ILO Working Paper 165

March / 2026

Gridded-Labour Market Data in Ghana using Remote Sensing and Random Forest

Authors / Jin Yan, Charpe Matthieu, Mei Yang, Li Zeshuo

ISBN 9789220432976 (print), ISBN 9789220432983 (web PDF), ISBN 9789220432990 (epub), ISBN 9789220433003 (html). ISSN 2708-3438 (print), ISSN 2708-3446 (digital)

https://doi.org/10.54394/00033744

Information on ILO publications and digital products can be found at: www.ilo.org/research-and-publications

ILO Working Papers summarize the results of ILO research in progress, and seek to stimulate discussion of a range of issues related to the world of work. Comments on this ILO Working Paper are welcome and can be sent to charpe@ilo.org.

Authorization for publication: Schmidt, Dorothea

ILO Working Papers can be found at: www.ilo.org/research-and-publications/working-papers

Suggested citation:
Yan, J., Matthieu, C., Yang, M., Zeshuo, L. 2026. *Gridded-Labour Market Data in Ghana using Remote Sensing and Random Forest*, ILO Working Paper 165 (Geneva, ILO). https://doi.org/10.54394/00033744

Abstract

This study presents high-resolution (0.005) gridded labor market data, generated by downscaling district-level census data for Ghana using random forest algorithms and remote sensing. It addresses the lack of spatially disaggregated labor market data by mapping 17 employment categories – including age, gender, skills, status, sectors, unemployment, and NEET. Auxiliary data (64 variables) such as land cover, nighttime lights, infrastructure, and points of interest are integrated to capture demographic, economic, and participation factors. The model achieves high accuracy (R2 > 90% for most categories) and reveals significant spatial heterogeneity, with employment rates ranging from 10% to 98% across pixels. Results highlight urban-rural and North-South divides, as well as sectoral concentrations. Variable importance analysis underscores the role of built-up areas, nighttime light, road density, and vegetation health in predicting employment patterns, with specificity across different employment categories. The methodology advances beyond traditional GDP or population gridding by incorporating labor market complexity. Findings demonstrate the potential of machine learning and geospatial data to enhance socio-economic mapping in data-scarce contexts.

About the authors

Yan Jin, Associate Professor, School of Internet of Things, Nanjing University of Posts and Telecommunications & Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Nanjing 210023, China

Matthieu Charpe, Senior Economist, Employment Policy Department, International Labour Organization, 4 route des Morillons, 1211 Geneva, Switzerland

Yang Mei, School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Zeshuo Li, School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Table of contents

List of Figures

List of Tables

Introduction

Gridded population and GDP data are essential tools for effective government policy planning and international aid allocation. Gridded population data are crucial for health monitoring and urban development, enabling precise and targeted interventions (Stevens et al., 2015; Tatem, 2017; Leyk et al., 2019; Chen et al., 2019; Lebakula et al., 2025). Meanwhile, gridded GDP data allow policymakers to monitor economic development at sub-national levels, facilitating more informed and localized decision-making (Kummu et al., 2018; Chen et al., 2022; Jin et al., 2023; Wu et al., 2024).

Downscaling population statistics or socio-economic attributes differ significantly. Socio-economic attributes require associating a measure of economic activity with a population, reflecting the accumulation of physical and human capital, agglomeration effects, and the development and quality of political and economic institutions.

Gridded-socio-economic data requires to go beyond GDP, given its well-known limitations and the necessity to measure individual well-being through labour market indicators, human development indices, or poverty measures. However, there are few attempts to produce high-resolution poverty measures (Jean et al., 2016; Huber and Mayoral, 2024) or gender development indicators (Bosco et al., 2017) and there are no existing gridded data of the labour market.

In this paper, we propose the first set of high-resolution maps of labor market data, based on disaggregating district-level statistics from the 2021 housing and population census of Ghana at a 0.005-degree resolution. The rationale behind downscaling labour market statistics is the large heterogeneity at lower administrative levels. While the employment rate is 48% at the country level, the standard deviation is 11.2 percentage points at the district level, with a minimum of 16.5% in the district of Nabdam in the Upper East region and a maximum of 64.7% in the district of Ashaiman in the Greater Accra region (see section 2.1 and Figure 1).

One difficulty with high-resolution maps of the labor market is that no single indicator best describes it. While employment quantity, i.e., total employment, is the most discussed variable, labor market challenges are often related to specific subgroups (youth, older workers, females), underemployment (unemployment, NEET[1]), employment quality (skills, employment status), or sectors (agriculture, manufacturing, services). Employment is also often represented in rates necessitating estimates of working age population.[2]

The second difficulty is that the employment relationship depends on multiple factors that are challenging to capture via GIS and remote sensing data. Labor supply, i.e., the quantity of persons working or available to work, is directly related to population growth. However, additional factors must be considered. The level of economic activity is an important determinant of the labor market as it sets the level of labor demand. In particular, the challenge with high-resolution maps of GDP is producing a sectoral decomposition of GDP (Jin et al., 2023; Wu et al., 2024). Thus, downscaling labor market data requires combining approaches associated with population and GDP disaggregation.

[1] NEET stands for not in employment, education, or training.

[2] Employment rate is defined as total employment over working age population (15+). This work also includes the disaggregation of population categories at the pixel level (see section 4.1).

A third element central to the labor market is the decision to participate based on individual decisions. Individual decisions to participate in the labor market are driven by various characteristics such as education level, gender, family structure, years of experience, poverty, or sector of activity (see Aaronson et al. (2020); Klasen et al. (2021) on female labour market participation and Baah-Boateng (2013); Amankwah et al. (2022) on specific factors influencing labour market participation in Ghana).

The set of gridded-employment data is made of 17 employment related categories: total employment and its decomposition by age (youth, prime age, old), by gender (male, female), by skills (high skills, low skills), by status (self-employed, salaried), by sectors (agriculture, mining, manufacturing, industries and services), unemployment and NEET. The auxiliary data collected comprises 64 categories aimed at capturing demographic factors, the level of economic activity, and the decision to participate in the labor market. These data include GIS data, remote sensing data, and points of interest. The labor market statistics come from the employment module of the Ghana 2021 Housing and Population Census and follow the 2019 International Conference of Labour Statisticians' definitions. The model is trained using a random forest algorithm based on district-level labor statistics and auxiliary data. Pixel disaggregation is performed at a resolution of 0.005 degrees.

Methods for producing gridded population data have evolved rapidly, moving from simple areal weighting methods – which assume a uniform distribution within the target zone – to dasymetric approaches that leverage a wealth of GIS and remote sensing techniques to estimate population at the grid level (Dobson et al., 2000; Bhaduri et al., 2007). Dasymetric methods, in turn, employ a wide range of techniques, from statistical models (Briggs et al., 2007; Chen et al., 2019) to machine learning (Azar et al., 2010; Stevens et al., 2015; Wang et al., 2020). Global datasets, such as LandScan (Lebakula et al., 2025) and WorldPop (Tatem, 2017),[3] illustrate these advances. Beyond downscaling methodologies, bottom-up approaches based on micro-censuses can also contribute to the production of country-wide gridded population data (Wardrop et al., 2018; Leyk et al., 2019).

Attempts to go beyond demographics to include socioeconomic indicators often focus on GDP and poverty. Global gridded GDP data for multiple years rely on statistical techniques and auxiliary data that measure population and urbanization. Murakami et al. (2021); Murakami and Yamagata (2019) propose a global downscaling of GDP at a 10-year frequency, covering the period 1850{2100 (at 0.5° and 1/12° resolutions), using weights estimated with gradient boosting techniques. Wang and Sun (2022) project national and regional GDP data to the pixel level using a combination of nighttime light and population per pixel for 185 countries, also at a 10-year frequency starting in 2005.

GDP data for a single country or year often include a large set of auxiliary data and rely on machine learning. In a production function approach, GDP depends on demographics as well as the stock of capital and technological progress. This requires auxiliary data that measure economic activity, such as nighttime light and auxiliary data that measure economies of scale/agglomeration effects-captured through distance to infrastructure/ institutions. Chen et al. (2022) develop a deep learning framework to map China's GDP at high spatial resolution by integrating multiple geospatial big datasets, including nighttime lights, land use, road networks, and points of interest. A primary objective is to model sectoral GDP (agriculture, manufacturing, and services) to reflect the different sets of auxiliary data driving sectoral economic activities. The sectoral

[3] https://landscan.ornl.gov/; https://www.worldpop.org/

approach is motivated by the limitation of nighttime light to adequately capture agricultural GDP, as increases in agricultural production is mainly indirectly reflected in higher light intensity. Wu et al. (2024) propose a multiscale fusion residual network model that combines a wide range of remote sensing and point-of-interest (POI) data for the UMARYR region of China. Similarly, Jin et al. (2023) propose high-resolution maps of sectoral GDP in Thailand based on remote sensing and POIs, which measure natural characteristics, demographics, economic activity, and sectoral dimensions. The model is estimated using random forest area-to-area regression kriging.

Alongside GDP, modelling socio-economic attributes includes gridded poverty data. Jean et al. (2016) train a convolutional neural network model to estimate expenditure data by combining night-light data with daytime satellite images that can proxy poverty indicators, such as roofing material.

Similarly, Huber and Mayoral (2024) estimate a random forest model between the DHS asset index[4] and auxiliary data for 11 African countries. They then project consumption per capita at the grid level for 42 African countries. Bosco et al. (2017) produce gender-disaggregated development indicators such as female literacy, child stunting, and contraceptive use. They rely on georeferenced DHS surveys and model the relationship between these development indicators and auxiliary data using Bayesian geostatistical and machine learning models for Nigeria, Kenya, Tanzania, and Bangladesh.

Section 2 presents the materials and methods. Section 3 presents the model results, model evaluation, high-resolution maps and importance analysis. Section 4 presents an application of the gridded labour market data, looking at employment rates and urban employment. Section 5 discusses the data, methods and results. Section 6 concludes.

[4] Demographic and Health Surveys.

