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

Github Link

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

Agricultural dependent populations (ADP) in low-income countries have unique development needs that are different from the population as a whole. Monitoring these development needs and progress against policy targets such as the Sustainable Development Goals requires information on the spatial distribution of these populations, to identify geographic inequalities that may be masked by aggregated statistics over a larger region. Gridded distribution data over a regular spatial scale can also be more readily combined with earth observation and environmental data, overcoming the limitations of responding to natural events that do not align neatly with administrative boundaries. This study therefore proposes and evaluates a two-stage methodology for the spatial disaggregation of ADP into a uniform spatial grid, applied to the case study of India. Firstly, a dasymetric mask is used to identify population distributed across cropland areas. In the second stage, estimates are adjusted to match census data at an aggregated level through the iterative addition or subtraction of a buffer area to the initial estimate. Performance was evaluated against computation time, spatial resolution of output, and variation in buffer radius, and variations of the input parameters compared. Following feasibility analysis on a single test state, Karnataka, the method was scaled to estimate ADP for the whole of India. The results show that this method can be scaled across a large and diverse geographic area, with 628 out of 640 districts total (98%) meeting eligibility criteria and successfully mapped. The final raster product and underlying code have been published online to encourage application of the output in practice or in further research.

Declaration

I hereby declare that this dissertation is all my original work and that all sources have been acknowledged. It is 10,877 words in length.

Acknowledgements

Firstly, I would like to express my gratitude and appreciation for my supervisor Dr Fulvio Lopane, for his expertise, support, and guidance throughout the development and execution of this research project. I also extend this gratitude to Sophie Ayling, for valuable input on the study design, coding, and initial research topic.

This work, along with everything I have gained and experienced from my time at CASA, would not have been possible without the generous support of the Rae & Edith Bennett Scholarship, provided by my alma mater the University of Melbourne.

Lastly, I am deeply grateful for the support of CASA colleagues, friends, and my partner Jason.

TABLE OF CONTENTS

List of Figures

List of Tables

1. 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 risk 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 (Vanthof and Kelly, 2020), and of heightened vulnerability to drought, disaster events, and climate change. In the latest *State of Food Security* report, the Food and Agriculture Organisation (FAO) explicitly acknowledge the "megatrend" of urbanisation as a factor impacting food systems globally, and the need for greater understanding of the changing spatial distribution of agricultural production – and the communities involved – across the urban-rural continuum (FAO, IFAD, UNICEF, WFP, WHO, 2023).

Spatial disaggregation is a crucial component of understanding spatial distribution. Particularly, the disaggregation of demographic data to a uniform spatial grid format has become a well-established method for reporting spatial distribution. 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. This study is a novel addition to the field, in that no other research has applied spatial disaggregation methods to estimate the agricultural dependent population beyond just agricultural labourers. India as a case study allows the assessment of feasibility and performance of the methodology at a large spatial scale and across a diverse range of landscapes, in a region with a dominant agricultural sector that is a major target of development action (FAO, 2023).

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 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.

1.1 Research Question

Agricultural dependent populations (ADP) 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 dependent population in India be identified at a small area scale?

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

i. Review existing methods for spatial disaggregation of demographic data,
ii. Propose and evaluate a new method that combines dasymetric disaggregation and iterative extension (buffers), and
iii. Scale the method up to estimate the small area agricultural population for all of India.

1.2 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 that critique reliance on labour statistics 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 this may be compounded by increasing female participation in agricultural labour in the context of male outmigration from the sector (Pattnaik *et al.*, 2018; Slavchevska, Kaaria and Taivalmaa, 2019). However, in many contexts, as in the case of this study, agricultural labour participation is one of the few statistics that is reliably published and made available that provides an indication of the scale of agricultural dependence in a region. Additionally, 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 agricultural labour, with or without household dependents. Labour force participation is an established method for calculating various dimensions of dependence, particularly in the primary sector

(e.g. Natale *et al.*, 2013). In India, Swaminathan (2020) has used labour force statistics in conjunction with time-use surveys to assess agricultural dependence in rural areas, with a focus on gender. Notably, Indian labour force statistics record data for all work irrespective of wages received (excluding own housework), which includes unpaid labour in the operation of household and family farms (Chattopadhyay *et al.*, 2022). By capturing unpaid labour there is greater confidence that labour statistics provide a more accurate estimate of the true level of agricultural dependence in a region.

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.

1.3 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). Administratively, the country is divided into 28 states and 8 union territories, each of which are further subdivided into districts and smaller 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 km^2 and a population of 2 million (Table 1).

Table 1: Characteristics of 2011 Indian Census Districts (n = 640)

Measure	Population	Area (km^2)	Population Density (km^2)
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 (Balk *et al.*, 2019), this factor was omitted from the methodology for this study. Instead, rural classification derived from remotely sensed land cover data has been used, as discussed in Section 2.2.

The nation has several characteristics that prioritise it as a setting for this research. India's rural population are overwhelmingly reliant on agriculture as a source of livelihood, are at the crux of a demographic transition which is seeing increasing feminisation of the sector, and face continuing food insecurity and sustainability challenges (Marois, Zhelenkova and Ali, 2022; FAO, 2023). As an example for this case study, there is the potential application of spatial disaggregation of ADP to evaluate water tank rehabilitation for irrigation. 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.

2. Methodology

This section introduces the technical concepts behind spatial disaggregation and reviews existing methodologies, in response to Objective I. The reasoning behind the proposed hybrid dasymetric and buffer method are explained, and the methodology is then broken down into the data sources, study setting, how ADP has been defined, and the details of the buffer iteration process.

2.1 Spatial Disaggregation

Spatial disaggregation is a broad term which applies to the process of transforming data from a set of source zones, such as district polygons, 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.

A key advantage of spatially disaggregated 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), and overcome the limitations of responding to natural events or crises which do not align neatly with administrative or jurisdictional boundaries.

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 load 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).

Dasymetric mapping, rather, uses complementary ancillary data to divide an area into homogenous zones based on the variable of interest (Eicher and Brewer, 2001) thus overcoming the main weakness of areal weighting. For example, earth observation data can be used to identify land cover classes within a zone and derive a binary ‘mask’ where all pixels classed as cropland are retained and all other pixels removed or set to zero (Figure 1). An aggregated value that applies to the entire zone, such as count of crop labourers, can then be

proportionally distributed across the non-zero cells, to produce a more accurate estimate of the real spatial distribution of the population (Qiu *et al.*, 2022). The method is not limited to population counts, and Holt *et al.* (2011) argue that in theory it could be applied to the disaggregation of any type of sociodemographic data.

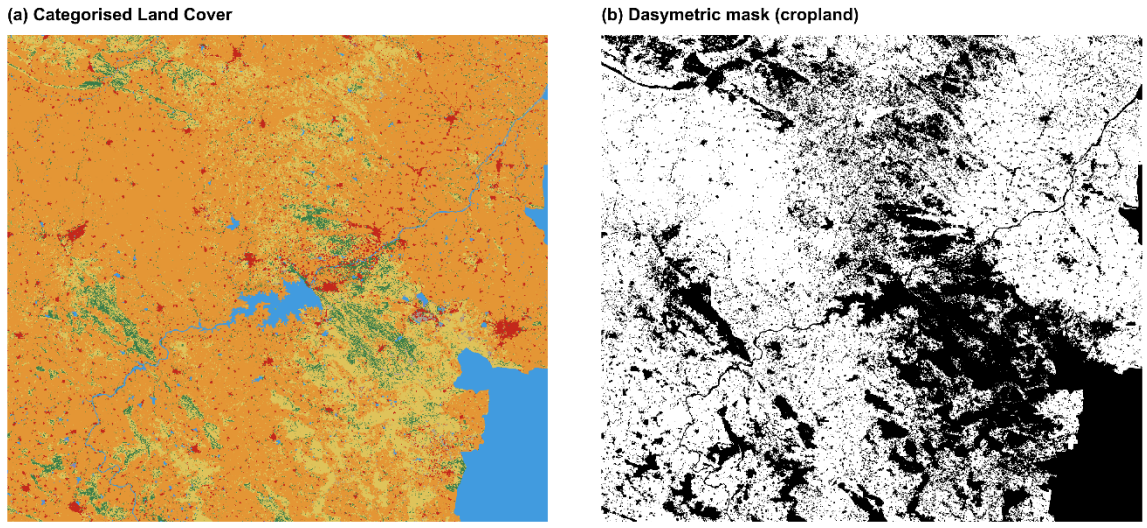

Figure 1: Illustrative diagram of dasymetric masking.

