NAVIGATE YOUR NEXT

Relationship b/w Population density and Call density in Malawi Ver-1 0

December 13, 2018

Infosys®
Navigate your next

Contents

- Objective
- Data Preparation
- Assumptions and implementation
- Results and interpretations
- Appendix
  - Complete Approach
  - Spatial weight matrix -Reference

Objective

- Goal of this step is to explore the relation of Call Details Records (CDR) from Airtel Malawi with population data obtained from Worldpop in Malawi.

➢ Motivation

- Computing real-time population density using CDR provides an efficient way to track human movement. This computation will be useful especially during disease outbreaks or natural disasters.

➢ Methodology

- A research paper titled “Dynamic population mapping using mobile phone data, PNAS early edition (2014)”* was studied to develop the baseline approach. Any changes relevant to Malawi were suitably made and is highlighted.

*-Reference :- http://www.pnas.org/content/111/45/15888

Data Preparation

Input Datasets

Dataset	Year
CDR dataset- (aggregated at day level for cell id)	2016
Tower Master	Towers present in 2016
Worldpop ppp 2015 adjusted V2 (interpolated for 2016)	2015
Administrative files	-

Approach to consider aggregated CDR dataset :- Average of sum of number of unique originating ids at each cell id at day level for the complete year. This approach is considered based on the following rationale

- Averaging the number of originating ids gives a representative number of call density at each cell tower for that year.
- Alpha and beta values obtained in a day level data is precisely useful when the model is used during any outbreaks.

Input Data Summary

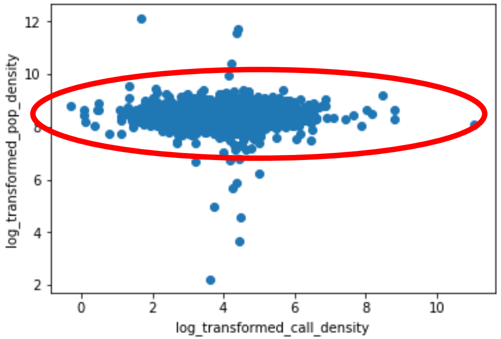
Descriptions	Values
# of ADM-2 units without towers in 10km boundary	5
# of unique cell ids	3,925
# of active subscribers	4,395,880
# data missing	APRIL, MAY

Assumptions

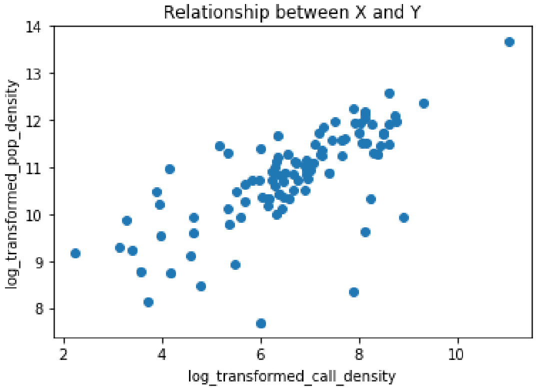
- All towers are assumed to have similar characteristic and have similar maximum coverage range (10km). Which might not be the case in reality.
- A2P numbers are considered as similar to P2P. This might impact the call density calculation.
- 611 towers in CDR dataset have missing location information. This might impact the call density calculation.
- 2015-PPP-ADJ v2 file interpolated for 2016 is used as a reference for 2016 population density since we don't have a ready reference for 2016 population density. From this method, obtained population density may not be actual 2016 population density.

Implementation

- Model could be implemented at ADM-3 or ADM-2 level.
- Relationship between log transformed population density and call density were studied at both the administrative level, as illustrated below. Below graphs refer to South region for the year 2016



ADM-3



ADM-2

- Similar results were obtained for North and Central regions as well.
- Since ADM-2 depicts a linear relationship between the variables we proceed with ADM-2 for our analysis.
- Note: - For analysis, distance based weight matrix is considered since polygon sizes are uneven.

