Robust Parliamentary Constituency Estimates
Geographic Data Science Approaches

JEFFREY C BLOSSOM, AKSHAY SWAMINATHAN, WILLIAM JOE, ROCKLI KIM, S V SUBRAMANIAN

This article is a response to Srinivas Goli's article "Unreliable Estimates of Child Malnutrition" (*EPW*, 9 February 2019) that had questioned the reliability of methodologies of Akshay Swaminathan et al's article "Burden of Child Malnutrition in India: A View from Parliamentary Constituencies" (*EPW*, 12 January 2019). The reliability and usability of the methodologies proposed by Swaminathan et al have been reiterated, emphasising that these can provide broad assessments at the parliamentary constituency level.

Parliamentary constituencies (PCs) represent an important geographic unit of governance and agenda-setting for health, nutrition, and development domains in India. Each of the 543 PCs has a representative member of Parliament (MP) who is responsible for representing the interests of the people living within the PC. Yet, data on key developmental indicators are not collected at the PC level. Instead, most existing data is available at the administrative unit of districts. Given the lack of direct correspondence between the PC and district boundaries, and the absence of data at the PC level, we recently proposed two methodologies in Swaminathan et al (2019) to estimate PC level burden of child malnutrition, using the available geographic information system (GIS) shapefiles and nationally representative data.

The first methodology involved building an indirect crosswalk between districts and PCs using boundary shapefiles; and the second involved aggregating individual-level data to a potential PC directly linked via the randomly displaced global positioning system (GPS) locations of the National Family Health Survey (NFHS-4) sampling clusters (Swaminathan et al 2019). In a subsequent article, we further refined these methodologies by applying precision-weighted estimations based on multilevel modelling to account for the multilevel data structure of the NFHS-4 and sampling variability, and presented PC estimates for child stunting, underweight, wasting, low birth weight, and anaemia (Kim et al 2019).

In a response to our article (Swaminathan 2019), Srinivas Goli (2019), while agreeing with the substantive importance of presenting developmental indicators

at the PC level, claimed that our estimates for child malnutrition indicators were "unreliable." This claim, however, was not supported by any empirical evidence. In turn, what was outlined by Goli (2019) was a list of concerns that could be examined for their veracity. Upon a detailed examination of these concerns, we have no reasons to believe that our PC estimates are "unreliable." Our original conclusion suggesting that, in the absence of "ground truth" data, our proposed methodologies can be used to provide broad assessments at the PC level remains unchanged.

We take this opportunity to elaborate and discuss the major concerns raised by Goli (2019). In doing so, we also further elucidate our methodologies to encourage future replication and application. We first provide a detailed explanation on how we created the indirect crosswalk between districts and PCs, followed by a section on discussing the direct methodology of linking clusters to PCs. Then, we present results from four different sensitivity analyses that collectively strengthen our initial findings. Next, we further assesses the validation of our estimates by applying our proposed methodologies to female literacy, an indicator for which census data are available to be aggregated to PC level and hence serve as the "ground truth." Finally, we conclude with a brief summary of our robust geographic data science approaches to estimate PC indicators.

Building Indirect Crosswalk

Potential methods for estimating district indicators at the PC level using advanced GIS tools require geographic data sets that map population, district borders, and PC borders. For our article (Swaminathan 2019), we used district borders representative of 2011 and PCs representative of 2014 as produced in shapefile format by the Community Created Maps of India (CCMA) project of DataMeet. Regarding these shapefiles, the CCMA states, "Data is not perfect, there are many errors both in data and boundaries." While there exists several district boundary data sets for different years in

GIS format, this was the only source that had PC boundaries. Knowing we were going to intersect the two shapefiles, we wanted districts and PCs that were as closely matched as possible where the boundaries were coincident. The CCMA's statement that "the external boundary of the district shapefile produced by this project was derived from the PC boundaries" (DataMeet) indicated there was an effort made to match at least the peripheral boundaries of districts and PCs. Visual inspection of the district and PC shapefiles in GIS software confirmed that a majority of the boundaries are indeed coincident where they should be, with a few exceptions.

