Need for mapping

Facilities location not always known

There is some existing data on names available

Maximizing the use of existing data to get desired rough cartography of facilities

# Data

**Facilities Masterlist:** The problem we are willing to tackle presents itself essentially in the form of a list of facilities. In the case we are talking about, this list obtained through the DHIS2 database used for collection and management of health data in Nigeria. Each facilities are located in a Ward, each of which are located in Local Government Areas (LGA) which are situated in States. The lowest level of cartography available at that point is the borders of LGA. The average surface of a LGA is XXX km2. To our knowledge, no complete data was available as to geographic borders of Wards.

**OpenStreetMap:** OpenStreetMap (OSM) is an open and collaborative mapping platform. Users can contribute mapping features of areas they know. This data can be accessed openly online. Additionally, OSM data can easily be downloaded and parsed using R package OSMAR. We downloaded all OSM data lying inside of the bounding box of LGA borders maps.

**Validation set:** In order to validate our methods, we had to create a validation set for a number of facilities. Some facilities are mapped in OSM, by projects such as XXX. We matched manually the facilities in OSM with facilities in our masterlist. XXX facilities could be matched and constitute our validation set.

# Methods

The central hypothesis for our approach is that most health facilities in the public system have a name attached to their geographical location. Using this intuition, we defined a sequence of algorithms to map facilities. Each step in the sequence potentially relaxes precision of location, and thus we applied these methods sequentially, only attending to map in each stage facilities that had not been mapped previously. The matching steps are as follow:

1. *Exact matching:* This first step is the most conservative. For each LGA, we attend to get an exact unique name match in the features names from OSM. If an exact match is returned, the facility is set as being at this location.
2. *Simplified name matching:* The second step was quite similar with the first, only stripping names of geographic features of their type indication. For example XXX hill was simplified into XXX. Here again, exact unique matching stratified on LGA was sought.
3. *Multi-match Centroid:* In the next step, we took into consideration facilities that returned multiple exact matches in the features. At that point, we set the location of the facility at the centroid of all features returning a match.
4. *Ward Centroid:* The fourth step was similar to the previous, but applied to facilities returning multiple partial matches in step 2.

Matching was made on entities identified as being places in OSM. The steps were applied sequentially. At each step, all entities in OSM data were kept, but only facility that had not yet been matched were kept in the facility Masterlist.

At step 1 and 2, facilities with multiple matches not considered matched, and the centroid was kept at step 3.

This analysis was carried out in R. R code is available on XXX.

# Results

Table XXX shows the results for each step of the matching algorithm. We evaluated the performance of each matching looking at the linear distance between matching location and actual location of health facilities in the validation set.

We see that exact matching of the first stage gives most of the early matching. Later on, the convex hull approach allows to geolocalize XX% of the facilities, without losing too much precision in the process.

|  |  |  |  |
| --- | --- | --- | --- |
| Step | N matched | Median precision on this step | % of matches < 5km |
| Exact Matching | 8680 |  |  |
| Simplified name Matching | 754 |  |  |
| Multi-match Centroid | 544 |  |  |
| Ward Centroid | 14328 |  |  |
| Overall | 24306 |  |  |

Table 1 - Matching result for each step

Tab 2 Final results by state

Tab 3 Results by facility type

Model 1: prediction validity

Regarder les prédicteurs dans les covariables et voir si ça améliore quelque chose.

# Discussion

These results are of great interest as they show that in a data poor situation, the availability of widely available data sources can be leveraged to improve knowledge and complete missing information. The methods used in this work are specific to the matter of interest, but can be modified to answer different needs and situation. The unavailability of exact ward borders drove us to explore convex hull solutions for the matching of most facilities. In other contexts, this step may not be usable, if low level border data are available. Conversely, our approach of name matching works mainly because of the way health facilities are named.

Working with