Interim Report

Estimation of health conditions in Cote D'Ivoire and Senegal using Call Detail Records (CDR)

Name: Opeoluwa Flora Fajemirokun
Supervisor: Licia Capra

Progress made to date

In the early stages of the project, I mainly focused on studying the Demographic & Health Surveys (DHS) survey data which contained relevant health information per household in Senegal and Cote D'Ivoire. This involved understanding the way the surveys were carried out, the information contained in the survey files, the file formats and which applications I would need to analyse the data. Once I got comfortable with the applications required – STATA (a statistical package) and qGIS (used for plotting spatial/geographic data), I started the pre-processing of the DHS data.

At the start of the pre-processing, using STATA, I was able to decide on the metrics I would be using to study malaria positivity rates (blood test & rapid test results), HIV rate (HIV test results), child mortality rate (number of children who died under age 3 months) and women's access to health (number of antenatal care visits during pregnancy). Prior to computing the spatial aggregates of the metrics at the various administrative units of analysis (levels 1 to 3), I needed to determine which method of counting the metric values per region and which method of aggregation were most accurate.

To compute these spatial aggregates, I used qGIS to determine for each cluster, the region it was contained in per administrative level. This was important as after calculating the rate of a metric (i.e. HIV rate) for all clusters, I could then determine the rate of that metric for each region. Following this, I computed Voronoi polygons for each survey cluster, this was necessary as clusters are not always contained in a region but could extend to multiple regions. Using Voronoi polygons, I would be able to determine the proportion of a metric in the regions in overlaps in. Using python scripts, for each metric, I calculated the spatial aggregates by counting the number of positive cases in each cluster where the clusters are represented as points (present in one region) and where the clusters are represented as Voronoi polygons (can overlap multiple regions). Using Kendall Correlation, I compared the results for each region at all administrative levels and determined that the high positive correlation meant that the use of Voronoi polygons would be sufficient and capture more information.

Following the decision on which metrics I would focus on, I computed the spatial aggregates of the metrics at the various administrative units of analysis available (administrative levels 1 to 3). However, to prior to computing these spatial aggregates, I needed to determine how to count the values corresponding to each metric per region – either by simply counting the number of occurrences in each row, normalising by a sample weight provided in the survey data or normalising by the cluster population. As a result, I proceeded to count using row counting as the other methods performed poorly at lower levels of granularity. I then proceeded to also perform the computations on the Senegal DHS data. To reduce repetition and redundancy, the analysis for Senegal data reuses the same functions I wrote for processing the Cote D'Ivoire data. This has led me to the next stage of the project where I currently stand – computing the Call Detail Records (CDR) metrics for the various administrative units (levels 1 to 3).

As with stage 1, I started research on the CDR metrics and what information I would need to extract from the data – data about cell phone towers and the number of calls sent over a period of 3 months. As this data was much larger than the DHS data and contained over 2.5 billion rows in total. I had severe performance bottlenecks due to the low computational power of my machine, as a result I looked into the use of a flexible Python parallel computing library – Dask to perform my initial computations. Prior to computation of the metrics, I cleaned the data by removing rows which contained unknown cell towers, to reduce the number of rows for the purpose of increasing performance.

The metrics I need to calculate are the strength of the cell towers per administrative area, gravity residual, network advantage and each administrative area's level of introversion. I have currently completed the computation for activity and I am currently working on the computation of network advantage.

Remaining work to be done before the final report deadline

- I need to complete the computation of the last two CDR metrics – gravity residual and introversion
- I need to build the models that will use the CDR metrics to estimate the DHS metrics
- I need to ensure that my code is written cleanly and avoids unnecessary repetition and redundancy