are made. These patterns and adjustments fall into two categories: 1) remove unnecessary characters such as periods and commas; and 2) remove apartment and suite information. For the latter, the address is simply truncated from the point the apartment information begins.[3]

For example:

- Before scrubbing: 1800 Roswell Rd., Apt #18B
- After scrubbing: 1800 Roswell Rd

See Appendix 1 for details on the address simplification rules.

3 To avoid problems with addresses in which the apartment number is listed before the street name, the pattern search is started in position 8 of the address.

Case Record Organization

The case records are divided into 11 categories based on the contents of the address (See Table 6.1). Addresses with a house number and street name, plus either a city and state, a ZIP Code, or all three, are regular "residential addresses". The geocoding software can handle conventional addresses of this type.

Table 6.1 - Categories of Addresses	
Residential Address	
PO Box	Rural Route
In care of (C/O)	Correctional institution
Institutional or medical facility	Shelter
Homeless	Insufficient (e.g. missing house number)
Out of State	Not provided

Non-standard addresses are not nearly as common but are found among the case records. Large numbers of addresses with PO Boxes and Rural Routes were experienced in the Phase I Georgia project. An enhanced geocoding approach developed for Phase II will be used to assign these cases to Census geography. This process is documented separately in the Enhanced Geocoding Methodology document.

Addresses such as "homeless", "crisis center", out-of-state addresses, etc., are also expected. All these kinds of addresses require special handling. For case records that are missing or have ambiguous address information, the Grantees will be provided with an opportunity to update the addresses. See Appendix 2 for complete details on the address categorization rules.

Geocoding Strategies

MapMarker falls into the class of geocoders known as relative matching systems. MapMarker software compares the house number, street name, city, state and ZIP Code components of an address against candidate addresses (i.e. possible matches) in the address dictionary. For each address component, it internally computes a 'score' to measure the degree of closeness of the match, and then sums the resulting scores for each candidate. It then uses the total score for each candidate to determine the best match or matches. The geocoder makes an 'exact' match when a candidate is the best match. If it finds no clear best match, then it does not geocode the address.[4] At this point, the software assigns each address a result code with information about the quality and precision of the match. These result codes are discussed below.

Relative matching systems allow the definition of a match to be tailored by altering the relative importance or 'criteria' attached to each address component score (e.g. street name score). While this flexibility allows the geocoder to be tuned to meet the project requirements, but it does imply trade-offs. By relaxing the weights, the number of geocoded records is increased –

4 For more information, please refer to page 50 of the MapMarker Plus 9.0 Users Guide.

but the chance of a false match and therefore an incorrect geocode is increased. Yet if criteria are tightened too much, the chance of a false match is eliminated while the overall geocoding match rate is lowered.

For legal services mapping, a strategy is selected that strikes a balance between the goals of assigning correct Census geographies while maximizing the number of geocoded case records, all the while employing only automated processes. This strategy also enhances the ability to review and understand the source of the geocodes assigned to each record.

This strategy involves making multiple geocoding passes against the case records, with each pass attempting to geocode records not matched in the previous pass. Each pass uses a different set of criteria for determining when a case address matches a record in the address dictionary. With the first pass, the most restrictions are applied, requiring an exact match on all address components. Such a scheme virtually eliminates the chance of false matches. Subsequent passes will relax the criteria to allow more records to be geocoded, while still minimizing the chance of false matches. By marking each case record with an indicator of the pass, how each record was geocoded thus recorded.

Five geocoding passes are used in this project. The first pass includes the most restrictive rules for a successful match. Each subsequent geocoding pass loosens the restrictions slightly to allow more matches. With every pass, the geocoder is always attempting to match the case address to a street address in the address dictionary. Only after all the possibilities for a street address match are exhausted, is MapMarker permitted to accept Postal ZIP Code matches in lieu of street address matches.

First Pass (assigned to Census Block):

- Tune the geocoder to require exact match on house number, street name, city name and ZIP Code.
- Do not let MapMarker choose a candidate when it finds multiple matching candidates in the Address Dictionary.
- Accept only street address based geocodes; reject Postal code (e.g., ZIP+4) based geocodes.
- If MapMarker cannot make a match in the Address Dictionary on the street address, then do not fall back to the ZIP Code or ZIP+4 centroid.

