The Indian Geographical Journal
Volume 96 (2) December - 2021, pp 01-19
ISSN 0019-4824

UTILITY OF GRIDDED POPULATION DATASETS IN AGGREGATING CENSUS DATA FOR NON-ADMINISTRATIVE GEOGRAPHIES OF INDIA

Lazar. A[1], N. Chandrayudu[2]

[1]Office of the Registrar General India, New Delhi
[1& 2]Dept. of Geography, College of Sciences, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.

lazar.rgi@gov.in

Abstract

Gridded datasets are very valuable for a wide variety of decision models in environmental, health, and climate change research. As the variables affecting the environment, health, and climate are contiguous, the census datasets published at administrative geographies often may not be usable and require recompilation. The open-access raster datasets of gridded population layers provide ample scope for aggregating census counts for non-administrative geographies. The existing gridded population datasets are prepared at global scale and suitable for national or regional scales, however, downscaling them for smaller non-administrative geographies is challenging and produce less accurate population counts. Thus, an attempt was made to analyse the utility of existing gridded population layers with census counts at different administrative levels (entire India, Karnataka State and Mysore city). Further, the fitness of use for different scales is also analysed. It is observed that the creation of gridded population layers using village/town level census counts along with the covariates improve the accuracy. The publication of census data as a gridded raster layer would greatly help researchers and planners to study the non-administrative geographies of India.

Keywords: Census data, Population grids, Areal weighting, Non-administrative geographies

Introduction

Humans are the most influential factor on earth's planet and will remain so (Palumbi, 2001). The carrying capacity of planet earth is likely to be challenged by the increasing population size, and these challenges pose a more significant threat to ecological balances (Hendry et al., 2017). It is estimated that the interaction of anthropogenic activities is so impactful that they trigger a series of changes and effects on their natural settings. The effects may be on localised, subregional, regional and global scales, depending upon the length and breadth of such collective activities (Gill and

Malamud, 2017). Increasing evidence points out that the effect of anthropogenic processes on the immediate environment can be both intentional and non-malicious, with differing magnitude and scale (RSUSNAS, 2020). It is essential to delineate the geographical spread of common disaster risks induced by these anthropogenic influences based on natural hazards' occurrence, frequency, and intensity and identify the population at risk. Through the evolution of humanity, natural features have often been used for demarcating the land borders, such as political, socio-cultural and legal boundaries. In general, the features like mountain ranges, valleys, watersheds, rivers, streams and major lakes were used as boundaries. During every Population and Housing Census cycle, the National Statistical Organizations (NSOs) produce immense data. The census results show how people and places change over time, how dense they are, where their distribution is likely to cause problems, and so on. The requirements for census count as per the administrative boundary systems for governance and service delivery are met through tabulation plans. However, NSOs do not publish aggregate population counts based on non-administrative boundaries. In addition to administrative boundary-based population counts, aggregated population counts based on natural boundaries is also vital for understanding the influence of human on natural resources. The natural boundaries are defined based on natural processes; in some cases, they stand for millions of years (e.g., physiography). However, they are subject to change due to temperature, rainfall, and soil, which were also used to define them (e.g., rainfall or drought regions). These changes are continuously monitored through scientific measures, and the extent of such boundaries is revised accordingly.

The geographies used for census counts often vary, and it is always a challenge to use them for temporal studies. To overcome this limitation, many methods were proposed in the past. Goodchild and Lam used areal interpolation for aggregating census track level counts to different zone level boundaries (Goodchild and Lam, 1980). Flowerdew et al. used ancillary variables to improve the accuracy of areal interpolation (Flowerdew et al., 1991) by applying Expectation-Maximization (EM) algorithm for the aggregating census counts for parliamentary constituencies. Reibel and Bufalino used street weighted interpolation for population estimation in incompatible zones (Reibel and Bufalino, 2005) by negating the homogeneity assumptions. Langford employed binary dasymetric mapping aided with ancillary covariates generated from multi-spectral satellite imageries (Landsat 7-Enhanced Thematic Mapper) to apply the perceived covariant influence on the presence or absence of population in each pixel (Langford, 2007). In recent studies by Stevens et al. and Lloyd et al., global scale remotely sensed data were used as population covariates, while aggregating pixel-level census counts (Stevens et al., 2015; Lloyd et al., 2017b). Leyk et al. (2019) analysed the global scale gridded population datasets and their fitness for use by comparing the accuracy matrices (Leyk et al., 2019). These recent studies at global scale were made possible due to the High-Performance Computing (HPC) capacities, which allowed researchers to use multivariate geostatistical analysis to model gridded layers by applying several ancillary variables to estimate the pixel level counts.

Gridded population datasets are prepared using various approaches to estimate population counts at grid cells (Lloyd et al., 2017a). Population details with regular grids are

the most appropriate model for assessing population distributions, as they offer several significant advantages over administrative geographies. It extends the utility of census data for various studies. The importance of accurate spatial datasets containing the population traits and distribution is critical in many research domains such as health, economic, and environmental fields across various temporal and spatial scales (Stevens et al., 2015). Despite its legacy of 15 uninterrupted censuses, the Census of India datasets are continued to get published as per administrative boundary systems. However, the research communities interested in non-administrative geographies like physiography, agroclimatic, soil region, watershed, river basin, and earthquake zone, often have to spend lots of time and energy for recompiling the census data as per their region of interest. There were many attempts such as WorldPop (2010, 2015 and 2020), Global Human Settlement Layer (1975, 1990, 2000 and 2015), Gridded population of the World (2000, 2005, 2010, 2015 and 2020) and Global Rural Urban Mapping (1990, 1995 and 2000) to create gridded population raster layers for meeting such demands. However, these attempts disaggregate coarser census geographies to arrive at finer resolutions, causing poor accuracy of census counts at smaller geographies. The present research reviews the utility of existing open-access gridded population datasets and suggests possible methods for publishing similar products using Census of India datasets.

Materials and Methods

Study Area

The present study was attempted at three levels viz., Country (India), State (Karnataka) and City (Mysuru) levels (Fig. 1). India lies entirely in the Northern Hemisphere and extends between 6°45' to 37°06' North latitudes and 68°07' to 97°25' East longitudes covering over an area of 3.28 million sq.km. (ORGI, 1988) and accounts for 2.42% of the world land area with a share of around 18% in the world population. As per the Census of India 2011, India consists of 28 States, 7 Union Territories (Fig.1a), 640 Districts, 5,924 Sub-Districts, 7,933 towns (4,041 Statutory Towns and 3,892 Census Towns) and 6,40,930 Villages. The total population as per Census 2011 is 1,21,05,69,573, out of which 68.8 per cent live in rural and the remaining 31.2 per cent in urban (ORGI, 2013).

The areal extent of Karnataka (Fig. 1b) in terms of latitudinal and longitudinal spread is approximately from 11°35' to 18°29' North latitudes and 74°03' to 78°35' East longitudes. As per Census 2011, The State is divided into 27 districts, 176 Sub-Districts, 347 Towns (including 127 Census Towns) and 29,340 Villages (including 1943 Uninhabited Villages). The total population of the State as per Census 2011 is 6,10,95,297, out of which 61.33 percent live in rural while the rest 38.67 percent in urban.

