

Visually Exploring Proximity Analyses Using SAS® PROC GEOCODE and SGMAP and Public Use Data Sets

Louise Hadden, Abt Associates Inc.

ABSTRACT

Numerous international and domestic governments provide free public access to downloadable databases containing health data. Two examples include the Demographic and Health Surveys which include data from Afghanistan to Zimbabwe and the Centers for Medicare and Medicaid Services' Compare data and Part D Prescriber public use files. This paper and presentation will describe the process of downloading data and creating an analytic data base which includes geographic data; running SAS®'s PROC GEOCODE (part of Base SAS®) using Tiger street address level data to obtain latitude and longitude at a finer level than zip code; and finally using PROC SGMAP (part of Base SAS®) with annotation to create a visualization of a proximity analysis.

INTRODUCTION

Data science, machine learning and natural language processing (NLP) are fields and tools in which interest and activity have grown exponentially in the past five years. Interest in data sources useful for demonstrating techniques has also increased accordingly. Freely available and reliable banks of data are highly sought after. This paper will briefly discuss several high quality and robust data sources for analytic work, and focus in depth on one data source and BASE SAS tools that can be used to graphically represent measures and trends.

This paper is relevant to SAS users at all levels and cross-industry. BASE SAS version 9 maintenance release 6 was used for production of examples for this report. There are additional enhancements to SAS Graphics procedures available in SAS version 9 maintenance release 7 that are worth further exploration, but M7 was not available at my site at the time of this paper's publication.

SEEKING OUT AND WORKING WITH "FREE" DATA

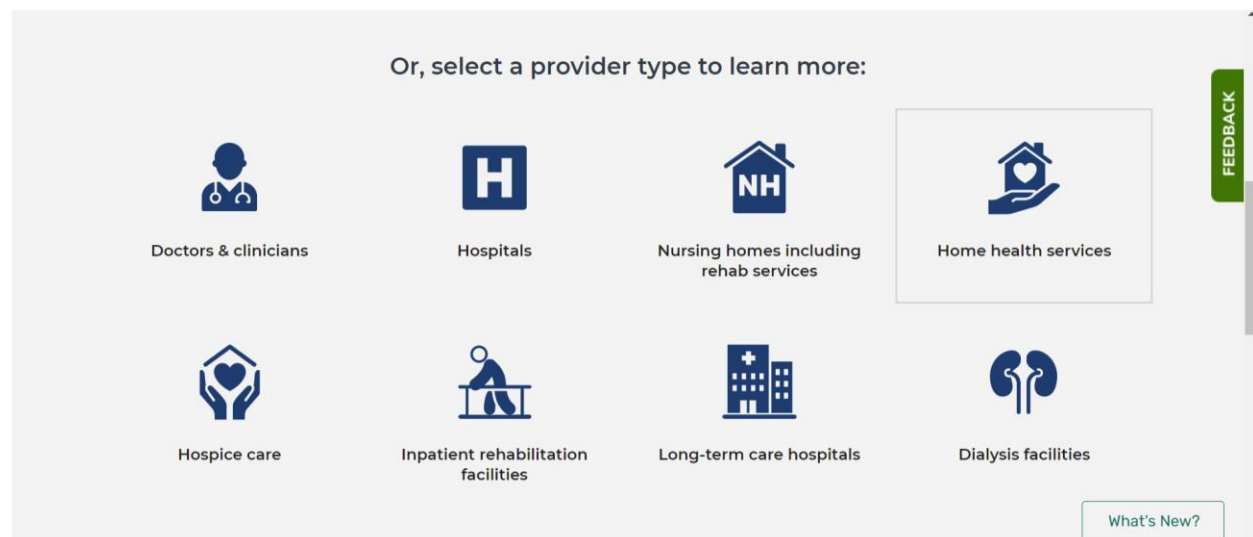
There are many resources for obtaining public use and freely available data, including scraping web sites with PROC HTTP, PROC LUA and the like, downloading and converting data on public use websites such as the US Census, CDC, CMS, DHS, and other domestic and international government websites. Sites such as Kaggle and Github can also be valuable to explore as they frequently hold data visualization and analytic competitions with supplied data. SAS also provides a number of sample data sets for targeted purposes in the SASHELP directory available to all licensed SAS users. Three public use resources will be discussed below, the international public use demographic and health surveys (DHS), National Health and Nutrition Examination Survey (NHANES), and CMS' Nursing Home Compare site (soon to be replaced by the Care Choices Experience (CCXP) website).

