

# Foundational Approach to optimize Allocation & Positioning of Health Posts in Malawi – Analysis Approach


Ver 1 1 1  
10/12/2018



# Contents

- 1 Discussion on Feedback
- 2 Foundational Approach Overview
- 3 Population Density Mapping
- 4 Short Term Population Movement Analysis
- 5 Long Term Population Movement Analysis
- 6 Assumptions

# Discussion on Feedback

#	Feedback	Feedback from	Infosys Response	Feedback incorporation status
1	Step 1 - While appropriate in densely clustered, urban areas, the size of these polygons should be capped in rural areas to the estimated maximum range of cell-phone towers (5-10km radius from node), in order not to under-estimate the call-density.	C/S	<ul style="list-style-type: none"> <li>As discussed and agreed with C/S and DIAL on 18-Sep, we will assume the max range of each tower as 10 km, refer appendix (slide #42) for more details.</li> </ul>	Completed
2	Step 1 - Administrative units outside of this range should be excluded from the analysis.	C/S	<ul style="list-style-type: none"> <li>Administrative unit without any mobile coverage will be excluded from our analysis. ~ 85% population will be covered if tower radius (range) is 10 km [Cooper/Smith - Erwin]. We have also validated and this value is 84.65%.</li> </ul>	Completed
3	Step 1 - C/S will conduct a parallel 'gap analysis' of these excluded populations to address potential equity concerns.	C/S	<ul style="list-style-type: none"> <li>C/S Gap analysis approach details have to be shared with Infosys</li> <li>We will utilize the gap analysis inputs in our analysis</li> </ul>	NA
4	Step 2a - $\text{Log}(P_j) = \alpha + \beta \log(D_j) + \delta_D + \epsilon_i$	C/S	<ul style="list-style-type: none"> <li>We will consider applying district fixed effect. While there are various factors like behavioral differences in cell-phone use (week day v/s week end, average call duration, A2P origin region, etc.) and demographic/ socio-economic indicators (mean age, education attainment, etc.), we can consider this based on data availability and demographics subject matter inputs from C/S.</li> </ul>	Completed
5	Step 2a - Test for Spatial Correlation using Moran's I and if reject the null, use a spatial lag model accounting for proximity of units, or cluster error terms.	C/S	<ul style="list-style-type: none"> <li>This will be addressed as part of the Spatial Correlation analysis performed during Regression Analysis. pysal python library will be used for Spatial Correlation analysis.</li> </ul>	NA 

# Discussion on Feedback Contd...

#	Feedback	Feedback from	Infosys Response	Feedback incorporation status
6	Step 2b - Spatial correlation issues addressed directly in the specification (see above). Recommend against sampling a coordinate with 100 km radius, as parameters will be region specific and cannot be extrapolated to the country as a whole.	C/S	<ul style="list-style-type: none"> <li>We may try both random and stratified sampling. Final decision of the sampling will be taken based on the results from the data.</li> </ul>	Completed
7	Step 2b - Recommend using bootstrap analysis, sampling with replacement 1000 times and using distribution of coefficients to infer mean and standard deviation of $\beta$ . Infosys can work with C/S on this.	C/S	<ul style="list-style-type: none"> <li>We may try both random and stratified sampling. Final decision of the sampling will be taken based on the results from the data.</li> </ul>	Completed
8	Step 2b - In parallel, suggest k-fold cross-validation, randomly sampling 70% of data k times and testing on remaining 30% according to best practice. Use R2 or MSE as indicator of predictive performance. Infosys can work with C/S on this.	C/S	<ul style="list-style-type: none"> <li>We have considered this, even though it is not explicitly called out in our approach.</li> </ul>	Completed
9	<p>Step 2c - Test sensitivity of <math>\beta</math> to different CDR data, including –</p> <ol style="list-style-type: none"> <li>Daily aggregated vs. night-time</li> <li>Calls sent vs. calls received</li> <li>Weekday vs. weekend</li> </ol> <p>&gt; Need agreement on what the principal specification is (ie night-time calls sent on a weekday) and which subsequent robustness tests will be reported. &gt; Across these agreed upon CDR variations, test and compare for predictive power as outlined in Step 2b.</p>	C/S	<ul style="list-style-type: none"> <li>For sensitivity testing of <math>\beta</math>, following data can be considered - <ul style="list-style-type: none"> <li>✓ Daily aggregated vs. night-time and</li> <li>✓ Weekday vs. weekend</li> </ul> </li> <li>“Calls sent vs. calls received” can’t be considered as it is not available.</li> </ul>	Completed

# Discussion on Feedback Contd...

#	Feedback	Feedback from	Response	Feedback incorporation status
<u>10</u>	Step 2b - Calculating population growth based on estimated growth rate $r$ > Once 2a and 2b complete, Infosys working with C/S can calculate $r$ using population estimates projected from Step 2a using multiple years of CDR data: $r = \ln(P_{2017}/P_{2016})$	C/S	<ul style="list-style-type: none"> <li>Please note that our approach is to analyze population movements using CDR and not calculate population estimate using CDR data as a proxy for 2019-2023.</li> <li>We had discussion with C/S on 10/03/2018 and they are agreed for the above point.</li> </ul>	Completed
11	Update Metric sheet for CDR data (Syed provided)	DIAL	<ul style="list-style-type: none"> <li>This is incorporated in metric sheet</li> </ul>	Completed
12	Long term / population migration approach (step 2), remove "List of events (shock)" and better to include the same in short term / dynamic population movement section	DIAL	<ul style="list-style-type: none"> <li>This was an error and is corrected now.</li> </ul>	Completed
<u>13</u>	Revisit approach for short term and long term movement	DIAL	<ul style="list-style-type: none"> <li>Approach for short term and long term movement was updated.</li> </ul>	Completed
<u>14</u>	How will population density growth be modeled?	DIAL	<ul style="list-style-type: none"> <li>This will be done based on clarification inputs from Worldpop</li> </ul>	Open
<u>15</u>	Will any additional aggregated MNO data (like age, churn rate etc.) be useful for analysis?	DIAL	<ul style="list-style-type: none"> <li>As agreed by C/S, district fixed effect is not required i.e. one common <math>\alpha</math> (alpha) and <math>\beta</math> (beta) for entire Malawi is not required. Rather this will be run at regional level i.e. North, Central &amp; Southern regions. So we will have 3 <math>\alpha</math> (alpha) and <math>\beta</math> (beta).</li> </ul>	Completed

