Memo on Geocoding at the Urban Institute
Eric Burnstein and Rob Pitingolo

November 2015

Summary Recommendation

The Mapping User's Group recommends that the Urban Institute adopt the use of the SAS Geocode procedure as a primary method of geocoding and ArcGIS online as a secondary method. This "hybrid" approach balances cost concerns with completeness concerns for researchers. For smaller files, the Census Geocoder and the Texas A&M Geocoding Services are easy to use and free.

Introduction

Since ESRI moved to a fee-based geocoding system, The Urban Institute has been without a standard and reliable procedure for translating addresses to geographic coordinates. In order to establish a standardized procedure, the Mapping Users' Group has tested six services that we had previously identified as ones that would potentially meet our needs. They are:

1. U.S. Census geocoder
2. Smarty Streets
3. ArcGIS Online
4. Sephamore system
5. Texas A&M Geoservices, and
6. SAS Geocode procedure (Proc Geocode)

Methodology

In order to test these systems, we used a dataset of 16,220 addresses across 32 states that were collected as part of an evaluation of the federal Strong Start program conducted in the Health Policy Center in Summer 2015. The addresses were scrubbed of identifying information such as names and demographic information, leaving only addresses and an anonymous identifier prior to transmission to the MUG team. The addresses were then run through the six geocoding services, in order to compare results.

The following chart shows the measures that were used to assess each category:

Table 1: Categories and Measures for Assessment

Category / Measure	Definition
Completeness	**Share of addresses that were geocoded**
The number and percentage of hits	A hit is defined as an address returned by the geocoding service with coordinates with full confidence, at the street segment or smaller area. NOTE: THIS DOES NOT NECESSARILY MEAN THAT THE MATCH IS CORRECT
The number and percentage of misses	A miss is defined as an address returned by the geocoding service with coordinates matched only to the centroid of the zip code, city, or higher-level geography; or an address returned with no coordinates matched.
Ease of Use	**The level of effort required to use this service across the Urban Institute**
Admin Privileges Needed?	Whether or not administrative privileges are needed to install, update or maintain any aspect of the service, requiring involvement of IT
Other IT assistance?	Whether IT would need to be involved in the operation of the service in any way beyond using administrative privileges
Set-up challenges	Any challenges associated with setting up the service that do not involve IT. These may involve initial setup, or steps required to process data prior to operating the service, such as putting the data into specific forms, or having assisting software installed/running
Maintenance challenges	Any non-IT challenges related to keeping the software up to date and usable long-term
Scalability	The level of ease or difficulty that would be involved in instituting the use of the service for all researchers across Urban
Category: User Friendliness	**The level of ease or difficulty for the individual using the service**
Interface	Whether the user interface for the service creates inefficiencies or challenges for the user
Intuitiveness	The degree to which a typical researcher at the Urban Institute otherwise unfamiliar with the service would be able to figure out its use
Speed	The amount of time to complete the geocoding
Category: Cost	**How much the service costs to obtain and maintain**
Category: Other factors	**Other important topics to consider**
Confidentiality	Many projects that require geocoding at Urban use confidential, individual-level data. This area is to note and assess any challenges to confidentiality that would come with use of the service.
Ability to append census geography	Often geographic analysis involves linking individual records to census data based on geography on the tract, zip code, or other geographic level. Several services include the internal capability of appending fields that include FIPS codes or other census geographic identifiers. In addition, a number provide the opportunity to attach basic demographic data for those geographies.
Added features?	This includes any other features unique to a given service that may provide value added for Urban Institute researchers.

This exercise was not intended to create a final "rank score" but rather to balance the specific pros and cons for each, and to come to a recommendation keeping all aspects in mind.

Overview of Services

1. U.S. Census Geocoder is a web-based service provided by the US Census Bureau, which uses TIGER base files to match addresses to a specific street segment.

- FORMAT: Web-based
- INTENDED USER: Individuals with occasional and limited geocoding needs

2. Smarty Streets is a web-based address verification and listing service which includes geocoding as a component of its product package. It also uses TIGER base files for its match, but it reports the centroid of the ZIP9 code associated with the address, which is generally block-level. This is slightly less accurate than the street segment approach used by other geocoders.