1 Materials and Methods

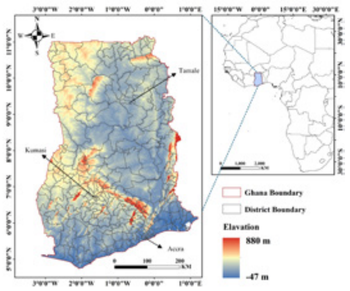
1.1 Study area and main labour market statistics

The study area is Ghana, which is composed of 16 regions and 261 districts. In 2021, the population of Ghana is 30.8 million and the working-age population (15+) is 19.9 million. Using the definition of employment set out in the 2019 International Conference of Labour Statisticians, total employment is 9.6 million, implying an employment rate of 48% (employment to working-age population). Employment differs greatly across age groups and gender. The employment rate of young persons (15-24) is 22.9% compared to 64.8% of prime-age workers (25-54) and 39.6% of older workers (>55). The employment gap between men and women is 8 percentage points. The unemployment rate is low due to the absence of unemployment insurance (3.6%). The share of workers with a low education level (basic and less than basic education) is widespread at 63.4%. This is reflected in the share of self-employed workers that account for the bulk of total employment (63.9%). Ghana's sectoral composition is characterized by a transition from agricultural employment (26.5%) to services (57%).

Mining is small in terms of employment (1.1%) but gold mining, in particular, is important to the economy of Ghana. Manufacturing and industries account for nearly 15% of total employment, a relatively high share compared to other sub-Saharan countries. In order to avoid a border effect when performing the downscaling, the auxiliary data described below are gathered for Ghana as well as for the neighboring countries districts.

▶ **Figure 1: Selected Labour Market Indicators**

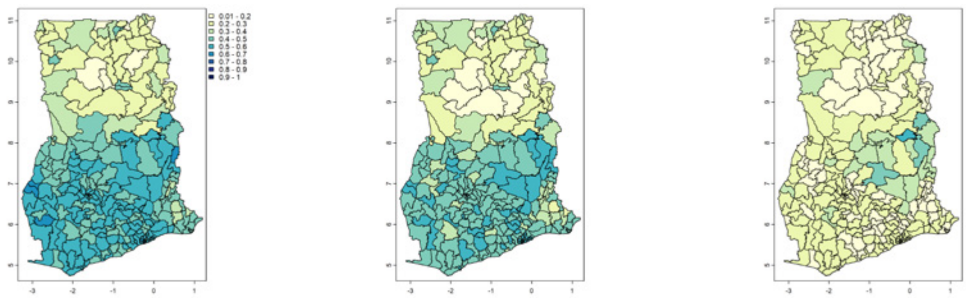
(a) Location of the study area



(b) Employment rate - districts

(c) Employment rate youth - districts

(d) Employment rate female - districts



1.2 Data description and processing

The methodology follows in this paper, data collection, data preparation, model estimation and the generation of gridded-labour market data is illustrated in the Figure 2. This study utilizes two types of datasets. The first type consists of census based population and labour market data at the districtlevel, sourced from the Ghana Statistical Services. These data are treated as the dependent variable.

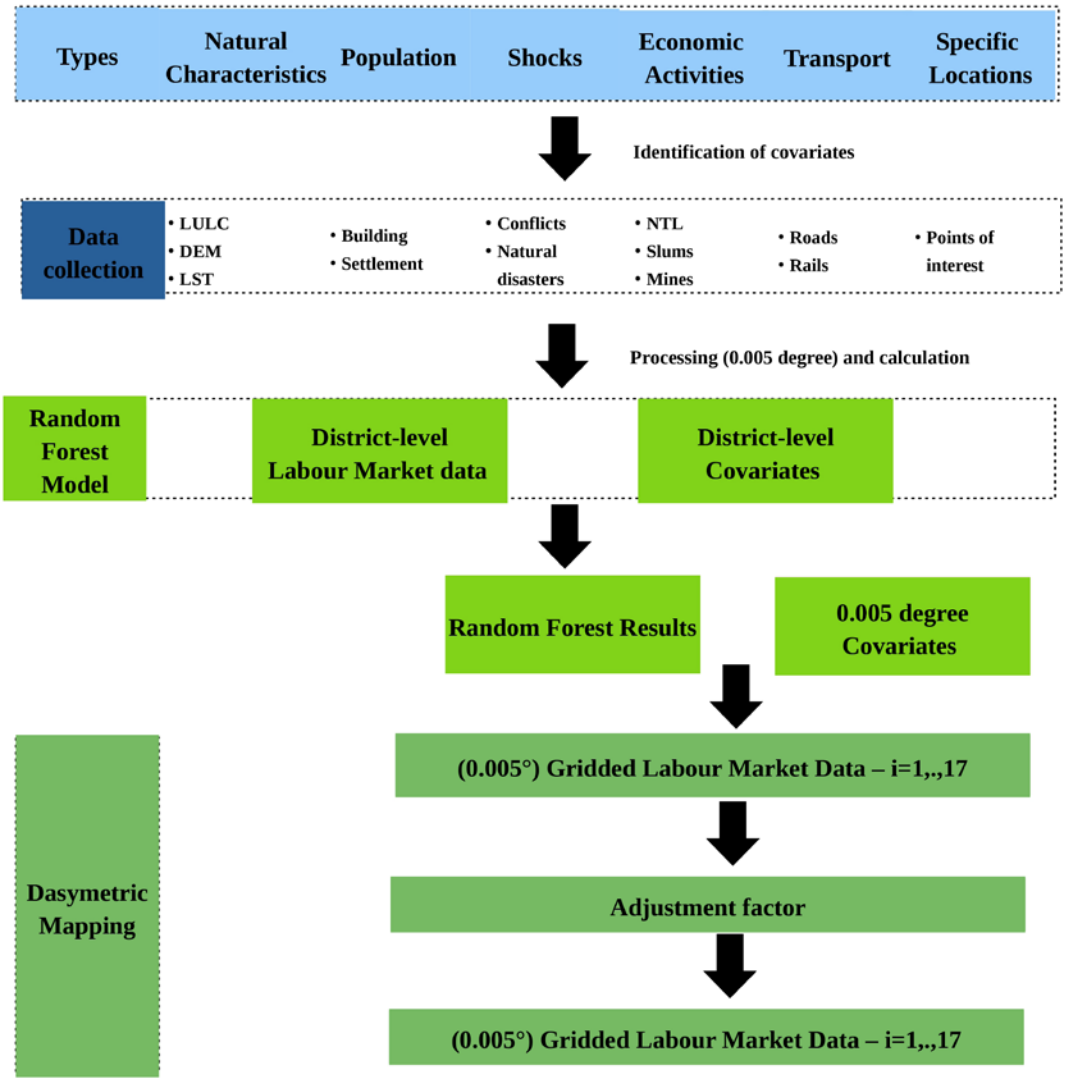
The second type comprises gridded auxiliary data at a resolution of 0.005 degrees, which are considered independent variables. Ghana Statistical Services provides a 10% representative sample of the Housing and Population census for the year 2021, together with the corresponding GIS-delineated administrative boundaries, national, regional and district levels. We compute 17 employment categories at district level using the 2019 International Conference of Labour Statisticians' definitions of employment categories and applying the population weight provided with the census data. These employment categories are total employment and its decomposition by age (youth, prime age, old), by gender (male, female), by skills (high skills, low skills), by status (self-employed, salaried), by sectors (agriculture, mining, manufacturing, industries and services), unemployment and NEET (Not in Education, Employment or Training). The age categories are as follows: total 15 and above, youth 15-24, prime age 24-54, older 55 and above. High skilled is defined as intermediate or advanced educational attainment, and low skilled is defined as less than basic and basic educational attainment.

Self-employment aggregates own-account workers, members of producers' cooperatives, contributing family workers, and the self-employed. Agriculture includes forestry and fishing; industry aggregates electricity, gas, water supply, and construction. Services include public and private services. NEET refers to youth not in employment, education, or training. While downscaling is performed on the employment categories in levels, the labour market situation of a country is best characterized in rates. We therefore compute 7 demographic categories in order to be able to produce high resolution maps of employment rates and related categories[5].

The auxiliary data are grouped into 5 categories. The first category primarily measures natural characteristics, closely following methods in the literature for downscaling demographic data, aiming to identify populated areas versus natural areas. ESA-WorldCover provides 9 categories of land cover classification at a 10m resolution, including tree cover, shrubland, grassland, cropland, built-up areas, sparse vegetation, permanent water bodies, herbaceous wetlands, and mangroves[6]. We generate a gridded set of auxiliary variables by calculating the percentage of different land cover types at a resolution of 0.005 degrees. Digital elevation model measures the slope and ruggedness the terrain. Copernicus digital elevation model has a resolution of 30m, based on which 4 measures are extracted: DEM, slope, ruggedness and proportion of slope ≤ 5 degrees. These data are resampled to a 0.005-degree grid. Vegetation is measured through the median of three index: the Enhanced Vegetation Index (EVI), the Normalized Difference Vegetation Index (NDVI) and the Near-Infrared Reflectance of Vegetation (NIRv). These indexes are made available from MODIS MOD13A1.061 product on a 16 days frequency and at a resolution of 500m. Median Land surface temperature are provided by MODIS MYD11A1.061 product. The frequency is daily and the resolution is 1km by 1km. Another measure of vegetation health and productivity is the Net Primary Production that measures the amount of carbon fixed by photosynthesis. It is one of the MODIS product (MOD17A3HGF.061) available at an annual frequency and at a 0.05 degree resolution. Annual Net Primary Production (NPP) is calculated by summing all the 8-day Net Photosynthesis (PSN) values from the MOD17A2H product for a given year. The PSN value represents the difference between Gross Primary Productivity (GPP) and Maintenance Respiration (MR). Rain fall estimates come from the CHIRPS global data. The resolution is 0.05 degree. The daily observation are summarized annually via minimum, maximum, mean and total precipitations. The Normalized Difference Water Index (NDWI) assesses water content in vegetation and soil. The corresponding MODIS product is MODIS/006/MOD09GA. The daily data are summarized annually with minimum, maximum and median. The resolution is 0.005 degree. Rivers and lakes are identified by the database HydroRIVERS and HydroLakes. The gridded data for the distance to the Longest Upstream Path and Downstream Path were used, based on the HydroRIVERS. Additionally, the river network density was calculated using the river network provided by the database. The nearest distance from the raster center points to the merged lake data was calculated using the lake polygon data. In addition, using OpenStreeMap, a measure of the distance from the pixel center to the nearest water categories (lake, canal, drain, river, stream) complements the HydroSheds data.

5 total population, working age population total and decomposed by age (youth, prime age, old), by gender (male, female).
6 11 categories minus "snow and ice", "moss and lichen" categories.