# Foundational Approach Overview

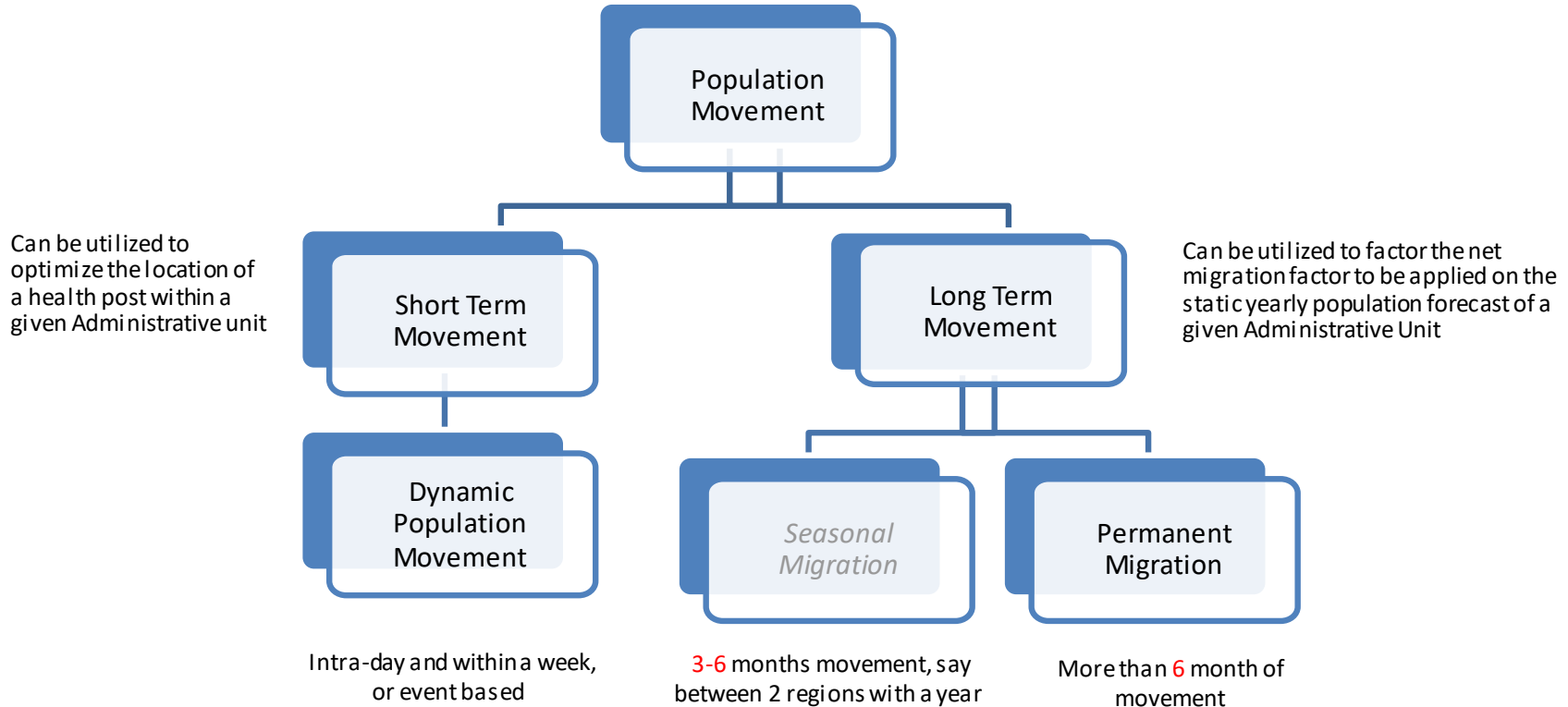
# Use Case Objective

Develop an analytical model to forecast / optimize the 900 health posts to be setup in Malawi over next 5 years [2019-2025] using the Anonymized CDR and Geospatial data (MNO Tower, Health Post) at Administrative Unit Level 2 (TA/GVH) .

There are 2 distinct levels of optimization involved as outlined below, and CDR data can provide insights for both

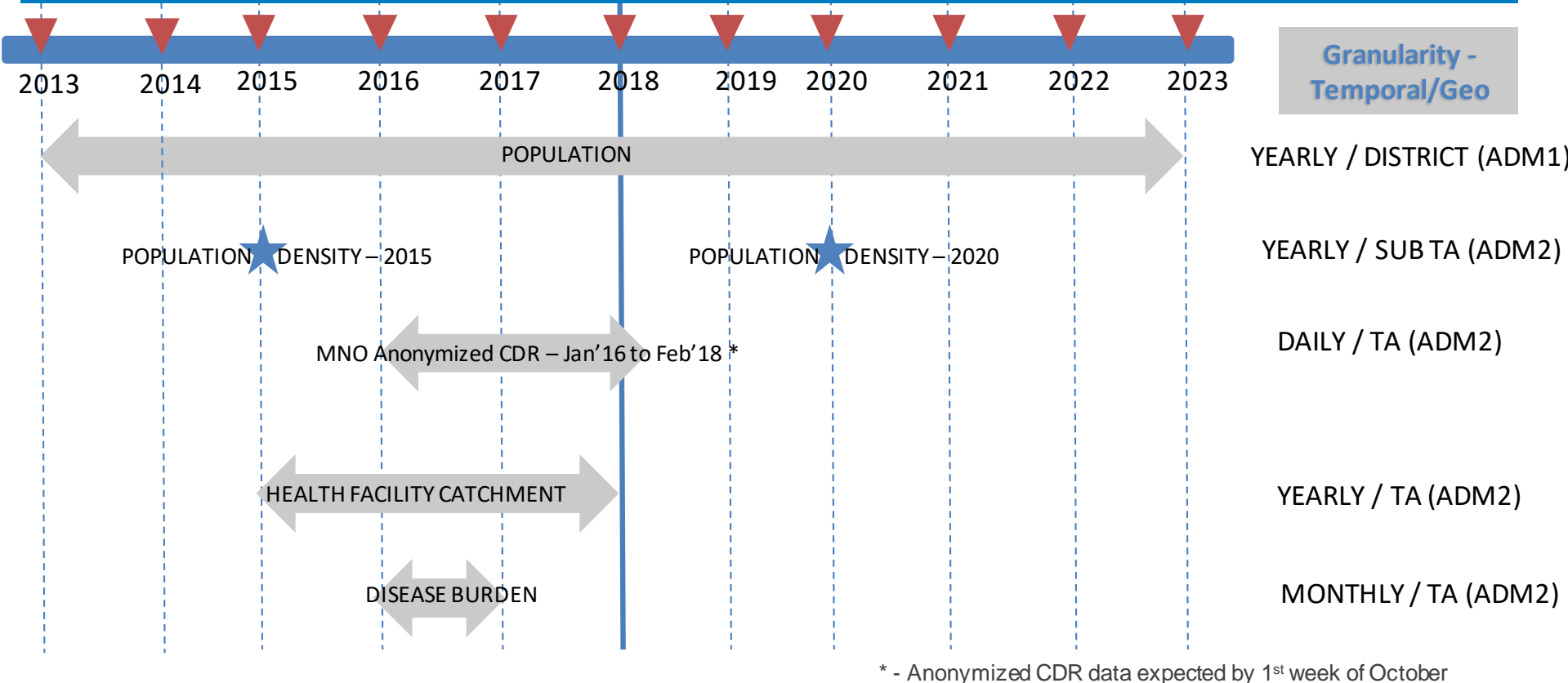
- Allocation of health posts to a given Adm Unit
- Positioning of health post within a given Adm Unit