(a) Categorised land cover map of an area of India, extracted from Dynamic World. (b) Dasymetric mask derived from image (a), representing only pixels classed as cropland (white) against all other classes (black).

An alternative method is the incorporation of the pycnophylactic, or 'mass-preserving', property, which requires that the sum of pixel estimates is equal to the supplied value of the source zone for the variable of interest (Malone *et al.*, 2012). Many global gridded population estimates, such as WorldPop, adjust estimates at a national or subnational scale based on United Nations population data (Stevens *et al.*, 2015), to ensure that gridded estimates align with these counts. This method also features in pycnophylactic interpolation, where the weighted average of a pixel's neighbours is used to iteratively smooth the population values in grid cells whilst ensuring the mass-preserving property is met, as a technique to lessen the effect of sharp changes in population density estimates at the boundaries of source zones. 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).

The method proposed for this study takes the strengths of these approaches by combining features of both dasymetric and pycnophylactic disaggregation. In most cases, dasymetric methods or hybrid methods combining dasymetry and pycnophylactic interpolation outperform simple areal weighting (You and Wood, 2006; Monteiro, Martins and Pires, 2018), and have thus become key features in the development of modern disaggregation models. World gridded population products such as the Global Human Settlement Layer, WorldPop, and

Esri's World Population Estimate all utilise ancillary data in some format to apply dasymetric constraint (Table 2). Advances in computational power and the availability of high quality earth observation data have supported a proliferation in spatial disaggregation studies over the last decade (Wardrop *et al.*, 2018), which have further increased the sophistication and variety of disaggregation methodologies.

Table 2: 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)	CIESIN[a]	Areal weighting	1km	Water bodies
Global Human Settlement Layer – Population (GHS-POP)	JRC[b] and CIESIN[a]	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	ORNL[c]	Smart interpolation	30 arcsec	Roads, Land cover, Built structures, Urban areas, Infrastructure, Climate, Topography, Elevation, Water bodies
World Population Estimate (WPE)	Esri	Dasymetric	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.

Although there has been extensive research and methodological development in the field of population disaggregation (Leyk *et al.*, 2019), 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 low birthweight, and in the Indian context produced an 'atlas' of 19 indicators nationwide at a 5km resolution (Pezzulo *et al.*, 2023).

However, 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). No studies were found that attempt

to estimate the spatial distribution of the entire agriculture dependent population, inclusive of non-workers and families. The importance of accounting for population beyond just working labourers highlights the novel contribution of this methodology, and hopefully acts as a catalyst for continued research and development in this area.

2.2 Data Sources

There are four key sources of data that form the input for this analysis: the Dynamic World land cover dataset, 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 Dynamic World 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 Dynamic World data from Earth Engine can be accessed on the Github repository.

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 (Table 3) 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 3: Global Human Settlement Layer – Settlement Model Grid (GHS-SMOD) Classification Rules (Schiavina, Melchiorri and Pesaresi, 2023)

Code	Class	Population Density (km^2)	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.

Tables from the Indian Census 2011 were used to calculate total population and agricultural dependent population at the state and district level. This census-estimated value was used as the 'ceiling' to which aggregated estimates were adjusted to align with, ensuring that final estimates from the model are roughly equivalent to official population estimates at higher administrative levels.

All analysis was conducted using Python v3.10.9, with additional visualisation completed using QGIS v3.26.3. Details of packages used, the python environment, and the code can be accessed from the project's Github repository (https://github.com/joepost/india_adp).

2.3 Study setting

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 ADP_C estimate. This methodology was then scaled up to estimate the ADP across all states.

Karnataka is the 8th largest state in India by population, with a 2011 population of more than 60 million people, and 6th largest in area, spread across 30 districts and 192,000 km^2 (Figure 2). 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.

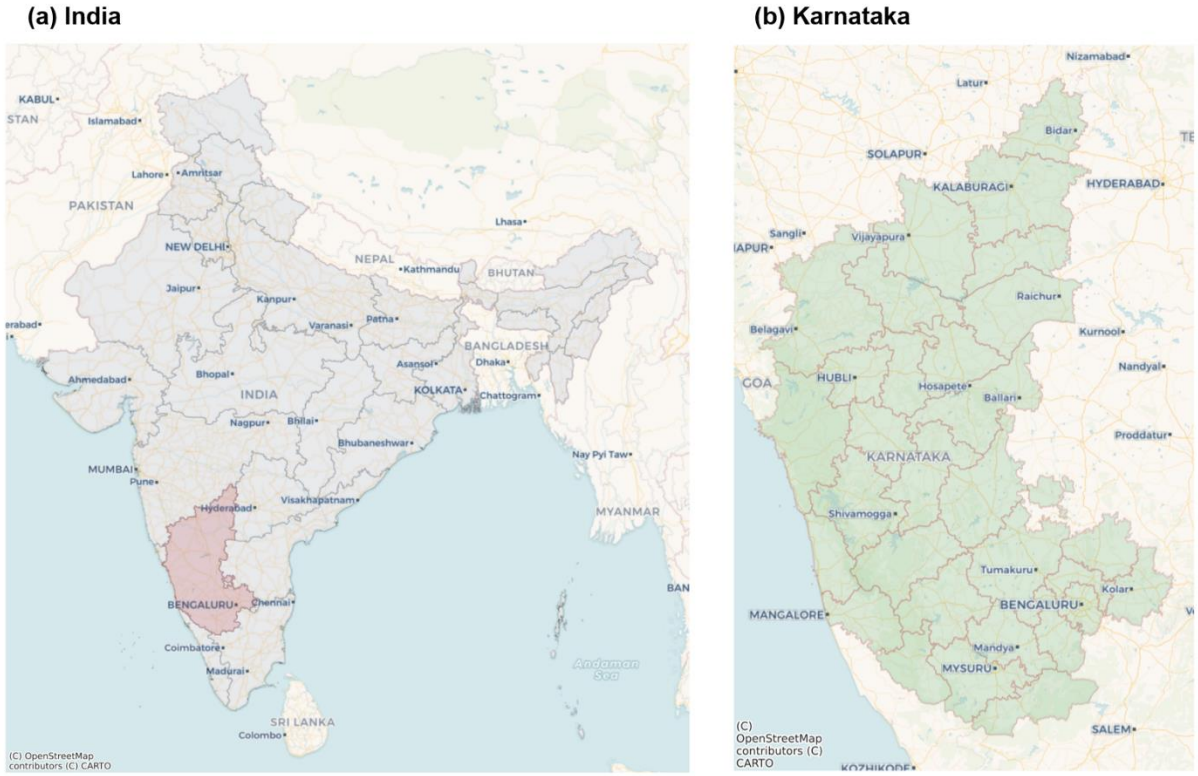

Figure 2: Reference maps of study setting.

(a) India, by State and Union Territories. The test state Karnataka is highlighted in green. (b) Karnataka, by District. Inset of the state in map (a). Maps have been generated using python *contextily* package.

2.4 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 from Dynamic World was used to create a dasymetric mask of cropland for each district in India. Gridded population estimates from WorldPop were then joined to cropland areas by intersection, and aggregated to the district level, to produce a base estimate of ADP that encompasses all inhabitants in crop landcover – this is referred to as the aggregated ADP, or ADP_A.

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 ADP_C. This estimate was used to validate the accuracy of the ADP_A, by calculating the difference between the district cropland population and the census-estimated agricultural population.