Figure 1 shows an area in Karnataka state where three district boundaries intersect. Notice the district and PC borders between Vijayapura (Bijapur) and Yadgir and Vijayapura and Kalaburagi (Gulbarga) are perfectly coincident. However, the border between Kalaburagi and Yadgir exhibits disparate district/PC boundaries, when in reality these probably should be coincident. How this data limitation affects the reliability of our estimates will be discussed after elaborating on the population data we chose to use.

Many population data sets exist for India in GIS format. For our analysis we wanted to model population ground conditions to most closely match the time period January 2015 to December 2016—the same year the NFHS-4 child malnutrition data were collected. We also desired a population data set with a fine enough spatial resolution to most closely represent the disparate nature of the PC and district boundaries. In choosing an appropriate population data set for our study, we analysed the strengths and limitations of three different data sets, presented below.

LandScan global population database 2016: It is a population estimate produced by the Oak Ridge National Laboratory using demographic census data and remote sensing imagery analysis in a dasymetric modelling approach. The model is tailored to match the data conditions and geographical nature of individual countries. LandScan represents

estimations of people per grid square in a raster data format at an approximately 1 kilometre (km) spatial resolution. This resolution varies from 960 metre (m) east to west by 960 m as measured in the extreme southern portion of Tamil Nadu province to 760 x 930 as measured in the extreme north of Jammu and Kashmir (J&K). Strengths of the LandScan population data are its complete coverage across India and temporal representation of the year 2016, which is within the time period of the NFHS-4. Limitations are that it presents population estimates, and not actual values, and at a spatial resolution of 900 m, it is in some locations too coarse to effectively model detailed variations exhibited by the PC boundaries.

AsiaPop 2015: It is a population estimate produced by the School of Geography and Environmental Science, University of Southampton, using demographic Census data, land cover remote sensing imagery analysis and a dasymetric modelling design. The model is adjusted to match the geographical conditions for individual countries in Asia for 2015 (Gaughan et al 2013). AsiaPop 2015 represents estimations of people per grid square in a raster data format at an approximately 100 m spatial resolution. This resolution varies from 96 m east to west by 96 m as measured in the extreme southern portion of Tamil Nadu province to 75 x 90 as measured in the extreme north of J&K. Strengths of the AsiaPop 2015 population data are its complete coverage across India and temporal representation of 2015, which is within the time period of the NFHS-4, and its spatial resolution of 90 m is granular enough to effectively model districts and PCs that exhibit a heterogeneous mix of both rural and urban areas. Limitations are that it presents population estimates, not actual values.

Census of India 2011 village points: It represents population counts for 6,37,848 inhabitant villages from the 2011 Census of India. Villages were located by ML Infomap, a geospatial mapping company based in New Delhi, by digitising village locations as vector points using GIS. Sources for the village mapping were from the Indian Revenue Village

Boundary maps and small-scale Census Atlas maps. Locations are verified using high resolution satellite imagery. Using the Village Census Code, the 2011 Census data was linked to the village locations.[1] Strengths of using this data for population modelling are that it has complete coverage for all of India, it is a true population count, and also includes population of children less than six years old. Limitations are that the population values are linked to discrete point locations at the centre of villages and they are from 2011, which is four–five years before the NFHS-4 data was collected.

Census of India 2011 village polygons: It also represents actual population counts for villages from the 2011 Census of India. Village boundaries were located by ML Infomap, by digitising village polygon boundaries (where available) from Indian Revenue Village Boundary maps and small scale Census Atlas maps. Locations were verified using high resolution satellite imagery.[2] Using the Village Census Code, the 2011 Census data was linked to the village locations. This data set does not include village polygons for the states of Nagaland, Mizoram, Meghalaya, Manipur, Arunachal Pradesh, and Andaman and Nicobar Islands. Strengths of using this data for population modelling are that it is a true population count and includes population of children less than six years old. Limitations of using this data are that the population values are linked to discrete point locations at the centre of villages, that they are from 2011, and not all of India is represented.