Results

- Year: 2016
- Region: South
- Weight Matrix: Distance based. Fixed Distance (Threshold based)
- Total administrative units: 107

Test	Log Alpha, Beta	R^2	P-Value (F-statistic P-Value)	Remarks
OLS	(4.5201145, 0.6242424)	0.2748	6.815e-09	OLS is significant. Proceed for Moran I's test
Spatial correlation (Moran's I) - Distance Based		-	0.10820	Not significant
Model accounted for spatial variation				Spatial correlation not observed using distance based weight matrix

Log(population density) = 4.5201145 + 0.6242424 * Log(Call density)

- For the test, run using distance based matrix on entire south region, we found Moran's I to be not significant.

Results contd...

- Year: 2016
- Region: Central
- Weight Matrix: Distance based. Fixed distance (Threshold based)
- Total administrative units: 95

Test	Log alpha, beta, Lambda	R^2	P-Value (F-statistic P-Value)	Remarks
OLS	3.3048248, 0.9145516	0.3701	6.185e-11	Proceed for Moran I's test
Spatial correlation (Moran's I) - Border Based			0.0210	Significant. Proceed with LM tests
Model accounted for spatial variation- Spatial error model	3.43561, 0.8599301, 0.492733	0.3701	Lambda (spatial error coeff p-value) =0.0046798	Coefficient is significant.

Log(population density) = 3.43561 + 0.8599301 *Log(Call density) - 0.492733 * W * ἐ, where, W- weight matrix and ἐ is vector of error terms .

- For the test, run using distance based matrix on entire central region, we found Moran's I to be significant.
- To account for spatial variation, among lag and error model, we found error model to be significant.
- So a spatial error model was developed for central region.

Results contd...

- Year: 2016
- Region: North
- Weight Matrix: Distance based, Fixed distance (Threshold based)
- Total administrative units: 48

Test	Log Alpha, Beta	R^2	P-Value (F-statistic P-Value)	Remarks
OLS	(4.0962260, 0.5951148)	0.2533	0.0002661	OLS is significant. Proceed for Moran I's test
Spatial correlation (Moran's I) - Distance Based			0.3440	Not significant
Model accounted for spatial variation				Spatial correlation observed using border based weight matrix

Log(population density) = 4.0962260 + 0.5951148 * Log(Call density)

- For the test, run using distance based matrix on entire north region, we found Moran's I to be not significant.

Data Preparation

Input Datasets

Dataset	Year
CDR dataset- (aggregated at day level for cell id)	2017
Tower Master	Towers present in 2017
Worldpop ppp 2015 adjusted V2 (interpolated for 2017)	2015
Administrative files	-

Approach to consider aggregated CDR dataset :- Average of sum of number of unique originating ids at each cell id at day level for the complete year. This approach is considered based on the following rationale

- Averaging the number of originating ids gives a representative number of call density at each cell tower for that year.
- Alpha and beta values obtained in a day level data is precisely useful when the model is used during any outbreaks.

Input Data Summary

Descriptions	Values
# of ADM-2 units without towers in 10km boundary	5
# of unique cell ids	6,721
# of active subscribers	5,740,154
# data missing	0

Results

- Year: 2017
- Region: South
- Weight Matrix: Distance based, Fixed distance (Threshold based)
- Total administrative units: 107

Test	Log Alpha, Beta	R^2	P-Value (F-statistic P-Value)	Remarks
OLS	(4.5120842, 0.5967817)	0.2529	3.388e-08	OLS is significant. Proceed for Moran I's test
Spatial correlation (Moran's I) - Distance Based			0.19659	Not significant
Model accounted for spatial variation				Spatial correlation not observed

Log(population density) = 4.5120842 + 0.5967817 * Log(Call density)

- For the test, run using distance based matrix on entire south region, we found Moran's I to be not significant.

Results contd...