As discussed above, ADP itself is a broad concept that is not well defined in the literature. Therefore, a series of 5 alternative ADP_c estimates were calculated, as outlined in Equations 1 to 5, to evaluate variation depending on definition.

$$[1] \quad ADP_{C1} = Cultivators_{Main} + Agricultural\ Labourers_{Main}$$

$$[2] \quad ADP_{C2} = Total\ Primary\ Sector_{Main}$$

$$[3] \quad ADP_{C3} = Cultivators_{(Main+Marginal)} + Agricultural\ Labourers_{(Main+Marginal)}$$

$$[4] \quad ADP_{C4} = Total\ Primary\ Sector_{(Main+Marginal)}$$

$$[5] \quad ADP_{C5} = ADP_{C3} * \left(\frac{Total\ Population}{Total\ Workers}\right)$$

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. ADP_{C1} and ADP_{C2} 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, ADP_{C3} and ADP_{C4} account for both main and marginal workers as agriculture dependent. Due to the often-seasonal nature of agricultural work, many labourers in the sector may be classed as marginal whilst still being functionally dependent on the work for their livelihood (Swaminathan, 2020).

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). ADP_{C1} and ADP_{C3} 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, ADP_{C2} and ADP_{C4} account for all workers within the agricultural sector.

Lastly, ADP_{C5} 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 labour force dependency ratio (Marois, Zhelenkova and Ali, 2022), which is the ratio of workers to non-workers (dependents), calculated as the total population divided by total workers, under the assumption that the ratio of workers to dependents is roughly equivalent in the agricultural sector as in the total population.

2.4 Validation of population estimates

The aggregated population estimate, ADP_A, was summarised at the district level and compared to district-level ADP_C 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 – adjusting the results to satisfy the pycnophylactic 'mass-preserving' property. 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.

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 were 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 ADP_A recalculated for this area. This process was repeated, at progressively scaling increments, until the difference between ADP_A and ADP_C was less than ±5% for all districts within a state. Equations [5] and [6] show the calculation of buffer radius according to an increase or decrease in distance, respectively:

$$[5] \quad r_n = 1.5\, r_{n-1}$$

$$[6] \quad r_n = \frac{r_{n-1}}{2}$$

Where r_n is the buffer radius at iteration n. For districts where ADP_A exceeds the threshold by greater than 5%, r is negative. A limit of $n \leq 10$ was set to prevent iteration continuing indefinitely in cases where the threshold is unable to be met. An overview of the process is shown in Figure 3.

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 ADP_A 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 deemed ineligible and removed from analysis.

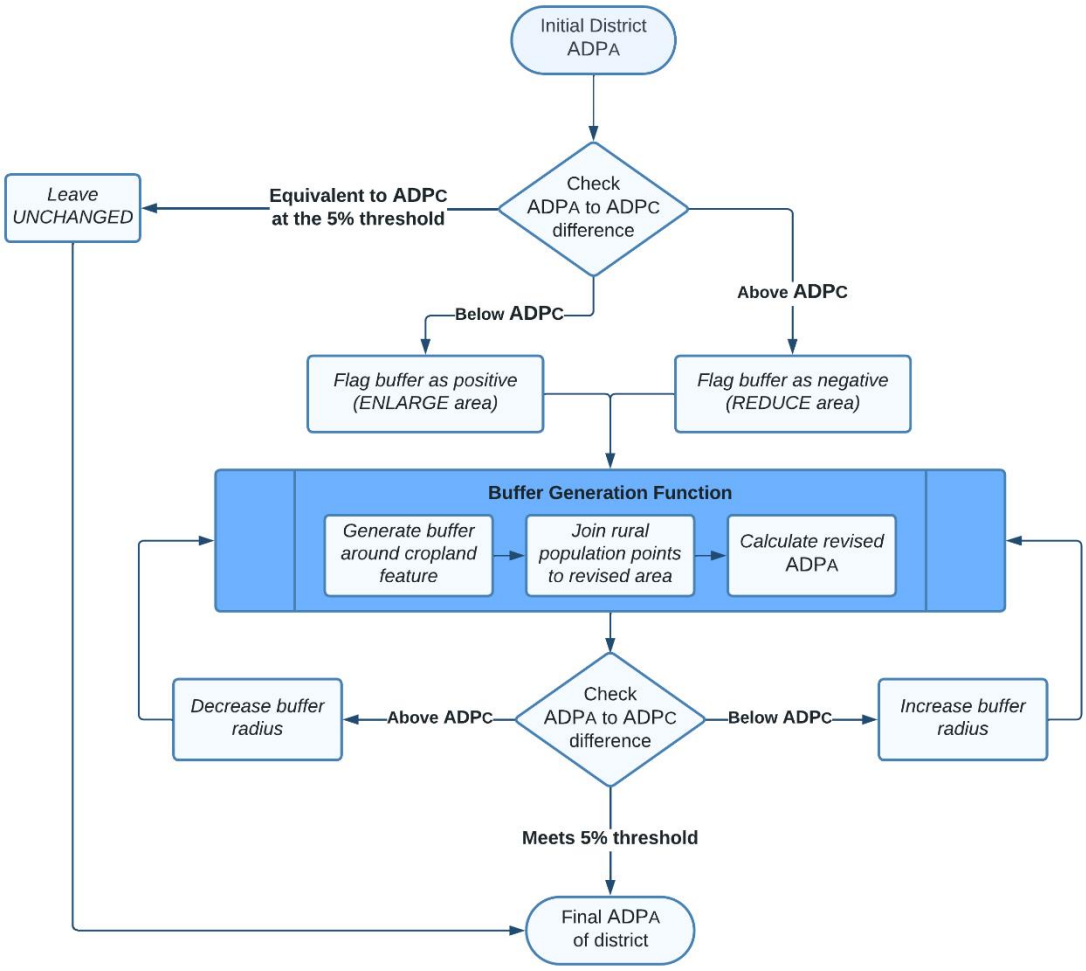

Figure 3: Overview of buffer iteration process.

Initial district ADP$_A$ represents the output from the first component of the method, described in Section 2.5. Each district continues through the iterative loop until the revised ADP$_A$ falls within the ±5% threshold. ADP$_A$ = aggregated agricultural dependent population; ADP$_C$ = census-estimated agricultural dependent population.

3. Results

The results for this study are presented in two main components. First, for the test case Karnataka, performance of the method under various parameters are shown, followed by results of the buffer iteration process and an example of the raster output. The second component then presents the results for the method scaled to India.

3.1 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 ADP_A and ADP_C was compared for each of ADP_{C1} to ADP_{C5}. For spatial resolution, performance was measured by computation time and ADP_A/ADP_C difference.

Of the five variations of ADP_C estimate, four produced a mean negative result – indicating that for most districts, the population residing within cropland areas exceeded the census-estimated agricultural population for the district overall. For ADP_{C5}, the reverse was true – the total population within cropland areas was less than the census-estimated agricultural population for the district overall, on average (Figure 4). ADP_{C5} also exhibited the lowest absolute mean (18.1) and the lowest variance of the five models (standard deviation = 17.8).

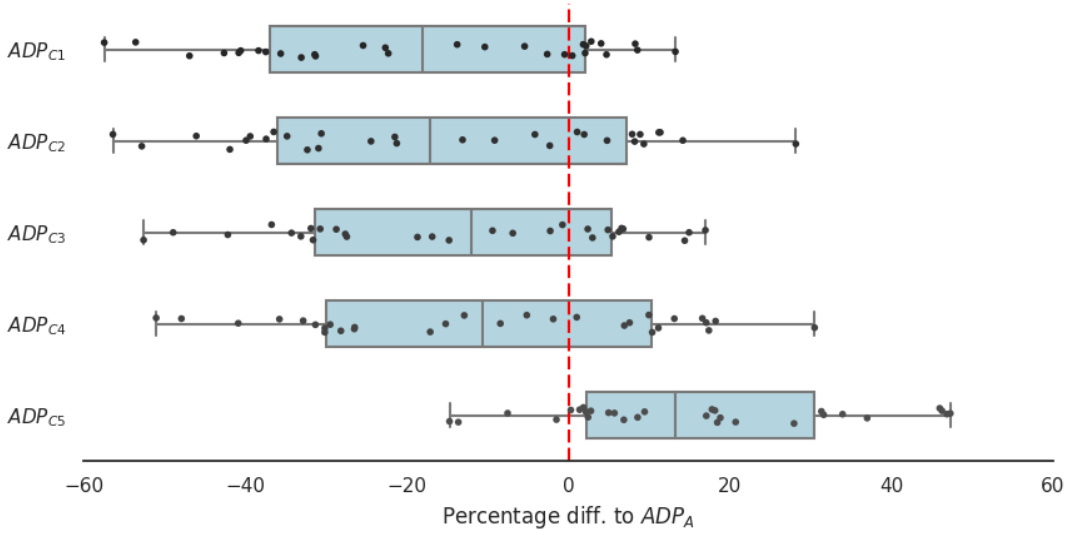

Figure 4: Distribution of ADP_A district estimates against census results, by ADP_C method, Karnataka.

