UCL GCRF-Project: User Manual

by
Jack Shipway

March 31, 2017

Contents

1 Introduction

A census is the procedure of systematically acquiring and recording information about members of a population[1], and is often used to infer a plethora of economic, social and health related indices about that population. It is infeasible to collect data from all population members, thus 'representative' (statistically meaningful) sampling takes place, in the form of Demographic and Health Surveys (DHS). In lesser developed countries this occurs infrequently, is expensive, and and is likely to be biased due to political objectives. War, famine, and terrorism also restrict data collection in parts or all of various countries.

Call Detail Records (CDRs) have the potential to address the problems faced by DHS collection. CDRs concern data recorded by telephone exchanges in the form of calls or texts transmitted through that particular exchange (or between exchanges). They are often aggregated and anonymised such that an individual's privacy is retained. Justification for the potential of CDRs is threefold: data coverage is maximal (everyone has a phone), it is continuous (power-failure aside) and up-to-date, and also plays such an important role in daily life (connection to others, information transmission, calling a doctor, the police, friends), that it is likely to be related in some way.

The implication is that the socio-economic status of an area can be estimated by CDRs. However, the extent to which this this is true, and the spatial granularity at which indicators can be estimated reliably, can differ tremendously. Examining the ways in which mobile phone activity fluctuates spatially and temporally offers insight into the structure of societies, and can lead to deriving a set of metrics correlated with socio-economic indicators.

Success will ultimately be defined if a low-cost, reliable, and spatially-fine alternative to DHS data can be achieved. This information could be used to inform decision-making by governments or aid organisations, to more efficiently allocate limited resources such as aid packages, development funding and government policy. An aid package could be sent to those in most acute need, a development scheme could target specific villages or towns rather than entire regions, and policy could be implemented to benefit those for whom it is intended to benefit. It could also be used to estimate the consequences of natural disasters, terrorist incidents or otherwise, and thus play a critical role in disaster planning.

Section 2 outlines existing work in the field, paying particular attention to estimating socio-economic indicators. This research applies a similar approach, but focuses specifically on health-indicators including malaria, HIV and child mortality rates in the Ivory Coast

[1] 'The Modern Census: Evolution, Examples and Evaluation', by Baffour, Bernard; King, Thomas; Valente, Paolo (2013)

and Senegal. Various CDR and DHS-derived indicators are discussed in section 3 and their relationships examined in Section 5. Section 4 discusses the importance of spatial-granularity, and details exactly how point-data is used to infer information about a region. Raw CDR and DHS data come at different spatial levels, and are collected over different periods of time. As such, a way in which to reliable aggregate those data, and compare them accurately needs to be determined precisely.

After identifying promising relationships, we use feature selection to build a hierarchical, stepwise linear regression model. We examine the extent to which our model is generalisable, thus determining whether our model can be used as a good estimator for health-indices.

Dynamic networks are discussed in the closing chapter. This work is more experimental and relatively open-ended. The thought behind it is that CDRs naturally define dynamic networks. They change through time; interactions between certain places may change from day to day or month to month. It seems too simplistic to disregard how the network changes through time, and so we discuss how to capture these temporal fluctuations.

2 Existing Work

I summarise two papers here that laid the foundations for this extentional work.

1. 'Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks', by Chris Smith-Clarke, Afra Mashhadi and Licia Capra.

2. 'Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates', by Chris Smith-Clarke, Afra Mashhadi and Licia Capra.

Paper 1 proposes the development and validation of a methodology that governments of developing countries can use to accurately estimate socioeconomic deprivation at a fine level of spatio-temporal granularity, and low-cost. Its chief objective is to analyse patterns of mobile phone users' collective behaviour, aggregated to as fine a level of granularity as possible, and define features that can be used as proxy indicators of poverty.

Another objective (maintained throughout all associated work) is upholding privacy of individual users. Any data used (particularly regarding call detail records) should be unobtrusively obtained from local network operators, whose penetration in developing countries is enough to be representative nationwide.

CDR data penetrates the entire region (rich and poor have access to mobile phones), although it has been highlighted that the more ubiquitous mobile phones become, the less

As mobile phones become more popular, it is hard to tell which areas are actually rich or poor.

underlying signal strength there will be, as there will be less to distinguish the wealthy and the impoverished. A secondary aim therefore is to strengthen the underlying signal, and get rid of as much noise as possible from the surrounding data. Sparsification techniques are being used to do this, though this is for later down the track. Finally, results should be interpretable to governments/NGOs→ timely and targeted interventions to combat poverty.

The first paper demonstrates such a correlation - using a Pearson's Moment Correlation Coefficient (PMCC) test with poverty level and activity of regions. The more active a region is in terms of mobile phone activity, the more wealthy it is. (At least, a 0.777 PMCC was achieved). This first attempt however, was at the level of 14 subnational regions in Ivory Coast, and 11 in Senegal. We really want to dig much, much further down to as fine a level as possible. We have the data for 1230 cell towers in the Ivory Coast, and 1666 in Senegal, suggesting that we could estimate features of the areas served by each of these cell towers. This is not feasible.

The only way to validate the model is against ground-truth data, which comes at a much coarser level of granularity. DHS data is essentially census data covering thousands of different variables, and intends on being representative of the wider population of the country. The country is split into 14 subnational regions, which are then further split into 341 (total) regions. These regions are sampled from, and a certain quantity of households (usually 10-20) are selected from each. Ah! You must think that we could theoretically work at the household level. This is not quite as easy as it seems. Firstly, the anonymity of individuals and thus, the households in which they reside, is of utmost importance.

This resulted in the DHS creators distorting the data slightly in terms of latitude/-longitude of households. In urban settings, households are displaced anywhere within a 2km radius of their actual location. In rural areas, houses are displaced anywhere up to a 5km radius, whilst 10% of them are displaced up to 10km away. This paper therefore, aggregates houshold data to the cluster level by taking various medians/indices to take into account the relative poverty of each region and how it might be distorted based on population etc. Features computed include total activity, introversion, network advantage and the gravity residual model, which are all explained in later sections. The idea however, is that these features are the most heavily correlated with the DHS data.

My first contribution has been to try and reproduce Chris' work, and achieve the same correlation values to the poverty level. The second objective is to perform the same analysis as related to health indices, rather than wealth/poverty. These health indices include HIV, Malaria, and infant mortality. Malaria is an interesting case as it tends to be less affected by external factors than the other two - i.e. poorer areas tend to have high levels of HIV, and

infant mortality due to a lack of education, contraception and medication. Mosquitoes however, are indiscriminate to wealth when they attack. Of course, certain regions are likely to prevent against mosquito bites more...

One of my contributions (Research Stream 1) takes this proposal and adapts it to health-related indices such as HIV and Malaria rates, rather than poverty level. It works under the same premise, although I extract my own set of indices from DHS data to run correlation tests with. The idea here therefore is to compute a value along the lines of 'propensity to catch malaria' in a given region, and then perform some kind of correlation test with CDRs. Perhaps it is the case that the higher call volume per person reuslts in a lower chance of contracting malaria, but this is what I want to find out.

Wealth and poverty level has My next contribution is temporal analytics. CDRs can naturally be interpreted as a network/mathematical graph, and so looking at the connections as they evolve through time may offer more insight into how the network is structured, and that perhaps certain cell towers are more important in the diffusion of information - that in turn, may be related to how well information about health is circulated. I would hypothesise that the areas with greater temporal centrality for instance, are more likely to be surrounded by healthier regions, as it is easier to spread this information.

I began by splitting each of the bi-weekly data sets into individual, easily-accessible files containing all data covering the finest level (hourly). This allows quick and effective temporal analyy

In terms of scale, the Ivory Coast data set covers around 280 million 'interactions', i.e. cell towers speaking to each other. Senegal is more like 1 billion, and so the data we are working with is of considerable size. I certainly recommend using some kind of harddrive when manipulating the data. Using UNIX commands in terminal is hepful as many of the files are unopenable. I.e. it is difficult to check visually that you are extracting the correct data!

My contributions so far have been to extract the data into a usable, temporal format from the enormous data files they come in. This allows me to quickly examine different hours of data, and aggregate it effectively, without needing to manipulate the massive data sets. My next contribution was to see how activity varied through time. I then categorised this into working hours versus non-working hours, or different phases of the day, and urban rural divide. Not much luck was achieved.

I am currently working on extracting health related indices for each administration level 4 clusters, and then comparing this to the aggregated call volumes per area. I should probably check whether the total activity of each region corresponds to what Chris managed to do, and the poverty levels for that matter.. In the same way that Chris tried to estimate

wealth and poverty of a region, I will try and build a model for health, and then look at rolling correlations for a more temporal analysis. The only problem I foresee with a rolling correlation data is that DHS data is static in the sense that it is collected 'over a year'. Collectively, it can only be measured as a single point, and does not measure a change over time. For this I would need multiple DHS data recordings, which as explained, are hard to come by in lesser developed countries.

3 Data

In this section I explain the type, raw format and size of each data set necessary to reproduce this work. Despite considerable overlap, data for both countries investigated differ in all aforementioned aspects, and also possess some unique attributes.

3.1 Call Detail Records (CDRs)

In this paper, CDRs specifically comprise total volumes of mobile phone calls between cell towers, aggregated to the nearest hour. Other features that can be recorded are duration of calls, number and length of SMS messages, though we do not analyse those. Network operators take many precautions to defend user privacy. Occasionally, individuals or research institutions can record their own data, although this is subject to legal challenges and tends not to be more invasive. Certain mobile operators (in our case, Orange, and Sonatel), have decided to release fragments of data as part of a development challenge.

3.1.1 Data for Development Challenge

- http://www.d4d.orange.com/en/presentation/endowment-and-panel/Folder/The-D4D-Challenge-is-a-great-success (submitted projects)

Four data sets were released as part of a Data for Development Challenge[2]in both the Ivory Coast (2012), and Senegal (2014). Call detail records in this context concern aggregated volume and duration of . In the case of Senegal (but not the Ivory Coast), we have the volume and length of SMS messages sent between receiver stations. [REF] have examined the relationship between the ratio of SMS messages sent to, with education[3]. For our purposes, we will not be examining this data since we do not have comparable data for the Ivory Coast.

[2]http://www.d4d.orange.com/en/Accueil
[3]research paper

3.1.2 Data Pre-processing

The data was collected for 150 days, from December 1, 2011 until April 28, 2012. The original set of CDRs contains 2.5 billion calls and SMS exchanges between around five million users. CDRs have the following standard format: timestamp, caller id, callee id, call duration, antenna code. The customer identifiers were anonymized by Orange Ivory Coast and all subsequent data processing was completed by Orange Labs in Paris. In order to have a homogeneous data sample, customers that subscribed or resigned from Orange during the observation period have been removed. Additionally, incoming and outgoing calls have been paired in order to eliminate double counts (i.e. an incoming call for an individual is an outgoing call for the correspondent). The provided datasets contain the geo-locations of cell phone antennas. Orange considers the exact antenna location as sensitive information and therefore the locations have been slightly blurred so as to protect Orange's commercial interests. For technical reasons, the antenna identifiers are not always available. Instead of removing the corresponding communications, the code -1 was given to antenna with missing identifier. This happens for a significant number of calls (about one in four). The datasets covers a total of 3600 hours. Due to technical reasons data is sometimes missing in the datasets; missing data covers a total period of about 100 hours.

3.1.3 Processed Data

- SET 1: For this dataset, the number of calls as well as the duration of calls between any pair of antennas have been aggregated to the nearest hour. Calls spanning multiple time slots are considered to be in the time slot they started in. Antennas are uniquely identified by a CT ID, and a geo-location. Data is available for the entire observation period. Communication between Orange customers and customers of other providers have been removed.

- SET 2: Individual movement trajectories can be approximated from the geographic location of the cell phone antennas during calls. Limited knowledge of an individual's trajectory is often sufficient for identification and the individual can then be traced during the entire observation period. Two obvious solutions to reduce the possibility of identifi- cation are to reduce the spatial resolution or to publish trajectories only for limited periods of time. Since long term observation data as well as trajectories with a high spatial resolution have interesting scientific applications, two different datasets are published in order to balance privacy protection and scientific interest. This dataset contains high resolution trajectories of 50,000 randomly sampled individuals over two-week periods. The original data has been split into consecutive two-week

periods. In each time period, 50, 000 of the customers are randomly selected and are assigned anonymized identifiers. To protect privacy new random identifiers are chosen in every time period. Time stamps are rounded to the minute.

- SET 3: In this dataset, the trajectories of 500,000 randomly selected individuals is provided for the entire observation period but with reduced spatial resolution (at a lower administrative level). The published dataset also contains the geographic center of the subprefectures.

- SET 4: This dataset allows the analysis of communication graphs. The dataset contains the communication subgraphs for 5,000 randomly selected individuals. For these individuals, communications within their second order neighborhood have been divided into periods of two weeks spanning the entire observation period. For constructing an ego-centered graph, one consider first and second order neighbors of the ego and communications between all individuals (we do however not include communications between second order neighbors). The anonymized identifiers assigned to the individuals are identical for all time slots but are unique for each subgraph. That is, a customer who is part of the communication graph of two different customers has a different identifier in the two graphs. We therefore have a total of 5,000 connected graphs in every time period. The egos have been given identifiers between 1 and 10,000 and neighbor labelling starts from 20,000. Phone calls that follow a public phone usage pattern have been excluded from the randomly selected individuals. In Ivory Coast, it is common for some mobile phone owners to provide their phone to people on the street for a fee. This usage is characterized by a large number of outgoing calls but little mobility. We have removed from our selection of egos the customers identified as public phone providers.

This research only considers SET 1. SET 2 and SET 3 are interesting as they could offer insight into the diffusion patterns of information spread, however that requires a considerable amount of additional research. [REF] have already looked into tracking malaria for example, by assigning a high/low risk classification, and determining the risk-propensity of a region by understanding the extent to which people travel from low risk to high risk areas, and vice versa.