# Estimating Population Movements through CDR Analysis





# Timescale of Available Data Coverage

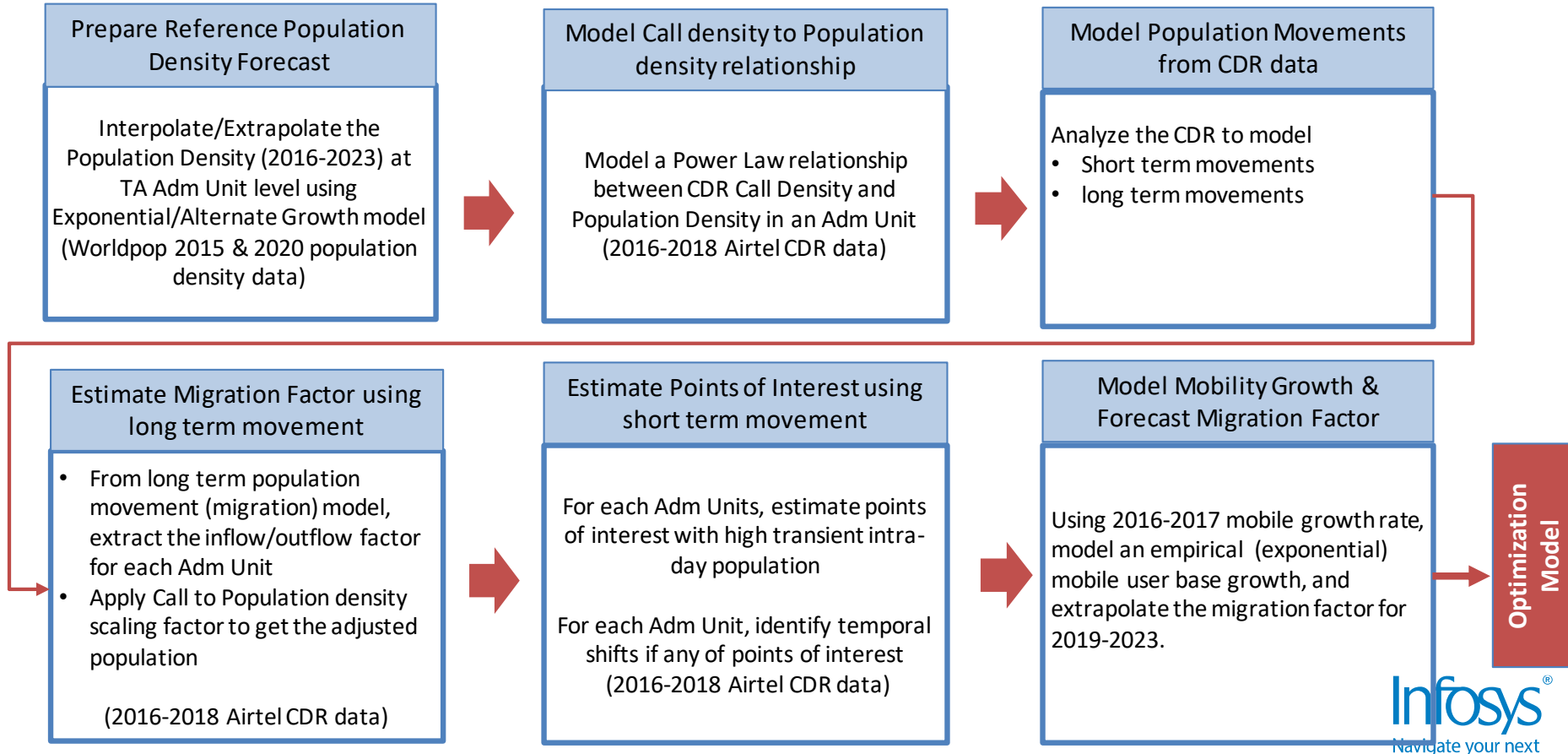


Non-Temporal Spatial  
Lookup/ Reference



DISTRICT (ADM1) / TA (ADM2) Shape files, Existing Health Post & MNO Tower Locations

# Foundational Approach - High Level Overview



# Population Density Mapping

# Population Density Mapping using Anonymized CDR - Approach

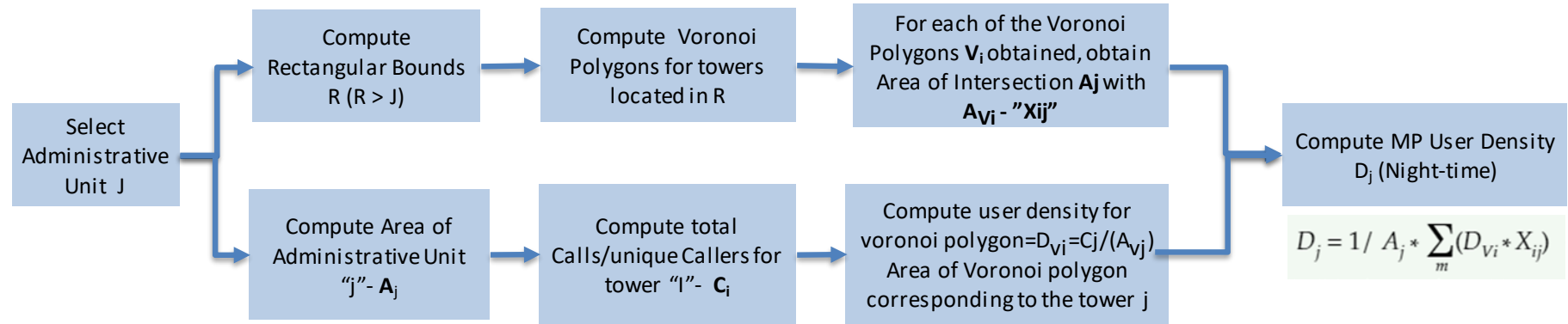
For scaling the population movements analyzed through Mobile CDR as a proxy, we need to establish a relationship between the mobile user density to the population density. This is done through two steps approach.



# Population Density Mapping Approach

## Step 1 – Compute Mobile User Density for each Administrative Unit

Obtain Voronoi polygons for towers and calculate overlap of administrative unit and polygons



Calculation of Mobile phone user density for each voronoi tower polygon

### Legend

Aj- Area of administrative "j"

Ci- Number of calls or unique callers at tower "i"

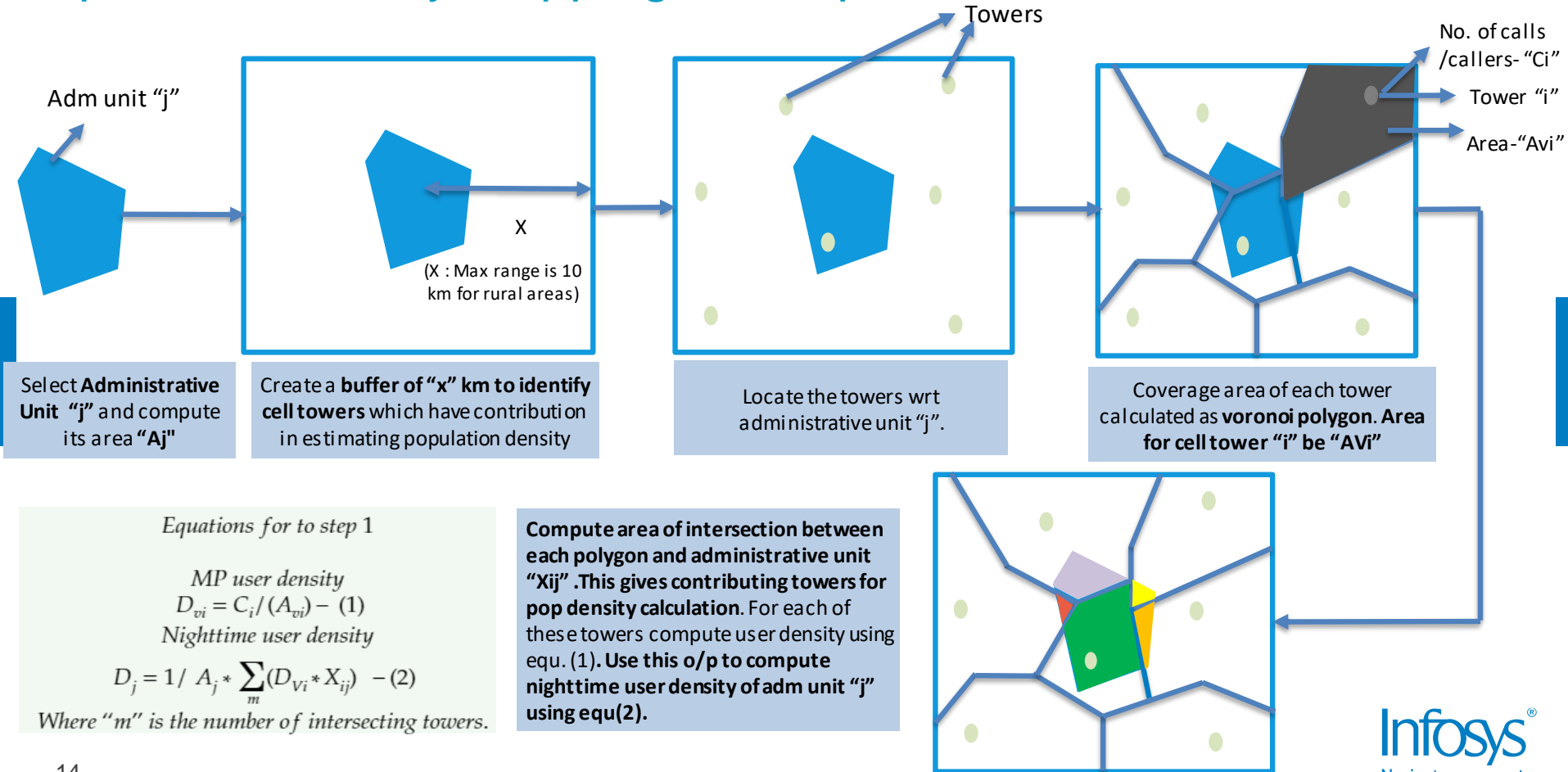
Vi- Voronoi polygon obtained for tower "i"