▶ **Figure 2: Methodology for gridded-labour market estimation**



The second category gathers auxiliary data identifying population location. The Open Buildings 2.5D Temporal Dataset identify building presence, building height and building proportion at a resolution of 0.5m. The GRID3 Settlement data provides yearly polygon at a resolution of 0.005 degree. The settlement polygons are matched to the area raster and the proportion of settlement is measured for each cell.

Large shocks that are likely to impact the local economy are identified via conflict data and natural disaster data. The Uppsala conflict data program identifies individual incident of lethal violence and provides an estimate of the number of casualties as well as the nature of the conflict. Given the small spatial resolution, the unit is the number of casualties within a 10km radius as well as the distance to the nearest conflict for the corresponding year. Natural disasters are measured via droughts, floods, earthquakes, and landslides. The Geocoded Disasters Dataset (GDIS) provides polygons and point estimates for various natural disasters. Droughts are measured over the period 1999-2018. Past floods are identified by superimposing both the Global

Flood Database and Geocoded Disasters (GDIS) Dataset over the period (2009-2018). Based on the Global Flood Database, a second measure of floods is constructed by identifying the pixels following into flooded area for the latest year available. The earthquake data is provided by the USGS Global Earthquake Dataset.[7] The landslide data is provided by the Global Landslide Catalog – NASA Goddard (1970-2019), and we extracted landslide records from 2000 to 2019.[8] For the above four types of natural disaster data, the distance from the center of each grid cell to the nearest disaster category is constructed on a 0.005-degree grid.

Local economic activity is estimated using annual nighttime light satellite imagery with a spatial resolution of 15-arc seconds. We utilize VIIRS data sourced from Black Marble, specifically focusing on snow-free radiance from all angles. This annual data is adjusted to account for sunlight, glare, moonlight, cloud cover, and aurora-related lighting effects. The data is measured in radiance units, and pixels with a value of 65535, as well as those flagged for low quality, are excluded from the analysis.

Beyond economic activity and employment level, income distribution and employment quality are determinants of some of the employment categories projected in this work. As it is rather complex to measure income distribution through geospatial data, a valid candidate is slums. There are no global slums data. However, Li et al. (2023) estimate slums in four sub-Saharan countries using urban morphology assuming that areas with tightly packed, small buildings often suggest a high occurrence of slums.

The final category includes auxiliary data that help identify sectoral dimensions as well as the density of infrastructures (such as roads and schools), which may capture certain factors driving labor market participation. The data on mines are derived from three sources: i) the Global Mining Footprint (Tang and Werner, 2023), ii) the USGS Mineral Industries and Related Infrastructure of Africa (Padilla et al., 2021), and iii) the Africa Major Mineral Deposits Dataset from the Regional Centre for Mapping of Resources for Development. The first database consists of polygons, while the other two are point data. For each pixel, a measure of the distance to the nearest polygon or the nearest point data is calculated. Transport infrastructure data are extracted from OpenStreetMap for the corresponding year. The roads data are divided into six categories: highway, major, minor, unsuitable for cars, very small, and unknown. For each category, density at the pixel level is constructed. For highways and major roads, a distance to the nearest road segment is produced. A similar measure is calculated for railways. Points of interests (POIs) are extracted from OpenStreetMap. These POIs are classified into 15 main types, including public, education, health, leisure, sports, catering, indoor accommodation, outdoor accommodation, shopping, money, tourism, fuel parking, transport, workplaces, and others. The density of these 15 categories is an important factor in downscaling labour market statistics, as their concentration can reflect both the sectoral composition of employment and capture the significance of specific factors, such as health, education, and religion, which may influence labour market participation decisions. The density is measured via a kernel density estimation with a bandwidth equal to 3000.

[7] https://developers.google.com/earth-engine/datasets/catalog/GLOBAL_FLOOD_DB_MODIS_EVENTS_V1
[8] https://gee-community-catalog.org/projects/landslide/

Table 1: Data source

Type	Dataset	Description	Source	Resolution
Natural characteristics	LULC DEM Vegetation LST NPP Precipitation Water indices River/Lake River/Lake	9 classes elevation/ruggedness slope / slope ≤ 5 degrees EVI,NDVI, NIRv (median) in Kelvin (median) gC/m²/year in mm (min, max, mean, total) NDWI (median) distance to LDU/LUP /density distance to 5 water categories	ESA WorldCover Copernicus MODIS MOD13A1.061 MODIS MYD11A1.061 MODIS MOD17A3HGF.061 CHIRPS MODIS MODIS/006/MOD09GA HydroSHEDS (HydroRIVERS/HydroLAKES) OSM	10m 30m 0.05 degree 0.01 degree 0.005 degree 0.05 degree 0.005 degree 0.005 degree line/poly-gons
Population	Building height Settlement	height (median) building proportion proportion	Open Buildings GRID3	0.5m 0.005 degree
Large shocks	Conflict Natural disaster	casualties within 10kn radius distance to nearest conflict Floods (distance) Flooded Droughts (distance) Earthquake (distance) Landslide (distance)	UCDP Global Flood Database Geocoded Disasters (GDIS) Dataset Global Flood Database GDIS USGS Global Earthquake Dataset Global Landslide Catalog	point data Point data (2009-2018) Point data (2009-2018) Point data (2018) Point data (1999-2018) Point data (1923-2024) Point data (1970-2019)
Economic Activity Income distribution	Nighttime lights Poverty	NTL, NLVI, NLVS, VCNLI Slums proportion	blackmarble Li et al. (2023)	0.005 degree 100m

Type	Dataset	Description	Source	Resolution
Sectoral	Mines	distance to nearest mines	Global mining footprint Tang and Werner (2023)	polygons
composition Transport		distance to nearest mines distance to nearest mineral distance to nearest rail distance to nearest roads density	USGS Padilla et al. (2021)	point data point data line
			Africa Major Mineral Deposits Dataset OSM	line line
	Rails Roads		OSM (highway, major roads) OSM (6 categories)	point data
Specific locations		kernel density	OSM (15 categories)	
	Points of interest			

This table presents the source of the auxiliary data used for the down-scaling of demographic and employment data.

In total, there are 64 auxiliary data. Each raster data has been resampled to a resolution of 0.5km by 0.5km. In addition, each raster is being normalized according to $\frac{x - min}{max - min}$, with *min* and *max* the minimum and maximum of data x. Normalization makes the different scales comparable and can help the machine learning algorithm to converge faster. In addition, the pixel population is restricted to populated pixel. A pixel is considered as populated if either nighttime light, building height, building proportion, settlement or slum is non-zero.

The POIs are transformed into a density layer with a resolution of 0.005 degree, for each category using a Kernel Density Estimation (KDE) following Wang et al. (2020); Jin et al. (2023). The KDE is estimated with the following bandwidths: 500m, 600m, 1000m, 1200m, 1500m, 3000m, 5000m and 8000m. By experimenting with various bandwidths, we noticed that increasing the bandwidth leads to a decrease in the peak value of the kernel density results. Since the trends in the kernel density distributions are consistent, and to prevent the bandwidth from being too large – thereby obscuring local features – it is typically advised to set the bandwidth to 1 to 2 times the raster resolution. This approach balances smoothing with the preservation of detail. Hence, we can select 1000 meters as the bandwidth for further modeling.

The gridding of labour market statistics is done via the downscaling of population density. Given the potential issues related to the changing size of the pixel with latitude changes, population density is measured as population divided by the number of pixels. At 0.005 degree resolution, there is only one pixel. However, the machine learning algorithm trains the data at the district level, an administrative level for which the number of pixel matters. The density formulation of our variables is also relevant for robustness check involving larger resolution such as 0.01 degree.

1.3 Downscaling methodology

We use a Random Forest model to downscale population and labor market data. The process is divided into three steps. The first step is to collect and process the data. As mentioned in the previous section, district-level population and labor market data serve as the dependent variables, while five categories of auxiliary variables are used as independent variables. Here, the district-level auxiliary variables are summarized from the 0.005-degree gridded auxiliary variables. The second step is model construction and training. district-level population, labor market data, and auxiliary variables are used as inputs to train the Random Forest model. The trained model is then applied to predict population and labor market data at the 0.005-degree grid level using the 0.005-degree gridded auxiliary variables as inputs. To achieve high-precision mapping, district calibration factors are introduced to match official census statistics. Finally, district-level population and labor market data are used to evaluate the accuracy of the 0.005-degree grid predicted data.

The specific representation of the random forest algorithm in this study is as follows. First, a random forest model $\int_{rf}^{d} 0$ is established at the district-level based on the equation $y^d = \int_{rf}^{d}(x_1^d, x_2^d, ..., x_k^d)$, where y^d denotes the population variables or labor market variables in a district-level, and x_k^d denotes the auxiliary variables at that location, with *k* representing the total number of auxiliary variables.

▶ Table 2: Tuning parameters that minimize RMSE

names	pop	wap	wap_y	wap_{pa}	wap_{old}	wap_w	wap_m	Et	Ety	Etpa	Etold	Etw
mtry	16	8	10	9	5	8	7	10	12	10	5	8
ntree	300	300	800	300	500	50	500	500	50	500	50	500
names	Etm	Ut	neet	Eths	Etls	Etse	Etees	Etag1	Etmi1	Etma1	Etind1	Etserv1
mtry	13	4	19	10	13	7	9	11	20	18	13	13
ntree	500	800	100	300	300	800	300	500	100	800	500	500

This table presents the combination of variables and trees that minimizes the RMSE in the tuning exercise for each of the population and employment categories.

Then, the random forest model $\int_{rf}^{d}()$ at the district-level is used to predict the target variable y^g at the 0.005-degree grid resolution, as shown by the equation $y^g = \int_{rf}^{d}(x_1^g, x_2^g, \ldots, x_k^g)$ where x_k^g denotes the auxiliary variables at the 0.005-degree grid cell. Finally, to match the district-level data, an adjustment factor is applied to correct the predicted target variables at the grid resolution, as shown by the equation $y_p^g = y^g \times \frac{y^d}{\sum_{j=1}^{n_i} y_{i,j}}$ where y_p^g is the corrected predicted variable, y^g is the predicted target variable at the grid resolution, y^a is the district-level target variable, *i* represents the district, *j* represents the pixel, and n_i denotes the number of pixels in district *i*. By repeating three steps, first building a random forest model at a coarse resolution, second applying the random forest model at a finer resolution, and third performing district-level correction on the predicted results, a target variable map at 0.005 degree resolution can be created.