- Year: 2017
- Region: Central
- Weight Matrix: Distance based, Fixed distance (Threshold based)
- Total administrative units: 95

Test	Log alpha, beta, Lambda	R^2	P-Value (F-statistic P-Value)	Remarks
OLS	3.1558757, 0.9425201	0.3707	5.925e-11	Proceed for Moran I's test
Spatial correlation (Moran's I) - Distance based			0.0261	Significant. Proceed with LM tests
Model accounted for spatial variation- Spatial error model	3.2885209, 0.88976,(-0.46802)	0.3707	Lambda (spatial error coeff p-value) =0.0349	Coefficient is significant.

Log(population density) = 3.2885209 + 0.88976 * Log(Call density) - 0.46802 * W * ἑ, where, W- weight matrix and ἑ is vector of error terms .

- For the test, run using distance based matrix on entire central region, we found Moran's I to be significant.
- To account for spatial variation, among lag and error model, we found error model to be significant.
- So a spatial error model was developed for central region.

Results contd...

- Year: 2017
- Region: North
- Weight Matrix: Distance based, Fixed distance (Threshold based)
- Total administrative units: 48

Test	Log Alpha, Beta	R^2,AIC	P-Value (F-statistic P-Value)	Remarks
OLS	(4.0101874, 0.5973179)	0.2575	0.0002321	OLS is significant. Proceed for Moran I's test
Spatial correlation (Moran's I) - Distance Based			0.36822	Not significant
Model accounted for spatial variation				Spatial correlation observed using border based weight matrix

Log(population density) = 4.0101874 + 0.5973179 * Log(Call density)

- For the test, run using distance based matrix on entire north region, we found Moran's I to be not significant.

Cross Validation-Results

- Cross validation is performed to assess the predictive performance of models and evaluate performance on test data.
- In our analysis we have used “leave one out” variant of k-fold cross validation

Region	Year	Method	Log (Alpha)	Beta	RMSE_test	RMSE_train	R2	Moran's I (P-value)
South	2016	random	4.52028	0.62426	4.11040	1.20794	0.30722	0.1082
		spatial-kfold	4.53637	0.64177	0.06418	1.23083	0.30626	
	2017	random	4.51978	0.61636	4.12308	1.22765	0.28443	0.1966
		spatial-kfold	4.53533	0.59581	0.09486	1.23869	0.26458	
Central	2016	random	3.48888	0.86640	7.40930	0.84458	0.48703	**0.0210 < 0.05**
		spatial-kfold	3.49828	0.87670	0.03492	0.83527	0.49870	
	2017	random	3.33914	0.89744	7.45335	0.83958	0.49309	**0.0261 < 0.05**
		spatial-kfold	3.34600	0.90931	0.11460	0.82809	0.50728	
North	2016	random	4.17027	0.64038	4.63286	1.20493	0.33710	0.3440
		spatial-kfold	4.21551	0.62061	0.02514	1.22275	0.34567	
	2017	random	4.07595	0.64226	4.63517	1.20031	0.34218	0.3682
		spatial-kfold	4.11655	0.62278	0.07922	1.21894	0.34974	

- The estimates for each of the methods (random, spatial k-fold) within each region are consistent across years . We proceed with “2016-random” estimates for each region to compute population densities.
- **Reference:-Spatial K-fold CV:- Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen & Jukka Heikkonen (2017) Estimating the prediction performance of spatial models via spatial k-fold cross validation, International Journal of Geographical Information Science, 31:10, 2001-2019**

Alpha, Beta and Power Equations

- The estimates for each of the methods (random, spatial k-fold) within each region are consistent across years. We proceed with “2016-random” estimates for each region to compute population densities.
- Region wise Alpha, Beta and Power Equation can be taken from below table

Region	Log (Alpha)	Beta	Power Equation
South	4.52028234	0.624258384	Log(population density) = 4.52028234 + 0.624258384 * Log(Call density)
Central	3.488878963	0.866401779	Log(population density) = 3.488878963 + 0.866401779 * Log(Call density)
North	4.170267484	0.640379476	Log(population density) = 4.170267484 + 0.640379476 * Log(Call density)

