Using SAS to Geocode Injury-Related Deaths in North Carolina

Authors: Bruce Nawrocki, Scott Proescholdbell, Shana Geary and Mike Dolan Fliss

BACKGROUND: Person, time and place are core epidemiological concepts. In North Carolina (NC), historical geographic units of analyses have been at the state and county levels, yet smaller geographic units might enable a better understanding of the patterns of injury deaths and a more robust response. This GIS project's purpose was meant to explore other geographic units available in NC that had not been utilized for injury surveillance before, such as census tracts and block groups, and political boundaries (US congressional districts, NC senate and NC house districts), and develop a production process to handle these in a standardized way.

METHODS: As input to our geocoding process, we used our NC death certificate files, examining the resident addresses of injury-related deaths. We wrote a series of data preparation and processing tasks in SAS code, incorporating them into a single process flow in SAS Enterprise Guide (EG). The tasks include geocoding with PROC GEOCODE to assign latitude and longitude values to addresses, then use PROC GINSIDE to assign each street-level geocoded death record into specific geographic units such aa census tract and block group, and US congressional districts. This allowed us to analyze and map injuries using these smaller geographic units, which allowed us to determine, for the first time, which areas experience the highest rates of injury death.

RESULTS: When we attempted to geocode 10,473 injury deaths for 2021, we were able to successfully match 91.4% of all cases at the street-address level, which provided us with additional geographical units, including census block/block-group/tract. A wide range of sub-state analyses were completed utilizing the new units. For example, we were able to determine which of NC's US congressional districts had the highest and lowest rates for unintentional drug overdose. We also compared rates between different census tracts and NC senate and house districts. Additional analysis investigated differences in demographics (sex, age group and race/ethnicity) within each of these geographic areas.

Details

For the purposes of this paper, we will focus on converting street addresses to geographic locations – that is, X (latitude) and Y (longitude) values, using the default World Geodetic System (WGS84). This allows us to plot these locations on a map. With a small amount of additional processing, we can aggregate these points into specific geographic areas, such as state congressional districts. With this information, we can determine which districts have the highest injury death rates overall and for specific causes of injury, such as violence, overdose, or falls.

We decided to geocode our data using SAS, since our data was all stored in SAS, we had already licensed base SAS, and we had some in-house SAS expertise. We started with a SAS dataset containing street address fields, then generated a new dataset adding X and Y values for those addresses using PROC GEOCODE, included in base SAS software.

Using PROC GEOCODE, you can find the X/Y values for various geographic areas, such as city, ZIP code and street address. Here's a summary:

GEOCODE can find X/Y for…	**Using this lookup dataset…**
US City (X/Y value of city center)	MAPSGFK.USCITY_ALL (included with Base SAS)
5-digit ZIP (X/Y value of ZIP center)	SASHELP.ZIPCODE (included with Base SAS)
U.S. Street Address (X/Y value of street address)	Base SAS provides a sample one for Wake County: SASHELP.GEOEXM For all US addresses, visit link[1] to get US address lookup datasets.
ZIP+4 codes	LOOKUP.ZIP4 (downloadable from link[1])

Create a Sample Address Dataset

Let's start with a sample SAS dataset containing street addresses (this is based on example SAS code[2]).

```
data addresses;
      infile datalines dlm='#';
      length address $ 30 city $ 25 state $ 2;
      input address zip city state;
      datalines;
123 Junk Street # 99999 # New York # NY
2630 Battleground Ave # 27408 # Greensboro # NC
940 NW Cary Parkway # 27513 # Raleigh # NC
5505 Six Forks Road # 27609 # Raleigh # NC
;
run;
```

Geocode ZIP Code Values

If you run PROC GEOCODE with no options, SAS will geocode based on ZIP Code values:

```
proc geocode data=addresses out=geocoded_byZIP;
run;
```

If we print the new dataset "geocoded_byZIP", we see this:

Y	X	M_OBS	_MATCHED_	address	city	state	zip
40.7142	-74.0064	85948	City	123 Junk Street	New York	NY	99999
36.1064	-79.8174	10720	ZIP	3003 Branchwood Dr	Greensboro	NC	27408
35.8054	-78.7977	10753	ZIP	940 NW Cary Parkway	Raleigh	NC	27513
35.8417	-78.6325	10829	ZIP	5505 Six Forks Road	Raleigh	NC	27609

The new fields added by PROC GEOCODE include:

- coordinate values Y (latitude) and X (longitude) from lookup dataset SASHELP.ZIPCODE
- M_OBS – the observation number (row) matched in lookup dataset
- _MATCHED_ - highest level match category.

Note that all rows were matched on ZIP, except the first row, which could only match on City, because the ZIP code for that observation was not found in lookup table. If neither ZIP nor CITY was matched, SAS will try to match at STATE level.

