Location Optimization of Healthcare Facility

Data Science Approach – Guidance Paper

Version 1.0

22-Jun-2018

Authors

Pushpak Banerjee, Infosys Advanced Analytics

Sudarshan Gopalan, Infosys Advanced Analytics

Infosys®

Navigate your next

Contents

Appendix

1 Use Case

Infosys®
Navigate your next

Use Case Description

Optimize the location of N1 health posts to be setup in a geographical region in next N2 years.

Geographical Region – Some geographical area that can be defined by GIS Vector Polygon

N1 – No. of Health Posts

N2 – No. of Years

Optimization Goal

Some of the critical goals that health care plans try to achieve are:

- Improvement in accessibility[1][6]
  - Spatial Factors [2]
    - Maximize Service Coverage
    - Minimization of Travel needs
  - Non-spatial Factors [3] [4] [5]
    - Demographic and Socio-economic factors [e.g. Age, Sex, Race, Social class, Income etc.]
- Minimization of burden on health facilities (to ensure uniform quality of service across facilities)
- Achieve better outcomes (e.g. reduction of infant mortality, maternal mortality etc.)

For planning the location of health care facilities we try to strike a balance between the above competing objectives.

Catchment Area

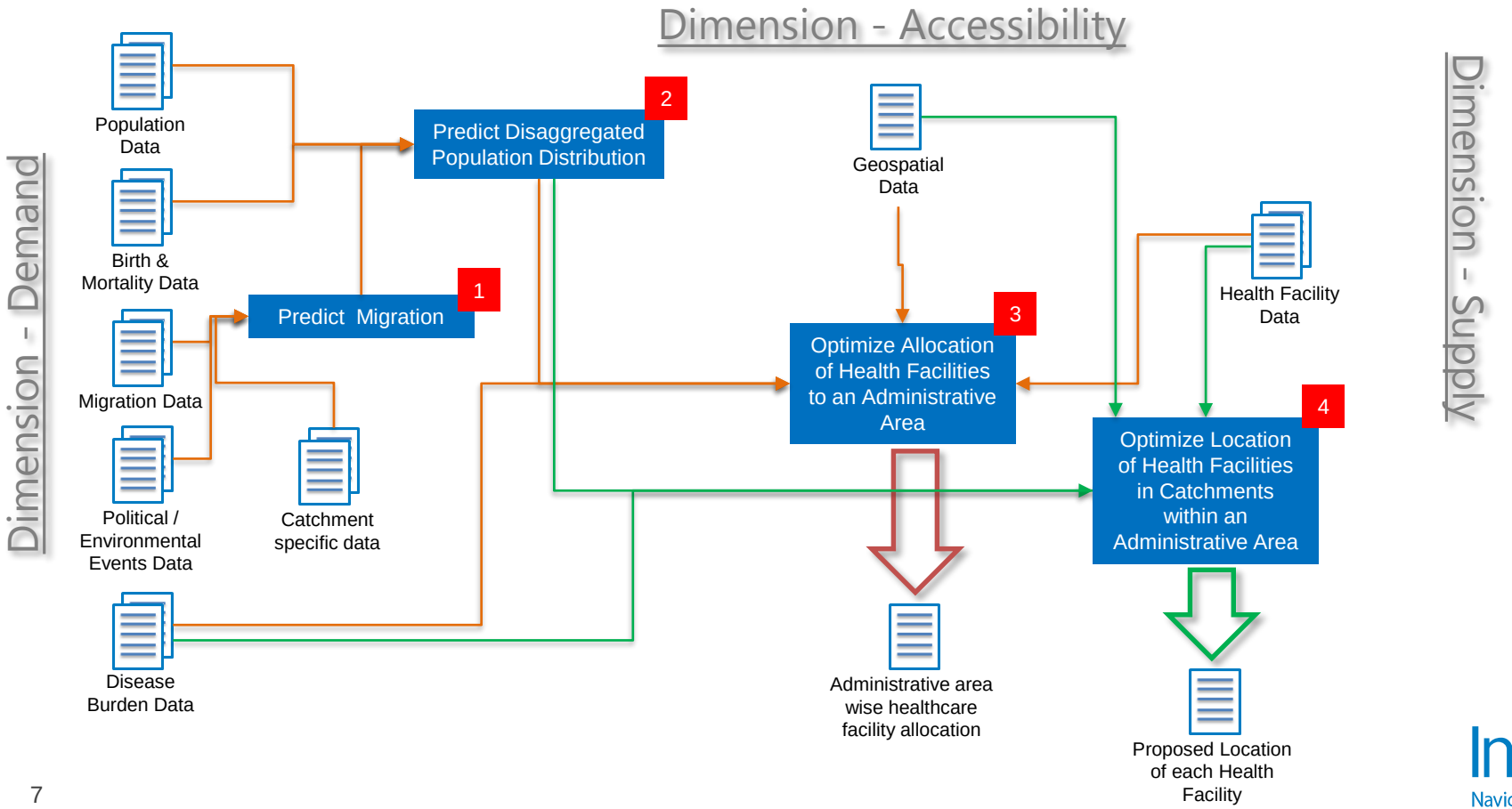
Geographical area around the health facility that includes or attracts the patient population who access its services. [7]