Avi- Area of voronoi polygon Vi

Xij- Area of intersection between area Aj and Avi

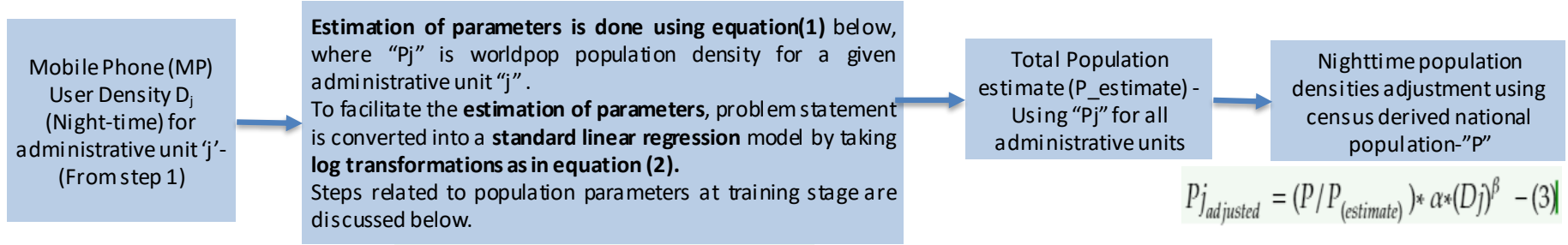
Dvi- Mobile phone user density –ratio of Ci/Avi

# Population Density Mapping >> Step 1 – illustration



# Population Density Mapping (without Fix Effect Factor) - Approach (contd..)

## Step 2.a – Magnify MP User Density against Census / RS derived population density using Power Law



### Parameters

$\alpha$  : Scale Ratio

$\beta$  : Super linear effect of population density (census) on night-time user density

- We would run the step 2.a region-wise (North, South, Central) resulting in 3 pairs of  $\alpha$  and  $\beta$ .

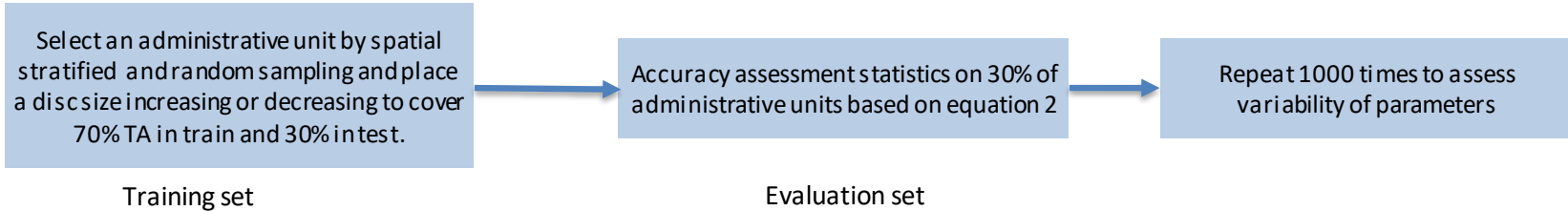
### Notes on availability of data

- Current availability of data for population in Malawi is at district level. In this specific case, population at administrative units is lower than districts in hierarchy (ADM-2(TA) and 3 (GVH) levels) is of interest.
  - Population data available at district level is not close to estimates obtained using Facebook HRSL file for the year 2015.
  - To estimate population density in Malawi, it is assumed that **Worldpop's 2015 estimates –ppp-adj-V2** version will be used instead of available district level data. Total population indicated for the year 2015 in **Worldpop pppadj\_V2** file is 17303306.73, which is very close to the value indicated in **Facebook HRSL file** (17201196.95) .
15. **Worldpop 2015 ppp V2 unadjusted** has total population count of 15799142.22.

[https://www.researchgate.net/publication/267449917\\_Dynamic\\_population\\_mapping\\_using\\_mobile\\_phone\\_data](https://www.researchgate.net/publication/267449917_Dynamic_population_mapping_using_mobile_phone_data)

# Population Density Mapping - Approach (contd..)

## Step 2.b – Cross validation procedures to derive “ $\alpha$ ” and “ $\beta$ ” parameters



- For Malawi, spatial resolution ( $\sqrt{\text{land area}/\text{total administrative units}}$ ) is around 5.5km (at ADM-3 level)\*\*.
- At the lowest level of administrative units, Malawi has around 3100 units.
- Land areas of Portugal and Malawi are comparable as well. Malawi has a land area of 4000 sq.km [<https://tradingeconomics.com/malawi/land-area-sq-km-wb-data.html>]
- While deciding % of data for train and test, we would also consider k-fold technique for cross-validation.
- Spatial correlation is studied using Moran I's test. Based on the test results we infer the correlation among units.
- With respect to sampling techniques (random/stratified), we may consider both the options. We will decide based on the results from the data

## Step 2.c – Sensitivity analysis of “ $\alpha$ ” and “ $\beta$ ” parameters

Study of population density estimations to population parameters

- Parameter “ $\alpha$ ” and “ $\beta$ ” will be tested for sensitivity using CDR data [Daily aggregated vs. night-time and Weekday vs. weekend CDR data]



# Estimating User Density Growth Rate

User Density growth rate can be estimated through

- Computed user density value for the years 2016 and 2017 based upon anonymized CDR data @ Administrative unit area

$$D_t = D_0 * e^{r * t}$$

$$r = \ln (D_{2017} / D_{2016})$$

## Legend

$D_t$  : User density for the year t

$D_0$  : User density for the initial year

$r$  : Annual user density growth rate (urban/rural)