We discard SET 4 as there is no discernible way to associate randomly selected, individual mobile phone activity with a geo-location (perhaps raising privacy concerns). Analysis of this network could reveal insights into the structure of social networks, however it is not possible to align users with socio-economic data, so it is set aside.

3.1.4 Data for Analysis: SET 1

In this section, I present a graphical representation of SET 1. Figure 3.1 illustrates the geographic locations of CTs within the Ivory Coast and Senegal respectively.

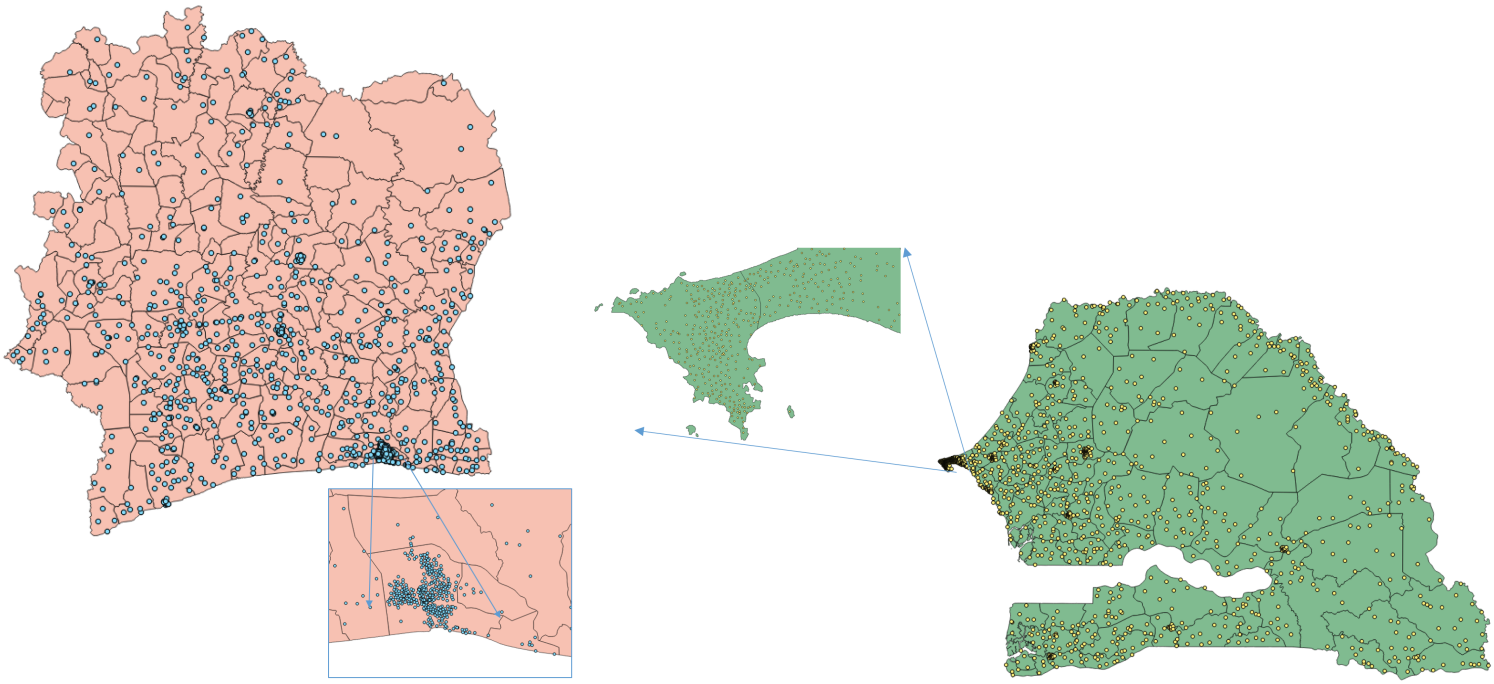


Figure (3.1): Geographic locations of CTs in Ivory Coast (L), and Senegal (R).

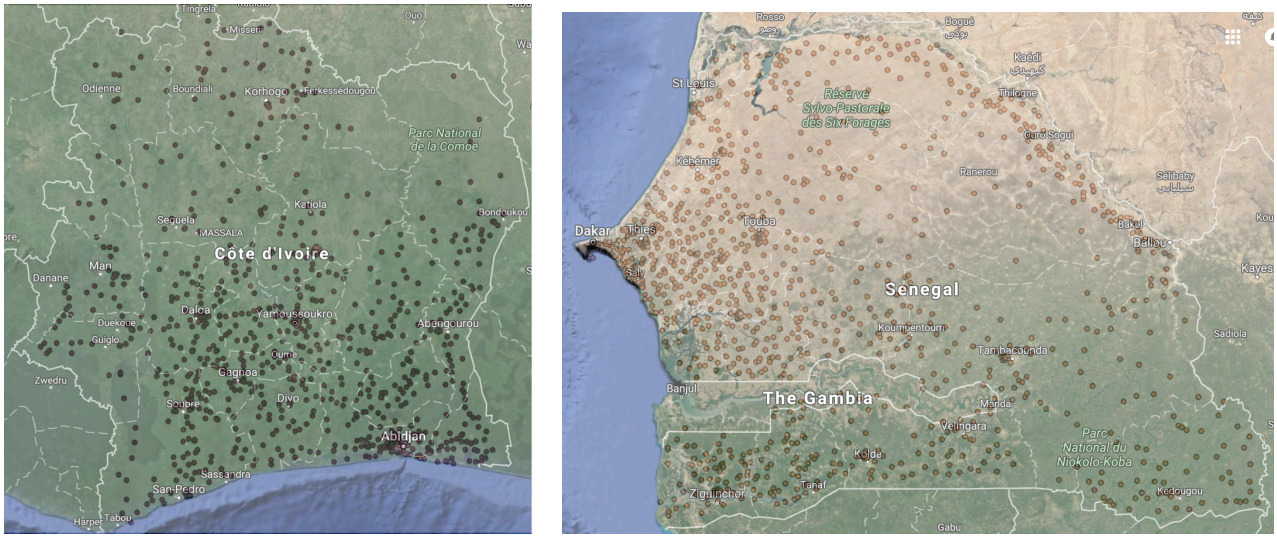


Figure (3.2): Geographic locations of CTs in Ivory Coast (L), and Senegal (R).

As shown, CTs appear clustered in certain areas (unsurprisingly, these are capital cities and other densely populated areas), and appear sparsely populated in more rural, mountaineous areas, as demonstrated by Figure 3.2. This is obvious in the Ivory Coast where national parks dominate the landscape in the north-east and the south-west. Senegal also has national parks in the central regions, but it is interesting to see that despite its mountaineous northern regions, there is a noticeable clustering of CTs near the border with Mauritania in the north.

We must not forget that Orange is also not the only network provider in the Ivory Coast or Senegal, and so these maps are not representative of the entire collection of CTs within each country. They do however, represent the largest portion of market share in terms of number of subscribers, and we have no reason to believe that certain locations are 'off-limits' to certain network providers due to legal restraints; at least, no reason to believe that these effects are important enough to distort results. We also have no reason to believe that there is some fundamental difference between people who opt to connect with the providers mentioned, and other network providers.

Ivory Coast

Raw data is provided in ten TSV formatted files, each containing two weeks worth of CDRs (~25 million rows), covering twenty weeks spanning the 5th December 2011 at midnight,

Datetime	Source	Target	Volume	Duration
05 12 11 09	1	2	1029	10841
05 12 11 09	2	1	281	1778

Table (1): A mock example of the format in which raw data is provided. Datetime is the nearest hour to which the aggregated number of calls (Volume) is recorded between pairs of CTs (Source, Target). Duration records the sum of the duration of all calls originating within that time-frame, in seconds.

to the 22nd April 2012 at 11pm. Data is aggregated to the nearest hour. If for example, there are 1000 calls routed through cell towers 1 and 2 at 9am, 05-12-11, this means that these calls occurred on the given date, between the hours of 8.30am and 9.30am but were aggregated to 9am.

A sample of data might look like table 1, where we see that at 9am on the 5th December 2011, 1029 calls originating at CT 1 and terminating at CT 2 were made. The sum of these calls lasted 10841 seconds. Also at that time, 281 calls were recorded between CTs 2 and 1, lasting 1778 seconds (highlighting the asymmetry of the network).

Each CT is associated with IDs from 1 to 1238. For data analysis purposes, it is easier to maintain a contiguous sequence of IDs, from 0 to 1239 to avoid indexing issues. It is straightforward to systematically remove unwanted data from a cleaned and analysed data set; it is not beforehand.

Some IDs are listed as having a geo-location but do not appear in the datasets provided, and some appear but do not seem to be monitored or even switched on (indicated by activity never deviating from zero). IDs that do not correspond to actively recording CTs include {573, 749, 1061, 1200, 1205, 1208, 1213}. Another exception is 1238 for which a geo-location is provided but activity remains at zero, suggesting that the CT exists but is not switched on or in use.

There are a few nodes for which no activity is seen throughout the duration of the monitored time frame. There are also some that seem to be 'switched off' until the final month, at which point, ~150 CTs are switched on. This is a particularly poignant point since it would be easy to assume an artificially low daily average activity rate (when measured over the entire data collection period), when in fact .

We can only reliably compare nodes that exist over the whole time frame analysed, and so we choose not to include these CTs in our analysis.

Another exception is *duplicated*CTs. IDs {233, 645, 737, 740, 743, 822, 900, 998, 1234, 1235} are given the same geo-locations as ten other CTs. There are either multiple cell

towers at each geo-location, or a recording mistake has occurred. To test this, I compared each pair of nodes to see if the sum of their total activities is equal. I found this to be the case, and so with (almost) absolute certainty we can discard them as duplicates. We say *almost* as there is a possibility that two CTs share incoming calls equally between them. The probability of all ten doing this precisely is extremely low, though we will not know for sure without visiting.

The final exception encountered is that of data for which it is unknown where a call originated from or terminated at (often both), but it is known that a call occurred. In this case, a '-1' is used as a substitute for the cell tower ID. For example, [05 12 11 09, -1, -1, 10] would indicate that there are ten calls between unknown cell towers at 9am on the 5th December 2011.

Senegal

This dataset is less 'exceptional' than that of the Ivory Coast. CT IDs are labelled from 1 to 1668, and are all active. There are no missing IDs nor duplicated CTs. There is also no need to remove calls between unknown CTs as none are provided. As far as raw data cleaning is concerned, there is no need to remove any CTs at this point. As we did for the Ivory Coast, we maintain a dummy ID for CT 0. This does not interact with other cell towers, but it is more efficient to perform analysis while regarding ths dummy variable as non-interactive than without.

The data spans one year from the 1st January 2012 to the 1st January 2013, arriving in ten monthly TSV files. This is the reasoning for country-specific functions in `process_raw.py`. Each 'month' for the Ivory Coast covers precisely 28 days, whereas each data set for Senegal covers the corresponding number of days per month (either 28, 30 or 31 - 2013 is not a leap year!).

3.1.5 Missing Data

One thing to consider is that the data are not contiguous[4]. Time is monotonically increasing (it never reverses), and cell tower interactions are ordered (activity between cell towers 1 and 1 comes before 1 and 2), however there are a number of 'missing' hours of data; in one case for the Ivory Coast, an entire day.

This could either be due to recording failures, power cuts, or system downtime. In terms of processing raw data, I had to rewrite the process_raw.py function to account for this. Initially, each line of data was read line by line until it was detected that the new

[4]I found this out the hard way

Feature	Ivory Coast	Senegal
Number of Cell Towers	1238	1668
Number of Calls (millions)	471	
Observation Period (Hours)	3600	8760
Population (millions)	20	
Area Covered (hectares (millions))	33	

Table (2): A feature summary of SET 1, for the Ivory Coast and Senegal.

hour had begun. After realising the data are not contiguous, I had to note the number of 'missed' hours, and store an empty file, at that particular time. I soon adopted a new approach - using pandas to group (exploiting efficient Map Reduce functionality) the data by hour, and subsequently labelling files and inserting missing data.

3.1.6 Geo-location accuracy

As explained, ccompanying the CDR datasets is a list of Latitude/Longitude coordinates of each CT. These are stored in bts_adm_1234.csv such that they can be easily loaded into geographic imaging software (GIS) or other mapping platforms[5]. The 'adm_1234' part of the filename refers to the administrative regions (Adm's) in which each CT resides (similar to counties in the United Kingdom), also provided in the csv file. This is explained further in Section 4. The accuracy of the provided coordinates is questionable, perhaps due to privacy concerns, and the extent to which displacement occurs is unknown to the authors.

3.1.7 Raw Data Analysis

Table 2 summarises some features of the CTs within the Ivory Coast and Senegal. In this section, we analyse the spatial and temporal distribution of these features, to gain a high-level understanding of the data and its dispersion.

Figure 8.1 visualises the distribution of total cell tower activity over both countries. In both cases, a small number of highly active towers are evident, versus a large quantity of less active towers. It is not possible to account for load-balancing and other call re-routing operations, and so we assume[6] that this is representative of the calling habits of subscribers

[5]We use QGIS as it is free to download, Mac-compatible, and easy to use

[6]Further research is required to establish the extent to which this can influence results

/Users/JackShipway/Desktop/civ_dist.jpg in the vicinity.	/Users/JackShipway/Desktop/sen_dist.jpg

Figure (3.3): Distribution of total cell tower activity. The grey plot is simply a reordering of IDs based on activity.

3.1.8 CDR Metrics

The first analytical process of this project is to derive metrics from CDR data that might be correlated with socio-economic indicators. Previous work [REF, REF] identified the following metrics as having a strong relationship with wealth and poverty intensity. Using the same approach, the objective of this paper is to establish similar relationships with indicators of *health*(Section 3.2).

