A method for spatial estimation of agricultural dependence

Using spatial disaggregation to identify agricultural populations in India.

**Joe Post**

August 2023

Supervisor: Dr Fulvio Lopane

Word count: XXXX

[Github Link](https://github.com/joepost/india_adp)

A Dissertation submitted in part fulfilment of the

Degree of Master of Science:

Urban Spatial Science

Centre for Advanced Spatial Analysis

The Bartlett Faculty of the Built Environment

University College London

## Abstract

/

## Acknowledgements

/

## Declaration

/

Table of Contents

[Abstract 2](#_Toc143291892)

[Acknowledgements 2](#_Toc143291893)

[Declaration 2](#_Toc143291894)

[1. Introduction 5](#_Toc143291895)

[1.1 Research Question 6](#_Toc143291896)

[1.2 Agricultural Dependent Population 6](#_Toc143291897)

[1.3 Indian context 7](#_Toc143291898)

[2. Methodology 9](#_Toc143291899)

[2.1 Spatial Disaggregation 9](#_Toc143291900)

[2.2 Justification 12](#_Toc143291901)

[2.3 Data Sources 12](#_Toc143291902)

[2.4 Study setting 14](#_Toc143291903)

[2.5 Computing Agricultural Dependent Population 14](#_Toc143291904)

[2.6 Validation of population estimates 16](#_Toc143291905)

[3. Results 17](#_Toc143291906)

[3.1 Comparison of methods 17](#_Toc143291907)

[3.2 Buffer iteration 19](#_Toc143291908)

[3.3 Scale method to India 19](#_Toc143291909)

[4. Discussion 22](#_Toc143291910)

[4.1 Establishing method parameters 22](#_Toc143291911)

[4.2 Interpreting buffer radius 22](#_Toc143291912)

[4.3 Limitations 22](#_Toc143291913)

[4.4 Utilising these findings in practice 23](#_Toc143291914)

[Appendix 24](#_Toc143291915)

[References 25](#_Toc143291916)

## Introduction

Agriculture is the single largest employer across the globe, as the source of income for 40 per cent of the world’s population (Kondylis *et al.*, 2023). In India, this share is even larger, with 52% of workers estimated to be dependent on agriculture for a living, rising to 70% in rural households, and predominantly in small and subsistence farms (Census of India, 2011; FAO, 2023). Agricultural populations in India typically face high rates of poverty and instability, and are identified by the World Bank as a key target for development funding (World Bank, 2023), especially in the context of increasing vulnerability due to the effects of climate change and increased variability of temperature and rainfall (Anand, Kakumanu and Amarasinghe, 2019).

To support effective, context-specific development, it is necessary to understand the spatial distribution of this agricultural population. In contrast to estimations of total population, which skews towards urban areas, rural and agricultural populations can provide an indication of demand on specific resources, such as water for irrigation, and of heightened vulnerability to drought, disaster events, and climate change (Vanthof and Kelly, 2020). Conversion of demographic data to a uniform spatial grid format has become a well-established method for reporting spatial distribution. An advantage of this output is that most earth observation data used to monitor climate and resources are produced in a grid (raster) format. By reporting demographic information in a comparable gridded format, rather than in large and spatially heterogeneous administrative divisions, it can be integrated with earth observation data to model human-environment interactions and dimensions of risk, access, need, and more (Freire *et al.*, 2020).

The United Nations Sustainable Development Goals (SDGs) Report explicitly acknowledges that “[t]imely, high-quality and disaggregated data”, in the context of broader statistical development, is a crucial component of delivering the SDGs and of supporting development work (United Nations, 2022, p. 3).

This study aims to produce a spatially disaggregated estimate of the agricultural population across India. It extends upon existing methodologies used to estimate total population and applies this to a specific demographic subset. The case study of India is designed to assess feasibility and performance of the methodology at a large spatial scale, comparative to partner research testing proof-of-concept in districts of Sri Lanka (unpublished). [Flesh out]

This introduction is broken into three parts. First, the research question and objectives for the study are presented. The second section provides an overview of the concept of agricultural dependent population, the implications of deriving agricultural populations from census or alternative data sources, and how this concept is relevant to research and development work in the case study context of India. The last section introduces spatial disaggregation methodologies, historical development, and applications, particularly regarding gridded population estimates of the world. The review is intended to highlight how this thesis addresses a gap in the literature and how the work is situated within the broader scholarship around spatial disaggregation of data.

### Research Question

Agricultural populations in low-income countries have unique development needs and face a set of risks in a changing climate future that are different from the population as a whole. Although there is a significant body of research globally on the development and estimation of spatially disaggregated population counts, this work thus far has not addressed the important subset that are the agricultural dependent population. Therefore, this study aims to answer the research question,

*How can the agricultural population in India be identified at a small area scale?*

To respond to this, the three objectives of this study are to:

1. Review existing methods for spatial disaggregation of demographic data,
2. Propose and evaluate a new method that combines dasymetric disaggregation and iterative extension (buffers),
3. Scale the method up to estimate the small area agricultural population for all of India, and
4. Assess the level of uncertainty in agricultural population estimates, and associated factors.

