SESUG Paper 110-2023

Geocoding in SAS®: The Basics

Keli Sorrentino, Julie Plano, Yale School of Public Health

ABSTRACT

Geocoding addresses is a common task required for a variety of projects across a range of industries. It isn’t necessary to become familiar with ArcGIS to geocode because PROC GEOCODE is included in SAS® beginning with version 9.4M5. The first time I needed to geocode addresses, I found it challenging to locate documentation that clearly described the steps to accomplish the task in SAS. I found bits of information in multiple places to finally complete my assignment. This paper will simplify the steps to get you geocoding your addresses quickly.

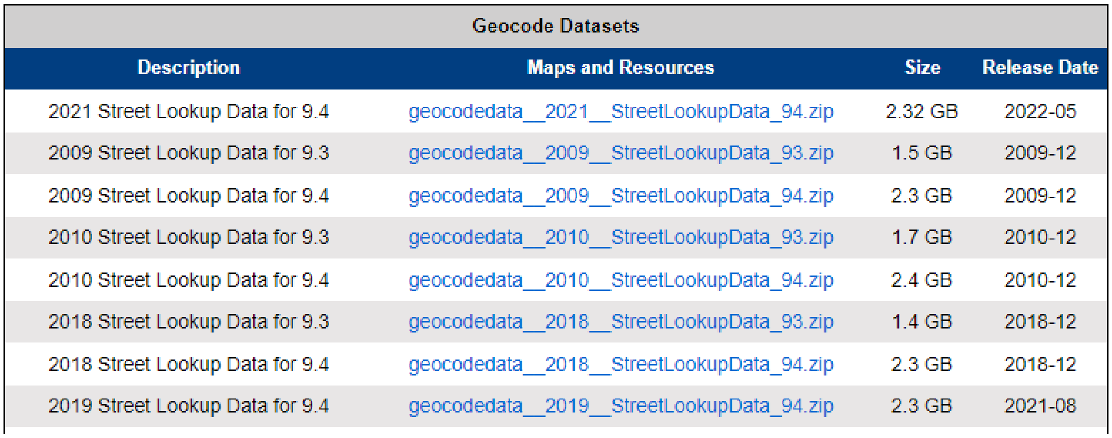
INTRODUCTION

Across many industries it is becoming imperative for a programmer to know how to geocode addresses. This doesn’t require the programmer to learn ArcGIS because geocoding can be done easily in SAS. There are three basic steps for geocoding in SAS. First, the required precision of geocoding is determined, and the appropriate lookup data is obtained. Next, the set of addresses needing to be geocoded is converted to a SAS file with appropriate variables. And finally, the PROC GEOCODE statement is adjusted to the needs of the project at hand and run.

GEOCODING PRECISION AND LOOKUP DATA

To begin, a programmer must determine the level of precision required for completing their specific geocoding task. This decision will be guided by the data available as well as the requirements of the project. SAS offers geocoding to city level, zip code level, and street level. If geocoding to street level precision, it is necessary to have street addresses in the data. Even with street addresses available, zip code or city level geocoding may suit the project fine. The examples in this paper are geocoded to street level.

Once the level of geocoding has been decided, the lookup data files required for the process must be downloaded. These files are preprocessed US Census Tiger Data files. Each zipped file is approximately 2GB and contains several CSV data files, a SAS program, and a ReadMe file. Once the lookup data is downloaded and saved, it can be reused for different projects but keep in mind addresses can change over time due to new street names or zip codes being split or added. It is always important to use the appropriate lookup data. There are options for lookup data from previous years which can be beneficial if the addresses being geocoded are from an older set of data while current address files should always be geocoded with the most recent data available (Display 1).



Geocode Datasets			
Description	**Maps and Resources**	**Size**	**Release Date**
2021 Street Lookup Data for 9.4	geocodedata__2021__StreetLookupData_94.zip	2.32 GB	2022-05
2009 Street Lookup Data for 9.3	geocodedata__2009__StreetLookupData_93.zip	1.5 GB	2009-12
2009 Street Lookup Data for 9.4	geocodedata__2009__StreetLookupData_94.zip	2.3 GB	2009-12
2010 Street Lookup Data for 9.3	geocodedata__2010__StreetLookupData_93.zip	1.7 GB	2010-12
2010 Street Lookup Data for 9.4	geocodedata__2010__StreetLookupData_94.zip	2.4 GB	2010-12
2018 Street Lookup Data for 9.3	geocodedata__2018__StreetLookupData_93.zip	1.4 GB	2018-12
2018 Street Lookup Data for 9.4	geocodedata__2018__StreetLookupData_94.zip	2.3 GB	2018-12
2019 Street Lookup Data for 9.4	geocodedata__2019__StreetLookupData_94.zip	2.3 GB	2021-08