Catchment Criteria	Rationale
Administrative boundary	Typically decided by the government or ministry of health
Distance (straight line or road network)	Based on the straight line distance from the facility. [8]
Travel time	Based on time to travel to the facility. [9]
Cumulative-case ratio for specific disease	Uses Patient Flow method. An geographical area is included in the catchment of a facility if the number of patients, with a specific ailment, visiting the facility is greater than a set minimum proportion. [8]

Key Challenges in determining catchment population [10] :

- People access facilities outside their catchment area based on socio-economic factors or specific needs
- Catchment areas overlap
- geographical boundaries between country administrative divisions and facility catchments are not compatible or coterminous

Use Case Solution Approach



Infosys® Navigate your next

Malawi Use Case - Definition

Use Case Definition[Note 1]

Develop an analytical model to forecast / optimize the 900 health posts to be setup in a region in next 5 years using the MNO and Geospatial data.

Geographical Region – Country/District – Malawi

N1 – No. of Health Posts – 900 Health posts

N2 – No. of Years – 5 Years

Catchment Area[Note 2]

For Malawi, all the analysis and optimization needs to be done at the level sub-administrative units of GVH / Traditional Authorities.

[Note 1]: Finalized in Meeting with Cooper Smith Jun 26, 2018

[Note 2]: Finalized in Meeting with DIAL & Cooper Smith Jul 31, 2018

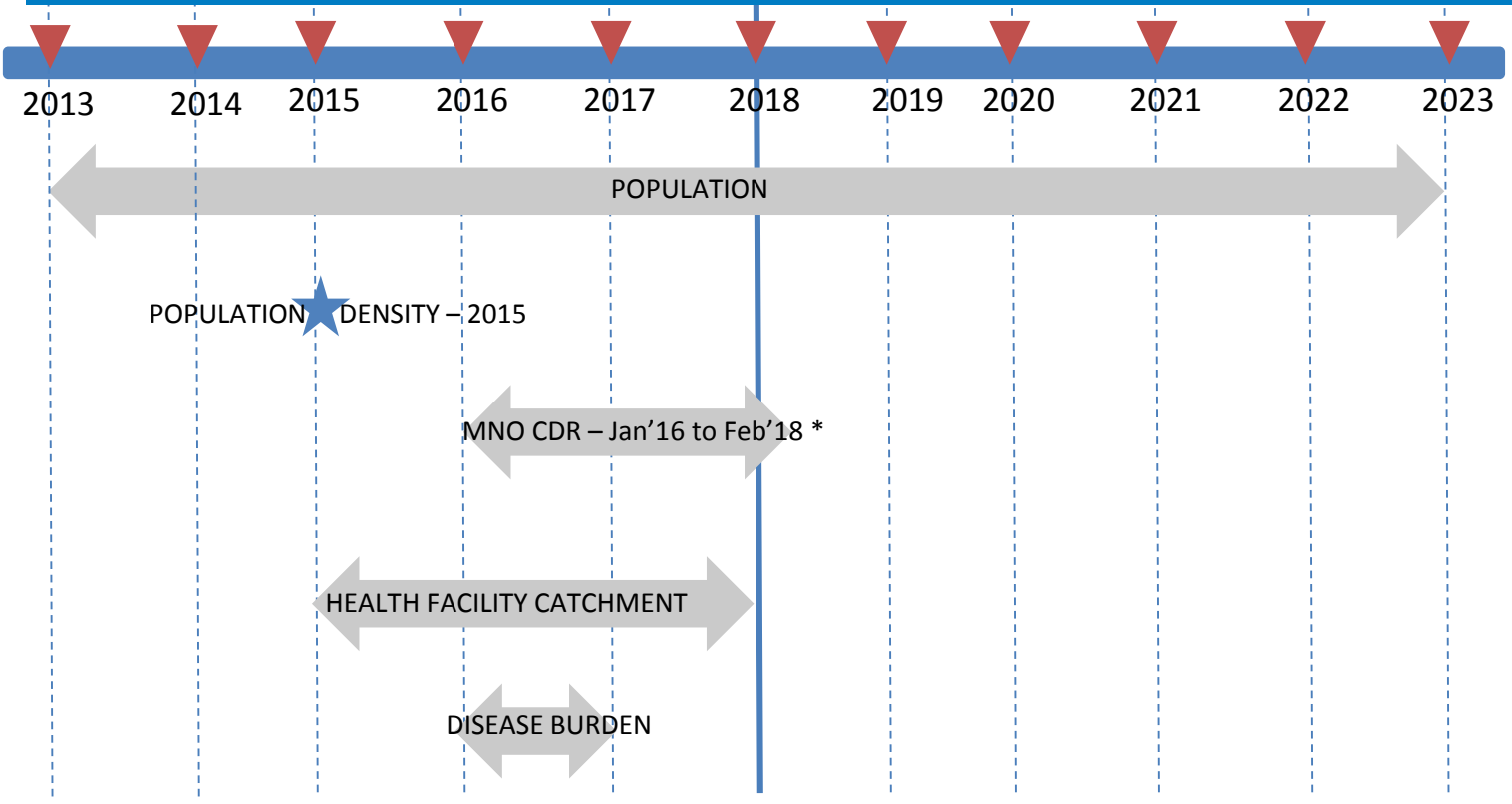
Malawi Use Case – Data Considered

Data Type	Source Used	Usage
Predict Migration		
Migration Data	CDR Data from Mobile Network Operators	CDR data will be used as a proxy of determining residence location and people migration.
Event Data	None Considered yet	This will be needed to determine migrations triggered by epidemics and geo-political shocks
Predict Disaggregated Population Distribution		
Aggregate Population	UNICEF Population projection	This will be used as the base population for prediction purposes.
Disaggregated Population	Facebook High Resolution Settlement Layer	This will be used to disagregate the Unicef projection to arrive at the population in smaller geo boundaries.
