Tufts GIS Tip Sheet

Geocoding Overview and Preparation

Overview

Geocoding is a process is the process of assigning locations to addresses to that they can be placed as points on a map, similar to putting pins on a paper map, and analyzed with other spatial data. The process assigns geographic coordinates to the original data, hence the name geocoding. It is also called address-matching. In a typical geocoding process, the data list might include an address like 508 W. 5th St. and a street centerline GIS data layer would have a street segment corresponding to the 500 block of West 5th Street. The result of geocoding would be a point placed somewhere along the even-address side of that street segment.

The geocoding process is described in **ArcGIS Desktop Help** (on the *ArcGIS Help* menu) under **Contents – Geocoding and Address Management**, and you should refer to that for details. This tip sheet provides an overview of the preparation process, which is very important and a bit tricky.

Note: if your data list already has x and y coordinates of some sort, you don't have to geocode - those *are* the geocodes. You can add them to a map by using the **Tools - Add XY Data** menu function in ArcMap.

Two sets of data are needed for the geocoding process - the **address data** that you want to place on a map, e.g., a list of addresses, and the GIS data layer that you will use as the geographic **reference layer**, e.g., a city's street centerlines layer or a parcel address point layer. Both data sets need to be prepared prior to geocoding.

About Address Data

Preparation of the address data set means formatting the information correctly so that a GIS software like ArcGIS can process it (called parsing). The data set should be in a database-compatible format like comma-delimited (.csv) or tab-delimited (.tab) text file or dBase (.dbf) or (if using ArcGIS 9.2) in an Excel worksheet. The address should be contained in a **single column** that contains the street number and street name, as well as the street's prefix direction (e.g., W.), street type (e.g., Blvd.), or suffix direction (e.g., W.), if any. Intersection descriptions (for example, "MLK Blvd. & Vine St.") can also be included in this field. (See the Proj_Add column in the database example on the next page.)

The next page shows an example of a table containing address information (from a US Department of Housing and Urban Development database of properties that received a low income housing tax credit). Note that several of these addresses may be unmatchable. Some records have no address, and some have multiple numbers in a single record (see the last row – “723 801 803 Pebble Beach”). Some have a street but no number. Some, like “West of Hwy 281” give a very generalized location. Some addresses are actually intersections, which is fine, except that the intersections are not consistently represented – in some cases there is an “& between street names (“State Highway 16 & Zanderson”), while others seem to have a “/” (“Hampton/J St”) or nothing at all (“SUNNYSIDE RD US HWY 181”) This example from Texas also includes road types found only in Texas (FM 81 stands for “Farm to Market Road 81”) which are not handled well by the standard geocoding functions found in ArcGIS.