After establishing the random forest model, we used a repeated k-fold cross-validation scheme to ensure valid predictions and further improve the model's performance through parameter tuning.[9] To train the data, the number of folds is *k* = 10 and the number of repeats is n=3. The database is split randomly into 10 subsets of equal size. For each of the 10 iterations, one fold is used for testing, and the remaining *k*-1 folds are used for training. This process is repeated 3 times, resulting in a total of 30 evaluations. The parameters to tune the model are defined by the number of trees to grow in the Random Forest. The number of trees takes the values 5, 10, 20, 50, 100, 300, 500, and 800. A second parameter is the number of variables randomly selected at each split, i.e., the process of dividing a node into two or more sub-nodes. This value ranges from 1 to 30. In the tuning process, the accuracy of each combination of parameters is assessed using the root mean square error, mean absolute error, and R-squared. Out of these three metrics, we rely on the root mean square error.

The result of the parameter tuning is presented in Table 2. The mtry parameter determines how many features are randomly selected as candidates for splitting at each node in a decision tree within a Random Forest. Introducing this randomness helps to reduce the correlation between individual trees, which enhances model robustness and lowers the risk of overfitting. The mtry parameter that minimizes the RMSE ranges from 4 (unemployment) to 20 (mining), with an average of 10 features selected across categories. The number of trees that minimizes the RMSE, in combination with the mtry parameter (ntree), ranges from 50 to 800, with an average of 410. The variables selected through the mtry parameter cannot be directly compared with those appearing in the variable importance (see discussion below), as the latter helps to identify which features are most important for making accurate predictions. However, there is an indirect connection: the algorithm tends to favor variables in mtry that produce efficient splits, thereby reducing node impurity.

[9] We use the R package, specifically the caret and randomForest libraries.

The optimal number of variables sampled for the population categories is similar, although sometimes slightly smaller, compared to the variables for the corresponding employment categories.[10] A difference is the number of trees, which is larger for the population categories than for the employment categories. The optimal number of variables sampled, as well as the number of trees, tends to increase for the remaining employment categories and sectoral employment categories. This is especially true for the latter. One explanation is that these categories measure underemployment and employment quality, while employment levels are closely related to population and demographics in low-income countries. This is particularly apparent for agriculture, which remains a subsistence activity in Ghana and is therefore closely related to population. In contrast, mining, industries, and manufacturing are sectors that require specific skills and competencies. The number of auxiliary variables and trees necessary to fit these categories may increase with the complexity of the category being estimated.

The estimation of the number of trees and variable selection is then incorporated into a Random Forest estimation. This model produces different metrics for assessing the quality of the estimation, including the mean square residual and the percentage of variance explained. Additionally, the model generates variable importance and node purity metrics to discuss the main auxiliary variables behind the different population and employment category projections. The resulting model is then used to predict the pixel values for the different categories. The predictions from all the individual trees are averaged to produce a more accurate and robust prediction.

[10] 15+, youth, prime age, old, women, men

2 Results

2.1 Model evaluation

Our two measures of estimation quality are mean square residual (MSR) and percentage of variance explained.[11] Across the different employment categories, the MSR ranges from 0.07 to 0.2. These low values indicate a good accuracy of the model. The model's predictions are closer to the actual observed values, suggesting a better fit of the model to the data. The good performance of the model is confirmed by the percentage of variance explained that exceed 90 percent to the exception of NEET (89%) and some of the sectoral employment (agriculture 68% and mining 67%). Seven of the 17 employment related categories achieve a percentage of variance equal to 94%. The relatively weaker performance of the model regarding agriculture is consistent with existing work attempting to downscale sectoral GDP data (Wu et al., 2024). This might be related to the need to gather additional auxiliary data to measure crops yields with more precision. Wu et al. (2024) measuring sectoral GDP data for the UAMRYR province in China achieve a R2 of 0.63 for agriculture, 0.65 for manufacturing and 0.74 for services using random forest.[12] We find as well no striking differences between categories that measure employment levels and categories that measure employment quality. Note that due to space limitations, the metrics regarding the population categories are reproduced in the appendix.

▶ **Table 3: Residual sum of square and percentage of variance explained**

	(1)	(2)		(3)	(4)
names	MSR	% of variance	names	MSR	% of variance
Et	0.16	0.94	Eths	0.18	0.94
Ety	0.13	0.92	Etls	0.15	0.92
Etpa	0.16	0.94	Etse	0.17	0.92
Etold	0.12	0.92	Etees	0.17	0.94
Etw	0.15	0.93	Etag	0.10	0.68
Etm	0.14	0.94	Etmi	0.07	0.67
Ut	0.10	0.92	Etma	0.16	0.91
neet	0.15	0.89	Etind	0.10	0.94
			Etserv	0.20	0.94

This table presents the mean square residual and the percentage of variance explained for each of the employment related categories.

2.2 Gridded labour market maps

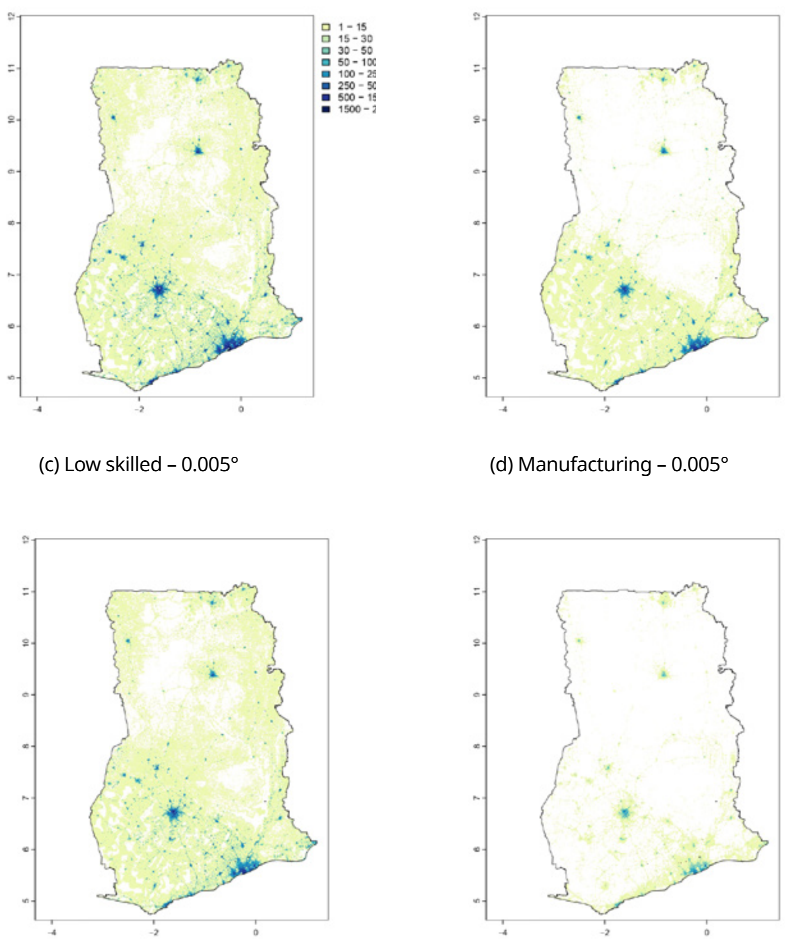
The gridded employment data are mapped in Figure 3 at a resolution of 0.005 degrees. Due to space limitations, we present only a selection of the 17 categories. The remaining categories can

11 $MSR = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,\ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

12 They achieve higher R^2 using Res-FuseNet model instead of random forest.

be found in the appendix, Figures 8 to 11. The different employment categories are represented in absolute values for the following breaks: (0, 15, 30, 50, 100, 250, 500, 1500, 2500). The data effectively represent the higher population density in the main urban agglomerations of Ghana. Urban areas are mostly located in the southern half of Ghana, between the towns of Accra and Kumasi. The higher population density also translates into higher employment values (Panel A). Employment is also concentrated along the coast of Ghana and along the main transport infrastructures connecting the cities. In the northern half of the country, employment is sparser and is concentrated around four main agglomerations: Tamale, Wa, Bolgatanga, and Bawku. The last two agglomerations are located close to the border with Burkina Faso.

▶ Figure 3: Spatial distribution of (selected) employment categories



(a) Employment – 0.005°

(b) Employees – 0.005°

(c) Low skilled – 0.005°

(d) Manufacturing – 0.005°

Beyond employment quantity, employment quality is a key dimension to depict local labour markets. We illustrate employment quality by plotting salaried employment in Panel (b). The distribution of salaried employment at the pixel level is similar to the distribution of total employment, reflecting population density. However, salaried employment is predominantly found in the largest urban areas. It is less prevalent in secondary cities as well as in the areas connecting these secondary cities. In the northern half of Ghana, salaried employment is present in limited areas.

In addition to employment status, we illustrate employment quality by mapping low-skilled workers (in panel (c)). The distribution of low-skilled employment mirrors that of total employment, confirming that a significant portion of the workforce has low educational attainment. A distinctive feature is that in city centers, low-skilled employment is less prevalent than total employment, reflecting the concentration of high-skilled jobs at the heart of the largest urban areas.

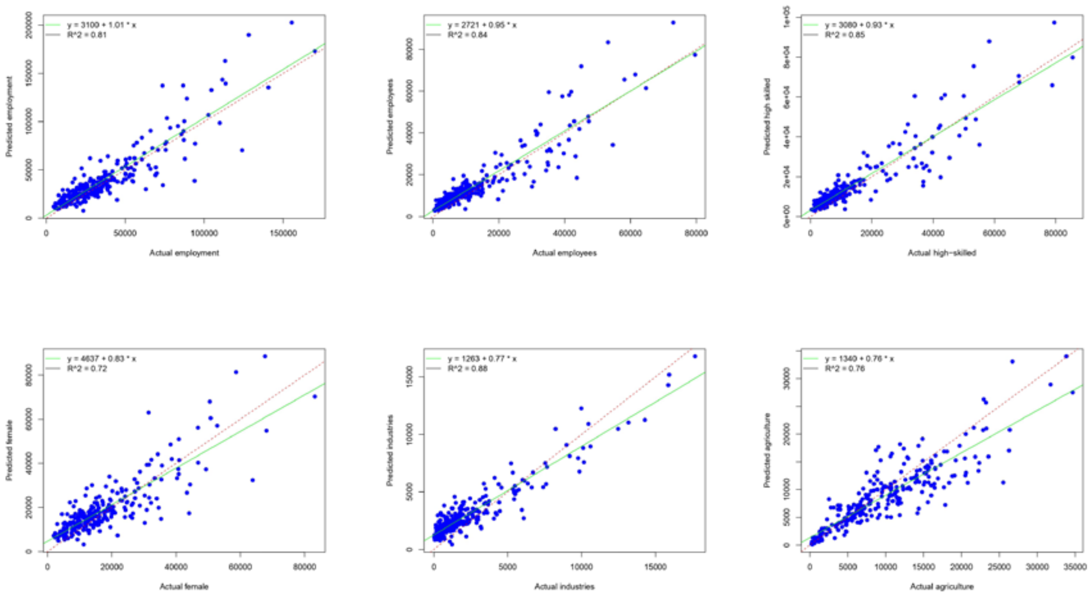
As a final illustration of the downscaling exercise, we present the mapping of manufacturing employment. We have previously shown that Ghana's economy is primarily based on services and agriculture, with industry and manufacturing employment remaining marginal. This is clearly evident in panel (d), where manufacturing employment appears nearly marginal across the country. However, manufacturing is not negligible in certain highly localized urban centers. This indicates that, although manufacturing employment is relatively small overall, there are areas with manufacturing specialization in Ghana.