Birth & Mortality	None Considered yet	To be reviewed for impact on Population projection.
Geospatial Data	Administrative shape files – District (ADM1), and TA/GVHs (ADM2)	Will be used to compute GVH/TA level population densities from Facebook High Resolution Settlement Layer.

Malawi Use Case – Data Considered (continued)

Data Type	Source Used	Usage
Optimize Allocation of Health Facilities to an Administrative Area		
Disease Incidence	Malawi Burden of Disease Data	Understand disease related patterns on demand for Health Care.
Health Facility Data with Catchment population	CH District data summary with Facility Catchment details [2015-2018]	Facility level data to understand supply of health care facility
Allocation Prioritization Rules/Constraints	<None considered yet – inputs such as Facility Capacity, Annual Budget, Regional Priorities etc. >	Will be used to compute GVH/TA level population densities from Facebook High Resolution Settlement Layer.
Health Facility Data	<TBD – data on New Facilities already under implementation>	This will be used to compute the total available supply in a future point in time
Optimize Location of Health Facilities in Catchments within an Administrative Area		
Geospatial Data	<TBD - Driving Routes and Travel time data>	This will be used to determine spatial accessibility for a proposed facility.

Malawi Use Case – Data Available



Granularity - Temporal/Geo

- POPULATION — YEARLY / DISTRICT (ADM1)
- POPULATION DENSITY – 2015 — YEARLY / SUB TA (ADM2)
- MNO CDR – Jan'16 to Feb'18 * — DAILY / TA (ADM2)
- HEALTH FACILITY CATCHMENT — YEARLY / TA (ADM2)
- DISEASE BURDEN — MONTHLY / TA (ADM2)

* - MNO CDR data expected by 1st week of September

Non-Temporal Spatial Lookup/ Reference › DISTRICT (ADM1) / TA (ADM2) Shape files, Existing Health Post & MNO Tower Locations

2 Data Science Methodology

Infosys®
Navigate your next

Scope & Methodology

Models to be developed

1. Predict Migration
2. Predict Disaggregated Population Distribution
3. Optimize Allocation of Health Facilities to an Administrative Area
4. Optimize Location of Health Facilities in Catchments within an Administrative Area

Methodology

CRISP-DM [11]

