



Mapping fine-scale socioeconomic inequality using machine learning and remotely sensed data

Nabin Pradhan ^a and Arun Agrawal ^{b,*}

^aSchool for Environment and Sustainability, University of Michigan, 440 Church Street, Ann Arbor, MI 48109, USA

^bKeough School of Global Affairs, O 308 Hesburgh Center, University of Notre Dame, Notre Dame, IN 46556, USA

*To whom correspondence should be addressed: Email: arun.agrawal@nd.edu

Edited By Erik Kimbrough

Abstract

Limited and missing socioeconomic data have made it nearly impossible to measure or estimate inequality consistently at fine spatiotemporal and jurisdictional scales, especially for lower- and middle-income countries. We deploy a novel data harmonization method that combines existing household survey data with freely available remotely sensed data and machine learning techniques to generate fine-scale socioeconomic inequality estimates across spatial and temporal scales for India. Our manuscript makes three important contributions. First, it identifies key remote sensing datasets that, in combination with nighttime luminosity, improve its predictive power to estimate measures of socioeconomic inequality. Second, it offers an analytical approach that reliably estimates the uneven distribution of socioeconomic conditions by harmonizing household assets and sociodemographic information that remotely sensed data at the village or similar geographic levels represent—the results achieve >84% prediction accuracy. Finally, it leverages a spatially cross-validated machine learning model with training and test datasets from two successive Demographic and Health Surveys to demonstrate how data gaps in socioeconomic inequality at subnational levels can be addressed. Our replicable approach has the potential to improve global inequality data, thereby supporting research and applications aiming to reduce socioeconomic inequality in the context of the Sustainable Development Goals.

Keywords: inequality, remote sensing, machine learning

Significance Statement

The unavailability of socioeconomic data at the subnational level has made it challenging to measure or estimate inequality consistently. Our study demonstrates a novel data harmonization method that combines existing household survey data with freely available remotely sensed data using machine learning to generate fine-scale inequality measures over time and space. Our replicable and generalizable analysis of socioeconomic inequality has the potential to address major gaps in the study of inequality.

Main

Nationally representative household surveys provide the most common foundation for measuring inequality, but they are expensive and time-consuming if used conducted regularly, frequently, and for representativeness at fine spatial or jurisdictional levels. Missing and patchy data at finer scales hamper analyses of inequality, patterns of change in it over time, and its relationship to socioenvironmental drivers and outcomes (1). The World Bank's World Development Indicators (WDI) database shows that over 65% of countries have Gini coefficient measures, the most commonly used indicator of income or consumption inequality, available fewer than six times between 2000 and 2022 (2). Specifically, the countries with unavailable inequality data predominantly belong to the lower- and middle-income countries (L&MICs)

(Fig. 1). Besides, inequality estimates available at the national level do not have a regular temporal frequency (3). The unavailability of inequality data at the lower jurisdiction level is more pervasive and affects redistributive or other policy decisions that affect well-being and inequality. Social assistance or safety net programs are an example (4). While economic and environmental indicators are relatively easy to identify based on existing data such as national censuses and remotely sensed data, the same is not true for inequality indicators because of the unavailability of spatially linked socioeconomic information at a fine scale (5, 6). Collecting such information through traditional household surveys in national censuses makes filling inequality data gaps expensive and time-consuming (7). Our study demonstrates the harmonization of survey data with remotely sensed data and the application of

Competing Interest: The authors declare no competing interests.

Received: July 26, 2024. **Accepted:** January 22, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

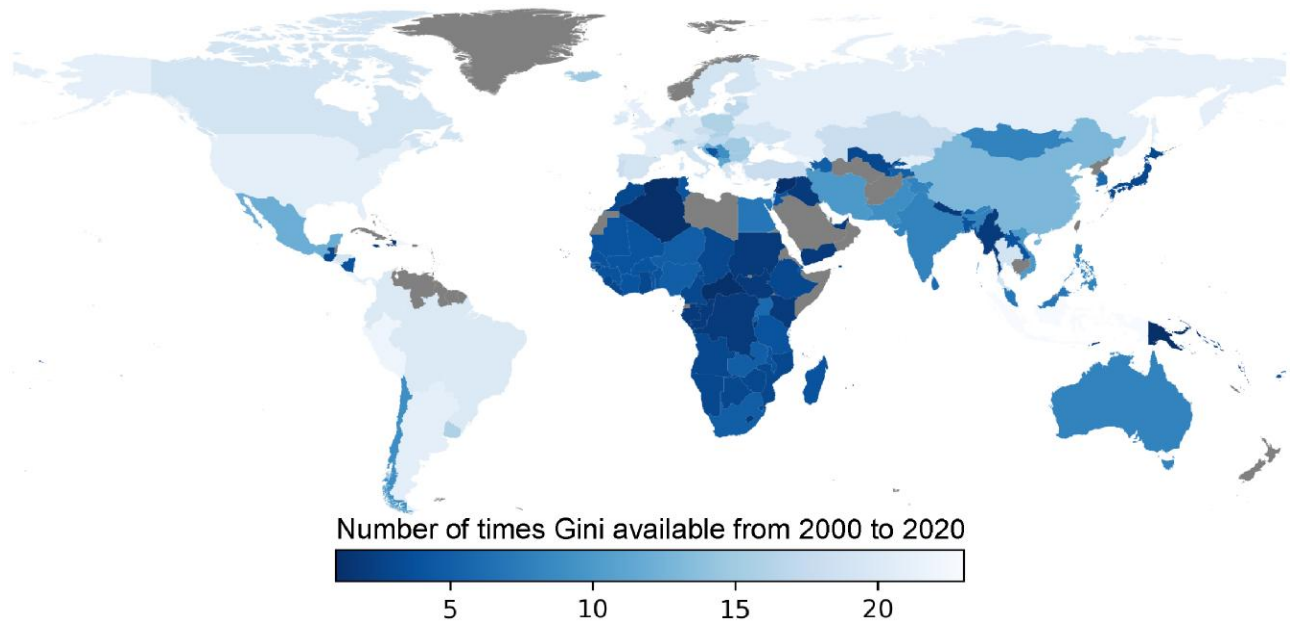


Fig. 1. Number of years country-level Gini coefficient available from 2000 to 2022. Lower frequency indicated by heavier shading intensity, and unavailability data indicated by neutral shade.

machine learning methods to construct measures of socioeconomic inequality and identify patterns in satellite data that best correlate with survey data. The proposed analytical method can be used to construct diverse inequality measures by combining freely available remote sensing data in conjunction with machine learning techniques.