Mysuru Municipal Corporation (M. Corp.), covering over 90 sq.km of area (Fig. 1c), is the fourth largest city in Karnataka State with a total population of 8,93,062, (ORGI, 2013). The latitudinal and longitudinal spread is approximately from 12°12'09" to 12°21'18" North latitudes and 76°35'56" to 76°42'20" East longitudes. Situated in the southernmost

part of the State, the city has been divided into 65 wards. Though the number of wards has remained unchanged since the 2001 Census, the extent of wards consistently changed between censuses.

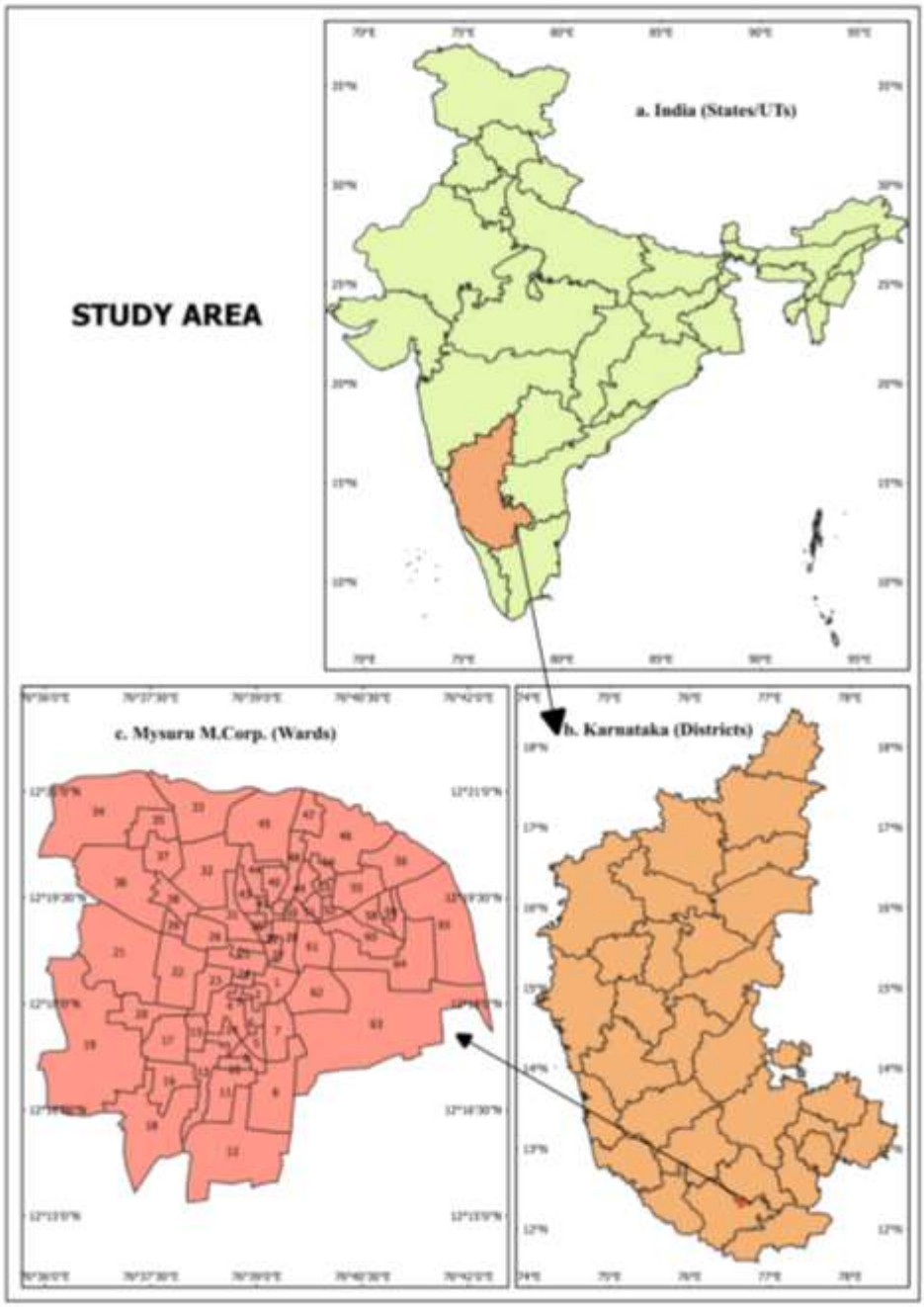

Fig.1: Administrative Units – a) India (States/UTs), b) Karnataka State (Districts), and c) Mysuru M. Corp (Wards)

Data Sources

Several initiatives have produced gridded population datasets (Table 1). In the present study, two open-source gridded population datasets, which were created using Census of India 2011 counts, namely WorldPop 2010 and Gridded Population of the World 2010 (GPW) have been downloaded from Center for International Earth Science Information Network (CIESIN, 2021) (https://sedac.ciesin.columbia.edu) and WorldPop (https://www.worldpop.org) portals. Both the datasets cover the entire country and are comparable to census counts from Census of India 2011. Each of these datasets provides different levels of accuracy and spatial resolution due to the modelling limitations. The WorldPop is a highly modelled gridded dataset created by employing random forest techniques based on multiple ancillary covariates for pixel level distribution of the census counts. It is also having the best spatial resolution of 100 m (3 arc sec) compared to other open-source gridded datasets (Stevens et al., 2015). On the other hand, the GPW datasets was produced at 1k resolution based on areal weighting without any ancillary variable. These raster datasets were clipped at three levels using ArcGIS Pro 2.8 with outline boundaries of India, Karnataka and Mysuru M. Corp. For spatial data analysis, district-level data for India, villages and town-level data for Karnataka State and ward level data for Mysuru have been considered, and the results were generated for each non-administrative geographies.

Table 1: Comparison of global level gridded population raster data sources

Open access Gridded data set	Produced by	Population Type	Method	Ancillary data	Modelling type and Resolution (Grid Cell Size) at Equator	File Types	Year of availability close to 2011 and Access Policy
Gridded Population of the World (GPW) ver. 4	CIESIN	Night-time population (population counted at place of domicile)	Areal weighted	Nil	Nil 1 km.	GeoTIFF, ASCII, and netCDF-4 format	2010 Open
Global Human Settlement Layer – Population (GHS-POP)	JRC and CIESIN	Night-time population (population counted at place of domicile)	Binary Dasymetric	Landsat-derived built up areas	Lightly Modelled 250 m and 1 km	GeoTIFF, Map service	2015 Open
World Population Estimate	ESRI	Mixed (45% of countries are night-time populations), based on available national population estimates, though progressing toward night-time with each new release	Dasymetric	Land cover for urban classes, road intersections, settlement locations, Landscape disturbance equating to human settlement	Highly modelled 250 m	Virtual Image Tile layer	2013 Commercial-ArcGIS Users
WorldPop	University of Southampton & CIESIN	Night-time population (population counted at place of domicile)	Random Forest	Settlement locations & extents, land cover, roads, built-up, health facility location, satellite nightlights, vegetation, topography, refugee camps	Highly modelled 100 m	GeoTiff	2011 Open
LandScan Global Population Datasets	Oak Ridge National Laboratory (ORNL)	Ambient Night-time population	Multi-layered, intelligent dasymetric, spatial modelling approach	Land cover, roads, slope, urban areas, village locations, and high-resolution imagery analysis	Highly modelled 1 km.	GeoTiff	2011 Commercial/Academic Research

Spatial Aggregation

The census counts were generated for selective non-administrative geographies through the zonal statistics algorithm of QGIS 3.16 by overlaying the vector layers on raster datasets. These non-administrative geographies are briefly elaborated in subsequent paras. A comparison of census counts at the district (India) and ward (Mysuru M.Corp) levels were also attempted for analysing the gridded layer fitness for use. Further, zonal statistics for agro-ecological regional level (as a test case) was calculated using vector layer (points) consisting of village/town level Karnataka State census counts and from that of gridded datasets to test the fitness of gridded population layers at a regional scale. Census counts at regional and administrative levels are tabulated and analysed using scatter plots to evaluate the linear relationship between raster datasets and the aggregates generated at regional, district and city scales using the JASP application (Jeffreys's Amazing Statistics Program).