Infosys® Navigate your next

Predict Migration – Key Data Considerations

Possible Data Sources	Processing Considerations	Insights to be extracted
CDR Data from MNO	- Segregate A2P data from Individual data - Address single ownership of multiple devices/multiple SIMs - Ability to link towers to geographical areas (administrative catchments or Facility catchments)	Spatio-Temporal movements (with snapshots across different time periods): - Permanent migrations - Multi-step migrations - Seasonal movements - Identify residential vs. commercial zones Identify Social Networks: - Identify communities - Community specific migration patterns Event specific (If Event Data leveraged): - Causal effects of Events on Migration patterns
Event Data from Google GDELT	- Large raw files. Will need to leverage Topic Modelling techniques to extract key events - Large data & computing requirement. May need Google Cloud based solution.	
Source and Destination zone Specific Data - Unemployment - Agriculture growth - Area - Weather - Infrastructure etc.	- The quality, granularity and time of data captured needs to be closely monitored for these data sources	

Predict Migration – Model Options

Approaches	Modelling Techniques
Macro Approaches	- Gravity Model - Radiation Model - Intervening Opportunities Model - Markov Chain migration model
Micro Approaches	- Diffusion migration models - Human Capital Models
Machine Learning Approaches	- XG Boost - ANN

3 Tools

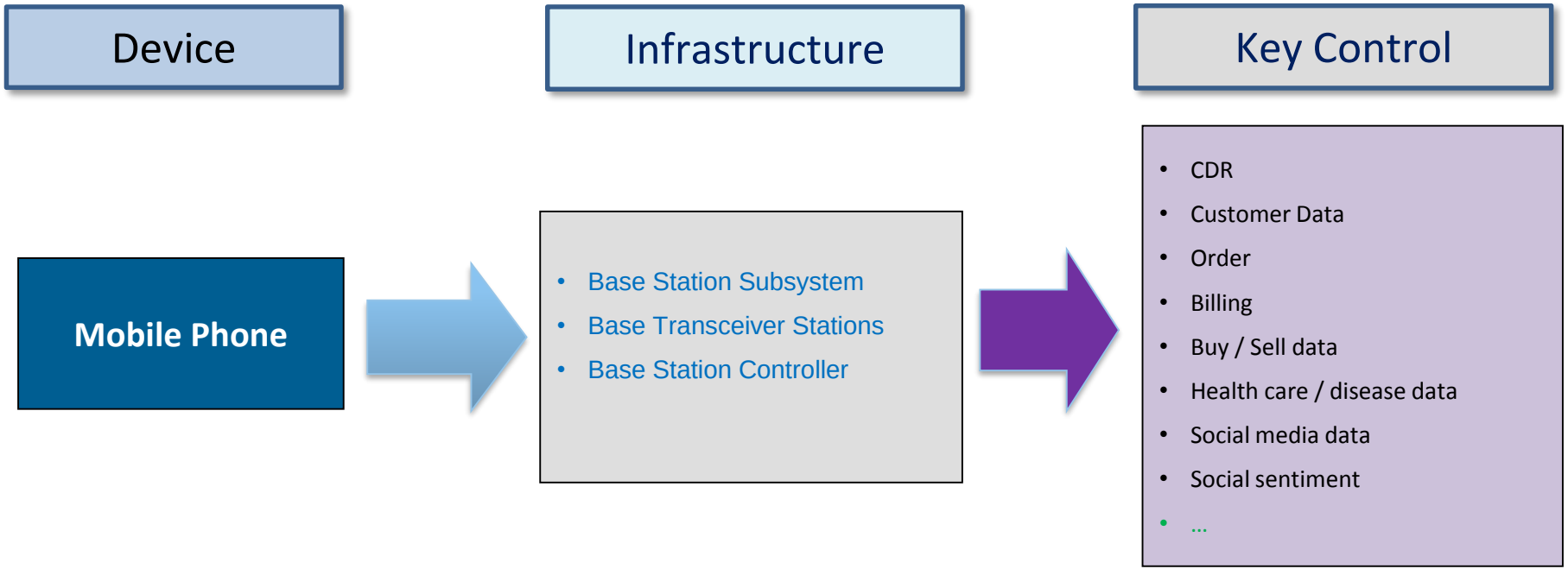
Analysis Tools

- Python for scripting
- PowerBI for visualization and exploratory analysis
- QGIS for GIS Analysis
- Neo4J/OrientDB for Graph/Network models (Migration Analysis)

Appendix

Infosys®
Navigate your next

What data is available from Mobile Networks?



Infosys®
Navigate your next

Opportunity Landscape - What insights are available from Mobile Data?