Since the village polygons do not have complete coverage of India, this data set was eliminated from our consideration. To evaluate the LandScan, AsiaPop, and village population data sets these were visualised on maps along with the PC and district boundaries (Figures 2 and 3).

The striking difference between Figures 2 and 3 is the difference in resolution between the LandScan 2016 and AsiaPop 2015 rasters. From these maps it is apparent that the AsiaPop 2015 is much better at modelling population fluctuations across short distances. Another observation is the meandering PC boundaries (black lines) cut across both

districts and villages. This evaluation led us to eliminate the LandScan population raster from consideration due to its coarse resolution. Village point locations were eliminated due to their discrete locational nature, representing all village population at one centroid point. This presents a problem where PCs cross through villages, causing all of the village population to be assigned to the PC in which the point is located, and zero village population being assigned to the remaining PCs the village overlaps with.

Therefore, we concluded that due to its high resolution that effectively models the heterogeneous nature of areas transitioning from urban to rural and its complete coverage of India and temporal harmony with the NFHS-4 data, the Asia-Pop 2015 would be the most appropriate population data set to use for our analysis. The AsiaPop 2015 is also widely used for other geospatial analyses.

In order to facilitate replication of our methodology, we further elaborate on the subsequent workflow involved in creating the indirect crosswalk to produce the PC-level malnutrition estimates:

1   The ArcGIS Intersect command was applied to the district and PC shapefiles, creating a new shapefile where each new polygon contained both the district ID and PC ID of the overlapping area.

2   Area was calculated for the new polygons using the Kalianpur 1975/India Zone IIa (EPSG:24379) coordinate system.

3   New areas were compared to the original district areas by dividing the new area by the total district area. Polygons with less than a hundredth of a percent of the original district area were deleted. These extremely small areas are "slivers" that act as noise, and are created from slight boundary inaccuracies between the district and PC shapefiles The ArcGIS Intersect command was applied to the district and PC shapefiles, creating a new shapefile where each new polygon contained both the district ID and PC ID of the overlapping area.

4   The percentage population for the intersected areas was calculated by performing an ArcGIS Zonal Statistics overlay with the intersected shapefile and the AsiaPop 2015 population raster to produce a total population per intersected polygon.

5   To determine the percentage of district population in each intersected polygon, the new population was divided by the total population of the district.

6   The district segments were matched to the NFHS-4 district level data using an ArcGIS Table Join based on the district ID.

7   Each district-level malnutrition percentage was multiplied by the total district population, to produce an estimated number of individuals in each district with a particular malnutrition state.

8   The number of individuals in each district with a particular malnutrition state was multiplied by the percentage of the district population contained in each intersected segment, producing an estimate of individuals in each district segment with a particular malnutrition state.

9   The intersected shapefile was summarised by PC ID, producing total individuals in each PC segment with a particular malnutrition state, and total population in each PC.

10   The total individuals in each PC with a particular malnutrition state was divided by the total population in each PC to produce estimates of percentage population with a particular malnutrition state for each PC.

We also take this opportunity to correct the mischaracterisations made by Goli (2019) on our indirect crosswalk methodology. Specifically, contrary to Goli's reading that the PC_District_Intersect

implicitly assumes homogeneity within a district, our methodology does not assume a homogeneous population within districts. The GIS Intersect command actually splits the geometry of the PCs and districts, creating new areas that are pieces of districts. Applying the population zonal statistics to each of these pieces accounts for the mix of urban and rural population distribution within districts. We also calculated areal percentages for the intersected data, chiefly for two reasons: (i) to identify and eliminate "sliver" polygons generated by slight boundary inaccuracies between the district and PC shapefiles; and (ii) by calculating percent area; this allows for additional, non-human data at the district level to be analysed at the PC level. For example, total forest or wetland or other natural features tabulated at the district level can be apportioned to the PC level using the area percentages.

