Response to Editor and Reviewers' Comments on Earth System Science Data manuscript ESSD-2025-260

Title: Geospatial micro-estimates of slum populations in 129 Global South countries using machine learning and public data

Overall Response: We would like to thank the editor and the two reviewers for their detailed comments and suggestions, which are very helpful for improving the quality of the manuscript. Please see our point-by-point response to all the comments and suggestions from the reviewers. Text marked in red denotes newly added or amended content.

Reviewers Comments:

Anonymous Referee #1

This study presents a standardized, machine learning-based approach to estimate slum populations at the neighborhood level across 129 Global South countries, using satellite imagery, household surveys, and gridded population data. The methodology aligns with the UN SDG 11.1 framework and demonstrates strong performance (R^2 = 0.82–0.96; RMSE = 4.85–10.47%), outperforming previous benchmarks. By addressing the limitations of government-reported data, this work provides the first comprehensive geospatial inventory of slum populations, offering critical insights for urban planning and humanitarian efforts.

This is a pioneering study that leverages satellite imagery to estimate slum populations at a regional scale, with significant potential for future applications in urban policy, research, and humanitarian aid. The manuscript is well-written, and the methodology is rigorous and clearly presented. I have only a few minor suggestions for the authors to consider before publication.

Response: Thank you for your encouraging feedback and affirming the value of our work. Please find our point-by-point responses below.

#1 The study employs a fine-tuned CNN model (ResNet) and XGBoost to classify slum households, which is innovative. However, the necessity of fine-tuning the CNN and performing extensive feature extraction is not entirely clear. It appears that classification might be achievable using the original satellite image bands without expanding the RGB inputs into 500+ features. Could the authors elaborate on the specific benefits of fine-tuning and high-dimensional feature extraction in this context? Additionally, did you evaluate the performance improvement compared to models using only raw image bands? Such comparison would help clarify the added value of the proposed approach.

Response: (1) We appreciate your thoughtful comment regarding the necessity of fine-tuning the CNN and performing extensive feature extraction. In our study, we fine-tuned the CNN to adapt it to our specific input setting—7-channel satellite imagery rather than the standard 3-band RGB. The incorporation of the additional spectral bands substantially increases the dimensionality and richness of the imagery, enabling the model to capture more nuanced and relevant features. The superior performance of the 7-band model compared to the RGB-only baseline confirms the benefits of this input enhancement [see part (2) in this response].

While pre-trained CNNs (e.g., ImageNet-based) provide a strong starting point for extracting general visual features, they are not optimized for the spectral characteristics critical to our task, particular those linked to deprived housing conditions. Fine-tuning enables the model to adjust low- and mid-level filters, better capturing spatial-spectral patterns indicative of slum household conditions.

Instead of using the CNN for classification, we extracted 512-dimensional feature from its penultimate layer. These features condense key spatial-spectral patterns relevant for identifying slum conditions. We then fed these features into an XGBoost regression model, which excels at modeling complex, non-linear relationships. This hybrid setup combines the strong feature extraction capability of CNNs with the flexibility and robustness of tree-based regression, yielding more accurate and robust predictions.

We have expanded the Discussion section to further explain the rationale and advantages of both CNN fine-tuning and high-dimensional feature extraction in our modeling framework (Lines 601-616), as follows:

5 Discussions

5.1 Robustness and external validation

We fine-tune the ResNet-34 model on 7-band satellite imagery to more effectively capture the complex spatial-spectral patterns associated with deprived housing conditions. The original ResNet-34, pre-trained on 3-channel RGB images, is optimized for extracting general visual features such as edges and textures. By fine-tuning, we adapt its early and intermediate filters to recognize context-specific visual cues—such as roof material, settlement density, and vegetation coverage—potentially indicative of slum-like environments, while exploiting the additional spectral information provided by multispectral inputs. Rather than using ResNet-34 for direct prediction, we extract 512-dimensional embeddings from its penultimate layer and use these as inputs to an XGBoost regression model. This hybrid architecture combines the representational strength of deep CNNs with the predictive robustness of gradient-boosted trees, enabling it to model complex, non-linear relationships in the data. The superior performance of this approach over RGB-only baseline models highlights the importance of multispectral fine-tuning and high-dimensional feature extraction for accurately modeling housing-related deprivation.

(2) We agree that evaluating model performance using only raw image bands is essential for clarifying the added value of our proposed approach. Following your suggestions, we have implemented baseline models that rely solely on raw image bands and compared their performance with our proposed approach (see Table 5, Results 4.4). The results show that our approach consistently outperforms the baseline,

with average increases in R^2 of 0.34 and reductions in RMSE of 6.81 units. Corresponding revisions to the Methods and Results sections are shown as follows (Lines 511-525):

4 Results

4.4 Robustness of our regional models

Table 5 reports the performance of nine regional models trained exclusively on RGB image bands. Across all regions, the RGB-only models consistently underperform relative to the 7-band models used in our main approach. On average, RGB-only models exhibit a 6.81-unit higher RMSE and a 0.34 lower R^2, indicating both reduced predictive accuracy and weaker model fit. The performance gap is particularly pronounced in regions such as West Africa and East Asia. These findings illustrate the limitations of relying solely on visible spectrum data for capturing slum-related features and underscore the value of incorporating additional spectral channels such as near-infrared and shortwave infrared to improve model robustness and predictive power.

Table 5 Regional model performance using only RGB image bands. ΔR^2 and ΔRMSE are the differences in model performance between the 7-band models and the RGB-only models.

Region/Income Group	Model	RMSE (%)	R^2	ΔRMSE (%)	ΔR^2
West Africa – Low Income	Model 1	12.98	0.58	−7.69	0.35
East Africa – Low Income	Model 2	12.65	0.53	−6.77	0.36
West Africa – Lower middle income	Model 3	15.00	0.52	−8.60	0.40
East Africa – Lower middle income	Model 4	14.51	0.75	−8.51	0.20
Africa – Upper middle income	Model 5	14.78	0.52	−5.86	0.31
West Asia – Lower middle income	Model 6	10.29	0.41	−5.09	0.43
East Asia – Lower middle income	Model 7	13.34	0.45	−7.47	0.44
Latin American – Lower middle and low income	Model 8	13.99	0.63	−7.63	0.26
Latin American – High and upper middle income	Model 9	13.84	0.49	−3.67	0.33

#2 Another concern relates to the quality of the ground-truth labels used for training and evaluating the model. In particular, for the demographic and health surveys utilized, did the authors perform any quality assessment or validation of the labeling process? Additional details on how label reliability was ensured would strengthen the

credibility of the results.

Response: Thank you for raising this important point. We fully agree that ensuring the reliability of ground-truth labels is crucial for robust model training and evaluation. In our study, label quality is ensured through three key measures. First, we rely on high-quality underlying DHS datasets. Second, before constructing the slum household indicator, we implement rigorous data cleaning procedures to remove inconsistencies and errors. Third, we evaluate the robustness of the indicator framework using two alternative weighting schemes. Details of these quality control procedures have been added to the revised manuscript (Lines 246-255), as follows:

3 Methodology

3.1 Framework of slum indicator

…. The DHS surveys' standardized instruments and rigorous quality control protocols ensure high reliability of the data, which in turn supports the validity of our labels. Before constructing the slum household indicator, we applied a comprehensive data cleaning process. Sub-indicator with missing values, "don't know" responses, or implausible entries were excluded. If all sub-indicators within a given dimension were missing, the corresponding household record was removed. Clusters with invalid GPS coordinates were also excluded. Additionally, to evaluate the sensitivity of the labels to the indicator weighting scheme, we implemented two alternative weighting scenarios (see Section 3.5 for more details).

Specific comments:

Figure 2. The numbers seem confusing. Shouldn't it be from f(x)1 to f(x)7 and tree1 to tree7, layer 1 to layer 7?

Response: Sorry for the incorrect numbering. We have corrected them and updated Figure 2 following the suggestions of both Reviewers. The updated figure 2 is shown as follows (Lines 304-307):

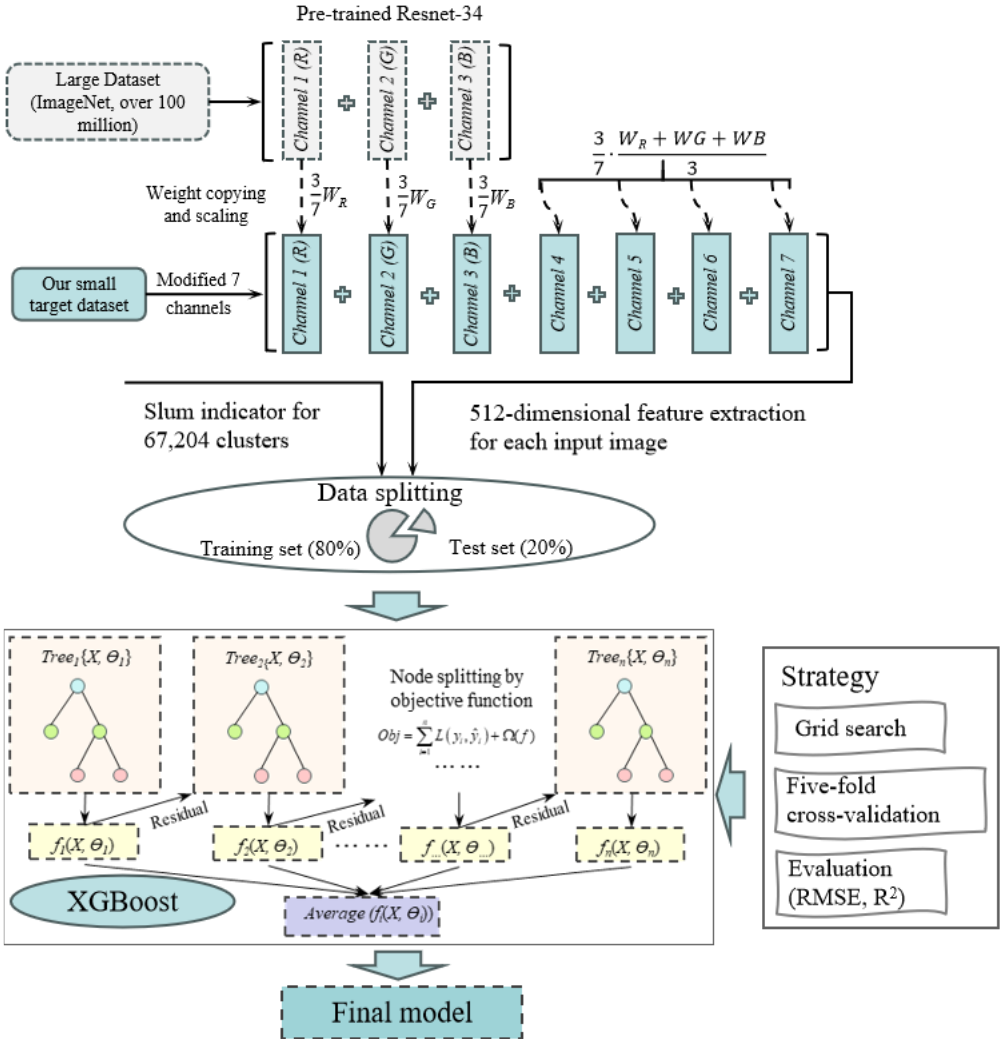

Figure 2 Schematic of the proposed model architecture. It integrates transfer learning by fine-tuning ResNet-34 to extract high-dimensional spatial–spectral features from multispectral imagery, followed by XGBoost regression for predictive modeling.