DHS data is obtained by request on their website (<https://dhsprogram.com/data/Using-Datasets-for-Analysis.cfm>). Their survey data covers a number of continents and countries outside of the US, from Africa to Latin America to Asia. The incredibly detailed surveys offer insights into numerous health topics, somewhat similar to the US Census. A variety of surveys are conducted in different countries on a rotating basis. Potential analysts are asked to review the available data and apply for use of specific surveys.

NHANES data, survey questionnaires and documentation may be freely downloaded from CDC's website, <https://www.cdc.gov/nchs/nhanes/index.htm>. In addition, those analysts

interested in nutrition and Usual Daily Intake analyses can review and download SAS analytic routines and instructions from the National Cancer Institutes's (NCI) website (<https://epi.grants.cancer.gov/diet/usualintakes/method.html#:~:text=The%20premise%20of%20the%20NCI,from%20those%20used%20for%20dietary>).

The Centers for Medicare & Medicaid Services (CMS) supports CMS Care at <https://www.medicare.gov/care-compare/>. CMS Care is replacing a set of "Compare" tools including Nursing Home Compare, which allows consumers to review and compare potential health related areas, including hospitals, home health agencies, nursing homes, etc.

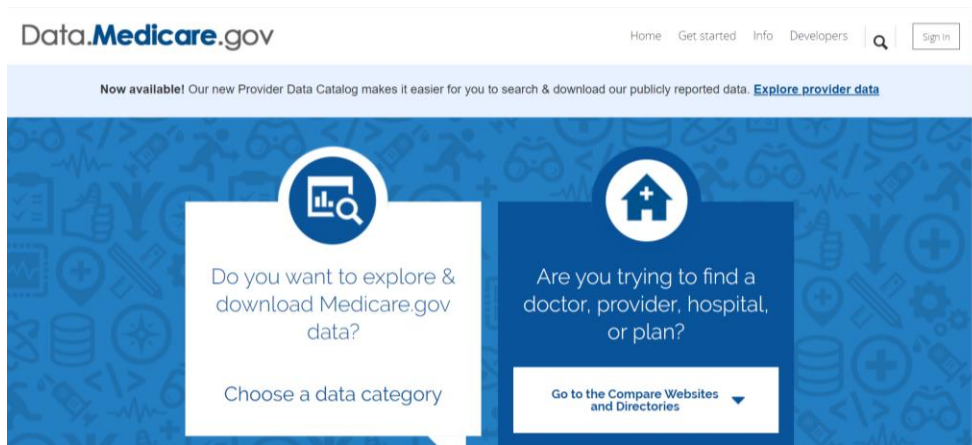


A variety of data elements from each health related area are freely available to the public for analytic purposes. Data sets that mirror the information backing CMS Care are available at <https://data.medicare.gov/>. Data sets in each realm that do not fall into the standardized "compare" area are available from <https://data.cms.gov/>, including data specific to COVID-19 outbreaks in nursing homes.

Examples used in this paper are derived from data from Data.Medicare.Gov and data.cms.gov. The author has used and published manuscripts using all three rich data sources described above, and encourages SAS users to avail themselves of these valuable resources, as well as others. The data that is provided in CSV format and documented in spreadsheets is meant to reflect the information that is provided on the interactive CMS website, but in a more convenient form for those who wish to analyze or track data. Many nursing homes and/or chain administrators download data to track their facilities' data across months, using the Technical User's Guide provided by CMS as a guide. It's a great way for providers to ensure that data that is being conveyed to CMS and the website is correct. It's also a great way for public reporters such as ProPublica, etc. to track performance of public entities. Data are free to download.