Non-Administrative Geographies of India

The boundaries are defined and used by humans to meet specific purposes, and they are most often the conceptual toolkit used by the researchers for their enquiries (Lamont and Molnar, 2002). The boundary system can be natural, for example, boundaries created based on topographic features like elevation, soil and vegetation etc., or can be artificial based on political authority such as administrative land regions, cultural or socio-economic. The boundary systems created based on physiographical features are more visible than the artificial ones. In this study, except those created for political authority, all remaining boundary systems are treated as non-administrative geographies. Among the non-administrative boundaries, the four important geographies, namely, physiographic regions, agro-ecological regions, river basins and earthquake zones (Fig. 2) were chosen to assess the utility of gridded population datasets. The characteristics of these non-administrative geographies are briefly discussed in the following sections.

Physiographic Regions: The physiographical divisions of India were first attempted by Holdich in 1904, which is somewhat a broad geographical zone of India based on geological information only (Holdich, 1904). An attempt by Stamp (1922-24) produced a more substantive division named as 'natural regions' based on physiographic structure for three macro-regions and 22 subregions (Stamp and Dudley, 1928). Later, the regional divisions proposed by Spate were empirically derived and divided India into 35 regions of the first order (under the three macro-regions and excluding the islands), 74 second-order divisions with 225 subdivisions (Spate and Learmonth, 1954). During the Census 1981 phase, more profound work on regional divisions based on physiography, geological structure, climatic conditions, and soils was attempted and mapping all these regions using village/town level boundaries was also completed. In this frame, four Macro, 28 Meso (State boundaries used to split the 4 Macro regions) and 101 Micro regions (grouping Meso regions by using district boundaries) were outlined for the country (ORGI, 1988). Sub-micro regions were delineated within this frame of micro-regions by considering the village/town

boundaries. The analysis for the present research was done using the four Macro regions (Fig. 2a).

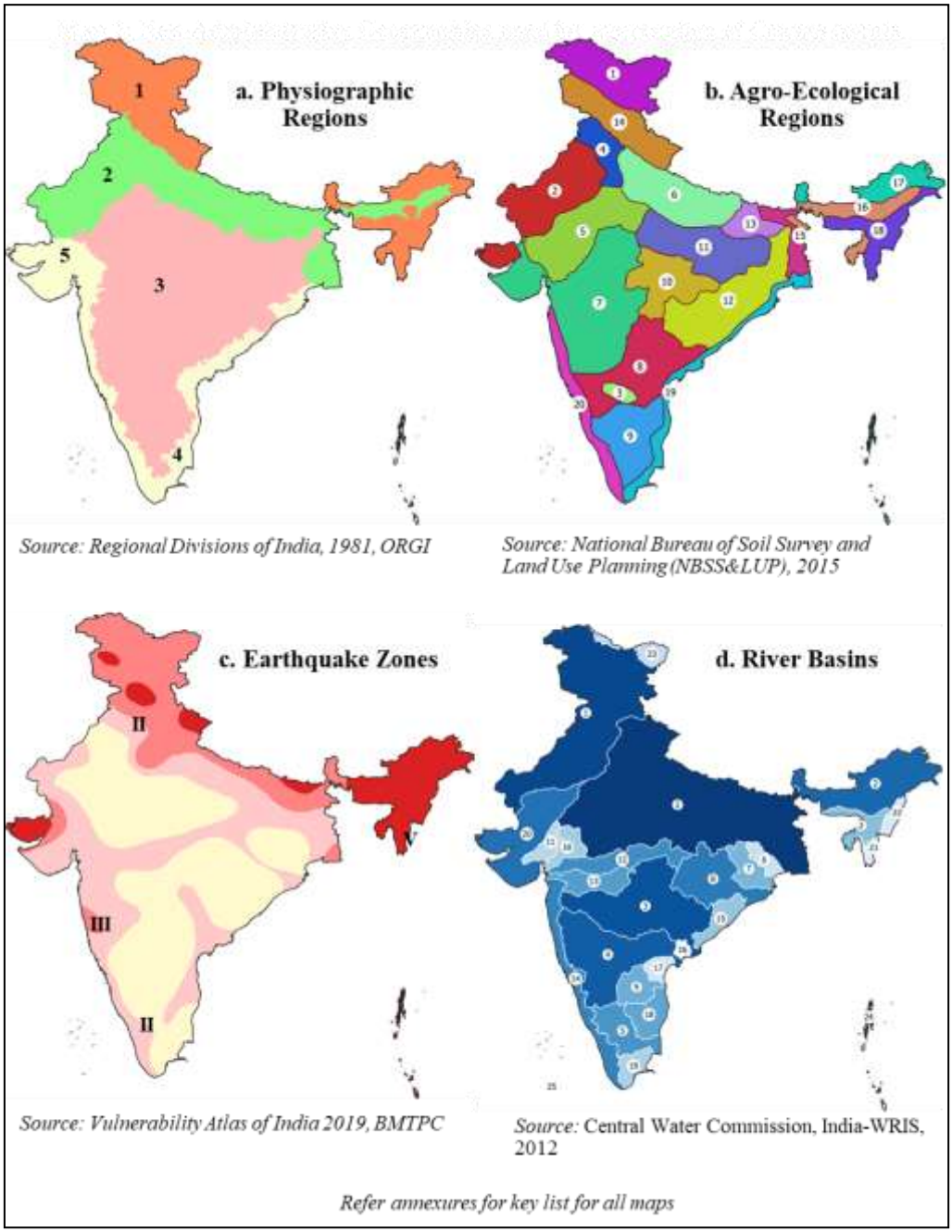

Fig. 2: Non-administrative geographies used for aggregation of census counts – a) Physiographic Regions, b) Agro-Ecological Regions, c) Earthquake Zones, and d) River Basins

Agro-Ecological Regions: The Agro-Ecological Regions (AER) are carved out based on agroclimatic conditions derived by superimposing climate parameters on landforms and soils. As the climate and soils are the modifiers of the length of the growing period, the agro-ecological regions are critical aspects for studies on agriculture-related issues. Krishnan and Singh (1968) delineated soil climatic zones by superimposing moisture index and mean air temperature isopleths on broad soil types of India (Virmani et al., 1980). The attempt by Murthy and Pandey (1978) used physiography, climate (rainfall and potential water surplus/deficit), major soils and agricultural regions. During 1988, under the Planning Commission, a similar attempt was made and 15 agroclimatic regions with 73 sub regions were delineated. The refined work of the National Bureau of Soil Survey and Land Use Planning (NBSS&LUP) has divided the country into 20 agro-ecological regions (AERs). The analysis for present research was done using the 20 AERs (Fig. 2b) defined by the NBSS&LUP (Mandal et al., 2014).