CDR Data

Type of Insights	Sample Insights	Possible Used Cases#
Mobility [Spatial-Temporal Movement Patterns]	• Home vs. Work location • Seasonal migration • Temporal gatherings • Significant locations / places	• Quantifying the impact of human mobility on malaria • Linking the Human Mobility and Connectivity Patterns with Spatial HIV distribution • Human mobility and communication patterns - A network perspective for malaria control
Social Networks [Connections / Strength of Connections]	• Health / Epidemic spreading • Relationships (friend, family, work, etc.) • Communication (face to face, mobile phone, email, etc.) • Communities detection	• Detecting Anomalies and Supporting Community to Ensure Healthy Society • Human contact and diffusion of Ebola Virus • Disease outbreak detection by mobile network monitoring
Socio-Economic [Literacy, Poverty]	• Population census • House hold income • Poverty level • Economic crisis	• Mapping poverty using mobile phone and satellite data • Human mobility and the spreading of diseases • Mapping and Measuring of social disparities using mobile phone subscribers data

- More details of each Use Case in Section 4

Challenges

Challenges faced with Mobile Data - Overview

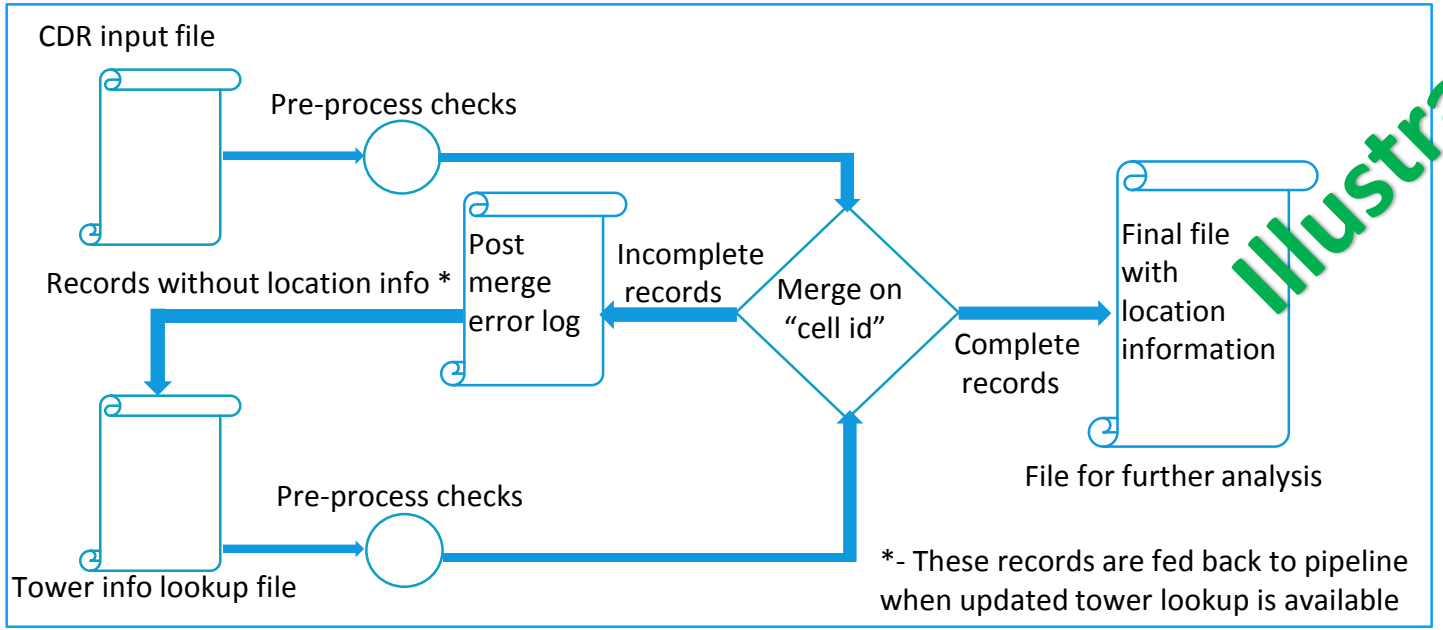
	Key Challenges	Solutions Explored
Ownership	• MNOs or Customers own the data (depending on the regulation) • Conflict between technology innovation and ownership • Missing model of ownership	• Privacy-preserving technical solutions • User control • Set-up a governance board (local and global)
Data Specific	• Availability of data • Data sources and integration • Data discontinuity • Data accuracy • Noisy Data • Individual and tele marketing data mixed	• Data handling (Volume of data) • Data preparation • Data quality report
Regulatory	• Identification of a person (violates privacy) • Data retention period • Data availability & accessibility • Adherence to country specific data protection laws	• Anonymization algorithms • Hashing identifiers • Access control • Statutory rights of access • Data privacy and protection