OBTAINING PUBLIC USE DATA

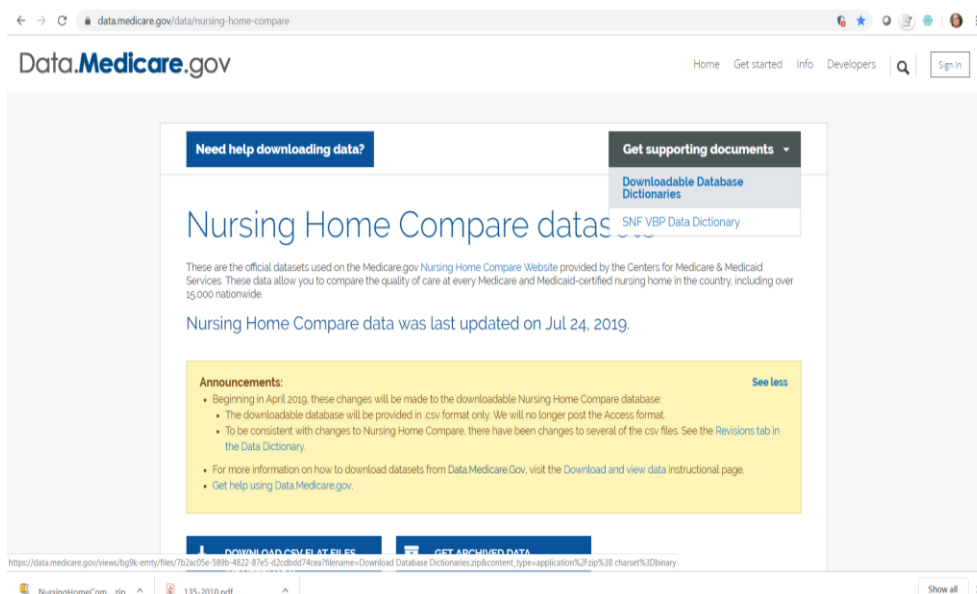
Navigate to Data.Medicare.Gov:



Choose your data category, in this case Nursing Home Compare data.



Download metadata by clicking on "Get supporting documents".



Download CSV file(s) by clicking on Download CSV Flat Files. The downloads should go directly into your downloads folder on Windows, or you will be able to save as, depending on your browser and operating system. There are archives, special views, and interactive visualization systems, but we're primarily interested in the data. You can also download data in other formats: click on a dataset in the list of datasets appearing at the bottom of

the page, click on Export button on the top right of the interactive dataset view and choose one of the following file formats: CSV (Comma Separated Values), Microsoft Excel (XLS), Microsoft Excel (XLSX), XML, RDF, and RSS. For this paper, I extracted the providerinfo_download file as well as the metadata spreadsheet. I've chosen the providerinfo file because it has geographic data as well as provider level data such as ratings, number of deficiencies, and so on.