Anonymous Referee #2

1. The authors take advantage of machine learning, satellite imagery, local surveys, and global population data to produce a gridded estimate of population in slum-like conditions for 129 Global South countries. The authors propose a generalized, regional modeling effort as country-specific data is varied in quantity and quality. There is certainly a need for this information. Humanitarian efforts, public health, and emergency response are among the countless uses that could benefit from these data. I believe this paper and the associated slum population maps have potential to be impactful for the community. However, I believe the paper has several major issues that should be addressed.

Response: Thank you very much for the excellent summary of the primary contribution of our research. We have made thorough revisions in line with your suggestions.

2. I have details below, but in summary, the paper could benefit from 1) Clearer and more detailed descriptions of the methods, particularly regarding combining and overlaying raster datasets with different resolutions. 2) Consistent definition and use of terms (e.g., clusters, settlement slices, neighborhoods) 3) Tables and figures along with their captions should be stand alone and require limited information from the text. Likewise, the text should not rely on content only described in a caption. Figures 1 and 2 are vital to the paper and care must be taken to make sure all details within them are consistent and clear. 4) Parts of the text should be reorganized into more appropriate sections.

I hope the authors take these notes in the constructive way they are intended. I believe the paper will be more impactful and useful with these major adjustments.

Response: We sincerely appreciate your constructive comments regarding the clarity of the methods section, the consistent definition and use of key terms, the revision of table and figure captions, and the reorganization of specific parts of the manuscript. In response, we have carefully revised the text to address each of these issues. The point-by-point responses to each of your concerns are presented below.

RE: Resolution and scale

3. The use of “cluster” and “cluster-level” throughout is confusing. Clusters are not defined in the paper, but I assume the authors are referring to DHS enumeration areas as “clusters”. I believe the authors are stating a pixel of size ~6.72 km is roughly equivalent to the ‘average’ size of a DHS enumeration unit (neighborhood or village). It should be clear when the authors are referring to irregularly shaped vector

boundaries (e.g., DHS clusters, villages) or a pixel with 6.72 km dimensions. For example, on line 156, you collect Landsat imagery centered on each cluster. Are you collecting imagery for each DHS cluster or is the imagery collected for every 6.72 km grid cell?

Response: (1) Thanks for your valuable comments. We acknowledge that our use of the terms "cluster" and "cluster level" was inconsistent and may have caused confusion. As you correctly noted, we refer to DHS enumeration areas as "clusters". The pixel size of ~6.72 km is roughly equivalent to the 'average' size of a DHS enumeration unit, thus we originally used the term "cluster level" to describe these standardized spatial units.

To improve clarity, we have standardized terminology throughout the manuscript. Specifically, we now use "cluster" exclusively when referring to DHS enumeration areas, and we provide a clear definition of clusters in the Section 2.1. The term "grid cell" is used when referring to uniform spatial units with a resolution of approximately 6.72 km. The added definition of cluster in Section 2.1 (Lines 137-138):

Here, a cluster refers to a DHS enumeration unit, approximately the size of a neighbourhood or village.

When referring to uniform spatial units of approximately 6.72 km resolution, we now use the term "grid cell".

(2) Regarding the original line 156, we apologize for the lack of clarity. There is indeed a slight difference in the image collection strategy between the modeling and mapping processes. In the modelling stage, the input image collection is centered on the geographic coordinates of each DHS cluster, using a standardized grid cell size of approximately 6.72 km to ensure spatial alignment with DHS cluster labels. In the mapping (prediction) stage, images are collected for every 6.72 km grid cell to produce spatially continuous predictions. We have revised the text accordingly to clearly describe this distinction, as follows (Lines 158-161, and 182-186):

2 Dataset description

2.2 Satellite imagery

We obtained publicly available Landsat-7 ETM+ (Collection 2, Tier 1), Landsat-8 OLI (Collection 2, Tier 1), and nighttime light imagery from the Google Earth Engine platform (Tamiminia et al., 2020), along with the Global NPP-VIIRS-like nighttime light dataset (Chen et al., 2021). ……The input images used for model training are centered on the geographic coordinates of each DHS cluster, using a standardized grid cell size

of approximately 6.72 km to ensure spatial alignment with the corresponding DHS labels. For map generation, images are collected for every 6.72 km grid cell across the study area, enabling the production of spatially continuous predictions.

4. In the abstract, the authors refer to these pixels as neighborhood level (line 27) and as cluster-level in the rest of the paper.

Response: Thank you for pointing this out. To ensure consistency, we have revised the terminology throughout the manuscript. Specifically, we have replaced "neighborhood level" in the abstract and "cluster level" in other sections with "grid-cell level" to consistently describe the spatial resolution of the generated maps (approximately 6.72 km). The revised abstract statement now reads (line 27):

Abstract

… Here, we develop a standardized bottom-up approach to estimate the slum population at the grid-cell level (~6.72 km resolution at the equator) for 129 Global South countries in 2018.

5. The authors go back and forth between cluster-level, 6.72km and 3.63 arc-minutes. A consistent approach would be clearer.

Response: Thanks for your comments. We now use "grid-cell level" throughout the manuscript when refer to the analysis and mapping resolution. The term "6.72 km" is retained only where the exact spatial resolution needs to be specified.

6. The authors also refer to "settlement slices". Are these also 6.72km grid cells or are they something else?

Response: Thanks for your comments. We apologize for the earlier ambiguity. The term "settlement slice" refers to built-up identified in the Global Human Settlement Layer (GHSL) within 6.72 km grid cells. To avoid confusion, we have removed this term and now use "grid cells" consistently. The revised sentence is as follows (Lines 450-453):

… The map is based on 6.72 km grid cells intersecting built-up areas with populations exceeding 1,000, in alignment with the urban-rural delineations provided by the Global Human Settlement Layer.

7. The authors say on line 521 that the resolution was carefully chosen based on several factors, including the characteristics of satellite imagery, algorithm architecture, and the truth-ground survey data. On line 673 the resolution was to protect privacy.

On line 530 the resolution was set because the model required 224x224 input images and 30 m * 224 = 6.72 km. My hunch is the explanation on line 530 is the primary reason for the output resolution. It is appropriate to discuss the pros and cons of this decision but be clear and consistent about your reasoning.

Response: (1) Thank you for pointing out the inconsistencies in our explanation of the spatial resolution choice. As stated in the original line 530, the primary reason for selecting a 6.72 km resolution is the requirement of our model architecture, which processes 224 × 224 pixel inputs derived from 30 m Landsat imagery. Other factors mentioned previously, such as the characteristics of the ground-truth survey data and the spatial displacement applied for privacy protection (the original lines 521 and 673), serve as supplementary justifications.

We have reorganized the relevant sections to provide a consistent explanation and to more clearly articulate the associated trade-offs, including model compatibility and spatial uncertainty due to DHS displacement. The revised Discussion section (Lines 659-669) now reads:

5 Discussions

5.2 Spatial resolution and dataset selection

…The mapping resolution is primarily determined by the technical requirements of the deep learning architecture and the native spatial resolution of the satellite imagery used. Specifically, the ResNet-34 model for feature extraction requires input images in a 224x224 pixel format (Wu et al., 2019). It is also crucial to ensure spatial and temporal consistency between the satellite images and survey data, which vary across years and countries. To enhance input richness, we prioritize satellite images with more spectral bands than standard RGB bands, such as Landsat imagery, which offers consistent temporal coverage, global reach, and multiple spectral bands at 30-meter resolution (Wulder et al., 2022). Based on these constraints and design considerations, the mapping resolution is set at 6.72 km at the equator.

(2) For the original Line 673 in the Conclusion section, since the determinants of spatial resolution are now fully addressed in the Discussion, we have removed the repeated explanation. The revised sentence (Lines 773-775) reads:

… The resulting maps represent the first comprehensive inventory of slum populations across Global South countries, produced at a spatial resolution of 6.72km at the equator.

8. There are several vector and raster datasets used throughout with different

resolutions. It would be helpful to consistently and clearly describe when and how these are converted or resampled (mean, mode, natural neighbor, bicubic, etc.)

- For example, I resampled the GHS_POP Population map by converting the raster to points and then aggregating (summing) those points to a new raster that aligned with your slum population map. For the 6.2 km pixel with centroid of (77.11267606, 23.16936109) I obtained a total population of ~16,324. The slum population value for this pixel is ~38,762. Did I resample differently or is it possible that your method can produce slum population estimates > GHS_POP estimates?

Response: Thank you very much for your insightful comments and for sharing your resampling approach. We fully agree that careful handling of spatial resolution and alignment is essential when integrating multiple datasets. In our workflow, the GHS-POP raster is first aggregated from its original 1 km resolution by summing adjacent cells to match the approximate scale of our target grid (~7 km, an integer multiple of the original resolution). All datasets are then projected to a common geographic coordinate system (WGS 1984). The aggregated GHS-POP raster is subsequently resampled to the target ~6.72km resolution using cubic convolution, with a snap raster applied to ensure precise alignment with our slum prediction map. Categorical datasets (e.g., GHS-SMOD, CGLS-LC100) are resampled using nearest-neighbor interpolation to preserve class integrity. We have added a detailed description of this process in the Methods (Lines 365-372):

3 Methodology

3.4 Out-of-sample predictions

…All auxiliary datasets are resampled to match the 6.72 km spatial resolution of the predicted grid cells. The GHS-POP raster is aggregated by summing adjacent cells to match the approximate scale of the target grid, and then resampled to a 6.72km grid using cubic convolution, with snapping applied to ensure alignment with the predicted slum indicator map. Categorical datasets (GHS-SMOD and CGLS-LC100) are resampled using nearest-neighbor interpolation to retain categorical integrity. All datasets are projected to WGS 1984 for spatial consistency.