Total volume is used as a metric in its own right, but we also consider more nuanced metrics, capturing different aspects of the network structure. All can be derived from the *adjacency matrix* defining by the network. The `cdr_metrics.py` script takes this adjacency matrix computed in `adj_matrix.py` as a parameter, and computes the metrics above.

- Total activity of a node is the sum of all incoming and outgoing calls.
  - Proportion of incoming versus outgoing activity = . This metric is interesting as it considers whether a region is
- Network Advantage
  - Median Degree.
  - Normalised Entropy
- Introversion
- Gravity Residuals
- Graph Metrics
  - Degree, closeness, betweenness and eigenvector centrality
  - Pagerank
- Temporal Metrics
  - Activity volatility
  - Temporal graph metrics

3.2 Demographic Health Surveys (DHS)

DHS data is usually collected annually and is incredibly comprehensive - thousands of aspects of individuals' lives and living situations are recorded. For the Ivory Coast, additional surveys were performed for HIV testing, though this is not the case for Senegal.

It is an expensive exploit, and certainly infeasible to examine everyone. Representative sampling is thus essential. To do this, experimenters initially

The DHS Program go to considerable lengths to enforce privacy measures. For example, geo-locations are distorted to preserve anonymity. The finest level at which geographic data is recorded is that of the sampled enumeration 'clusters'. A single centroid is calculated using the mean of all household geo-locations within that cluster. Then, the resultant centroid value is displaced within a 5km radius in urban areas, and up to 10km in rural areas. However, we are assured that the administrative region in which they reside will not change (I.e. a household is never displaced from one administrative region to another).

To access DHS data, you must either create an account at dhsprogram.com, or use the following details to login. You must then confirm the written description as the purpose of the download.

- Login: chris.smith@ucl.ac.uk
- Password: gcrf-project.

Once logged in, navigate to the 'Download Data' tab, and select the most recent census in the most suitable format. We use the SAV files in SPSS though SAS and Stata files are available. Again, geo-locations of the DHS clusters are provided in a separate file (in the list of possible downloads). Once the data has been succesfully imported into SPSS, one should compute a weight vector equal to the sample weight divided by 1000000 (1 million) - DHS does not deal with decimal data in collection. This can be coded in a 'New Syntax' window, by typing COMPUTE WGT = <Sample Weight Tag> / 1000000. After running, then type WEIGHT BY WGT, and the weight will automatically apply to the dataset. Sample weight tags are usually identified as V005 or HV005, though check as this is different for each dataset. This must be done for each dataset individually (i.e. household sample weight, HIV sample weight, and so on). The weights are then applied to samples as a way to align values with UN estimates[7]

For the Ivory Coast, the data is collected at a level of 341 subregions - i.e. the country is stratified into 341 different regions, and within those regions, certain villages or collections of people are sampled from at random.

[7]More information available at:

3.2.1 DHS Metrics

It is our proposal that CDRs are are in some way correlated with socio-economic data (in our case, health indicators) and can therefore be used as a proxy for estimating them. Data is examined in two stages: firstly, I searched through various metrics recorded and identified their header tags, and then removed all of the unnecessary data. In the end, I would be left with a data set containing each respondent's household ID, cluster ID, and then relevant metrics such as age, HIV testing status, accessibility to water and so on.

There are six metrics that we seek to capture, listed next. These can be derived from five different data sets available to download from the DHS program - `household.sav`, `Individual.sav`, `aids − hiv.sav`, `children.sav`, and `birth.sav` - although HIV data is not available for Senegal. An explanation into which data sets are used to derive each metric can be found in Appendix B.

1. Malaria: positivity rate as a result of examining blood test results.
2. Malaria: positivity rate as a result of examining rapid test results.
3. Child Mortality Rate: proportion of children under the age of five who died, relative to the total number of children born over then observation period.
4. HIV rate: positivity rate - number of positive cases found out of the total number of people surveyed.
5. Women's Access to Health - an aggregated score of the difficulties women face when attempting to get treatment.
6. Immunity Against Preventative Disease - an aggregated rating of how well equipped households are to deal with preventative disease.

Diseases can broadly be split into two types; we first define each type of disease, and then discuss the potential importance of these differences when analysing network structure. There are also two interesting metrics that can be used to examine the - incidence and prevalence rate. Both , but incidence rate captures, offering greater insight into how regions are changing through time, whilst incidence rate fails to tell us whether

Definition 1. Communicable Diseases

Diseases that spread from person to person or animal to person. The transfer can be airborne, through contact with contaminated surfaces, or direct contact with blood, feces, or other bodily fluids. Rabies, HIV, malaria, influenza, and athlete's foot are some examples.

Definition 2. Non-communicable Diseases.

Non-communicable diseases are medical conditions that cannot be passed from person to person. Heart disease, diabetes, cancer and asthma do not 'spread' in the same way that communicable diseases do. Other examples that are particularly relevant include sleeping sickness, some tropical diseases and child mortality.

Definition 3. Incidence Rate

The number of new cases..

Definition 4. Prevalence Rate

refers to the number of new cases

Malaria Rate

Malaria is a communicable disease, . It is for instance does not discriminate . However, you will find that one of the poorest ethnicities in "' are well-protected against Malaria due to their natural tendency to have sickle cell anaemis. There are two methods to test for the presence of Malaria.

1. Blood Test: The Gold-standard: . False positive/negative rates sit at, allowing us to . It is also capable of identifying the particular strain of Malaria if present.

2. Rapid Test: A cheap yet effective alternative to the blood test. Particularly useful when . It cannot distinguish between strains (and therefore severity). The false positive/negative rates however, are higher than the blood test, coming in at , and respectively.

HIV Rate:

This is a simple dataset - `HIV.sav` contains the cluster ID, household ID, sample weight, and test result for every person who agreed to be tested for HIV. It does *not* contain the results of every member in the household. It might for instance, provide three data points of people within the same house, but there might be seven people in the house. This is a known problem that the DHS organisation understand, and warn that it can introduces bias into our model. The very essence of 'asking' if people want to be tested may be biased - those who are afraid of the results may not agree, and those who know they are already positive (or negative), may choose not to. Also children may not want to, despite knowing that it may be passed down via parents or via blood (i.e. non-sexually). On the other hand, if only two people are sampled from a house of five, and those two tested positive,

there is reason to believe the other three might test positive, however we will never know if the positivity rate for the household is 2:5 or 5:5, or somewhere between.

The most valuable information is of course whether their HIV test returned positive or negative. This is recorded as a 0 or a 1 for negative and positive respectively. One must take care to account for 6, 7, and 8s in the recorded data, meaning that the test was inconclusive, something or something, One cannot simply 'sum' by column headers, explaining the complexity in `dhs_metrics.py`. I exported this data to a csv and extracted value counts using a dictionary-based structure.

We are only able to compute the prevalence rate as we do not have access

Unfortunately, this data is not available for Senegal, and so we will not be able to cross-compare results. It is however, an interesting avenue to pursue since preliminary analysis indicates that it is strongly correlated with CDR metrics.

3.2.1.1 Child Mortality

Disease Prevention

Covering a range of preventable diseases, this metric aims to capture the immunity or vulnerability of a household to particular diseases.

Female Health Access

This final metric is more exploratory in nature, but it is hypothesised that some measure of 'difficulty' of women attaining healthcare, either for themselves or for their children, is correlated with CDR data. Information diffusion . [REF] for example, demonstrated a strong correlation between those households with a radio, and their ability to access healthcare. We would like to examine similar relationships with mobile data.

3.2.2 Assumptions, Limitations and Improvements

DHS data is collected between late 2012 and late 2013 for the Ivory Coast and over 2010 for Senegal, though I am essentially treating it as a single point in time, aggregating over the entire observation period. CDR data covers 2012/13. Population covers 2010 and 2014, between which the actual population of the Ivory Coast grows by around two million, and a smaller but significant amount in Senegal. The actual populations in 2012/13 should lie somewhere in between. However, it is difficult to take this into account since different areas grow in population at different rates. I therefore chose to examine both to see if there is any notable difference. In an ideal world, our datasets would overlap, or at least be available at the same spatio-temporal granularity.

An interesting area for future research could start with determining 'the optimal time-lag between comparing data sets?'. So far, we have compared data sets that , however it might be the case that CDR data is correlated with DHS data in three years time. It is also

Secondly, in an ideal world it would be sensible to direct all calls through the nearest cell towers assuming unlimited capacity of the towers; this simply does not reflect real life. It is also reasonable to believe that cell towers differ in reach, capacity and maintenance requirements. We know that CTs occasionally re-route calls through alternative CTs (load-balancing), though it is impossible to tell which towers do this and when. This paper workds under the assumption that the effect of this is negligible, and should already be accounted for during aggregation - cell towers that need to re-route calls are likely to exist in areas of high capacity (urban areas) - calls will not be re-routed via isolated, rural cell towers.

We know for a fact (even in developed countries) that certain areas are not covered by cell tower reach, and that there will be people. It would be more sensible to normalise metrics by the number of subscribers that we know enter network covered areas on a day to day basis. It would be nice to combat this using some kind of probabilistic model that takes into account the distance, strength and current stress of current cell towers, and also the movement of people in and out of network coverage; at this moment, that information is simply not available.

4 Spatial Granularity

To reliably correlate CDRs with DHS data, we need to define a level of spatial granularity at which metrics can reliably be compared. We employ administrative regions (Adm's, also referred to as subprefectures, arrondissements, or simply, regions) for this purpose - essentially being political boundaries of a sort, analogous to counties within the United Kingdom. We hypothesise that these boundaries provide some sort of physical and socio-economic split between regions.

Shapefiles are vector-based graphics to be read by GIS software and in our case, define the boundaries of each administrative region. As they are geo-tagged, it is possible to project a vector layer into the same coordinate reference system as the CTs and DHS Clusters, and thus locate them within Adm's. To follow this procedure we need access to shapefiles, of which there are four levels. Level 4 is the most fine-grained, but still only defines the boundaries of 141 regions in the IC, and 200 in Senegal. A summary of how many regions are defined by each shapefile layer is given in Table 3. It also provides the

Administrative Level	Ivory Coast	Senegal
1	14	14
2	33	45
3	111	225
4	191	431

Table (3): The number of administrative regions defined by each level of spatial-granularity for which we have geo-locations, for both the Ivory Coast and Senegal.

number of areas with which we have geo-locations, and can thus identify them spatially. Unfortunately, geo-data is missing for a large portion of DHS clusters, reducing the number of data points available for comparison.

* **An important tradeoff:** As we aggregate to coarser levels of granularity, more data points are included in the aggregation step, improving the ability to detect outliers and/or reduce the effect of them. However, this reduces the number of samples available for comparison. One objective of this research is to achieve results at a fine level of spatial granularity, and so we would ideally like to remain at as fine a level of granularity as is feasible.

4.0.1 Proportional Representation

CTs are not to our knowledge, limited to serving people within a particular administrative boundary. Their 'reach' is defined by their strength, load capacity, and ... I need to find a way to compute how much of the activity of a single cell tower, belongs to each of the administrative regions that it may serve. Again, aligning with previous work, I assume that cell towers are homogeneous in that they are similarly powerful. I am not taking into account their actual 'reach', load-balancing, and physical or geographical factors affecting the transmission of calls, which absolutely introduces bias into our model. For example, towers in the city may be more powerful in that they may be able to handle more calls than those in rural areas, however they may not need to have such a far reach, since there are many other cell towers in the region. If one cell tower is particularly overloaded, it may transfer some of its load to a neighbouring cell tower and spread the load (load-balancing). Again, it is unknown to what extent this happens (it is assumed that it does). Finally, it is assumed (rather overtly), that a call is picked up by the nearest tower. I.e, someone calling next to cell tower 1, will have his call diverted through cell tower 1. It is highly unlikely that a call will be transmitted via a cell tower at the other end of the country, however it is unknown exactly how calls get allocated. Overall, I think that this affect will

be negligible since we are comparing regions - in which there may be many cell towers that share load. They are likely to be evenly spread within that region, and so the uncertainty kind of cancels out..

but this is not possible as the data are collected at different spatio-temporal granularities. By this, we mean that different quantities of samples are collected at different areal units, at different points in time. To recap, CDR data is recorded from ~1220 uniquely located cell towers for the Ivory Coast between December 2011 and April 2012, and ~1650 for Senegal between January 2012 and January 2013. Refer to <INSERT FIGURE> for their geographic positioning. DHS data is considerably coarser; collected at ~191 cluster points for the Ivory Coast, and ~400 for Senegal, although again, many of those points are unusable, resulting in ~140 u, and ~200 for Senegal. Section <INSERT SECTION> details how these points were originally selected and data collected.

For now, we omit the temporal aspect of our data by aggregating over all time-periods. DHS data is irregularly collected over a given year, and there simply aren't enough samples to reliably look at rates of change. There exist many ways in which to aggregate spatial data; six models are proposed next, each offering a unique perspective.

Theoretically, we could operate at the level of the Voronoi[8] regions of each cell tower. However, our aim is to estimate some feature and compare that to a ground-truth data to see how accurate our estimations were. DHS data is used as ground-truth data which, as explained in Section 3, is not collected at the level of CT voronoi regions. It would be possible to employ various interpolation techniques to estimate at this level, but we opt against using it as there are too many underlying factors (physical and social variance across regions) distorting estimations. This is a well-known problem in spatial analysis; we address this in Section 7.

4.1 Aggregation

With a cleaner data set and having understood the distributions of CDR and DHS metrics graphically, we would now like to cross-compare derived metrics. We have established levels to which we would like to aggregate data, but the ways in which data can be aggregated range in power and complexity. The following lists six different methods, all offering a unique perspective of the data,

We aggregated the CDR data by ... . We might consider computing the DHS data for the same Voronoi regions, however, as section .. details, and so we cannot be sure. We can however, be sure that the data point will never cross administrative boundaries - locations

[8] All points closer to that cell tower than any other cell tower

are displaced within these boundaries. Therefore, it seems appropriate to aggregate to the administrative regions in which each DHS cluster point lies. This poses further problems, as the finest level of granularity at which, is that of 'administrative level 4', representing 140 subnational regions in the Ivory Coast, and 'administrative level 3' for Senegal. I therefore aggregate both CDR and DHS metrics to these administrative levels, so that eventually, I will have a CDR metric per adm_4, and DHS metric per adm_4.