	PROVNUM	PROVNAME	ADDRESS	CITY	STATE	ZIP	PHONE	COUNTY_SSA	County_name	OWNERSHIP	BEDCERT	RESTO
156	015373	HENRY COUNTY HEALTH AND REHABILITATION FACILITY	212 DOTHAN ROAD	ABBEVILLE	AL	36310	3345952241	330	Henry	Government - County	142	103.2
157	015374	CAMDEN NURSING FACILITY INC.	210 PONDEROSA DRIVE	CAMDEN	AL	36726	3346824231	650	Wilcox	For profit - Corporation	95	67.9
158	015375	ANNISTON HEALTH AND REHAB SERVICES	P.O. BOX 1825	ANNISTON	AL	36207	2562364949	070	Calhoun	Non profit - Other	85	41.6
159	015376	DIVERSICARE OF WINFIELD	144 COUNTY HWY 14	WINFIELD	AL	35594	2054874211	460	Macon	For profit - Corporation	123	111.7
160	015378	MARSHALL MANOR NURSING HOME	3120 NORTH STREET	GUNTERSVILLE	AL	35976	2569826561	470	Marshall	For profit - Corporation	91	86.5
161	015379	MOBILE NURSING AND REHABILITATION CENTER	7020 BRUNS DRIVE	MOBILE	AL	36695	2516391588	480	Mobile	For profit - Partnership	120	109.2
162	015381	WILLOWBROOKE CT SKILLED CARE CTR WESTMINSTER VLG	500 SPANISH FORT BLVD	SPANISH FORT	AL	36527	2516267007	010	Baldwin	Non profit - Corporation	60	43.9
163	015382	CANTERBURY HEALTH CARE FACILITY	1720 KNOWLES ROAD	PHENIX CITY	AL	36869	3342910485	560	Russell	For profit - Corporation	137	126.7
164	015383	WOODLAND VILLAGE REHABILITATION AND HEALTHCARE CEN	1900 OLIVE STREET	CULLMAN	AL	35055	2567391430	210	Cullman	For profit - Corporation	149	133.7
165	015386	ADAMS NURSING HOME	1555 HILLABEE STREET	ALEXANDER CITY	AL	35010	2563290847	610	Tallapoosa	For profit - Corporation	88	65.5
166	015388	EAST GLEN	53 MEDICAL PARK DRIVE EAST	BIRMINGHAM	AL	35235	2058331777	360	Jefferson	For profit - Corporation	108	101.9
167	015390	CAPITOL HILL HEALTHCARE CENTER	520 SOUTH HULL STREET	MONTGOMERY	AL	36104	3348342920	500	Montgomery	For profit - Corporation	284	256.1
168	015391	JACKSONVILLE HEALTH AND REHABILITATION, LLC	410 WILSON DRIVE SOUTHWEST	JACKSONVILLE	AL	36265	2564357704	070	Calhoun	For profit - Corporation	167	161.2
169	015392	CYPRESS COVE CARE CENTER	200 ALABAMA AVENUE	MUSCLE SHOALS	AL	35661	2563914330	160	Colbert	For profit - Corporation	90	74.6
170	015393	CROWNE HEALTH CARE OF MONTGOMERY	1837 UPPER WETUMPKA ROAD	MONTGOMERY	AL	36107	3342648416	500	Montgomery	For profit - Corporation	185	165.5
171	015396	COLONIAL HAVEN CARE & REHABILITATION CENTER	616 ARMORY STREET	GREENSBORO	AL	36744	3346243054	320	Hale	For profit - Corporation	97	85.9
172	015397	WINDSOR HOUSE	4411 MCALLISTER DRIVE	HUNTSVILLE	AL	35805	2568378595	440	Madison	For profit - Corporation	117	95.4
173	015398	MONROE MANOR HEALTH & REHABILITATION CENTER	236 WEST CLAIBORNE STREET	MONROEVILLE	AL	36460	2515752648	490	Monroe	For profit - Corporation	84	79.3
174	015400	MEADOWVIEW NURSING CENTER	7300 OLD HIGHWAY 78 EAST	PELL CITY	AL	35128	2056405212	570	St. Clair	For profit - Corporation	59	50.6
175	015402	EL REPOSO NURSING FACILITY	260 MILNER CHAPEL ROAD	FLORENCE	AL	36534	2567572143	380	Lauderdale	Non profit - Church related	60	58.8
176	015403	PALM GARDENS HEALTH AND REHABILITATION, LLC	3104 DAUPHIN SQUARE CONNECTOR	MOBILE	AL	36607	2514502800	480	Mobile	For profit - Corporation	100	86.0
177	015404	BARON HOUSE OF HUEYTOWN	190 BROOKLANE DRIVE	HUEYTOWN	AL	35023	2054912905	360	Jefferson	For profit - Corporation	50	33.0

