Paper 118-2012

Applications of PROC Geocode and Incorporation of Census Block-Level Data

John Havens-McColgan, Yale University, New Haven, CT, US

ABSTRACT

PROC GEOCODE is a powerful tool in SAS® with applications in epidemiological studies that involve geographically restricted exposures and health outcomes. This paper describes the use and application of PROC GEOCODE, specific to street-level geocoding. Preliminary steps are reviewed, and results are presented with a focus on applications of PROC GEOCODE in an observational study for which data at a census block level must be incorporated. Applications discussed include determining eligibility where requirements for the study are geographically limited, predicting the size of a population of eligible subjects in a geographic area, analyzing subject demographic characteristics, and comparing information to census data to ensure representative sampling.

INTRODUCTION

In environmental health research, matching exposure to health outcomes in space and time is critical. One method to accomplish this is using the SAS Geocode procedure to match street addresses with other location-linked data, e.g. by Census Block and Tract. Many important demographic co-variates are available at the Census Block level.

The purpose of this paper is to describe a method of matching addresses with census block and tract numbers using the Street Geocoding procedure in SAS, then describe how to apply the results. In the context of subject recruitment where eligibility is geographically restricted, a respondent would only be eligible for the study if he/she resided in Census Blocks that contained certain exposures (e.g., sources of pollution under study). The first goal of using PROC Geocode is to determine eligibility for the study. Due to the large number of potential participants, a systematic way to screen for eligibility that is both accurate and efficient is needed, and PROC Geocode satisfies these requirements. One main advantage of PROC Geocode is that it uses exact address matching and does not rely on spatial geocoding, which introduces potential errors. Thus, matches between census block and address information are precise and can be replicated repeatedly without issue. Geographic and spacially matched data can be incorporated easily and at different points over the course of a study.

This method and application was originally developed through work on The National Children's Study at the CT Study Center at Yale University. Eligibility for the study was based on pregnancy status and residence within geographical segments based on Census Blocks. Original investigation into this method was a result of preparatory work for the beginning of the study. It was later expanded and used for other projects.

BEFORE GEOCODING

Before geocoding, preparation of the dataset is necessary. Ensure that there are separate fields for:

- The street address (e.g. 123 Day St.)
- A unit/apartment number (e.g. Apt 12)
- City
- State
- Zip

In this application of the geocode procedure, the unit apartment number is not used because census blocks do not divide apartment buildings. Thus, a street address is sufficient. It may be useful to retain apartment number information, but it is necessary to separate it from the street address.

The next step is to check the source data. PROC Geocode can handle alternative road names, but there are instances in which addresses may not be properly geocoded. Common sources of error are: misspelled street names, wrong street type (e.g. 'avenue' instead of 'drive'), and incorrect zip codes. Using an address lookup tool at this point is a way to handle address problems prior to using PROC geocode. Such tools include using the Google Earth program, which features an extremely flexible address matching mechanism, and using the geocode 'rematch' functionality of Arc GIS.

Addresses refined in this way are recorded in a separate set of fields to maintain integrity of the dataset. When ready to geocode, the following fields should present:

- Original street address
- Original unit/apartment number
- Original City
- Original State
- Original Zip
- Refined Street Address
- Refined City
- Refined State
- Refined Zip

PREPARING THE REFERENCE DATASET

This is an important step, and has been documented in other SAS papers. It is necessary to create a reference dataset so that study data can be matched to the census blocks. Once the appropriate census reference year has been decided, follow the procedures listed here <http://support.sas.com/rnd/datavisualization/mapsonline/html/geocode.html > to create the necessary file.

USING PROC GEOCODE

Once the datasets are available, continue with the geocoding procedure. This paper will focus on street-method geocoding due to the nature of the data used and the specificity needed to assign census blocks.

The SAS code, below, fulfills several functions. The first step is to import the file from the format used to check address consistency. In this example, the dataset was exported to ArcMap and used the native rematching process to resolve most issues. Next, in the example application, eligibility for the study was confined to a particular county location, and recruitment efforts were confined to this area. All inquiries for the study from individuals residing out of the county were set aside and examined, to ensure that resources were not being spent to recruit ineligible individuals. Third, street level geocoding occurs, and census block (BLOCKCE00) and tract (TRACTCE00) are specified as output variables.

```
PROC IMPORT OUT= address
            DATAFILE= "C:\ address.dbf"
            DBMS=DBF REPLACE;
     GETDELETED=NO;
RUN;

data geocode outofcounty;
set address;

if ARC_City in (*LIST OF ELIGIBLE TOWNS*) then output geocode;
else output outofcounty;
run;

libname x 'C:\geocoding\';
proc geocode
method = street
addressvar = ARC_Street
ADDRESSCITYVAR=ARC_City
ADDRESSZIPVAR=finZip
ADDRESSSTATEVAR=ARC_State
data = geocode
out = x.results
lookupstreet=x.newhaven_m
attribute_var = (BLOCKCE00, TRACTCE00);run;
```