$t$  : # of year between the initial year and input data and the year being projected

$D_{2017}$  : User density for the year 2017

$D_{2016}$  : User density for the year 2016

# Short Term Population Movement

# Use Case: Dynamic Population Movement– Approach

Approach- Two steps approach is suggested for dynamic Population movement

**Step 1** – Finding out the trajectory at individual level

**Step 2**- Tracing out the aggregate trajectory from individual level and analyze population mobility motifs

<b>Inputs</b>	<ul style="list-style-type: none"><li>Day wise anonymized CDR data aggregated at unique subscriber level</li></ul>	<ul style="list-style-type: none"><li>Individual trajectory recognition for each of the unique subscriber for a month</li></ul>
<b>Key Activities</b>	<ul style="list-style-type: none"><li>Sort the day wise call records for each of the unique subscribers according to the time stamp attached to them and map these calls to the respective towers. This will give us the daily trajectory of an individual.</li><li>Extract the stay points and divide move/ stay segments based on the parameters and threshold applied on the CDR data</li><li>From the daily trajectory of each of the individual identify the most frequent trajectory for each one of them over a considerable period of time (e.g. over a month)</li></ul>	<ul style="list-style-type: none"><li>From the individual level information aggregate to a administrative unit</li><li>Identify the most visited trajectories per administrative unit.</li><li>Identify frequently visited stay points per administrative unit.</li><li>Perform mobility motifs analysis for dynamic population movement</li></ul>
<b>Deliverables</b>	<ul style="list-style-type: none"><li>Individual trajectory recognition for each of the unique subscriber for a month</li></ul>	<ul style="list-style-type: none"><li>Most visited trajectories per Administrative unit</li><li>Most frequently visited stay point per Administrative unit</li></ul>

# Dynamic Movement Analysis

**OD Matrix** - Construct mobility matrix using ADM Unit - tower, tower-tower, tower – ADM Unit aggregated statistics, to analyze mobility between ADM Units

**Trajectory Analysis** - Analyze aggregated reconstructed trajectories between ADM units and within ADM unit to identify zones/places of interest, through home locations, stay points, movement routes, mobility motifs.

# Suggested Metric/KPI for Anonymized CDR data

Reference	As a	I want	So that	Applicability in Malawi Use case	Metric / KPI	Remarks	Feasibility	Relevance	Rationale for Relevance
A	Demand Side Agency (DSA)	to better understand the movement of population following a shock, such as an epidemic outbreak or natural disaster	I can make better informed decisions.						
	DSA	to understand where a population is moving to immediately following a shock	I can better support their humanitarian needs.	Yes	# of calls / callers during shock period	Assuming we would use events from Google projects-GDELT. One time events such as earthquakes or seasonal events such as floods, disease outbreaks along with dates and impact areas.	Yes	Yes	Movements associate with recurring shocks, floods, disease outbreaks helps better facility.
A2	DSA	to understand the movement of population during the period following a shock	I can monitor the situation, measure the impact of decisions and continue to make decisions or predictions.	Yes	# of calls / callers post shock period		Yes	Yes	
A3	DSA	I can see when and if people are returning to the affected location following a shock	I can understand the medium- to long-term impact on a region.	Yes	# of calls / callers at the shock location pre and post shock		Yes	Yes	
A6	DSA	To know the population of a region at a specific time	Measure the number of people exposed to, or at risk from, a given situation	Yes	# of calls / callers at that specific time for a given location	using population density approach	Yes	Yes	To decide the size and operation hours facility
A7	DSA	to see current population estimates based on mobile data for a particular area	I can use draw informed conclusions about the impact of a shock in a specific region.	Yes	we are already doing	we are already doing	Yes	Yes	To understand facility burden in an area
A10	DSA	to see population movement data over a specific time period	I can monitor how population movements have changed e.g. in response to interventions.	Yes	# of calls / callers at a location for a given time frame		Yes	Yes	In long term to optimize number of health based on population at a location. During understand number of temporary health required.
A11	DSA	to see predicted population movement estimates	I can prepare for an epidemic or disease outbreak, limit its severity and restrict its spread.	Yes	Identify seasonal population movement due disease outbreak	Based on previous disease outbreak population movement, future movements can be estimated.	Yes	Yes	For better operation planning in surrow facility. Understand temporary health fa
A12	DSA	to segregate application and human related CDR data	I can differentiate between application and human related CDR data	Yes	# of A2P contacts / day and # of P2P contacts / day		Yes	Yes	more accurate estimation of population Malawi
A13	DSA	to understand the distribution of CDR data across SMS and voice	I can differentiate between SMS and voice data in CDR	Yes	# of SMS / day and # of voice calls / day		Yes	Yes	better understanding on the distribution data from visual representation
A14	DSA	to understand SMS and voice data distribution in CDR data as per town or district	I can get an overview of # of CDR data as per town or district	Yes	# of SMS / district and # of voice calls / district		Yes	Yes	better understanding on the distribution data from visual representation
A15	DSA	to understand daily CDR data distribution	I can get an overview of day wise CDR data distribution as per town or district	Yes	# of calls / day for each of the town or district		Yes	Yes	better understanding on the distribution data from visual representation
A16	DSA	to understand day wise unique subscriber's distribution	I can get an overview of day wise CDR data distribution as per unique subscribers	Yes	# of calls from unique numbers / day for each of the town or district		Yes	Yes	better understanding on the distribution data from visual representation
A17	DSA	to understand total no of visited location daywise for each unique subscriber	I can get an understanding about the individual mobility pattern	Yes	# of unique location visited / day for each of the unique subscriber		Yes	Yes	better understanding on the distribution data from visual representation
A18	DSA	to understand the CDR data distribution across a day	I can get the overview on CDR data distribution across Time-of-the-day	Yes	# of calls at specific time points / day		Yes	Yes	better understanding on the distribution data from visual representation
A19	DSA	to understand cell towers distribution	I can get an overview of the distribution of cell towers as per town or district	Yes	# of cell towers / town or district		Yes	Yes	better understanding on the distribution data from visual representation
A20	DSA		I can get the overview of the distribution of cell ids						better understanding on the distribution



Metrics-KPIs for  
CDR

Infosys<sup>®</sup>  
Navigate your next

# Anonymized CDR data analysis

Type of Analysis on CDR data	Insights generation	Level of data aggregation	Relevance to usecase
Identification of application and human related CDR data	Segregation of application and human related CDR data	Raw CDR data	Yes
Distribution of CDR data as per SMS/Voice	SMS and voice data distribution in CDR data	Raw CDR data	Yes
Overview of # of CDR data as per Towns / Districts	SMS and voice data distribution in CDR data as per town /district	Raw CDR data	Yes
Day-wise CDR data distribution as per Towns / Districts	Daily CDR data distribution	Day level aggregation	Yes
Day-wise unique sender Id as per Towns / Districts	Day wise unique senders distribution	Day level aggregation	Yes
Average total unique location (daily)	Total unique location visited day-wise	Day level aggregation	Yes
Unique originating number of CDR from Districts	Details of unique originating location	Day level aggregation	Yes
Time of the day -wise CDR data distribution as per Towns/Districts	CDR data distribution as per time of the day	Hourly level aggregation	Yes
Distribution of Cell Towers as per Towns / Districts	Cell towers distribution	Day level aggregation	Yes
Burden of cell id across Towns / Districts	Details on Burden of cell	Day level aggregation	Yes
Busiest cell id across Towns / Districts	Top most busiest cell id	Day level aggregation	Yes
Distribution of Health Facilities and Towers	Health Facilities locations coverage by Towers	Day level aggregation	Yes

