The Surpluse People Project

Improving the Geographic Variables

Alex Montgomery, DataFirst

October 2016

DataFirst

1 Introduction

Attempts were made to encode all of the geographic variables in the dataset and match them to a database of South African administrative areas as they were at the time of Census 2011. This proved difficult at times for two broad reasons. First, as the overview report of the SPP notes:

> "Where place names were asked for, the answers given were often found to be uninterpretable and, although these were coded for the computer, these codes and the tables incorporating them should be disregarded. Regional SPP groups were left to interpret this information manually where they could and wished to. The rest of the information obtained through this questionnaire appears generally sound, although there were fluctuations between and within the survey area for a number of variables (Surplus People Project (South Africa), 1983, p. 44)."

Second, it was not always possible to get a clean match to the (relatively) current constructed municipal demarcation database. There were a number of reasons for this which depended on the variable in question. Some issues idiosyncratic to each variable will be described later in this document.

2 General Strategy for Encoding Geography Variable

First, all values of the geographic variable were put into upper case and trimmed of leading, trailing and internal blanks. Unique values were then matched into groups based on their similarities. This grouping involved a custom written program that made extensive use of Reij (2010). The program matches values together using their Levenshtein edit distance, which is essentially a metric that captures how many edits it requires to make one string like another.

We then attempt to merge this with four different variables that demarcated administrative boundaries at the time of Census 2011. These levels are:

- Small Place
- Main Place
- Local Municipality
- District Municipality

The reason for attempting to merge on multiple levels of geographic disaggregation is that entries for the variable were captured from respondents who tended to give them at varying degrees of accuracy. For the most part matches were made on small place or main place. The clean matches were then coded as a set of four alternative geographic variables that gave respondent location (with the caveat that this is relative to 2011 demarcations) as accurately and unambiguously as possible

Some manual oversight was necessary, however, as the there are multiple parts of South Africa with the same place name. For example, there is an Athlone (main place) in the City of Cape Town, eThekwini, and Umgungundlovu (local municipalities). These ambiguities meant that place names were not encoded using the geography match in all cases.

This isn't to say that ambiguities were always left unencoded. Sometimes, matching was done on a discretionary basis. For example, a variable in one of the datafiles, q1a_21, had numerous responses with varied spellings of "Welkom". There are only two possible matches for Welkom in the geography database. One of these is a large town with a population of roughly 200,000 in 2011 (Statistics South Africa, 2015). The other is a small area in the Northern Cape with a population of 380. In this case, the assumption was made that the respondent meant the larger Welkom. It may be possible to make a guess as to the correct choice based on other corroborating information attached to the respondent/household, but this process has been deferred for later versions of the data as it may benefit from end user input.

After completing all the "good" matches and excluding the irreconcilably ambiguous ones attention was focused on the non-matched values and figuring out why they weren't caught by the fuzzy match done initially or the exact match done afterwards. The most likely explanation was usually that there were only idiosyncratic spellings of the place which were too far in terms of Levenshtein edit distance to register as a match. Some issues of ambiguity also arose because multiple levels of geographic disaggregation had the same name. That is, a main place may have the same literal name as a small place. This would match twice with the geographic variable value and register as an ambiguity despite communicating the same thing. The policy here was to manually choose the match with the smaller of the matched areas.

The remaining values were done on a discretionary basis. Fortunately, Google's fuzzy matching of strings is far more sophisticated and could often locate the unmatched response to a South African small place, main place, local municipality, or district municipality. For the more obscure small areas portions of the unmatched value were searched on the South African geography database and assigned that place name if there weren't multiple matches. The original and matching strings are all given in the data so users are empowered to scrutinise the matching process.

The still remaining non-matches for the variable were then appraised. In some cases it was because the place given was outside of South Africa. In other cases, it was because of ambiguity in the question leading to responses that were not places in the database. For example, q1a_21 asked, "where was your last place of work?" Some respondents took that to mean the name of the company or individual for whom they worked. This resulted in responses like "3M ATLANTIS" (a stationery manufacturer - note how we can still get the place name here) or "John Streabal" (assumed to be a person? a defunct store?). Obviously when companies were involved the derived geographic variables were set to missing.

With some variables, e.g. aa_areaname in the household level datafiles, the places given no longer exist. Matching these places to current demarcations did not work. To illustrate, some values for the variable have values "Released Area 33". This was not in our database, nor could it be found on Google Maps. Some further investigation yielded the following information about the place from two separate newspaper article sources:

> "I was born and lived for nearly 40 years in the Phoenix settlement established by Mahatma Gandhi, my grandfather, in an area now known as Bambayi in Inanda, KwaZulu-Natal. The apartheid government designated it Released Area No 33, I think because the area was not declared for any particular race group, so people of all races lived there an island in apartheid South Africa. Phoenix was a 100-acre property, with the adjoining settlements of Ohlange, established by Dr John Langalibalele Dube, and Shembe, established by Nkosi Isaiah Shembe (Gandhi, 2014)."
>
> "Many people have lived in Inanda for a number of years have contributed to the community and have permanent jobs in Durban. Already over 20 residents have been charged under the 1936 Trust and Land Act and the 1951 Illegal Squatting Act. The policy to evict people from RA 33 (Released Area 33) in Inanda contradicts earlier government statements to develop the area. (South African Students Press Union, 1982)"

When some of the alternative place names given above were matched against the 2011 database geography databases it become clear that the demarcations were not congruent (in the literal, geometric sense) to the old area. There is no longer a current administrative boundary that matches the one carved into the "former homeland" in which it was located. A decision was made to simply encode the area names variables and leave geographic matching at the user's discretion.

References

GANDHI, ELA. 2014. Twenty years of democracy: Much has changed, but not enough. April.

REIJ, J. 2010. strgroup: Stats module to match strings based on their Levenshtein edit distances.

SOUTH AFRICAN STUDENTS PRESS UNION. 1982. Thousands will be evicted from Inanda homes. Auguest.

STATISTICS SOUTH AFRICA. 2015. *South African Census 2011.*

SURPLUS PEOPLE PROJECT (SOUTH AFRICA). 1983 (January). *Forced Removals in South Africa.*