After this process, the dataset 'x.results' looks like this with a proc print:

```
                  Block   Tract
   Y         X      CeO0   CeO0   M_ADDR               M_CITY  M_STATE  M_ZIP  M_OBS  MATCHED_

41.286   -72.754    1001  184600  1 Day Dr                                 6405   3507   Street
41.555   -72.774    2004  171700  4 Mountain Rdg                           6450   8915   Street

_STATUS_        _NOTES_         _SCORE_        FID_1    Status                  Score     type

 Found      AD ZC NM TY            60           1077      M                    100.00      A
 Found      AD CT ST NM TY         55             52      M                    100.00      A

Match_addr

1 Day Dr, Madison, CT, 06405
4 Mountain Rdg, Clinton, CT, 06450

Addr_type    ARC_Street          ARC_City     ARC_State            ARC_ZIP

Address      17 Eastwood Drive   Branford     CT                   06405
Address      4 Pine Tree Ridge   Meriden      CT


strCHA_1            strCHACity           strCHAStat   strCHAZip

1 Day Dr            Madison               CT          06405
4 Mountain Rdg      Clinton               CT
```

Output 1. PROC PRINT Output after running PROC Geocode (note: output has been modified to protect potential PII, addresses and matches are not accurate)

In the output, latitude and longitude and census block and tract number have been added. There is also information about how well the addresses matched. A lower score indicates a poorer match. Since the purpose of this geocoding method is precise matching, a follow-up to the actual geocode is to find those addresses that did not geocode properly:

```
proc print data = x.results;
where _SCORE_ <= 50;
run;
```

This allows a manual review of the addresses that are problematic. These issues can be resolved by:

1. Refining the reference dataset or using another from a different source.
2. Ensuring proper spelling and extensions of the street addresses used.
3. Checking for alternate registered names for the address in question.

For more information on using PROC geocode, see the paper referenced below.

APPLICATIONS OF RESULTS

Ultimately, the purpose of PROC geocode is to utilize information that can now be matched with addresses. This encompasses anything that can be identified in a Census Block and incorporated into analysis. Eligibility for a study based on residence inside of certain census blocks can be determined as follows:

```
data x.results;
set x.results;

ELIG = 0;

if TRACTCE00 EQ 142300 AND BLOCKCE00 in (4000) THEN ELIG = 1;
else ELIG = 0;
run;

proc print data = x.results noobs;
where elig = 1;run;
```

The subsettings IF loop is continued until all eligible geographic areas are included. This provides a systematic way to produce the addresses or study ID's of all eligible respondents.

The next step in preparing a dataset for analysis in an air pollution study, for example, is to match participants with exposure data. All that is required is to develop the datasets so that a match-merge based on Census Block and then Census Tract is possible. Depending on the data involved, specific environmental data (e.g., pollution data collected by the EPA in specific locations) can be matched with individual subject's health data by location. An alternative to using Census Blocks is to conduct a similar process using the Latitude/Longitude provided by the Geocode procedure. Although this is not an easy way of matching Census data, other applications are possible, such as distance calculations which would allow determination of health risk based on distance to a stationary hazard, such as a superfund site.

Another application of this method is to match respondent data with demographic data provided on the census block or tract level. In this application, a file with demographic data collected on the Census Block level for target years can be matched with respondent records. Once all respondents had been assigned to a census block using a sort and merge, it is possible to generate a summary of the demographic characteristics at each geographic level of interest: block, tract, zipcode, city and county. These could then be compared to census information to ensure representative sampling. Thus, under-recruitment or potential study bias can be identified and corrected.

This method could also be used to examine characteristics of interest to develop estimates of the size and location of a potential study population. For example, for a study on birth outcomes where pregnant women would need to be recruited, previous Census data could be used to develop lists of addresses in census blocks containing high levels of births. This could help direct the recruitment effort to particular areas. In fact, once an area has been defined, PROC geocode can be used to target mass mailings by using postal lists then selecting only addresses within or directly adjacent to this area.

CONCLUSION

PROC Geocode has applications beyond mapping. In epidemiologic studies where geographic location of potential subjects is critical (for example in studies of environmental exposures), PROC Geocode can be helpful in identifiying

areas of largest numbers of potential subjects, determining geographic eligibility of volunteers, monitoring sample demographics during recruitment, and matching subjects to local exposures.

REFERENCES

"PROC GEOCODE: Now with Street-Level Geocoding." SAS Global Forum 2010 paper and example source code download. SAS Institute Inc. http://support.sas.com/rnd/papers/.

RECOMMENDED READING

- SAS/GRAPH 9.2 Documentation: http://support.sas.com/documentation/onlinedoc/graph/index.html
- SAS® For Dummies®

CONTACT INFORMATION

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

Name: John Havens-McColgan
Enterprise: Yale University, CPPEE
Address: 1 Church St., 6th Floor
City, State ZIP: New Haven, CT, 06510
Work Phone: 203-764-9795
E-mail: john.havens-mccolgan@yale.edu