Earthquake Zones: As per the seismic zoning map of the country (BMTPC, 2019), the total area is classified into four seismic zones (Medvedev–Sponheuer–Karnik scale). Zone V is seismically the most active region, while zone II is the least. Approximately 11% area of the country falls in zone V, 18% in zone IV, 30% in zone III and remaining in zone II. For the present study, the analysis of the gridded population layer was done as per earthquake zone boundaries delineated in the Vulnerability Atlas of India, 2019 (Fig. 2c).

River Basins: The rivers are the soul of civilisation and a vital geophysical ecosystem engine that support and sustain life on earth. The term "river basin" encompasses many sub-systems such as surface (soil and land resources), subsurface water resources, wetlands and associated ecosystems, including those coastal and nearshore marine systems. The systematic delineation of river basins in India was attempted in 1949 by the erstwhile Central Waterways, Irrigation and Navigation Commission (now Central Water Commission). CWC has delineated 20 river basins, comprising 12 major river basins and 8 composite basins, using Survey of India (SOI) toposheets and contour maps. CWC revised the basin's boundaries and classified 32 basins, 94 sub-basins, and 3448 watersheds in the latest available data. The study used River Basin Atlas of India (Fig. 2d) sourced from the India-WRIS portal (India-WRIS, 2012) for river basin level analysis .

Comparisons of gridded datasets

The gridded population datasets were compared with the Administrative Atlas of India 2011 published by Office of the Registrar General, India, for ascertaining fitness of their use. The Gridded Population of World (GPW) (Fig. 4a) was created using simple areal weighting method to distribute the census counts proportionately and equally across the all grids (at given scale) (Fig. 3a). Whereas Global Human Settlement Layer (Fig. 4c) and World Population Estimates (Fig. 4d) are derived weights based on built-up and land use through dasymetric techniques (Fig. 3b & 3c). Similarly, the WorldPop dataset was created using dasymetric model based on statistically derived weights from multiple ancillary

variables for assigning census counts to each of the grids (Fig. 3d). As the gridded population layers (WorldPop and GPW) used Census of India 2011 counts, these two datasets were selected for the study for comparison. The village / town layers (Fig. 5b polygon centroids) sourced from Karnataka State Remote Sensing Application Centre were used to calculate zonal counts for agro-ecological regions (Fig. 5a) and used to compare with gridded datasets at state level. The ward level boundaries of Mysuru M. Corp. were used to compare ward level census counts from gridded datasets. It is important to note that variations may arise in population counts between GPW and WorldPop raster datasets due to limitation of zonal statistics algorithm. The over counts are possible due to the cells along the shared borders of zonal boundaries are also get counted more than once (called as edge effect), producing over representation of cell values across shared zonal boundaries. There is also possibility for over counts due to misalignment of boundaries used for creating gridded layers.

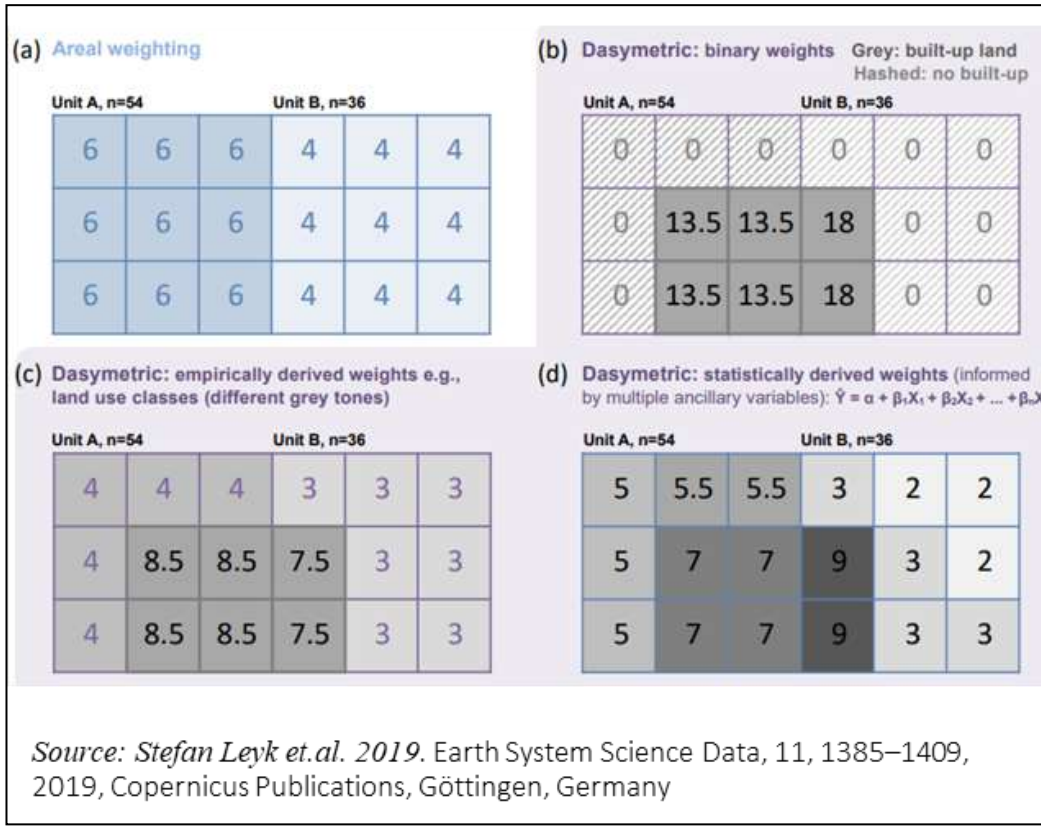

Fig. 3: Different statistical approach for assigning census population counts to grid cells – a) Areal Weights, b) Dasymetric: Binary Weights, c) Dasymetric: Empirically derived Weights, and d) Dasymetric: Statistically derived Weights