HUD_ID	PROJECT	PROJ_ADD	PROJ_CTY	PROJ_ST	PROJ_ZIP
TXA1987033	CROSSWINDS APARTMENTS	200 METHODIST DR	COMMERCE	TX	75428
TXA1987034	DEKALB APARTMENTS	900 W FRONT ST	DE KALB	TX	75559
TXA1987035	DUPLEX AT 3905 CHASE CIRCLE	3905 CHASE CIR	AUSTIN	TX	78721
TXA1987036	EAGLE LAKE GARDEN VILLAGE	1300 VILLAGE GARDEN DR	AZLE	TX	76020
TXA1987037	FLORESVILLE SQUARE APTS	SUNNYSIDE RD US HWY 181	FLORESVILLE	TX	78114
TXA1987038	GOLDEN HELMET APTS	2121 52ND ST	DALLAS	TX	75216
TXA1987039	GREGORY PLACE	912 CHURCH	GALVESTON	TX	77550
TXA1987040	HAMPTON HEIGHTS	HAMPTON/J ST	SAN ANTONIO	TX	78200
TXA1987041	HARRINGTON ROAD 1	FM 2560 & HARRINGTON RD	SULPHUR SPRINGS	TX	75482
TXA1987042	HARRINGTON ROAD 2	FM 2560 & HARRINGTON RD	SULPHUR SPRINGS	TX	75482
TXA1987043	HARRINGTON ROAD 3	FM 2560 & HARRINGTON RD	SULPHUR SPRINGS	TX	75482
TXA1987044	HETH APTS		LINDEN	TX	75563
TXA1987045	JACKSONVILLE SQUARE LTD	1200 ANDREWS ST	JACKSONVILLE	TX	75766
TXA1987046	JOURDANTON SQUARE	STATE HWY 16 & ZANDERSON	JOURDANTON	TX	78026
TXA1987047	LA ALAMEDA		SAN BENITO	TX	78586
TXA1987048	LA ALAMEDA SUBDIVISION	WEST OF HWY 281	PHARR	TX	78577
TXA1987049	LAVACA LANDING APARTMENTS	RT 4	HALLETTSVILLE	TX	77964
TXA1987050	LOS FRESNOS APARTMENTS	PT ISABEL BROWNSVILLE ST	LOS FRESNOS	TX	78566
TXA1987051	LYTLE APARTMENTS	FM 81 & I 35	LYTLE	TX	78052
TXA1987052	MAGNOLIA PLAZA		MAGNOLIA	TX	77355
TXA1987053	MESQUITE WOODS		HITCHCOCK	TX	77563
TXA1987054	NEW CANEY OAKS APTS	US 59	NEW CANEY	TX	77357
TXA1987055	OLD CATHOLIC DIOCESE BUILDING	1923 MARKET ST	GALVESTON	TX	77550
TXA1987056	ORANGE GROVE APARTMENTS	RT 1	ORANGE GROVE	TX	78372
TXA1987057	PERKINS		MISSION	TX	78572
TXA1987058	RANGERVILLE PARK SUBDIVISION	RANGERVILLE RD	HARLINGEN	TX	78552
TXA1987059	REFUGIO APARTMENTS	405 OSAGE ST	REFUGIO	TX	78377
TXA1987060	REGENCY APARTMENTS	1100 ALCOA DR	PORT LAVACA	TX	77979
TXA1987061	SCHNEIDER APTS	120 S RUSSELL	PAMPA	TX	79065
TXA1987062	SOMERVILLE PLAZA	SOMERVILLE RD	SOMERVILLE	TX	77879
TXA1987063	SPRINGHILL APARTMENTS	4830 RAY BON	SAN ANTONIO	TX	78218
TXA1987064	STUART APTS	COUNTRY PARK/LANGSTON	MONT BELVIEU	TX	77580
TXA1987065	TAFT GARDENS LTD	HWY 181	TAFT	TX	78390
TXA1987066	TAFT TERRACE	HWY 181	TAFT	TX	78390
TXA1987067	TIERRA DORADA SUBDIVISION		MISSION	TX	78572
TXA1987068	TOWN OAKS APRTMENTS	120 KENNEDY ST	KENEDY	TX	78119
TXA1987069	TRAILS	723 801 803 PEBBLE BEACH	GARLAND	TX	75043

This table is, in fact, a good example of many of the issues that come up when dealing with addresses. Ideally you should have as complete an address as possible, but without apartment information. The geocoding process will likely be able to handle missing street types like St. or Blvd., but it will go easier the more complete your address information is. Apartment information should go in a separate field if you need to maintain it for your information purposes - it will not be part of the geocoding process. It is also useful to have city, zip code, and state in your address table, plus all the attributes you need for a particular project (e.g., for a grocery store table, name, chain, type, size, annual sales, etc.). At a minimum, include the zip code if possible – this will be a big help.

The more carefully you format your address list with geocoding in mind, the better the geocoding process will work, so take some time doing this, and plan ahead if you know you will be geocoding. It is important that your data formatting is consistent throughout the database. If you include intersections for some addresses instead of street numbers, always use the same connector (e.g., &) and use the complete street names (e.g., Burnet Rd. & W. 51st St, not Burnet & W. 51st.). Make your zip code information a text (string) field – if they are numeric, leading zeros will be lopped off (same goes with telephone numbers or ID numbers like FIPS codes). Other tips: for your column names, follow dBase compatible formats - no spaces or odd characters in the field name, and a maximum of 10 characters.

If you are using a spreadsheet to create this data set, make the first row the field names, and start your actual address records on the second row. Do not put in other formatting or rows or columns, e.g., no titles, or spacer rows. Only enter the field names and actual data records. ArcGIS 9.2 can read Excel files directly. If you are using an earlier version of ArcGIS, you must save the file to a comma-delimited text file (.csv) or dBase file (.dbf).