2.3 Actual vs estimated district statistics

In the absence of finer administrative units in the publicly available census data, it is only possible to compare the actual versus predicted labour market variables at the district level (Fig 4). The predicted variables are presented before the implementation of the district adjustment. The actual district statistics are based on census data, and the predicted values result from the aggregation of individual pixel values for the corresponding district layer.

Despite the large difference in scale between districts and the 0.005-degree pixels, the correlation between actual and predicted values is satisfactory. For the selection of the four categories displayed below, the R^2 is equal to 0.81 for employment, 0.84 for employees, 0.85 for high-skilled employment, and 0.88 for industries. The regression coefficients are close to one: 1.01, 0.84, 0.85, and 0.88, respectively.

▶ **Figure 4: Actual vs estimated statistics - district level - selected categories**



Regarding employment, employees, and high-skilled employment, the dispersion around the 45-degree line (in red) increases with the population of the districts. For low-population districts, the points are close to the 45-degree line and are distributed both below and above it. As the population increases, the data points are distributed further away from the red line. In addition, the model underestimates district-level employment for districts with mid-sized populations and then overestimates employment for districts with high populations.

For industries, the pattern is slightly different. The dispersion is smaller. However, the model overestimates industry employment in districts without industrial specialization, while it underestimates industry employment in districts with industrial specialization. A similar pattern emerges for agriculture employment. As the absolute size of the district increases, the model under-predicts agriculture employment.

Due to space limitations, only six categories are plotted above. For the remaining categories, sectoral employment performs well in general, with R2 values for manufacturing, and services at 0.6, and 0.85, respectively. The main explanation for these good performances at sectoral level is that the auxiliary data are straightforward and directly related to land use, land cover, and vegetation health index for agriculture, while the other sectors are linked to urban-related auxiliary data such as impervious surfaces and transport networks.

Regarding employment status and skills, self-employment and low-skilled employment display smaller R^2 values compared to employees and high-skilled employment (0.65 and 0.70). This is mostly explained by increasing dispersion between actual and predicted values as the size of the district increases.

Lastly, there is a 0.1 percentage point difference between the R^2 values of male and female employment, as well as a 0.2 and 0.3 percentage point difference between youth and prime-age employment, and between old workers and prime-age employment, respectively. Interestingly, these categories are the most sensitive to decisions to participate in the labour market. The factors driving labour participation, discussed in the introduction, are often related to family structure and education, which are more difficult to capture with auxiliary data.

2.4 Importance analysis

Importance analysis offers a method for evaluating the contribution of each feature to the model's accuracy. This helps identify the most influential variables in making predictions and supports feature selection and model interpretation. The metric used is “increase in mean square error”, which assesses the contribution of each auxiliary variable to the accuracy of the model.[13]

Regarding total employment, three types of auxiliary data stand out (see Figure 5). Built-up related variables capture population density and relate to the labour supply aspect of employment – i.e., there are jobs where there are people. Variables such as nighttime lights and slums capture the demand side of the employment relationship and its composition (i.e., the quality of employment). Lastly, road density appears prominently and captures different dimensions central to the employment relationship. Road density reflects population density, as the road network tends to increase with population. It also reflects a sectoral dimension of employment, as there are more roads in urban areas specialized in services, manufacturing, and industry. Road density may also reflect the density of public infrastructure in general, which can serve as a proxy for employment quality (skills and status). Lastly, in economic geography, economic activity benefits from proximity to roads, as this reduces transport costs. This is particularly relevant for sectors subject to agglomeration effects, such as manufacturing.

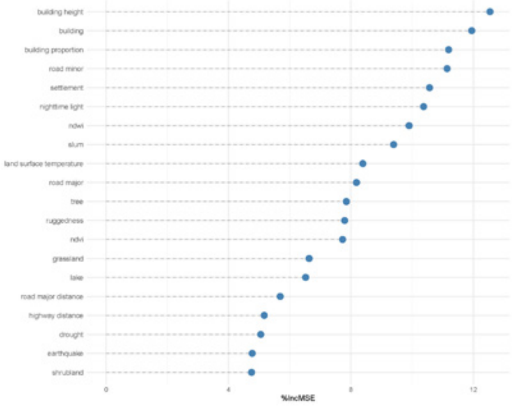
In contrast, employment in agriculture is primarily influenced by auxiliary data reflecting vegetation, such as land-use and land-cover inputs, including shrubland and herbaceous cover. Variables indicating vegetation health, such as the Normalized Difference Water Index (NDWI), also appear central to the model. Climate change variables – particularly droughts – emerge as the third most important factor, as the lack of precipitation directly impacts crop yields. Mining employment is driven by variables capturing the distance to mines and the distance to railways, as rail infrastructure is often constructed to transport mining extractions, along with road infrastructure (Jedwab and Moradi, 2016).

Employment in manufacturing, industry, and services is associated with impervious surfaces, road infrastructure, and nighttime lights. This reflects the concentration of these sectors in urban areas.

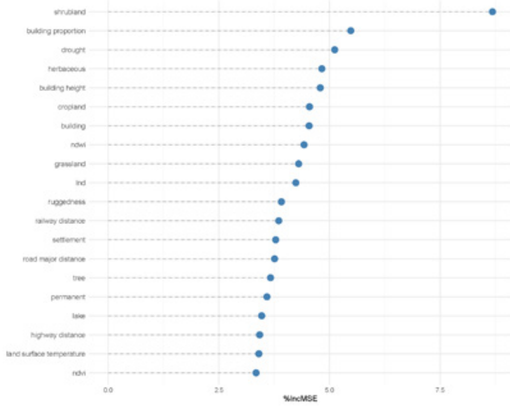
The service sector, in particular, is influenced by the location of public services in densely populated areas. Manufacturing and industry are also shaped by transportation costs and agglomeration economies, which are effectively captured by these different categories of auxiliary data.

13 A second measure, available upon request is increase node purity. It indicates the contribution of each auxiliary variable to the purity of the nodes. The two measures produce similar results.

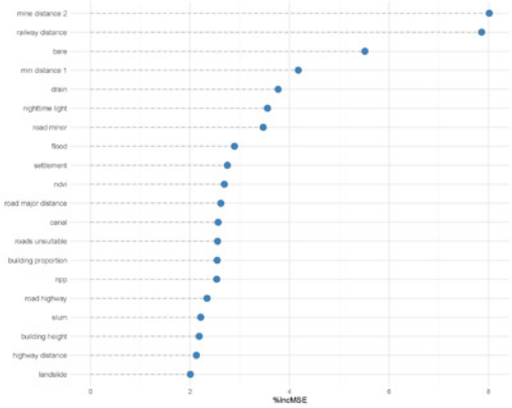
▶ Figure 5: Importance analysis



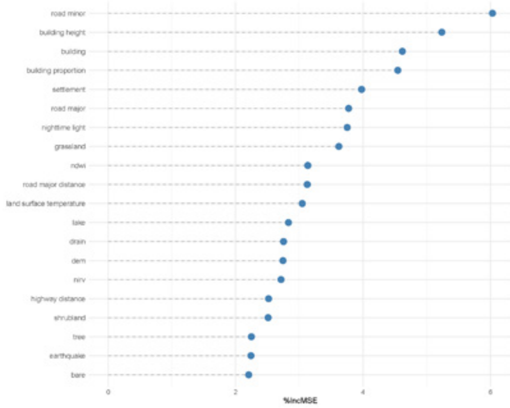
(a) Employment



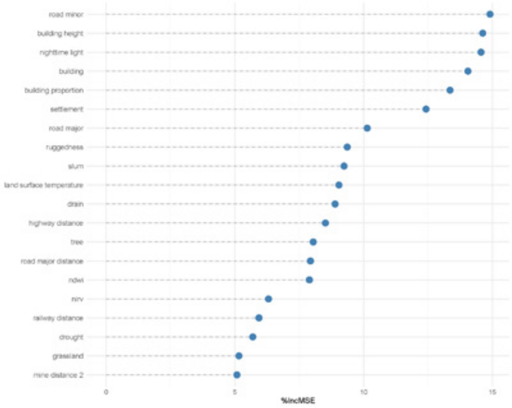
(b) Agriculture



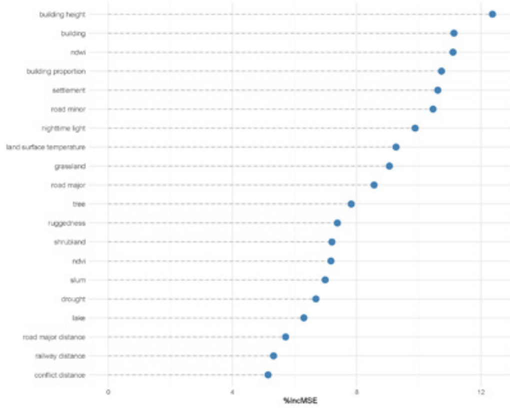
(c) Mining



(d) Manufacturing



(e) High skilled



(f) Low skilled

Regarding the quality of employment, high-skilled employment is driven by factors that reflect its concentration in urban areas – namely, built-up related categories, nighttime lights, and road infrastructure – regardless of the measure considered. In contrast, low-skilled employment is prevalent both in subsistence service sectors within urban areas and in subsistence farming in rural areas. This dual presence explains the mix of built-up related variables and vegetation-related variables. For instance, the Normalized Difference Water Index (NDWI) ranks as the third most important factor in terms of mean squared residuals. Land surface temperature and grassland also appear among the top ten variables in terms of importance.