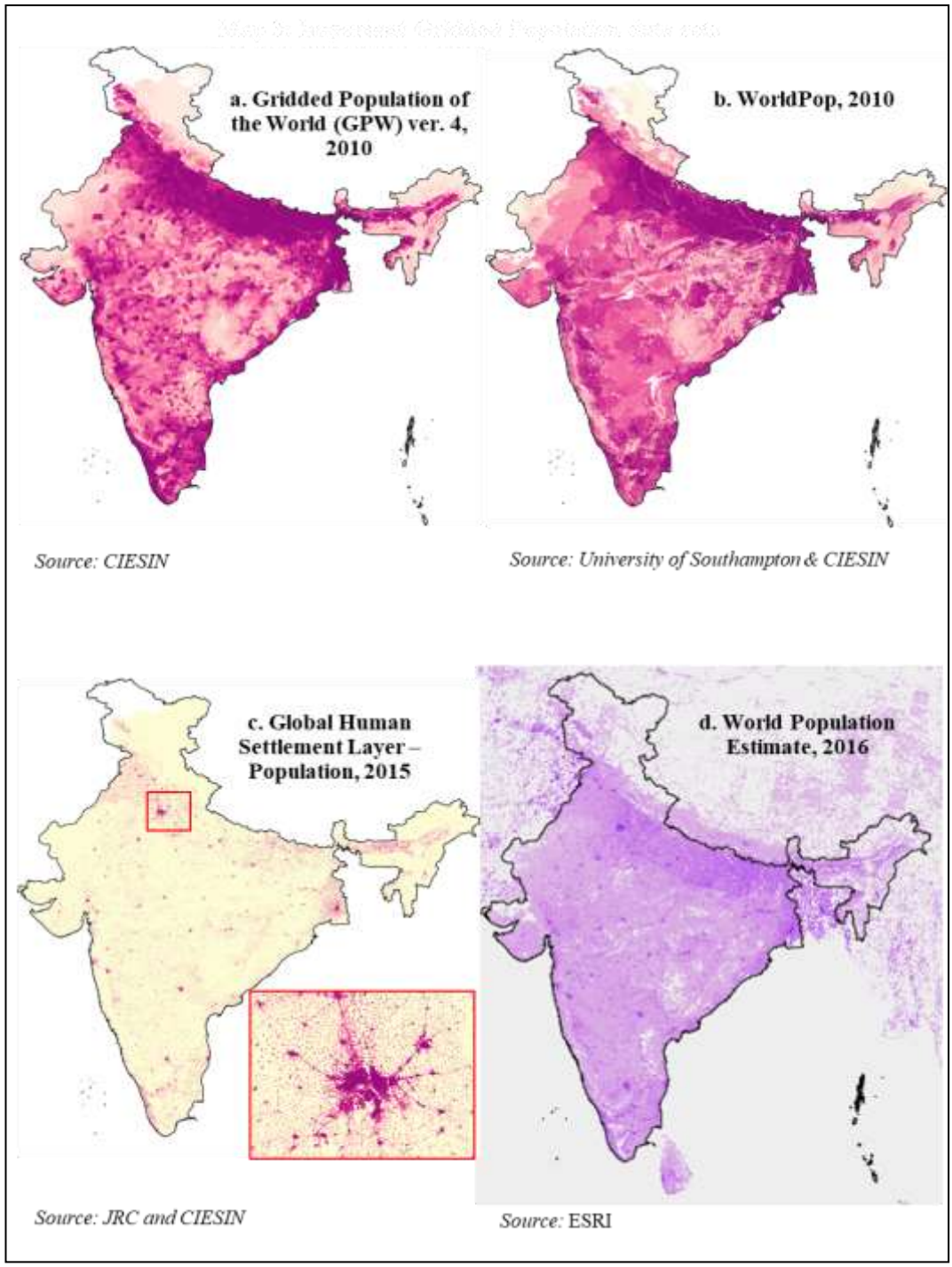

Fig. 4: Gridded population datasets of India – a) Gridded Population of the World, b) World Population, 2010, c) Global Human Settlement Layer – Population, 2015, and d) World Population Estimate, 2016

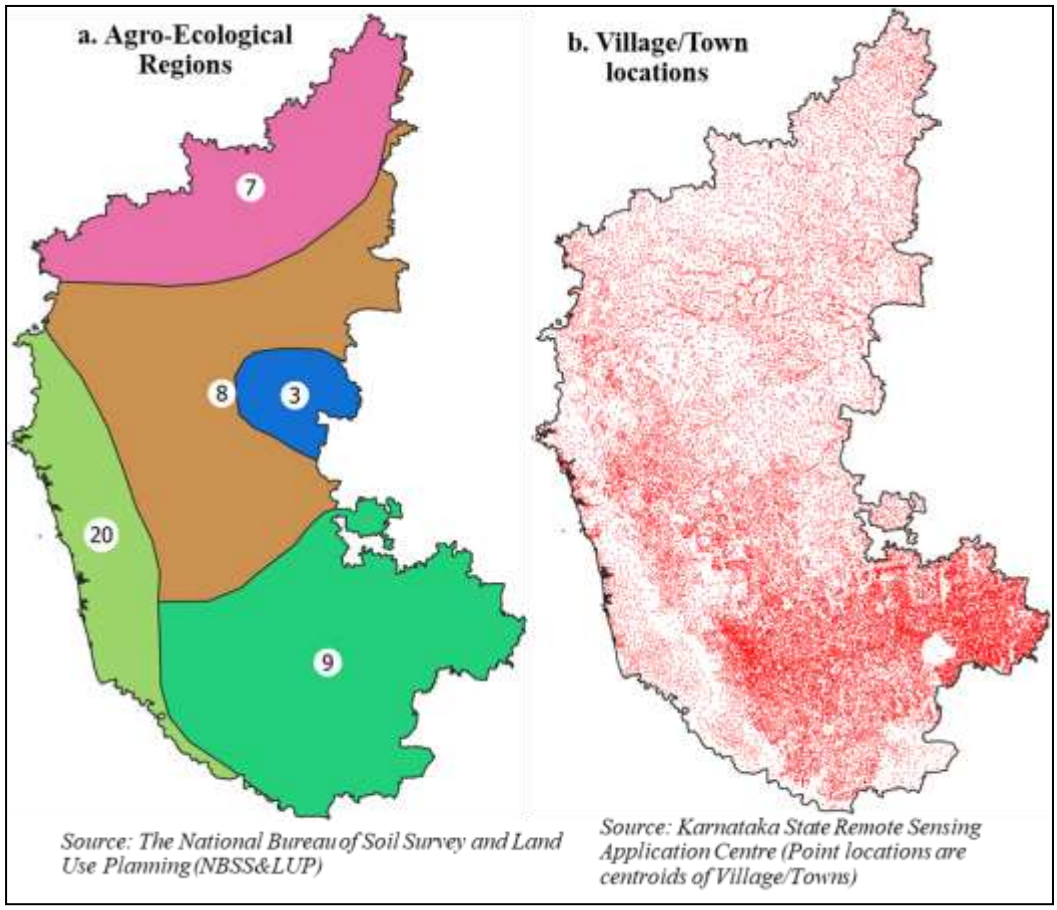

Fig. 5: (a) Agro-ecological regions of Karnataka State and (b) Location (centroids) points of villages and towns in Karnataka State

Results and Discussion

Aggregates of census counts for physiographic regions: The aggregates of Census count for four of the non-administrative geographies are presented in Tables 2 to 5. The highest proportion of the population lives in the Great Plains (GWP: 40.63%, WorldPop: 41.07%) and the Deccan Plateau (GWP: 36.20%, WorldPop: 36.18%). These two physiographic regions constitute 2.2 million sq.km of area (Table 2) and more than 3/4th of population of the country. The smallest proportion of population live in the Northern Mountain region (GWP: 4.72%, WorldPop: 4.29%). The difference in counts between GPW and WorldPop is less than half a percent and the highest is observed in the Northern Mountain region (0.43%)

Aggregates of census counts for earthquake zones: About 6.4% of people live in very high earthquake zone V (GWP: 6.46%, WorldPop: 6.42%), and near about 20% of people live in moderately high earthquake Zone IV (Table 3). The remaining 73% people

live in Zone III and II, which accounts for 72 percent (1.37 million sq.km.) of area in the country. The difference in census counts between GPW and WorldPop is less than half percent across all zones, and the highest could be observed in Zone II (0.30%).