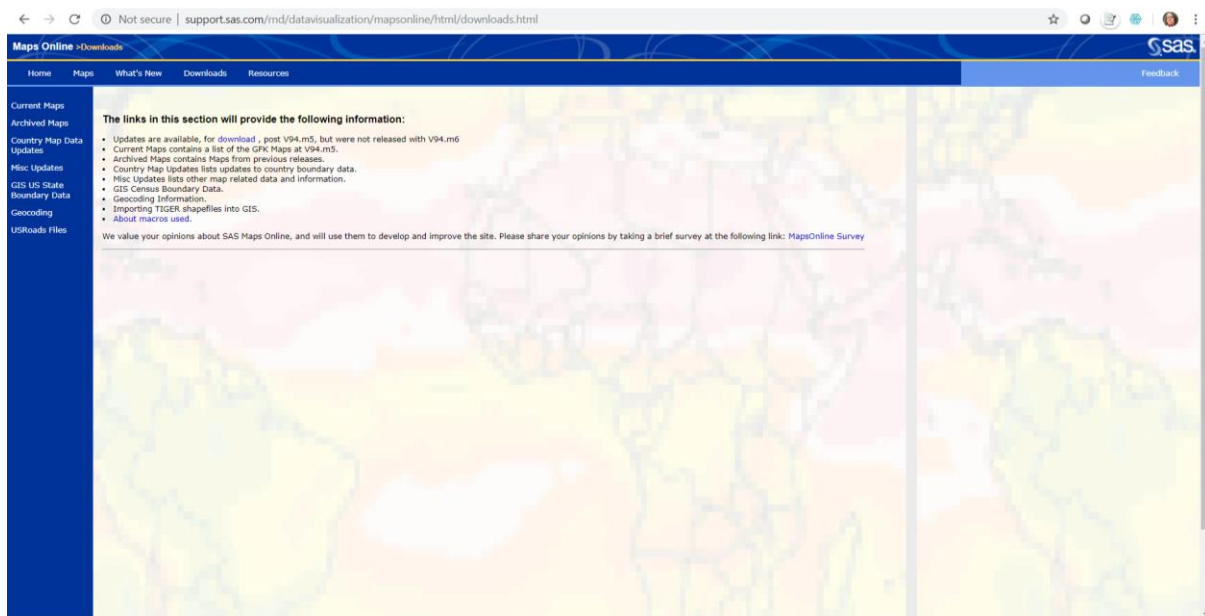
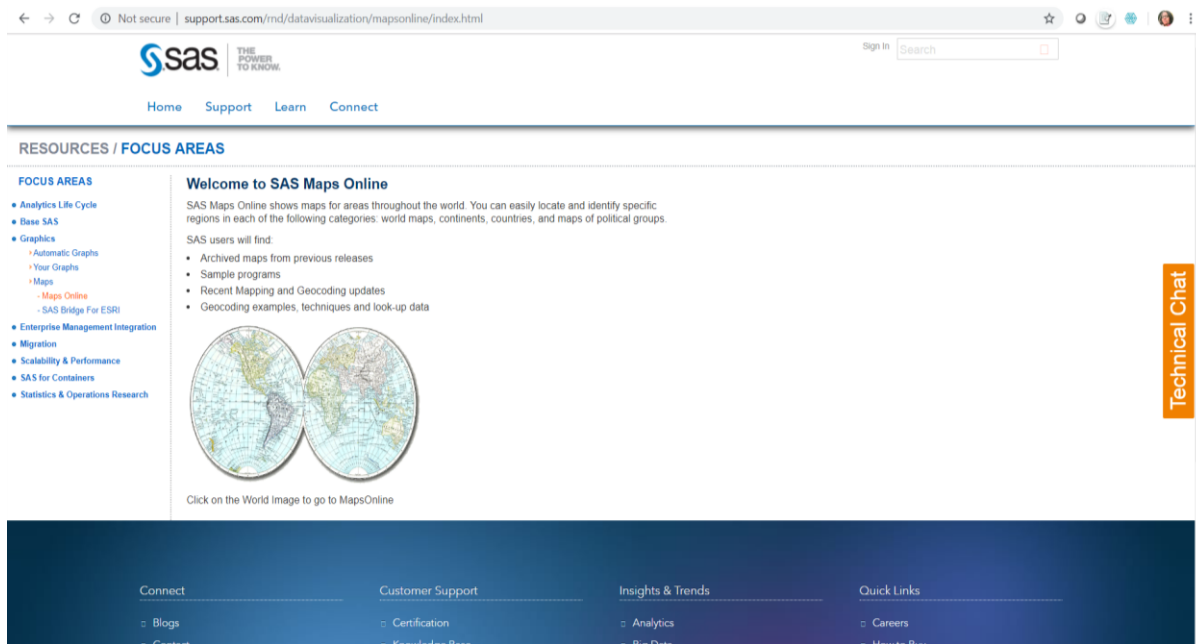
PROC GEOCODE

PROC GEOCODE has been available in SAS/GRAPH since Version 8.2, and recently became available in BASE SAS with a number of other tools in Version 9.4 Maintenance Release M5. SAS geocoding is available at the city, zip, zip+4, street and custom levels. It produces x and y coordinates for matches between your data and the lookup table. SAS provides city and zip level datasets in the SASHELP folder, and very limited street level data set in SASHELP – a county in NC which just might be Wake County. SAS also provides a link to files required for street level geocoding and more on SAS MAPSONLINE.