Percentage difference is calculated the difference in percentage points between ADP_C (census-estimated agricultural dependent population) as a proportion of total population and ADP_A (aggregated agricultural dependent population) as a proportion of total population. Each point represents a single district.

Of the three raster input datasets, Dynamic World 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 Dynamic World and WorldPop. All analysis has been performed on a single Windows Computer 64-bit operating system with 16.0 GB RAM. As expected, 1km:1km (Dynamic World:WorldPop) performed the fastest (2.2 minutes), followed by 1km:100m (46 minutes), and 100m:1km (115 minutes). Although 100m:100m would be expected to 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. The transition between 100m to 1km resolution in Dynamic World data was found to have the greatest impact on variation in results; therefore, 100m cropland data was prioritised for the scaled method.

3.2 Buffer iteration

Comparing the results of the buffer process at 1km and 100m spatial resolution of cropland, there was a near 10-fold reduction in spread for the latter (Figure 5). At 1km input, there was a mean buffer radius of 674 (standard deviation of *1,028*), compared with 124 (*116*) at 100m resolution. This resulted in a much lower range and removed extremes of buffer radius in each direction. There was also a halving in the proportion of negative buffers – from 21% to 10% of districts, at 1km and 100m respectively.

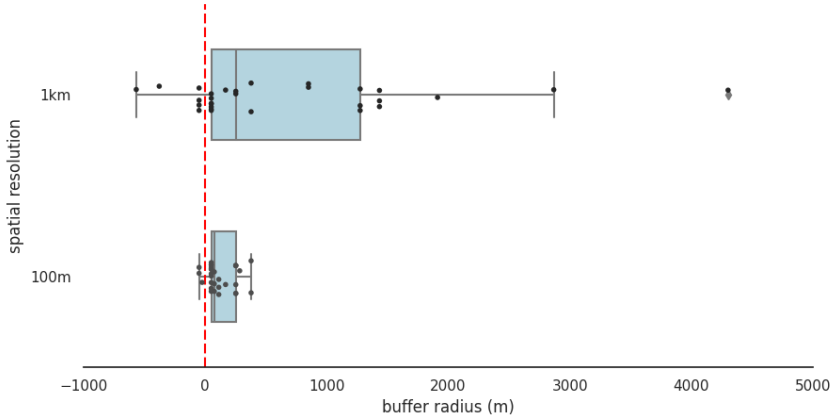

Figure 5: Distribution of buffer radius estimates by spatial resolution of input cropland data, Karnataka.

Spatial resolution of WorldPop and GHS-SMOD data stable at 1km for both analyses. Each point represents a single district.

As shown in Figure 6, the agricultural population in Karnataka is evenly distributed across the central, north, south, and east of the state. A sparsely populated barrier runs across the west of the state, roughly along the path of Sahyadri mountain range, leading to a dense cluster of agricultural population wedged along the west coast. Additional clusters can be identified in the hinterland of major urban centres, such as Bengaluru in the southeast, Mysuru (Mysore) in the south, and along the corridor between Hubballi and Belagavi in the northwest.

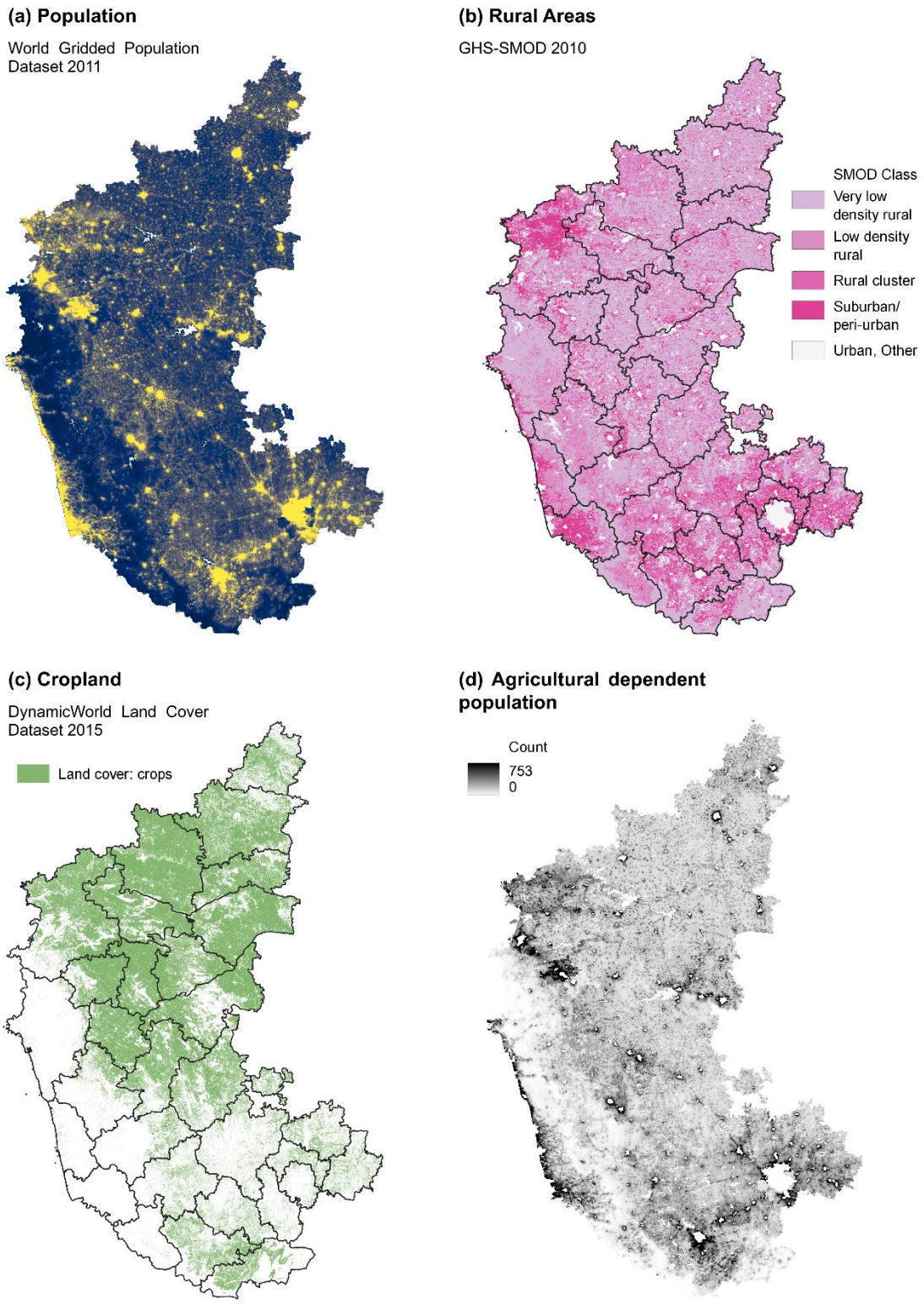

Figure 6: Spatial distribution of inputs and results, Karnataka.