### Agricultural Dependent Population

The concept of an agricultural population and agricultural dependence is referenced somewhat often in the literature, but rarely is the topic addressed directly. Zarkovich *et al.* (1976) decades ago explored the statistical challenges of defining agricultural populations, in the context of enumerating agricultural labourers and landholders residing in urban areas, and the inverse challenge of accounting for farmland residents who do not participate in agricultural labour. Other studies have highlighted the complexity of gender, noting that women’s work in farming (often unpaid) has historically been systematically underestimated in labour force statistics (Dixon, 1982), and that female participation in agricultural labour is increasing in the context of male outmigration from the sector (Pattnaik *et al.*, 2018; Slavchevska, Kaaria and Taivalmaa, 2019). However, there is no universal definition of agricultural population or how it should be calculated.

In this project, agricultural dependence is primarily understood from the lens of employment in the primary sector. As recorded in the 2011 census, any worker who participated in agricultural labour in the preceding year is considered, with or without their dependents. See section 2.5 for a detailed explanation of how agricultural dependence has been calculated.

A more rigorous exploration of how best to define agricultural dependent population is outside the scope of this study. However, it is noted that further research into developing a systematic definition of agricultural dependence would improve the transferability of the findings, and for comparison of findings across different settings, by ensuring that interpretation of the term is consistent across studies.

### Indian context

India, the subject of this study, is one of the world’s largest countries by area, the third-largest economy, and is expected to become the most populous country before the end of 2023 (United Nations in India, 2022). [Morph this paragraph into an explanation of why India has been chosen for case study?] The nation has several advantages as a case study for this research. Firstly, India is a major target for development funding and has a , and provides …

The nation is divided into 28 states and 8 union territories, each of which are further subdivided into districts and smaller administrative divisions, variously termed *tehsil*, *taluka*, or *mandal* (Government of India, 2012). Census data for most socioeconomic indicators, including labour statistics required to calculate agricultural dependence, are published at the district level, the most recent being conducted in 2011. Districts vary significantly in size, population count, and population density, but on average cover just under 5,000 km2 and a population of 2 million (Table 1).

Table : Characteristics of 2011 Indian Census Districts (n = 640)

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Population | Area (km2) | Population Density (km2) |
| Mean | 1,891,961 | 4,948 | 936 |
| Minimum | 8,004 | 9 | 1 |
| 25th percentile | 817,861 | 2,297 | 207 |
| Median | 1,557,367 | 3,798 | 373 |
| 75th percentile | 2,583,551 | 6,235 | 719 |
| Maximum | 11,060,148 | 45,674 | 36,155 |

In census enumeration, urban areas are divided into four classes – wards, outgrowths, statutory towns, and census towns, the latter being legally rural settlements that have been designated as urban. The 2011 census estimates that 31% of India’s population reside in urban areas, however this is predicted to be a significant underestimate (Balk *et al.*, 2019). Classification of urban versus rural has implications for the estimation of population based on land cover classification. Because the census designates urban/rural status using an administrative method, which is not stable across census collections and is not systematic across the country, this factor was omitted from validation calculations. Rural classification derived from remotely sensed land cover data has been prioritised instead, as discussed in Section 2.3.

In India, particularly in southern states of Andhra Pradesh, Tamil Nadu, and Karnataka, small scale irrigation has historically been managed through tank systems – traditional water storage reservoirs designed to harvest and store rainwater and surface runoff (Mialhe, Gunnell and Mering, 2008). In many areas, as farming practice has increasingly transitioned to a reliance on groundwater extraction, these tanks have become degraded and are not functioning at their peak (Anand, Kakumanu and Amarasinghe, 2019). Rehabilitation of these degraded tanks is a relatively cheap and effective way to improve water security for agriculture in local communities, and improved irrigation can benefit cropping intensity and subsequently reduce pressure on forest cover being converted into cultivated land (Meiyappan *et al.*, 2017). Locating which tanks are in areas of high demand (high agricultural population) provides an evidence base to direct development efforts in areas to maximise impact.

## Methodology

*Add in intro text for the methodology section; what will be covered, how it ties the narrative from introduction section. E.g. Section 2.1 (Spatial disaggregation), 2.2 (Justification), 2.3 (Presentation).*

### Spatial Disaggregation

Spatial disaggregation is a broad term which applies to the process of transforming data from a set of source zones into target zones, such as a raster grid, at a finer level of spatial resolution. There is considerable interest in the process across both academic literature and in policy, particularly applied to estimating resident population at fine spatial scales, as this has important implications for service planning and delivery (Deichmann, 1996), disaster preparation and response (Schneiderbauer and Ehrlich, 2005), monitoring international development goals (Tuholske *et al.*, 2021; United Nations, 2022) and the implementation of public health interventions (Viel and Tran, 2009; Tatem, 2022), among others.

On a global scale, the spatial disaggregation of administrative census data has been used to develop world gridded population estimates, providing regularly sized comparable population estimates across regions. Early iterations of this, such as the Gridded Population of the World (GPW) version 1 (Tobler *et al.*, 1997), have inspired a variety of contemporary global grid models, each utilising a specialised methodology and with particular strengths and limitations, bolstered by advances in computational power and the availability of high quality census and earth observation data (Wardrop *et al.*, 2018). Table 2 provides a summary of current global grid models and their key characteristics.