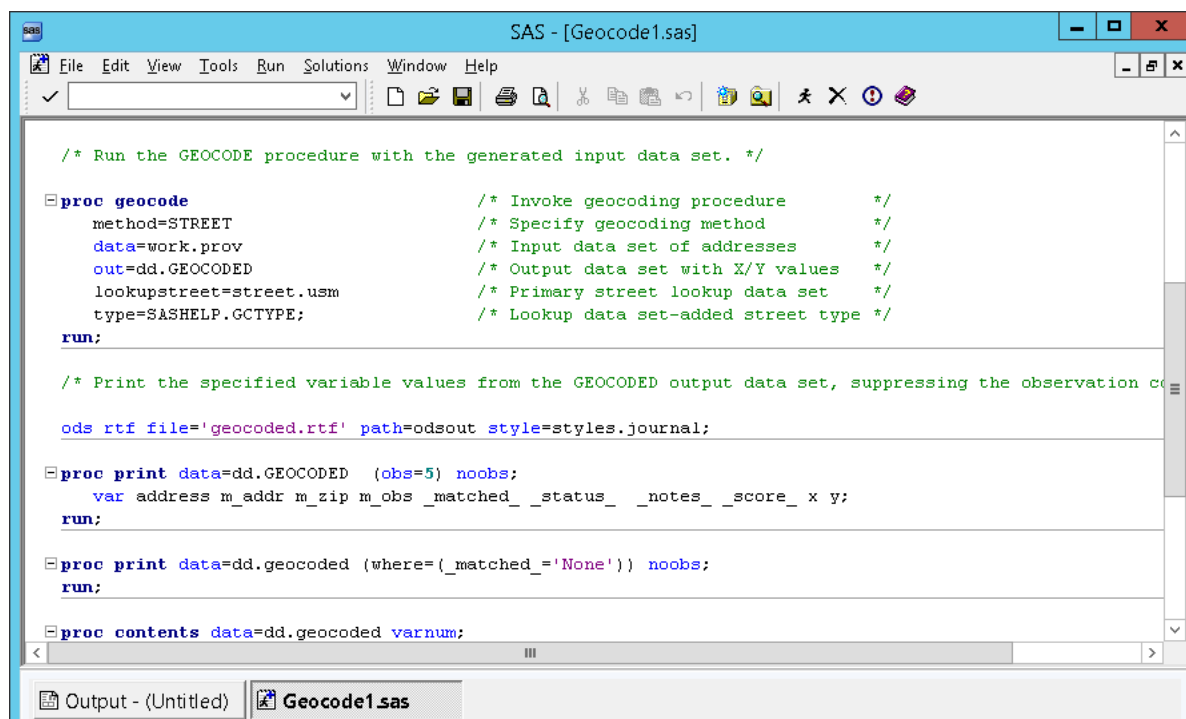
Basic syntax for the GEOCODE procedure is shown below:

```
PROC GEOCODE /* Invoke geocoding procedure */
method=STREET /* Specify geocoding method */
data=work.prov /* Input data set of addresses */
out=dd.GEOCODED /* Output data set with X/Y values */
lookupstreet=street.usm /* Primary street lookup data set */
type=SASHELP.GCTYPE; /* Lookup data set-added street type */
run;
```

To compare providers, we want to have the finest possible level of X,Y matching. It's not perfect – but better than zip level. The one county's worth of street level coding in SASHELP won't cut it. SAS MAPSONLINE has the answer. For those of you who don't know what SAS MAPSONLINE is, it's a page within support.sas.com that is dedicated to mapping in SAS, with examples, downloadable data, links for more data, and great information. First click on the map to enter, then click on Geocoding in the left hand menu.



Scroll down then click on the appropriate street level link for your version of SAS. In order to download, you'll need to log in to your SAS account, and then go get a cup of coffee unless you have amazing throughput because the street level file is huge.



Specify the method of geocoding, input and output data sets, where the lookup file can be found, and an additional street name data set translator. Not everything gets matched. Given enough time, you can clean up the less than perfect matches. For this example, 92% were at matched at street level, with another 7% at zip, the next level up.

DATA PREPARATION

Once we have our response data “geocoded”, we need to go to the next step. It’s easier to visualize large amounts of data if you can categorize the information beforehand. For example, with approximately 15,500 Medicare/Medicaid nursing homes in the United States, a given measure might have over 15,000 values (allowing for some missingness). In addition, you need to consider the unit of measure – is it the provider address, CBSA, county or state that you want to visualize?