Direct Methodology

The second method we had proposed in our original article (Swaminathan 2019) involved aggregating individual-level data to a potential PC linked via randomly displaced GPS locations of the NFHS-4 sampling clusters. We requested and downloaded the GPS cluster locations from the Demographic Health Surveys

(DHS 2016) Programme for India. The cluster locations in this data file contained a "LATNUM" field listing the cluster's latitude coordinate in decimal degrees and a "LONGNUM" field which lists the cluster's longitude coordinate in decimal degrees. Using the ArcGIS addxy command, these cluster locations were converted into a shapefile for use in GIS. In order to ensure respondent confidentiality, the DHS/NFHS randomly displaced the GPS latitude/longitude positions such that urban clusters were displaced up to 2 km and rural clusters were displaced up to 5 km, with 1% of the rural clusters displaced up to 10 km.

The displacement was restricted so that the points stay within the same district. This gave us the confidence to overlay district boundaries with each cluster to determine its district. Specifically, we performed an ArcGIS Spatial Join from each cluster location to the districts shapefile. This GIS command does a "point in polygon" test, and calculates the district each cluster falls in, saving this information into a new column of the cluster attribute table. We then performed an ArcGIS Spatial Join between the cluster shapefile and the PC boundaries. This determined which PC each cluster fell into. After determining the PC that each cluster fell into, we summarised the cluster populations to create a "sample population" for each PC. Prevalence for the malnutrition indicators was then computed, that is, number of individuals with each condition divided by the total number of individuals in the PC.

We recognise that this methodology may misclassify some clusters to fall into incorrect PCs due to the random

displacement of GPS coordinates. Figures 4 and 5 illustrate possible cluster/PC misclassifications. In Figure 4, the southernmost rural cluster point (red dots) on this map is at risk of misidentification due to inaccurate PC boundary. It falls in a Firozabad PC, but visual inspection of the location compared to the satellite imagery reveals it should be in Agra. In Figure 5, rural cluster point in Rajgarh PC is 0.4 km from the PC border with Guna, risking classification in the wrong PC.

Yet, the resulting measurement error will most likely be random when aggregated. Given the highly consistent PC estimates for child malnutrition indicators when comparing the indirect crosswalk and direct methodologies ($r = 0.92$ for stunting, $r = 0.92$ for underweight, $r = 0.84$ for wasting, and $r = 0.89$ for anaemia) (Swaminathan 2019), as well as the relative simplicity of the direct methodology, we encourage the latter approach when GPS coordinates for survey clusters are available to be linked to PC boundaries.

Sensitivity Analyses

Prompted by other issues raised by Goli (2019), we conducted the following sensitivity analyses and can conclusively say that our original findings remain robust. First, we identify PCs that share the same boundary as the districts and compare the estimated child malnutrition prevalences for PCs to the values given in the NFHS-4 district reports. Second, we use child population (0–6 years) instead of total population for the indirect crosswalk and re-estimate the PC level malnutrition indicators. Third, we apply sampling weights at the district level. Fourth, we summarise findings from our recent study (Kim et al 2019) where we incorporated precision-weighting to account for small samples. For brevity, we conduct these sensitivity analyses for stunting only. We also note that ranking and prevalence of all child malnutrition indicators for all PCs are available upon request to the authors as noted in the corrected Appendix of our original publication, that is, Swaminathan (2019).

PCs and districts with 'matching' boundaries: One way to test the validity

of our model is to analyse our malnutrition estimates in PCs that exhibit identical boundaries with a district. To find these exact matches, we compared the total area of each PC with the total area of each intersecting district. There were no PC/district combinations that exhibited a zero percent difference in area, due to the boundary inaccuracies described above. Knowing that in reality there are many PCs and districts that are identical, we visually inspected all PC/district area comparisons that exhibited less than a 4.0% difference in area. This revealed 28 PCs that were identical with a district boundary. In an ideal scenario, our estimated child malnutrition prevalence rates in these PCs should perfectly match the district-level NFHS-4 data, and indeed this is what we found (Table 1).