I must also take into account the population of , and so each CDR feature will need to be normalised by the population of their administrative region.

To aggregate the CDR data, I had to assess the controbution of each cell tower to a particular location. This was made difficult by the fact that many cell towers overlap administrative boundaries. A proportional amount of activity must be aggregated. Following the methodology outlined in 7.3.3, resulted in 4 files containing CDR metrics aggregated to the administrative region, normalised by the population of each region.

In QGIS, an intersection (built-in algorithm) of the voronoi and administrative regions were performed. This formed a , from which it is possible to determine the proportion that each cell tower contributed to each administrative region. VoronoiPop.py extracts this data, and along with total_activity.py, and activity of each cell tower is partitioned according to the proportion of it belonging to each administrative region. I will highlight here how this.

Now that our dependent and independent variables are on the same scale and have been normalised appropriately, correlation analysis can begin.

This is complicated by the fact that voronoi regions overlap administrative regions to different degrees. I need to compute the proportion of each feature derived, that corresponds to a given administrative region.

I have 140 administrative regions here, and 1238 Voronoi regions. These Voronoi regions may reside entirely within an administrative region (there are a lot in Abidjan, the capital), or they may overlap. In the case of these overlaps, I need to compute the proportion of this feature attributed to each region, and thus the populations within each of these regions. To do this, I compute the intersection of the Voronoi cells and the shapefile, and then cross-reference this with the raster file, giving me the populations of each of the mini regions. To reiterate, I am assuming that the mobile activity of a region is proportionally split amongst CTs in its vicinity.

Model 1

The simplest model, CTs and DHS clusters are assigned to the administrative level in which they reside geographically. A simply average (mean and median) is then taken of

all data points within each ADM. As granularity increases, we notice certain regions as missing CTs or DHS clusters, or both! This is not surprising, but means that we must exclude those data points from our analysis. Another point to notice is that as granularity increases, fewer points are used in calculating the average. This can be a problem because a regional rate may be determined by a single anomalous point, whilst numerous valid data points lie just outside the border. Model 1 also does not take into account the range of CTs. This is not a problem in urban areas as a higher proportion of the region has network coverage. However, in rural areas, we do not account for the proportion of population with network coverage. We also do not account for individual movement - people who travel (work or otherwise), are . It is difficult to capture such a metric without analysing movement patterns. For instance, a CT might be located in a sparsely populated region, but if a large number of people travel there for work, the actual number of people using the network is higher than this metric would suggest.

In terms of DHS data, we must consider the number of people surveyed per region before computing our metrics. If for example, we have two clusters A and B with rates 0.5 and 0.9 respectively, a naive average of these two clusters is 0.7 ((0.5+0.9)/2). However, this does not take into account the number of people surveyed. We would be more confident in a rate of 0.5 if it came from 500 positive samples in 1000 tests, than 9 samples in 10 tests. To combat this, we sum the individual samples beforehand and then compute the rate, as if the samples were collected at the same location. So to combine areas with 500 positive samples from 1000 tests, and 9 positive samples from 10 tests, we would compute a rate as 509/1010, yielding an aggregated rate of 0.504. This is considerably lower than the rate of 0.7 calculated earlier, reflecting the additional confidence we have in testing in area A.

Model 2

A drawback of Model 1 is that data points can be heavily influenced by outliers, particularly in large regions of low activity. In this model, we aggregate data to the centroid of each cluster, using inverse-distance weights. Another problem is that in rural areas, whilst more points are taken into account, the centroid may by chance be located close to a highly or lowly active CT. This - suggesting that there is no particular reason for selecting the centroid of the cluster as the point at which to aggregate.

Model 3

This model uses a K-nearest neighbours approach. DHS clusters are left unchanged, whilst 'k' surrounding CTs are aggregated using an IDW method to the DHS cluster points. This means that we have, however a large portion of these are zero.

Model 4

Model 5

Again, slightly more experimental, but we define a grid of equal size squares, and overlay this onto our country. Within each cell, we perform a k-nearest neighbours distance weighted analysis within a certain distance, and compare these results - the benefit here is that we have essentially generated a consirable amount of data points which may prove more insightful correlations, however we will have to be carefeul only to include areas for which - mountaineous regions for example, simply do not contian, but hopefully this will be taken into account. Interpolation - this goes against our initial proposition, but is worth testing.

Model 6

Models 1 through 5 are tested at all four adminstrative levels separately. Model 6 is experimental in that it is tested at overlapping levels of granularity. The underlying hypothesis is that urban areas have a tendency to have greater populations, , resulting in obvious outliers (such as Figure 1.1.). If it were possible to zoom in on these urban areas, split them into smaller areas, we would hope to see less warping of results, and less of an influence on outliers from the urban versus rural divide.

What format is the data in. What are the different attributes? Timestamp, call volume, call duration (how highly are these correlated, expect them to be highly, maybe look at rolling correlation here??? YES). What are reasonable values for attributes to take, are they integers or.. ? Look for outliers, how to test for and deal with them? What about missing values, how to deal with missing values and how to smooth data if there is missing data. Should I do this or should I just leave it? Look at the distribution of each variable. What does the data actually represent? What does it mean. How is it distributed? How can I incorporate location inti the variables? MAybe I could do some kind of regression on location of the cell tower?

4.2 Normalisation

Our model would not be particularly useful if it were not possible to compare regions with different populations. Certain features (such as total call volume) are dependent on the number of people in a given region - activity in an area with a million people will be higher than an area with a few thousand. As such, a 'per person' metric (for example, calls per person) is necessary for regions with different populations to be reliably compared. To

isolate that signal, call volume is normalised by the population density of the region in question.

4.3 Spatial Autocorrelation

X's law, that regions closer together exhibit similar characteristics can also be taken into account. Our data are not randomly dispersed and our intuition concerning physical geography would lead us to expect clusterings of similar data points in cities for example. This is however, slightly more complex than it may seem. There are areas within cities that are different. Very poor areas located within wealthier areas, and also some cases of very wealthy areas amongst poorer areas. These are potentially some of the more important , but it is difficult for our model to recognise them as their values are almost aggregated out of any aggregation. They could in fact quite easily be considered outliers. In future work, we would like to achieve some way of boosting the signal strength of these spatial anomalies, but for now, progress only with a spatially lagged variable that accounts for a region's neighbours, when this information is available.

We incorporate a spatially lagged variables . This can be provided when a certain level of information is known neighbouring regions prior to analysis. We use Moran's index..

4.4 Miscellaneous Data

Provide links to google/usaid data sets potentially providing rainfall estimates etc.

4.4.1 Climate

Brief intro to what climate data could add. Malaria for instance, is

I have not analysed climate data in detail. I have however collected rainfall and temperature data covering the time period in question. I have examined the correlation between . In the future I would like to obtain climate data at a much more fine-grained level, and see whether or not mobile phone activity is in any way related to the climate. Eventually, it would be interesting to see whether there is some kind of response to natural disasters that occur in the form of activity spikes or otherwise. A fundamental objective of this research is to provide governments with evidence upon whicht aid and development regimes to regions in most need. If we can analyse areas at most need of aid after certain natural disasters based on mobile phone data, we could have a cheap and reliable source of information.

4.4.2 Raster

Raster files are tagged image files - often depicting a map of a country or region with some form of data attached, at a given resolution. I am interested in the spatial distribution of population - files of this kind can be found for both countries at www.worldpop.com. For the IC, there are two files covering 2010 and 2014. This highlights a problem - four years is a long time. Census data for instance, is available in 2005/06 and 2012/13 for the IC, and . Results where we need to normalise by population, will always be limited by this time lagged variable. The population of the Ivory Coast has increased rapidly since 1960 - from around 3.5m people, to over 22m in 2014 and is still climbing. Senegal began with a similar population in 1960, and increased at a slower rate to a population of m in 2014.

The spatial distribution of population growth is unclear - there may be certain regions accounting for a larger share of the increase than others.

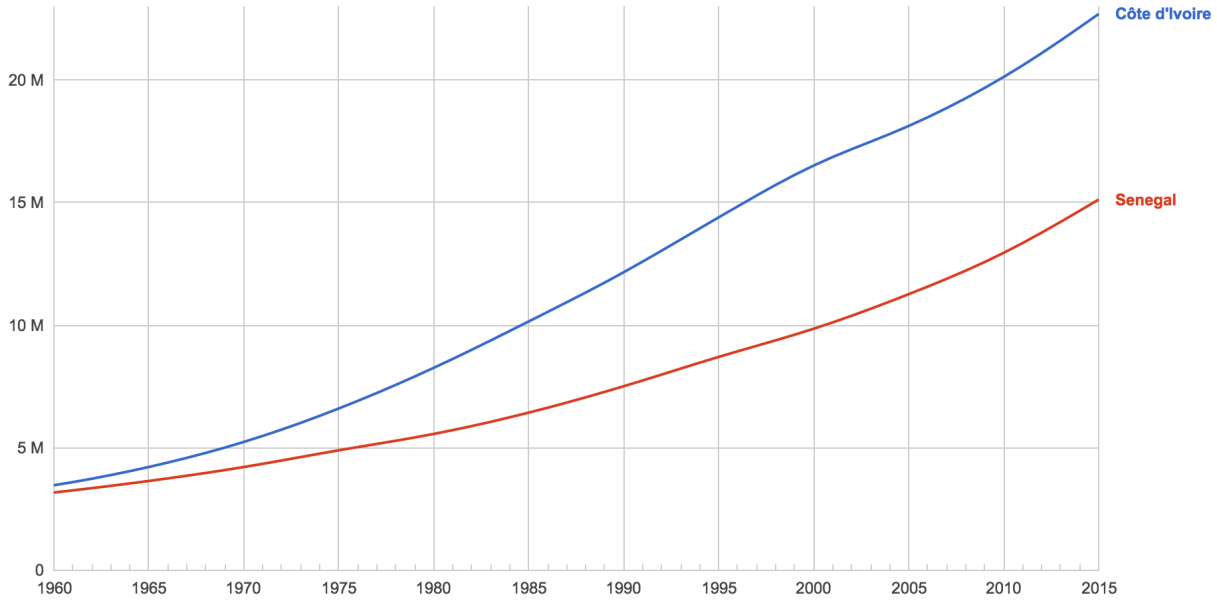


Figure (4.1): Population of Senegal and the Ivory Coast since 1960

The reason for which we need to do this is because our metrics are heavily dependent on the number of people per region - volume of phone calls is (almost certainly) going to be higher for a region with twice the population of another. There are exceptions of course (this might be true for an urban versus a rural area), and so we wish to normalise our data in a way that allows us to compare regions, independently of their population. As such,

we divide all of our metrics by the regional population, to essentially gain a 'per person' average output.

5 Methodology

In this section I outline our key objectives, the processes by which our hypotheses can be tested and confirmed or not.

5.1 Objective

We would like to obtain meaningful predictions of socio-economic indicators (in this paper, health indicators) at as fine a level of spatial-granularity as possible. We could view this as an optimisation problem - starting with a budget, a policy, or combination of both, we would like to find the strategy (or set of strategies), that minimises or maximises some objective. More formally, with a budget B, and strategies S in S, we would like to optimise . Some imaginary scenarious might include:

- Minimising the rate of malaria contraction by efficiently allocating a limited number of bednets or vaccinations amongst the areas that could benefit the greatest.
- Minimising HIV contraction rates or the spreading of disease by identifying villages or towns at greater risk, and directing policy change or medicine towards them
- Maximising the economic returns as a result of a policy change.

5.2 Hypothesis

The underlying hypothesis behind achieving such objectives is that metrics derived from CDR data are highly correlated with health indicators derived from DHS data, and thus can be used as predictors for them. By providing justification for those correlations, we believe it is feasible to construct a model capable of predicting health indicators at a fine level of spatial granularity, based on historical CDR data. The utility of our model comes from its ability to successfully predict the consequences of implementing a particular policy in a particular region. This can therefore be used as a tool for Governments to allocate resources efficiently - it does not guarantee that it will be used, however.

To do this, we would need to repeatedly compare two regions of similar constitution; one in which a given policy has been implemented, and one in which it has not. If, for example, our model can reasonably predict the outcome of introducing such a policy, we could analyse the combinations of strategies that lead to the maximal results, thus

allocating resources in a more efficient manner. Our model need not be precise, merely accurate and convincing enough to distinguish good from bad policy.

5.2.1 Deriving CDR Metrics

Figure 8.4 shows part of an adjacency matrix constructed from the CTs, describing the number of calls routed through each pair of CTs. The matrix is not symmetric, since a CT can be the source of fewer calls to another CT than it receives in return. It is also common for calls to originate and terminate at the same CT (imagine calling a neighbour).

From this matrix, constructed in adj_matrix.py, we derive , by passing it as a parameter in cdr_fundamentals.py. This constructs a pandas data frame containing each CT ID, the four administrative regions in which it resides (as 'Adm_1', 'Adm_2' etc) and the values of their respective metrics, storing the result as cdr_fundamentals.csv. A range of distributions can be seen,

5.2.2 Deriving DHS Metrics

Figure 8.2 shows the distribution of malaria prevalence, as obtained by. We immediately notice that there are regions in which there is a rate of 0 - i.e. no positive cases are found. This could be due to the region being particularly immune, or simply down to circumstance. This information is important however in our treatment of the distribution of positivity rates.

To account for these differences, it is important to take into account sample size, giving us some measurement of confidence in the rate we obtain. If there are 1000 people in a region, 800 are tested, and 100 return positive, we can be more confident that we achieved a positivity rate of 1/8, than if 1 person out of 8 sampled, in a region of 1000 people. Trying to transmit this level of confidence has been posed - solved by..?