Table 2: Aggregate census counts from gridded population data sets for physiographic regions of India

Physiographic Regions	Area	Aggregate population			
		GPW	Percentage share	WorldPop	Percentage share
Northern Mountains	5,42,819	5,82,82,328	4.72	5,42,48,996	4.29
The Great Plains	7,30,116	50,13,10,467	40.63	51,88,99,623	41.07
Deccan Plateau	15,20,788	44,65,88,010	36.20	45,70,85,023	36.18
Eastern Coastal Plains and Islands	1,89,519	9,34,13,365	7.57	9,50,55,985	7.52
Western Coastal Plains and Islands	2,90,926	13,41,68,782	10.87	13,81,49,696	10.93

Table 3: Aggregate census counts from gridded population data sets for earthquake zone of India

Earthquake Zone	Area	Aggregate population			
		GPW	Percentage share	WorldPop	Percentage share
II	13,73,944.32	40,07,60,240	32.48	41,41,89,957	32.78
III	9,83,973.32	50,24,97,084	40.73	51,26,70,815	40.58
IV	5,50,630.66	25,08,40,213	20.33	25,54,38,386	20.22
V	3,65,641.38	7,96,69,538	6.46	8,11,43,918	6.42

Aggregates of census counts for agro-ecological regions: Among the agro-ecological regions, the proportion of people living in the Northern Plain Middle Gangetic Plain is 16% and hot semiarid with moderately deep black soils region is 12%. Together, these two regions constitute more than 1/4th of population in the country (Table 4). Other significant proportion of population (between 5-10%) live in regions, which are hot semiarid spread over Deccan and hot subhumid spread over the coastal areas.

Aggregates of census counts for river basins: Among the river basins (Table 5), the Ganga basin alone constitutes around 44% of population in the country, and constitute 1/4th of the geographical area of the country (25.88%). Other river basins with significant share of population are Krishna (7%), Indus and West flowing rivers from Tapi to Kadri (around 6.5%) and Godavari Basins (6%). Remaining 20 basins have proportion of population less than 5 percent and all of them sum up to around 30 percent of population of the country and thus, top five river basins alone constitute 70 percent of country's population.

Table 4: Aggregate census counts from gridded population data sets for agro-ecological regions of India

AER Code	Agro-Ecological Region Name	Area in Sq.Km.	Aggregate population			
			GPW	Percentage share	WorldPop	Percentage share
1	Western Himalayas, cold arid eco-region with shallow skeletal soils	1,77,707	14,94,050	0.12	14,52,676	0.11
2	Western plains and Kutch Peninsula, hot arid ecoregion with desert saline soils	2,54,302	3,61,23,607	2.93	3,75,88,078	2.98
3	Deccan plateau, hot arid eco-region with mixed red and black soils	21,398	54,32,006	0.44	55,76,032	0.44
4	Hot semi-arid ecoregion with alluvium-derived soils	68,710	5,52,71,974	4.48	6,09,58,922	4.82
5	Deccan plateau, hot arid eco-region with mixed red and black soils	2,25,847	8,02,71,130	6.51	8,12,01,886	6.43
6	Northern Plain Middle Gangetic Plain with hot semiarid to sub humid climate and alluvial and Tarai soils	2,10,767	19,82,73,220	16.07	20,02,69,014	15.85
7	Hot Semiarid with moderately deep black soils	4,90,714	14,85,08,357	12.04	15,55,97,894	12.32
8	Deccan Plateau, hot semiarid eco-region with mixed red and black soils	2,48,400	7,17,70,347	5.82	7,52,77,639	5.96
9	Deccan Plateau, hot semiarid eco-region with red loamy soils	1,74,563	7,51,26,347	6.09	7,76,30,102	6.14
10	Hot sub humid eco-region with moderately deep black soils	1,55,184	4,03,13,460	3.27	4,10,88,435	3.25
11	Eastern Plateau (Bundelkhand Upland) hot sub humid eco-region with red and yellow soils	2,18,466	6,81,94,299	5.53	6,82,09,366	5.40
12	Eastern Plateau, hot sub humid eco-region with red and lateritic soils	2,66,098	6,85,11,715	5.55	7,15,24,174	5.66
13	Northern Plains (Lower Gangetic) hot, sub humid ecoregion with alluvial soils	57,540	7,17,58,182	5.82	7,29,69,089	5.78
14	Western Himalayas, warm to hot sub humid to humid sub montane shallow and skeletal hill soils	1,52,370	3,08,47,048	2.50	2,68,39,344	2.12
15	Bengal basin, hot, sub humid eco-region with loamy to clayey alluvial soils	51,929	7,47,24,799	6.06	7,59,77,269	6.01
16	Assam and North Bengal Plain, warm humid to per humid eco-region with alluvial soils	94,175	4,52,96,222	3.67	4,73,35,927	3.75
17	Eastern Himalayas, warm per humid eco-region with shallow and skeletal red soils	86,841	41,56,869	0.34	34,88,424	0.28
18	North Eastern hills (Purvanchal), warm per humid ecoregion with red and yellow soils	99,690	1,25,79,947	1.02	1,20,56,419	0.95
19	Eastern Coastal Plains and Island of Andaman and Nicobar, hot sub humid	1,17,721	7,34,46,547	5.95	7,48,18,023	5.92
20	Western Ghats Coastal Plains and Western Hills with red and lateritic and alluvium derived soils	1,01,768	7,16,66,950	5.81	7,35,84,363	5.82

Note: Due to edge effect, the population counts shall exceed the total population of the country.

Table 5: Aggregate census counts from gridded population data sets for the major river basins of India

WRIS Code	Name of the River Basins	Area in Sq.Km.	Aggregate population			
			GPW	Percentage share	WorldPop	Percentage share
1	Indus Basin	4,67,069	8,33,22,851	6.75	8,21,58,512	6.50
2a	Ganga Basin	8,39,786	53,84,94,209	43.65	55,49,63,879	43.92
2b	Brahmaputra Basin	1,92,743	4,11,82,969	3.34	4,11,26,851	3.26
2c	Barak & Others	49,083	1,00,87,177	0.82	99,19,183	0.79
3	Godavari Basin	3,10,257	7,35,57,884	5.96	7,45,85,567	5.90
4	Krishna Basin	2,60,212	8,54,56,361	6.93	8,91,32,362	7.05
5	Cauvery Basin	85,865	4,23,64,386	3.43	4,38,08,981	3.47
6	Subarnarekha Basin	26,721	1,20,21,764	0.97	1,22,52,814	0.97
7	Brahmani & Baildoni Basin	53,497	1,42,51,203	1.16	1,44,27,585	1.14
8	Mahanadi Basin	1,44,300	4,02,78,734	3.26	4,00,96,246	3.17
9	Pennar Basin	55,075	1,16,86,357	0.95	1,26,75,950	1.00
10	Mahi Basin	39,659	1,64,87,177	1.34	1,61,96,850	1.28
11	Sabarmati Basin	31,853	1,69,78,424	1.38	1,72,36,113	1.36
12	Narmada Basin	95,612	2,07,40,284	1.68	2,08,76,296	1.65
13	TapiBasin	65,524	2,26,64,944	1.84	2,19,19,670	1.73
14	West Flowing River from Tapi to Kadri	1,14,027	7,96,40,162	6.46	8,37,91,018	6.63
15	East Flowing Rivers between Mahanadi and Pennar	48,477	1,69,32,186	1.37	1,73,57,052	1.37
16	East flowing rivers between Godavari and Krishna	10,691	49,57,686	0.40	54,15,342	0.43
17	East flowing rivers between Krishna and Pennar	23,723	67,52,406	0.55	66,56,078	0.53
18	East Flowing Rivers between Pennar and Cauvery	64,096	4,03,78,277	3.27	4,17,69,403	3.31
19	East Flowing Rivers between Cauvery and Kannyakumari	38,312	1,69,83,703	1.38	1,72,17,482	1.36
20	West Flowing Rivers of Kutch and Saurashtra Including Lun	1,91,492	3,49,07,526	2.83	3,61,54,765	2.86
21	Minor Rivers draining into Myannmar and Bangladesh	12,645	5,05,198	0.04	5,25,889	0.04
22	Minor rivers draining into Myanmar	17,627	27,42,369	0.22	27,77,954	0.22
24	Drainage Area of ANI	6,842	3,37,271	0.03	3,45,273	0.03
25	Drainage Area of Lakshadweep	30	40,624	0.00	52,210	0.00