About Reference Data

To geocode data, you must have a GIS reference layer available to act as your reference layer. Your choice of reference data is very important and will affect the accuracy and completeness of your results. Street centerlines are often used as a reference layer. Well-formatted street centerline GIS data layers have separate fields for street name, street's prefix direction, street type, and suffix direction as appropriate (some streets don't have suffix or prefix directions). They will also have four address range fields indicating *From address left* and *To address left* (e.g., 1100 and 1122), *From address right* and *To address right* (e.g., 1101 and 1123). This address range is what allows an address to be pin-mapped onto the street network by indicating house numbers on both sides of the street. If you need to match by both address and zip code, your reference layer should also have fields for zip code on the left side of the street and zip code on the right side of the street. The TIGER road data is formatted in this way.

Errors can arise from several factors. First, the positional accuracy of the streets may be off - this is affected by source scale as well as digitizing error. Also, the accuracy of address information in the street attribute table may be wrong or limited. Imagine the address you are geocoding is 1214 E. 44th St. Some street centerline files (e.g., TIGER) often give "generic" ranges in the attribute table, for example1200-1298 on the even side of a street and 1201-1299 on the odd side. Using this kind of data set, the point will be placed near the beginning of the street on the even side. But if the street address range is in reality is 1200-1214 and 1201-1215, the point for 1214 should in fact be placed at the opposite end of the street.

One easily available source for street centerline GIS data layers for any area in the US is the US Census' TIGER roads data set. These can be downloaded from the Census Bureau itself (http://www.census.gov/), or from the Geography Network (http://www.geographynetwork.com/freeresources.html) - the Geography Network has a more user friendly interface. Note that TIGER roads files are downloaded county by county, so you may have to merge files to create a reference data layer for more than one county (to merge, in ArcMap, add all the data layers to be merged and then open *ArcToolbox – Data Management Tools – General* and choose *Merge*).

But the TIGER road data is digitized from 1:100,000 scale maps and are often less accurate than road data produced by local government agencies. In the past the address ranges have also been generic for many localities, although this is improving with new releases. Look at the following examples from Houston, Texas. Figures 1 and 2 illustrate different geocoding reference files in Houston, focusing on a suburban area west of downtown as an example. In Figure 1 below, the street centerlines from the TIGER/Line files appear in yellow, while the streets from the Greater Harris County (GHC) 911 network appear in black. The TIGER/Line streets in this area may be anywhere up to 300 meters off, they frequently do not represent the true shape of streets and blocks, and they are missing in some cases compared to the aerial photo and the GHC-911 street centerlines. The Census TIGER roads data was created for purposes of aiding the decennial census, while the GHC 911 street network was developed to aid emergency dispatchers. It is thus not surprising that the GHC 911 streets appear more accurate than the Census Bureau data. It seems obvious that using the GHC-911 street centerline data would give a more accurate geocoding result. Figure 2 shows parcel point data for the same area – addresses geocoded to these points would be even more accurately located. But these type of data sources may not always be available or be the best choice – advantages and disadvantages of some different reference data sources are listed in Table 1.

FIGURE 1 – COMPARISON OF TIGER/LINE ROADS (yellow) AND GHC-911 STREET CENTERLINES (black)	FIGURE 2 – PARCEL ADDRESS POINTS WITH GHC-911 STREET CENTERLINES
	

TABLE 1 – ADVANTAGES AND DISADVANTAGES OF GEOCODING REFERENCE FILES

Parcel Address Points

Advantages	Disadvantages
• Typically allows more accurate placement of residential location than street centerline geocoding (parcel positional data is often very good, e.g., +/- 5 meters or less) • If owner name is present, may allow a validity check	• May need to contact individuals within agencies to get most up to date data • May not be available, or may cost a substantial amount of money • Address data may not be formatted in a way that directly fits standard GIS geocoding capacities • Data ends at jurisdictional boundaries • Data files tend to be very large

Street Centerlines from Local Jurisdictions

Advantages	Disadvantages
• Potential to be more up to date (often yearly updates, sometimes quarterly) • Often adequate accuracy to meet city infrastructure needs (typically +/- 10 meters or less)	• May need to contact individuals within agencies to get most up to date data • Accuracy often not documented • Streets often end at jurisdictional lines that don't match study boundaries • Street formatting may not match standard GIS geocoding capabilities • May not support topological network analysis • May not have address information to support geocoding (e.g., neither MassGIS nor the City of Boston provide address ranges in their road centerline GIS data)