2.5 Coarser resolution and quality of the model

The pixel-level projection of the labour market variables has a resolution of 0.005 degrees, approximately 0.5 km by 0.5 km near the equator. This is a granular resolution that may be considered overly fine, given that local labour markets are often best captured at coarser resolutions, such as 1 km by 1 km or 5 km by 5 km. For this reason, the model is estimated at a coarser resolution of 0.05 degrees (approximately 5 km by 5 km). The quality of the estimation remains high. However, both the mean squared residual (MSR) and the percentage of variance explained deteriorate. On average, the MSR increases by 13 percentage points across all employment sub-categories, with this increase primarily driven by sectoral categories such as manufacturing, industry, and services. The decline in the percentage of variance explained is more limited, averaging around 1 percentage point. One possible explanation is that the auxiliary data used for disaggregation lose explanatory power when averaged at a coarser resolution. For example, road density measured at 0.5 km by 0.5 km resolution is more informative than when aggregated to 5 km by 5 km. Due to space limitation the MSR and % of variance for the different employment sub-categories are displayed in the appendix Table 4.

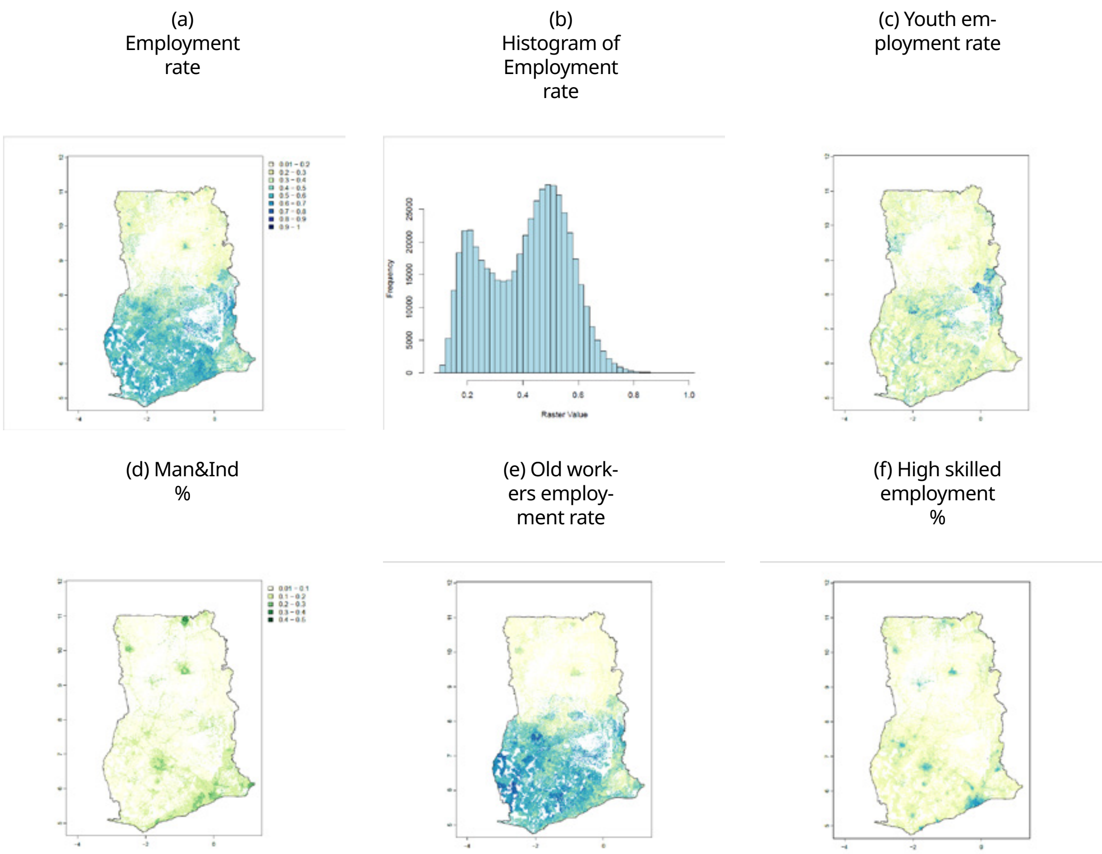
3 Applications

3.1 Employment rates, sectoral percentages

The downscaling exercise undertaken in this work concerns labour market categories in levels. However, in order to better describe employment at the local level, measuring employment rates is important as it gives a sense of employment in proportion to the working-age population (defined as 15+). To build these categories, this work also disaggregates population categories at the pixel level.[14] These categories are population total, working age population total, youth, prime age, old, male and female. The quality of the estimation for the population categories are as good as the quality of the estimation for the employment categories see Table 5 in the appendix. The heterogeneity of employment rates increases with the granularity of the data. Employment rate is 48% countrywide. The standard deviation is 11.2 percentage points at the district level with a minimum of 16.5% and a maximum of 64.7%. While relatively low at the aggregate level, some districts in Ghana display employment rates that are as high as in high-income countries. At the pixel level, the standard deviation increases to 15 percentage points, and employment rates fluctuate between 0.1 and 0.98 (Figure 6 Panel A).

[14] These categories are population total, working age population total, youth, prime age, old, male and female. The quality of the estimation for the population categories are as good as the quality of the estimation for the employment categories. For each categories we list the MSR and R2 in brackets: pop (0.16; 0.92), wap (0.15; 0.93), wapy (0.15; 0.92), wappa (0.14; 0.94), wapold (0.14; 0.92), wapw (0.14; 0.93), wapm (0.13; 0.94). In addition, the percentage of pixels with aberrant values, i.e., rates greater than 1, is not di_erent from zero for all categories except for female employment rate (4%), which may reect either an overestimation of female employment or an underestimation of female working-age population.

▶ Figure 6: Spatial distribution of (selected) employment rates and percentages



Regarding the geographic distribution of employment rates, while employment in levels reflects the prevalence of urban areas, employment rates describe a North-South divide. Employment rates are high in Southern Ghana, whether in rural or urban areas. In the Northern part of the country, employment rates are much lower except in the main 4 urban centers mentioned previously. Employment rates are particularly low in the North-East (Figure 6 Panel A). At pixel level, employment rates follow a binomial distribution centered around 0.2 and 0.49 (Panel B).

Across age categories, the distribution of employment rates displays different patterns for young workers and older workers (Panels C and D). Youth employment rates are low in urban areas, reflecting the lack of employment opportunities for young workers in urban areas. However, youth employment rates exhibit high values in the central area of Ghana, and in particular in the Central-Eastern area. High employment rates for youth are also located in very granular areas in the West. The employment rates of older workers are lower in the southern urban centers compared to the northern urban centers. In terms of geographic pattern, older workers have high employment rates in the Western part of Ghana; the higher the rates, the closer to the border with Ivory Coast.

The last two illustrations map high-skilled employment and the sum of manufacturing and industrial employment as percentages of total employment (Panels E and F). Regarding high-skilled employment shares, while high-skilled employment is concentrated solely in urban areas in absolute terms, the share of high-skilled workers appears to be relatively significant in secondary cities, along the coast, and between urban centers. Similarly, manufacturing and industrial employment is less concentrated in a handful of urban centers when measured as a percentage of total employment rather than in absolute values. In urban areas, manufacturing and industrial employment can account for a substantial percentage of total employment. While on aggregate the share remains small, at local level high percentages signal the existence of specialization centers.

3.2 Labour market at city level

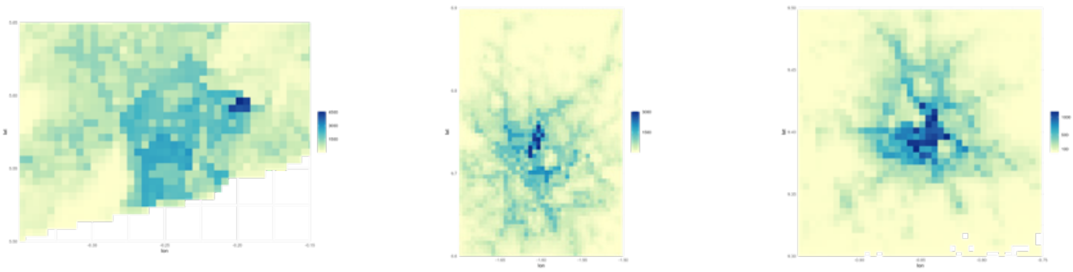
Gridded employment data enables us to produce a detailed panorama of local labour markets for any village, town, or city. In this section, we provide an illustration at the agglomeration level. Figure 7 zooms in on the cities of Accra, Kumasi, and Tamale. The model distinguishes between city centers, which concentrate employment, and city peripheries, which contain fewer jobs. Accra exhibits a high density of employment, along with a heterogeneous distribution within the agglomeration. In Kumasi and Tamale, the model identifies a core area where employment is concentrated, with jobs in the peripheries clustering around road infrastructures.

▶ **Figure 7: Employment distribution for Accra, Kumasi and Tamale**

(a) Accra – 0.005°

(b) Kumasi – 0.005°

(c) Tamale – 0.005°



In order to measure the distribution of employment between urban and rural areas, we overlay our gridded labour market statistics with the agglomeration layer produced by Africapolis.[15] Africapolis defines urban agglomerations as a continuum of built-up areas with less than 200 meters between buildings and constructions and a total population of more than 10,000 inhabitants. Their definition is based on a spatial approach, in contrast to existing definitions of urban areas that rely on administrative criteria. Using the agglomeration layer from Africapolis, we identify 129 agglomerations with populations exceeding 10,000 inhabitants. For these cities, we aggregate the gridded population data to generate labour market statistics. Focusing on the five largest agglomerations — Accra, Kumasi, Tamale, Takoradi, and Koforidua —we find that they account for 32% of the total population but 38% of total employment, 53% of high-skilled jobs, 50% of high-skilled employees, 42% of manufacturing employment, 53% of service employment, and 41% of the unemployed. This shows a bias towards urban areas that concentrates most of the employment categories. Taken as a whole, the 129 cities with a population greater than 10 000 habitants concentrates 49% of the population but 68% of high skilled employment and 58% of manufacturing employment.