# Anonymized CDR data analysis (contd...)

Type of Analysis on CDR data	Insights generation	Level of data aggregation	Relevance to usecase
Distribution of Health Facilities and Towers wrt administrative areas	Health Facilities locations coverage by Towers	Day level aggregation	Yes
Location of cell Tower at TA/GVH level	Location of cell Tower at TA/GHV level	Day level aggregation	Yes
Most Visited Locations in the highest populated Administrative Unit	Most visited locations at administrative unit	Day level aggregation	Yes
Trip distance distribution	Trip distance pattern	Day level aggregation	Yes
Radius of gyration	Quantifying human mobility patterns	Day level aggregation	Yes
# of visited location distribution	Visited location pattern	Day level aggregation	Yes
Frequently visited locations	Frequently visited location pattern	Individual level aggregation	Yes
# of re-visited location distribution	Re-visited location pattern	Day level aggregation	Yes
Average time spent distribution (min)	Time spent pattern for voice	Individual level aggregation	Yes
Proportion of home interactions	Proportion of home interaction with other interactions	Individual level aggregation	Yes
Churn Rate	Provides insight into the growth or decline of the subscriber base	Individual level aggregation	Yes

# Long Term Population Movement



# Use Case: Population Migration – Approach

This approach proposes a mobile CDR based approach (mobile phone MP) to obtain population migration of a given tower level.

Approach- Three steps approach is suggested for Population Migration

**Step 1-** Aggregation of CDR data @ month level as per tower and unique users.

**Step 2-** Understand population movement by netflows and proportion of inflow/outflow

**Step 3-** Inference on change in user density and compute population migration

<b>Inputs</b>	<ul style="list-style-type: none"><li>• Anonymized CDR data</li><li>• Aggregation rules for CDR data @ day level</li></ul>	<ul style="list-style-type: none"><li>• Aggregated CDR data at month level as per tower and unique user</li></ul>	<ul style="list-style-type: none"><li>• Details of change in unique subscriber list at tower level.</li><li>• Proportion of Inflow/Outflow information month wise for each district.</li></ul>
<b>Key Activities</b>	<ul style="list-style-type: none"><li>• Identify A2P and P2P SMS</li><li>• Create script to do aggregation of CDR data at month level for each tower and each unique user</li><li>• Perform aggregation of CDR data at month level for towers and unique users</li></ul>	<ul style="list-style-type: none"><li>• Identify new and inactive unique subscriber count at each tower. Calculate netflow change of population at each tower</li><li>• Calculate proportion of inflow/outflow at higher ADM levels for each month</li></ul>	<ul style="list-style-type: none"><li>• Identify administrative units with floating population movements.</li><li>• Scaling user density at tower over months to population density</li><li>• Identify locations with consistent inflow and outflow of populations</li></ul>
<b>Deliverables</b>	<ul style="list-style-type: none"><li>• Aggregated CDR data at month level as per tower and unique users</li></ul>	<ul style="list-style-type: none"><li>• Details of change in unique subscriber list at tower level.</li><li>• Proportion of Inflow/Outflow information month wise for each district.</li></ul>	<ul style="list-style-type: none"><li>• Locations with favorable conditions to stay (using proportion of inflow) and difficult to stay (using proportion of outflow) and floating population (using inflow and outflow)</li><li>• Population density of units month wise</li></ul>

# Proportion of Inflow and Outflow based population movement analysis

## Rationale:-

- A higher proportion of inflow month over month at a district/TA may be due to favorable living conditions at that location.
- A higher proportion of outflow month over month at a district/TA may indicate not so favorable living conditions at that location.
- A high proportion of inflow and outflow indicate floating population movement.

## Approach:-

New connections and inactive numbers are identified while computing inflow/outflow proportion.



## Terms:-

Proportion of inflow:- For a given month “m” and district “d”, it is the ratio of (population moved into district “d” from other districts in month “m”)/(total population outside district “d” in month “m-1”)

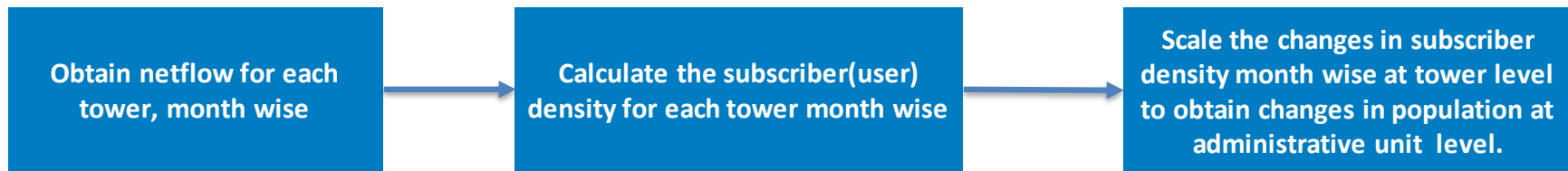
Proportion of outflow:- For a given month “m” and district “d”, it is the ratio of (population moved out of district “d” to other districts in month “m”)/(total population in district “d” in month “m-1”)

# Population migration at tower/Administrative unit level

**Rationale:-** To understand the population migration at administrative units over long term

## **Approach:-**

- Net flow(outflow-inflow) of users is calculated for each tower month wise.
- The observed net flow change over each month is used to obtain the change in subscriber density for each tower month wise.
- Change in subscriber density is used as proxy to sense the change in population over months at the tower level.
- Changes in subscriber density is used to obtain base population estimates at administrative unit levels.
- This is scaled appropriately for each administrative unit to obtain population estimates for the unit.



## **Terms:-**

“Netflow” for month “m” for tower “j” is defined as (# of subscribers in month “m-1” + (# of subscribers inflow from other regions for month “m”) - (# of subscribers outflow from tower “j” to other locations in month “m”).

Subscriber density- (Number of subscribers)/ (Area of polygon for tower)

# Assumptions

# Assumptions

- In this approach it is assumed that all cell towers are similar\* wrt transmission power, antenna heights, elevation, structural and environmental characteristics. This ensures that the approach is less computationally intensive.
- For computational convenience 7 pm to 7 am is assumed as the night time and from 9 am to 6 pm its is assumed as the working hours or the most productive hours of the day for an individual.
- A phone user's home location is identified as the most frequently communicated tower during nights of weekdays and weekends over the study period.
- Each of the unique subscribers/users comes back to their base/home location at the end of the day.
- Approximately 85% of population will be covered if mobile tower range is 10 km.