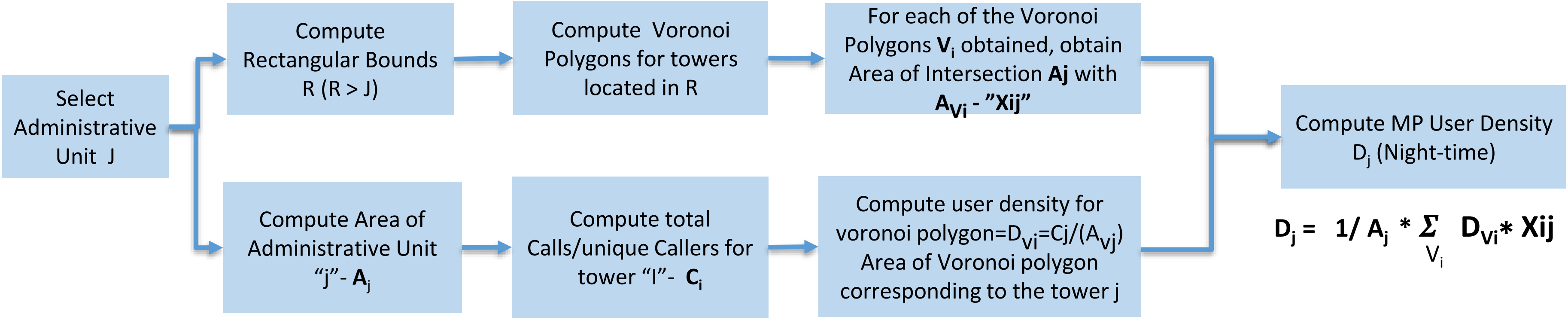
Appendix

Infosys®
Navigate your next

Population Density Mapping - Approach

Step 1 – Compute MP (mobile phone) User Density for each Administrative Unit

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



Select Administrative Unit J

Compute Rectangular Bounds R (R > J)

Compute Voronoi Polygons for towers located in R

For each of the Voronoi Polygons $\mathbf{V_i}$ obtained, obtain Area of Intersection **Aj** with $\mathbf{A_{Vi}}$ **- ”Xij”**

Compute Area of Administrative Unit “j”- $\mathbf{A_j}$

Compute total Calls/unique Callers for tower “I”- $\mathbf{C_i}$

Compute user density for voronoi polygon=D_{Vi}=Cj/(A_{Vj}) Area of Voronoi polygon corresponding to the tower j

Compute MP User Density D_j (Night-time)

$$\mathbf{D_j = 1/ A_j * \sum_{V_i} D_{Vi} * Xij}$$

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”