We also wish to clarify that the estimated slum population must be lower than the GHS-POP values, as it is calculated by multiplying GHS-POP by the predicted slum household percentage (strictly less than 1). In our raster-based method, the value at the centroid is approximately 79,389 for the 6.72km grid. The discrepancy with your result likely arises from differences in aggregation methods and spatial resolution—your calculation may have been based on a 6.2km grid rather than 6.72km. More

importantly, this location lies at the urban fringe, intersecting multiple grid cells. In such cases, point-based aggregation is highly sensitive to boundary effects.

At large spatial scales, we observed that many point fell near the edges of adjacent grid cells, making their assignment sensitive to even minor changes in grid alignment or projection parameters. To address this, we adopt a raster-based aggregation approach: summing values within consistently defined raster blocks and applying a snap raster during resampling minimizes edge effects and ensures more stable, reproducible results.

Organization.

9. The order of 3.2 and 3.3 seems off. On line 296 you produce the 'final model' from the entire dataset after determining the optimal hyperparameters. In 3.3 there are 9 models with 5-fold cross validation. Is the cross-validation part of the grid search? Its not clear how you use the cross-validation results to get to the final 9 models mentioned in line 325. Do you retrain on the entire dataset after the cross-validation? Reorganizing these sections would make this clearer.

Response: Thank you for your constructive comments. We apologize for the confusion between Sections 3.2 and 3.3. In the revised manuscript, we have reorganized these sections to improve clarity and coherence. Specifically, Section 3.2 now introduces the overall model framework, while Section 3.3 provides a detailed description of model training, cross-validation, and testing. The description of the grid search procedure has been moved from Section 3.2 to Section 3.3 to better refelct the workflow.

Regarding your specific questions: 1) Is cross-validation part of the grid search? Grid search is a systematic procedure for evaluating all specified hyperparameter combinations to identify the optimal set. Cross-validation, in turn, estimates the model performance for each hyperparameter combination. While cross-validation is not inherently part of grid search, is the two are commonly combined—cross-validation is applied to each candidate hyperparameter set, and the set with the highest average performance is selected as optimal. We have clarified this relationship in Section 3.3.

2) Do you retrain on the entire dataset after cross-validation? A model's parameters can be divided into hyperparameters and learned weight parameters. Cross-validation is used to determine the optimal hyperparameters. Once selected, the model is retrained on the combined training and validation set using these hyperparameters, thereby updating the weight parameters based on the fully available data. We have also made this workflow explicit in the revised manuscript.

Accordingly, all training- and validation-related content (including the grid search description in the original Section 3.2) has been consolidated in Section 3.3. The

revised text (Lines 332-346) reads:

3 Methodology

3.3 Model training, cross validation and testing

…We perform a grid search to identify the optimal combinations of hyperparameters and apply L_2 regularization to the loss function to prevent overfitting. The tuned hyperparameters include the learning rate (learning_rate), maximum tree depth (max_depth), fraction of training data used for building each tree (subsample), fraction of features per tree (colsample_bytree), regularization term controlling tree complexity (gamma), and number of boosting rounds (n_estimators). Table 3 lists the candidate values considered. For each hyperparameter combination, we conduct 5-fold cross-validation to train the models and compute average validation performance. The combination with the highest average performance is selected as optimal. We then retrain the model on the entire training and validation dataset using these optimal hyperparameters to update its weight parameters. The final model is evaluated on the hold-out test set to assess generalization. All training, cross-validation, and testing steps are conducted independently for each of the nine regional groups, following the defined data partitioning scheme, yielding nine distinct final regional models.

10. Section 3.4 This is the final part of the methods and should be expanded on. "We apply satellite imagery and nighttime light data to grid cells" is vague. Same for "we integrate the GHS-POP dataset with the map of cluster-level slum indicator."

Response: Thank you for your helpful comments. We have now expanded Section 3.4 to provide more detailed descriptions of our methods, including how satellite imagery and nighttime light data are applied to grid cells, and how the GHS-POP dataset is integrated with cluster-level slum indicators. The revised section (Lines 354-381) reads:

3 Methodology

3.4 Out-of-samples predictions

Using our final models, we predict slum indicators at the grid-cell level across 129 countries in the Global South. Collected satellite imagery and nighttime light data are first divided into 224×224 pixel slices. The GHS-SMOD and CGLS-LC100 datasets are then applied as spatial masks to classify urban, suburban, and rural areas, and to exclude grid-cells dominated by bare ground or sparse vegetation. Based on population estimates from the GHS-POP dataset, any image slice covering an area with a total population below 1,000 people is excluded to avoid unreliable out-of-sample

predictions.

The remaining image slices are processed by fine-tuned ResNet-34 models to extract 512-dimensional features, which are subsequently passed to the final regional XGBoost models to generate slum indicator predictions for each 6.72 km grid cell. All auxiliary datasets are resampled to match the 6.72 km spatial resolution of the predicted grid cells. The GHS-POP raster is aggregated by summing adjacent cells to match the approximate scale of the target grid, and then resampled to a 6.72km grid using cubic convolution, with snapping applied to ensure alignment with the predicted slum indicator map. Categorical datasets (GHS-SMOD and CGLS-LC100) are resampled using nearest-neighbor interpolation to retain categorical integrity. All datasets are projected to WGS 1984 for spatial consistency.

To produce the final map of population living in slums or slum-like conditions, we integrate the GHS-POP dataset with the predicted slum indicator map. In the absence of household-level population statistics at the grid-cell level, we assume that slum and non-slum households have equal average household sizes within each 6.72-km grid cell. Under this assumption, the proportion of slum households corresponds to the proportion of the population living in slums or slum-like conditions. Accordingly, the resampled and spatially aligned GHS-POP population data are multiplied by the predicted slum indicator values to generate the final grid-cell–level map of slum populations.

11. There is no section 3.5

Response: Thank you for pointing this out. Section 3.6 was misnumbered and has been corrected to 3.5 to ensure consistent numbering.

12. The methods, results, and discussion bleed into each other at various parts. There are more than these examples. Please check throughout.

Response: Thank you for your valuable comments. We have thoroughly reviewed the entire manuscript to ensure a clearer separation between the Methods, Results, and Discussion sections.

- 363-368 – Move to methods.

Response: These sentences have been moved to Methods section 3.5 (Lines 396-400).

- 380-386 – Move to discussion.

Response: These sentences have been moved to Discussions Section 5 (Lines 590-597).

- 401-402 – Move to methods. Refine how? Which land cover types do you mask? It is not clear what you are doing with the land cover data.

Response: We have moved these sentences to the Methods section 3.4 and clarified

how the land cover dataset is used (Lines 356-358):

The GHS-SMOD and CGLS-LC100 datasets are then applied as spatial masks to classify urban, suburban and rural areas, and to exclude grid-cells dominated by bare ground or sparse vegetation.

- 431-437 Move to methods. Also how do you implement the hierarchical classification from very low to very high. Thresholds?

Responses: We have moved these sentences to Methods Section 3.4 and clarified the thresholds. The corresponding revisions are as follows (Lines 382-390):

We further analyze the local shares of slum population in Global South countries at a resolution of 6.72km. This metric enables cross-country and cross-regional comparisons while accounting for differences in population size. To capture the degree to which local populations lack adequate infrastructure and housing, we implement a hierarchical classification of slum population share, ranging from very low to very high concentration. Specifically, concentration levels are defined by the proportion of the slum population within each grid cell: very low (< 10%), low (10%-20%), moderate (20%-40%), high (40%-60%) and very high (> 60%).

- 457-461 Move to discussion.

Response: We have moved this part to Discussion Section 5.3 (Lines 713-719).

- Section 5.3 is a mixture of methods, results, and discussion

Response: Thank you for your helpful suggestion. We have reorganized the original section 5.3 by moving the methods-relevant part into Section 3.5 (Lines 417-421), the results-related part into Section 4.5 (Lines 555-585), discussions into Section 5.1 (Lines 625-655). The sections and figures have also been renumbered accordingly.

Accuracy and Robustness

13. How did you calculate RMSE? On line 367 the slum indicator (i.e., percentage of households in slum-like conditions) is the ground truth and your method (i.e., total population living in slum-like conditions) is the predicated value. How did you get these two values on the same scale? Did you assume percentage of households was the same as the percentage of population? Did you scale population based on average household size?

Response: Thanks for your valuable comments. The RMSE reported in line 367 is computed by comparing the ground-truth slum indicator—defined as the percentage of households living in slum-like conditions within each cluster—with the model-predicted slum indicator, which is expressed on the same percentage scale. This ensures both the ground truth and the predicted values are directly comparable for

RMSE calculation.

As you rightly noted, our framework makes an assumption when converting the model-predicted percentage of slum households into estimates of slum populations. Due to the absence of data on slum versus non-slum household size data at the grid level, we assume that average household sizes are comparable within each 6.72 km grid cell. Evidence from the 2011 Indian National Census supports this assumption, showing that household size distributions for slum and non-slum households are remarkably similar. For example, 1–3 member households comprise 29.0% of urban households and 28.1% of slum households, while 6–8 member households account for 20.6% and 22.2%, respectively. Under this assumption, the percentage of slum households is treated as equivalent to the percentage of the population living in slums or slum-like conditions.

We have clarified the RMSE calculation, explained the conversion from slum households to slum population, and explicitly discussed this underlying assumption in the revised manuscript, as follows (Lines 373-381, 396-400, 727-737):

3 Methodology

3.4 Out-of-samples predictions

To produce the final map of population living in slums or slum-like conditions, we integrate the GHS-POP dataset with the predicted slum indicator map. In the absence of household-level population statistics at the grid-cell level, we assume that slum and non-slum households have equal average household sizes within each 6.72-km grid cell. Under this assumption, the proportion of slum households corresponds to the proportion of the population living in slums or slum-like conditions. Accordingly, the resampled and spatially aligned GHS-POP population data are multiplies by the predicted slum indicator values to generate the final grid-cell-level map of slum populations.

3.5 Evaluation approaches robustness analysis

We evaluate model performance using two metrics: the coefficient of determination (R^2) and root mean squared error (RMSE). The household-based slum indicator calculated from DHS data serves as the ground truth, while our method generates the predicted slum household indicators.

5 Discussions

5.3 Implications and future research

…In this study, we assume that the proportion of slum households is equivalent to the proportion of the population living in slums or slum-like conditions. This is a

practical approximation given the lack of the household size statistics at the grid-cell scale. Evidence from the 2011 Indian National Census (Chandramouli, 2011) supports its plausibility, showing minimal differences in household size distribution between urban and slum households. For example, 1-3 member households comprise 29.0% of urban households and 28.1% of slum households, while 6-8 member households account for 20.6% and 22.2%, respectively. Nonetheless, we acknowledge this as a potential limitation. Future work incorporating spatially explicit household size data could further improve the accuracy of slum population estimates.