(a) Population, derived from WorldPop Gridded Population Dataset 2011 at 1km resolution. (b) Rural areas, derived from the Global Human Settlement Layer – Settlement Model Grid Dataset 2010 at 1km resolution. *Urban/Other* includes water bodies and other non-inhabitable areas. (c) Cropland, derived from DynamicWorld Land Cover Dataset 2015 at 100m resolution. (d) Spatial distribution of agricultural dependent population. Buffers validated against ADP_{C5} (census-estimated agricultural dependent population).

Buffer results, when assessed spatially, show a clustering pattern of low and negative radii in the north of the state, and higher radii in a band of districts running from west to southeast (Figure 7a). When tested for the presence of spatial autocorrelation, the map produced a Moran's I of 0.4 (p-value < 0.001), indicating that the radius values cluster together more than would be expected under the null hypothesis of random spatial distribution. Local Moran's I was then calculated by district, identifying 'hot spots' and 'cold spots' of high and low buffer radii, respectively. In Figure 7b, *High-High* indicates high-radius districts surrounded by other high-radius districts, that are more similar than would be expected under a situation of random spatial distribution. *Low-High* indicates the inverse; a low-radius district surrounded by high-radius districts, that are more dissimilar than expected. Significance is calculated as $p < 0.05$.

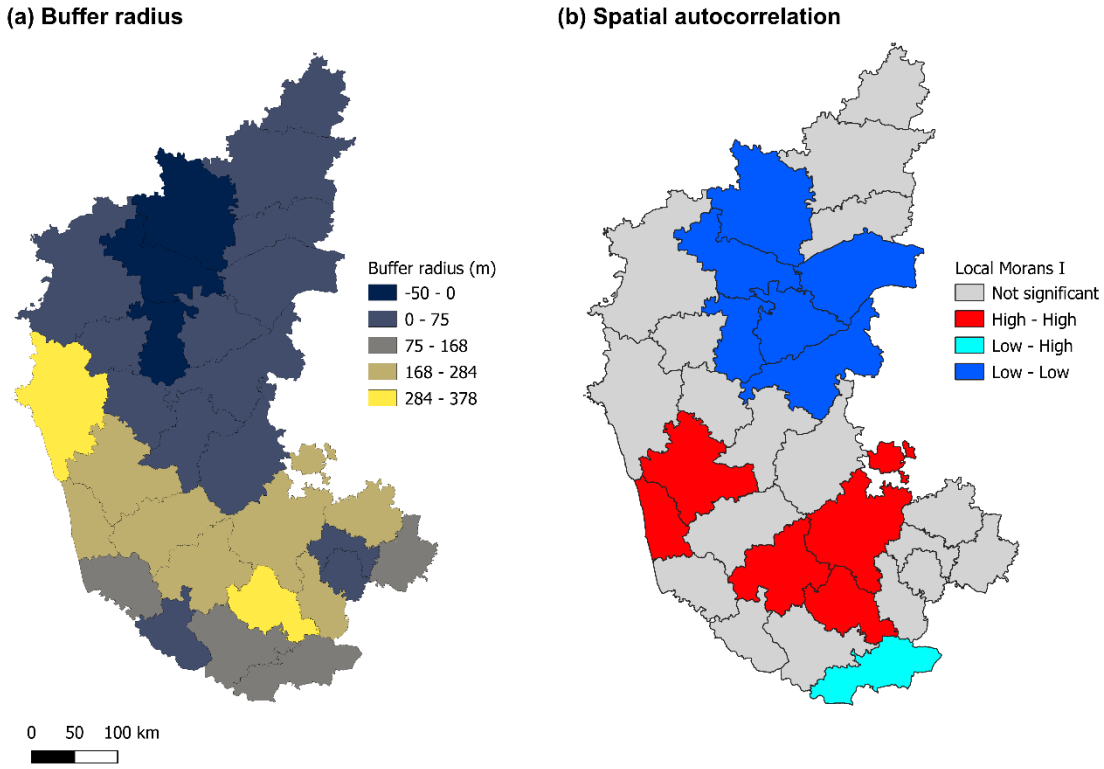

Figure 7: Spatial distribution maps of buffer radius, Karnataka.

(a) Buffer radius by district, Karnataka. Based upon ADP_{C5} as validation threshold. (b) Spatial autocorrelation of buffer radius by district. Calculated by Local Moran's I, at significance $p<0.05$.

3.3 Scaling method to India

The method was scaled to the whole of India by iterating the method individually over each Indian state, and two of the union territories (Jammu & Kashmir, and Ladakh). The remaining six union territories (Andaman & Nicobar Islands, Chandigarh, Dadra & Nagar Haveli & Daman & Diu, National Capital Territory of Delhi, Lakshadweep, and Puducherry) were excluded due

to being metropolitan territories or small islands, with minimal agricultural cropland. ADP_{C5} was used for validation of ADP_A and calculation of buffer distance, with cropland at a spatial resolution of 100m.

Computation time at 100m cropland resolution varied widely by state, depending on area, from less than a minute for Goa to 1,731 minutes (28.8 hours) for analysis of Maharashtra. At the 1km resolution, for comparison, time difference was negligible with no state taking longer than 6 minutes to complete the analysis. Figure 8 below compares the runtime for a selection of states – note that the x-axis (Runtime) is presented on a log scale.

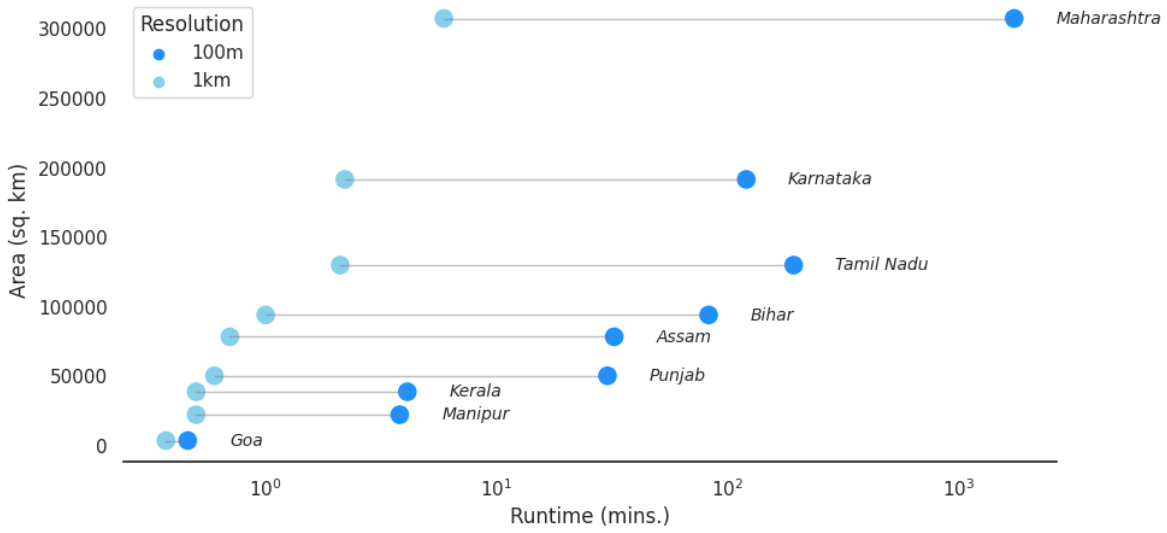

Figure 8: Computation time by state area and spatial resolution.

Sample of total states. Spatial resolution refers to cropland (Dynamic World) data. Runtime is presented on a log scale. Resolution of WorldPop and GHS-SMOD data stable at 1km for both analyses.

Once all the states were completed at the 100m level, buffer radius results for the full sample size of districts were compared against four key characteristics: (a) total population, (b) ADP_{C5} as a percentage of total population, (c) percentage of area classed as rural, and (d) percentage of area classed as cropland (Figure 9). The results suggest a trend of decreasing buffer radius with increasing total population and cropland area and increasing buffer radius with increasing proportions of ADP_C and rural area.

