Geocoding of the Afrobarometer

Eivind Hammersmark Olsen

February 2, 2016

This file describes the geocoding of the Afrobarometer for "Mining and local corruption in Africa".

Overview of the geocoding process

Raw files are in the "Data/Afrobarometer/Raw data" folder, with .7z files containing what we received from Afrobarometer (the contact person is Ms. Carmen Alpin). The contained .sav SPSS files were converted to Stata .dta format using StatTransfer. For Round 2 there is one file for each country, so these were appended into one dataset in "append.do".

The geocoding process relies on a user-written Stata program called "geocode3", which permits batch geocoding. (The program has been removed from SSC, so it is included here for reference.) It connects to the Google Maps API V3, sends placenames to the server, and retrieves coordinates for those placenames in json format. At the time of geocoding, the API had a limit of 2500 requests per IP adress per day, so the operation was split into parts.

Round 5 was geocoded country by country due to different definitions of levels of administration in this round (level 1 through 5). It is not known whether the program still permits batch geocoding through the API, because it seems that Google may have put some addtional restrictions on their API since our geocoding was undertaken in 2014/2015.

Some observations in Sierra Leone are matched to enumeration areas in "ea-match_sierraleone.do". This data was collected from `http://www.sl-wash.org/`. The website seems to no longer be operational.[1]

[1] See also `http://www.washlearningsl.org/sierra-leone-waterpoint-report-review-version-2012/`.

The final output of the geocoding for round X is the file "rX_merged_geocoded.dta", which includes the geocodes and all survey questions. These are the source files for the construction of the analysis dataset. The file also includes information about georeferencing quality in "match_quality". We only use observations with a match quality of "townvill" or "ea", although for some (small) countries, matching at level 3 or level 4 is likely precise enough for this application.

Specifics of the geocoding

All do-files are included in "Data/Afrobarometer/Geocoding/Round X" for reference. The geocoding was in some sense a trial and error process, in an attempt to understand the Google maps search algorithms, and the outcome of this process culminated in the included do-files.

Variable definitions

country Name of country, as reported by Afrobarometer.

region Highest level administrative division, right below country level, sometimes called province. Corresponds to level 1 in GADM for most countries.[2]

district Second level administrative division, below region/province.

townvill Town, village or urban subdivision (borough or quarter).

Round 2, 3, and 4 Rounds 2-4 have a similar structure on location information, so they were geocoded almost identically. First, we collapse the data on townvill, region, district and country, to make the algorithm search faster. We then concatenate *townvill* with *region* and *country* into the variable "town_region_country", with a space delimiter. (*geocode3* automatically replaces spaces with "+" signs, to conform with the Google API.) We also concatenate into a variable "town_district_country". These two variables form the primary search keys in the geocoding.

The two search keys are cleaned, by replacing non-unicode characters identified using *charlist.ado* with their non-accented counterparts (e.g. 'é' is replaced by 'e'). The dataset is split into two, with the first 2,500 observations in one dataset, and the rest saved in another dataset, to facilitate geocoding from different IP-adresses.

geocode3 is run on *town_region_country*. Some observations are geocoded to the wrong country, so these are returned back to "missing" status. The rest of the

[2] http://www.gadm.org.

observations are marked with *match_quality* = townvill. For those respondents that didn't get coordinates in the first attempt, we run the *geocode3* program again, this time on *town_district_country*. We do not search only using the town name and country, to prevent Google from having to pick from too many duplicated town names. (We are not sure about how Google selects from duplicates.)

Round 5 The round 5 dataset has a slightly different structure than rounds 2-4. In addition, this geocoding was performed at a much later stage, due to round 5 being provided to us much later than round 2-4. Hence, the keys used to geocode respondents were slightly different. In place of *town_region_country* we used *town_lvl1_ctry*, where "lvl1" corresponds to the first level administrative division in the Afrobarometer, below region/province level. For some countries this level corresponds to level 1 in GADM, in other it corresponds to level 2 in GADM.[3]

For respondents that were not assigned coordinates in the first round of geocoding, we ran the algorithm using *lvlX_ctry* as well, where X is level 2, 3 or 4, depending on the existence of these levels in a particular country. The coordinates we got from this search were marked with "level X", and not used in the analysis dataset.

Sierra Leone and South Africa For South Africa we have access to maps of census enumeration areas for censuses in 1996, 2001 and 2011, as polygons in GIS shapefiles. Enumeration areas are the primary sampling unit of the Afrobarometer, and gives us certain locations of respondents, at least for enumeration areas that are relatively small. It also provides close to 100% retrieval of coordinates, because we can match on the unique enumeration area code provided in the Afrobarometer location information. Centroids of enumeration areas in the 1996 census were used for Afrobarometer round 2, and centroids from the 2001 census were used for rounds 2.5-5. The South African respondents are assigned these coordinates in "02a_prenearstat.do", before the geographical matching with the mines.

Around 320 respondents in Sierra Leone were also given coordinates based on enumeration area centroids, based on data from `http://www.sl-wash.org/`.

[3] Respondents located in "Random villages" were given location level 1 + country in the concatenated variable, which means that these were matched at the district level, but still have *match_quality* = townvill. This does not affect too many respondents, however.