- FORMAT: Web-based
- INTENDED USER: Organizations doing mass mailing

3. ArcGIS Online continues to provide their internal geocoding service, now on a credit-based system. The service is also available through ArcGIS Online and as an API.

- FORMAT: Internal to desktop software, or web-based
- INTENDED USE: broad-based ArcGIS users, researchers, government, etc.

4. Semaphore ZP4 is a DVD-based system designed for mass mailing. In addition to a geocoder, it provides an address validating and correcting service. As with Smarty Streets, coordinates are based on ZIP9 (here called ZIP+4) codes. However, instead of returning a specific coordinate, it provides a range of x and y coordinates, describing, presumably, the edges of the ZIP+4 zones matched with each record.

- FORMAT: DVD-ROM / Desktop PC
- INTENDED USER: Organizations doing mass mailing

5. Texas A&M Geoservices provides a web-based geocoder based on locally-developed algorithms, using TIGER base files along with the Boundary Solutions 2012 National Parcel File, allowing for parcel-specific[1] results.

- FORMAT: web-based
- INTENDED USER: researchers

6. SAS Geocode Procedure allows the user to define the base data, and match addresses to geographic coordinates from within SAS. Primary use is with TIGER base files. The 2014 TIGER release was used in this test.

- FORMAT: Internal to SAS desktop software
- INTENDED USER: SAS users, researchers

[1] The term "parcel" is used by the Texas A&M Geocoding service, but the geocoding process does not attach local parcel IDs.

Assessment Results by Category

Completeness and Consistency Measures

Based exclusively on number of matches, ArcGIS provides the highest number of hits at 15,828, while the SAS Proc Geocode provides the lowest at 13,002. As mentioned above, this does not mean that the ArcGIS Online system necessarily provides more *correct* matches than the other services, but rather that it provides matches for more of the addresses. However, based on ArcGIS's scoring system, 15,828 of these matches have a score of 100, meaning that they are the highest quality match available. In fact, 99% (16,604) records matched with a score of 90 or above. Sixty-five percent of the addresses were geocoded by all 6 systems. A full breakdown of the match rates by geographic type is included as an appendix to this memo.

The following table shows the number and percent of hits and misses for each geocoding service.

Table 2: Completeness Results

Service	Hits		Misses	
	Number	Percent	Number	Percent
Census Geocoder	13,134	81.0%	3,086	19.0%
Smarty Streets	14,370	88.6%	1,850	11.4%
ArcGIS Online	15,828	97.5%	392	2.4%
Semaphore ZP4 System	13,720	84.6%	2,500	15.4%
Texas A&M Geocoding	13,727	84.6%	2,493	15.4%
SAS Proc Geocode	13,002	80.2	3,218	19.8%
Hit Overlap	10,547	65.0%	5,673	35.0%

In addition to the accuracy, we also examined the average z-score for each service, using only the addresses that returned a hit across all services for consistency. This was calculated by finding the average latitude and longitude results for each record from the six services, standardizing their differences into z-scores, and taking the mean of those across all records. This measure can be interpreted as how much, on average, the results for each service differ from the overall average, in terms of standard deviations. The following table shows the latitude and longitude z-scores for each service. A score of zero would mean no variation from the average, where a score of 1 would indicate a full standard deviation difference, on average.

Table 3: Consistency Results

Service	Average Z-score	
	Latitude	Longitude
N=10,547		
Census Geocoder	.005	.011
Smarty Streets	.008	.012
ArcGIS Online	.02	.04
Semaphore ZP4 System	.001	.005
Texas A&M Geocoding	-.004	.008
SAS Proc Geocode	-.001	.007

ArcGIS results differed from the mean scores, on average, slightly more than any of the other services, for latitude and longitude. With the tests completed for this memo, we cannot provide a definitive explanation for this. One possible explanation for this difference could be the base map file used. Public geocoding systems often use the publicly available Tiger file. Esri, however, is known for using its own proprietary file. Other differences may be the algorithms used to resolve non-standard addresses, such as abbreviations such as "MLK", missing quadrants, and the decision rules each service uses on offset of the coordinates relative to the street segment.

Ease of Use Measures

In this area, each service has areas of ease and difficulty.

The Census Geocoder requires no IT involvement in setup and is easily scalable due to its web-based format and no requirement for a user account. Although costs will be discussed in the following section, it should be noted that because this program is free, it does not require maintenance of credits or an account balance, which removes a challenge to maintenance and scalability. However, the service accepts limited file formats, and limits submissions to 1,000 records at a time, requiring the user to subdivide and reconcatenate large datasets. In addition, results are returned without column headings.