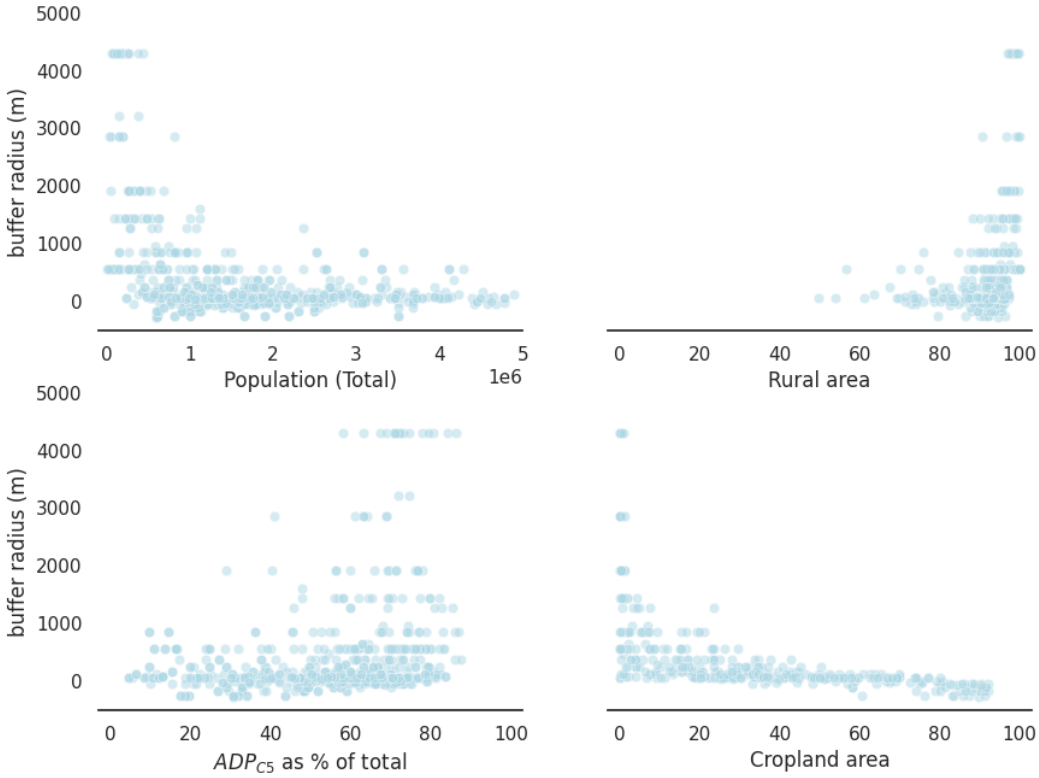

Figure 9: Buffer radius by selected district characteristics, India.

(a) Total population of district (in millions). (b) ADP_{C5} (census-estimated agricultural dependent population) as a percentage of total population. (c) Rural area as a percentage of total district area. (d) Cropland area as a percentage of total district area. Each point represents a single district.

3.4 Ineligible districts

After calculating an initial ADP_A for each district, a set of conditions were applied to remove any that were ineligible for further analysis. These were districts that had one of (i) minimal rural area/population, (ii) minimal cropland area/population, or (iii) where the ADP_C exceeded the total rural population (Table 4). Of 640 districts total, 12 met the exclusion criteria (1.9%).

The districts removed include a mix of major cities (Chennai, Mumbai, Hyderabad, Kolkata) with little to no rural area or rural population, and highly unurbanized regions with low population and very low areas of cropland, in the northeastern states of Mizoram, Nagaland and Sikkim. More unexpected were districts such as Dhubri in Assam, which had large rural and cropland areas, but where the ADP_C still exceeded the total rural population.

Table 4: Demographic and land use characteristics of ineligible districts.

State	District	Pop. (Total)	Pop. (Rural)	ADP_{C5}	% of Land Area		Reason ineligible
					Crops	Rural	
TN	Chennai	4,646,732	-	71,926	1.0	-	No rural population
MH	Mumbai	3,085,411	175	28,346	1.4	1.1	ADP_{C5} exceeds total rural population
JH	Godda	1,313,551	1,014,663	1,076,054	41.8	89.9	ADP_{C5} exceeds total rural population
AS	Dhubri	1,949,258	1,044,305	1,096,769	36.1	76.8	ADP_{C5} exceeds total rural population
TL	Hyderabad	3,943,323	-	143,158	-	-	No rural population
WB	Haora	4,850,029	615,687	698,162	34.9	29.6	ADP_{C5} exceeds total rural population
WB	Kolkata	4,496,694	5,225	71,275	1.9	1.7	ADP_{C5} exceeds total rural population
WB	Purba Medinipur	5,095,875	2,467,615	2,794,297	45.9	66.6	ADP_{C5} exceeds total rural population
CT	Narayanpur	139,820	111,121	114,487	1.2	99.4	ADP_{C5} exceeds total rural population
MZ	Saiha	56,574	56,017	28,678	0.08	99.5	Less than 1% cropland
NG	Zunheboto	140,757	134,025	100,205	0.02	98.8	Less than 1% cropland
SK	West District	136,435	93,780	95,175	0.07	89.6	Less than 1% cropland

ADP_{C5} = census-estimated agricultural dependent population; TN = Tamil Nadu; MH = Maharashtra; JH = Jharkhand; AS = Assam; TL = Telangana; WB = West Bengal; CT = Chhattisgarh; MZ = Mizoram; NG = Nagaland; SK = Sikkim.

4. Discussion

This study aimed to answer the research question, *How can the agricultural dependent population in India be identified at a small area scale?* A hybrid method was proposed, combining the dasymetric masking of population in agricultural areas with an iterative buffer process to adjust scope areas based off census data. Results show that this method is feasible to scale to large geographic areas (the Indian subcontinent), depending on time and resource context, providing a valuable first step in the literature in the spatial disaggregation of agricultural populations.

4.1 Establishing method parameters

Five variations of ADP_C calculation were compared, and of these ADP_{C5} was selected as the most appropriate method to scale, predominantly for two reasons. Firstly, ADP_{C5} showed the lowest absolute mean and lowest standard deviation of all the models, indicating that the initial ADP_A estimates were closer to the census validation value, and that there were fewer districts with extreme differences (positive or negative). Absolute mean has been used here instead of the mean to assess the magnitude of difference between ADP_A and ADP_C initial estimates, without being impacted by the 'direction' of that difference. This is intended to account for cases with very large positive and negative differences, which would balance each other out and could result in a mean close to zero that masks the magnitude of difference.

Secondly, ADP_{C5} is conceptually the most appropriate given the research question, as it accounts for the broader population who may be agriculture dependent, not only the working population. The results reflect this, in that ADP_{C5} was the only model where the census-estimated ADP exceeded the population within cropland areas, on average, because it is accounting for a larger share of the total population. The four equations which account for labourers alone inherently assume that [1] the non-working dependents of agricultural labourers are not themselves agriculture dependent, and [2] there are no non-working population residing within cropland areas. Neither of these assumptions are reasonable, and accounting strictly the labour force in this way has long been criticised for the significant population that it leaves out (Zarkovich, Bosnich and Anichich, 1976). However, various models have been included in the analysis to assess how the change in definition impacts the results outcome.

4.2 Computation load

In addition to testing a methodology, Objective III of this study involved scaling the method to assess feasibility across a large spatial scale. The results highlight an important cost-benefit decision when deciding the spatial resolution of input data. Although higher spatial resolutions

produce more confidence in results (with lower variation in buffer radius), they also require much longer computing times, especially for larger areas.

In comparing running the model at 100m WorldPop resolution compared with 100m cropland resolution, the latter required a much longer computation time. However, investigation found that reducing cropland to a 1km resolution led to a large components of cropland area being removed, as cropland in many parts of Karnataka is small scale and highly fragmented. This pattern is likely to be repeated across many parts of India, particularly in sparsely populated and highly rural regions in the northeast, such as Mizoram, Sikkim, and Nagaland. The model performed better, in terms of lower variation in buffer radii, when these smaller fragments of cropland were captured. Therefore, 100m spatial resolution for cropland was prioritised in the scaled model.