References:

Chandramouli C. Housing stock, amenities and assets in slums-Census 2011. New Delhi: Office of the Registrar General and Census Commissioner; 2011.

14. The paragraph starting at 349 can use more explanation. It is confusing to introduce the new weighting scenarios and then in the next sentence say "we also compare the performance of regional models with country-specific models" I assume the country-specific models for the comparison follow the "equal weight" (0.33, 0.33, 0.33) scenario but it's not clear with the previous sentence. This should be stated.

Response: Thank you for pointing this out. This has now been explicitly stated to improve clarity and avoid confusion. The revisions are as follows (Lines 414-417):

Third, we also compare the performance of regional models with country-specific models, which are all constructed using the equal-weight scenario (0.33, 0.33, 0.33), to identify the strengths and limitations of applying a generalized model across different region/income groups.

15. How did you calculate variation? Is it pixel-by-pixel and then aggregated to country, percent difference between the two models, or something else?

Response: Thank you for your comments. The variation is calculated by aggregating the grid-level predictions to the national level and computing the percentage differences between the Basic Service (or UN-Habitat) and Equal Weight scenarios. To improve clarity, we have replaced "variation increment" with "percentage change" and clarified the calculation in the Methods (Line409-412):

We then calculate the percentage change in slum population estimates by aggregating the grid-level predictions to the country level, and computing the percentage difference between each alternative scheme and the equal-weight scenario.

16. Section 5.3 – A table with a list of assessment cities/regions, the number of

estimates for that city/region, and citations would be useful. This is in regards to my comment below for Figure 7. It would clarify how many estimates you have for Rio de Janeiro and why is it a range in 7a and a single value in 7b. Are these from different sources, etc.

Response: Thanks for your valuable comments. To clarify, we have now added a table (Table 6) listing all assessment cities/regions, the number of available estimates for each, and their corresponding citations (Lines 571-572).

Figure 7a presents scaled-up estimates derived from individual slum-level sample extrapolations reported in various studies, whereas Figure 7b shows city-level statistics directly obtained from the literature. This difference explains why Figure 7a displays multiple values (a range) for each city, while Figure 7b shows only a single estimate per city. A more detailed explanation is provided in our response to comment on Figure 7 below.

Table 6 Comparison of model-derived slum population estimates with scaled-up figures from sampled slums and published city-level statistics, circa 2015.

City	Our estimate (million)	Scaled-up estimates (million, from sampled slums in other literature)	City-level statistics in other literature (million)	References
Rio de Janeiro	3.29	0.96, 1.17, 1.4, 1.46, 1.54, 1.58, 1.91, 2.39, 3.98	2.00	Breuer et al. (2024); Trindade et al. (2021);
São Paulo	2.09	1.3, 1.41, 3.02, 5.67, 8.62	2.16	Breuer et al. (2024); Trindade et al. (2021)
Cape Town	1.45	1.08, 1.13, 1.27, 1.29, 1.38, 1.54, 1.55, 1.89, 2.05, 2.15, 2.2, 2.8	1.23	Breuer et al. (2024)
Cairo	2.56	0.01, 0.3, 0.71, 3.75, 4.55, 7.47	/	Breuer et al. (2024); Sabry (2009);
Mumbai	2.70	1.65, 1.68, 5.2	/	Breuer et al. (2024); Taubenböck and Wurm (2015)
Dhaka	1.85	0.64, 0.76, 1.87, 2.69, 2.73, 3.43, 3.83, 3.84, 4.81, 4.84, 6.61	1.32	Breuer et al. (2024); Patel et al. (2019)
Johannesburg	0.80	/	0.6	Trindade et al. (2021)
Hyderabad	2.28	/	2.3	Trindade et al. (2021)

Tables

17. Table 1 – Nice

18. Table 2 – Nice

Response: Thank you for your encouraging feedback on Tables 1 and 2.

19. Table 3 – Consider adding more information to title (e.g., to optimize an XGBoost model, etc.). What are best_num_boost_rounds and Param_grid? These descriptions should either be expanded or explained in the table caption.

Response: Thanks for your comments. We have expanded the title of Table3 and added descriptions of best_num_boost_rounds and Param_grid in the table caption. The revised version (Lines 348-351) reads:

Table3 Hyperparameters used in the grid search for optimizing the XGBoost model. Param_grid indicates the ranges of values explored for each hyperparameter. Best_num_boost_rounds denotes the number of boosting iterations that achieved the highest validation performance.

20. Table 4 – Make each line a unique Region/Income Group (e.g., East Africa – Low Income, West Africa – Low Income). You can leave how you defined these in Supplemental, but this layout makes it appear as though you made 2 models for one combination (Africa/Low income) as opposed to 1 model for 2 different regions.

Response: Thanks for your valuable suggestion. We have updated Table 4 accordingly to make each line a unique Region/Income Group (Line 438).

Table 4 Model performance of slum indicator in Global South.

Region/Income Group	Model number	RMSE (%)	R^2
West Africa – Low Income	Model 1	5.29	0.93
East Africa – Low Income	Model 2	5.88	0.89
West Africa – Lower middle income	Model 3	6.40	0.92
East Africa – Lower middle income	Model 4	6.00	0.95
Africa – Upper middle income	Model 5	8.92	0.83
West Asia – Lower middle income	Model 6	5.20	0.84
East Asia – Lower middle income	Model 7	5.87	0.89
Latin American – Lower middle and low income	Model 8	6.36	0.89
Latin American – High and upper middle income	Model 9	10.17	0.82

Figures

21. Fig 1 – It is difficult to follow the progression of satellite imagery through the ResNet Algorithm, into the XGBoost algorithm, and ultimately to the Slum population

map. It is also difficult to tell if the slum indicator is the response variable for the ResNet algorithm, the XGBoost Algorithm, or both. Consider reorganizing to make this clearer. It looks like only 80% of the landsat images go to the ResNet Algorithm. Is that accurate? The "Results Analysis" panel is not very informative. It could likely be removed or consolidated to a single box to make room for clearer descriptions in the other sections.

Responses: Thank you for your valuable comments. We have updated Figure 1 to more clearly illustrate the workflow progression. Satellite imagery and other inputs for 53 countries are first processed through the ResNet model for feature extraction. The 80%/20% split refers specifically to the division of data into training and testing sets for the XGBoost model. In addition, the "Results analysis" panel has been removed and replaced with a "Robustness analysis" panel. The revised Figure 1 (Lines 218-221) is shown below:

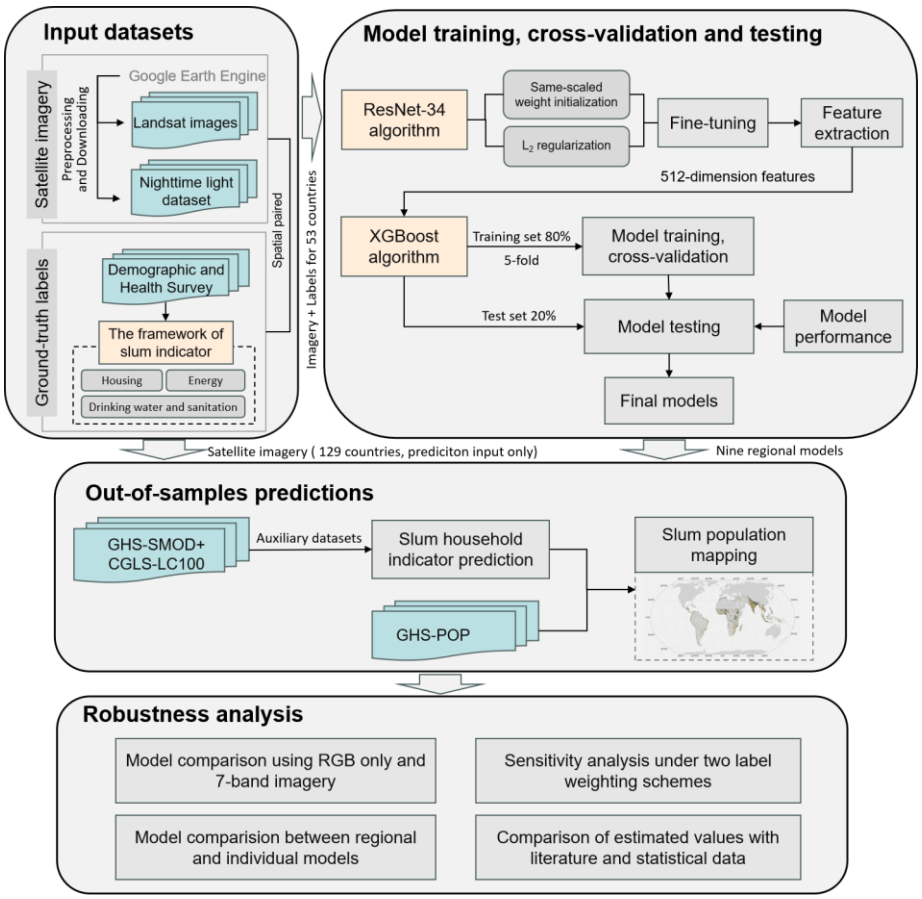

Figure 1 Workflow of this study. Blue backgrounds represent datasets, yellow backgrounds indicate the indicator framework or model architecture, and gray backgrounds denote specific processes or procedural steps.

22. Fig 2 – I interpret the "weight copy" arrows as using all the weights from ResNet,

but you only used those for RGB. The others were averaged from the RGB (Line 256). Then, are they all scaled by 3/7 or only the non-RGB values? Since there are only 7 layers it might be best to expand the ellipses and describe each layer along with the weights used. You introduce a "Global Average Pooling Layer" in the figure that is not described or explained in the text. Where are the 512-dimensional features in this figure or figure 1? All "trees" are identical and have the same label (Tree1{..}). Is this intentional? It would be helpful to have the response variable (Slum indicator for 67,204 clusters) documented in this figure or figure 1. Clearer descriptions of your workflow and the input/output of each step would be more useful than the bottom portion of the figure illustrating XGBoost decision tree splits. This is documented elsewhere.

Responses: Thank you for your valuable comments. (1) Clarification of pretrained weights: We have revised Figure 2 to indicate that only the RGB channels use pretrained weights from ResNet, while the non-RGB channels are initialized with the average of the RGB weights. All weights are then scaled by 3/7, as described in the text (Lines 276-277). The ellipses have been expanded to display all 7 input channels, and the annotations have been updated accordingly.