Studies that try to address economic well-being data gaps find that its measures are strongly associated with nighttime luminosity (8–12). Furthermore, recent studies show that nighttime lights are a reliable predictor of poverty and can distinguish economic activity at fine spatial scales (4, 7). These studies reveal the potential of nighttime lights and machine learning approaches to estimate the level of economic growth, poverty, and other human development indices (13, 14). However, there is limited evidence at best to clarify the relationship between the uneven distribution of socioeconomic well-being and nighttime lights to match local-, subnational-, and national-level estimates of inequality based on data from social surveys (15). To address these knowledge and data gaps, we focus on assessing the association between nighttime lights and the uneven distribution of socioeconomic well-being. We use household assets, including durable goods, nondurable goods, and socioeconomic conditions, as proxies for socioeconomic well-being and use the Gini coefficient to estimate the uneven distribution of socioeconomic inequality (16). Additionally, we identify other remotely sensed datasets for sociodemographic characteristics, land cover, and remoteness to better predict socioeconomic inequality.

Our approach is based on a novel data harmonization framework that integrates household surveys and satellite using machine learning approach to estimate a reliable proxy for inequalities in socioeconomic well-being at multiple spatial and temporal scales. Understanding the uneven distribution of socioeconomic condition within a community is crucial to enable decision-makers to design targeted policy interventions that reduce inequality. Without attention to how different policy measures and socioeconomic forces that affect well-being also simultaneously affect inequality, changes in levels of equality have little likelihood of being influenced purposively (17, 18).

Nighttime lights as a proxy for estimating inequality

Using data from 84 countries, we found that nighttime luminosity is an important predictor of income or consumption inequality. However, the analysis shows large differences between nighttime lights-based and survey-based Gini estimates of consumption, income, and assets at the national scale (Figs. S1 and S11). For this analysis, we obtained national-level Gini index for 2015 from the WDI database (Table S1), selecting countries with available data for the study year. To calculate the nighttime lights-based Gini, we used the lowest level administrative boundaries available in the geoBoundaries database (19). We calculated the average nighttime luminosity using the annual Visible and Infrared Imaging Suite (VIIRS) dataset (20). Finally, we computed nighttime lights-based Gini coefficient for each country, using the lowest administrative level as the unit of analysis and compared with WDI's Gini index. Similarly, we conducted state-level analysis in India to compare between survey-based Gini using household asset holdings and nighttime lights-based Gini. The analysis also shows sizable differences between the two indices (Fig. S10A).

The analysis at various spatial scales suggests that nighttime lights can serve as a potential predictor of various inequality indicators (21). However, generating a reliable proxy for survey-based socioeconomic inequality requires consideration of additional covariates. These covariates could be demographic, socioeconomic, land use related, remoteness of locations, among other factors (22). It also requires addressing other possible reasons for the mismatch between nighttime lights-based Gini and survey-based Gini: differences in sample sizes and measurement strategies researchers have employed to assess the uneven distribution of income and consumption. For example, the WDI's Gini index is computed using household income or consumption data, while others have used wealth inequality using assets and net assets (23). Our analysis uses household assets as a proxy for socioeconomic inequality. Specifically, our analytical approach utilizes household-level socioeconomic data derived from nationally representative Demographic and Health Surveys (DHS) conducted in

India, integrated with multiple layers of remotely sensed data within a machine learning framework (Fig. 2).

Given the unavailability of comprehensive current socioeconomic inequality data in India, this analysis contributes to understanding of the distribution of socioeconomic well-being at the subnational level. The novel inequality database generated from this analysis can support efforts to prioritize areas and groups for actions that help reduce inequality. Existing studies on inequality in India have contributed to understanding inequality based on incomes and consumption (24–27). However, these studies are not comprehensive in terms of spatial and temporal coverage. We addressed these gaps related to socioeconomic inequality by harmonizing social survey and remote sensing data that approximates a village in rural areas and a ward in urban areas (7).

Estimating socioeconomic inequality

Two successive nationally representative DHS surveys in 2015–2016 and 2019–2021 provided an opportunity to test and validate estimated socioeconomic inequality at fine spatial and temporal resolutions. These datasets include a distribution of ~22% of the poorest households, 21% poorer, 20% middle-income, 19% richer, and 18% of the wealthiest households, totaling over 600,000 households across more than 28,000 geographic clusters in both survey rounds. Each geographic cluster corresponds to a small village or community, comprising ~23 (Fig. S2) households (7). We leveraged the spatial location (Fig. 2A) of each cluster to estimate socioeconomic well-being conditions (28). We employed the principal component analysis method to estimate the socioeconomic well-being index for each household (29, 30). Table S2 provides a comprehensive list of variables utilized in the computation of the socioeconomic well-being index. To account for intergroup variation in household-level assets, we standardize the household socioeconomic well-being index on a 0–1 scale instead of employing a quantile approach. The choice to normalize the socioeconomic index to a 0–1 scale aligns with how asset-based indices like the International Wealth Index assess material well-being. Normalizing to a 0–1 range enables a clear, comparable framework across households by setting bounds where a score

of 1 represents maximum material well-being and a score of 0 represents the absence of these assets and services (16). This approach allows consistent comparisons across households with varying asset profiles and simplifies the interpretation of socioeconomic inequality. We used Lorenz curves (Fig. S3) and the Gini coefficient to measure and estimate the uneven distribution of socioeconomic well-being at the cluster level (31). The Gini coefficient is calculated as the area between the Lorenz curve and the line of perfect equality divided by the total area below the line of perfect equality (32). To benchmark against the national-level Gini coefficient, we computed country-level socioeconomic inequality and compared them with WDI's income or consumption Gini coefficients in 2015 and 2019. Our analysis shows that the Gini coefficients from the survey data and the WDI for income or consumption are closely aligned at the country level (Fig. S4).