Recommendations

Infosys®
Navigate your next

Configurable Data Pipeline

Data Pipeline for CDR (call details records)



Illustrative

Data pipeline-Advantages

Streamlines the entire process of providing inputs to obtain output. It provides great flexibility. Any particular part of the code can be altered without disturbing the main file / any other part of the programs.

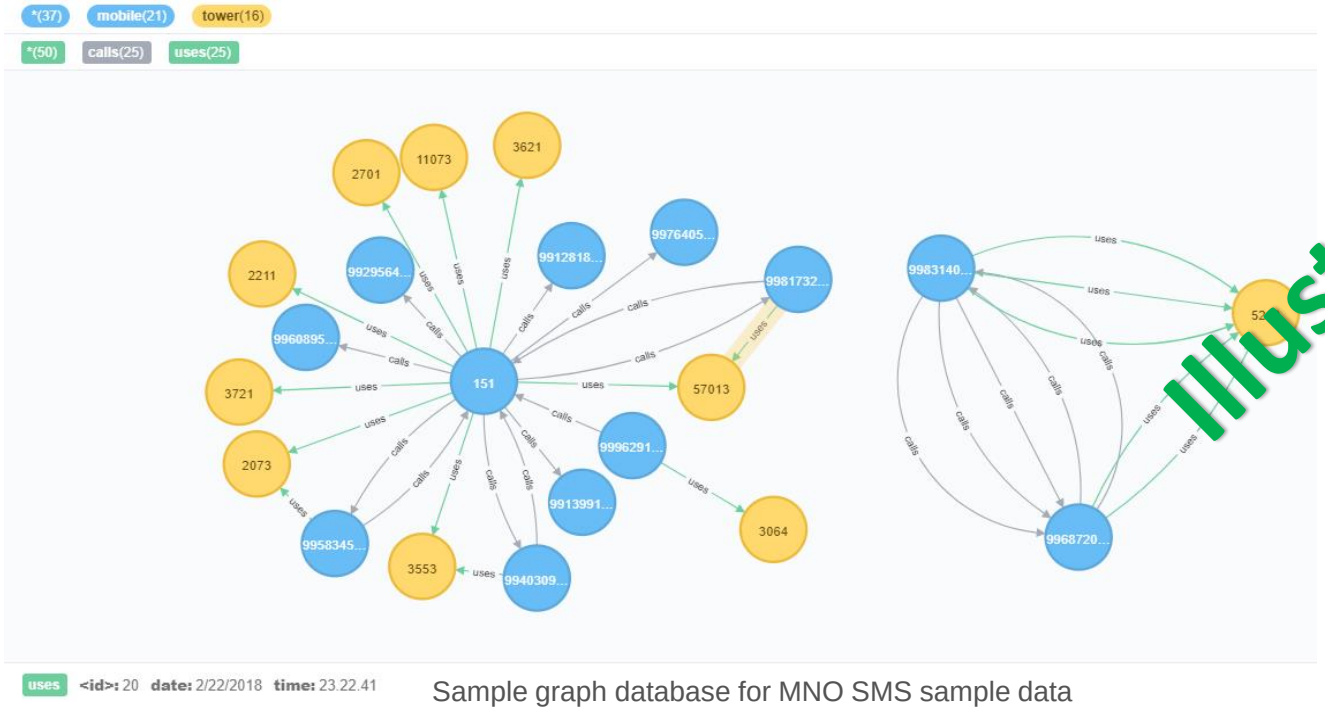
CDR data pipeline:- At the first stage CDR and tower lookup are taken as input files. They are passed for pre-processing checks. Post those checks, they are merged . The merged file obtained is used for further analysis such as finding P2P communications, A2P communications. Also among P2P communications based on the objective from local events to urbanization can be identified.