Second Pass (assigned to Census Block):

- Require an exact match on house number.
- Relax the requirement for an exact match on the street name, thereby allowing for spelling mistakes.
- Relax requirement for an exact match on the city name and ZIP Code, thereby allowing for missing, incorrect or out of date ZIP Codes.
- Do not let MapMarker choose a candidate when it finds multiple matches. Accept only street address based geocodes. In addition, do not fall back to ZIP Code and ZIP+4 centroids.

Third Pass (assigned to Census Block):

- Require exact match on house number.
- Relax requirement for an exact match on the street name, city name and ZIP Code.
- If MapMarker finds multiple, close matching candidates in the Address Dictionary, allow it to accept the first of the multiple matches. If the candidates consist of a mix of street address based geocodes vs. ZIP+4 based geocodes, then have MapMarker use the street address based geocode.
- Do not fall back to ZIP Code and ZIP+4 centroids if MapMarker cannot make a street address match.

Fourth Pass (assigned to best Census possible level):

- Require exact match on house number.
- Relax requirement for an exact match on the street name, city name and ZIP Code.
- Fall back to ZIP Code, ZIP+2, or ZIP+4 based centroids, allow MapMarker to choose the best available.

Fifth Pass (assigned to best Census possible level):

- Geocode to ZIP Code, ZIP+2, or ZIP+4 based centroids for case records without street addresses.

To illustrate the impact of these geocoding rules, MapMarker is used to geocode a collection of two test addresses with multiple variants. Table 6.2 shows the geocoding result for each variant, and the number of passes that were required for a successful match. Pass #1, the most restrictive of the rules, requires a complete and correct address. MapMarker would geocode two of the addresses in this test using these rules. Pass #2 allows some errors and missing data but still geocodes the records correctly. Pass #3 relaxes the rules further, allowing for less precise addresses. Finally, passes 4 and 5 allow MapMarker to fallback to Postal centroid based geocodes.

Table 6.2 – Example Geocoding By Pass				
Original Address City, State, ZIP Code	Success Pass	Geocoding Result Code	Assigned Census ID	Assigned Longitude / Latitude
15 Glynn Ave **Jekyll Island, GA 31527**	1	S5HPNTSCZA	13127-001000-5044	-81.408022 / 31.070307
15 Glynn *Rd*, Jekyll Island, GA	2	S5HPN-SC-A	13127-001000-5044	-81.408022 / 31.070307
15 *Glenn* Ave 31527 (missing city and state)	2	S5HP-TS-ZA	13127-001000-5044	-81.408022 / 31.070307
15 *Glenn* Ave, Brunswick, GA (alias for Jekyll Island)	2	S5HP-TS--A	13127-001000-5044	-81.408022 / 31.070307
15 *Gynn* Brunswick, GA 31527	4	Z1	13127	-81.41022 / 31.066280
500 N Beachview Dr, **Jekyll Island, GA 31527**	1	S5HPNTSCZA	13127-001000-5037	-81.405701 / 31.061812
500 Beachview Dr **Jekyll Island, GA 31527** (missing directional)	3	M5H-NTSCZA	13127-001000-5037	-81.405701 / 31.061812
500 ***S*** **Beachview Dr** Jekyll Island, GA (no ZIP Code)	2	S5HPNTSC-A	13127-001000-5055	-81.417601 / 31.030989
31527 (ZIP Code only, no address)	5	Z1	13127	-81.41022 / 31.066280

(Addresses in **bold** are valid. Mistakes in the entered addresses are in *italics*.)

Interpreting MapMarker Result Codes

MapMarker assigns a result code to every case record it geocodes. The result code indicates the success or failure of the geocoding process, and conveys information about the quality of the match and the precision of the assigned geocode. Result code values consist of 10 characters, each of which indicates specific information about the match.

Result Code Categories

To help interpret the codes, MapMarker divides them into three major categories: geocodes based on a single close match, those based on the best match from multiple candidates and those based on postal centroids. The first character of the result code indicates the category, with values of S, M and Z respectively. See summary table below.