(2) Feature extraction updates: The Global Average Pooling layer has been removed, and the 512-dimensional feature vector is now explicitly labeled in the figure. The response variable (slum indicator for 67,204 clusters) has also added to improve clarity.

(3) Corrections and enhanced labeling: The labeling error in tree identifiers have been corrected. We have also provided clearer descriptions of the workflow and explicitly indicated the input and output at each step.

The revised Figure 2 is shown as follows (Lines 304-307):

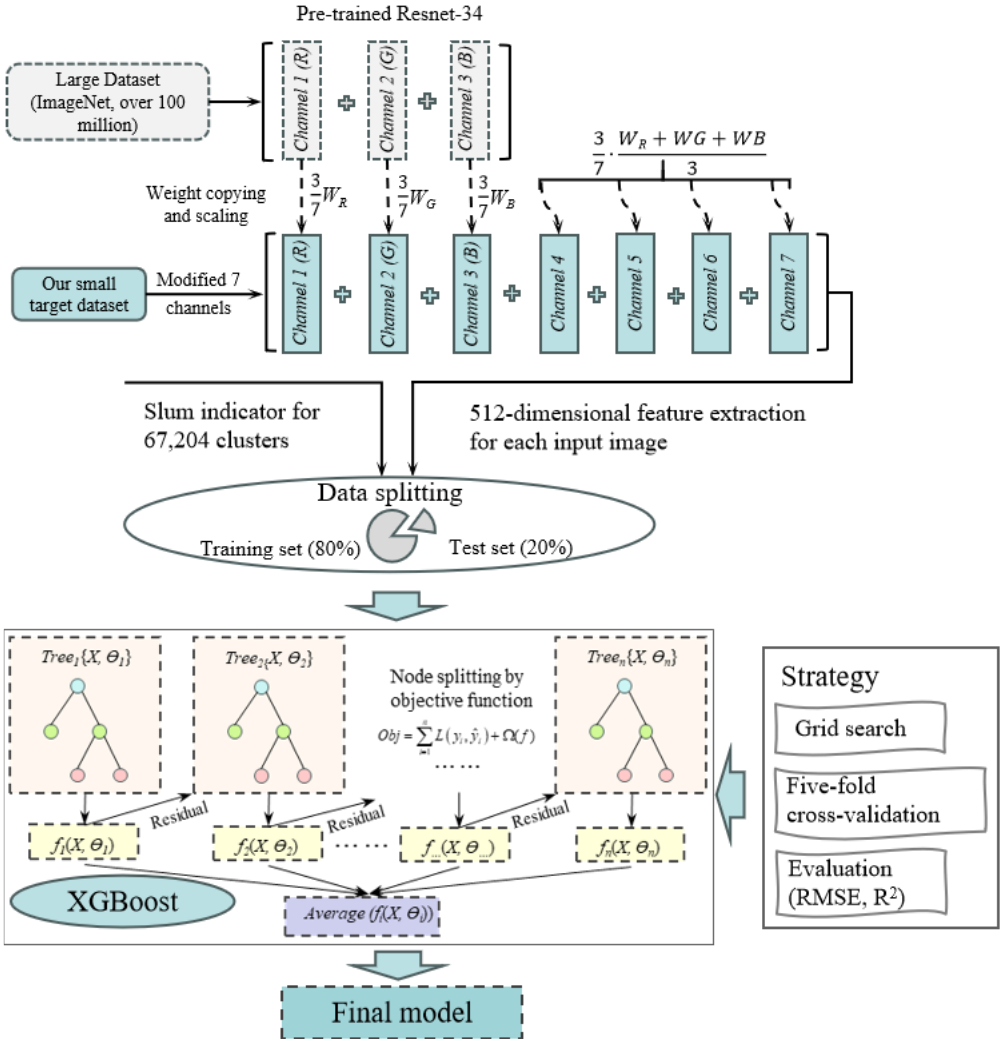

Figure 2 Schematic of the proposed model architecture. It integrates transfer learning by fine-tuning ResNet-34 to extract high-dimensional spatial–spectral features from multispectral imagery, followed by XGBoost regression for predictive modeling.

23. Fig 3 – Check title. These models are across geographical regions and income groups (line 304-5) not income only.

Responses: Thank you for pointing this out. We have revised this title to "Figure 3 Model performances across different geographical regions and income groups".

24. Fig 4a - Nice

Response: Thank you for your positive feedback.

25. Fig 4b – Nice. The bar charts may be more informative as a stacked bar chart like Fig 5. I concede that incorporating the different scale of Europe and Central Asia may preclude this suggestion, so I leave it the authors discretion.

Response: Thank you for the helpful suggestion. We initially considered using a stacked bar chart similar to Fig. 5. However, due to the relatively small values for Europe and Central Asia compared to other geographic regions, it was difficult to display all regions clearly on the same axis. To ensure visual clarity and preserve comparability, we chose to retain the separated bar format in Fig. 4b.

26. Fig. 6 – It is unusual to introduce a figure like this in discussion. These belong in results. Figure title should be more objective and avoid the authors interpretation.

Response: Thank you for raising this important concern regarding the placement and objectivity of Figure 6. We have addressed this by reorganizing the manuscript: Figure 6 and its associated result descriptions have been moved to the Results section (Section 4.4) (Line511-552), while the relevant interpretation is retained in the Discussion (Section 5.1) (Line617-624). In addition, we have revised the title and caption of Figure 6 to ensure a more objective and descriptive presentation. The updated results section 4.4 (Lines 512-552) reflects these changes. The updated Figure 6 caption is as follows:

Figure 6 Differences in slum population estimates across two alternative weighting schemes and comparison of model performance between regional models and country-specific models. Panels (a) and (b) show the percentage change in slum population estimates under the Basic Services scenario and the UN-Habitat scenario, respectively. Panels (c) and (d) compare model performance between regional and country-specific individual models using RMSE and R^2 metrics, respectively.

27. Fig 7 – Same as Fig 6 re: section. The caption needs more explanation and titles for each subplot.

Response: Thank you again for highlighting this important point regarding the placement and more detailed explanation. We have reorganized this section accordingly by moving Figure 7 and its associated result descriptions to the Results section (Section 4.5) (Line555-585), while the relevant interpretation has been retained in the Discussion (Section 5.1) (Line625-655). In addition, we have also expanded the explanation of each subplot. More details are presented below in response to your specific comments:

7(a) The asterisks should be explained in the caption. What are the n for each boxplot? Are (a) and (b) related; are the "Other literature" values in (b) the averages of (a) and if so why are some cities in one but not the other? If not, its confusing to have some cities (e.g., Rio de Janeiro) as a range in (a) and single value in (b). More caption information would help clarify this.

Response: We apologize for the confusion caused by the lack of details in the original Figure 7 caption and appreciate your valuable comments. Panels (a) and (b) present mainly external estimates derived using different methods and are therefore not directly related. Panel (a) shows scaled-up estimates based on data from multiple individual slum areas, resulting in a range of values for each city. Panel (b), by contrast, presents single-value estimates obtained from official statistics or literature-reported city-level totals. This explains why some cities (e.g., Rio de Janeiro) appear with a range of values in panel (a) but with a single value in panel (b).

We have revised the caption to clarify this distinction and added explanations for the asterisks in panel (a) (representing our estimates), and the sample size (*n*) used in each boxplot.

7(c) – What are each dot? A country, select cities?

Response: In Figure 7c, each dot represents a country in the Global South. This has been explicitly stated in the revised caption.

7(d) – Should the legend entry for Other literature be UN-Habitat? Should UN-Habitat be in the label for the x-axis?

Recommend hiding the top and right spines in (a) and (c) to match the other two plots.

Response: Following your suggestion, we have updated the legend entry to "UN-Habitat" and removed "UN-Habitat" from the x-axis label, which now reads "Shares of population living in slums." We have also removed the top and right spines in panels (a) and (c) to match the styling of the other plots.

The updated Figure 7 and its revised caption are as follows:

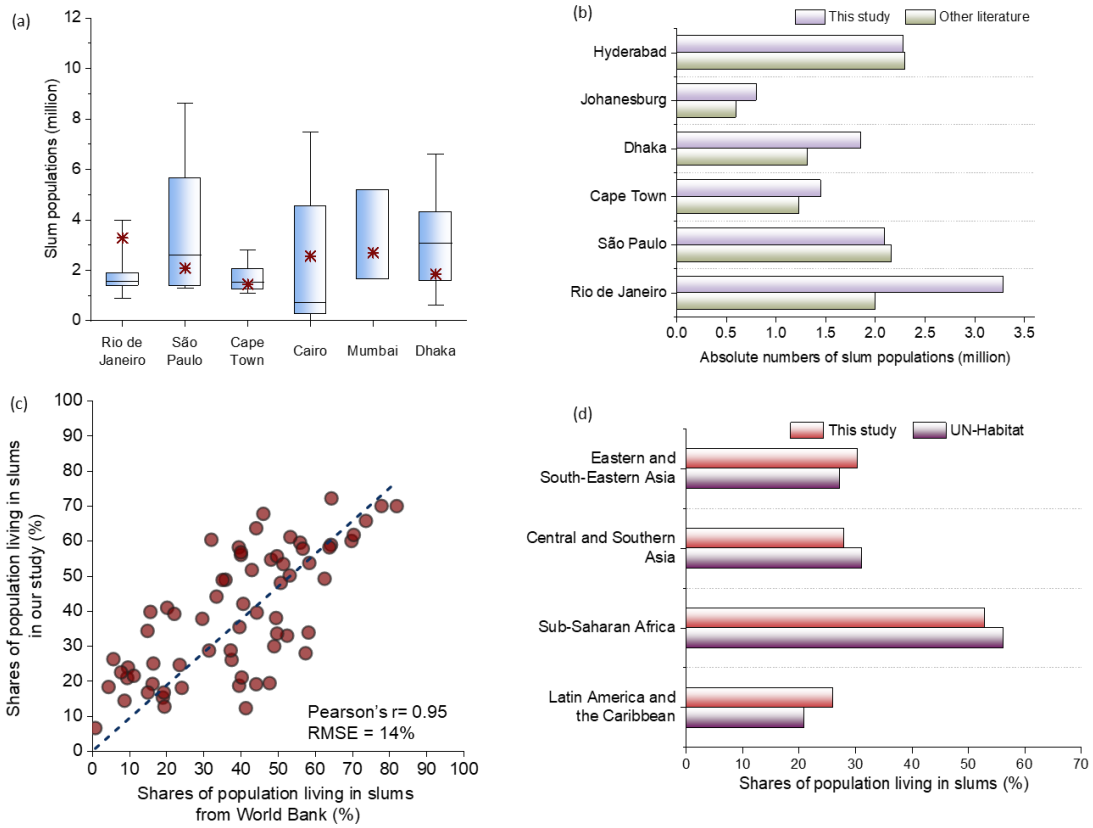

Figure 7 Comparison of our slum population estimates with data from previous literature and official statistics, presented as both absolute numbers and relative shares. (a) Comparison with scaled-up values from previous studies for six cities: Rio de Janeiro (n=9), São Paulo (n=5), Cape Town (n=12), Cairo (n=6), Mumbai (n=3), and Dhaka (n=11). Scaled-up values are derived by extrapolating slum population counts from multiple sampled slums to the city level. Asterisks indicate estimates from this study. Box plots display the median (central line), interquartile range (box), and minimum-maximum range (whiskers). (b) Comparison with official city-level statistics for the same six cities. (c) Comparison of country-level slum population shares between our estimates and World Bank data; each dot represents a Global South country. (d) Comparison of regional-level slum population shares between our estimates and UN-Habitat data.