Table : Selected World Population Grid Datasets, adapted from Leyk et al. (2019)

| Dataset | Source | Method | Spatial Resolution | Ancillary data layers |
| --- | --- | --- | --- | --- |
| Gridded Population of the World (GPW) | CIESINa | Areal weighting | 1km | Water bodies |
| Global Human Settlement Layer – Population (GHS-POP) | JRCb and CIESINa | Dasymetric | 250m | Built structures |
| WorldPop | University of Southampton | Statistical/ Dasymetric | 100m, 1km | Roads, Land cover, Built structures, Urban areas, Night-time lights, Infrastructure, Climate, Topography, Elevation, Water bodies |
| LandScan Global | ORNLc | Smart interpolation | 30 arcsec | Roads, Land cover, Built structures, Urban areas, Infrastructure, Climate, Topography, Elevation, Water bodies |
| World Population Estimate (WPE) | Esri | Dasymetric redistribution | 150m | Roads, Land cover, Urban areas, Water bodies |

a Centre for International Earth Science Information Network; b Joint Research Centre of the European Commission; c Oak Ridge National Laboratory.

The most straightforward method of spatial disaggregation is areal weighting, where data from the source zone (such as the total population of a district) is evenly distributed across the gridded cells within it. Areal weighting benefits from low computational power and no requirement for ancillary data. However, this approach assumes that populations are evenly distributed across administrative regions, which is rarely the case (Qiu *et al.*, 2022), necessitating the development of more complex models which can incorporate knowledge from additional sources. A common approach is dasymetric mapping, which divides the area into homogenous zones based on the variable of interest (Eicher and Brewer, 2001). For example, remote sensing data can be used to identify water bodies and other non-inhabitable areas within a zone, and limit these cells to a value of zero, creating a ‘mask’. Population count can then be proportionally distributed across the non-zero cells, to produce a more accurate estimate of the real spatial distribution of population. This method is referred to as binary dasymetric mapping or binary masking (Qiu *et al.*, 2022). WorldPop products refer to variations of this method as *top-down constrained*, where population is distributed only across grid cells identified as containing built settlements, as opposed to *top-down unconstrained*, where population is distributed across all cells (Stevens *et al.*, 2015).

\*Add section discussing the limitations of existing applications, and findings from review papers (e.g. Comber 2019)

An alternative method is the incorporation of pycnophylactic interpolation, or the ‘mass-preserving’ property, which requires that the sum of pixel estimates is equal to the supplied population of the source zone or polygon (Malone *et al.*, 2012). Tobler (1979) described the process of pycnophylactic smoothing, where the weighted average of a pixels neighbours is used to iteratively smooth the population values in grid cells whilst ensuring the mass-preserving property is met, as a method to lessen the effect of sharp changes in population density estimates at the boundaries of source zones (figure 1). This approach relies on the assumption of Tobler’s ‘First law of geography’, that near things are more related than things that are far apart (Tobler, 1970). In most cases, when compared directly, dasymetric methods or hybrid methods combining dasymetry and pycnophylactic interpolation outperform simple areal weighting (You and Wood, 2006; Monteiro, Martins and Pires, 2018).

A picture containing screenshot, colorfulness, cube, design

Description automatically generated

Figure : Illustrative diagram of pycnophylactic interpolation, from Deichmann (1996, p. 33)

Alternative to these ‘top-down’ methods of spatial disaggregation, ‘bottom-up’ approaches for small area estimation can be used to produce gridded population estimates. These methods are designed for estimating population distribution in the absence of high-level source data such as a national or regional census, or when such data is out-of-date or known to be inaccurate (Wardrop *et al.*, 2018). However, bottom-up estimation requires the completion of tailored geo-located field surveys, and are generally viewed as complementary to traditional census enumeration in data-poor settings (Leyk *et al.*, 2019).

Although there has been extensive research and methodological development in the field of population disaggregation, there are fewer studies that extend these methods to estimate additional demographic or socioeconomic characteristics beyond population count or density, despite the methodologies being broadly similar. An early study by Eicher and Brewer (2001) showed the potential for dasymetric mapping to map age structure and housing value in the United States, and more recently novel data sources have been utilised, such as Point of Interest property data in Singapore (Szarka and Biljecki, 2022), to estimate elderly populations at the neighbourhood scale. The WorldPop research unit regularly produce national and regional gridded maps across health and social indicators, such as vaccination coverage, and in the Indian context produced an ‘atlas’ of 19 indicators nationwide at a 5km resolution (Pezzulo *et al.*, 2023). In assessing agricultural populations, only one relevant study was identified, which estimated the proportion of primary sector labourers at the parish level in Portugal using a hybrid method of dasymetric mapping and pycnophylactic interpolation (Monteiro, Martins and Pires, 2018) adapted from work by Malone *et al.* (2012).

The spatial disaggregation of population into fine spatial scales is a rich field of research and has benefitted from extensive methodological development and innovation alongside improved data quality through censuses, surveys, and increasingly available earth observation data. However, there is a clear gap in applying these methods to estimate agricultural populations, particularly in the context of developing regions such as rural India, where such data can provide an indication of the local water demand and development need.

### Justification

*After introducing the concepts and alternative approaches in LR sections above, introduce the chosen method for my study and explicitly justify why, linking to the discussion previous.*

[Description of analysis tools and environment – python version, packages, computer RAM] All analysis was conducted using Python v3.10.9. Details of packages used, the python environment, and the code can be accessed from the Github repository for the project (<https://github.com/joepost/india_adp>).