Comparisons of gridded population datasets with census counts: The comparisons of gridded layers done by matching of census counts published by ORGI (ORGI, 2013) at the levels of districts (India) and wards (Mysuru M. Corp).

The dot plots created using ORGI District level counts and counts derived from gridded datasets exhibit relatively strong relationship (GPW: 0.9672, WorldPop: 0.9726). However, there are more outliers in GPW dataset than in the WorldPop (Fig. 6a & 6b). This is so because, the GPW was created using areal interpolation without covariates, whereas the WorldPop was created using multiple ancillary variables for estimating and distributing census counts.

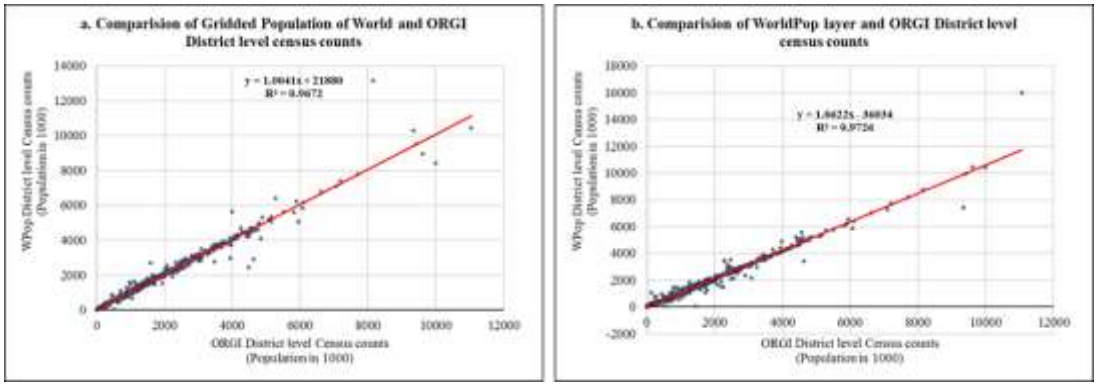

Fig. 6: a) Comparison between ORGI district level census counts and gridded population layers, and b) Comparison between ORGI district level census counts and World population layer

The ward level aggregates created from both GPW and WorldPop exhibit insignificant relationship (Fig. 7a & 7b) with ward level census counts from ORGI. Though insignificant, yet the WorldPop shows slightly better fit (R^2: 0.0202) than GPW (R^2: 0.0071) due to better pixel level resolution. As the distribution of census counts for the WorldPop dataset was created based on the built-up, land cover and other covariates, have profound control on the probability of assigning Census counts for known urban clusters. Yet, using these two datasets for ward level geographies that are smaller than sub-district level geographies may not yield accurate results, hence not fit for use.

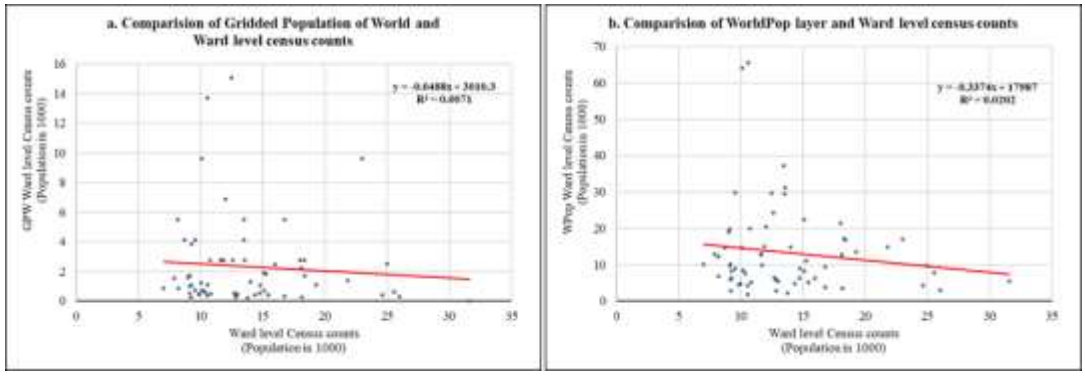

Fig. 7: a) Comparison of ORGI Mysuru M.Corp. ward level units census counts with the gridded population layers, and Comparison of ORGI Mysuru M.Corp. ward level units census counts with the World population layer

Improving accuracy of aggregates

CIESIN advices to keep the aggregate geographies either equal to that of geography used for modelling the grids or at least one step higher. As the sizes of the non-administrative geographies are higher than the sub-district level, both datasets are good enough to produce the census counts for non-administrative geographies of higher orders

i.e., above the size of districts. However, smaller geographies need better granularity of administrative geographies for creating gridded population raster datasets. In addition to increasing the granularity of lower-level administrative geographies, the accuracy can be further improved by using ancillary covariates, while distributing the census counts at village / town levels. As the granularity increases along with use of covariates such as built-up, land use, road density and night-time light data derived from remote sensing platforms, it is feasible to create more accurate gridded population datasets for India. By producing gridded population datasets with good accuracy, the population data gap can be filled in various domains of research such as climate change, public health, disease modelling, agriculture, disaster risk, impact assessment and so on.