Harmonizing remotely sensed covariates

We used the Gini coefficient in each DHS cluster as the reference data or outcome variable and harmonized remotely sensed data product as covariates or features for our machine learning model. The DHS provides spatial information of cluster locations with a displacement error of 2 km for urban clusters and 5 km for rural clusters. To account for these displacement errors, we applied buffers to approximate the location of each household within each geographic cluster (7). We then extracted pixel values for the identified remote sensing covariates within these buffers. The remote sensing covariates were selected based on their hypothesized importance for estimating socioeconomic conditions and their availability at 500-m spatial resolutions from 2015 onward: the included variables represent socioeconomic, demographic, land use, and remoteness (Table S3). We rely on existing literature to select seven layers of covariates for our analysis. These include nighttime lights (11, 21) for economic activities, crop area (33–35) as land-use indicator, and normalized difference vegetation index (NDVI) (36–38) as proxy agricultural yield. We have used population (39, 40), urban-built, and human footprint (41–43) as representing sociodemographic factors and the nearness to the national highway and administrative center (44, 45) as indicators of remoteness. All satellite imagery used in

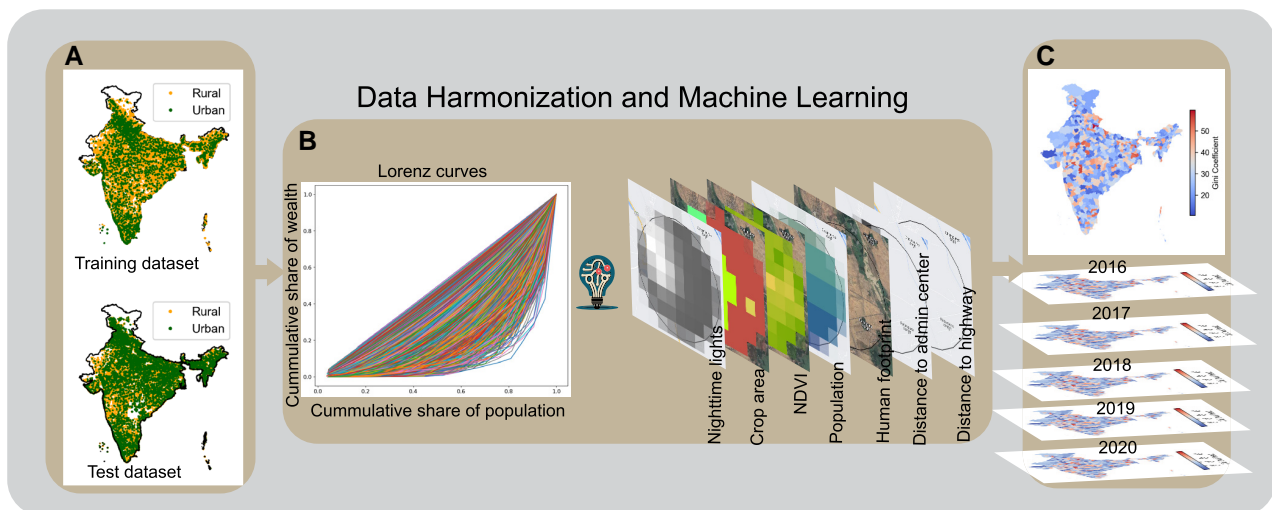


Fig. 2. Analytical framework for data harmonization and machine learning approach to estimate remote sensing-based socioeconomic inequality. A) DHS 2015–2016 (number of clusters = 28,000) is the training dataset and DHS 2019–2021 (number of clusters = 30,000) is the test dataset. B) Lorenz curves to estimate the uneven distribution of household socioeconomic well-being at the cluster level. Remotely sensed covariates harmonized at 2- and 5-km buffers around urban and rural clusters. C) Predicted Gini coefficient using machine learning model from 2016 to 2020.

this study is reprojected and resampled at 500-m spatial resolution to make the spatial scale consistent before data harmonization (Table S4).

Cross-validation machine learning model

We have implemented the spatial cross-validation method to train the random forest model, allowing us to split the data into multiple training and test datasets within each cross-validation loop (46). We utilize a grid search strategy to identify the best set of hyperparameters and optimize the model performance (47). In this study, we used 5-fold cross-validation and district as spatial units. One-fold constitutes ~5,600 clusters (Fig. S5). In each iteration of the cross-validation loop, one of the folds is used as test data, and the remaining folds are combined to create the training dataset. The spatial cross-validation approach ensured that our model was evaluated on different subsets of data to assess our model's applicability to unseen data (48). Additionally, we performed a robustness check using the XGBoost method with cross-validation and train and test data partition method. The results of cross-validation models are available in Tables S5 and S6 in the [supplementary material](#).

Results

Nighttime lights as a predictor of socioeconomic inequality

We used nighttime lights as the only predictor to estimate socioeconomic inequality at the cluster level, aiming to assess its explanatory power. The test results indicate that nighttime lights alone provide limited explanatory power, with an R-squared of 26% between surveyed and predicted Gini using nighttime lights in 2015–2016, coupled with a mean squared error (MSE) of 0.014. Similarly, for the 2019–2021 data, the R-squared further decreased

to 22%, with a MSE of 0.013. These findings suggest that nighttime light alone is inadequate to accurately predict socioeconomic inequality (Fig. S6).

Nighttime lights and sociodemographic factors as predictors of socioeconomic inequality

In addition to nighttime lights, we have incorporated other satellite-based sociodemographic factors to enhance the model's accuracy. Our random forest machine learning model finds a strong correlation between the survey-based and predicted Gini coefficient at the cluster level. The MSE on the test dataset is 0.0082, the MSE of the full dataset is 0.0031, and the R-squared shows an 84% correlation between observed and predicted Gini (Fig. 3A). The results indicate that nighttime light is the most significant covariate, contributing (0.326) of the mean decrease in impurity in the model's predictive performance. Other important variables include crop area (0.185), settlement footprint (0.140), NDVI (0.102), distance to the national highway (0.095), total population (0.091), and distance to the nearest administrative center (0.060). Using the same model specification, we trained the DHS 2019–2021 dataset which shows similar results. The MSE for the test dataset is 0.0076, while the MSE for the full dataset is 0.0027. The R-squared value of 84% correlation indicates a stronger correlation between survey-based and predicted Gini coefficients (Fig. 3B). The mean absolute error (MAE) of 4.78% in 2015–2016 and that of 4.47% in 2019–2020 indicate strong model performance, showing that the average prediction error is <5% compared with the range of target values of full datasets.