# Appendix

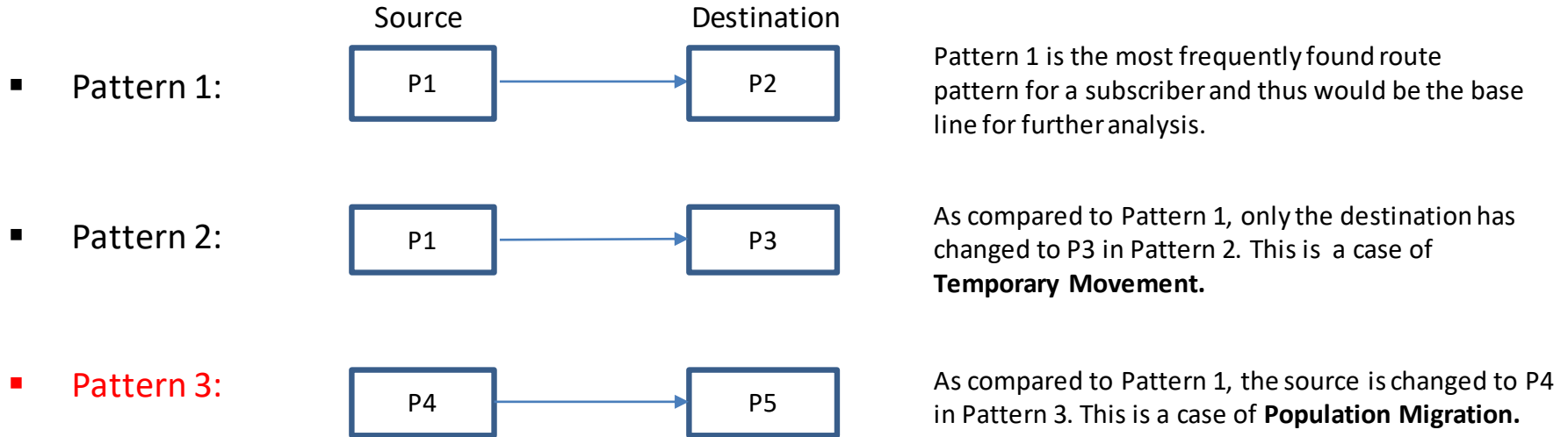
# Dynamic Population Movement – Methods

- **Route Pattern Analysis** - Track movements of people through a given area at a particular time, through home locations, stay points, movement routes and timing of mobile phone users.
- **Human Mobility Motifs** - Analyze the daily movements using the Mobility Motifs paradigm (ref: Schneider et al).
- **Individual Mobility Patterns** - Explore the statistical properties of the population's mobility patterns using a truncated Levy flight model (ref: Barabasi et al).
- **Analysis Based on Rank of Locations** - Analyze the daily movements on the basis of the # of times an individual visited to a place.
- **Analysis Based on Radius of Gyration** - Characterizing human mobility patterns and human daily travel with periodicity and regularity, the radius of gyration for an individual can be predicted and this can be used to track their future movement path.

<http://humnet.scripts.mit.edu/wordpress2/wp-content/uploads/2010/10/J.-R.-Soc.-Interface-2013-Schneider-.pdf>  
<http://barabasi.com/f/250.pdf>

# Method - Route Pattern Analysis

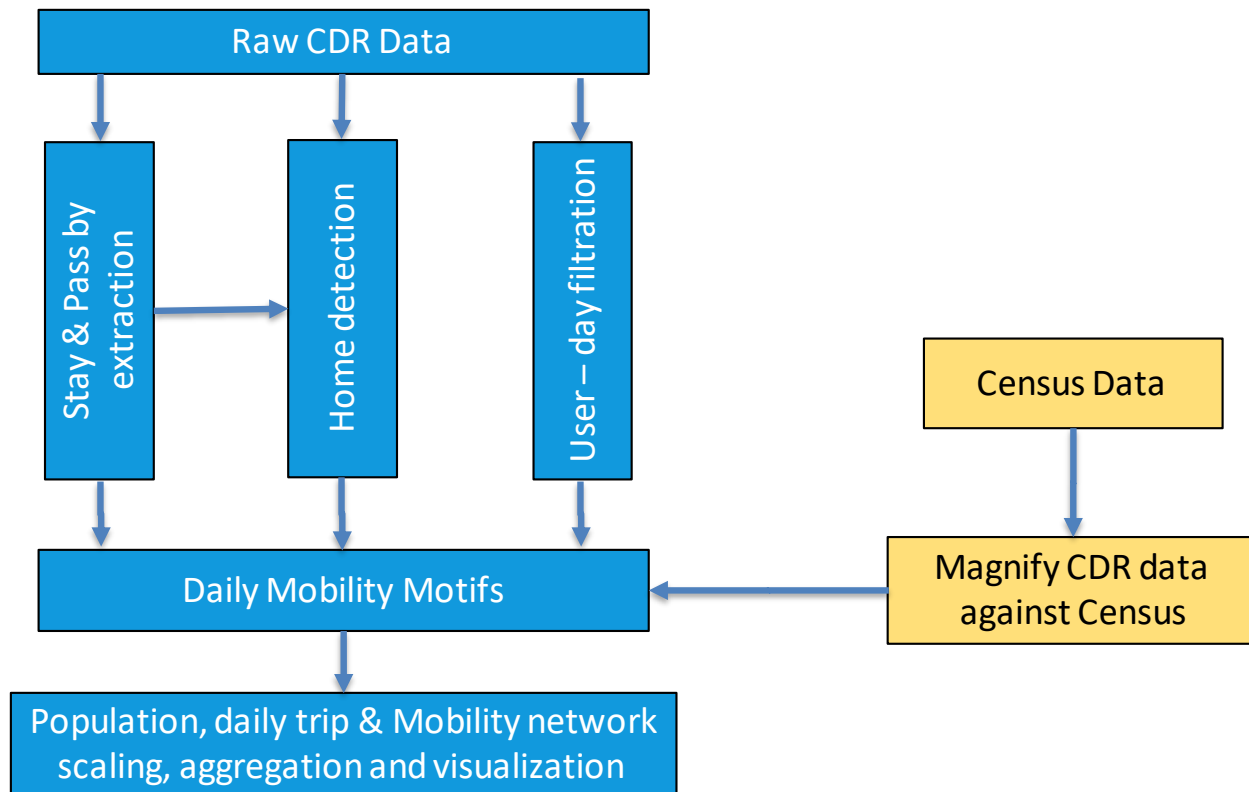
- Identify the frequent travelled routes for each subscriber based on a pre-defined threshold for # of occurrences.
- The analysis would be based on the frequent route patterns as follows:



All these analysis can be done using at least 1 month of CDR data. For these patterns, the aggregation of CDR data will take place at day level for a subscriber.



# Method –Human Mobility Motifs and Individual Mobility Patterns



Legends

Population Estimation & Mobility Pattern step

Same step as Population Density Mapping

Daily Mobility Motifs together with magnification factor can be used to estimate dynamic population at any location for a given time frame.

# Method –Population Estimation and Mobility Pattern (contd...)

- **Stay & Pass by extraction:** Parsing the mobile phone CDR data to extract anchor points (i.e., stay locations) of individual daily travel and to differentiate stops from pass-by points is the first stage before identifying individual mobility networks.
- **Home detection:** An individual always departs from home and returns home by the end of a day (assumption), for planning purpose, it is important to understand individuals' mobility patterns from their home and a phone user's home tower is identified as the most frequently communicated tower during nights of weekdays, and weekends over the study period.
- **User – day filtration:** It's important to sample user-days when the mobile device is used frequently enough. Separate observations on weekdays from those on weekends, as mobility patterns can be different. Here we only focus on records on weekdays.
- **Daily Mobility Motifs:** With the extracted stay locations, users' identified home location, we will be able to identify the mobility networks. It has been observed that not all users necessarily have the first stay-point (or potential stay-point) starting from their home tower, or the last stay point (or potential stay-point) ending at their home tower every day in the filtered trajectory, due to the passive nature of the CDR data. In order to form a complete tour for each user, we make the assumption that, if a user travels in a day, s/he starts the first trip from, and ends the last trip at home.
- **Population, daily trip & Mobility network scaling, aggregation and visualization:** After filtering the user-day samples, it is important to verify that users with more phone usage events do not have systematic differences in travel behavior. Therefore, we examine the relationship between the filtered users' cell phone usage patterns and their daily travel patterns.
- **Magnify CDR data against Census:** Reallocated points are aggregated by each census of population census. The magnification factor is the number of people a user represents and can be computed by comparing real population of a census zone with the aggregated number of mobile phone users.