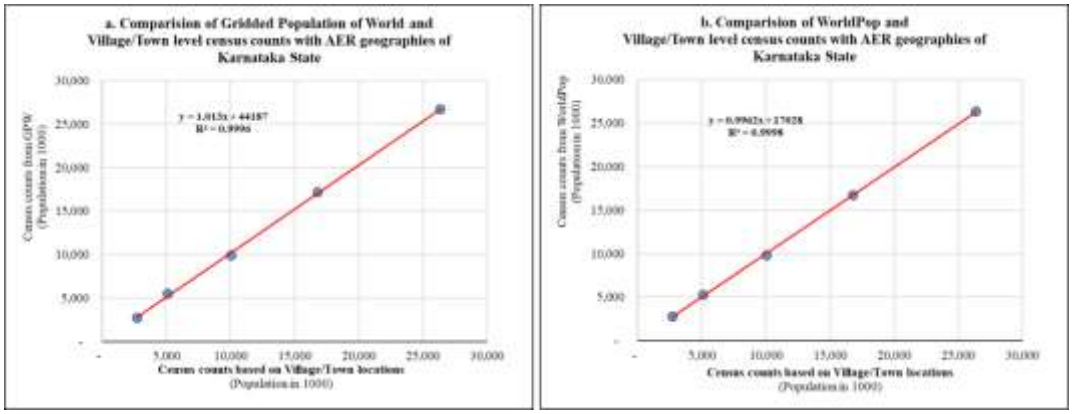

Fig. 8: a) Comparison between village/town level census counts with AER geographies of Karnataka State and gridded population of World, and b) Comparison between village/town level census counts with AER geographies of Karnataka State and World population layer

Conclusion

The evaluation of existing open access gridded population datasets for recompiling census counts for non-administrative geographies at different scales, attempted through zonal summary statistics algorithms in QGIS application. The comparisons of GPW (resolution 1K) and WorldPop (resolution 100 m) shows that there is a strong linear relationship (R^2 0.9334) between these datasets even though the latter was created using simple areal weighting method. However, the comparison of aggregates from gridded population datasets at district, city, village/town and ward level reveals that there are accuracy issues below the sub-district levels.

However, these accuracy issues can be addressed by creating gridded population layers using village/town level census counts along with the covariates. This approach is demonstrated at village/town level for the different agro-ecological geographies of Karnataka State. The results show perfect linear relationship (GPW: R^2 0.9996, WorldPop: R^2 0.9998) with the census counts (Table 6). It is feasible for the ORGI to generate such datasets for the larger interest of the research communities instead of merely providing

standards census tables. The gridded datasets can eliminate the amount of time and effort required for recasting census data for non-administrative boundaries. In addition, the preparation of micro-level gridded datasets can enhance the utility of census counts and will be very useful for all micro-level studies using non-administrative geographies.

Table 6: Comparison of GPW, WorldPop and actual census counts for agro-ecological regions of Karnataka State

Region Code	Agro-Ecological regions	Aggregated Population created from			Variation from aggregates of Village/Town level units (ORGI)	
		WorldPop	GPW	Village/Towns	WorldPop	GPW
3	Deccan plateau, hot arid eco-region with mixed red soil and black soils	27,54,962	27,06,542	27,39,572	-0.6	1.2
7	Hot Semiarid with moderately deep black soils	98,21,713	99,44,554	1,00,27,006	2.0	0.8
8	Deccan Plateau, hot semiarid eco-region with mixed red and black soils	1,67,40,429	1,71,72,664	1,67,92,564	0.3	-2.3
9	Deccan Plateau, hot semiarid eco-region with red loamy soils	2,63,31,258	2,67,35,340	2,63,66,025	0.1	-1.4
20	Western Ghats Coastal Plains and Western Hills with red and lateritic and alluvium derived soils	52,39,992	54,87,572	51,06,978	-2.6	-7.5
Total		6,08,88,354	6,20,46,672	6,10,32,145	0.2	-1.7

Note: Due to edge effect, the population counts shall exceed the total population of the State

References

1. BMTPC (2019). *Vulnerability Atlas of India (Third Edition).* Building Materials & Technology Promotion Council (BMTPC), New Delhi.
2. CIESIN (2021). *Documentation for the Gridded Population of the World, Version 4*. Center for International Earth Science Information Network (CIESIN), Columbia University. (Retrieved from https://sedac.ciesin.columbia.edu/binaries/web/sedac/collections/gpw-v4/gpw-v4-documentation-rev11.pdf)

3. Dudley, S.L. (1928). The natural regions of India. *Geography*, 14(*6*), 502-506.
4. Flowerdew, R., Green, M. and Kehris, E. (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science,* 70(*3*), 303-315.
5. Gill, J.C. and Malamud, B.D. (2017). Anthropogenic processes, natural hazards, and interactions in a multi-hazard framework. *Earth-Science Reviews*, 166(*2017*), 246-269.
6. Goodchild, M.F. and Lam, N.S. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing,* 1, 297-312.
7. Hendry, A.P., Gotanda, K.M. and Svensson, E.I. (2017). Human influences on evolution, and the ecological and societal consequences. *Philosophical Transactions B*, 1-13. (https://doi.org/10.1098/rstb.2016.0028)
8. Holdich, T.H. (1904). *The Regions of the World-India.* Oxford University Press, London.
9. India-WRIS (2012). *River Basin Atlas of India.* WRIS and RRSC-West, NRSC-ISRO, Jodhpur.
10. Krishnan, A. and Singh, M. (1968). Soil climatic zones in relation to cropping patterns. *Proceedings Symposium on Cropping Patterns*, Indian Council of Agricultural Research, New Delhi, 172-185.
11. Lamont, M. and Molnar, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, 28, 167-195.
12. Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems,* 31(*1*), 19-32.
13. Leyk, S., ….. and Pesaresi, M. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data,* 11(*3*), 1385–1409.
14. Lloyd, C.D., Catney, G., Williamson, P. and Bearman, N. (2017a). Exploring the utility of grids for analysing long term population change. *Computers Environment and Urban Systems*, 66, 1-12.
15. Lloyd, C.T., Sorichetta, A. and Tatem, A.J. (2017b). High resolution global gridded data for use in population studies. Scientific Data, 4-170001, DOI: 10.1038/sdata.2017.1.
16. Mandal, C., ….. and Thakre, S. (2014). Revisiting agro-ecological sub-regions of India – A case study of two major food production zones. *Current Science,* 107(*9*), 1519-1536.
17. Murthy, R.S. and Pandey, S. (1978). Delineations of agro-ecological regions of India. *In Paper presented in Commission V, 11th Congress of ISSS*, Edmonton, Canada, 19-27.
18. ORGI (1988). *Regional Divisions of India-A Cartographic Analysis.* Office of the Registrar General India, New Delhi.
19. ORGI (2013). *Primary Census Data Highlights - India.* Office of the Registrar General India, New Delhi.
20. Palumbi, S.R. (2001). Humans as the World's greatest evolutionary force. *Science*, 293(*5536*): 1786-1790. DOI: 10.1126/science.293.5536.1786.
21. Reibel, M. and Bufalino, M.E. (2005). Street-weighted interpolation techniques for demographic Street-weighted interpolation techniques for demographic. *Environment and Planning A Vol(37)*, 129-139.

22. RSUSNAS (2020). *Climate Change-Evidence and Causes- Update 2020 - An overview from the Royal Society and the U.S. National Academy of Sciences (RSUSNAS).* The National Academy of Sciences, Washington and The Royal Society, London.
23. Spate, O.H.K. and Learmonth, A.T.A. (1954). *India and Pakistan: A general and regional geography.* Methuen, London.
24. Stevens, F.R., Gaughan, A.E., Linard, C. and Tatem, A.J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, 10(*2*): e0107042.
25. Virmani, S.M., Sivakumar, M.V.K. and Reddy, S.J. (1980). Climatic classification of semi-arid tropics in relation to farming systems research. *Consultants' Meeting on Climatic Classification,* ICRISAT (International Crops Research Institute for the Semi-Arid Tropics), Patancheru, 59-88.