Several features were tested to maximise the computation efficiency of the overall process. After initially testing feasibility in QGIS, the method was translated to run entirely in python, particularly with the use of packages *Rasterio*, *GDAL*, *Shapely*, *Fiona*, and *GeoPandas*. Time intensive input/output operations were streamlined to use the Geofeather file format, which produced considerable speed advantages over traditional shapefiles. However, results show that at the 100m scale the largest states still had runtimes between 2 – 10 hours, and that runtime increased superlinearly with increasing area. Future applications should consider the trade off between scale, accuracy and time cost when selecting the spatial resolution of input data.

4.3 Interpreting buffer radius

To interpret buffer radius in terms of the results output, it is necessary to acknowledge that the methodology for this study is built upon two main assumptions.

i. The agricultural dependent population are geographically within or proximal to cropland areas, where they source their livelihoods.
ii. The agricultural dependent population do not reside in high density urban areas, and surplus labour is drawn from rural settlements surrounding cropland.

The predominance of smallholder farming in the Indian agriculture sector supports these assumptions, as most agricultural labour is performed locally in and around farming villages (FAO, 2023). 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.

To assess the accuracy of ADP_A estimates, the buffer radius should be interpreted in the context of the two assumptions outlined 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 part of the 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 these cases should be interpreted with caution, and always in the geographic and socioeconomic context of the district in question.

As observed in the comparison between district buffer distributions at 1km and 100m cropland resolutions, increasing the spatial resolution input tends to reduce the buffer radius, especially at the extremes. This trend reflects findings from a study by Zhang *et al.* (2014) on the effects of spatial resolution on estimates of crop acreage, with decreasing spatial resolution leading to lower accuracy and higher standard deviation. There are likely multiple contributing factors to this. At higher spatial resolutions smaller fragments of features can be identified, which is important for agricultural fields in smallholder farms, which are often situated within heterogeneous landscapes (Brown *et al.*, 2019). Secondly, the buffers are generated over a finer polygon that more accurately represents the true spatial distribution of cropland, reducing the scale of redundancy in iterative buffer estimates.

4.4 Spatial distribution of buffer radius

The paired choropleth map and Local Moran's I hotspots map of Karnataka in Figure 7 tell a similar story – there is clear geographic clustering of high and low buffer radii. The Moran's I analysis provides a statistical validation for this clustering, identifying 'hotspots' of high radius clusters and 'coldspots' of low radius clusters. Placing these clusters in geographical context, particularly with local knowledge and for those working in agriculture, policy and development in this region, can help to tease apart the potential factors that are driving this effect.

Scatter plots of buffer radius against four district characteristics were produced to investigate likely factors that may affect the required buffer size in a district. Buffer radii tend to be most extreme, roughly indicative of higher uncertainty, in districts with low population, high proportion of the population classed as agricultural dependent, high percentage of rural area, and low percentage of cropland area. A complete analysis on the factors influencing buffer radius is outside the scope for this study; however, further research on this methodology should investigate the nature and statistical strength of these associations, to better understand the factors that affect buffer radius and by extension the spatial estimation of ADP. Topography, climate, and economic structure are all factors which may be associated with the spatial distribution of buffer radii.

4.5 Common factors of ineligible districts

Urban districts, low cropland districts, and districts where the ADP_C exceeded the total rural population were removed, as the maximum value of ADP_A is equal to the total rural population. Thus, the buffer iteration process (which accounts for only rural population) would not ever meet the census estimate. 7 districts were removed due to this last criterion –of these, 2 were major cities, and 4 had an ADP_C to rural population difference of less than 5%, meaning these districts would still theoretically meet the threshold with a buffer applied across the entire rural area of the district. However, encompassing the entire rural area of a district irrespective of its proximity to cropland is likely to violate assumption (ii) described above. Hence, caution should be used if trying to draw inference from these districts.

In the case of predominantly rural states which had minimal (<1%) cropland area, there are likely two contributing factors. First, the states themselves have a low population density and are dominated by uncultivated rural land. Second, due to population sparsity there may be greater fragmentation of cropland, particularly small-scale agriculture such as household farms, which are not identified by the Dynamic World land cover layer. In applying this methodology to these areas, it is advised to reduce the area under analysis and consequently increase the spatial resolution of ancillary data. Dynamic World, for example, can be extracted to the 10m resolution (Brown *et al.*, 2022), and the coarse GHS-SMOD layer could be replaced with a finer grain building footprint layer – such as the GHS built-up surface layer, also produced at the 10m level (Pesaresi *et al.*, 2013).

4.6 Limitations

As mentioned in Section 1.2, the lack of a consistent and well-established definition for agricultural dependence is a key limitation of this study. The census-estimated calculation for this study, ADP_{C5}, is designed to account for both agricultural labourers and their dependents who are unrepresented in labour statistics. However, relying on the ratio of total workers to non-workers is a crude proxy for estimating the number of non-working dependents. The definition relies on the assumption that agricultural households are of similar size and structure to the average household across the whole population. There are significant differences in household structure and number of dependents between urban and rural areas, and agriculture and other industries (Marois, Zhelenkova and Ali, 2022), but the high proportion of rural and agricultural workers in India is likely to skew the overall average closer to the true value for this subset. By performing each analysis at the district level, the method is also designed to capture variations in disparate districts – such as highly urban or highly rural areas, which will have very different labour force dependency ratio patterns.

This analysis could be tailored further if more sophisticated demographic data were available – such as detailed information on the age structure of the workforce, with older adults in India overrepresented in agriculture (Chattopadhyay *et al.*, 2022), on the prevalence of unpaid agricultural work, with Indian women significantly underrepresented in labour force statistics (Swaminathan, 2020), and on the household structure in rural and agricultural families distinct from the population as a whole.

There are also weaknesses in relying on census and administrative data for validation of disaggregated population estimates. Despite being recognised as the authoritative source for most countries, census statistics suffer from uncertainty that is often poorly understood and overlooked, and in disaggregation these errors can propagate and impact the results of downstream analyses (Freire *et al.*, 2020). Gridded population data can also be a source of error, and at small spatial scales inherently have higher uncertainties. Nilsen et al. (2021) argue for incorporating Bayesian modelling to quantify uncertainty in geostatistical estimates, and a similar approach could be a valuable addition to this method.

Ideally, the method would be modelled on input layers that share the same input resolution. Although cropland data (Dynamic World) and gridded population data (WorldPop) are both made available up to a 100m resolution, the GHS-SMOD dataset is limited to 1km. Analysis at this coarser scale is likely to increase overall variation in buffer radii, as observed in the comparison between scales of cropland data. Alternative sources of urban/rural classification data could mitigate this effect in future research.

In addition to spatial resolution, the temporal resolution of the input data also dictates how meaningful output estimates will be. India has one of the fastest growing populations globally, and is simultaneously undergoing a rapid process of urbanisation and demographic change (Gu, Andreev and Dupre, 2021; Marois, Zhelenkova and Ali, 2022). Therefore, the spatial form and demographic structure of cropland and rural areas in India are going to be considerably different now compared to when the last population census was conducted, over a decade ago. The completion of the upcoming Indian census, scheduled to begin collection in January 2024 (*The Times of India*, 2023), will provide an important update to these and other census-derived population estimates.

4.7 Transferability

The method has been developed on open source software, and with typical computing resources, to maximise the reproducibility for use in other settings. All the code for analysis can be accessed from the project's Github repository, including details of the python working environment, packages required, and their versions.

The methodology is general enough that it can in theory be transferred to other study regions with only minor adjustment. Each of the spatial inputs – WorldPop, Dynamic World, and GHS-SMOD – are released as global layers, meaning that they could be easily reproduced in another setting. Assessment of how the performance compares given different landscapes, socioeconomic structures, and baseline data availability would improve confidence of the approach an understanding of the strengths and limitations.