Geocode Street Address Values – Using SAS Sample Street Lookup File

Let's geocode by **street address**, first using the sample lookup dataset SASHELP.GEOEXM, which includes **only Wake County addresses** (because SAS Institute is in Wake County in Cary, NC):

```
proc geocode
   method=STREET                /* Specify geocoding method          */
   data=addresses               /* Input data set of addresses       */
   out=geocoded_WakeCounty      /* Output data set with X/Y values   */
   lookupstreet=SASHELP.GEOEXM /* Wake County street lookup dataset */
   type=sashelp.GCTYPE;         /* Lookup table for street types     */
run;
```

Note: SASHELP.GCTYPE contains various spellings of street types, such as STREET, STR, STRT and ST, and how they should be standardized → ST. You can add a customized street type dataset if you like.

If we print selected fields from the new dataset "geocoded_WakeCounty", we see this:

Y	X	M_ADDR	_MATCHED_	_STATUS_	_NOTES_	_SCORE_	address
40.7142	-74.0064		City	City/State Match	CT ST	10	123 Junk Street
36.1064	-79.8174		ZIP	ZIP match	ZC	15	3003 Branchwood Dr
35.8184	-78.7950	899 NW Cary Pkwy	Street	Found	AD ZC ENDNM DP TS NOCTM	77	940 NW Cary Parkway
35.8562	-78.6412	5505 Six Forks Rd	Street	Found	AD ZC NM TS	75	5505 Six Forks Road

When geocoding by Street, additional geocoding-related fields are added:

- M_ADDR – street address match value from lookup dataset
- M_CITY, M_STATE, M_ZIP (not shown above) – match values from lookup dataset
- _STATUS_ - type of match found
- _NOTES_ - code values contribute to _SCORE_ total
  - AD – street name matched (+20 to _SCORE_ total)
  - CT – City name matched (+5)
  - NM – House number matched on correct side of street (+10)
  - TS – Street type suffix matched (+20)
  - ZC = 5-digit ZIP matched (+15)
  - Other _NOTES_ values described in link[2] (**Street Geocoding Note Values** section)
- _SCORE_ -- comparative accuracy of match

Notice only the last two rows _MATCHED_ by Street (they are the only ones in Wake County). All the others matched either at ZIP or CITY, as before.

Although we only reference SASHELP.GEOEXM in the above SAS code, PROC GEOCODE is reading all these files from the SASHELP library:

- GEOEXM – Street names and ZIP codes in Wake County, NC– links to GEOEXS
- GEOEXS – census geographic areas – links to GEOEXP
- GEOEXP – X (longitude) and Y (latitude) values

Geocode Street Address Values – Using SAS US Street Lookup File

Now, let's geocode again by **street address**, but this time use the SAS supplied **TIGER U.S. Census lookup tables**.

To run this code, there's some preliminary work you must do, including first downloading a .ZIP file from the SAS website[1]. This .ZIP file contains several CSV files, a SAS program, and a ReadMe file with instructions on how to run the SAS program to import the CSV files and create these output SAS datasets with U.S. TIGER data:

- USM – Street names and ZIP codes in US – links to USS
- USS – census geographic areas – links to USP
- USP – X (longitude) and Y (latitude) values

Note: These files are very large. If you only need one or two states' worth of street addresses, you can't easily subset these files. Instead, SAS Institute recommends downloading specific state TIGER Census files directly. See Appendix for more details.

```
libname lookup 'C:\Geocode\Lookup';

proc geocode
   method=STREET            /* Specify geocoding method        */
   data=addresses           /* Input data set of addresses     */
   out=geocoded_byStreet    /* Output data set with X/Y values */
   lookupstreet=Lookup.usm /* US street lookup data set       */
   type=sashelp.GCTYPE;     /* Lookup table for street types   */
run;
```

If we print selected fields from the new dataset "geocoded_byStreet", we see this:

Y	X	M_ADDR	_MATCHED_	_STATUS_	_NOTES_	_SCORE_	address
40.7142	-74.0064		City	City/State Match	CT ST	10	123 Junk Street
36.1090	-79.8246	3003 Branchwood Dr	Street	Found	AD ZC NM TS	75	3003 Branchwood Dr
35.8183	-78.7974	940 NW Cary Pkwy	Street	Found	AD ZC NM DP TS NOCTM	85	940 NW Cary Parkway
35.8562	-78.6412	5505 Six Forks Rd	Street	Found	AD ZC NM TS	75	5505 Six Forks Road

All the addresses matched at the Street level, except for the first row, which is a bad address, so it only matches at the City level (or ZIP level if it contains a valid ZIP code).

Assigning Census Geographic Areas

With a slight change to the above SAS code, we can retrieve the United States Census geographic areas – blocks, block groups and tracts – for those addresses that successfully found a geocode match at the "Street" level. As a reminder, census blocks are grouped into block groups, which are grouped into tracts. All these values are stored within the USS lookup dataset.