Analysis of uncertainty in estimates

To assess the uncertainty in predictions, we employed the bootstrap resampling method with 1,000 iterations to estimate the variability of the model's predictions (Fig. S7). This approach

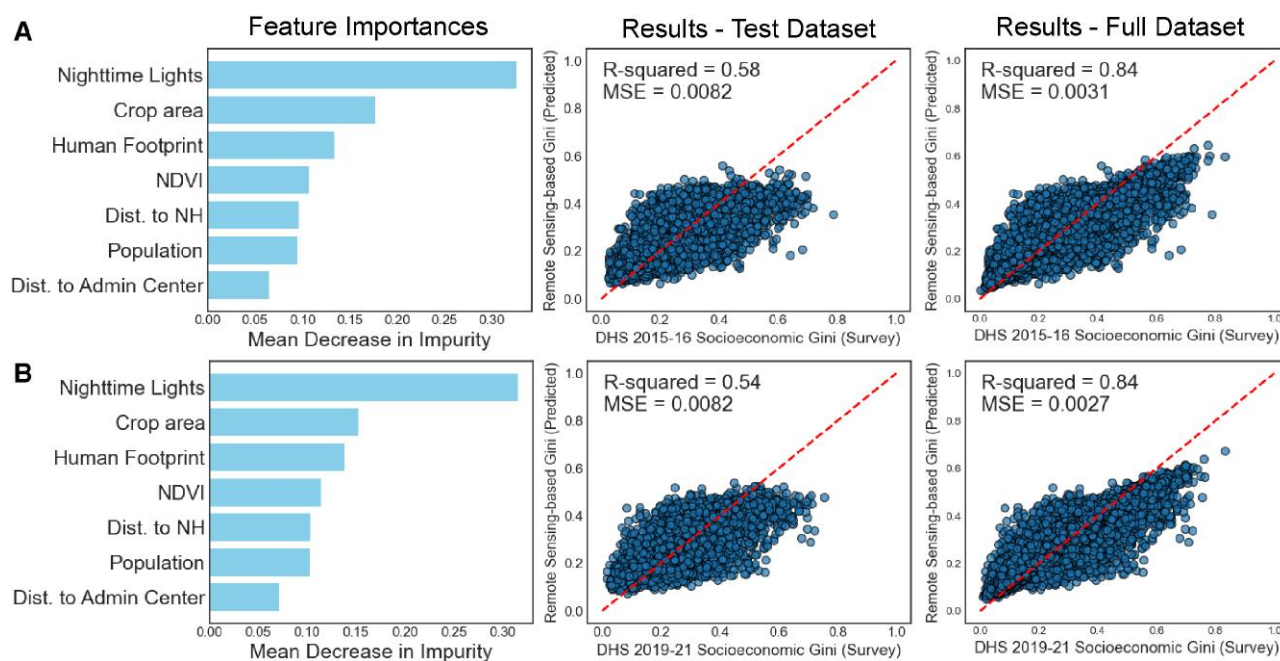


Fig. 3. Machine model performance summary. A) Model performance of test dataset (2015–2016) and B) model performance of full dataset (2019–2021). Left panel: feature importance. The x-axis shows the mean decrease in impurity, and the y-axis shows the feature importance in descending order for the 2015–2016 model. Middle panel: test dataset results. The x-axis shows the cluster-level estimated Gini based on household socioeconomic well-being, and the y-axis shows predicted Gini coefficients derived from remotely sensed data. Right panel: full dataset results. The x-axis shows the cluster-level estimated Gini using household socioeconomic well-being, and the y-axis shows the predicted Gini coefficients based on remotely sensed data.

allowed us to calculate 95% CIs for the prediction errors, with limits of [0.0264, 0.0277]. The narrow range of these intervals indicates that the model's prediction errors are consistent and exhibit limited variability. These findings underscore the model's robustness while accounting for the inherent uncertainty in its estimates.

Feature contributions

We derived SHapley Additive exPlanations (SHAP) values to analyze the machine learning model and elucidate the contribution of each variable in predicting socioeconomic inequality (Fig. S8). SHAP provides a comprehensive summary of the contributions made by each feature or input variable used in a machine learning model (49, 50). The SHAP plots for both DHS 2015–2016 and DHS 2019–2021 consistently demonstrate that higher levels of luminosity correspond to positive SHAP values, while lower luminosity levels exhibit negative SHAP values. This suggests that households residing in well-lit areas tend to have higher predictions of socioeconomic inequality. Conversely, the relationship observed for crop area is inverse, wherein higher crop areas are linked with lower predictions of socioeconomic inequality. Additionally, the results indicate that areas with high population density are associated with higher levels of socioeconomic inequality. Variables such as total population, NDVI, distance to highways, and proximity to administrative centers exhibit SHAP values clustered around 0, suggesting that these factors have minimal impact on predicting socioeconomic inequality.

Validations

Model performance

We validated the reliability of our framework and model by analyzing the difference between the survey-based and predicted Gini indices. Overall, at the cluster level, the model performance shows <5% MAE. We further analyzed aggregated Gini indices at the district level to examine spatial patterns across the country. We calculated the 90th percentile of the absolute differences in Gini indices, identifying the top 10% of districts out of 640 with the largest disparities between the survey-based and predicted values (Fig. S9). The analysis of districts showed that positive differences were mostly found in more developed districts with higher socioeconomic inequality ($N = 45$), while negative differences were observed in less developed districts with lower socioeconomic inequality ($N = 19$), particularly in the northeastern regions. Finally, we analyzed the extreme values with residuals above 0.05, and the scatter plot (Fig. S10) in the [supplementary material](#) shows a few districts that exceed this threshold. This analysis indicates that, overall, the model produces reliable estimates that align with survey data, with relatively moderate performance in a few cases. Possible reasons for these discrepancies include differences in the representativeness of the survey data and variations in the satellite data's ability to capture socioeconomic in regions with limited development or low luminosity. Additionally, we conducted a state-level bivariate analysis to validate the results by comparing the predicted Gini generated from the machine learning model, with the survey-based socioeconomic Gini. The validation showed a significant reduction in the gaps between the two measures (Fig. S11B).

Temporal validation

To validate the model across temporal scales, we used the trained model of DHS 2015–2016 to predict cluster-level Gini using DHS

2019–2021 geographic cluster and vice versa. Since the geographic locations of the clusters are different for both rounds, the test results provide a robust validation of our cross-validation models across temporal scales. The DHS 2019–2021 model performs better on DHS 2015–2016 data with a 71.62% correlation between observed and predicted Gini and the DHS 2015–2016 model finds a 68.49% correlation on DHS 2019–2021 data (Fig. S12). The higher predictive performance of the DHS 2019–2021 model could be due to larger sample sizes and changes in covariates over time (7). Additionally, we compared the model performance between the rural and urban clusters (Fig. S13), block and district (Fig. S14), and the results revealed a strong correlation between observed Gini and remote sensing-based Gini district and block. Finally, we filled temporal socioeconomic inequality data at the district level from 2016 to 2019 (Fig. 4).