5.2.3 Other Metrics

A multitude of features are recorded in DHS data, some more useful than others. For example, as in [CHRIS WORK], each cluster ID is assigned two interesting categorical variables. Urban versus Rural, and Capital versus Non-Capital. This information is interesting as it offers something in common. It is also used to identify why certain outliers are so, in Section 5.4.2. It may for example, be more appropriate to model urban areas as a single entity, and rural areas as a different entity entirely. As we will show in Section 9, there is a considerable difference in total activity in urban and rural areas, and so we should expect there to be differences in other metrics derived from those data. For now

however, we consider all points without classification, as we would rather a single model that can distinguish between urban and rural areas successfully, than two different models.

5.3 Correlation Testing

Here we determine health indicators that are most acutely correlated with CDR metrics. For instance, it might be true that the lower the connectivity of CTs in a given region, the higher the HIV rate. It is entirely plausible that more highly connected areas benefit from greater information diffusion, and thus lower HIV rates as a result of greater access to preventative treatment, care, and information. We would like to try and demonstrate such relationships - in the subsequent sections I present side-by-side comparisons of metric correlations for both countries examined, following Model 1. As the most primitive model, it is not expected to achieve exceptional scores, particularly as we increase spatial granularity since sample scores will be increasingly influenced by anomalous results

5.3.1 Correlation Tables

I apply a Pearson's Moment Correlation Coefficient (PMCC) test to the CDR and DHS metrics, to identify the existence and strength of relationships between them. I do this for all CDR metrics versus all DHS metrics, across all administrative levels. A PMCC test measures the linear dependence between two variables, and the correlation coefficient ρ can take on value between [-1, 1], where a score of 1 represents a perfectly positive correlation, and -1 indicates a perfectly negative correlation. A score of 0 suggests no such correlation.

To calculate ρ, one computes the formula $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where $cov(X, Y)$ defines the covariance of A and B, and σ_X and σ_Y represent the standard deviations of A and B respectively. I present some key results here, with Figure 5.1 summarising the wide-ranging results in a correlation matrix. We also show how correlations tends to fade as spatial granularity increases. This is expected, but will prove to be a challenge in addressing the objective of estimating socio-economic indicators at a *fine* level.

Figure (5.1): Correlation matrices between CDR and DHS metrics, presenting the correlation coefficient, p-value, and confidence interval achieved through performing a PMCC test. Each sub-figure represents a given administrative level.

5.3.2 Scatter Plots

Scatter Plots

Figure (5.2): Adm_level 1....

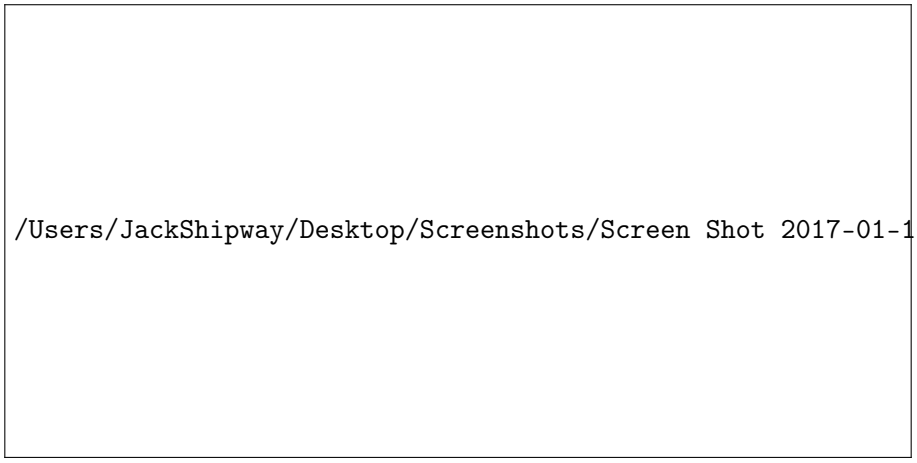


Figure (5.3): Adm_level 2....

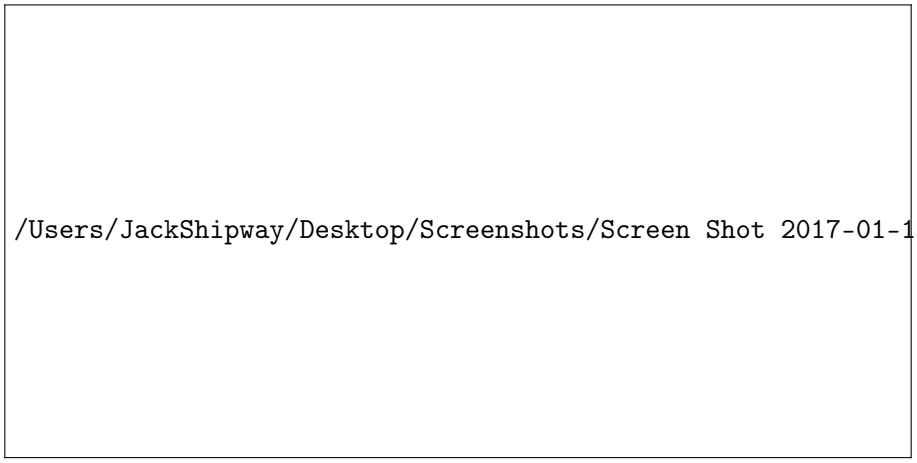


Figure (5.4): ....

5.3.3 Categorising Data

In the previous section, it was suggested that it might be necessary to treat certain samples differently from others. Here, we present similar plots, but highlight the difference when samples are grouped in a particular way. Groupings that we use include {urban, rural}, {capital, non-capital}, {upper 50% wealth, lower 50% wealth}, {upper 50% extreme poverty, lower 50% extreme poverty. Figure 5.5 shows the difference between each category, which if plotted on their own, may reveal more insightful correlations.

As we see in Figure .., grouping variables in this way reveals that there may be more complexity in the relationships between variables than anticipated.

5.4 Linear Models

Satisfied that there exists some relationship between CDR and DHS metrics, we now wish to test whether the assumptions upon which a linear relationship is well-founded, are satisfied by our data.

5.4.1 Outliers and Transformations

Figure .. is . It would however be naive to exclude this data point from our analysis as it may be a perfectly valid data point! In fact, upon further inspection, it can be revealed that these points represent the capital cities - Abidjan and .. for Senegal and Ivory Coast respectively. If we take this analysis. These points indicate that Model 1 is perhaps too simplistic to take into account the plethora of extraneous features of each region.

5.4.2 Multicollinearity

amongst independent variables must be tested for, as it can dramatically influence the coefficient estimates of a multiple regression. There are certain features that we can infer to be highly correlated, such as the total volume and total duration of calls. However, some are not so obvious. The correlation matrix in Figure <INSERT> gives a PMCC score, p-value and confidence interval for the independent variables. In general, IVs with a PMCC score above 0.7 are not used in the same model. Another test is that of the condition number when fitting a linear model. If the condition number is sufficiently high,

5.4.3 Heteroscedasticity

5.4.4 Statistical Tables

Comparing results over differnet countries, do we see the same variables selected? To an extent yes, but some manual intervention is required. The following table indicates the models/features that we will use in our regression analysis in the next section. They have the best potential based on our statistical analysis and so we wish to see how generalisable they are by performing a multivariate linear regression.

Spatial-Autocorrelation measures the level of association between neighbouring regions. I.e, is it true that areas that are physically closer to other areas, similar in the results that they give? Some ways to test this include Moran's I, Geary or Gamma index.

Spatial Lag is an independent variable that is computed . Essentially it means that an area's response variable is highly dependent on neighbouring regions, and can be used as a feature. This is however, only possible when some prior information is known about a region, either through census data or otherwise.

5.5 Model Selection for Multivariate Linear Regression

Satisfied with IVs that have been scrutinised, transformed and , I would like to select individual features . This is done by forward selection, whereby variables are selected in a way that optimises the adjusted-R^2 value of the fit. This value describes the variance within the response variable that can be described by the IVs. I am now going to compare whether similar IVs are selected for particular response variables, across both countries.

5.6 Hierarchical Stepwise-Regression Model

With a model containing selected IVs, we would like to build a hierarchical stepwise regression model.

This is achieved through repeated selection of randomised data points, fitting a linear model, and computing the . We also test over varying train:test set proportions. Satisfied that our model , we would like to test its generalisability. The fundamental objective of this project is to be able to predict various socio-economic indicators of areas, using CDR metrics. It is also an objective to find out by how much the model is improved (if at all), when adding our CDR-derived features to existing (supposedly easy to obtain data such as

population density). If we can significantly improve the accuracy of a model, governments might wish to use our data to inform decision making.

Figure <INSERT> shows the results of 1000 iterations of a Linear Regression model, selecting varying proportions of randomised data points as training and test sets respectively.

Dep. Variable:	y	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	-0.003
Method:	Least Squares	**F-statistic:**	0.6787
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	0.412
Time:	22:44:17	**Log-Likelihood:**	-10.237
No. Observations:	100	**AIC:**	24.47
Df Residuals:	98	**BIC:**	29.69
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	0.4086	0.060	6.861	0.000	0.290 0.527
x1	0.0810	0.098	0.824	0.412	-0.114 0.276

Omnibus:	13.631	**Durbin-Watson:**	1.829
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	5.530
Skew:	0.318	**Prob(JB):**	0.0630
Kurtosis:	2.039	**Cond. No.**	4.75

Figure (5.5): Statistics...

6 Results Discussion

It would seem that there are relatively small improvements when CDR features are used in addition to baseline models. Spatial Lag plays an important role, indicating that neighbouring . I would now like to investigate more nuanced models whereby a more reliabel model can be established..

7 Conclusion

It is pleasing to see strong correlations between derived health indicators, particularly as the metrics used are so simplistic. They are easy to quantify, straightforward to derive from

real data, and provide a strong foundation upon which to develop more complex models.

Despite not achieving exceptional results, and . Overall, results indicate that the relationships between health indicators and call detail records do not exist purely by chance. However, to isolate and highlight these relationships requires a reduction in dataset noise. There are simply too many. An investigation into non-linear models would also be shrewd as it is quite clear through examples in Section 5 that linear relationships exists to a degree. However, as soon as those data

8 Improvements and Future Work

8.1 Alternative aggregation models

The first improvement I would seek to address is testing different models of aggregation. There is reason to believe (through experimentation) that distance weighted metrics, or taking a k-nearest neighbours approach could considerably improve the reliability of aggregated values. It also means that areas in which there are fewer data points will be assigned a value, although we should bear in mind just how confident we can be in that value. We have shown that fairly strong spatial autocorrelation exists within both datasets, indicating that areas near each other exhibit similar features than with those further away. There are many uncontrollable socio-economic reasons that could influence results. A range of new models have been suggested, and should not be difficult to implement given the first example.

8.2 Sparsification methods

It would be interesting to explore various ways of reducing dataset noise. The only difficulty would be highlighting and

8.3 Other metrics

Climate data is especially interesting - explain why, provide links to data.

A paper [2] on PubMed, suggested a way in which to understand DHS data a little more. For instance, the current rate of Malaria would be derived from the previous rate (which we have access to from previous census data), plus some additional variables. The previous rate is defined by a rate before that , plus something else. Being able to quantify this relationship might offer more insight into how malaria rates increase or decrease, and also the effectiveness on policy.

I have only considered one time frame, so it would be interesting to perform analysis on previous time frames and in exactly the same way, perform correlation testing. It might be the case that there is a stronger correlation when one of the variables is lagged by a certain amount. We don't expect the correlation to change in nature; rather the strength to improve or disprove until a point at which the lag is optimised. Using this optimised lag, we could be sure to use data for .

Analysing historical data in this way might also lead to better ways in which to appreciate the change in metrics, rather than simply their current value. This remains a difficult task since data is only available for particular years, and in countries where war or famine is rife, not available at all.

9 Temporal Analytics

Here, I am trying to derive new features that are related to the way in which cell tower activity evolves through time. I am currently aggregating the data in a way that does not take this temporal aspect into account, and certainly would not notice that certain cell towers 'turn on' and are active at different times of day, different days of the week, or months of the year. This is why it is so fundamental to extract the data into timesteps, rather than aggregating. This defines a network or mathematical graph, containing nodes (cell towers), that may exist at different points in time (i.e if they are active), and edges between them (calls passing through them), which also may exist at different points in time (i.e. cell tower 1 may call cell tower 2 at 9am, so there is an edge existing between these nodes at this time, but not at 12pm). Essentially, I am assuming that all interactions take place at 'the same time', which clearly masks any underlying signal. First results suggests that activity changes dramatically over the course of a year. It is natural to see that activity drops off at night, but identifying days of the week where there is more activity, offers insight into the underpinnings of the network that we otherwise would not identify.

A major step of this project would be to implement all of the data in some sort of Spark GraphX network that can quickly analyse certain things (I am learning to do this). Betweenness centrality for instance has a temporal equivalent, and . Fundamentally, once I've identified that there is indeed some kind of signal relating to wealth (already confirmed) and health indices, looking at how these correlations change over time is the next logical step.

9.1 Temporal Granularity

I wanted to test different levels of temporal granularity, to see how activity changes over the course of the day. Looking at the hourly level is too granular - there is no noticeable difference between this level. Essentially, the law of diminishing returns is at play here - it is not worth the extra effort (time) to analyse at the hourly level, when virtually the same signal is found at the 12-hour level. We found after trial and error, that comparing working hours and non-working hours is the most useful. I.e, looking at activity that occurs between the hours of 6am and 9pm, and 9pm to 6am gives the best split.