### Data Sources

There are four key sources of data that form the input for this analysis: the Dynamic World land cover dataset (DW), the Global Human Settlement Layer – Settlement Model Grid (GHS-SMOD), the WorldPop gridded population estimates, and tables from the Indian Census 2011.

Dynamic World is a global-scale, high resolution (up to 10m), land use land cover (LULC) dataset that is freely released as a Google Earth Engine Image Collection, available up to near real-time and historically from 2015 onwards (Brown *et al.*, 2022). The dataset is trained using semi-supervised deep learning from Sentinel-2 imagery and classifies pixels to 1 of 8 types: Water, Trees, Grass, Flooded Vegetation, Crops, Shrub & Scrub, Built Area, Bare Ground, and Snow & Ice. For this study, the DW layer was extracted from Google Earth Engine as a composite image aggregated over the period 1st January 2020 to 1st January 2021, selecting the most frequently occurring class label for each pixel over the specified period. The script used to extract DW data from Earth Engine can be accessed on the Github repository. [Add section on why the 100m resolution chosen].

The Global Human Settlement Layer is a set of several datasets that present the spatial distribution of urbanisation and human presence across the world, developed by the Joint Research Centre of the European Commission. The GHS-SMOD is an extension of the settlement layer that applies the Degree of Urbanisation methodology (Eurostat, 2021) to classify pixels into an urban/rural typology on the basis of population density, size, and contiguity () at a 1km spatial resolution in 5-yearly epochs. For this study, the three rural classes in addition to suburban/peri-urban areas were combined into a single class, labelled ‘rural’, for use as a mask for population data. GHS-SMOD data was downloaded for the year 2010 from the European Commission GHSL website (<https://ghsl.jrc.ec.europa.eu/ghs_smod2023.php>), to align closest with the 2011 Indian Census data.

Table : Global Human Settlement Layer – Settlement Model Grid (GHS-SMOD) Classification Rules (Schiavina, Melchiorri and Pesaresi, 2023)

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Class | Population Density (km2) | Definition |
| 30 | Urban Centre | >1,500 | Contiguous grid cells (4-connectivity) that has at least 50,000 inhabitants in the high-density cluster. |
| 23 | Dense urban cluster | >1,500 | Contiguous grid cells (4-connectivity) that has at least 5,000 inhabitants and less than 50,000. |
| 22 | Semi-dense urban cluster | 300 – 1,500 | Contiguous grid cells (8-connectivity) that has at least 5,000 inhabitants in the cluster and is at least 3km away from other urban clusters. |
| 21 | Suburban or peri-urban | 300 – 1,500 | All other cells that belong to an urban cluster that do not meet the criteria for Urban centre, Dense, or Semi-dense urban cluster. |
| 13 | Rural cluster | <300 | Contiguous grid cells (8-connectivity) that has at least 500 and less than 5,000 inhabitants in the cluster. |
| 12 | Low density rural | 50 – 300 | A cell with more than 50 inhabitants that is not part of an urban or rural cluster. |
| 11 | Very low density rural | <50 | A cell with less than 50 inhabitants that is not part of an urban or rural cluster. |
| 10 | Water | - | Cells where more than 0.5 share covered by permanent surface water that are not populated nor built. |

Gridded population data were downloaded from WorldPop as 1km resolution United Nations adjusted estimates for 2011, using an unconstrained top-down method. This method uses national administrative and census datasets as a ‘ceiling’ from which small area estimates are disaggregated using Random Forest machine learning modelling (Stevens *et al.*, 2015). An unconstrained method estimates population count over all land squares globally, in contrast to a constrained model which applies a mask to restrict population estimates to only grid cells that are predicted to contain built settlements. For this study, an unconstrained method was chosen under the assumption that global settlement datasets may not capture all potential built areas in sparsely populated and rural areas (Freire *et al.*, 2020), especially in the context of a low-data setting such as rural India.

[Introduce India census 2011 Data Source] Tables from the Indian Census 2011 were used to calculate total population and agricultural dependent population at the state and district level.

### Study setting

[Paragraph concept: explain that Karnataka used as a test state] As outlined in Objective III, this study aims to estimate the agricultural dependent population of India at a small area scale. A single test state, Karnataka in Southern India, was selected to trial the methodology for Objective II, comparing computation time at different spatial resolutions and performance results for each ADPC estimate. This methodology was then scaled up to estimate the ADP across all states.

[Add paragraph introducing triptych figure; short description of each] Karnataka is one of the largest states in India, with a 2011 population of more than 60 million people spread across 30 districts and 192,000 km2. The state was chosen as a test state for analysis due to its large area and population, covering a diverse landscape from coast to interior, and a population density similar in scale to Sri Lanka where the prototype of this study has been conducted. WorldPop, DW, and GHS-SMOD extracts for Karnataka are shown in Figure [2]. The state capital of Bengaluru (formerly Bangalore), a metropolitan area of approximately 11 million people, is clearly visible in the southeast corner of the state, and the urban region takes up close to the entirety of the district area.