Abstract

28. Line 32 -… *outperforming previous benchmarks*. I don't recall mention or reference to previous benchmarks in this paper.

Response: Thank you for pointing this out. We have removed this statement.

Supplement

29. Figure S1 – This is a slightly difficult map to interpret. I would suggest emphasizing the 129 countries in the Global South (or de-emphasizing the non-129), and perhaps a gradient based on the number of surveys in each country. This would highlight the countries with more, less, and no training data. However, I leave this to the authors discretion.

Response: Thank you for your thoughtful comment and valuable suggestion. In response, we have updated Figure S1 to highlight the 129 countries in the Global South, applied a gradient color scheme reflecting the number of DHS survey clusters in each country, and added a hatched pattern to indicate countries without available DHS data. The updated figure S1 is shown below.

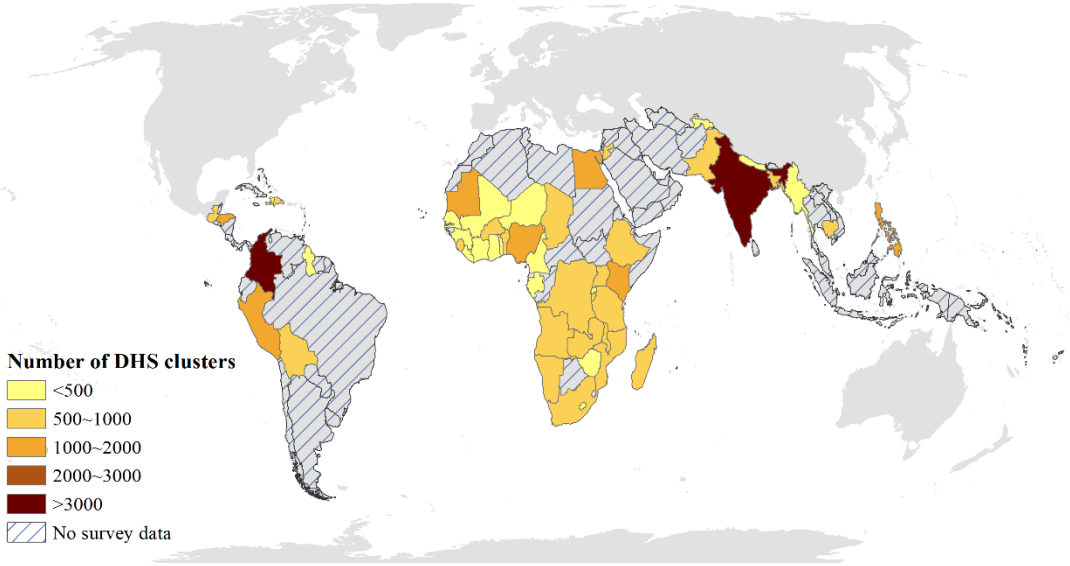

Figure S1 Number of clusters from Demographic and Health Surveys in 53 Global South countries. The highest legend category (>3,000 clusters) includes only Colombia (n = 4,868) and India (n = 30,170).

Minor Comments

30. 142 – deprive? Should this be derive?

Response: Sorry for the typo. Correction has been made (Line 145).

31. 158 – 161. Seems to be missing 1 or more citations.

Response: Thank you for noting this. We have added supporting references to these statements. The revised sentence (lines 161- 164) now reads:

The Landsat imagery series provides the world's longest-running collection of consistently acquired, high-resolution Earth observation data and have been widely

applied in studies such as urban growth monitoring (Patel et al., 2015; Wulder et al., 2022).

References:

Patel, N. N., Angiuli, E., Gamba, P., Gaughan, A., Lisini, G., Stevens, F. R., Tatem, A.J., Trianni, G. (2015). Multitemporal settlement and population mapping from Landsat using Google Earth Engine. International Journal of Applied Earth Observation and Geoinformation 35, 199–208. https://doi.org/10.1016/j.jag.2014.09.005.

Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C. J., Masek, J. G., Hermosilla, T., White, J. C., Belward, A. S., Schaaf, C., Woodcock, C. E., Huntington, J. L., Lymburner, L., Hostert, P., Gao, F., Lyapustin, A., Pekel, J.-F., Strobl, P., Cook, B. D. (2022). Fifty years of Landsat science and impacts. Remote Sensing of Environment 280, 113195. https://doi.org/10.1016/j.rse.2022.113195

32. 172 – 174. Why is it particularly suitable for long-term trends? Either provide citations of its use or explain.

Response: Thank you for your comment. We have clarified the reasons and added the relevant citation. The revised sentence (Lines 175-179) now reads:

This NPP-VIIRS-like dataset is particularly suitable for analyzing long-term demographic and socioeconomic trends, as its cross-sensor calibration between DMSP-OLS (2000 – 2012) and NPP-VIIRS (2013 – 2018) removes inter-sensor bias and drift, yielding a radiometrically consistent time series (Chen et al., 2021).

References:

Chen, Z., Yu, B., Yang, C., Zhou, Y., Qian, X., Wang, C., Wu, B. and Wu, J. (2021). An extended time-series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration. *Earth System Science Data Discussions* 13, 889-906.

33. 249 – what are fine-grained mapping tasks?

Response: The term “fine-grained” refers to high-resolution mapping tasks that capture detailed spatial features beyond those represented in coarse-scale mapping. This terminology is widely used in the literature—particularly in fields such as urban mapping and population disaggregation, especially when combined with machine learning methods (e.g., Du et al., 2021; Yang et al., 2018).

References:

Du, S., Du, S., Liu, B. and Zhang, X., 2021. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network

and object based approach. *Remote Sensing of Environment*, *261*, p.112480.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J. and Wang, L., 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European conference on computer vision (ECCV)* ,420-435.

34. 280 - saying XGBoost is renowned needs some citations.

Response: Thanks for noting this. We have added two citations, as follows (Lines 300-303):

XGBoost is renowned for its high computational efficiency and strong predictive performance, making it particularly suitable for large datasets and high-dimensional feature spaces (Nielsen, 2016; Ramraj et al., 2016).

References:

Nielsen, D. Tree boosting with XGBoost: Why does XGBoost win "every" machine learning competition? Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway. http://hdl.handle.net/11250/2433761,2016.

Ramraj, S., Uzir, N., Sunil, R., and Banerjee, S. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications 9, 651–662, 2016.

35. 305 – A brief explanation of how you set/acquired the income levels for each country would be useful. A map of the 9 geographic/income regions would be helpful but is entirely optional.

Response: Thank you for your helpful comments. The classification of countries by income level follows the World Bank income group definitions, and we have added a brief explanation with the appropriate references. Since the nine geographic/income regions are already clearly defined in the supplementary material (Table S3), we did not include an additional map. The revised sentence (Lines 312-313) now reads:

The classification of countries by income level follows the definitions provided by the World Bank (World bank, 2022).

References:

World Bank. World Bank country and lending groups https://datahelpdesk.worldbank.org/knowledgebase/articles/906519, 2022.

36. 328 – The 129 counties are from UN Finance Center? This should be introduced on first mention of the 129 counties. Is there a citation?

Response: Thank you for your helpful comments. We have clarified the source of the

129 countries at their first mention in the revised manuscript and added the appropriate citation. The revised sentence (Lines 110-112) now reads:

The 129 countries in the Global South are identified according to the list provided by the United Nations Conference on Trade and Development (UNCTAD, 2025).

References:

United Nations Conference on Trade and Development. Countries, All Groups Hierarchy,UNCTADstat.https://unctadstat.unctad.org/EN/Classifications/DimCountries_All_Hierarchy.pdf, 2025.

37. 673 – 675. RMSE values ranging from …*when compared to our slum indicators derived from DHS surveys?* Consider expanding on the text a bit to make the conclusion more impactful.

Response: Thank you for your helpful comments. Following your suggestion, we have slightly expanded this part of the text to make the conclusion more impactful. The revised sentence (Lines 775-777) now reads:

The models exhibit strong performance and robustness, achieving RMSE values of 5.20%-10.17% and R^2 values of 0.82-0.95 when validated against slum indicators derived from DHS surveys.

Anonymous Referee #3

This study proposes a standardized and scalable bottom-up approach for estimating the high-resolution spatial distribution of slum populations in Global South countries. This approach integrates household survey data, satellite imagery, and gridded population data, and employs machine learning and transfer learning techniques (e.g., ResNet-34 and XGBoost). It generates slum population distribution maps with a spatial resolution of approximately 6.72 km (at the equator) across 129 Global South countries. However, the current manuscript is insufficient in its discussion of the accuracy and reusability of this geospatial data product, and there are the following issues that need clarification:

Response: We sincerely appreciate your constructive comments regarding the accuracy and reusability of our data product. In response, we have expanded the discussion to address the spatial accuracy, visualization validation and application scenarios. Please see our point-by-point responses below.

1. What standards are used to define slums? When conducting mapping for Global South countries, how is the consistency of slum feature extraction ensured across different regions and data sources to guarantee the accuracy of the product?

Response: In this study, we define the slum population following UN-Habitat's operational definition of a slum household (UN-Habitat, 2021): a household lacking one or more of the following: (i) access to improved water sources, (ii) access to improved sanitation facilities, (iii) sufficient living area, (iv) durable housing, or (v) security of tenure. Because the last criterion is particularly difficult to assess (UN-Habitat, 2003), we exclude security of tenure from our final definition. These details are clarified in Methods (Section 3.1) and the indicator framework is summarized in Table 2.

