# CASA dissertation: OVERVIEW Planning Document

Student Number: 22186878

**Guidelines from CASA Dissertation Handbook**

*The literature review should evaluate existing research, demonstrate contrasting and/or similar views whilst highlighting research gaps. Synthesise previous work / policy documents and provide a narrative through it whilst trying to show where the research gap is / where your question fits in. Don’t just list what authors have done in the past (e.g., Smith 2009 did x but Jones 2008 did y then Frank 2010 did x). Try to end the literature review with a concluding paragraph that concisely summarises everything within it and states what your work is going to contribute or address.*

*Think of this section as providing a story about what everyone else has done (whilst also showing issues / research gaps) and then what you are going to do.*

## Table of Contents

### Introduction

1. World Bank Water Project: Sri Lanka
2. Purpose of this study
   1. Present a broad scope, outside just the water tanks. Knowledge of agricultural populations in particular ties significantly into understanding food security, especially in areas that are at risk of changing climate and natural disasters.
   2. Literature review part 1 (embedded): Agricultural dependent populations (concept, relevance, research background
   3. Literature review part 2 (embedded): Indian context, particular study areas
3. Research Question + Objectives
   1. RQ: *How can the agricultural population in India be identified at small spatial scales/fine spatial resolution?* 
      1. Objective 1: Review existing methods for spatial disaggregation of demographic data
      2. Objective 2: Propose and evaluate a new method that combines dasymetric disaggregation and iterative extension (buffers)
      3. Objective 3: Scale the method up to subcontinent (all of India)

### Methodology

1. Literature review part 3 (embedded): Spatial disaggregation
   1. What does this term mean?
   2. World gridded population
   3. Examples of spatial disaggregation methods; their strengths and weaknesses.
   4. Binary dasymetric mapping: the main method to be used
2. Literature review part 4 (embedded): Scaling Geocomputation \*\*TBD if this section will be relevant\*\*
   1. Challenges/Limitations of computation at large spatial scales
   2. Options for overcoming this; discuss pros and cons of different approaches
3. Justification of chosen method
4. Presentation of chosen method
   1. Data Sources
      1. WorldPop
      2. GHSL
      3. DynamicWorld
      4. Indian Census
      5. Other?
   2. Analysis (step-by-step summary)
      1. Calculate the agricultural dependent population at a district-level, from census data (include relevant equations)
      2. Calculate the agricultural land (vector) at a district-level, from Dynamic World model
      3. Calculate the rural landscape (vector) at a district-level, from Global Human Settlement Layer
      4. Convert WorldPop raster into a vector geometry of gridded points
         1. Clip the gridded population points to rural landscape (from GHSL): RuPoints. This will be used as the set of points included in the buffer iteration.
         2. Clip the gridded population points to the agricultural land boundaries (from DynamicWorld): AgPoints. This will be used as the baseline estimate for distribution of ADP.
      5. Calculate aggregated AgPoints population estimate by district
      6. Validate AgPoints estimates against the district-level ADP calculated in Step (i).
      7. Use iterative buffer process to grow/retract AgPoints space across RuPoints space until validation threshold is met.

## Introduction

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

Justification

This study is a novel addition to the field as it extends upon existing methodologies used to estimate total population and applies this to the estimation of the agricultural dependent population. Additionally, the case study of India is designed to assess feasibility and performance at a large spatial scale, comparative to partner research testing proof-of-concept in districts of Sri Lanka (unpublished). Understanding the distribution of agricultural population in a region will provide a more accurate estimate of local demand on water resources.

Stretch out ‘scale’ component further; key aspect of study

Text

## Literature Review

### Introduction

This literature review 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 second section introduces spatial disaggregation methodologies, historical development, and applications, particularly regarding gridded population estimates of the world. Finally, the review highlights how this thesis addresses a gap in the literature and how the work is situated within the broader scholarship around spatial disaggregation estimates.

### Agricultural Dependent Populations

Agriculture represents 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 (REF), 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 total population, which skews towards urban areas, estimating rural and agricultural populations can provide an indication of demand on specific resources, such as water 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, 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.

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.

### 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). 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 are published at the district level, the most recent being conducted in 2011.

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.

### 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 1 provides a summary of current global grid models and their key characteristics.

Table 1: 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 1: 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).

### Conclusion

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.

## 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.

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.

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).

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.

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).

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.

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.