The other web-based services (Texas A&M, Smarty Streets and ArcGIS Online) share the benefit of not needing IT assistance with installation. However, both Texas A&M and ArcGIS Online require users to purchase credits which are used to pay for the service, on a per-record basis. Maintenance would require IT or another body monitoring credits in order to ensure there are enough available when they are needed. In addition, in order to access the same pool of credits, Smarty Streets and Texas A&M services would require that all users at Urban log in under the same credentials, making tracking credit usage and accounting to specific projects difficult, and potentially meaning that only one person could be logged into the system at a time. This also presents a challenge for accounting and how to bill credits back to specific UI project codes. The ArcGIS system allows for multiple users to be connected to the same central account, but this would still require IT or another body monitoring credit levels and granting access to new users.

The Semaphore system does not require credits and does not charge per record geocoded. However, due to its nature as a DVD-ROM based program, IT assistance would be required for installation. In addition, a single subscription only pays for a single DVD, which includes all base data, and requires monthly updating. Sharing a single disc to update the base data across Urban would create significant challenges to scaling.

Finally, the SAS-based geocode procedure does not require IT assistance to install, although for scalability, IT involvement in placing the base files into a broadly accessible space on the SAS or other server would require some front-end effort (files are approximately 12GB in size). In addition, the procedure would require some skill in SAS programming. Maintenance involves a yearly update of the base files with the annual release of new TIGER files.

The following table summarizes observations regarding the ease of use of the tested geocoding services.

Table 4: Ease of Use Results

Service	Ease of Use				
	Admin Privileges Needed?	**Other IT assistance?**	**Set-up Challenges**	**Maintenance Challenges**	**Scalability**
Census Geocoder	no	no	- Addresses must be in CSV with standardized headings - maximum of 1000 records at a time	none	Universal access
Smarty Streets	no	no	none	none	All users would share the same log-in
ArcGIS Online	no	- maintaining credits/ payment - administering urban general account	Establishing credentials for all users and providing access to Urban account	Controlling expenditure of credits and ensuring availability of credits when needed	Minimal uptake issues due to program and web-based format; potential budgeting problem
Semaphore ZP4 system	yes	no	Access to installation DVD	Necessary monthly updating of addresses	Difficult due to monthly updates
Texas A&M Geocoding services	no	Maintaining account, credit access	none	Requires a balance of credits	Web-based, easily accessible; budgeting may be difficult
SAS Proc Geocode	no	Potential placement of base files on server (~12GB)	Base reference files are large (3-4 files of 1-2 GB)	Yearly update of reference files	Challenges due to size of files and lack of universal SAS usage

User Friendliness Measures

As is visible in this table, the services vary greatly in terms of the type of interface, level of intuitiveness and the speed at which they execute the geocoding. The slowest service – the census geocoder—included the time required to subdivide and reconcatenate the address list, which is necessary due to the service's 1000-address limit per submission. The fastest services by far were Smarty Streets and ArcGIS Online. It should be noted that the SAS geocode procedure ran on a local PC drive during testing – running it off of a central server could be significantly faster.

The complexity of the Semaphore system cannot be overstated; while repeated use would be relatively easy, the initial time needed to learn the terminology, organization and operation of the program is a significant weakness. While the ArcGIS online geocoding service will be familiar for users of ArcGIS, it could create challenges if people not trained in ArcGIS were to attempt to geocode.
The following table summarizes observations regarding the user friendliness of the tested geocoding services.

Table 5: User Friendliness Results

Service	**User Friendliness**		
	Interface	**Intuitiveness**	**Speed**
Census Geocoder	Simple but documentation and instruction are hard to find; output does not have column headings	low	30 minutes
Smarty Streets	cut-and-paste into fields on webpage	easy web interface	<1 minute
ArcGIS Online	easy for those familiar with ArcGIS	no - must read directions	1 minute
Semaphore ZP4 system	GUI, but complex	Complex and technical; requires significant time with documentation	5 minutes
Texas A&M Geocoding services	easy	easy web interface	22 minutes
SAS Proc Geocode	familiar for SAS users	familiar for SAS users	7 minutes 56 seconds

Cost Measures

Comparing the true cost of the geocoding services is challenging due to complex price structures. As such, the following table categorizes the services by price structure, cost scale, and the cost per geocode, where applicable.