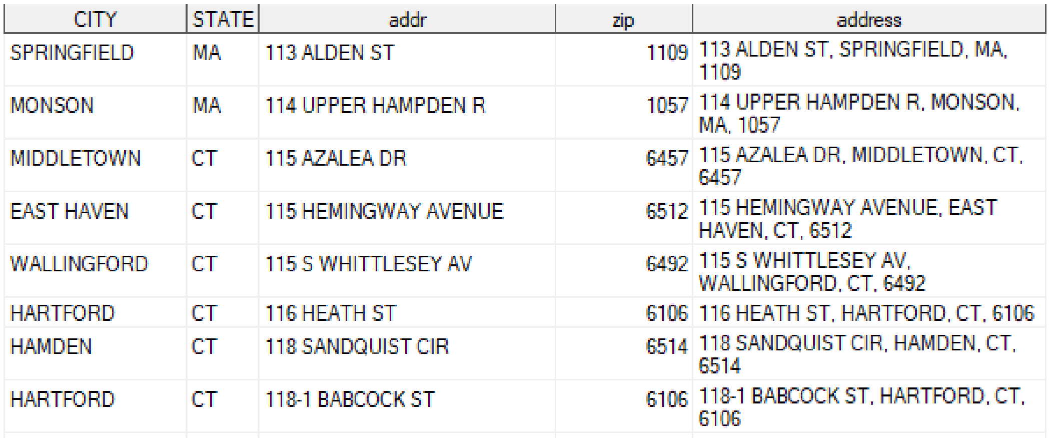
Display 1. Some of the Lookup data options for Street Level Geocoding

Next steps include unzipping the lookup data file, opening the import program provided by SAS, and updating the PATHIN and PATHOUT macro variables for the user's storage requirements. These are the only two variables which require updating by the user for the provided program to import the CSV lookup data and save the produced SAS datasets. Be aware approximately 16GB of disk space is needed for the final data sets.

```
/***Update the macro variable paths***/
%let PATHIN=C:\Geocode; /*Location of saved, unzipped, Street Lookup Data*/
%let PATHOUT=C:\Geocode\Data; /*Location to save the SAS geocoding datasets*/
```

PREPARING THE ADDRESS DATA FOR GEOCODING

Project data containing the addresses needs to be imported to SAS (if it isn't already a SAS file). The data should be inspected for any cleaning required prior to geocoding. For example, the address data needs to be in distinct variables which define the main street address, city, state, and zip code. Single variable strings which include the street address, city, state, and zip will not work (Display 2). Address attributes such as apartment or floor numbers should be removed from the primary address and made a separate variable or deleted. FIPS codes can't be used for geocoding in SAS. If the data for city and/or state is in FIPS or other numerical form, those data will need to be reformatted to city names and state names or abbreviations. While SAS will geocode addresses without zip codes, the process will be more effective with zip codes included. PROC GEOCODE has default variable names which are "address", "city", "state", and "zip". The variable names can be adjusted during the preparation phase, or the dataset variable names can be specified in the PROC GEOCODE statement.



CITY	STATE	addr	zip	address
SPRINGFIELD	MA	113 ALDEN ST	1109	113 ALDEN ST, SPRINGFIELD, MA, 1109
MONSON	MA	114 UPPER HAMPDEN R	1057	114 UPPER HAMPDEN R, MONSON, MA, 1057
MIDDLETOWN	CT	115 AZALEA DR	6457	115 AZALEA DR, MIDDLETOWN, CT, 6457
EAST HAVEN	CT	115 HEMINGWAY AVENUE	6512	115 HEMINGWAY AVENUE, EAST HAVEN, CT, 6512
WALLINGFORD	CT	115 S WHITTLESEY AV	6492	115 S WHITTLESEY AV, WALLINGFORD, CT, 6492
HARTFORD	CT	116 HEATH ST	6106	116 HEATH ST, HARTFORD, CT, 6106
HAMDEN	CT	118 SANDQUIST CIR	6514	118 SANDQUIST CIR, HAMDEN, CT, 6514
HARTFORD	CT	118-1 BABCOCK ST	6106	118-1 BABCOCK ST, HARTFORD, CT, 6106

Display 2. Address Variable Examples. It is important to have unique variables for each portion of the address. The variable "address" is an example of a string SAS® will not geocode. Note: CT and MA have zip codes with leading zeros which are not required for PROC GEOCODE to work.

WRITING THE PROC GEOCODE STATEMENT

Once the data are prepared, the PROC GEOCODE statement is very straight-forward. The base syntax for PROC GEOCODE is:

PROC GEOCODE DATA=*address_file*; run;

When this statement is run, SAS will geocode the address file to zip code or city accuracy. Using optional arguments will provide better outcomes. The syntax we recommend is:

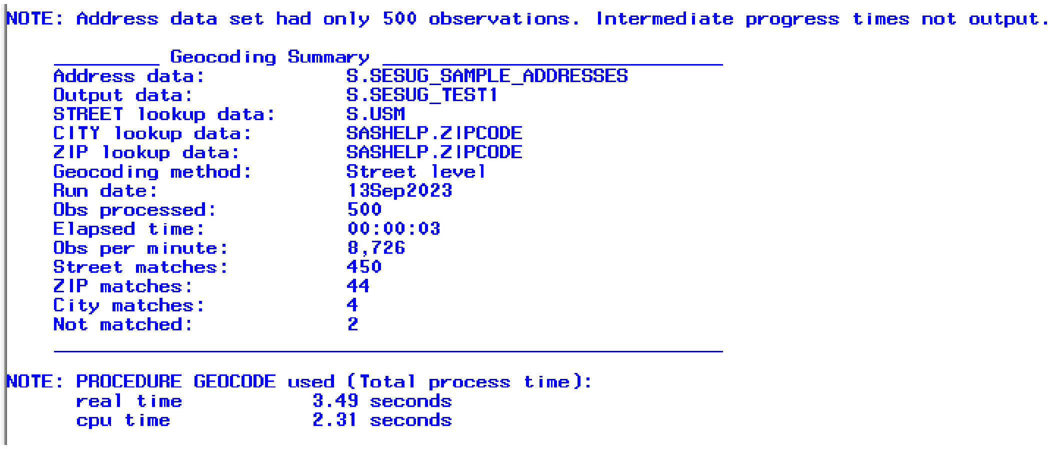
```
proc geocode
 method=street
 data=&addresses /*Source of address data*/
 out=&out1 /*Storage location for geocoded output*/
 lookupstreet=&LUS /*path to the lookup data file (usm.sas7bdat)
    saved from the ImportCSVfiles step*/
 lookupcity = sashelp.zipcode /*preloaded sashelp file*/
 addressvar = addr; /*Example of specifying the variable if not
     "address"*/
run;
```

In this example, addressvar shows how the name of the variable can be specified if it does not have the default name "address" in the dataset. There are similar arguments for the other address variables called addresscityvar, addressstatevar, and addresszipvar. There are also attribute variables that can be easily added to the output. These attributes may be helpful for the project and include - but aren't limited to - block, tract, side (of street), and county. The code below shows the inclusion of attribute variables:

```
proc geocode
 method=street
 data=&addresses
 out=&out2
 lookupstreet=&LUS
 lookupcity = sashelp.zipcode
 addressvar = addr
 attributevar = (block, tract);
run;
```

RESULTS

The first place to check the results of the PROC GEOCODE is in the log where a Geocoding Summary is produced (Display 3). This summary provides a range of information including the number of addresses geocoded to each level. Even with street geocoding chosen, not every address will be geocoded to that level, and it is important for the user to assess how well the procedure worked for their data. If the percentage of matching to street level is low, it is necessary to re-evaluate the original address dataset for issues which may be leading to inaccurate geocoding.



```
NOTE: Address data set had only 500 observations. Intermediate progress times not output.

    _________ Geocoding Summary _____________________________
    Address data:            S.SESUG_SAMPLE_ADDRESSES
    Output data:             S.SESUG_TEST1
    STREET lookup data:      S.USM
    CITY lookup data:        SASHELP.ZIPCODE
    ZIP lookup data:         SASHELP.ZIPCODE
    Geocoding method:        Street level
    Run date:                13Sep2023
    Obs processed:           500
    Elapsed time:            00:00:03
    Obs per minute:          8,726
    Street matches:          450
    ZIP matches:             44
    City matches:            4
    Not matched:             2
    _________________________________________________________

NOTE: PROCEDURE GEOCODE used (Total process time):
      real time           3.49 seconds
      cpu time            2.31 seconds
```

Display 3. Geocoding Summary found in the SAS® Log

The output dataset includes all the variables from the original address dataset . In addition to the X and Y coordinates of the geocoded address, there are several other useful variables produced. These include variable values for each matched segment of the address from the lookup data and a score of the relative accuracy of the geocoding outcome. (Display 4)

M_ADDR	Char	Matched value from the lookup data
M_CITY	Char	Matched value from the lookup data
M_OBS	Num	Row number of the matched address in the lookup data
M_STATE	Char	Matched value from the lookup data
M_ZIP	Num	Matched value from the lookup data
X	Num	Geocoded Longitude (DEGREES)
Y	Num	Geocoded Latitude (DEGREES)
MATCHED	Char	Level of match – Street, Zip, City
NOTES	Char	See “Street Geocoding Note Values”
SCORE	Num	Relative accuracy of match
STATUS	Char	Type of match

Display 4. SAS® generated output variables

CONCLUSION

Geocoding in SAS can be a relatively simple procedure. Preparing the data for analysis may be the biggest hurdle. A critical piece of the process is determining the level of geocoding required by the project and downloading the required lookup data. Once the data is clean and in proper formats, the PROC GEOCODE statement is simple to adapt to a range of scenarios. Using PROC GEOCODE provides a summary of the data run and easy to understand output variables.

CONTACT INFORMATION

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

Keli Sorrentino
Yale School of Public Health
203-764-9185
Keli.sorrentino@yale.edu