Mobile usage segregation

- Data pipeline provides a mechanism where raw input files are made ready for further analysis.
- Type of CDR segregation required depends on the end objective that we wish to achieve.
- Processed CDR files from data pipeline covers both Application to person SMS (A2P SMS) and person to person SMS (P2P SMS).
- If the end objective of this exercise is to understand the society and its dynamics, steps must be taken to remove A2P SMS from the dataset to avoid spurious analysis.
- A2P SMS broadly speaking contains bulk sms, sms related to order confirmation, OTPs, voting lines via sms etc. These SMSes must be identified and removed from the analysis.
- Once only P2P SMS is retained, level of aggregation of dataset spatially and temporally depends on our end objective.
- The level of aggregation gets closely associated with the metrics used in the analysis. Call frequency is useful when aggregation is of shorter time period while unique callers from a specified location may be used even in case of longer time period.
- Various types of analysis and associated type of aggregation -
    - Social interaction, location of home-work mappings : These analysis are best suited with individual data without aggregation. Since it is worked on individual level data this might be very computationally intensive.
    - Event detection:- Call frequency from a location hour-wise/day-wise may be analyzed to see a deviation from normal situation to identify events/ gatherings. If a pattern is identified these events may be predicted as well (e.g. frequent communicable disease may be predicted). This requires data at smaller temporal aggregations (hour wise/day wise) and spatially may be at tower level.
    - Urbanization, migration:- This analysis considers long term time horizon. For this analysis we might not need data in granularity. ,By this we mean, urbanization is a long term phenomenon so, data at weekly/monthly aggregation at site/town level for long periods of time might help us analyze it.

Graph based Data and Analytics Technology adoption

- Neo4j is the world's leading open source graph database management system. In order to leverage data relationships, we use graph database that stores relationship information.
- Neo4j is highly suitable for storing data that has many interconnecting relationships. This is where graph databases can make a huge difference.
- Using Neo4j for storing CDR (Call Detail Records) data, helps in better visualizing the connections between two mobile phone users.



Illustrative

Sample graph database for MNO SMS sample data

Infosys® Navigate your next

Use Case Details

Infosys®
Navigate your next

Use Case: Quantifying the impact of human mobility on malaria

Description

An increase in migration could lead to increased short term movement that can have important health effects in the destination. To examine whether there is a link between the movement data and the malaria data. CDR data is used to study population movement.

Value Delivered

Malaria is seasonal, with almost all cases falling between July and December. This is largely driven by the rainy season which occurs during that time period. For many of the locations, we do not see a strong link because there are few cases of malaria. In the locations where there are more case of malaria, there does seem to be a pattern of bumps in movement followed by bumps in malaria.

Inputs

Malaria patient details for the year 2013 - total # of malaria reported cases : 345,889 and # of death reported (malaria) : 815.

CDR data - # of calling 146,352 individuals between January 1, 2013 and December 31, 2013.

Malaria Data – From middle of February 2013 to the end of November 2014.

Solution Approach (including Algorithms)

EDA- Total Call distribution Day wise

People Calling per Day as Percent of Sample Living in a Particular area

Heat map of Malaria Prevalence

Effect of Number of People Returning after Visiting a High Malaria Place for 3-14 days

Multiple linear regression

Insights Created

The analyses show an increase in malaria due to people returning home, but no effect on the places that people visit.

It has been established a link between short term movements around the country and increased malaria.

References

Quantifying the Effect of Movement Associated with Holidays on Malaria Prevalence Using Cell Phone Data - D4DChallengeSenegal_Book_of_Abstracts_Scientific_Papers.pdf

Use Case: Linking the Human Mobility and Connectivity Patterns with Spatial HIV distribution

Description

The HIV pandemic has devastating effects on entire human population in Africa. Ivory Coast has a generalized HIV epidemic with the highest prevalence rate in the West African region, 3 percent. Although the prevalence rate appears to have remained relatively stable for the past decade, nowadays there are several studies which declare that this number is even increasing, especially due to war conflicts. Deeper understanding of epidemics can help stop this trend and find ways to suppress it.

Value Delivered

Study showed how crude real world data can be used for significant knowledge extraction. We addressed the problems of HIV/AIDS spatial distribution prediction by analyzing human activity and mobility in the area of extracted features. However, the results leave a lot of room for improvement, especially in the field of defining the features which affect disease transmission the most.

Inputs

HIV Spatial Distribution

Mobile Data Sets

Solution Approach (including Algorithms)

Graph Representation - Pairwise connectivity of regions by measuring the of communications and migrations between them.

Ridge Regression

Insights Created

Migrations of people are higher during the working hours, as well as one hour before and after work.

Migrations of people are higher during weekend due to the lack of specific contents in their own environment.