TIGER/Line Street Centerlines (US Census Bureau)

Advantages	Disadvantages
• Uniform across jurisdictional lines and nationally • Street address formatting works well with standard GIS geocoding capacities • Available online for free download • Robust database design, tested, uniform, supports topological network analysis	• Not up to date • Digitized from 1:100,000 scale maps originally – positional accuracy varies widely but +/- 300 meters is not unusual • Placement of address point is approximate

Note that there are a number of private data vendors that sell street centerline data for navigation and geocoding, as well as a number of private geocoding services that will allow you to process address data. All such sources and services should be carefully evaluated to see if using them results in the required accuracy for a project. In the Tufts GIS lab, we have a license to StreetMap USA, a street centerline file for the entire US. You should test the StreetMap USA product and carefully evaluate the results if you use it on a project.

You can also match your data only to a zip code if you desire. For statewide or nationwide data sets, the zip code may be the only information you have or mapping to a generalized zip code boundary may be good enough for your needs. In this case you will need some kind of zip code points (centroid - center point of a zip code) or polygon GIS layer to act as reference. The US Post Office which creates and maintains zip codes for mail purposes does not maintain this data for various reasons, but the US Census has an approximation of zip code areas that it calls *Zip Code Tabulation Areas* (ZCTAs) that you can download from the Census Bureau by state (http://www.census.gov/geo/ZCTA/zcta.html). But understand that these are only approximations for census purposes and do not reflect actual zip code areas and are not kept up to date. Note that many private data vendors also sell zip code GIS information and offer services for zip code mapping.

Preparing Reference Data by Creating an Address Locator

The GIS reference layer, e.g., a street centerlines or zip code polygon layer, needs to be prepared by creating what ArcGIS refers to as an "address locator". This process essentially indexes a reference layer, much like indexing a book. You create the **Address Locator** using **Arc Toolbox**. Note that if you are using the StreetMap USA street centerlines as your reference file, an Address Locator already exists in the same folder as the street data on the Tufts server. (M:\\Country\USA\ESRIDataMaps906\streetmap_USA\streets)

It is very important that you are familiar with your reference data layer before you create an Address Locator. You need to understand which fields contain the necessary information for creating an Address Locator – e.g., which fields contain street name, street type, street prefix direction, address ranges, zip codes, etc. You will map this information to an Address Locator Style in ArcGIS. There are several commonly used styles with which you should be familiar:

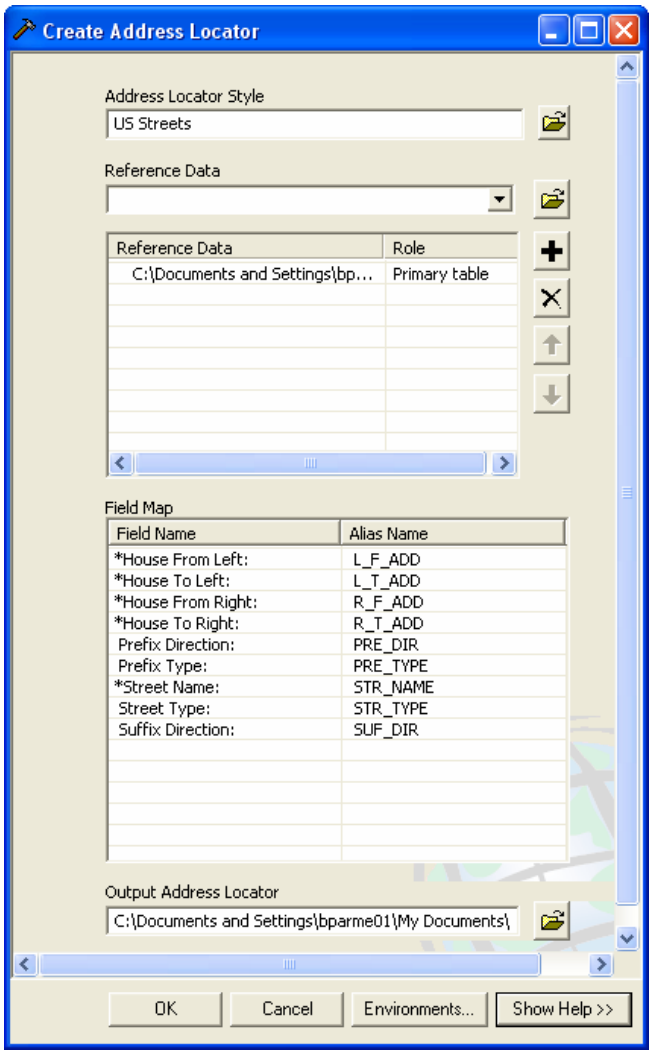
- US Streets – use this if you are matching addresses to a street centerline file with address ranges in a single town and you don't need or don't have zip code information
- US Streets with Zone – use this if you are matching addresses to a street centerline file with address ranges in a single town and you do have zip code information (always try to have zip code information)
- US Streets with City, State, Zip – use this if you are matching addresses to a street centerline file that includes multiple towns and/or states and you have city, state, and zip information in your address file
- US One Address – use this if you are geocoding using parcel address point or polygons as your reference file (choose US One Address with Zone if you have zip code information as well)