Table 1: Comparing PC Stunting Estimates Derived from Indirect Crosswalk and Direct Methodologies to the District Stunting Estimates from NFHS-4 (across 28 PCs/districts with identical boundaries)

PC/District with Identical Boundaries	State	PC Stunting %, Indirect Estimates	PC Stunting %, Direct Estimates	District Stunting %, NFHS Report
Araria	Bihar	48.4	48.6	48.4
Gopalganj	Bihar	35.6	36.8	35.6
Begusarai	Bihar	44.9	45.0	44.9
Arrah (Bhojpur)	Bihar	43.5	43.9	43.5
Nalanda	Bihar	54.1	54.4	54.1
Surguja	Chhattisgarh	37.3	32.3	32.3
Jamnagar	Gujarat	27.9	29.7	27.9
SabarKantha	Gujarat	50.6	51.5	50.6
Anand	Gujarat	45.3	46.9	48.2
Jamshedpur (Purbi Singhbhum)	Jharkhand	39.3	41.0	39.3
Vijayapura	Karnataka	44.9	44.8	44.9
Davangere	Karnataka	46.4	48.3	46.4
Chhindwara	Madhya Pradesh	33.6	33.5	33.6
Jajpur	Odisha	30.3	29.8	30.3
Sundargarh	Odisha	37.2	36.6	37.2
Jalandhar	Punjab	28.2	29.6	29.3
Kanyakumari	Tamil Nadu	17.2	18.2	17.2
Krishnagiri	Tamil Nadu	24.7	22.4	25.1
Thoothukudi	Tamil Nadu	21.2	21.1	21.2
Rampur	Uttar Pradesh	46.0	45.2	46.0
Shahjahanpur	Uttar Pradesh	49.3	49.4	49.3
Mathura	Uttar Pradesh	42.1	40.3	40.8
Domriaganj (Siddarth Nagar)	Uttar Pradesh	54.3	56.4	57.9
Firozabad	Uttar Pradesh	46.0	44.2	44.0
Maharajganj	Uttar Pradesh	45.9	54.6	53.3
Mirzapur	Uttar Pradesh	49.1	48.5	49.1
Fatehpur	Uttar Pradesh	52.4	52.5	52.4
Unnao	Uttar Pradesh	44.7	46.9	46.5

For stunting, the correlation between PC estimates and NFHS-4 district estimates was $r = 0.982$ for indirect crosswalk estimates and $r = 0.995$ for direct

estimates. The difference in stunting prevalence between PC estimates and the NFHS-4 district estimates was less than 1.0% for 20 out of 28 PCs/districts, with the largest difference being 7.4% for Maharajganj using indirect crosswalk estimates and 2.7% for Krishnagiri using direct estimates.

Using child population for indirect crosswalk: In our original article, the population calculations involved in the indirect crosswalk methodology represented all age groups. We had intentionally used the total population to encourage our apportioning method to be applied to other various health and development indicators that affect the general population. However, given that malnutrition indicators we had investigated are relevant for population of children only, we performed a sensitivity analysis to generate PC estimates for stunting using the 0–6 year old population from the 2011 Census. Of note, children surveyed in NFHS-4 were under five, and hence do not perfectly match with the child population defined in the census. Moreover, our model does not account for the differences in age distribution among the neighbouring districts. Nevertheless, the resulting stunting estimates from this sensitivity analysis remained identical to the estimates we had generated using the total population ($r = 0.98$) (Figure 6).

Applying sampling weights: The use of sampling weights made minimal difference at the district level, and hence our PC estimates generated from unweighted district level data are unlikely to be affected. For instance, the correlation in weighted

versus unweighted district estimates was above 0.99 for stunting (Figure 7).

Precision-weighting: In our more recent work (Kim et al 2019), we used multilevel models to generate PC estimates based on precision-weighted predicted probabilities of child undernutrition indicators at the cluster level (that is, villages in rural areas and census enumeration blocks in urban areas). This methodology is well known to provide a technically robust and efficient framework to generate small area estimations by accounting for sampling variability (Goldstein 2011; Jones and Bullen 1994). In our comprehensive assessment of PC estimates using different statistical modelling (precision-weighted versus none) and methodologies to identify PC membership (direct versus indirect crosswalk), we found very high consistency with r = 0.92–0.99 for stunting (Kim et al 2019).