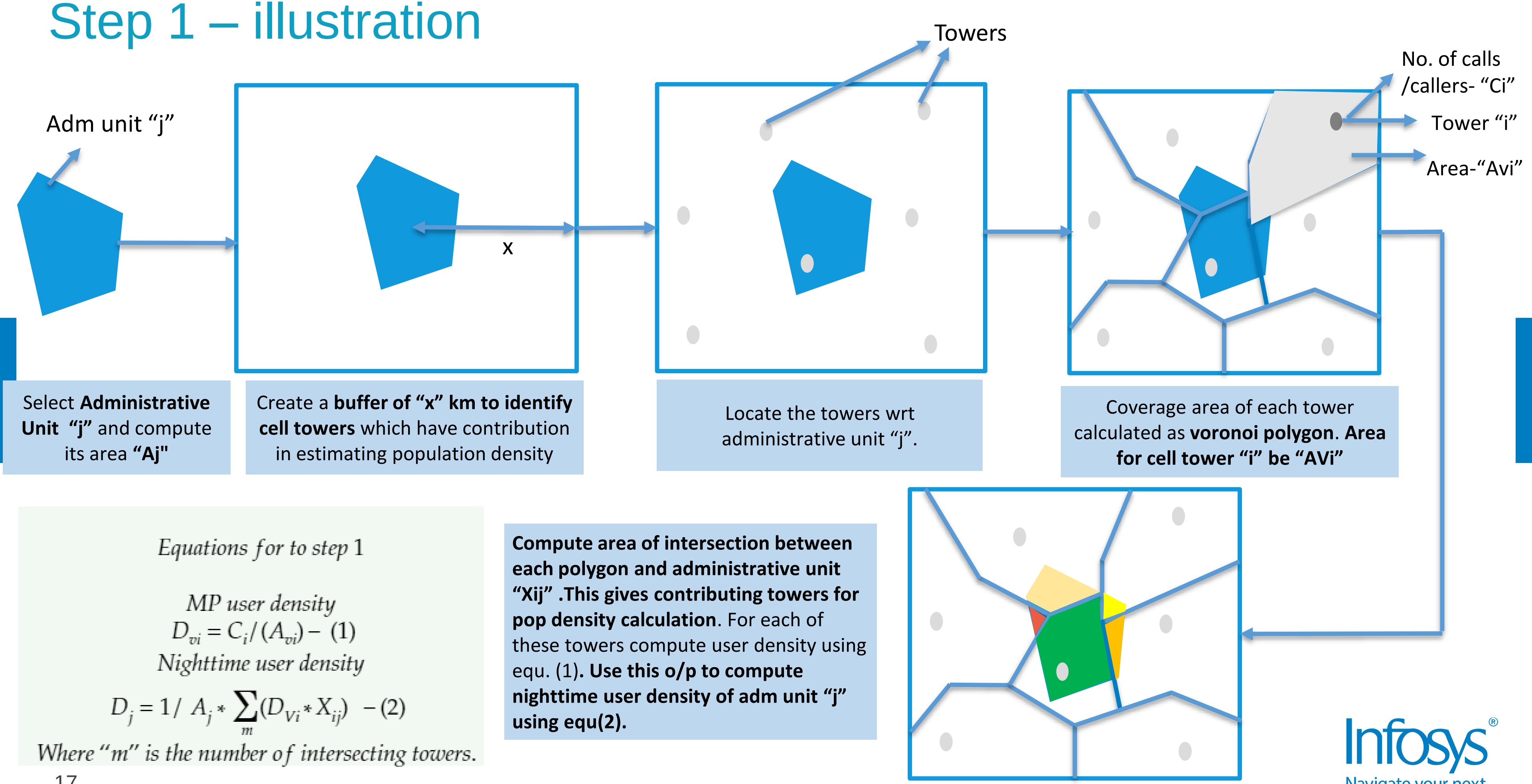
Vi- Voronoi polygon obtained for tower “i”

Avi- Area of voronoi polygon Vi

Xij- Area of intersection between area Aj and Avi

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

Step 1 – illustration



Select **Administrative Unit "j"** and compute its area **"Aj"**

Create a **buffer of "x" km to identify cell towers** which have contribution in estimating population density

Locate the towers wrt administrative unit "j".

Coverage area of each tower calculated as **voronoi polygon**. **Area for cell tower "i" be "AVi"**

Compute area of intersection between each polygon and administrative unit "Xij" .This gives contributing towers for pop density calculation. For each of these towers compute user density using equ. (1)**. Use this o/p to compute nighttime user density of adm unit "j" using equ(2).**