9.2 Volatility

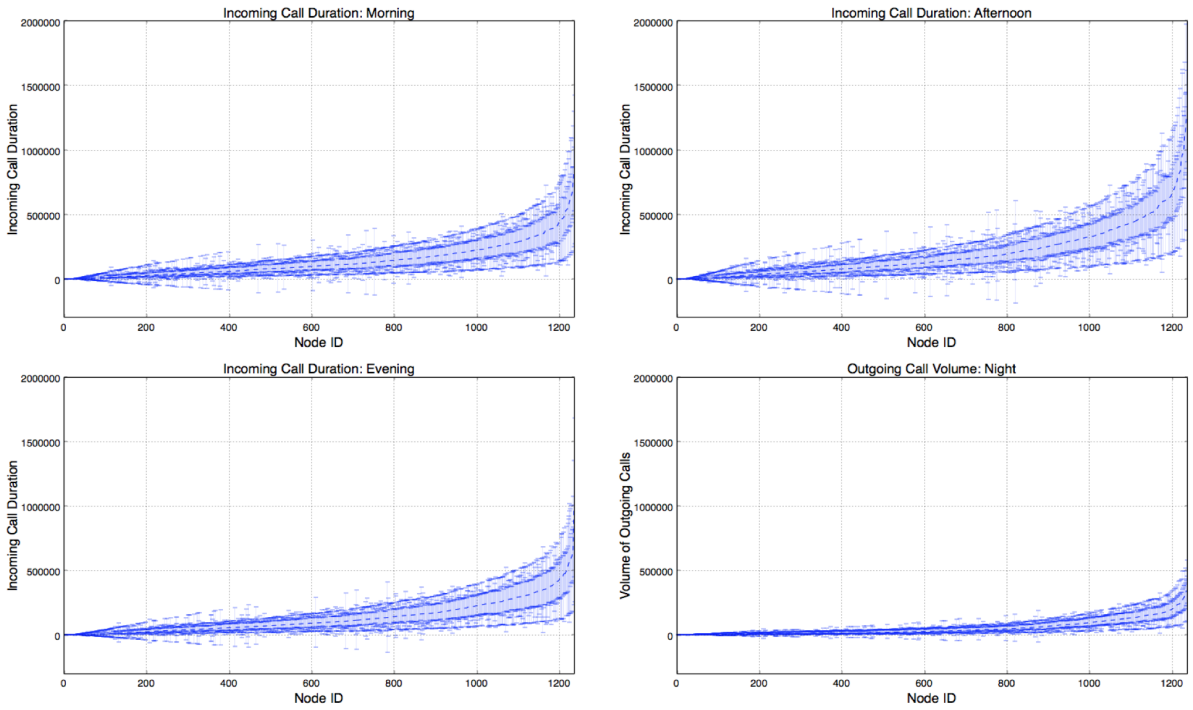


Figure (9.1): .

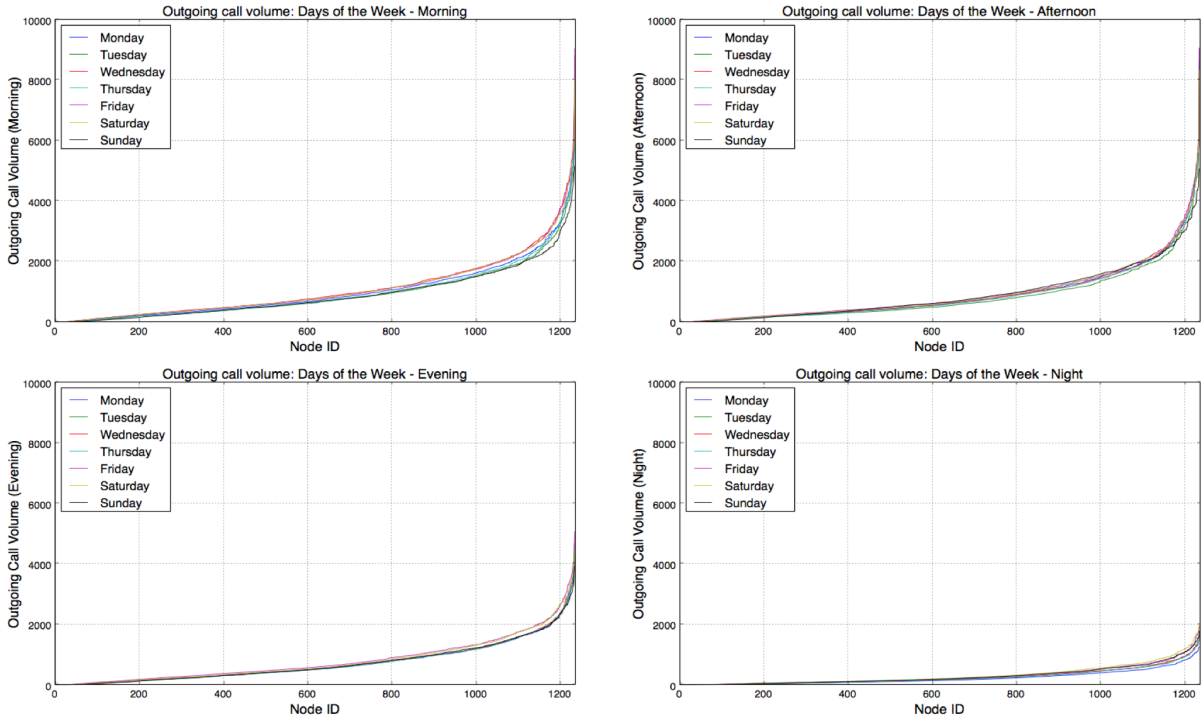

Outgoing call volume: Days of the Week - Morning
Outgoing Call Volume (Morning)
Node ID
Outgoing call volume: Days of the Week - Afternoon
Outgoing Call Volume (Afternoon)
Node ID
Outgoing call volume: Days of the Week - Evening
Outgoing Call Volume (Evening)
Node ID
Outgoing call volume: Days of the Week - Night
Outgoing Call Volume (Night)
Node ID
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
2000
4000
6000
8000
10000
0
200
400
600
800
1000
1200

Figure (9.2): .

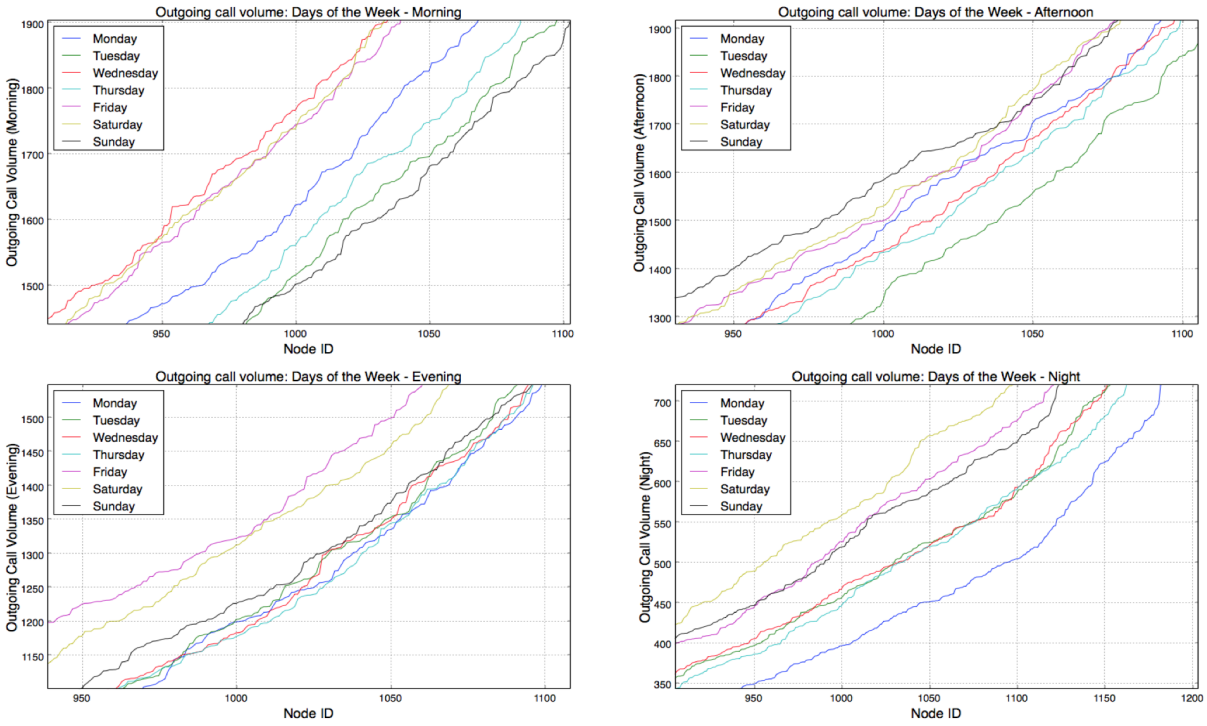

Outgoing call volume: Days of the Week - Morning
Outgoing Call Volume (Morning)
Node ID
Outgoing call volume: Days of the Week - Afternoon
Outgoing Call Volume (Afternoon)
Node ID
Outgoing call volume: Days of the Week - Evening
Outgoing Call Volume (Evening)
Node ID
Outgoing call volume: Days of the Week - Night
Outgoing Call Volume (Night)
Node ID
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

Figure (9.3): .

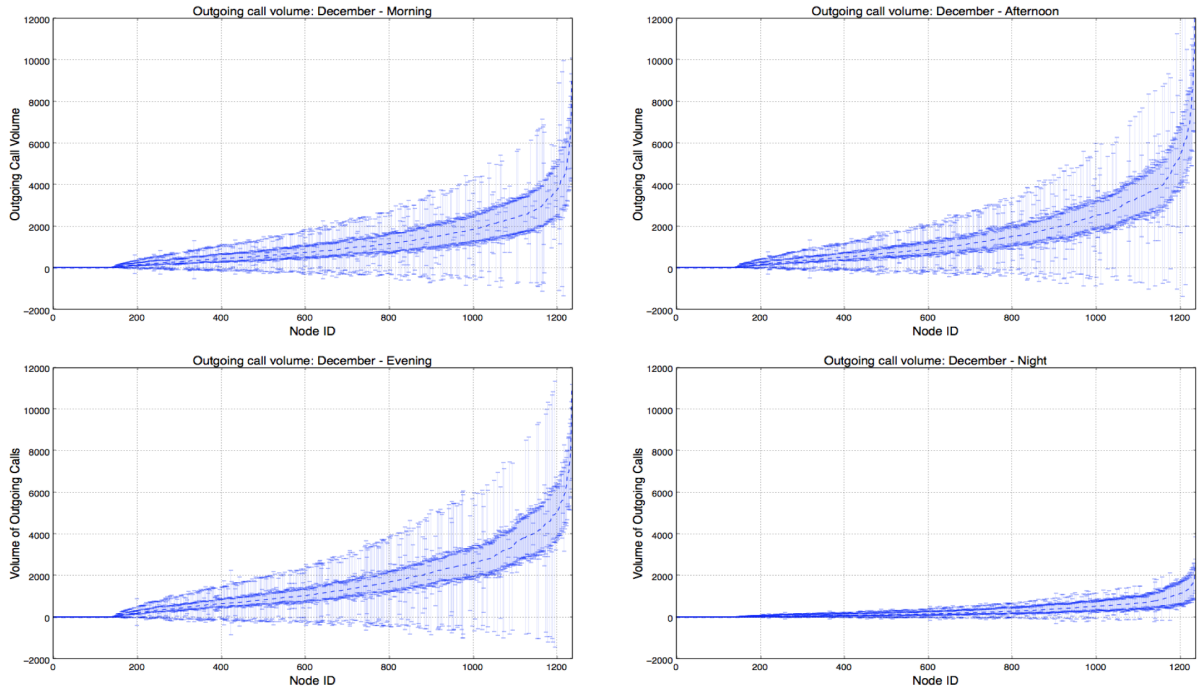


Figure (9.4): .

Discovering Inactive Cell Towers

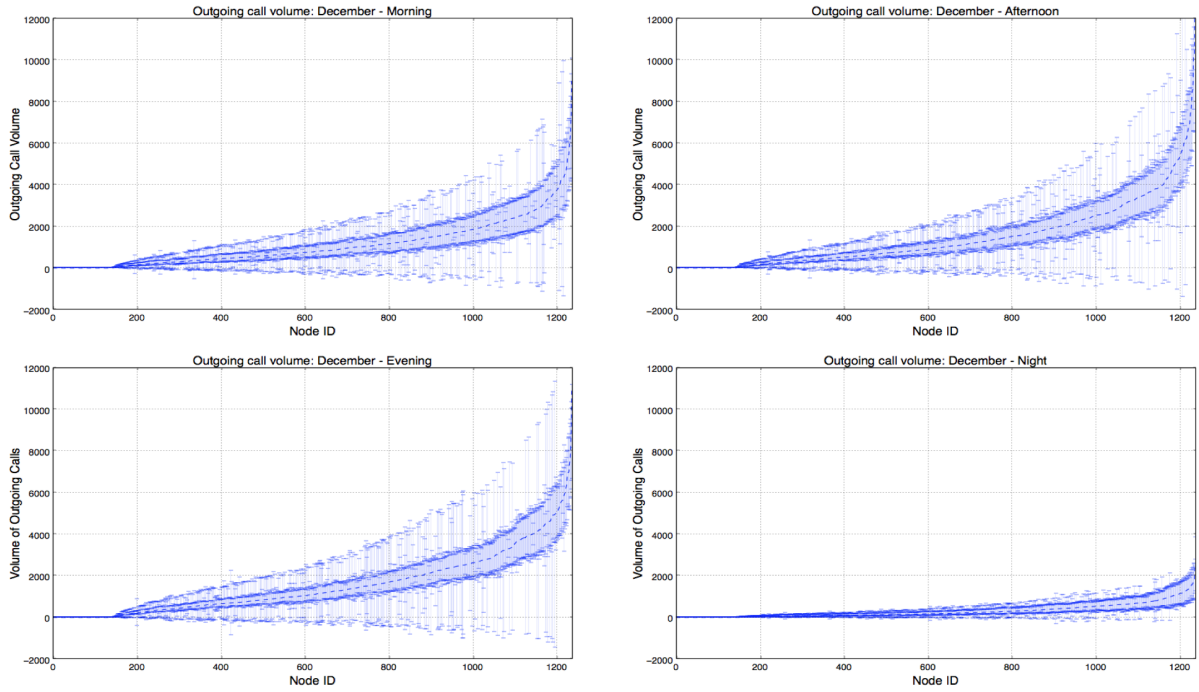


Figure (9.5): .