Validation of Methodologies

A formal validation of our PC estimates for child malnutrition indicators necessitates census data on anthropometry and haemoglobin measures of all children in India linked to PC identifiers. In the absence of such data, we sought to validate our methodologies by applying them to a key developmental indicator, and a strong predictor of child malnutrition, for which census data are available. We have a sense of "ground truth" on the extent of female literacy, which can be aggregated to the PC level, from the 2011 Census. The NFHS-4 also collected information on literacy for all surveyed women, enabling us to generate PC estimates by using the proposed indirect crosswalk and direct linkage methodologies. Comparison of our estimates to the PC-level proportion of literate females from the census indicated an incredibly high correlation of r = 0.96–0.97 for both indirect crosswalk and direct methodology (Figure 8). Although there are some differences between the census data and NFHS-4 survey in terms of population coverage (all females older than six years in the census versus 15–49 year olds surveyed in NFHS-4) and time of data collection (2011 for census versus 2015–16 for NFHS-4), this exercise clearly

demonstrates the validity of our proposed methodologies to generate PC level data.

In Conclusion

The analyses presented here further support that the two methodologies proposed in our earlier articles (Kim et al 2019; Swaminathan 2019) provide a robust assessment of child malnutrition at the PC level. It is well-recognised that monitoring data on population health and well-being at the PC level is important to increase political accountability and to effectively design and evaluate policies and programmes. In the absence of identifiers for PCs in the current surveys and census data, we present two realistic methodologies using GIS that produce robust PC-level estimates given the currently available data in India. We have further elaborated on the two methodologies to aid application of our work on other diverse indicators of population health and development. We are optimistic that this increased awareness of PC-level data will lead to better policy decisions and overall leadership among PCs in India.

REFERENCES

DHS (2016): "India: Standard DHS, 2015–16," DHS Program: Demographic and Health Surveys, https://www.dhsprogram.com/data/dataset/India_Standard-DHS_2015.cfm.

Gaughan, A E, F R Stevens, C Linard, P Jia and A J Tatem (2013): "High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015," *PLoS One,* Vol 8, No 2, e55882, doi: 10.1371/journal.pone.0055882.

Goldstein, H (2011): *Multilevel Statistical Models,* United Kingdom: Wiley.

Goli, S (2019): "Unreliable Estimates of Child Malnutrition," *Economic & Political Weekly,* Vol 54, No 6, pp 64–67.

Jones, K and N Bullen (1994): "Contextual Models of Urban House Prices: A Comparison of Fixed- and Random-Coefficient Models Developed by Expansion," *Economic Geography,* Vol 70, No 3, pp 252–72.

Kim, R, A Swaminathan, R Kumar, Y Xu, J C Blossom, R Venkataramanan, A Kumar, W Joe and S V Subramanian (2019): "Estimating the Burden of Child Malnutrition across Parliamentary Constituencies in India: A Methodological Comparison," *SSM-Population Health,* Vol 7, No 100375, doi:10.1016/j.ssmph.2019.100375.