### Computing Agricultural Dependent Population

To respond to Objective II, a novel approach to estimating small area spatial distribution of agricultural population has been proposed and tested. First, classified land cover imagery (DW) was used to create a binary mask of cropland for each district in India. Gridded population estimates from WorldPop were then joined to cropland areas, to produce a base estimate of ADP that encompasses all inhabitants in crop landcover – this is referred to as the aggregated ADP, or ADPA.

A district-level estimate of ADP was separately calculated from Indian Census data, using a combination of total population counts and count of employment by industry – referred to as the census ADP, or ADPC. As discussed above, ADP itself is a broad concept that is not well defined in the literature. Therefore, a series of 5 alternative ADPc estimates were calculated, as outlined in Equations 1 to 5, to evaluate variation depending on definition.

[1]

[2]

[3]

[4]

[5]

In Indian census collections, labourers are divided into one of two employment classes: main or marginal. Main workers receive their primary source of income, or are employed predominantly, in a given industry sector. Marginal workers receive some income from a given industry but work in that industry for less than 6 months overall in the census year. ADPC1 and ADPC2 assume that only main workers, who are primarily employed in agriculture for more than 6 months in a year, should be accounted as agriculture dependent. Alternatively, ADPC3 and ADPC4 account for both main and marginal workers as agriculture dependent. Due to the often-seasonal nature of agricultural work, it is reasonable to assess that many labourers in the sector may be classed as marginal whilst still being functionally dependent on the work for their livelihood (REF).

Within the agricultural sector, workers are divided into three classes: Cultivators, Agricultural Labourers, and Primary Sector Other (including plantation, livestock, forestry, fishing, hunting and allied activities). ADPC1 and ADPC3 are designed to include only agricultural workers who are employed in cropland cultivation. Defining agriculture dependence as cropland dependence is logical when using cropland LULC data as a mask for the spatial distribution of ADP. For comparison, ADPC2 and ADPC4 account for all workers within the agricultural sector.

Lastly, ADPC5 is designed to account for the significant non-working population who are not captured in the other estimates. The count of main and marginal cropland workers is multiplied by the ratio of total workers to total population, under the assumption that the ratio of dependents to workers is roughly equivalent in the agricultural sector as in the total population.

### Validation of population estimates

The aggregated population estimate, ADPA, was summarised at the district level and compared to district-level ADPC estimates. Where the difference between estimates as a proportion of total population exceeded ±5%, an iterative buffer process was implemented to enlarge or reduce the size of the mask area containing the agricultural population. This process assumes that, where an agricultural population is not entirely captured within the cropland area, the rural population in adjacent non-cropland areas are the most likely source of agricultural labour. A projected coordinate reference system (CRS) of EPSG:9999 was set for all distance and area calculations.

To ensure that increasing buffers do not encompass adjacent urban areas, where estimates would be influenced by high counts of inhabitants that have a low likelihood of working in the agricultural sector, only rural population points are included in the buffer calculation. Rural population points were calculated by joining gridded population estimates from WorldPop to rural and peri-urban areas derived from the GHS-SMOD layer. Buffers were implemented at 50m distance around cropland polygons and the ADPA recalculated for this area. This process was repeated until the difference between ADPA and ADPC was less than ±5% for all districts within a state.

Each calculation was performed at the district level, as this is the smallest area scale at which administrative data is available. Buffers have therefore been restricted to district administrative boundaries, to ensure that ADPA calculations only account for population within the district of analysis, to align with the population used for validation. Districts that have no rural and/or cropland area, such as those containing major cities, were removed from analysis. A list of these districts is shown in Table [X].

[Insert logic pathway (possibly flowchart)? See page 35 of Mahfouz thesis]

## Results

From Adam: suggestion to structure whole paper around 5 or 6 key figures. These then become the focal point for discussion.

Table : List of proposed results (figures/tables) for inclusion

|  |  |  |  |
| --- | --- | --- | --- |
| Subsection | Fig. | Description | Linked Objective |
| 3.0 Raw inputs (Potentially incorporate into methodology) | t3.01 | Table of population characteristics of all census districts in India |  |
| f3.01 | Triptych of Karnataka showing rural areas (GHSL), cropland areas (DW), and population distribution (WorldPop) 🡪 See . |  |
| t3.02 | Table of Karnataka districts, showing district size, population, pop density, WorldPop estimate, base difference (WorldPop vs census) |  |
| 3.1 Comparison of methods | f3.11 | Plot comparing performance of ADP1-5 🡪 distribution plot (boxplot/violin plot?) | Objective II |
| t3.11 | Paragraph (or matrix) presenting results of computation time analysis for test state of Karnataka | Objective II |
| f3.12 | Table comparing ADP estimates at 100m and 1km input; analysis of result performance, compared with computation time figure above  (if possible, in time frame) | Objective II |
| f3.13 | Point plot of Karnataka districts, showing ADPA, ADPC, and difference | Objective IV |
| 3.2 Buffer results | t3.21 | Table of districts ineligible for buffer (ADPc5 > WorldPop rural), and their characteristics | Objective II & IV |
| f3.21 | Choropleth map of buffer radius by district | Objective II |
| f3.22 | Scatterplots of buffer radius by (i) APDa, (ii) ADPc, (iii) Total Population, (iv) Rural area, (v) Cropland area | Objective IV |
| f3.23 | Spatial autocorrelation of buffer radius by district | Objective IV? |
| 3.3 Scale method to India | f3.31 | Faceted distribution plot comparing ADPA, ADPC and difference (parallel to f3.13) | Objective III |
| f3.32 | Raster map of India showing ADP built from the cropland + buffer area, coloured by cell population count | Objective III |
| f3.32b | Raster map of ADP built from 1km inputs (to compare results to 1km:100m output) | Objective II & III |
| f3.33 | Plot of buffers distribution by state (if possible, in time frame) | Objective III |