Working versus Non-working Hours

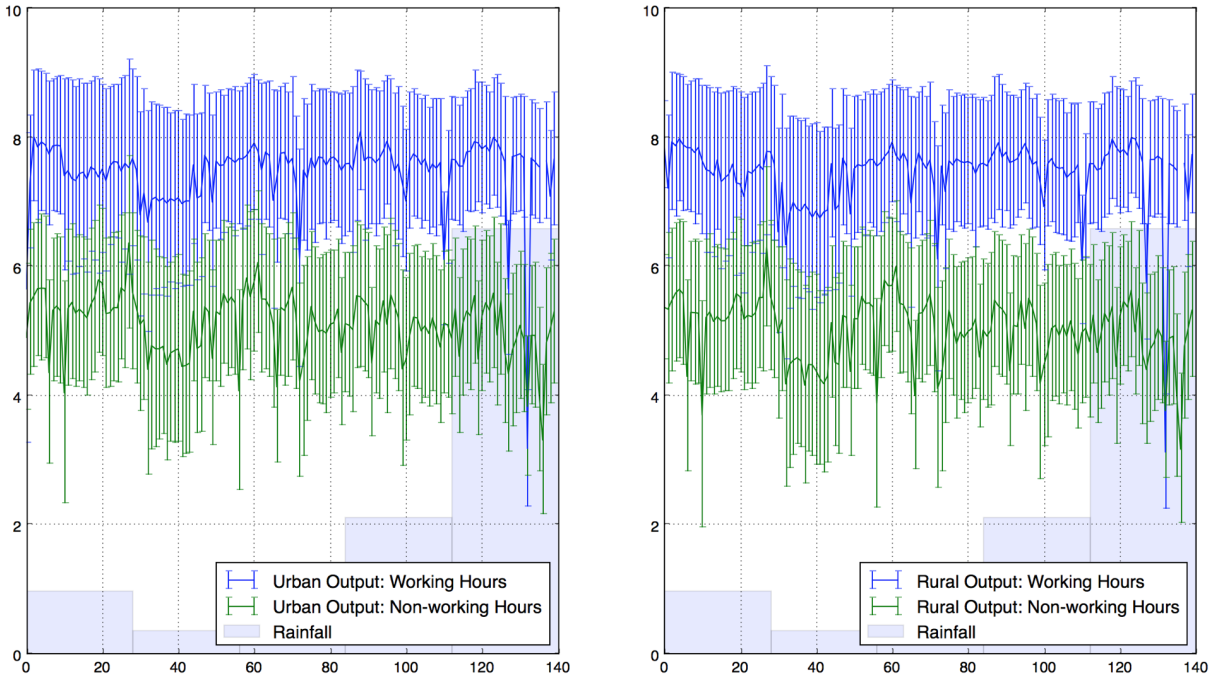


Figure (9.6): .

10 References

11 Appendices

Filesystem

In this section I outline how to structure your filesystem to access code and data - especially important if using an IDE (Pycharm). There is considerable overlap between both countries analysed, although there are some key differences that I will highlight.

Directory Structure

Retrieve the GitHub repo at www.github.com/ShipJ/GCRF and use that as your home directory - always be sure to only commit/push code changes; the data itself should not be publicly available.

With this document, I provide an empty file structure that should be added to the cloned GitHub repo, and populated with the necessary data. It is set up to ignore files in

the data directory and will therefore not commit/push them to the online repo by default. Files that are needed before data processing include:

1. SET1.tsv - for both the Ivory Coast and Senegal. These can be found on the remote server (see 'Downloading Data' for instructions). They should be placed in data/raw/<country tag>/cdr.

2. malaria.sav, child_mortality.sav - for both the Ivory Coast and Senegal. hiv.sav, to be placed in data/raw/<country tag>/dhs.

3. malaria.csv, hiv.csv, child_mortality.csv etc, should be obtained through using SAS or SPSS on the files downloaded from the DHS Program above. I have provided the exact files used, but you are free to select other variables if they are relevant and process those.

4. ct_adm_1234.csv - the geographic locations of cell towers, both in terms of their latitude/longitude, and their respective administrative regions. This was obtained through using QGIS and various point-in-polygon operations but requires some manual relocation of towers. I explain how to do this in the QGIS section below, but also provide the csv's that I used.

5. dhs_adm_1234.csv - the geographic locations of dhs clusters. Again, latitude/longitudes are provided as well as their administrative regions. The coordinates are found in a file provided by the DHS program, but I manually added the administrative regions into the same csv so that they correspond.

6. Shapefiles are for use within QGIS. They provide the geographic boundaries of administrative regions within countries . They were downloaded from http://www.diva-gis.org/gdata, but I also provide the same ones that I used.

7. Raster files are also for use within QGIS. They were downloaded from http://www.worldpop.org.uk/, but I provided the specific ones I used.

8. intersect_pop.csv was constructed after intersecting the administrative regions (shapefiles above), with the voronoi regions of cell towers (QGIS has an algorithm for this). This file is provided. It allows us to compute the proportion of cell tower activity occurring within each administrative region because certain cell towers evidently serve more than one region. It should be stored in data/processed/<country tag>/pop.

9. distance_adm1.csv, etc, give a distance matrix of the distance between administrative regions

10. All of the timestamped files - eg. 2011-12-05-00.csv - are created after running process_raw.py on the raw data. data/interim/<country tag>/cdr should therefore populate itself.

11. All of the adjacency matrices are created when running adj_matrix.py on the pre-created timestamped files above, and are stored in data/processed/<country tag>/cdr/adjacency.

12. master.csv data files are created when running

Files

Downloading data It is easy to use ssh and scp to transfer the files from the remote server - to view the files, with your login details, type 'ssh <username>@uova.cs.ucl.ac.uk' into terminal. You will be prompted for your password, then directed to a directory, type 'cd ..' to navigate back into the home folder, and the data can be found under datasets/d4d. To retrieve data, go back into a terminal screen (not in the remote repo!), and type scp username@uova.cs.ucl.ac.uk:/home/datasets/d4d/... <local directory path>. Warning: They are large files.

Data Files are too large to permanently store in memory - I suggest downloading them once, processing the data into smaller, timestamped files providing easy access and ability to quickly insert/sample from them. The time-stamped files can again be stored once for the purposes of feature extraction, and then stored elsewhere (or deleted).

Processing Instructions

* Ivory Coast: Download SET1.zip from the server (instructions in User Manual) and place into raw/civ/cdr. Proceed to run process_raw_cdr.py, using 'civ' as the keyword argument.

* Senegal: Similarly to above, grab the data set (SET1) from the server, but then manually expand it using manually in terminal or otherwise, and place the expanded files into raw/sen/cdr. For some reason there are problems when automating the expansion procedure that I cannot resolve currently. Use the keyword argument 'sen' to process the raw data. This data set also included quantities of SMS messages in separate files, but these are not included in our research.

QGIS

QGIS is a great piece of software for Mac, used for plotting spatial/geographic data. If you are on Windows there are a host of other different programs you can use such as ArcGIS etc - the good thing about QGIS is that it is free, and functionality-wise, there is negligible difference in terms of what a paid piece of software can do versus this one; at least for the reasons that we require. It is actually fairly straightforward to use and play with once you know what you're doing. It easily plots spatial data - provided you have a csv file containing longitude and latitude of various points (eg. cell towers), regions (shapefiles, we'll get to that). Just need to be a little bit careful with the type of coordinates used. I don't think it matters because we are only looking at. If however, you choose to plot some of these points, there is what's called the UTM coordinate transform, which is supposedly a better representation of where they would be on a real map. However, this just complicates things - QGIS does all of this for you.

Essentially, the steps that one would take in order to do any kind of analysis come in the form of adding layers to the project. These layers can be rearranged and moved about later, but first I'll talk through the different kinds of layers that I've used. Adding a 'Raster Layer' is one type - this is where you would search for your raster file of the country, and simply insert it. The result would be an outline of the country, and if you hover over different points in the grid (sometimes right click is necessary), it will give you the population at that particular point and any other associated features. If you look at the toolbar at the top, you can select the hand which will allow you to move the grid, scroll in and out etc. If you select the hand with the information mark, this will allow to select certain areas of the layer, or drag to select whole regions.

Next are shapefiles - these are boundaries that are drawn corresponding to subprefecture regions within a particular country. Again, I am relying on online sources for these and the finest level of spatial granularity that I can find are 191 subprefectures. This is problematic since DHS data is sampled from 350 'subprefectures', i.e. these 191 regions are split into a further 160 regions, meaning that some aggregation of areas will be necessary. Obviously, and again, this boils down to time constraints, if it were feasible to 'draw' these boundaries, it would add immensely to our model, meaning that we wouldn't have to aggregate as much, and the spatial granularity at which we operate is as low as possible (our main objective). Shapefiles are simply the lines drawn defining the regions, however I don't know where these regions are - if it were a case of copying a map, I could add in a bunch of points, and draw the boundary, but I have no information on this, and so cannot make a judgment.

The shapefiles go all the way up to country level, to subnational level, down to 191

regions (for Ivory Coast), and ... (for Senegal). They all cover the same area however, and so should overlap each other, and the raster file, exactly.

Finally comes data points - such as cell towers. I have extracted the longitude, latitude and ID number of each of the cell towers and put this in a csv file. You than want to add a 'Delimited text layer', fill out the relevant information, and plot. Again, the points should all be within the boundary regions - if they form a country shape, but don't overlap, you've probably got latitude and longitude round the wrong way (experience..).

These are all of the layers you might need. Now I'm going to talk about relating each of these layers, and combining them etc. You might wish to find out which region each of the cell towers belongs to. This comes under the category of 'add polygon attribute to point'. A 'polygon' in this case literally means an area/region. I can do all sorts of summaries like, if I have a list of activity per cell tower, and want to know the activity per region, I could sum the activity of all cell towers in the region - but how do I know which region each cell tower belongs to? - Add polygon attribute to point, and then group by region ID.

It makes everything easy, but my workflow tends to use QGIS to get any spatial attribute of the data, and for plotting results (spatially), and Python for manipulating the data sets in the first place.

Extraneous Results

Code

The Pandas library in Python has a fast and easy 'read_csv' function which allows me to manipulate this data (in the form of a dataframe) efficiently. Likewise, numpy's 'load' is similar, and for large matrices, performs operations very quickly.

There are Python packages that can read SAV files directly but upon testing a few different methods, none were found to be particularly user-friendly.

Population Per Region

Here I describe how to compute the population per voronoi region pertaining to each cell tower. I then compute the proportion of this population that belongs to each administrative region, allowing me to normalise the metrics I compute (such as number of calls, number of HIV cases), by population to achieve a 'per person' metric. I do this at the cell tower level rather than at the administrative level as the voronoi regions of many towers overlap the administrative boundaries, particularly as the granularity of the regions increase. The

difference is negligible at the coarser levels, but can influence finer levels. For instance, if one cell tower covering one million calls serves two regions with a 51:49 split, we would want to attribute 51% of those calls to region A, and 49% to region B. Not 100% to the larger regions. One million can have a dramatic influence on the results.

Population Per Intersect.csv: Contains the population of each cell tower voronoi region associated to each administrative region - to be used for normalising CDR data at each level of granularity.

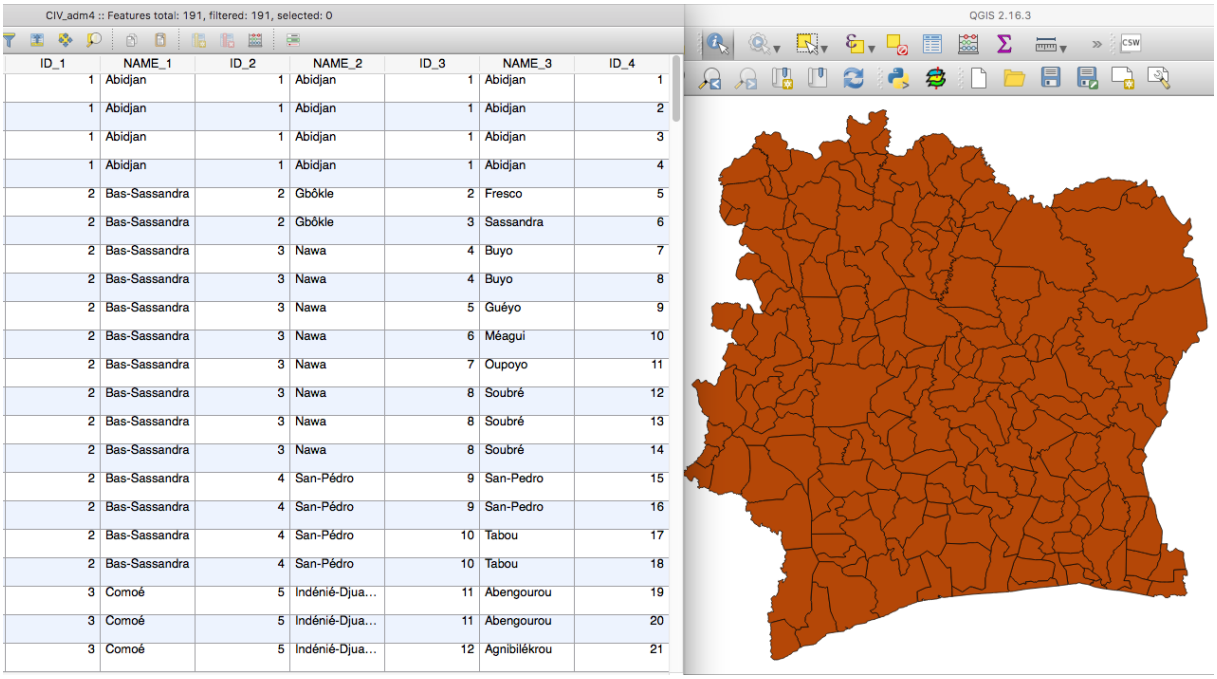
Population_Per_Adm_1234.csv: Contains the population of each administrative region - to be used for normalising DHS data. This is an estimation based on CHRIS NEEDS TO TELL YOU THIS, as there are no shapefiles at this level of granularity.