Swaminathan A, R Kim, Y Xu, J C Blossom, J William, R Venkataramanan, A Kumar and S V Subramanian (2019): "The Burden of Child Malnutrition in India: A View from Parliamentary Constituencies," *Economic & Political Weekly*, Vol 54, No 2, pp 44–52.
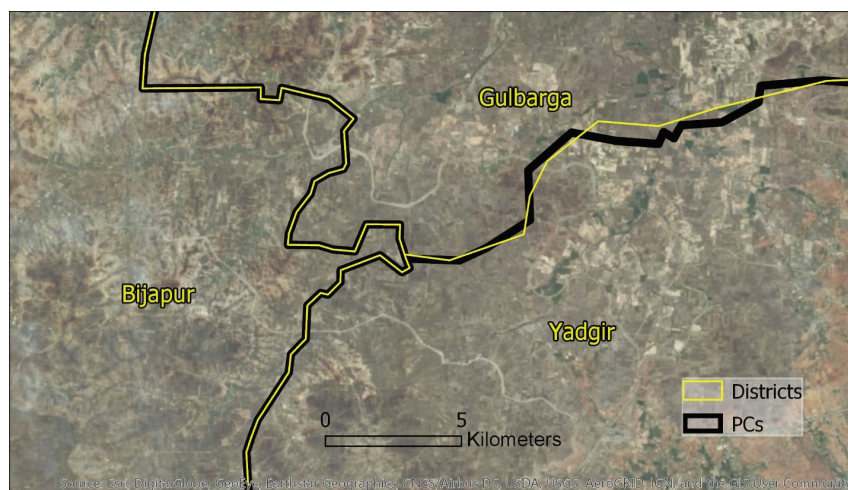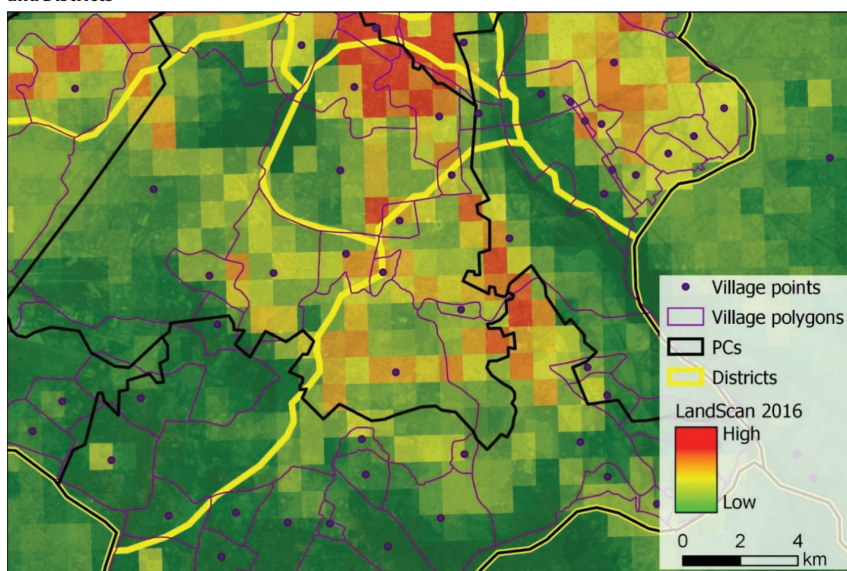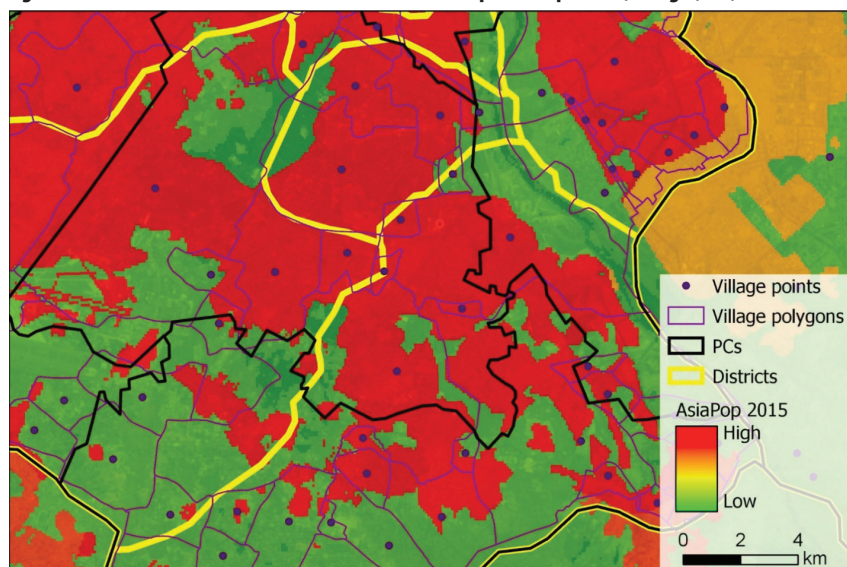
Figure 1: District and PC Boundary Comparison

Figure 2: An Area in Southern New Delhi Shown with LandScan 2016 Population, Villages, PCs, and Districts

Figure 3: An Area in Southern New Delhi Shown with AsiaPop 2015 Population, Villages, PCs, and Districts

