**Ope – Analysis of Call Detail Records (CDR)**

Knowledge of the distribution of wealth and poverty in developing countries country is often acquired by means of manually collected household survey data (e.g., Demographic and Health Surveys). However, the cost associated with this method is such that poorer nations can only run such surveys every 10 years or more, and on sample sizes rarely large enough to provide statistically significant estimates for small geographical units, such as municipalities and villages. In our recent work, we have described a method to aggregate mobile phone data at cell tower level, and to extract features from such aggregated data that can then be used as proxy indicators of poverty levels. Using telecommunication data made available by the Data for Development Challenge for two developing countries (namely, Cote D'Ivoire and Senegal), we have quantitatively demonstrated the existence of a strong correlation between these new features and poverty indicators derived from costly census data. The goal of this project is to expand this methodology, to produce estimates about health conditions in these countries – specifically, studying the incidence of malaria cases, child mortality rate, and women’s access to health.

**Essential background reading**

* V. Blondel et al. “Data for Development: the D4D Challenge on Mobile Phone Data”. URL: <https://arxiv.org/abs/1210.0137>
* C. Smith-Clarke, A. Mashhadi and L. Capra. "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks". In 32nd ACM SIGCHI Conference. Toronto, Canada. April 2014.

URL: <https://dl.acm.org/citation.cfm?id=2557358>

* C. Smith-Clarke and L. Capra. "Beyond the baseline: Establishing the value in mobile phone based poverty estimates". In 25th International World Wide Web Conference (WWW 2016). Montreal, Canada. April 2016.

<https://dl.acm.org/citation.cfm?id=2883076>

**[Optional] CDR data and challenge goals / outcomes**

* <http://www.d4d.orange.com/>
* Book of abstracts for 2013 and 2015 challenges (shared electronic copies)

**Data**

* CDR data directly provided
* Geographic boundaries at different administrative levels provided
* DHS data available from<https://dhsprogram.com/data/available-datasets.cfm> (login and password separately provided)

**Project main steps**

1. Preprocessing of DHS data
   1. Decide what DHS Health data to study (e.g., malaria positivity rates, both using blood test results and using rapid test results; child mortality rate; HIV rate; women’s access to health; immunity against preventive disease) and collate raw data from DHS website using login/passwd provided
   2. Compute spatial aggregates of the above indices, at the various spatial (administrative) units of analysis available (geographic boundaries provided), and normalized by population size
2. Preprocessing of CDR data
   1. Decide what CDR metrics to study (e.g., activity, network advantage, introversion, gravity residuals, graph metrics, etc.)
   2. Compute from given raw data, and for the various spatial (administrative) units of analysis available, again normalized by population size
3. Build models that use CDR metrics to estimate DHS variables.
   1. Set possible benchmarks (population density, spatial auto correlation)
   2. Use hierarchical stepwise regression models, at each levels of spatial granularity, and measure overall model fit and significance of each metric. [study assumptions behind regression models; be careful about possible need for metrics transformation and normalization, about multi-collinearly issues, an test for heteroscedasticity].