The first step is to look at our data. Despite the title of the presentation, proximity analysis does not necessarily mean geographic proximity. Our provider file also has categorical, continuous and binary variables as well as some geographic information. For most visualizations, we are going to want to take that geographic data and use methods to assign discrete numbers of values of other variables to geographic entities such as state, county or zip code. Mapping associated providers (in our case) allows us to explore groups of providers in a way that a table might not immediately provide.

One method we use frequently in preparing data for mapping is PROC RANK. PROC RANK will rank records in ascending order into a specified number of groups. For mapping, our research indicated that viewers can’t really distinguish more than 5 different colors in a heat map, so we chose 5. There are some cons to PROC RANK – there may be issues with “tie-breaking” and your groups may not be equal in size.

```

PROC RANK DATA = state_sum
      GROUPS = 5
      OUT = stateranks_providers
      (KEEP=state providers:);
VAR ProviderR;
RANKS Provider_5 ;
RUN;

```

Rick Wicklin of SAS has come up with an amazing guide to binning methods – including PROC RANK – in his blog, <https://blogs.sas.com/content/iml/2019/08/07/essential-guide-binning-sas.html>. I highly recommend this blog, which presents a number of different options for grouping observations in sas, and offers some good ideas for visualizing “binned” data.

For distance calculations, there are SAS-supplied methods to obtain geographic distance calculations.: the geodist function calculates the distance between two pairs of latitude and longitude, accounting for the earth’s curvature – as the crow flies. The %geodist macro options include calculations output in miles, kilometers, degrees and radians. Through geocoding, we’ve added latitude and longitude to our provider records. These can now be used to calculate distance between providers and each other, and to other points of interest. This allows comparison of providers who are geographically adjacent, or not, to add other potential influences on quality measures. We can also use the zipcitydistance function using the zip codes available on the provider file (the function uses the sashelp.zipcode file which has zip code centroids – or we can pull the centroids off ourselves.)

MATCHMAKING, AKA PROC SGMAP

The hot new SAS mapping procedure PROC SGMAP is adding capability with every release. I don’t have M7 yet, but there are some amazing new options, which are bringing PROC SGMAP closer to what can be achieved with (many) years of experience with SAS/GRAPH maps and PROC GMAP. It has also brought a lot of functionality which used to be reserved for SAS/GRAPH users to BASE SAS – including PROC GEOCODE.

PROC SGMAP was introduced in SAS 9.4M5 as an extension of (ODS) Graphics techniques to render maps and then overlay plots such as text, scatter, or bubble plots. Starting with SAS 9.4M6, the SGMAP procedure added the GRADLEGEND and SERIES statements, and numerous options. The CHOROMAP statement is at production level. There’s even more in 9.4M7.

After your map data sets are obtained and prepared, use the SGMAP procedure to create maps and then overlay plots such as bubble, scatter, series, or text plots. With the enhanced SGMAP procedure, automatic legends are now generated, and the option that disables them is provided. Continuous or discrete legends are now possible, as well as legend customization, grouping capability, and the ability to customize polygon borders and series [plot](#) lines. A very useful addition to the SGMAP procedure (also available in other SG procedures) is a discrete attribute (non-geographic) map data set, which effectively acts like a hash table or format to control an entity’s appearance.

Below follows an example of using PROC SGMAP on our geocoded public use data. We construct five separate bubble plots with nursing homes with one star, two star, etc. ratings and overlay on a county map. This demonstrates the power and attractiveness of maps created with the new SGMAP procedure.

```

ODS HTML path=odsout body="&name..htm"
  (title="MA Nursing Home Facilities: Overall Ratings") style=htmlblue;