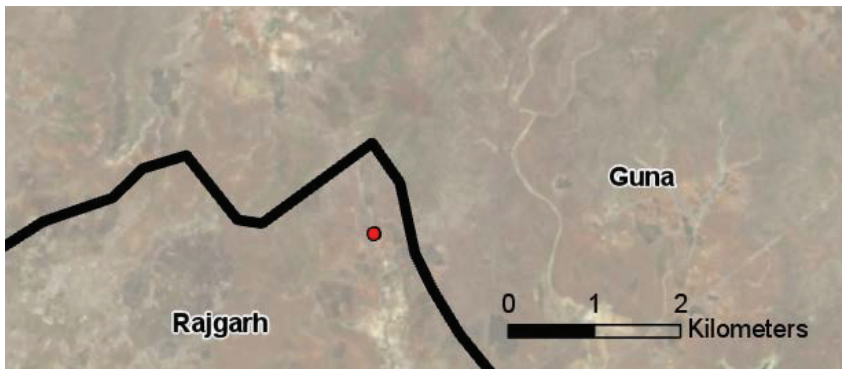**Figure 4: Exemplification of Potential Misclassification of a Cluster and PC Using Direct Methodology (Case 1)**

Figure 5: Exemplification of Potential Misclassification of a Cluster and PC Using Direct Methodology (Case 2)

Figure 6: Correlation between Indirect Crosswalk Estimates Stunting Using Child Population (<6 Years) and Total Population
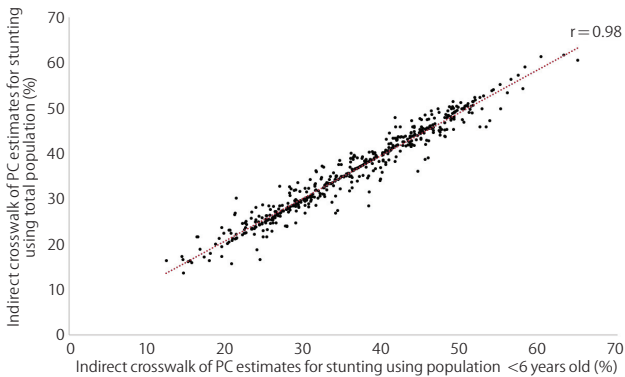

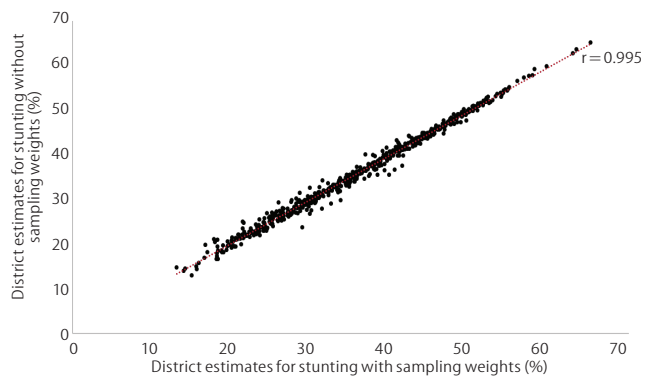**Figure 7: District Estimates for Stunting before and after Applying the Sampling Weights**

Figure 8a: Validation Test Comparing PC Female Literacy from Census 2011 to Indirect Crosswalk Estimates
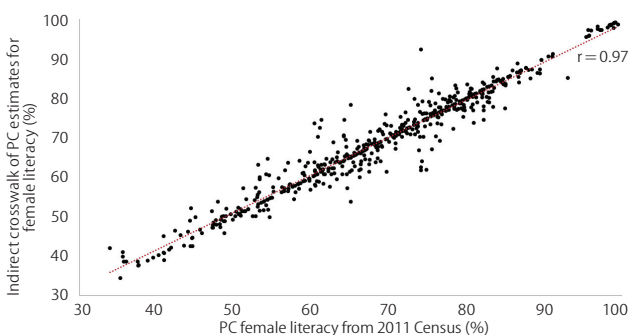

Figure 8b: Validation Test Comparing PC Female Literacy from Census 2011 to Direct Methodology