Spatial validation

We used spatial interpolation to validate estimated socioeconomic inequality using household socioeconomic well-being index at the district (administrative level 2) and block (administrative level 3). The spatial interpolation followed three steps for the district- and block-level validation: (i) we computed Gini coefficients at the district and block levels using household socioeconomic well-being index, (ii) we aggregated cluster-level Gini coefficients using average at the district and block levels, and (iii) we performed bivariate analysis to compare between the observed Gini and aggregated. We find a strong correlation between the estimated Gini and aggregated Gini. The R-squared value between the block-level socioeconomic Gini and interpolated Gini is 84%, and it is 95% at the district level (Fig. S14). The district-level correlation value is higher than the block- and cluster-level correlation values due to the size of observations and noise in the data. We have ~44 clusters per district and five clusters per block in the DHS 2014–2015 dataset. Figures S17–S20 present the distribution of clusters by state, district, and block. All the 640 districts of the 2011 Indian census have at least five clusters, whereas out of 7,272 blocks, DHS clusters could be mapped with 5,429 blocks.

Robustness check

Analysis of socioeconomic well-being levels as the outcome variable

To perform the robustness check using levels of socioeconomic, we categorized them into poorest, poorer, middle, richer, and richest levels. We computed the average socioeconomic for each cluster and applied the same model specifications along with remotely sensed covariates. The test results show an R-squared of 62% for the test model and 86% for the full dataset (Fig. S15). However, the test model shows a higher MSE of 0.45 compared with 0.16 for the full dataset and an MAE of 7.52%. Although the model gives a better R-squared, the error percentages are higher than those observed in the main results. The results indicate that nighttime light is the most significant covariate, contributing (0.286) of the mean decrease in impurity in the model's predictive performance. Other important variables include crop area (0.202), settlement footprint (0.142), NDVI (0.105), distance to the national highway (0.091), total population (0.110), and distance to the nearest administrative center (0.061).

Alternate machine learning model

We performed robustness checks using the XGBoost model. The comparison of model performance metrics of random forest and

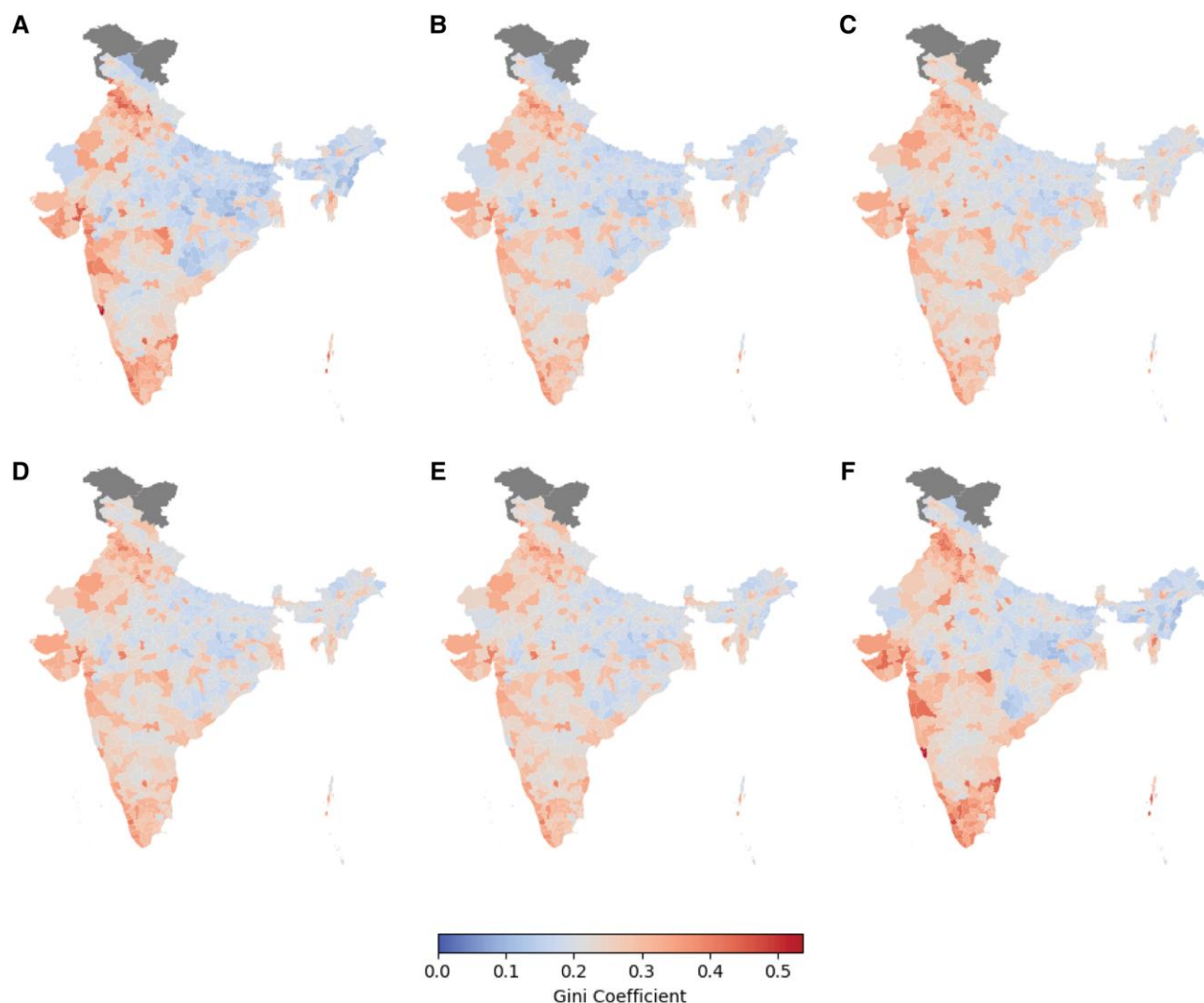


Fig. 4. District-level socioeconomic inequality maps from 2015 to 2020. A) Survey-based socioeconomic Gini coefficient (2015), B) predicted socioeconomic Gini coefficient (2016), C) predicted socioeconomic Gini coefficient (2017), D) predicted socioeconomic Gini coefficient (2018), E) predicted socioeconomic Gini coefficient (2019), and F) survey-based socioeconomic Gini coefficient (2020). Regions with lower Gini coefficients are represented with shades associated with lower intensity, higher Gini coefficients are shown with darker shades. Areas with unavailable data are represented in neutral shade.