# Method - Analysis Based on Rank of Locations

- For all the districts we can find out the rank  $L$  when each location is ranked on the basis of the number of times an individual was recorded in its vicinity.
- Theoretically it has been found out that people devote most of their time to a few locations, although spending their remaining time in other places, visited with diminished regularity.
- The same kind of analysis can be done on Malawi and this might be helpful in finding out the population movement analysis.

# Method - Analysis Based on Radius of Gyration

Gyration radius of individual's trajectory plays a key role in characterizing human mobility pattern. Because of the importance for quantitatively understanding human mobility patterns, it's essential to find out the factors responsible for the gyration radius of trajectories. Different kind of analysis can be done on radius of gyration:

- As the radius of gyration (RG) has a strong impact on travel distance distribution over all users, we can do a graphical comparison of average rg across all the districts of Malawi to obtain meaningful insights regarding the travel pattern of individuals in those districts
- Analyze the variability of RG with respect to the distance between an individual's home and work location (we can assume the most frequent pattern from the route pattern analysis as the route between an individual's home and work location if we consider the data only for weekdays)
- Analyze the variability of RG with respect to maximum daily travel distance of an individual

# CDR Data Aggregation – High Level Entity Model

Aggregated Entity	Key Processing Steps	Features Extracted
Daily_CDR_Summary	Get_Daily_CDR_Summary	User_Id, date, month, year, # of calls, # of towers, max_used_tower, max_used_tower_callcount
Daily_Tower_Summary	Get_Daily_Tower_Summary	Tower_Id, date, month, year, Hourly_Traffic, Total # calls, Total # of unique callers
Monthly_Caller_Summary & User_Master	Update_Master_User* Filter_A2P_CDR Map_User_Home_Tower	User_Id, Month, Year, Home Tower, Avg Daily Calls. A2P_Flag (update in User_Master)
Monthly_Tower_Summary & Tower_Master	Update_Master_Tower_List Detect_Home_Tower_Changes Calculate_Monthly_Flow	Tower_Id, Month, Year, Average_Daily_Calls (by event type) # home users # visitors Monthly_Ingres (detect change in home tower) Monthly_Egres (detect change in home tower) Avg_Hourly_Traffic (23 columns)

\* This can be reused from Caller\_Anonymization

# What is Dynamic Population Movement ?

**Population dynamics** is the study of how and **why populations** change in size and structure overtime. **Important** factors in **population dynamics** include rates of reproduction, death and migration.

## Characteristics of Dynamic Population Movement:

- It has been found that in contrast with the random trajectories predicted by the prevailing Le´vy flight and random walk models, human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time independent characteristic travel distance and a significant probability to return to a few highly frequented locations.
- After correcting for differences in travel distances and the inherent anisotropy of each trajectory, the individual travel patterns collapse into a single spatial probability distribution, indicating that, despite the diversity of their travel history, humans follow simple reproducible patterns.
- This inherent similarity in travel patterns could impact all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and agent-based modelling.

# Approach – Human Mobility Pattern study

- **Trip distance distribution (Levy flight form)**

$$p(r) = r^{-\beta}; \beta = 1.59; \quad r : \text{distance}$$

- **Radius of gyration**

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r}_i^a - \vec{r}_{cm}^a)^2}, \quad \text{where } \vec{r}_i^a \text{ represents the } i = 1, \dots, n_c^a(t) \text{ positions recorded for user } a$$

and  $\vec{r}_{cm}^a = 1/n_c^a(t) \sum_{i=1}^{n_c^a} \vec{r}_i^a$  is the center of mass of the trajectory.

- **# of visited location distribution**

$$f(N) = e^{-(\ln N - \mu)^2 / (2\sigma^2)} / (N\sigma(2\pi)^{1/2})$$

$\mu = 1, \sigma = 0.5$

N : # of locations

- **# of re-visited location distribution**

$$s(t) = t^{\mu}; \mu = 0.6; t = \text{time}$$

# Individual Trajectory Based Population Movement Analysis

## Approach:



## Terms:

- **Trajectory data:** Is a sequence of time-stamped points,  $P = (p_1, p_2, \dots, p_n)$ , where  $p = (\text{ID}, \text{time}, \text{latitude}, \text{longitude})$  and  $n = \text{a total number of points}$ .
- **Stay point:** This is a geographical reference to a place where a user stayed over a time threshold ( $tt$ ) within a distance threshold ( $dt$ ). In a trajectory, stay point is characterized by a set of consecutive points  $P = \{p_m, p_{m+1}, \dots, p_n\}$ , where  $\forall m < i \leq n, \text{Distance}(p_m, p_i) \leq dt, \text{Distance}(p_m, p_{n+1}) > dt$  and  $\text{Time Interval}(p_m, p_n) \geq tt$ . Therefore,  $s = (x, y, ta, tl)$ , where  $x, y$  are a centroid location of those points in a stay point.



# Population density- Notes on availability of data

Dataset	Availability
World pop	2010, 2015, 2020 , 2008 (Census)
CDR dataset	2016,2017

- World pop data is available for the years 2010, 2015 and 2020, while CDR dataset is available for 2016 and 2017.
- World pop used 2008 census for Malawi and projected it for the years 2010, 2015 and 2020 using an exponential function.
- Using the 2015 worldpop adjusted file, we obtain population estimates for the year 2015, but it doesn't match with the available data of CDR, which is available only for 2016, 2017.
- **How to overcome this data interval availability issue**
- Using the exponential growth approach from world pop and referred paper\* obtain population estimates for 2016 and 2017. Check whether the estimates are inline with the 2020 world pop estimates.
- Using CDR for 2016, 2017 and worldpop estimates for the same time period from previous step, obtain a relation between cell phone mobile user density and population estimates. (Estimate alpha and beta).
- For the year 2018, since CDR data would be unavailable, capture mobile penetration in Malawi or similar countries (through research articles or information) and use it to extrapolate the cell phone usage in 2018.
- Use the extrapolated CDR information for 2018 to compute population density estimates for the year 2018. Check if the obtained estimates are in line with world pop 2020 estimates.

# DIAL- Justification for Cell-Tower Coverage Radius

The working range of a cell site depends on a [number of factors](#) including

- Height of antenna over surrounding terrain, establishing line of sight
- The [frequency](#) of signal in use,
- The transmitter's rated power.
- Reflection and absorption of radio energy by buildings or vegetation.

In urban areas, masts are built close to each other to handle the heavier demand in terms of users, [1-2 km apart](#). In rural areas, the *maximum* range of a mast depends on technical considerations, notably the ability of a low-powered personal cell phone to transmit back to the mast. Under ideal conditions (tall mast and flat terrain), a GSM signal can technically travel [up to 70 km](#). Under more realistic conditions, with hilly terrain and potential obstacles, the maximum distance can be [as little as 5 km](#).

To capture these considerations, the Cooper/Smith gap analysis allows for different radii, with a minimum of 5km in rural areas to reflect worse case conditions, 10km as the median value of expected range, and 20km as a realistic upper bound.