[Introduction text on breakdown of Results chapter – sections 3.1-3.3]

### Comparison of methods

Due to the availability of input data at different spatial resolutions and alternative methods for estimation of ADP from Census data, variations of the overarching method were tested on Karnataka and the results compared. For ADP calculations, the difference between ADPA and ADPC was compared for each of ADPC1 to ADPC5. For spatial resolution, performance was measured by computation time and ADPA/ADPC variation.

Of the five variations of ADPC estimate, four produced a mean negative result – indicating that for most districts, the population residing within cropland areas exceeded the estimated agricultural population for the district overall. For ADPC5, the reverse was true – the total population within cropland areas was less than the estimated agricultural population for the district overall, on average (Figure 2). This difference is logical, as ADPC5 is intended to capture the broader agricultural population (inclusive of the dependents and families of agricultural labourers) and is therefore comparatively larger, whereas the four initial models account for only the labourers themselves. ADPC5 also exhibited the lowest variance of the five models (standard deviation = 17.8).

A diagram of a graph

Description automatically generated

Figure : Distribution of ADPA district estimates against census results, by ADPC calculation method, Karnataka. Percentage difference is calculated the difference in percentage points between ADPC (census-estimated agricultural dependent population) as a proportion of total population and ADPA (aggregated agricultural dependent population) as a proportion of total population. Each point represents a single district.

[Paragraph on computation time results] Of the three raster input datasets, DW is available at 10m, 100m, and 1km; WorldPop is available at 100m and 1km; and GHS-SMOD is available at 1km resolution. Computation time was tested across combinations of 1km and 100m input for DW and WorldPop. As expected, 1km:1km (DW:WorldPop) performed the fastest at [X time], followed by 1km:100m [X time], 100m:1km [X time], and 100m:100m [X time]. Although 100m:100m would have the highest accuracy, the exponential increase in computation load rendered it infeasible, within the bounds of this study, for scaling to the national level. Where 100m:1km and 1km:100m had roughly equivalent computation time, investigation found that reducing DW to a 1km resolution led to a large components of cropland area being removed.

[Add discussion of use of feather files to improve speed of processing? I/O operations]

Example of computation time matrix (t3.12):

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source |  | Dynamic World | |
|  | Spatial resolution | 1km | 100m |
| WorldPop | 1km | Mean time | Karnataka: 30 minutes + 90 minutes |
| 100m | Mean time | Mean time |

### Buffer iteration

[Introduce results of buffer process for Karnataka]

### Scale method to India

Text

A screenshot of a graph

Description automatically generated

[Insert Figure of ADP box plots; consider adding legend to boxplot that specifies a shorthand definition of each ADP]

A green and black map

Description automatically generatedA pink and black map

Description automatically generatedA map of a city

Description automatically generated

Figure : Placeholder for triptych of (a) Population, (b) Cropland, and (c) Rural areas, State of Karnataka, India.

## Discussion

Introduction text for section; breakdown + summary

### Establishing method parameters

[Paragraph topic: Why was ADPc5 selected for the scaled method?] ADPC5 selected as the most appropriate method to scale.

[Discussion points] Computation time and considerations of input resolutions.

### Interpreting buffer radius

[Paragraph topic: outlining the major assumptions of the methodology, and how this impacts interpretation] Be clear on the 2 major assumptions of the methodology and results:

1. The agricultural dependent population are geographically proximal to cropland areas, where they source their livelihoods.
2. The agricultural dependent population do not reside in high density urban areas, as these areas are unlikely to be proximal to cropland (point 1). Realistically, it is acknowledged that a fraction of ADP may reside in high density urban areas, but this fraction is so small that it does not meaningfully impact the analysis.

[Paragraph topic: Interpreting buffer radius] How should we interpret different buffer radius? Comes back to the central assumptions above; a large radius in either direction (positive or negative) challenges the validity of these assumptions. If a radius is very large, it suggests that the ADP travel long distances to their place of work and are therefore not proximal to cropland (to a certain extent). If the radius is very subtractive, it suggests that population on the edge of cropland zones are not ADP, and that there are therefore large areas of cropland where the population is systematically non-agricultural. Neither of these are likely to be true, and so interpretation must be cautious.

Should also look for case examples of areas with a large radius close to large urban centres; are there districts where a buffer skips over or ‘engulfs’ an urban area? What would this suggest?

Link results to action – how are the results useful for the stated goals? What would they mean for policy makers?