Table 6: Cost Results

Service	Price structure	Cost scale	Cost per geocode
Census Geocoder	Free	n/a	free
Smarty Streets	Price levels for number of geocodes available per month	250=free/month 500 =$20/Month 1000 =$30/month 5000 = $50/month 10,000=$80/month 25,000=$200/month 50,000=$300/month 100,000=$500/month unlimited=$1000/month	$0.04/geocode or less
ArcGIS Online	Users purchase credits, which are then used to pay for geocodes at a rate of 1,000 geocodes for 40 credits. Unlimited use also available, with prices for single user or server access.	1000 credits = $100 Full server access = $75,000/year Single user access = $6,000/year	$0.004/geocode (1,000 geocodes cost 40 credits)
Semaphore ZP4 system	Monthly or annual subscription	$99/month	n/a
Texas A&M Geocoding services	Users purchase credits, which are then used to pay for geocodes at a rate of 1 geocode per credit, sold in monthly batches. Partner Program for nonprofits: free geocoding in batches of 2,500 (manually replenished)- must put public attribution of services on website	2,500 credits = free 50,000 credits = $95/month 100,000 credits =$180/month 40,000 credits=$560 unlimited =$1000/month	$.002-$.0014/geocode
SAS Proc Geocode	Free	n/a	free

Based on price, Smarty Streets is by far the most expensive geocoder on a per-unit basis at $0.04 per geocode, particularly because the monthly purchase structure means that ensuring that there would be enough geocodes available would likely require buying more than would be used. While ArcGIS Online is a less expensive option in terms of the per-unit price, the cost for unlimited access is the highest.

Of the paid options, the Texas A&M geoservices provide the best value for their price. They also offer a partner service that allows for unlimited geocodes free, but does so by releasing 2,500 credits to partners at a time. The free credits do not automatically replenish; the user must add the credits to their account manually. No free credits are guaranteed in the future.

Other Factors

Confidentiality

Breach of confidentiality was initially a concern when geocoding datasets with personally identifying information using web-based services due to the risk of interception during transmission or use by the service providers. However, further inquiry and conversations with members of the Urban Institute Institutional Review Board revealed that there is no reasonable risk of a breach of confidentiality when using these services, so long as address fields are separated from all other data except a non-traceable identification number. Users should create a new ID with no intrinsic meaning for this purpose, if not already available in the dataset. The non-web-based services reviewed do not pose any threat to confidentiality other than standard risks of working with the data inside the Urban Institute.

Ability to append census geography and additional features

The following table summarizes observations regarding the ability to append census geographic identifiers (to aid in use for summarizing and linking to other data) and any other features of the services that could be of use to the Urban Institute.

Table 7: Other Features

Service	Ability to append census geographic identifiers	Additional features?
Census Geocoder	yes	choice of source geography.
Smarty Streets	no	address cleaning
ArcGIS Online	no	works internally to ArcGIS
Semaphore ZP4 system	yes	address cleaning
Texas A&M Geocoding services	yes	options to allow for "ties" / has online mapping tool incorporated
SAS Proc Geocode	yes	Works internally to SAS

While the Census Geocoder, Semaphore, Texas A&M, and the SAS geocode procedure allow for the automatic appending of census geographic identifiers, Smarty Streets and ArcGIS Online do not.

None of the services have additional features that are essential. Both Semaphore and Smarty Streets offer address cleaning, as they are designed for mass mailing, not research. As an academic service, Texas A&M provides an online mapping tool in addition to a more nuanced matching system that allows for greater control over the algorithm and "ties" in matching when an address is equally closely matched to two options. ArcGIS online and Texas A&M both offer online mapping tools in addition to the dataset.

Perhaps the most significant additional features are in SAS Proc Geocode and ArcGIS, in that the services work within the desktop programs, allowing the user to move seamlessly from geocoding to mapping in ArcGIS or further data manipulation or analysis in SAS.

Comparison and Assessment

A number of services appear ill-suited for use at the Urban Institute. These include the Semaphore due to its challenging interface, the need for IT involvement for set-up and difficulty with scalability due to its distribution on a DVD-ROM, and Smarty Streets, which provides similar services to the other web-based services, but with a less flexible price structure.