XGBoost models is presented in Table S7. We selected the random forest model by comparing the MAE as a percentage of the predicted range between the random forest and XGBoost models on the test data. The random forest model reported an MAE of 9.09%, while the XGBoost model reported 9.32% on the test data.

Discussion

The lack of inequality data at the subnational level has made it impossible to measure and track inequality effectively, hampering efforts to reduce within-country disparities, especially in meeting global targets for L&MICs. Our analysis offers an analytical framework that integrates remotely sensed data with machine learning to estimate inequality at finer jurisdictional, spatial, and temporal scales compared with those available from survey or census data. The novel data harmonization process combines cluster-level socioeconomic inequality with remotely sensed data. The framework is computationally efficient and was implemented using freely available open-source software. Despite the diverse demographics and socioeconomic conditions across rural and urban areas in India, the cross-validated

machine learning model produced inequality estimates closely aligned with socioeconomic inequality measures derived from household survey data. The cluster-level dataset produced from this analysis is comparable with the survey data at local, subnational, and national levels of inequality in India. To validate the results at a temporal scale, we leveraged the household survey data from two survey rounds. The results revealed a consistent relationship between remotely sensed and survey data between the two survey rounds. The association of nighttime lights with economic activities is well established, and evidence is also available for analyzing socioeconomic and poverty through daytime satellite imagery (9, 11). Our study shows how nighttime lights data in combination with other remotely sensed covariates provide more reliable estimates of socioeconomic inequality. The developed approach thus bridges data gaps across spatial and temporal scales, employing cost-effective and scalable machine learning techniques.

We address the challenges reported in prior studies, particularly the mismatch of inequality that is based solely on nighttime lights (21). In addition to nighttime lights, our data harmonization approach includes several other remotely sensed layers that

explain demographic, socioeconomic, and land-use characteristics related to inequality (26). The analysis presents a reliable method for estimating socioeconomic inequality from satellite imagery, despite the fundamental challenges in measuring socioeconomic conditions remotely (51). For instance, it is challenging to distinguish the assets between wealthy and poorer households using nighttime lights and other remotely sensed data. However, we can distinguish the wealthy and poorer neighborhoods by harmonizing data at a finer spatial scale. We trained the machine learning model using the socioeconomic well-being distribution at the cluster level which helped the model “learn” the distribution of demographic and socioeconomic characteristics from satellite data. The data harmonization approach contributes to a broader application of free available remote sensing data and machine learning models to estimate inequality and many other similar social indicators at a fine spatial scale. Our analysis provides a reliable proxy of socioeconomic inequality that can be combined with several other drivers to analyze the causes and effects of inequality by identifying variations in its persistence and changes in relation to a suite of social, economic, and environmental factors. The analysis in our study contributes novel methods to generate socioeconomic measures of inequality globally and their application to address other Sustainable Development Goals related to inequality.

Limitations

The study uses satellite data at the DHS geographic cluster level as reference (ground truth) data to train the model. Each cluster consists of about 23 households, with the Gini coefficient used to measure the distribution of socioeconomic well-being. The sample size used in this study may lead to undersampling and inadequate representation of the socioeconomic well-being distribution within each cluster. This can introduce bias, causing the model to smooth out extremes and potentially underpredict the Gini coefficient. Additionally, generalizing this model to other geographic context beyond India and developing further socioeconomic indices will require careful selection of satellite-based covariates and assessment of the performance of machine learning models used in conjunction with these covariates.

Materials and methods

Dataset

DHS is the most extensive household survey on population, health, nutrition, and socioeconomic data covering more than 90 countries worldwide. It provides the approximate location of each household at the geographic cluster level with a displacement error of 2 and 10 km for urban and rural areas (7). The dataset used in this study comprises 28,000 clusters with 600,000 households in DHS 2015–2016 and 30,000 clusters containing 636,699 households in DHS 2019–2021. One geographic cluster constitutes 20 to 25 households (Fig. S1). The study uses the location coordinates of each DHS cluster with 2- and 5-km buffers for urban and rural areas to extract covariates from remotely sensed data.

Machine learning model

Developing socioeconomic indices at scales becomes possible with the advancement of open-source geospatial computing platforms, free access to satellite imagery, and machine learning models. These socioeconomic indices have helped monitor

several socioeconomic and environmental issues through time and space. However, satellite data need ground truthing and data labeling to extract meaningful information before we use them in machine learning and deep learning models (7). Data labeling is not feasible when we develop large-scale development indices through time and space. In this study, we used each cluster as the ground truth or reference data of socioeconomic well-being distribution and extracted pixel values for each cluster. The study used cluster-level Gini coefficient from the DHS survey as the outcome variable and covariates from remotely sensed data. Table S3 presents the list of covariates and measurement strategies.

Implementation approach

We combine DHS survey data and satellite imagery, executing four steps to estimate socioeconomic inequality at the subnational levels. First, we compute DHS cluster-level socioeconomic inequality; second, we process and extract features from satellite imagery; third, we train the machine learning models to estimate socioeconomic inequality; and finally, we create inequality estimates at the block, district, and state levels for India.

DHS cluster-level socioeconomic inequality

Socioeconomic well-being index

Household-level socioeconomic information constitutes dwelling characteristics (material used for floor, roof, wall, and toilet facility), durable assets (television, refrigerator, motorcycle, mobile, and bike), utilities (access to clean drinking water and cooking fuel), and socioeconomic (dependency ratio, agricultural land, livestock, below-poverty card holder and bank account) to estimate household-level assets accumulation. We normalized the socioeconomic well-being index using the following equation:

Normalized socioeconomic index (NSI)

$$= \frac{(\text{First components} - \text{Minimum value of first component})}{(\text{Maximum value of first component} - \text{Minimum value of first component})} \times 100$$

Socioeconomic inequality

We compute the Gini coefficient to measure geographic cluster-level socioeconomic inequality. Socioeconomic inequality is calculated at the national level for benchmarking with the World Bank's national estimates and further at the state, district, and subdistrict levels to validate the model at finer spatial scales.

$$G = 1 - 2 \times \sum [(i - 1) \times X_i] / (n \times \sum X_i)$$

where X is the normalized socioeconomic index, i represents the rank of the observation (1 to n), X_i represents the value of the variable X at rank i , and \sum denotes the summation operator, adding up the values for all observations.