The result codes fall into three major categories, as indicated by the first letter of the code.

- S: The geocoding software found a single close match for the case record to a record in the software's Address Dictionary. This is the best possible result.
- M: Best match out of multiple candidates, meaning that the case record matches against several records in the address dictionary. (This is not an unusual occurrence because it is often the case that the address dictionary will contain a record for the street address plus a separate record for the ZIP+4 of the same street address.)
- Z: ZIP Code centroid match, which means the address of the case record did not match against anything in the address dictionary but the ZIP Code or ZIP+4 did match.

Result codes in the category of S means that MapMarker found a single best match for the address in its address dictionary. In other words, the particular combination of the house number, street name, city, state and ZIP Code exists in the address dictionary, thus confirming a valid address in terms of Postal standards. Furthermore, MapMarker was able to match the address to a single best candidate in the dictionary, thus virtually eliminating ambiguities in the match.

When MapMarker assigns a result code of M, it was able to match the address to multiple candidates in the address dictionary, but no one candidate stood out as the best match. Perhaps the address was missing the directional. If for example MapMarker were presented with 1 Main St when in fact 1 N Main St and 1 S Main St were valid, MapMarker would not know which to pick.

If MapMarker cannot resolve the house number and street address to any candidate in the address dictionary, but it can resolve the Postal information, either the ZIP Code or ZIP+4, then it can geocode the address to that level. In these cases, the result code begins with a Z.

Positional Accuracy

The second character indicates the positional accuracy of the geocode, specifically the map coordinate assigned to the address. Ideally, geocodes would always be based on street addresses, but this is not always possible. When an address level geocode is not available, then MapMarker can default to ZIP Code, ZIP+2 or ZIP+4 centroids.
In the best case, MapMarker assigns a map coordinate calculated from the street address. Considering only the first two characters of the result code, MapMarker would assign a result code of "S5" when it calculates the geocode at the street address level. S5 geocodes are the most desirable because they would be closest to the actual location. If MapMarker can match the address to the address dictionary but the dictionary contains only a ZIP+4 level geocode, then it would assign a result code of S3. If it cannot find the address in the dictionary, but it does find the ZIP Code, then it would assign a result code of Z1. (See Table 6.3).

Table 6.3 – Geocoding Result Codes		
Category	Positional Accuracy	Location of map coordinate
(1st character)	(2nd character)	
S, M		S = Single address match, M = Multiple address candidates match
	1	Centroid of a residential ZIP Code
	2	ZIP+2 centroid
	3	ZIP+4 centroid
	4	Center of the street segment
	5	Street address position
	6	Centroid of a unique (point) ZIP Code
	X	At street intersection
	0	No coordinate available (which is rare)
Z		Z = ZIP Code match
	1	ZIP Code centroid
	2	ZIP+2 Centroid
	3	ZIP+4 Centroid
	6	Centroid of a unique (point) ZIP Code
	0	No coordinate available (which is rare)

Address Dictionary Match Codes

The remaining eight characters of the result code describe how closely the components of an input address matched an address in the address dictionary (See Table 6.4). The characters appear in the result code in the order given in the following chart. Dashes in the result code string represent any non-matched components.

Table 6.4 – Address Match Codes		
Component	Description	Example
H	House number	
P	Street prefix	West
N	Street name	Pike
T	Street type	ST
S	Street suffix	NW
C	City name	Clarksburg
Z	ZIP Code	26301
A or U	Address Dictionary Used [5]	A

5 An "A" indicates the match was found in MapMarker's address dictionary as opposed to a user specific, custom dictionary, as indicated by a "U."

For example, if MapMarker exactly matches 3333 K ST NW, Washington DC, 20007 to a candidate in the address dictionary and to a street address based map coordinate, it would return a result code of S5HPNTSCZA.

Postal Centroids

How ZIP Code centroids are positioned requires some explanation. Residential ZIP Codes typically have boundaries. For these ZIP Codes, the centroid is weighted towards the population center of the ZIP Code. In certain circumstances, special ZIP Codes can be specific to a PO Box or single location such as a large apartment building. If the address of the Post Office or location is both known and map-able to the street address level, then the centroid relates to that specific location. Otherwise, the centroid of the enclosing boundary (geographic center) ZIP Code is used.