[15] https://africapolis.org

4 Discussion

While the downscaling of demographic variables or GDP is typically restricted to a handful of variables—such as total population, age subcategories, or sectoral GDP—one contribution of this work is to perform downscaling across 17 employment categories. The main motivation is that employment is a complex phenomenon, combining aspects of quantity, under-utilization (unemployment or NEET), quality (skills and status), and sectoral distribution. This complexity is especially evident in a middle-income country, where employment challenges go well beyond the overall employment level. Variables such as population (built-up areas), aggregate demand (nighttime lights), and access to infrastructure emerge as important determinants of the various employment categories.

We have observed that, beyond the good performance of the model and the accuracy of its projections, categories sensitive to labour market participation decisions could be further improved. In the literature, family structure, education level, and the decision to work or study are important factors influencing the employment of youth, older workers, and women. Our choice of auxiliary data could be further refined to better capture these dimensions. In this respect, points of interest (POIs) do not appear highly ranked in terms of variable importance. This might be due to the heterogeneous coverage of POIs in Ghana between major cities and rural areas.

Another point of discussion is the difference in scale between the existing employment data at the district level and the granularity of the projection at a resolution of 0.005. Section 3.3 discussed the strong performance of the projection model in predicting district-level statistics. However, starting from a more granular administrative level, such as councils, could improve the model's accuracy.

Model improvement could also rely on alternative estimation methods. In this paper, the method relies on random forests. Alternative estimation techniques—such as XGBoost, multiple linear regression, support vector regression, or Res-FuseNet—could help improve the performance of the pixel-level projections. One source of model improvement lies in its ability to better predict high-population areas. This paper has discussed the increasing dispersion between actual and estimated values as the size of the area increases. This is a common result in high-spatial-resolution population and GDP data. Jin et al. (2023) show that area-to-area kriging models tend to better estimate large population units. Similar challenges have been noted in other high-resolution downscaling studies. For example, U.S. state-level occupational employment data have been downscaled to Census tracts, producing tract level estimates for over 800 occupations (Wang et al., 2024). Deep-learning methods using satellite imagery have also been applied to map fine-scale socioeconomic variables, including population and employment size, in developmentally uneven areas (Ahn et al., 2023). These examples demonstrate both the possibilities and limitations of fine-grained spatial disaggregation, reinforcing the potential for methodological improvements in pixel-level projections.

Our analysis focused on residential areas when selecting built-up areas, leaving questions of commuting and job location for future research. Overall, these results show that high-resolution employment data can provide valuable information for local labour policy and planning. Despite some limitations in fine-scale predictions, the model delivers reliable estimates across most employment categories and spatial units

5 Conclusion

This study introduces a pioneering approach to generating high-resolution (0.005) gridded labor market data for Ghana, addressing a critical gap in spatially disaggregated socio-economic datasets. By leveraging random forest algorithms, remote sensing, and 64 auxiliary variables—including land cover, nighttime lights, infrastructure, and points of interest—we successfully downscaled district-level census data into 17 employment categories, capturing age, gender, skills, status, sectors, unemployment, and NEET.

The model achieves high accuracy ($R^2 > 90\%$ for most categories) and uncovers substantial spatial heterogeneity in Ghana's labor market. Pixel-level disaggregation provides a novel, granular perspective, revealing distinct patterns in employment distribution. Employment is highly concentrated in the main urban agglomerations, particularly in the southern half of Ghana, between Accra and Kumasi, as well as along the coast and major transport infrastructures. In contrast, the northern half of the country exhibits sparser employment, focused around four key agglomerations. Salaried employment is predominantly linked to urban areas, where either public administration or manufacturing and industrial activities are located. Similarly, low-skilled employment is less prevalent in city centers compared to total employment, reflecting the concentration of high-skilled jobs in the core of major urban areas. While manufacturing employment remains marginal at the national level, it is not negligible in certain localized urban centers, indicating the presence of specialized manufacturing hubs in Ghana. An analysis of employment rates reveals wide variation, ranging from 10% to 98% across pixels, and highlights a pronounced North-South divide. The gridded data also underscore the prominence of urban employment.

Variable importance analysis underscores the role of built-up areas, nighttime lights, road density, and vegetation health in predicting employment patterns, with sector-specific nuances—such as the influence of vegetation indices on agricultural employment and infrastructure on manufacturing. While the model performs robustly, challenges remain in capturing labor market participation decisions, particularly for youth, women, and older workers, where family structure and education play pivotal roles but are less reflected in geospatial data.

The methodology advances beyond traditional GDP or population gridding by incorporating the complexity of labor markets, offering a scalable framework for data-scarce contexts. Future work could refine auxiliary data selection, explore alternative machine learning techniques (e.g., stepwise random forest, random forest area-to-area regression kriging, XGBoost or Res-FuseNet), and integrate finer administrative units or additional auxiliary data to further improve precision.

While censuses include an employment module, labor markets are more effectively captured through specialized surveys, such as labor force surveys or household income and expenditure surveys. These surveys offer detailed labor market indicators—such as informal employment or wage/income categories—and their higher frequency enables the potential generation of gridded labor market data on an annual or bi-annual basis. However, their smaller sample sizes often limit representativeness at the district level. One solution is to leverage the geographic coordinates (latitude and longitude) of surveyed households, where publicly available, to produce high-resolution gridded labor market data.

The gridded labor market maps offer policymakers a valuable resource to design targeted interventions that address local employment challenges and promote more inclusive economic development in Ghana and similar contexts

References

Aaronson, D., R. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, and K. Schulze (2020, 08). The effect of fertility on mothers' labor supply over the last two centuries. *The Economic Journal 131* (633), 1–32.

Ahn, D., M. Song, S. Lee, Y. Choi, J. Kim, S. Park, H. Yang, and M. Cha (2023). Fine-grained socio-economic prediction from satellite images with distributional adjustment. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, New York, NY, USA, pp. 3717–3721. Association for Computing Machinery.

Amankwah, A., P. Castaing, N. Owoo, and A. Palacios-Lopez (2022). Labor market participation and employment choice in ghana : Do individual personality traits and gender role attitudes matter ? *World Bank Policy Research working paper 10664.*

Azar, D., J. Graesser, R. Engstrom, J. Comenetz, R. M. L. Jr., N. G. Schechtman, and T. Andrews (2010). Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in haiti. *International Journal of Remote Sensing 31* (21), 5635–5655.

Baah-Boateng, W. (2013). Determinants of unemployment in ghana. *African Development Review 25* (4), 385–399.

Bhaduri, B., E. Bright, P. Coleman, and M. L. Urban (2007). Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal 69* (1/2), 103–117.

Bosco, C., V. Alegana, T. Bird, C. Pezzulo, L. Bengtsson, A. Sorichetta, J. Steele, G. Hornby, C. Ruktanonchai, N. Ruktanonchai, E. Wetter, and A. J. Tatem (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface 14* (129), 20160825.

Briggs, D. J., J. Gulliver, D. Fecht, and D. M. Vienneau (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment* 108 (4), 451–466.

Chen, Y., G. Wu, Y. Ge, and Z. Xu (2022). Mapping gridded gross domestic product distribution of china using deep learning with multiple geospatial big data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15, 1791–1802.

Chen, Y., R. Zhang, Y. Jin, and Z. Xia (2019, 10). Downscaling census data for gridded population mapping with geographically weighted area-to-point regression kriging. *IEEE Access* 7, 1–1.

Dobson, J. E., E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley (2000). Landscan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing 66* (7), 849–857.

Huber, J. D. and L. Mayoral (2024, Mar). Economic development in pixels: The limitations of night-lights and new spatially disaggregated measures of consumption and poverty. Working Papers 1433, Barcelona School of Economics.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353 (6301), 790–794.

Jedwab, R. and A. Moradi (2016). The permanent effects of transportation revolutions in poor countries: Evidence from africa. *The Review of Economics and Statistics 98* (2), 268–284.

Jin, Y., Y. Ge, H. Fan, Z. Li, Y. Liu, and Y. Jia (2023). Mapping gross domestic product distribution at 1 km resolution across thailand using the random forest area-to-area regression kriging model. *ISPRS International Journal of Geo-Information 12 (12)*.

Klasen, S., L. Tu-Thi-Ngoc, J. Pieters, and M. Santos-Silva (2021). What drives female labour force participation? comparable micro-level evidence from eight developing and emerging economies. *The Journal of Development Studies 57* (3), 417–442.

Kummu, M., M. Taka, and J. H. A. Guillaume (2018). Gridded global datasets for gross domestic product and human development index over 1990–2015. *Scientific Data 5* (1).

Lebakula, V., K. Sims, A. Reith, A. Rose, J. McKee, P. Coleman, J. C. Kaufman, M. Urban, C. Jochem, C. Whitlock, M. Ogden, J. Pyle, D. Roddy, J. Epting, and E. A. Bright (2025). Landscan global 30 arc-second annual global gridded population datasets from 2000 to 2022. *Scientific Data 12*, 495.

Leyk, S., A. E. Gaughan, S. B. Adamo, A. de Sherbinin, D. Balk, S. Freire, A. Rose, F. R. Stevens, B. Blankespoor, C. Frye, J. Comenetz, A. Sorichetta, K. MacManus, L. Pistolesi, M. Levy, A. J. Tatem, and M. Pesaresi (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data 11* (3), 1385–1409.

Li, C., L. Yu, F. Oloo, E. G. Chimimba, O. Kambombe, M. Asamoah, P. D. Opoku, V. W. Ogweno, D. Fawcett, J. Hong, X. Deng, P. Gong, and J. Wright (2023). Slum and urban deprivation in compacted and peri-urban neighborhoods in sub-saharan africa. *Sustainable Cities and Society 99*, 104863.

Murakami, D. and Y. Yamagata (2019). Estimation of gridded population and gdp scenarios with spatially explicit statistical downscaling. *Sustainability 11* (7).

Murakami, D., T. Yoshida, and Y. Yamagata (2021). Gridded gdp projections compatible with the five ssps (shared socioeconomic pathways). *Frontiers in Built Environment 7.*

Padilla, A., D. Otarod, S. Deloach-Overton, R. Kemna, P. Freeman, E. Wolfe, L. Bird, A. Gulley M. Trippi, C. Dicken, J. Hammarstrom, and A. Brioche (2021). Compilation of geospatial data (gis) for the mineral industries and related infrastructure of africa. U.S. Geological Survey data release.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015, 02). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE* 10 (2), 1–22.