```

```
ods graphics / noscale imagefmt=png imagename="&name"
width=1200px height=900px noborder;

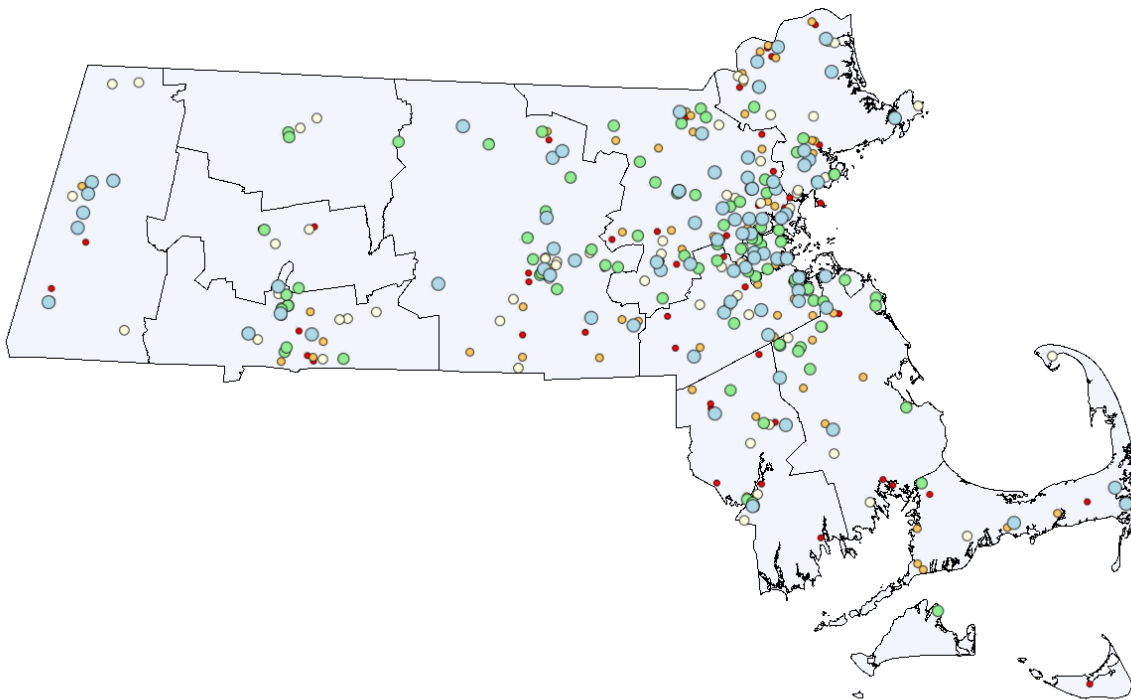
title1 color=gray33 height=24pt "Providers by Overall Rating Status (1-5
Stars) August 2019";
title2 color=gray33 height=18pt "Red: 1 Star Orange: 2 Stars Yellow: 3 Stars
Green: 4 Stars Blue: 5 Stars";

proc sgmap mapdata=my_map (drop = lat long) maprespdata=my_counties
plotdata=my_data noautolegend;
choromap counter / mapid=id lineattrs=(color=black) /* transparency=.5 */ ;

bubble x=x1 y=y1 size=or1 / fillattrs=(color=red) group=or1 transparency=.8
bradiusmin=3px bradiusmax=11px ;
bubble x=x2 y=y2 size=or2 / fillattrs=(color=lightorange) group=or2
transparency=.8 bradiusmin=4px bradiusmax=12px ;
bubble x=x3 y=y3 size=or3 / fillattrs=(color=lightyellow) group=or3
transparency=.8 bradiusmin=5px bradiusmax=13px ;
bubble x=x4 y=y4 size=or4 / fillattrs=(color=lightgreen) group=or4
transparency=.8 bradiusmin=6px bradiusmax=14px ;
bubble x=x5 y=y5 size=or5 / fillattrs=(color=lightblue) group=or5
transparency=.8 bradiusmin=7px bradiusmax=15px ;
run;
```

Providers by Overall Rating Status (1-5 Stars) August 2019

Red: 1 Star Orange: 2 Stars Yellow: 3 Stars Green: 4 Stars Blue: 5 Stars



CONCLUSION

SAS has always been at the forefront of data visualization, from the days of SAS/GIS, ESRI bridge to SAS, SAS/GRAPH, and now with the SG procedures, SAS Visual Analytics, and SAS Viya. The SG procedures, PROC GEOCODE and new M7 enhancements such as ATTRMAPS and annotation macros available in BASE SAS continue to evolve. In conjunction with public use data and educated data preparation techniques, SAS allows the creation of some fantastic custom visualizations. I hope you will feel inspired to explore on your own.

ACKNOWLEDGMENTS

Many thanks to Robert Allison, Cynthia Zender, Dan Heath, Ed Odom, Sanjay Matange, Nat Wooding, Barbara Okerson, Troy Hughes, and Mike Zdeb, for their unfailing patience, inspiration and technical assistance!

Code provided upon request.

CONTACT INFORMATION

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

Louise Hadden
Abt Associates Inc.
Louise_Hadden@AbtAssoc.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.