In Postal terms, a ZIP+4 usually corresponds to a collection of house numbers along a street, typically less than 10. The ZIP+4 centroid represents the center point of the section of the street on which these houses are found. If two houses belong to the same ZIP+4, then both would have the same ZIP+4 centroid. In addition, a ZIP+4 could represent a collection of apartment units in a particular building. The centroid would represent the address of the building and all units in the same building would have the same ZIP+4 centroid.

Census Geography Assignments

The quality of the match and the geocoding precision determines the level of Census geography (e.g. Tract, Block Group, Block, etc.) that MapMarker can assign. MapMarker will assign to a Census Block only those records geocoded with the highest precision (See Table 6.5). Those with a lower precision are assigned to a Census Block Group, Tracts or even only a County if the precision is low.

Table 6.5 – Census Assignments

Highest Census Geography	Required Result Code	Comments
Census Block (15 character Census ID)	S5	Best outcome
Census Block Group (11 character Census ID)	S3	An S3 result was matched to a ZIP+4 centroid, which will fall inside a single Block Group. However, it is possible that the ZIP+4 segment will cross into another Block Group.
Tract or County	S1, S2 Z1, Z2, Z3	It is also possible that MapMarker will not return any Census geographies for these ZIP Code and ZIP+2 centroids.
None	S4	

Conclusion

This geocoding strategy maximizes the number of case addresses that can be geocoded to the Census geographies. It provides detailed tracking information on each case record, allowing us to verify the geocoding results, both overall and at the individual case level. Moreover, it does all this entirely with automated processes, thereby forming the foundation for a repeatable system.

Geocoding Summary

A geocoding summary with geocoding results tables will present the geocoding statistics separately for each grantee as well as a table showing the geocoding statistics for the overall project. These summaries are shown in Appendix B.

Appendix 1

Address Simplification Rules
(In SQL language)

Remove “unit” data from address (code snippet only)

```
Case
        When PATINDEX ('%Apt%', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Apt%', Street)-1)

        When PATINDEX ('%Suite%', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Suite%', Street)-1)

        When PATINDEX ('%Bldg%', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Bldg%', Street)-1)

        When PATINDEX ('%Building %', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Building%', Street)-1)

        When PATINDEX ('%Floor %', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Floor%', Street)-1)

        When PATINDEX ('%Room %', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Room %', Street)-1)

        When PATINDEX ('%Space %', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Space %', Street)-1)

        When PATINDEX ('%Unit %', Street) > 7 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, PATINDEX ('%Unit %', Street)-1)

        When CHARINDEX ('#', Street, 7) > 0 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, CHARINDEX ('#', Street, 6)-1)

        When CHARINDEX ('(', Street, 7) > 0 AND IsNumeric(left(LTrim(Street),1)) = 1
                then left(Street, CHARINDEX ('(', Street, 6)-1)

        Else LTrim(RTrim(Street))
End,
```

Remove periods

```
Update Case_Geocodes
        Set Street_in = Replace(Street_in, '.', Space(1))
                where CHARINDEX ('.', Street_in, 1) > 0
```

Remove commas

```
Update Case_Geocodes
        Set Street_in = Replace(Street_in, ',', Space(1))
                where CHARINDEX (',', Street_in, 1) > 0
```

Appendix 2

Rules for Categorizing Address

(The "search string" is used in a SQL "where" clause, as in "Where Street like @Search_string.")

Address Category	Search String
PO Box	%PO % %POB % %POBox % %P O % %Post Office% %PMB% %General Delivery%
Rural Routes	%Rural Route% RR % % Rt % % Rte %
C/O	%C/O%
Correctional	%jail% %correctional% %sheriff% %prison% %detention%
Institutional	%nursing% %hospital% % hosp % %Convalescent% %Health% %Retirement% %personal care% %Veterans% %VA % %Assisted% %Treatment% %Long Term%
Shelter	%crisis% %rescue% %shelter% % army % % house% %YMCA% %YWCA% % motel% % hotel% %confidential%
Homeless	%Homeless% %no home% none% no address%
Not Provided	%Unknown% n/a% na% enter% ?%