Data Required

- Raster files: 2010 and 2014 for the Ivory Coast,

The main idea is to intersect the voronoi regions of each cell tower with each shapefile layer. Then, intersect that layer with a raster file containing the population of the entire country to compute the population of each of the regions, providing us with the proportion of total population served by each cell tower, attributed to each administrative region. A small number of CTs are located in the sea near the coastline and their 'nearest' administrative region must be identified manually. To do this, a a 'Point in Polygon' operation is performed in QGIS. This will add the features of the region to each node within it (i.e. it's ID number). If this is not done, then they will be given a value of 0 as the ID of the administrative region in which they reside. This is not a real region... For the Ivory Coast, you need to look at points {55, 121, 125, 170, 688, 758, 760, 1051, 1104}, and change their administrative region (in the generated CSV file) ID to {19, 24, 24, 1, 24, 1, 27, 18, 78}.

Step 1: Open a new QGIS file. Firstly we want to add the shapefile containing the regional boundaries of each country. Layer -> Add Layer -> Vector Layer -> Browse -> RegionalBoundaries/Adm_4.shx. To simplify things, you only need to use layer 4 (rather than 1, 2 and 3), because it contains all of the other layers within it. I.e. it provides the level 1, 2 and 3 administrative regions that each level 4 region is in, and so on.



ID_1	NAME_1	ID_2	NAME_2	ID_3	NAME_3	ID_4
1	Abidjan	1	Abidjan	1	Abidjan	1
1	Abidjan	1	Abidjan	1	Abidjan	2
1	Abidjan	1	Abidjan	1	Abidjan	3
1	Abidjan	1	Abidjan	1	Abidjan	4
2	Bas-Sassandra	2	Gbôkle	2	Fresco	5
2	Bas-Sassandra	2	Gbôkle	3	Sassandra	6
2	Bas-Sassandra	3	Nawa	4	Buyo	7
2	Bas-Sassandra	3	Nawa	4	Buyo	8
2	Bas-Sassandra	3	Nawa	5	Guéyo	9
2	Bas-Sassandra	3	Nawa	6	Méagui	10
2	Bas-Sassandra	3	Nawa	7	Oupoyo	11
2	Bas-Sassandra	3	Nawa	8	Soubré	12
2	Bas-Sassandra	3	Nawa	8	Soubré	13
2	Bas-Sassandra	3	Nawa	8	Soubré	14
2	Bas-Sassandra	4	San-Pédro	9	San-Pedro	15
2	Bas-Sassandra	4	San-Pédro	9	San-Pedro	16
2	Bas-Sassandra	4	San-Pédro	10	Tabou	17
2	Bas-Sassandra	4	San-Pédro	10	Tabou	18
3	Comoé	5	Indénié-Djua...	11	Abengourou	19
3	Comoé	5	Indénié-Djua...	11	Abengourou	20
3	Comoé	5	Indénié-Djua...	12	Agnibilékrou	21

Figure (11.1): Step 1 - administrative level 4 shapefile with attribute table also encapsulating adm_1, adm_2 and adm_3.

Step 2: Now insert the locations (coordinates) of each cell tower. Layer -> Add Layer -> Delimited Text Layer -> Browse -> CellTower_Adm_1234.csv. Set the coordinate field to latitude and longitude, and insert. Each of the cell towers should now appear atop the shapefile. Notice that some of the points (five or six at the bottom of the image) are not within the regional boundaries; rather, they are lost at sea somewhere. I have taken the liberty of adding the nearest administrative region ID to each of them manually. Had this not been done, each cell tower would have been given an administrative region ID of 0, which screws up calculations. This has already been taken into account in CellTower_Adm_1234.

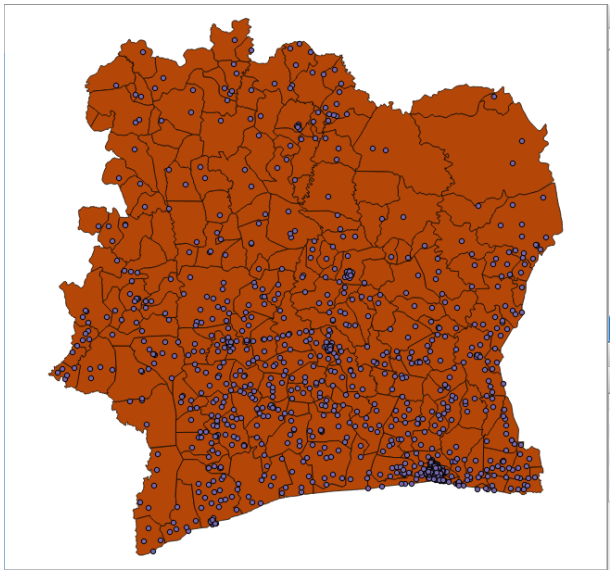


Figure (11.2): Step 1 - administrative level 4 shapefile with attribute table also encapsulating adm_1, adm_2 and adm_3.

Step 3: Now I want to compute the voronoi region of each cell tower. That is, all points (an area) closer to that cell tower than any other. QGIS does not provide a way to compute this using the national boundary as a breakpoint, and so the voronoi regions end as a square surrounding the region. Use Vector -> Geometry Tools -> VoronoiPolygon as the algorithm, and it will output the results for you.

Step 4: Notice however, that there are a few regions not covered by the voronoi areas - these need to be extended such that the entire raster file is covered (otherwise part of the population might be missed). To do this, right-click on the administrative shapefile and toggle editing mode to 'on'. Now select the 'Node Tool', allowing you to select regions on the map. Double tap will add a new node to any of the voronoi regions you choose,

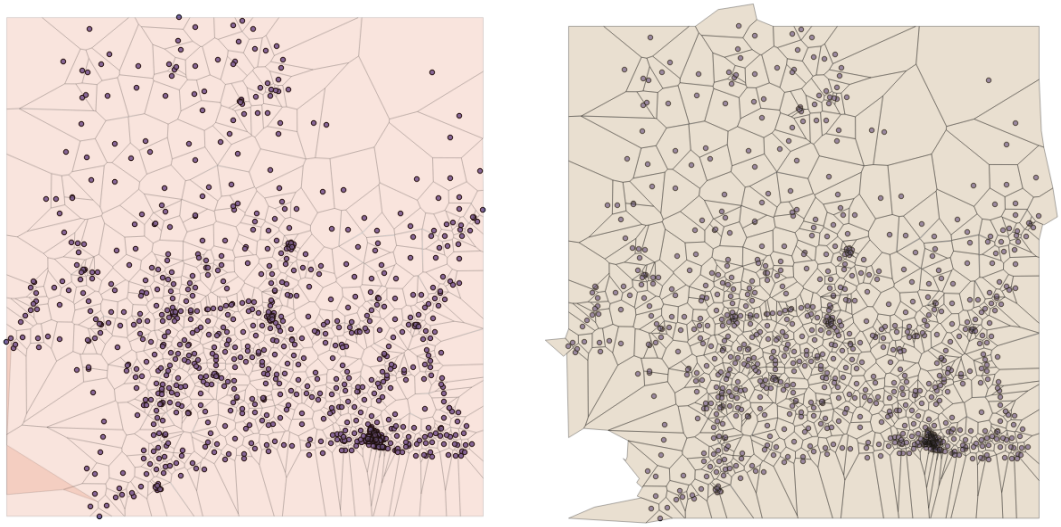


Figure (11.3): Steps 3 and 4 - extending the voronoi polygon areas manually.

Step 5: We now wish to intersect our newly constructed voronoi polygon file with the administrative level 4 shapefile (since this covers level 1, 2 and 3). To do this, select Vector -> Geoprocessing Tools -> Intersection -> Select correct files. A new file will be created that you can save as you like. There are 2377 rows, which makes sense since there should be at least 1238 regions (for each cell tower), plus around 1000 corresponding to overlapping regions.

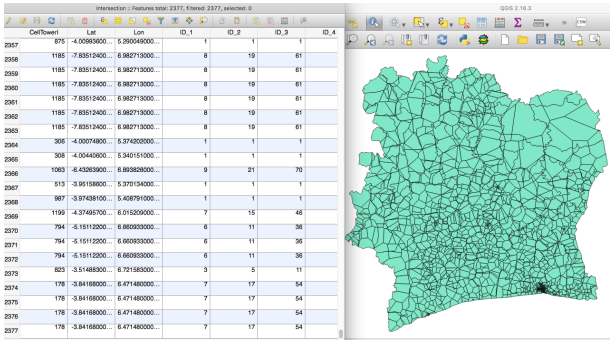


Figure (11.4): Step 5 - intersection of the voronoi regions with the regional shapefile.

Step 6: Now we want to intersect this with the population raster file, giving us a population per region. Start by inserting the raster files: Layer -> Add Layer ->

Raster Layer -> Pop_2010.tif and Pop_2014.tif.

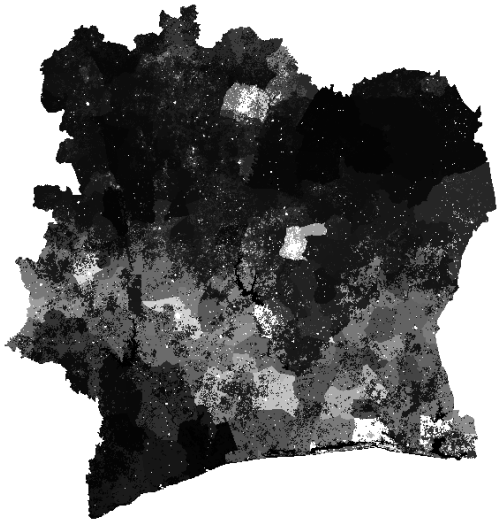
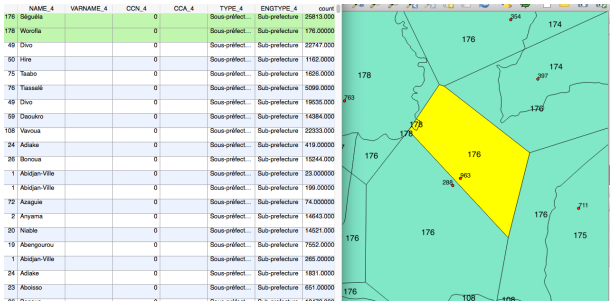


Figure (11.5): Step 6 - Raster file.

Step 7: Rather than manually performing an intersection and adding the population to each section, QGIS has a set of functions under 'zonal statistics' that performs everything efficiently. Select Raster -> Zonal Statistics making sure to set the attributes file as the intersection file from Step 5. This will append a count of the population to the intersection file - save this to a csv after performing the zonal statistics for both raster files.

Step 8: We can check that this has provided us with what we want in the following figure. We see that the voronoi region corresponding to cell tower 963 spans over administrative regions 176 and 178. Again, assuming that CDR data is split according to the proportion of it overlapping each administrative region (perhaps unlikely), we see that 20 out of 26984 people correspond to. It feels a more natural way to compute the population like this since we are computing metrics per voronoi area - it makes sense to compute the population per voronoi region, and then aggregate afterwards, rather than aggregating, and then dividing by the population of each region.



	NAME_4	VARNAME_4	CCN_4	CCA_4	TYPE_4	ENGTYPE_4	count
176	Séguéla		0		Sous-préfect...	Sub-prefecture	25813.000
178	Worofla		0		Sous-préfect...	Sub-prefecture	176.00000
49	Divo		0		Sous-préfect...	Sub-prefecture	22747.000
50	Hire		0		Sous-préfect...	Sub-prefecture	1162.0000
75	Taabo		0		Sous-préfect...	Sub-prefecture	1626.0000
76	Tiassalé		0		Sous-préfect...	Sub-prefecture	5099.0000
49	Divo		0		Sous-préfect...	Sub-prefecture	19535.000
59	Daoukro		0		Sous-préfect...	Sub-prefecture	14384.000
108	Vavoua		0		Sous-préfect...	Sub-prefecture	22933.000
24	Adiake		0		Sous-préfect...	Sub-prefecture	419.00000
26	Bonoua		0		Sous-préfect...	Sub-prefecture	15244.000
1	Abidjan-Ville		0		Sous-préfect...	Sub-prefecture	23.000000
1	Abidjan-Ville		0		Sous-préfect...	Sub-prefecture	199.00000
72	Azaguie		0		Sous-préfect...	Sub-prefecture	74.000000
2	Anyama		0		Sous-préfect...	Sub-prefecture	14643.000
20	Niable		0		Sous-préfect...	Sub-prefecture	14521.000
19	Abengourou		0		Sous-préfect...	Sub-prefecture	7552.0000
1	Abidjan-Ville		0		Sous-préfect...	Sub-prefecture	265.00000
24	Adiake		0		Sous-préfect...	Sub-prefecture	1631.0000
23	Aboisso		0		Sous-préfect...	Sub-prefecture	651.00000

Figure (11.6): Step 1 - administrative level 4 shapefile with attribute table also encapsulating adm_1, adm_2 and adm_3.

Step 9: Using Python or otherwise, we now need to sum the populations of people within each administrative region, by grouping the data and summing. I did this by, resulting in ...

In summation, I have extracted the population (as of 2010 and 2014) covered by each cell tower under consideration.