References

Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast, May 1-3, 2013 Linking the Human Mobility and Connectivity Patterns with Spatial HIV distribution , K.GAVRIC, S.BRDAR, D.CULIBRK, V.CRNOJEVIC

Use Case: Human mobility and communication patterns - A network perspective for malaria control

Description

Objective was to assess the importance of mobility and communication patterns for malaria control efforts. Based on mobile phone usage behavior, geographic networks reflecting patterns of human movement (based on where individuals made calls) and patterns of communication was constructed.

Value Delivered

The hypothesis of a highly cost-effective reduction in mean malaria prevalence nation-wide by targeting sub-prefectures of high mobility/communication can be tested using simulation models of both disease transmission and information dissemination, and the costs and benefits of different intervention strategies for the control of malaria under a variety of scenarios can also be predicted.

Inputs

Mobile phone usage dataset - 500,000 randomly selected mobile phone users from Dec 1, 2011 to April 28, 2012.

Malaria presents a particularly large burden estimated 2.5 million suspected malaria cases and1 44,000 malaria infections requiring hospitalization in 2011 among its 21million inhabitants.

Solution Approach (including Algorithms)

EDA for-

Population density and mobility by sub prefecture

Malaria prevalence and mobility by sub prefecture

Population density and communication by sub prefecture

Population density and communication by sub prefecture

Insights Created

A total of 16.9 million movements between sub-prefectures were observed, with each user making an average of 33 such movements over the observed period.

Approximately 63% of movements were reciprocal.

We also compared the prevalence of malaria between sub-prefectures with strong connections, through either human movement or communication

References

Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast, May 1-3, 2013

Human mobility and communication patterns in Cote d'Ivoire : A network perspective for malaria control, E.A. ENNS, J.H.AMUASI

Use Case: Detecting anomalies and supporting community to ensure healthy society

Description

Detecting anomalies and supporting community to ensure healthy society

Value Delivered

For anomaly detection, anomaly was defined as deviation from regular call volume. For each tower, VAR model was applied with lagged dependent variables and exogenous variables to estimate regular pattern. Although, explanatory variables are limited to holiday, month, and hour dummy that are easily defined, model fit was satisfactory.

Inputs

Call Detail Records (CDR) from January 2013 to December 2013.

Solution Approach (including Algorithms)

As call and SMS shows distinct pattern, an anomaly is defined as "deviation from regular pattern". To detect anomalies, call (SMS) volume were related to previous call volumes and exogenous variables. Four time series variables: volume of incoming and out going volume for call and SMS, they are explained by lagged four variables and exogenous variables.

Insights Created

As observed, call and SMS shows distinct pattern, anomaly is defined as "deviation from regular pattern". To detect anomalies, call (SMS) volume were related to previous call volumes and exogenous variables. Four time series variables: volume of incoming and out going volume for call and SMS were considered.

References

D4DChallengeSenegal_Book_of_Abstracts_Scientific_Papers.pdf, Detecting anomalies and supporting community to ensure healthy society Yutaka Hamaoka

Use Case: Mapping poverty using mobile phone and satellite data

Description

In 2015, approximately 700 million people lived in extreme poverty. Poverty is a major determinant of adverse health outcomes including child mortality and contributes to population growth, societal instability and conflict. Eradicating poverty in all its forms remains a major challenge and the fist target of the sustainable development goals. To eradicate poverty, it is crucial that information is available on where affected people live.

Value Delivered

Models employing a combination of CDR and RS data. However, RS-only and some CDR-only models performed nearly as well good. The fine granularity of the resultant poverty estimates shows the predicted distribution of poverty for all three measures

Inputs

Poverty data of Bangladesh through – 2011 Bangladesh DHS; 2014 FII survey with data collection on PPI and national household surveys conducted by Telenor group between Nov, 2013 and Mar, 2014.

CDR data between Nov, 2013 and March, 2014.

Solution Approach (including Algorithms)

Covariate selection- Prior to statistical analysis, all CDR and RS covariate data were log transformed for normality. Bivariate Pearson's correlations were computed for each pair pf covariates to assess multicollinaraity and for high correlation.

Hierarchical Bayesian geostatisy tocal models to predict three povery metrics at unstamped locatiosnacross the population.

Spatial autocorrelation in the data.

Insights Created

Models employing a combination of CDR and RS data. However, RS-only and some CDR-only models performed nearly as well good. The fine granularity of the resultant poverty estimates shows the predicted distribution of poverty for all three measures

References

Mapping poverty using mobile phone and satellite data J.R.Soc. Interface 14: 201606901

THANK YOU

Infosys®
Navigate your next