Northwestern
GLOBAL POVERTY
RESEARCH LAB

ipa
INNOVATIONS FOR POVERTY ACTION

Geospatial Tools for Creating Sample Frames

Case Study: "Proyecto Mi Barrio" Phone Survey, Medellín, Colombia

Researchers often have a database of addresses as a starting point for sampling design. They often want to sample from some geographic unit like a neighborhood. To do this, they need geospatial data, which is a set of coordinates that represent the boundaries of the geographic unit, and software capable of locating these coordinates. This is called geolocation. This brief describes the performance of two tools that IPA Colombia used to sample respondents based on geolocated data. The Google Geolocation API and QGIS outperformed ArcGIS in terms of geolocation accuracy by a substantive margin.

Motivation and Recommendations

Research teams has several software options for the two necessary steps in the process of creating a sample frame by geolocating addresses and converting geolocations into GPS points. The chosen software's geolocation accuracy, and the chosen statistical package's spatial data management tools, can have a large impact on sample frames, response rates, and results. Besides this accuracy, the researcher may need to consider costs (which depend on the number of geolocation points) and ease of use.

In the Medellín context, the research team identified two top recommended packages to test: Google Geolocation API with QGIS and ArcGIS. When it came time to match GPS points to spatial units, the team also identified two options for geographic merging: st_join for R and geoinpoly for Stata. Both performed evenly in this context.

Cost and privacy are common concerns with geolocation. QGIS and R are free to use, while Stata and ArcGIS must be licensed; the Google Geolocation API costs USD $0.0004-0.0005 per address depending on volume; ArcGIS is priced similarly at USD $0.0004. Both options were considered secure in this context. But all software should be evaluated from the perspective of your governing IRB's data security protocol, especially as addresses are sensitive data.

Results

The research team randomly selected 100 addresses from its potential respondent database, and then geolocated these hundred addresses using both ArcGIS and QGIS/Google API. The team then manually searched for these addresses using Google Maps, and recorded the GPS locations to establish a baseline for comparing the results of the software's results. Table 1 below describes the distribution of geolocation error by software used, in meters, showing that the Google API and QGIS was more accurate than ArcGIS by more than 50 percent, although both located addressed on average to within 4.5 meters.

Table 1: Descriptions of error rates (meters) between geolocation options and true location

Software	N	Mean	Standard Deviation	Min.	25th Percentile	Median	75th Percentile	Max
Google API (QGIS)	99	2.15	3.59	0.01	0.02	0.10	0.10	19.67
ArcGIS Pro	100	4.40	4.06	0.01	0.39	3.83	7.28	19.14

After the team geolocated the addresses, they tested the performance of two options for performing a "spatial join", which in this case meant matching the GPS points to the spatial units designated for surveying. While there were several options to conduct this process, the team decided to verify the performance between `st_join` for R and `geoinpoly` for Stata. Both tools had no differences in performance for this task.

This document was made possible by the work of David Cerero and Juan Pablo Mesa-Mejía.
IPA's phone survey methods case studies are part of a series on best practices on implementing surveys using computer-assisted telephone interviewing (CATI) and other remote survey modes. These case studies are made possible with the generous support from and collaboration with Northwestern University's Global Poverty Research Lab (GPRL).