We recognize that slum characteristics vary across countries and regions. To ensure consistent feature extraction across the Global South, we adopt the following measures:

1) Standardized definition and indicators: a widely recognized, operational framework from UN-Habitat.
2) Transfer learning with enriched inputs: ResNet pretraining on >100 million ImageNet images to capture broadly representative visual features, combined with expanding inputs from 3 RGB bands to 7 bands (including near-infrared, etc.) to enhance spatial–spectral richness.
3) Regional stratification for generalization: models stratified by geographic region and national income group to improve transferability and reduce

domain shift.

4) Robust data foundations: publicly available Landsat imagery and comprehensive, representative DHS household data for ground truth.

Together, these choices promote consistent and accurate extraction of slum-related features across diverse settings and highlight the scalability advantages of our framework for large-area applications. We have added a dedicated discussion on consistency in slum feature extraction in the revised version as follows.

Discussions

…

Slum characteristics vary across countries and regions (da Fonseca Feitosa et al., 2021; do Nascimento et al., 2025), making consistent feature extraction at large spatial scales challenging. We address this systematically across the workflow, from data selection to model design and feature engineering. At the core we adopt transfer learning with Resnet-34 pre-trained on over 100 million ImageNet samples and adapt it to 7-channel multispectral patterns associated with deprived housing conditions. While the original RGB-pretrained ResNet-34 is optimized for general features (e.g., edges and textures), fine-tuning enables early and intermediate filters to learn context-specific cues, such as roof material, settlement density, and vegetation coverage, while leveraging the added information from near-infrared and other bands. The combination of broadly learned representations and increased spectral richness provides a robust foundation for identifying spatial and morphological signatures of slum environments. The approach outperforms RGB-only baseline, highlighting the value of multispectral fine-tuning and high-dimensional feature extraction for modeling housing-related deprivation.

During the model development phase, we constructed regionally stratified models based on geographic location and national income level, rather than separate country-specific models, to improve generalizability and ensure consistent feature extraction across heterogeneous contexts. In addition, the use of publicly available Landsat imagery and DHS ground-truth data provides a uniform, reproducible basis for cross-country analysis. Collectively, these design choices enable consistent and accurate slum feature extraction at scale and underpin the production of a harmonized slum population product.

…

References:

Da Fonseca Feitosa, F., Vasconcelos, V. V., de Pinho, C. M. D., da Silva, G. F. G., da Silva

Gonçalves, G., Danna, L. C. C., & Lisboa, F. S. (2021). IMMerSe: An integrated methodology for mapping and classifying precarious settlements. Applied geography, 133, 102494.

Do Nascimento, G.A., Giannotti, M., Regueira, T.A. and Tomasiello, D.B., 2025. Identifying slum areas: A multidimensional analysis leveraging with explanatory machine learning techniques. Sustainable Cities and Society, 131, 106645.

UN-habitat. Urban indicators database. https://data.unhabitat.org/pages/housing-slums-and-informal-settlements, 2021

UN-Habitat. The challenge of slums: global report on human settlements 2003. Management of Environmental Quality: An International Journal 15, 337-338, https://doi.org/10.1108/meq.2004.15.3.337.3 , 2004.

2. Visualization is an important basis for verifying product accuracy. Can the authors provide visualized comparison results of selected regional cases? By comparing with known data or real-world conditions, further verification of the product's accuracy could be achieved.

Response: We appreciate your valuable suggestion to include visual comparisons. Following your advice, we have added comparisons between our slum population estimates and observed conditions for four representative Global South cities, including (a) Nairobi, Kenya; (b) Cape Town, South Africa; (c) Medellín, Colombia; and (d) Dhaka, Bangladesh. We overlay the gridded slum population estimates on high-resolution satellite imagery and Google Earth imagery. The resulting maps show consistent spatial correspondence between areas with high estimated slum populations and known locations of dense slums areas or informal settlements, supporting the reliability of our mapping results. These visualizations are included in the revised main text as Figure 6 (pasted below for the reviewing convenience).

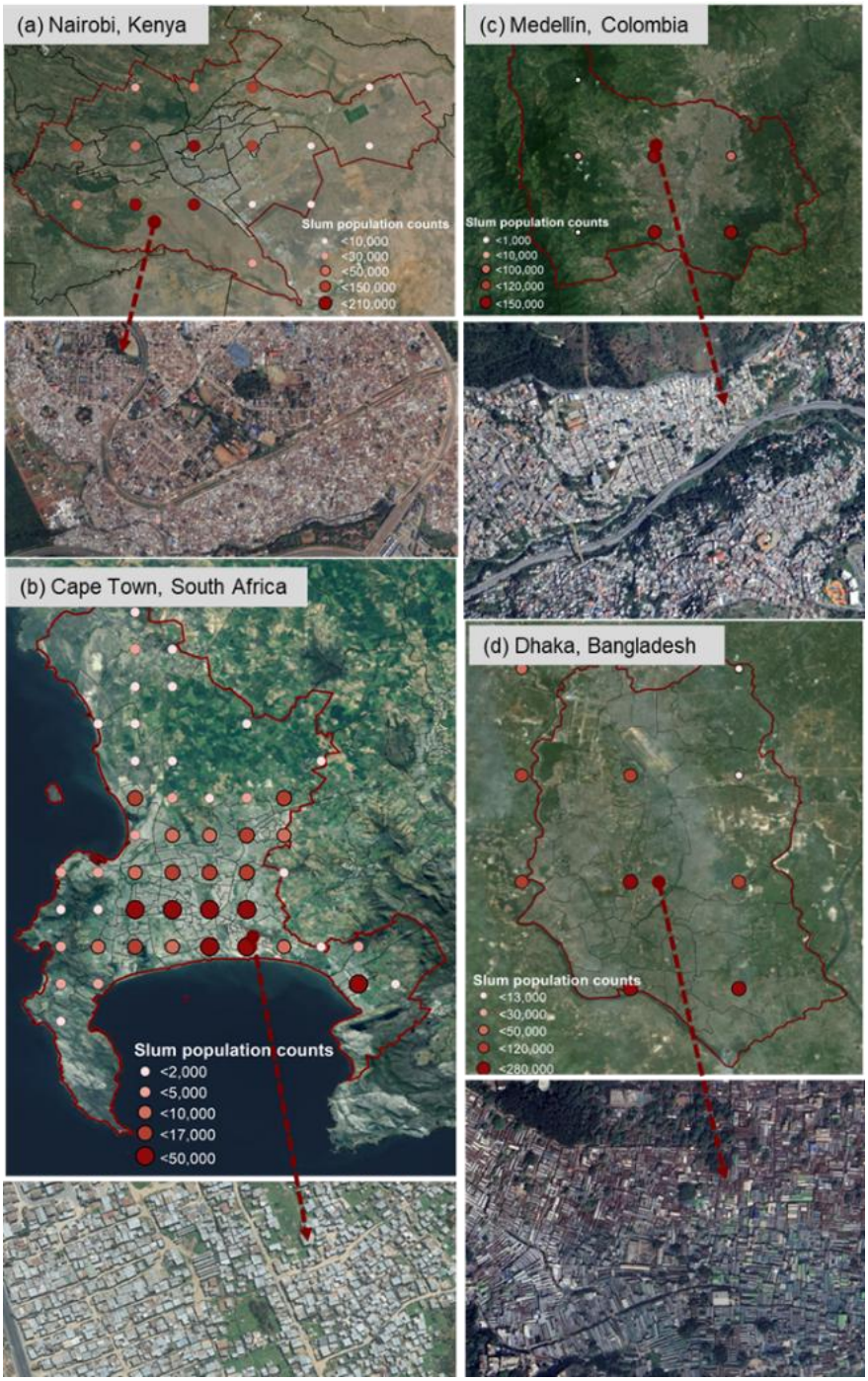

Figure 6. Showcases of visual comparisons between estimated slum population distribution and conditions observed in Google Earth imagery for four representative Global South Cities: (a) Nairobi, Kenya, (b) Cape Town, South Africa, (c) Medellín, Colombia, and (d) Dhaka, Bangladesh. Estimated population counts are shown as circular markers overlaid on Landsat basemaps, with each marker representing a 6.72 km × 6.72 km grid cell. Red arrows indicate locations where high estimated slum populations align with visible informal settlements in Google Earth, providing visual validation of the product.

3. Compared with traditional slum population estimation methods and existing relevant data, what advantages does the current product have in terms of spatial continuity? How to demonstrate these advantages from both theoretical and practical application perspectives?

Response: Thanks for your valuable comments. We agree that the advantage of spatial continuity better highlights the strengths of our product. Traditional approaches to estimating slum populations typically rely on (1) census or household survey aggregation or (2) the overlap of morphological slum boundaries with gridded population maps. While census and survey data provide official, well-validated estimates, they are often available only at coarse administrative levels or as sample-based estimates, limiting spatial detail and coverage. Likewise, combining mapped slum boundaries with population grids can pinpoint settlement locations but is constrained by the discrete nature of those boundaries and the uneven availability of high-quality maps—largely limited to a few city hotspots (e.g., Cape Town, Mumbai, Dhaka).

By contrast, our framework produces spatial continuous estimates, meaning that slum population is estimated for every 6.72 km grid cell covering inhabited areas. This does not imply continuously designated slum locations; rather, it yields a wall-to-wall grid that (i) captures local variation and spatial gradients, overcoming the aggregation effects of administrative units and survey areas; (ii) does not depend on the existence of pre-mapped slum boundaries, allowing coverage in regions lacking such data; and (iii) can be dynamically updated as new imagery and ancillary data become available. In practice, this enables flexible aggregation across multiple spatial scales (neighborhood, city, national), supports integration with research on urban inequality, sustainability, environmental exposure, and spatial justice, and provides a scalable, data-driven foundation for human-centered urban planning, targeted resource allocation, and evidence-based policy design.

In line with response to your comment #7, we have added discussion of these theoretical and practical benefits and outlined specific application scenarios in Discussion (pasted below for the reviewing convenience).

Discussions

The product advances a human-centered mapping framework, prioritizing populations living in inadequate housing rather than focusing solely on the physical delineation of slum settlements. Theoretically, it exposes gaps in progress toward the Sustainable Development Goals by providing spatially continuous, gridded estimates that directly quantify the number of people affected by deficits in water, sanitation,

and electricity. These wall-to-wall estimates reveal local variation and spatial gradients that are often obscured by administrative boundaries or survey aggregates. Because the approach does not depend on pre-defined morphological slum boundaries, it supports population estimation in regions lacking boundary data and can be dynamically updated as new datasets become available.