- Zip 5 Digit – use this if you ONLY have zip code information and you are geocoding to a zip code centroid or polygon reference layer

For more information about each style, go to ArcGIS Desktop Help and search for Address Locator Style.

To create an Address Locator for your reference data layer:

1. Open **ArcToolbox - Gecoding Tools** and double-click on **Create Address Locator**
2. Click on **Show Help** in the bottom left corner of the dialog box for context sensitive help (you can click on the *Help icon* in that box for detailed instructions)
3. Click on the folder next to *Address Locator Style* to choose a style.
4. Under **Reference Data**, use the pull-down or Folder icon to select the GIS reference data you wish to use (e.g., street centerline, parcel points, zip code polygons)
5. Under "**Role**," click and select "**Primary table**" or "**Alias**." Primary is the most common choice. Alias is used for when a place has an address (90 Congress St.) but is also known as something else (Government Center)
6. The **Field Map** table should then populate after you make your selection - the field map links required information to fields in the reference file's attribute table. Check to ensure everything is filled in correctly and fill in blanks as needed.
7. Under Output Address Locator, give the new file a name and store it on a drive to which you have write access (e.g., C on your personal computer: H, or P at the Tufts GIS Lab).



Note: If a red "X" appears at any point, don't panic. It just means that another step needs to be completed. It will also appear if you try to use the M:\ drive or other drive to which you don't have write access

Geocoding a list of addresses

Once you have prepared your address data and created a geocoding service using your reference GIS data layer, you are ready to do the actual geocoding. There are good instructions for geocoding in **ArcGIS Desktop Help** under **Contents - Geocoding and Address Management -Geocoding a Table of Addresses**. You should refer to these instructions for the rest of this process.

Typically, only some percentage of your addresses will actually find a match. Some will remain unmatched. For these, there is a re-matching process described well in **ArcGIS Desktop Help** under **Contents – Geocoding and Address Management - Re-matching a Geocoded Feature Class.** But before you do the re-matching process, you should spend time carefully examining the addresses that didn't match (indicated by a "U" in the status field of the geocoded results). There can be many reasons for a failure to match. The address in your list may be misspelled or be in a wrong format, or the street centerline file may have problems (e.g., be out of date, list a name that for a street that is different from the same street in your address list - e.g., MLK Blvd, instead of 19th St. or Martin Luther King, Jr. Blvd or I-93 N versus Interstate 93 North), or the address ranges may be incorrect or missing for a street segment.

Also note that TIGER files typically do not contain street address ranges for rural areas or small towns, thus addresses in these areas cannot be matched against the TIGER files.

You should also check the addresses that did match. They may have matched incorrectly for various reasons. You always need to do a data check for any processes that you perform in GIS!

Note that when there are multiple records for a single address (or zip code), all the points will be placed one on top of the other at that point. It will look like just a single point, but if you click on it with the information tool, you will see all the records come up. If you select it with the selection tool, all the records will be selected. If you summarize or join the geocoded data, all the records for that point will be processed.