### Limitations

[Discussion points]:

* Proxy calculation for ADPC5; could improve this by finding the average household composition or number of dependents for agricultural labourers (as opposed to the crude ratio of workers to non-workers), Could further break this down to improve accuracy: is the ratio different for main/marginal workers? Is the ratio different for different age groups (likely)
* Extending on point above; the purpose of this study is not to interrogate the definition of ADP. Rather, a method is tested and presented that is open to modification in the presence of a more sophisticated definition or calculation of ADP.
* Add paragraph more generally on the limitations of population data (see Freire et al. 2020); population statistics suffer from uncertainty that is often poorly understood or overlooked. These errors can propagate and impact the results of downstream analyses.
* Temporal alignment: DynamicWorld is only available as far back as 2015, whereas the other datasets are aligned with the Census year (2011). This difference likely to have some effect, due to the rapid scale of growth/development in India as a whole – would lead to considerable differences in cropland areas over the four-year period.

### Utilising these findings in practice

Understanding where people live, and the social and economic characteristics of those populations, is core to providing adequate, efficient, and targeted services and investment.

## Appendix

TEST: Formatting of figure with table side-by-side

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A diagram of a graph  Description automatically generated  Figure : Distribution of ADPA district estimates against census results, by ADPC calculation method, Karnataka. Percentage difference is calculated the difference in percentage points between ADPA as a proportion of total population and ADPC as a proportion of total population.  TO CONSIDER: Add map of India (labelled states/Union territories) and map of Karnataka (labelled districts) as Appendix | |  |  | | --- | --- | |  | mean (*st.dev*) | | ADPC1 | -18.1 (*21.0*) | | ADPC2 | -15.6 (*23.0*) | | ADPC3 | -13.4 (*20.4*) | | ADPC4 | -10.4 (*22.4*) | | ADPC5 | 15.6 (*17.8*) | |

## References

Anand, S., Kakumanu, K.R. and Amarasinghe, U.A. (2019) ‘Use of Remote Sensing and GIS for Identifying Tanks and Rehabilitation Benefits to the Rural Areas’, *Journal of Rural Development*, 38(1), p. 55. Available at: https://doi.org/10.25175/jrd/2019/v38/i1/121801.

Balk, D. *et al.* (2019) ‘Urbanization in India: Population and Urban Classification Grids for 2011’, *Data*, 4(1), p. 35. Available at: https://doi.org/10.3390/data4010035.

Brown, C.F. *et al.* (2022) ‘Dynamic World, Near real-time global 10 m land use land cover mapping’, *Scientific Data*, 9(1), p. 251. Available at: https://doi.org/10.1038/s41597-022-01307-4.

Census of India (2011) ‘B-04 Main Workers classified by Age, Industrial Category, and Sex’. Available at: https://censusindia.gov.in/census.website/data/census-tables (Accessed: 30 May 2023).

Deichmann, U. (1996) *A Review of Spatial Population Database Design and Modeling*. Santa Barbara, CA: National Centre for Geographic Information and Analysis. Available at: https://escholarship.org/uc/item/6g190671 (Accessed: 28 February 2023).

Dixon, R.B. (1982) ‘Women in Agriculture: Counting the Labor Force in Developing Countries’, *Population and Development Review*, 8(3), pp. 539–566. Available at: https://doi.org/10.2307/1972379.

Eicher, C.L. and Brewer, C.A. (2001) ‘Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation’, *Cartography and Geographic Information Science*, 28(2), pp. 125–138. Available at: https://doi.org/10.1559/152304001782173727.

Eurostat (2021) *Applying the degree of urbanisation: a methodological manual to define cities, towns and rural areas for international comparisons : 2021 edition*. LU: Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2785/706535 (Accessed: 4 August 2023).

FAO (2023) *Food and Agriculture Organization of the United Nations (FAO) in India*. Available at: https://www.fao.org/india/fao-in-india/india-at-a-glance/en/ (Accessed: 6 June 2023).

Freire, S. *et al.* (2020) ‘Enhanced data and methods for improving open and free global population grids: putting “leaving no one behind” into practice’, *International Journal of Digital Earth*, 13(1), pp. 61–77. Available at: https://doi.org/10.1080/17538947.2018.1548656.

Government of India (2012) *Census of India 2011: Administrative Atlas*. Delhi, India: Office of the Registrar General and Census Commissioner. Available at: https://censusindia.gov.in/census.website/data/atlas# (Accessed: 1 June 2023).

Kondylis, F. *et al.* (2023) *Agriculture*, *World Bank: Development Impact Evaluation (DIME)*. Available at: https://www.worldbank.org/en/research/dime/brief/agriculture (Accessed: 1 June 2023).

Leyk, S. *et al.* (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), pp. 1385–1409. Available at: https://doi.org/10.5194/essd-11-1385-2019.

Malone, B.P. *et al.* (2012) ‘A general method for downscaling earth resource information’, *Computers & Geosciences*, 41, pp. 119–125. Available at: https://doi.org/10.1016/j.cageo.2011.08.021.

Meiyappan, P. *et al.* (2017) ‘Dynamics and determinants of land change in India: integrating satellite data with village socioeconomics’, *Regional Environmental Change*, 17(3), pp. 753–766. Available at: https://doi.org/10.1007/s10113-016-1068-2.

Mialhe, F., Gunnell, Y. and Mering, C. (2008) ‘Synoptic assessment of water resource variability in reservoirs by remote sensing: General approach and application to the runoff harvesting systems of south India’, *Water Resources Research*, 44(5). Available at: https://doi.org/10.1029/2007WR006065.