The most important consideration for applying this methodology to a new region is the availability of demographic data, for use in calculating the ADP_C reference values – which forms the validation for calibrating buffer radius. This data does not have to be restricted to censuses but should be representative at a sufficiently low spatial unit (e.g., the district level in India), and have information on the agricultural labour force and agricultural and rural households. Additionally, the scale of the target area will impact the appropriate spatial resolution for data inputs, and consequently the computation time required to perform the analysis.

4.8 Opportunities

Understanding where people live, and the social and economic characteristics of those populations, is core to providing adequate, efficient, and targeted services and investment. A map of spatial distribution of agricultural population means that for any given point or area within the study extent, the ADP around that location can be quickly estimated. This has particular benefit, for example, in projects such as prioritising reservoir rehabilitation in Southern India (Vanthof and Kelly, 2020), where point data of tank location would allow comparison of demand across a set of tanks, and thus prioritise those which will have the greatest benefit for development.

Looking forward, the field of remote sensing is in a phase of rapid development and expansion, as research, industry and government capitalise on the opportunities opened by advances in satellite technology, data availability, and the increasing sophistication of machine learning models. In just the last few months, both Google and Microsoft have released large data updates for South Asia to their high resolution, open access, building footprint datasets (see, for example, the dataset introduced in Sirko *et al.*, 2021). New sources of data such as these, alongside cloud-based analytical software such as Google Earth Engine and Microsoft Planetary Computer, represent further opportunities to tackle this research problem in novel and innovative ways. The methodology presented here should be viewed as a first step in testing the concept of mapping agricultural populations and provides an opportunity for refining and extending these insights further.

To maximise accessibility for research and collaborative efforts, the ADP map generated for all of India has been published to this project's Github repository, available as a GeoTIFF file. This file can be freely downloaded for use in further research or policy applications and is easily integrated into common GIS platforms such as QGIS or ArcGIS.

5. Conclusion

As a subset of the broader population, agricultural dependent populations in India have unique service and development needs, sit at the forefront of policy efforts to address food security/insecurity, and face increased vulnerability due to a changing climate and more frequent disaster events. Understanding the spatial distribution of this population, beyond district level administrative regions, provides a basis for tailoring service delivery, resilience planning, and crisis response. Disaggregated data is also essential for the monitoring of development goals and agendas, identifying small-scale geographic inequalities that may be masked by aggregated statistics over a larger region, in the ethos of the Sustainable Development Goals' 'leave no-one behind' agenda (United Nations, 2022).

This research project therefore set out to address a gap in current approaches for the spatial disaggregation of sociodemographic data, proposing and testing a methodology specific to the estimation of agriculture dependent populations at a small area scale, applied to the case study of India. The proposed method capitalises on existing, global, open data sources and open source software to ensure reproducibility as a priority. Although computation load varied considerably by state, the method was shown to be feasible for application to a very large area within the constraints of a typical resource setting. The low rate of district ineligibility, at less than 2%, also indicates that the method is transferable across a diversity of geographic settings.

The development of this methodology has highlighted several critical gaps in the literature that would benefit from future research in this topic. First, a thorough exploration of how best to define agricultural dependence, including a set equation for deriving this metric from labour force and household statistics, would increase confidence in estimates and capture more of the variation within the population. Second, an investigation of the factors that affect buffer radius, to uncover the patterns behind spatial clustering of high and low radii districts and provide insight into how much of this difference reflects true spatial patterns, compared with systematic over- or under-estimation due to unintended error. This research may also identify additional data sources that could be incorporated into the method to mitigate these error effects and improve accuracy.

Appendix

Supervisor Meetings

Date (Attendance)	Description
30/03 (FL, SA, JP)	First meeting to discuss scope of proposed project, direction, logistics, etc. Introduction to partner project – Sri Lanka water tanks with World Bank.
24/04 (FL, SA, JP)	Coding session focused on addressing issues in the Sri Lanka tanks analysis. Discussion of approaches to address these issues; how they may arise in India project.
23/05 (FL, JP)	Discussion of key dimensions of project – how to define agricultural dependence, proposals of research questions, draft Table of Contents structure. Set date for submission of Literature Review draft.
08/06 (FL, JP)	Comments/feedback on draft literature review.
12/07 (FL, JP)	Review updated draft Table of Contents. Discuss analysis issues (temporal and spatial alignment of input datasets, logic behind including rural areas in buffer analysis, transition of method away from QGIS to run independently in python).
22/08 (FL, SA, JP)	Review core features of methodology and results, and each of the output figures; opportunities for improvements to figures; explanation of ADP_C variation rationale; high level discussion of important content for Discussion section.

FL = Fulvio Lopane (Supervisor); SA = Sophie Ayling (Additional supervisor; PhD student); JP = Joe Post (Author).

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.

Brown, P.R. *et al.* (2019) 'Constraints to the capacity of smallholder farming households to adapt to climate change in South and Southeast Asia', *Climate and Development*, 11(5), pp. 383–400. Available at: https://doi.org/10.1080/17565529.2018.1442798.

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).

Chattopadhyay, A. *et al.* (2022) 'Insights into Labor Force Participation among Older Adults: Evidence from the Longitudinal Ageing Study in India', *Journal of Population Ageing*, 15(1), pp. 39–59. Available at: https://doi.org/10.1007/s12062-022-09357-7.

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).

FAO, IFAD, UNICEF, WFP, WHO (2023) *The State of Food Security and Nutrition in the World 2023: Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum*. Rome, Italy: FAO (The State of Food Security and Nutrition in the World (SOFI), 2023). Available at: https://doi.org/10.4060/cc3017en.

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).

Gu, D., Andreev, K. and Dupre, M.E. (2021) 'Major Trends in Population Growth Around the World', *China CDC Weekly*, 3(28), pp. 604–613. Available at: https://doi.org/10.46234/ccdcw2021.160.

Holt, J.B., Lu, H. and Yang, X. (2011) 'Dasymetric Mapping for Population and Sociodemographic Data Redistribution', in *Urban Remote Sensing*. John Wiley & Sons, Ltd, pp. 195–210. Available at: https://doi.org/10.1002/9780470979563.ch14.

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.

Marois, G., Zhelenkova, E. and Ali, B. (2022) 'Labour Force Projections in India Until 2060 and Implications for the Demographic Dividend', *Social Indicators Research*, 164(1), pp. 477–497. Available at: https://doi.org/10.1007/s11205-022-02968-9.

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.

Natale, F. *et al.* (2013) 'Identifying fisheries dependent communities in EU coastal areas', *Marine Policy*, 42, pp. 245–252. Available at: https://doi.org/10.1016/j.marpol.2013.03.018.

Nilsen, K. *et al.* (2021) 'A review of geospatial methods for population estimation and their use in constructing reproductive, maternal, newborn, child and adolescent health service indicators', *BMC Health Services Research*, 21(1), p. 370. Available at: https://doi.org/10.1186/s12913-021-06370-y.

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.

Pesaresi, M. *et al.* (2013) 'A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results', *IEEE Journal of Selected Topics in Applied Earth Observations*

and Remote Sensing, 6(5), pp. 2102–2131. Available at: https://doi.org/10.1109/JSTARS.2013.2271445.

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.

Sirko, W. *et al.* (2021) 'Continental-Scale Building Detection from High Resolution Satellite Imagery'. arXiv. Available at: https://doi.org/10.48550/arXiv.2107.12283.

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.

Swaminathan, M. (2020) 'Contemporary Features of Rural Workers in India with a Focus on Gender and Caste', *The Indian Journal of Labour Economics*, 63(1), pp. 67–79. Available at: https://doi.org/10.1007/s41027-020-00210-z.

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.

The Times of India (2023) 'Ground work for Census 2021 to start from Jan 2024', 5 July. Available at: https://timesofindia.indiatimes.com/city/goa/ground-work-for-census-2021-to-start-from-jan-2024/articleshow/101496121.cms?from=mdr (Accessed: 23 August 2023).

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.

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.

Zhang, M. (2014) 'The Effects of Spatial Resolution on the Maize acreage estimation by Remote Sensing', *IOP Conference Series. Earth and Environmental Science*, 17(1). Available at: https://doi.org/10.1088/1755-1315/17/1/012052.