Equations for to step 1

MP user density

$$D_{vi} = C_i / (A_{vi}) - (1)$$

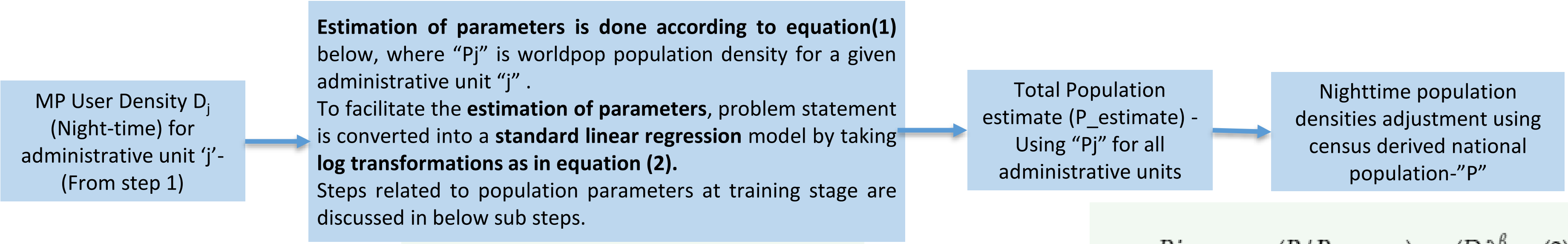
Nighttime user density

$$D_j = 1/\ A_j * \sum_{m} (D_{Vi} * X_{ij}) \ - (2)$$

Where "m" is the number of intersecting towers.

Population Density Mapping - Approach (contd..)

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



MP User Density D_j (Night-time) for administrative unit ‘j’- (From step 1)

Estimation of parameters is done according to equation(1) below, where “Pj” is worldpop population density for a given administrative unit “j” .
To facilitate the **estimation of parameters**, problem statement is converted into a **standard linear regression** model by taking **log transformations as in equation (2).**
Steps related to population parameters at training stage are discussed in below sub steps.

Total Population estimate (P_estimate) - Using “Pj” for all administrative units

Nighttime population densities adjustment using census derived national population-”P”

$$Pj_{adjusted} = (P/P_{(estimate)}) * \alpha * (Dj)^{\beta} \quad -(3)$$

$$Pj = \alpha * (Dj)^{\beta} \quad -(1)$$

$$Log(Pj) = Log(\alpha) + (\beta * log(Dj)) \quad -(2)$$

Parameters

α : Scale Ratio

β : Super linear effect of population density (census) on night-time user density

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 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.
- In case of Malawi, to estimate population density, 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 2015 in **Worldpop ppp adj_V2** file is 17303306.73, which is very close to the value indicated in **Facebook HRSL file** (17201196.95) .
- **Worldpop 2015 ppp V2 unadjusted** has total population count of 15799142.22 .

Spatial weight Matrix

- Population density is considered to be continuous variable. (Reference:- https://gis.stackexchange.com/questions/39196/is-population-density-considered-continuous-data-and-why)
- Inverse distance:- Most appropriate for continuous data. For large datasets it is very computationally intensive. We need to setup a threshold to control the mapping of number of neighbors. This threshold is subjective.
- Polygon Contiguity- is effective when polygons are similar in size and distribution. In our case this is not applicable since polygon area and population density are skewed at ADM-2 level. This is the easiest and computationally less intensive weight matrix.
- Zone of indifference:-Fixed and inverse distance concept used together. Not suitable for large datasets. Will not try. Complex and may not be accurate since we need to decide on the values of Fixed distance threshold and inverse distance rate.
- KNN:-When features are highly skewed we evaluate features with eight of its neighbors (rule of thumb). In our case, population density at ADM-2 level is skewed. We may try this approach since it is not computationally intensive and readily available in PYSAL. For KNN threshold fixed distance keeps varying based on the administrative unit.
- Fixed distance:-Useful when polygon sizes are uneven. Here we need to provide a fixed cut-off which is subjective. In our case polygons are unevenly sized at ADM-2 level. We can decide the fixed distance threshold such that each unit has at least 1 neighbor.

- Reference for the approach :- http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/spatial_statistics_toolbox/modeling_spatial_relationships.htm
- Reference paper:- https://hal.archives-ouvertes.fr/hal-01544823/document

THANK YOU

Infosys®
Navigate your next