By default, census area fields are not returned by PROC GEOCODE. To retrieve them, use the ATTRIBUTE_VAR option, specifying which additional fields (from the USS lookup dataset) you want to add to the output dataset. Census area fields include:

- Block
- BlkGrp
- Tract
- CountyFP – County FIPS code

This code saves three census area fields to the output dataset - geocoded_census:

```
proc geocode
   method=STREET
   data=addresses
   out=geocoded_census
   lookupstreet=Lookup.usm
   type=sashelp.GCTYPE
   attribute_var = (BLOCK, BLKGRP, TRACT);
run;
```

If we print selected fields from “geocoded_census”, we see this:

Y	X	Block	BlkGrp	Tract	M_ADDR	M_CITY	M_STATE	M_ZIP	M_OBS	_MATCHED_
40.7142	-74.0064	.	.	.		New York	NY	.	85948	City
36.1081	-79.8281	1018	1	12508	2630 Battleground Ave	Greensboro	NC	27408	1319999	Street
35.8183	-78.7974	1002	1	53512	940 NW Cary Pkwy	Cary	NC	27513	1987950	Street
35.8562	-78.6412	2000	2	53716	5505 Six Forks Rd	Raleigh	NC	27609	7388344	Street

Assigning Districts

Once you have street-address X and Y location values, you can then determine how these points fit within any geographic area, if you have a shape file-set containing the outlines of those areas. Luckily, there's a SAS procedure – PROC GINSIDE – to help you do that.

In our example, we would like to see which NC house congressional district each X/Y point belongs in. Using our NC house congressional district shape file-set, we first convert the .SHP file-set into a SAS map dataset (a one-time process):

```
%let NCHouseShapeFile = C:\Shapefiles\NCGA_House_2022.shp;
libname maps "C:\Geocode\Data\Maps";

* Shape Files all have same ID field - district;
proc mapimport datafile="&NCHouseShapeFile" out=maps.NCHouse;
     g_NCHouse = district; * Rename ID field to g_NCHouse;
run;
```

Now, we call PROC GINSIDE, referring to our geocoded dataset and map dataset:

```
proc ginside
   data=geocoded_byStreet /* Input file with X/Y values     */
   map=maps.NCHouse       /* File with shape outlines       */
   out=DistrictValues;    /* Output dataset                 */
   id g_NCHouse;          /* Lookup value from MAP= dataset */
run;
```

If we print selected fields from the new dataset "DistrictValues", we see this:

Y	X	g_NCHouse	address	city	state	zip
40.7142	-74.0064	.	123 Junk Street	New York	NY	99999
36.1090	-79.8246	61	3003 Branchwood Dr	Greensboro	NC	27408
35.8183	-78.7974	49	940 NW Cary Parkway	Raleigh	NC	27513
35.8562	-78.6412	34	5505 Six Forks Road	Raleigh	NC	27609

Note that districts, and other geographic areas, may change over time.

In Summary

Using SAS geocoding procedures such as PROC GEOCODE and GINSIDE has allowed us to analyze data in ways we had been unable to in the past. We have always been able to provide reports based on ZIP or county, but as a result of this new process, we are now able to report on additional geographic areas such as legislative districts and census tracts.

APPENDIX

How to create a subset of USM/USS/USP files

Overall Process: Download the Tiger/Line files from the Census Bureau for a particular state and use %TIGER2GEOCODE to create the lookup data from the downloaded data.

Read more about %TIGER2GEOCODE here: https://documentation.sas.com/doc/en/pgmsascdc/v_030/grmapref/p1fh3zqvaqtlgan15yqko1g4t838.htm - it includes these instructions:

1) Start here: https://support.sas.com/en/knowledge-base/maps-geocoding/geocodes.html
2) Click "Street Geocoding Downloads". Login with your SAS login. Select the link for "Code To Import US TIGER files for 9.4", which will download %TIGER2GEOCDE SAS macro in .zip file.

To get one specific state's TIGER files:

1) Go here: https://www.census.gov/cgi-bin/geo/shapefiles/index.php Select "Access our FTP site for additional downloading options". Or you can view this site within File Explorer: ftp://ftp2.census.gov/geo/tiger/
2) Navigate to the most recent TIGERyyyy folder (currently, that's TIGER2021).
3) Select the EDGES folder. Download the .zip file for your state, based on FIPS code xx. For example, the North Carolina data points are in these files: **tl_2021_37XXX_edges.zip**, because NC's FIPS code is "037". Other states have a unique FIPS code.
4) Repeat the above process with the FACES, FEATURNAMES and PLACE folders.
5) Unzip all the files into a single folder, modify and run the %TIGER2GEOCODE SAS macro.

You can unzip all the .zip files in a single folder by running this Windows PowerShell command (after changing to the folder containing all the ZIP files (by using the CD command):

Get-ChildItem *.zip | Expand-Archive -DestinationPath "C:\output-foldername"

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the primary author at:

Bruce Nawrocki
North Carolina Division of Public Health, Injury and Violence Prevention Branch
5505 Six Forks Rd, Raleigh NC 27613
E-mail: bruce.nawrocki@dhhs.nc.gov

REFERENCES – See superscripts in document above

1. For an overview of the geocode process in SAS and download necessary files -- http://support.sas.com/rnd/datavisualization/mapsonline/html/geocode.html

2. More information on PROC GEOCODE and SAS coding examples here: http://documentation.sas.com/?docsetId=grmapref&docsetVersion=9.4_01&docsetTarget=n02y3yabtlqatsn16gp2fo51yo7p.htm&locale=en

3. More information on PROC GINSIDE and SAS code examples here: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grmapref/p1frskc294tbapn1wwumr6m4d3ka.htm