Tang, L. and T. T. Werner (2023). Global mining footprint mapped from high-resolution satellite imagery. *Communications Earth & Environment 4*, 134.

Tatem, A. J. (2017). Worldpop, open data for spatial demography. *Scientific Data 4*, 170004.

Wang, S., S. Agrawal, E. A. Mack, N. Kalani, S. R. Cotten, C.-H. Chang, and P. T. Savolainen (2024). Downscaling occupational employment data from the state to the census tract level. *Applied Geography 170*, 103349.

Wang, T. and F. Sun (2022). Global gridded gdp data set consistent with the shared socioeconomic pathways. *Scientific Data 221*, 2052–4463.

Wang, Y., C. Huang, M. Zhao, J. Hou, Y. Zhang, and J. Gu (2020). Mapping the population density in mainland china using npp/viirs and points-of-interest data based on a random forests model. *Remote Sensing 12* (21).

Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. J. Tatem (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences 115* (14), 3529–3537.

Wu, N., J. Yan, D. Liang, Z. Sun, R. Ranjan, and J. Li (2024). High-resolution mapping of gdp using multi-scale feature fusion by integrating remote sensing and poi data. *International Journal of Applied Earth Observation and Geoinformation 129*, 103812.

A Appendix

Table 4: Model evaluation coarser resolution - 0.05°

	(1)	(1)
names	MSR	% of variance
Et	0.21	0.92
Ety	0.23	0.89
Etpa	0.21	0.92
Etold	0.24	0.91
Etw	0.21	0.92
Etm	0.20	0.92
Ut	0.29	0.91
neet	0.21	0.87
Eths	0.27	0.93
Etls	0.20	0.91
Etse	0.22	0.90
Etees	0.29	0.92
Etag	0.19	0.71
Etmi	0.76	0.75
Etma	0.36	0.89
Etind	0.29	0.92
Etserv	0.28	0.93

This table presents the mean square residual and the percentage of variance explained for each employment categories for a resolution of 0.05.

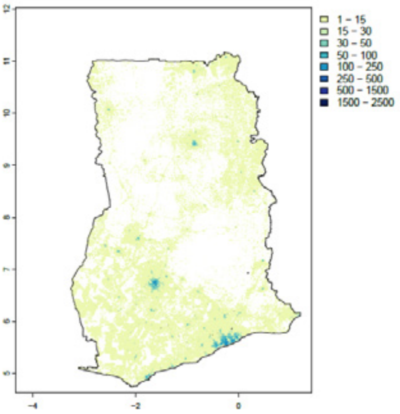
Table 5: Model evaluation population categories

	(1)	(1)
names	MSR	% of variance
pop	0.16	0.92
wap	0.15	0.93
wap y	0.15	0.92
wap pa	0.14	0.94
wap old	0.14	0.92
wap w	0.14	0.93
wap m	0.13	0.94

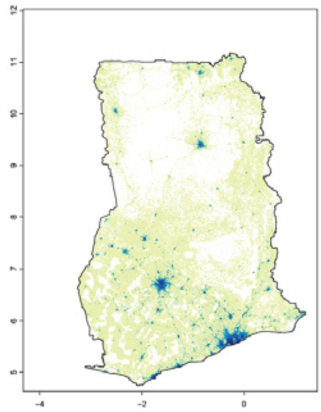
This table presents the mean square residual and the percentage of variance explained for each population related categories.

▶ Figure 8: Spatial distribution of (selected) employment categories

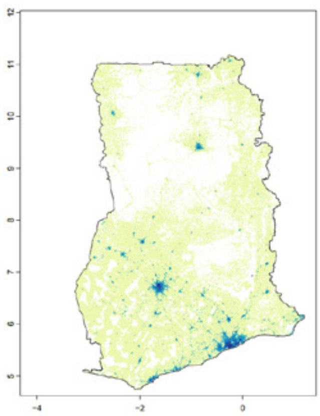
(a) Youth – 0.005°



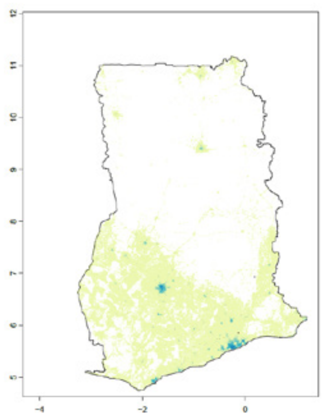
(b) Prime age – 0.005°



(c) Old-workers – 0.005°

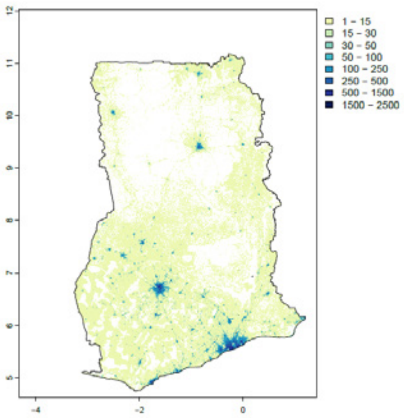


(d) Women – 0.005°

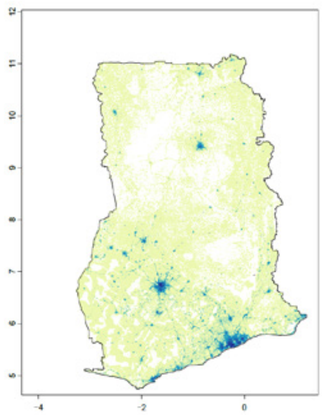


▶ **Figure 9: Spatial distribution of (selected) employment categories**

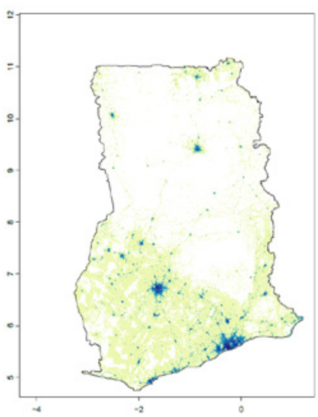
(a) Men – 0.005°



(b) Self-employment – 0.005°



(c) High-skilled – 0.005°



(d) Agriculture – 0.005°

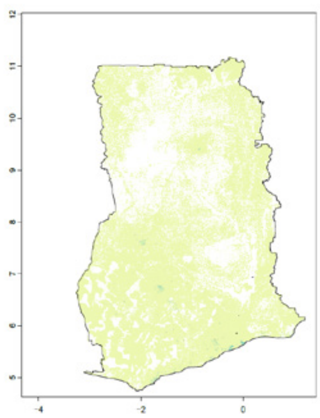
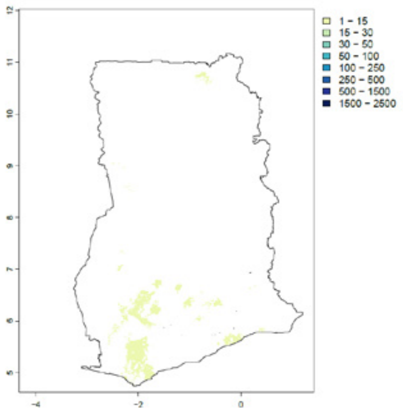
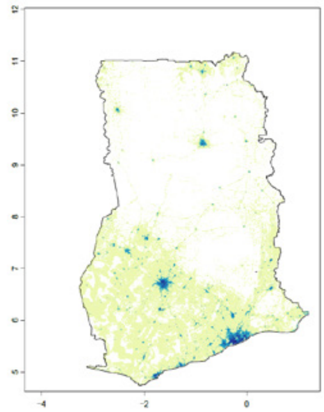


Figure 10: Spatial distribution of (selected) employment categories

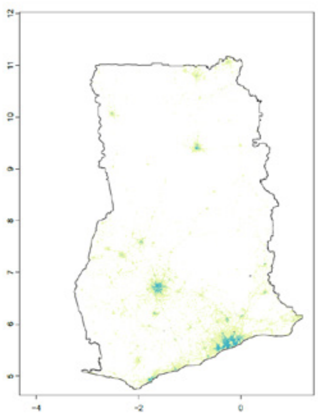
(a) Mining – 0.005°



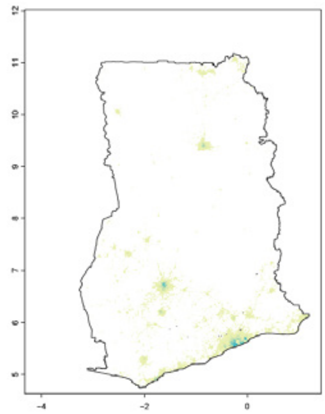
(b) Services – 0.005°



(c) Industry – 0.005°

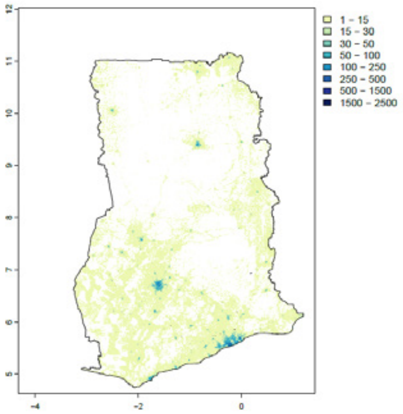


(d) Unemployment – 0.005°



▶ **Figure 11: Spatial distribution of additional employment categories**

(a)NEET – 0.005°



Acknowledgements

We would like to thank Sara Elder, Lorenzo Guarcello, Luca Fedi, Marie Urban, Carter Christopher, Viswadeep Lebakula, Andrew Tatem, Chris Nnanatu and Yacouba Diallo for useful comments and discussions. We would like to thank Yves Perardel and Vipasana Karkee for the exchange regarding the labour market module of the housing and population census. We are grateful to Sandra Ofosuapea from the Ministry of Labour of Ghana and to Johnson Kagya and Anthony Oduro-Denkyirah from Ghana Statistical Services. We would like to thank Mathieu Chenut for editing support.

Advancing social justice, promoting decent work

The International Labour Organization is the United Nations agency for the world of work. We bring together governments, employers and workers to improve the working lives of all people, driving a human-centred approach to the future of work through employment creation, rights at work, social protection and social dialogue.

Contact details

Employment Policy Department (EMPLOYMENT)

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland
T +41 22 799 7861
employment@ilo.org
www.ilo.org/employment

ISBN 9789220432983