Extract features from satellite imagery

The study uses 2015 and 2020 satellite imagery for DHS 2015–2016 and DHS 2019–2021 survey rounds. All the satellite imagery used in this study is reprojected and resampled at 500-m spatial resolution to make the spatial scale consistent before feature extraction. Image acquisition, processing, and feature extraction are done using a combination of geospatial toolkits comprising Google Earth Engine (GEE), QGIS, and Python (46, 52, 53).

Nighttime lights

The study extracts the annual average radiance of nighttime light data from the VIIRS Day Night Band annual series with cloud-free and other atmospheric corrections by removing outliers (20). The spatial resolution of the product is 15 arc seconds (~500 m at the equator) with global coverage (180W, 75N, 180E, 65S). The study uses the average radiance for each cluster, indicating the luminosity intensity.

Settlement footprint

The study draws human settlement data from the World Settlement Footprint 2015 (WSF 2015) (54). The data product has a spatial resolution of 10 m (0.32 arcsec) generated using optical and radar satellite imagery. The study resampled the image at a 500-m spatial resolution and computed the total settlement area for each cluster.

Population

WorldPop provides gridded population count data at 100-m resolution for individual countries. We reprojected and resampled the data at 500-m resolution and extracted the total population for the study years (55).

Normalized difference vegetation index

We used USGS Landsat 8 Collection 2 data to calculate NDVI for each cluster. We have performed the image process and reprojected and resampled at 500-m spatial resolution. We calculated the normalized difference between Band 5 (near infrared) and Band 4 (red) surface reflectance.

Land cover

The study draws the total crop areas from Moderate Resolution Imaging Spectroradiometer Land Cover Type (MCD12Q1) version 6.1 data products at a 500-m spatial resolution.

Distance to nearest highway and administrative center

The study obtained India's latest national highway data and computed the nearest distance from each cluster. Similarly, the distance from the cluster location to the nearest administrative center is calculated using India's census 2011 subdistrict shapefile.

Train machine learning models

We trained the random forest model performance employing a cross-validation approach and tuned the hyperparameters using a grid search strategy. We executed three steps: (i) the dataset was divided into 5-folds using district as a spatial unit, (ii) a random forest regressor was created and trained on the training dataset, and (iii) model performance was evaluated for each fold using MSE, SD of MSE, and R-squared values. We followed the same steps for a robustness check using XGBoost and train and test partition approach.

Create inequality estimates at block and district levels

We overlaid the geographic clusters of DHS 2015–2016 and DHS 2019–2021 using block and district boundary and aggregated using cluster-level average for each administrative level. Figures S17–S20 in the [supplementary material](#) present the distribution of cluster by administrative units.

Acknowledgments

The authors thank the editor and reviewers for their comments and suggestions. N.P. thanks Ashwini Chhatre, Ines Ibanez, Pamela Jager, Meha Jain, and Andrew Jones for their feedback on the initial version of the manuscript. N.P. also acknowledges Ian McCallum, Fernando Orduna-Cabrera, Juan Carlos Laso Bayas, and Roman Hoffmann from the International Institute for Applied Systems Analysis and Esther Greenwood from Eawag and ETH Zurich for their input during the Young Summer Scientist Program in 2023, supported by the National Academy of Sciences, USA.

Supplementary Material

[Supplementary material](#) is available at PNAS Nexus online.

Funding

The authors declare no funding.

Author Contributions

N.P. was involved in conceptualization, data curation, formal analysis, validation, investigation, visualization, methodology, writing—original draft, and writing—review and editing. A.A. was involved in conceptualization, supervision, investigation, methodology, writing—original draft, project administration, and writing—review and editing.

Data Availability

The authors declare that the data used in this study are free and publicly accessible. <https://www.dhsprogram.com/methodology/survey/survey-display-541.cfm>. Data analysis was performed using Python, GEE, and QGIS. <https://github.com/easresearch/inequality>.

References

- Smith RJ, Rey SJ. 2018. Spatial approaches to measure subnational inequality: implications for sustainable development goals. *Dev Policy Rev.* 36:O657–O675.
- World Bank Open Data. World Bank Open Data. [accessed 2023 Nov 23]. <https://data.worldbank.org>.
- Solt F. 2020. Measuring income inequality across countries and over time: the standardized world income inequality database. *Soc Sci Q.* 101:1183–1199.
- Smythe IS, Blumenstock JE. 2022. Geographic microtargeting of social assistance with high-resolution poverty maps. *Proc Natl Acad Sci U S A.* 119:e2120025119.
- Galimberti JK, Pichler S, Pleninger R. 2023. Measuring inequality using geospatial data. *World Bank Econ Rev.* 37:549–569.
- Suss J, Kemeny T, Connor DS. 2024. GEOWEALTH-US: spatial wealth inequality data for the United States, 1960–2020. *Sci Data.* 11:253.
- Jean N, et al. 2016. Combining satellite imagery and machine learning to predict poverty. *Science.* 353:790–794.
- Doll CNH, Muller J-P, Elvidge CD. 2000. Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *Ambio.* 29:157–162.
- Chen X, Nordhaus WD. 2011. Using luminosity data as a proxy for economic statistics. *Proc Natl Acad Sci U S A.* 108. 8589–8594.