The Census Geocoder and the Texas A&M geocoding service partner plan provide free service that is within the average range of accuracy. However, each suffers from the limitation of only accepting a relatively small number (1,000 for the Census Geocoder, 2,500 for the Texas A&M service) of addresses at a time.

For SAS users, the Geocode procedure provides a viable option, in cases where high levels of matches are not essential, or where time is available to check and clean unmatched records. As a component of SAS, Proc Geocode poses no additional cost. It does not require administrative privileges for installation, and after running an initial script to link the base files to the program, operates in a similar fashion to other procedures. Working with IT staff, the base TIGER files could be saved onto a partition of the SAS server, making the procedure usable without downloading the base reference files. Although it had the lowest hit rate in our selection, the relative ease and low costs involved with its use are significant benefits.

It is possible that SAS Proc Geocode may not be an option at times, either because of a need for very high match levels, or because a staff member with basic SAS proficiency is unavailable. In these cases, the paid services at ArcGIS Online or Texas A&M Geocoding services would be the best option. Both services allow for large-batch geocoding. Texas A&M is the least expensive of the paid services, and offers options for greater control. Conversely, ArcGIS Online provides easy integration into the mapping workflow, and possibly higher match rates and greater accuracy.

The following chart shows particular areas of strength and weakness for each service.

Table 8: Comparison of Results

	Census Geocoder	**Smarty Streets**	**ArcGIS Online**	**Semaphore ZP4 system**	**Texas A&M Geocoding services**	**SAS Proc Geocode**
number of "hits"	low	average	high	average	average	low
admin privileges needed?	no	no	no	yes	no	no
other IT assistance?	no	no	yes	no	no	yes
set-up challenges	moderate	none	high	high	none	moderate
maintenance challenges	none	none	moderate	high	moderate	low
Scalability - to all of Urban	easily scalable	easily scalable	partially scalable	difficult to scale	easily scalable	scalable
Interface	some challenges	simple	some challenges	complex	simple	some challenges
Intuitiveness	low	high	very low	very low	high	n/a
Speed	slow	fast	fast	fast	slow	moderate
Cost	free	high	High	moderate	moderate	Free
Confidentiality	no issues	no issues	no issues	no issues	no issues	no issues
Ability to append census geography	yes	no	no	yes	yes	yes
Added features?	some	some	highly useful	some	highly useful	highly useful

Limitations and Future Exploration

It should be noted that these tests were only performed on one dataset. Moving forward, Mapping Users Group managers will informally monitor Proc Geocode and Esri results for future geocoding efforts to determine if long-term outcomes differ from this initial attempt.

In the future, we would like to explore more tests of accuracy, and to better understand the low overlapping hit rate. We want to test other systems, including those that that pull from the Google Maps API, such as R, an open source statistical package, with a geocoding function similar to SAS (the geocode() function from the ggmap library).

The Google's Geocoding API has a limit of 2,500 free requests per day, but offers a paid option to submit more records. There may be other open source GIS options or locally-developed tools like the Master Address Repository's Geocoder for the District of Columbia or Geosupport in New York City.

We will also be developing some tips for cleaning addresses to improve your hit rate and accuracy of placement.

Conclusion

In light of this analysis, the Mapping User's Group recommends the use the of the SAS geocode procedure as a primary geocoding tool for files greater than 2,500 records.

SAS Geocode requires minimal financial burden as it is internal to software already in use within Urban, and utilizes public data sources. We plan to prepare a well-documented sample program and to explore with IT the possibility of storing the base datasets in a central location easily accessible to all Centers. SAS Geocode comes with some limitations: specifically, the tested "hit rate" was lower than other tools. However, this may vary based on the cleanness of the inputted addresses.

We recommend using Esri's ArcGIS online geocoding system whenever a project necessitates extremely high levels of completeness. In cases where the ArcGIS online geocoding system is necessary, researchers will need to include the cost of geocoding in the project budget. The Mapping User's Group recommends keeping an account with credit available (or easily obtainable) at ArcGIS Online. We would further advise that each center designate one person to establish a center-wide login rather than have accounts from several individual staff members.

For smaller files, the Census Geocoder and the Texas A&M Geocoding Services are easy to use and free.