Monteiro, J., Martins, B. and Pires, J.M. (2018) ‘A hybrid approach for the spatial disaggregation of socio-economic indicators’, *International Journal of Data Science and Analytics*, 5(2), pp. 189–211. Available at: https://doi.org/10.1007/s41060-017-0080-z.

Pattnaik, I. *et al.* (2018) ‘The feminization of agriculture or the feminization of agrarian distress? Tracking the trajectory of women in agriculture in India’, *Journal of the Asia Pacific Economy*, 23(1), pp. 138–155. Available at: https://doi.org/10.1080/13547860.2017.1394569.

Pezzulo, C. *et al.* (2023) ‘A subnational reproductive, maternal, newborn, child, and adolescent health and development atlas of India’, *Scientific Data*, 10(1), p. 86. Available at: https://doi.org/10.1038/s41597-023-01961-2.

Qiu, Y. *et al.* (2022) ‘Disaggregating population data for assessing progress of SDGs: methods and applications’, *International Journal of Digital Earth*, 15(1), pp. 2–29. Available at: https://doi.org/10.1080/17538947.2021.2013553.

Schiavina, M., Melchiorri, M. and Pesaresi, M. (2023) ‘GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030).’ European Commission, Join Research Centre (JRC). Available at: https://doi.org/10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA.

Schneiderbauer, S. and Ehrlich, D. (2005) ‘Population Density Estimations for Disaster Management: Case Study Rural Zimbabwe’, in P. van Oosterom, S. Zlatanova, and E.M. Fendel (eds) *Geo-information for Disaster Management*. Berlin, Heidelberg: Springer, pp. 901–921. Available at: https://doi.org/10.1007/3-540-27468-5\_64.

Slavchevska, V., Kaaria, S. and Taivalmaa, S.L. (2019) ‘The feminization of agriculture: evidence and implications for food and water security’, in J.A. Allan (ed.) *The Oxford Handbook of Food, Water and Society*. Oxford University Press.

Stevens, F.R. *et al.* (2015) ‘Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data’, *PLOS ONE*, 10(2), p. e0107042. Available at: https://doi.org/10.1371/journal.pone.0107042.

Szarka, N. and Biljecki, F. (2022) ‘Population estimation beyond counts—Inferring demographic characteristics’, *PLOS ONE*, 17(4), p. e0266484. Available at: https://doi.org/10.1371/journal.pone.0266484.

Tatem, A.J. (2022) ‘Small area population denominators for improved disease surveillance and response’, *Epidemics*, 40, p. 100597. Available at: https://doi.org/10.1016/j.epidem.2022.100597.

Tobler, W. *et al.* (1997) ‘World population in a grid of spherical quadrilaterals’, *International Journal of Population Geography*, 3(3), pp. 203–225. Available at: https://doi.org/10.1002/(SICI)1099-1220(199709)3:3<203::AID-IJPG68>3.0.CO;2-C.

Tobler, W.R. (1970) ‘A Computer Movie Simulating Urban Growth in the Detroit Region’, *Economic Geography*, 46, pp. 234–240. Available at: https://doi.org/10.2307/143141.

Tobler, W.R. (1979) ‘Smooth Pycnophylactic Interpolation for Geographical Regions’, *Journal of the American Statistical Association*, 74(367), pp. 519–530. Available at: https://doi.org/10.2307/2286968.

Tuholske, C. *et al.* (2021) ‘Implications for Tracking SDG Indicator Metrics with Gridded Population Data’, *Sustainability*, 13(13), p. 7329. Available at: https://doi.org/10.3390/su13137329.

United Nations (2022) *The Sustainable Development Goals Report 2022*. New York, NY: United Nations. Available at: https://unstats.un.org/sdgs/report/2022/.

United Nations in India (2022) *UN India Annual Report 2021*. New Delhi, India. Available at: https://india.un.org/en/195240-un-india-annual-report-2021 (Accessed: 30 May 2023).

Vanthof, V.R. and Kelly, R.E.J. (2020) ‘Earth Observation at Finer Scales is Critical to Farming Communities Facing Increased Water Shortages Over the Next Decade’, in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3716–3718. Available at: https://doi.org/10.1109/IGARSS39084.2020.9324327.

Viel, J.-F. and Tran, A. (2009) ‘Estimating Denominators: Satellite-Based Population Estimates at a Fine Spatial Resolution in a European Urban Area’, *Epidemiology*, 20(2), pp. 214–222.

Wardrop, N.A. *et al.* (2018) ‘Spatially disaggregated population estimates in the absence of national population and housing census data’, *Proceedings of the National Academy of Sciences*, 115(14), pp. 3529–3537. Available at: https://doi.org/10.1073/pnas.1715305115.

World Bank (2023) *Agriculture and Food*, *World Bank*. Available at: https://www.worldbank.org/en/topic/agriculture/overview (Accessed: 22 May 2023).

You, L. and Wood, S. (2006) ‘An entropy approach to spatial disaggregation of agricultural production’, *Agricultural Systems*, 90(1), pp. 329–347. Available at: https://doi.org/10.1016/j.agsy.2006.01.008.

Zarkovich, S.S., Bosnich, S. and Anichich, Z. (1976) ‘Agricultural Population’, *International Statistical Review / Revue Internationale de Statistique*, 44(2), pp. 283–288. Available at: https://doi.org/10.2307/1403288.