- 10 Elvidge CD, Baugh KE, Zhizhin M, Hsu F-C. 2013. Why VIIRS data are superior to DMSP for mapping nighttime lights. *APAN Proceedings*. 35:62.
- 11 Levin N, et al. 2020. Remote sensing of night lights: a review and an outlook for the future. *Remote Sens Environ*. 237:111443.
- 12 Bickenbach F, Bode E, Nunnenkamp P, Söder M. 2016. Night lights and regional GDP. *Rev World Econ*. 152:425–447.
- 13 Head A, Manguin M, Tran N, Blumenstock JE. 2017. Can human development be measured with satellite imagery? Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, Lahore, Pakistan, 2017. ACM. p. 1–11.
- 14 Yeh C, et al. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nat Commun*. 11:2583.
- 15 Zucman G. 2019. Global wealth inequality. *Annu Rev Econom*. 11: 109–138.
- 16 Smits J, Steendijk R. 2015. The international wealth index (IWI). *Soc Indic Res*. 122:65–85.
- 17 Cruz M, Foster JE, Quillin B, Schellekens P. 2015. *Ending extreme poverty and sharing prosperity: progress and policies*. Policy Research Note, PRN/15/03. Washington (DC): World Bank.
- 18 Kuhn M, Schularick M, Steins UI. 2020. Income and wealth inequality in America, 1949–2016. *J Polit Economy*. 128:3469–3519.
- 19 Runfola D, et al. 2020. geoBoundaries: a global database of political administrative boundaries. *PLoS One*. 15:e0231866.
- 20 Elvidge CD, Baugh K, Zhizhin M, Hsu FC, Ghosh T. 2017. VIIRS night-time lights. *Int J Remote Sens*. 38:5860–5879.
- 21 Mirza MU, Xu C, Bavel BV, van Nes EH, Scheffer M. 2021. Global inequality remotely sensed. *Proc Natl Acad Sci U S A*. 118: e1919913118.
- 22 Pokhriyal N, Jacques DC. 2017. Combining disparate data sources for improved poverty prediction and mapping. *Proc Natl Acad Sci U S A*. 114:E9783–E9792.
- 23 Davies JB, Sandström S, Shorrocks A, Wolff EN. 2011. The level and distribution of global household wealth. *Econ J (London)*. 121:223–254.
- 24 Tyagi A. 2023 Mar 25. Income inequality in Indian states. *EPW*. 58(12).
- 25 Singhal A, Sahu S, Chattopadhyay S, Mukherjee A, Bhanja SN. 2020. Using night time lights to find regional inequality in India and its relationship with economic development. *PLoS One*. 15: e0241907.
- 26 Chancel L, Piketty T. 2019. Indian income inequality, 1922–2015: from British Raj to Billionaire Raj? *Rev Income Wealth*. 65:S33–S62.
- 27 Revathi E, Awasthi IC, Reddy BS, Madan Aditi. Intersecting paths of sustainable development, urbanization, and women's empowerment. 22nd Annual Conference of Indian Association of Social Science Institutions (IASSI), CESS, Hyderabad, India, 2–4 November 2023. Springer.
- 28 The DHS Program. Country Main. [accessed 2023 Mar 9]. https://dhsprogram.com/Countries/Country-Main.cfm?ctry_id=57.
- 29 The DHS Program. Research Topics—Wealth Index. [accessed 2023 Mar 9]. <https://dhsprogram.com/topics/wealth-index/>.
- 30 McCallum I, et al. 2022. Estimating global economic well-being with unlit settlements. *Nat Commun*. 13:2459.
- 31 Haughton J, Khandker SR. *Handbook on poverty and inequality*. World Bank, Washington, DC, 2009.
- 32 Sitthiyot T, Holasut K. 2020. A simple method for measuring inequality. *Palgrave Commun*. 6:1–9.
- 33 Benami E, et al. 2021. Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nat Rev Earth Environ*. 2:140–159.
- 34 Paliwal A, Jain M. 2020. The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. *Front Sustain Food Syst*. 4:25.
- 35 Thenkabail PS, Schull M, Turrall H. 2005. Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data. *Remote Sens Environ*. 95: 317–341.
- 36 Jain M, et al. 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. *Environ Res Lett*. 12:094011.
- 37 Milesi C, et al. 2010. Decadal variations in NDVI and food production in India. *Remote Sens (Basel)*. 2:758–776.
- 38 Sarmah S, Jia G, Zhang A. 2018. Satellite view of seasonal greenness trends and controls in South Asia. *Environ Res Lett*. 13: 034026.
- 39 Szreter S. 2015. Wealth, population, and inequality: a review essay. *Popul Dev Rev*. 41:343–354.
- 40 Boo G, et al. 2022. High-resolution population estimation using household survey data and building footprints. *Nat Commun*. 13:1330.
- 41 Venter O, et al. 2016. Global terrestrial Human Footprint maps for 1993 and 2009. *Sci Data*. 3:160067.
- 42 Alsamawi A, McBain D, Murray J, Lenzen M, Wiebe KS. The inequality footprints of nations; A novel approach to quantitative accounting of income inequality. In: Alsamawi A, McBain D, Murray J, Lenzen M, Wiebe KS, editors. *The social footprints of global trade*. Springer, Singapore, 2017. p. 69–91.
- 43 Kaushik B. 2005. UNU-WIDER: working paper: globalization, poverty and inequality. [accessed 2024 May 24]. <http://www.wider.unu.edu/publication/globalization-poverty-and-inequality>.
- 44 Banick RS, Kawasoe Y. 2019. Measuring inequality of access: Modeling physical remoteness in nepal. World Bank Policy Research Working Paper 8966. [accessed 2024 May 24]. <https://papers.ssrn.com/abstract=3433473>.
- 45 Hasan S, Wang X, Khoo YB, Foliente G. 2017. Accessibility and socio-economic development of human settlements. *PLoS One*. 12:e0179620.
- 46 Pedregosa F, et al. 2011. Scikit-learn: machine learning in python. *J Mach Learn Res*. 12:2825–2830.
- 47 Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *J Mach Learn Res*. 13(2).
- 48 Roberts DR, et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 40:913–929.
- 49 Lundberg SM, Lee S-I. 2017. A unified approach to interpreting model predictions. In: Guyon I, et al., editors. *Advances in neural information processing systems* (Vol. 30). Red Hook (NY): Curran Associates. p. 4765–4774.
- 50 Molnar C. Interpretable machine learning. Lulu. com, 2020.
- 51 Agrawal A, et al. 2022. From environmental governance to governance for sustainability. *One Earth*. 5:615–621.
- 52 Gorelick N, et al. 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ*. 202:18–27.
- 53 QGIS Development Team. 2002. QGIS geographic information system. QGIS Association, QGIS Project. <https://www.qgis.org>.
- 54 Marconcini M, et al. 2020. Outlining where humans live, the World Settlement Footprint 2015. *Sci Data*. 7:242.
- 55 Open Spatial Demographic Data and Research. WorldPop. [accessed 2023 Nov 23]. <https://www.worldpop.org/>.