Practically, this continuous representation enables flexible aggregation across multiple spatial scales (e.g., neighborhood, city, national) and provides a robust, data-driven foundation for human-centered urban planning, targeted resource allocation, and evidence-based policy design. For example, planners can optimize the spatial distribution of schools, clinics, and parks to improve accessibility and equity for vulnerable populations, thereby supporting inclusive urban development. These estimates also inform resource allocation and emergency response planning aligned with actual population needs. When combined with spatial information on environmental hazards, such as floods, heat islands, or air pollution, the gridded product becomes a critical tool for resilience planning and adaptive governance, enabling social protection and welfare programs tailored to those at greatest risk.

4. This study mentions the issue of underestimation of the actual slum population size. What strategies does the current product adopt to solve this problem? To what extent can it improve the underestimation? Can both quantitative verification (e.g., the proportion of error reduction, the degree of proximity between estimated values and actual values) and qualitative verification (e.g., the rationality of the method, the completeness of data coverage) be provided?

Response: Thank you for raising this concern about strategies to address the underestimation. Many prior approaches estimate slum populations by overlapping mapped (morphological) slum boundaries with population grids at the settlement level. Recent studies show these methods can underestimate slum populations relative to literature or official statistics, largely due to inaccuracies in slum geometries and uncertainties in the underlying population datasets.

1) What strategies does the current product adopt to solve this problem?

Response: We take a direct estimation approach by predicting the proportion of slum households within each enumeration area, thereby reducing errors introduced by post-hoc scaling. Our strategy has three components: (i) Modelling design: we define slums from a household perspective rather than relying on binary morphological maps (which are unavailable at scale). This household-based framework provides a more direct basis for estimating slum populations. (ii) Ground-truth labeling: we use DHS surveys for household labels. DHS employs standardized instruments and rigorous

quality controls, providing reliable and valid measures of slum households. (iii) Population dataset selection: we adopt GHS-POP, which refines population estimates along the urban gradient using human settlement information and share a production framework with other datasets used in this study, thereby reducing uncertainties from heterogeneous sources. These points are elaborated in the revised Discussion to clarify how our method mitigates underestimation, as follows:

Discussions

…

Previous studies have addressed slum population underestimation by scaling single-slum estimates to match literature-reported values and then applying these ratios to adjust aggregated population totals (Breuer et al.,2024; Thomson et al., 2021). In contrast, we adopt a direct estimation approach that predicts the proportion of slum households within each enumeration area. First, we operationalize slums from a household perspective by constructing a household-level indicator, expressed as the share of slum households among all households in an enumeration area, and use deep learning to link this indicator directly to features extracted from remote sensing imagery. This design avoids reliance on dichotomous morphological slum maps (which are rarely unavailable at scale) and provides a more flexible, generalizable framework. Second, we use high-quality DHS survey data as labels; their standardized instruments and rigorous quality-control protocols yield valid household-level indicators essential for accurate prediction. Finally, for population distribution, we employ GHS-POP, which refines population estimates along the urban gradient and share a production framework with the GHSL datasets used in this study, thereby reducing uncertainties arising from heterogeneous sources.

2) To what extent can it improve the underestimation? Can both quantitative verification (e.g., the proportion of error reduction, the degree of proximity between estimated values and actual values) and qualitative verification (e.g., the rationality of the method, the completeness of data coverage) be provided?

Response: Our results include both quantitative and qualitative verification.

(A) quantitative verification: We validate the predicted slum household indicators against DHS-derived ground-truth indicators using RMSE and R^2. We also compare our slum population estimates with: (i) scaled-up city-level estimates derived from multiple individual slum samples in the literature, (ii) official city statistics, (iii) country-level statistics reported by World Bank, and (iv) regional-level statistics reported by UN-Habitat. Across these benchmarks, our estimates fall within the range of literature-based scaled-up values and show strong alignment at city, country, and regional scales

(Figure 3, Table 6 and Figure 7 in the revised version, which are presented below for review convenience).

(B) qualitative verification. We discuss the methodological soundness and data coverage of our framework, including the rationale for key dataset choices—DHS surveys (for standardized, quality-controlled household labels), Landsat imagery (for globally consistent, multispectral inputs), and GHS-POP (for human-settlement–informed population distribution). We also provide robustness analyses that explain how these selections support consistent performance and broad geographic applicability.

Results

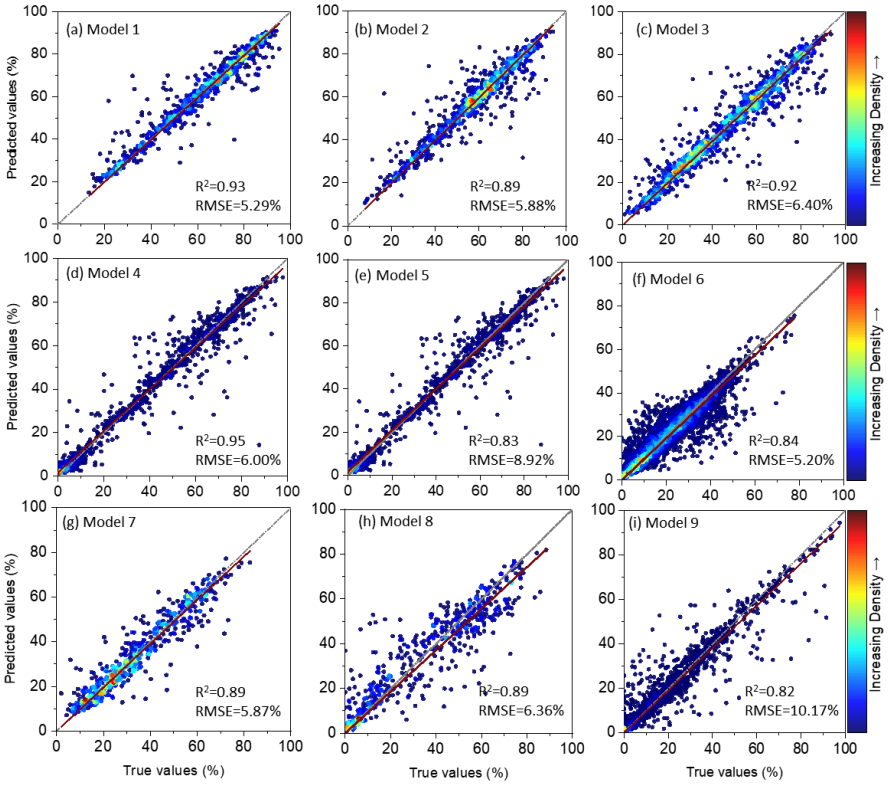

Figure 3 Model performances across different geographical regions and income groups. The true values represent the slum indicator calculated from DHS data using our slum framework, while the predicted values are produced by our machine learning-based models. For each model, a linear regression line is displayed alongside a reference line with a slope of one for comparison.

Table 6 Comparison of model-derived slum population estimates with scaled-up figures from sampled slums and published city-level statistics, circa 2015.

City	Our estimate (million)	Scaled-up estimates (million, from sampled slums in other literature)	City-level statistics in other literature (million)	References
Rio de Janeiro	3.29	0.96, 1.17, 1.4, 1.46, 1.54, 1.58, 1.91, 2.39, 3.98	2.00	Breuer et al. (2024); Trindade et al. (2021);
São Paulo	2.09	1.3, 1.41, 3.02, 5.67, 8.62	2.16	Breuer et al. (2024); Trindade et al. (2021)
Cape Town	1.45	1.08, 1.13, 1.27, 1.29, 1.38, 1.54, 1.55, 1.89, 2.05, 2.15, 2.2, 2.8	1.23	Breuer et al. (2024)
Cairo	2.56	0.01, 0.3, 0.71, 3.75, 4.55, 7.47	/	Breuer et al. (2024); Sabry (2009);
Mumbai	2.70	1.65, 1.68, 5.2	/	Breuer et al. (2024); Taubenböck and Wurm (2015)
Dhaka	1.85	0.64, 0.76, 1.87, 2.69, 2.73, 3.43, 3.83, 3.84, 4.81, 4.84, 6.61	1.32	Breuer et al. (2024); Patel et al. (2019)
Johannesburg	0.80	/	0.6	Trindade et al. (2021)
Hyderabad	2.28	/	2.3	Trindade et al. (2021)

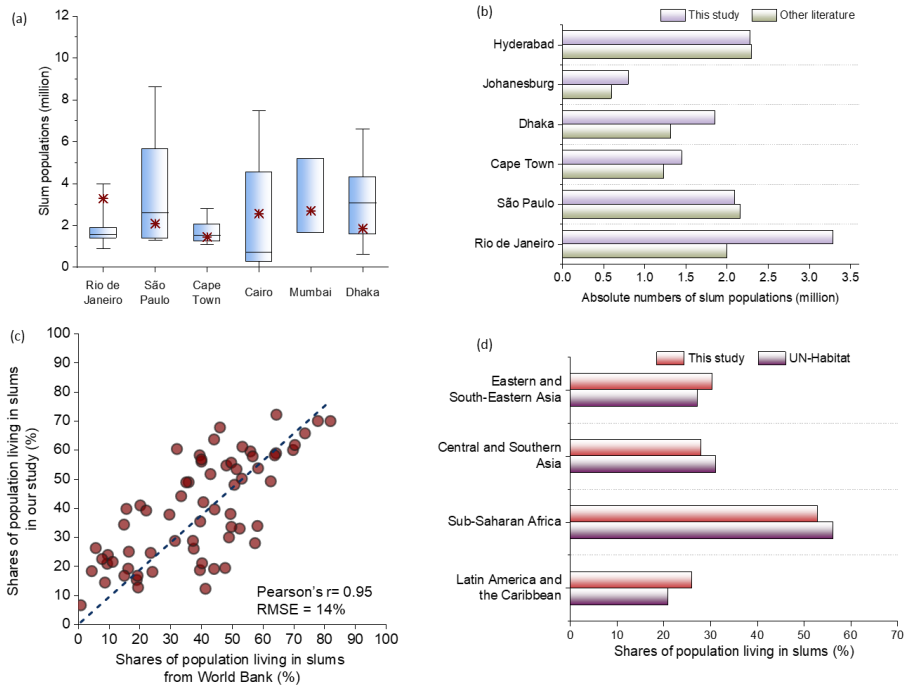

Figure 7 Comparison of our slum population estimates with data from previous literature and official statistics, presented as both absolute numbers and relative shares. (a) Comparison with scaled-up values from previous studies for six cities: Rio de Janeiro (n=9), São Paulo (n=5), Cape Town (n=12), Cairo (n=6), Mumbai (n=3), and Dhaka (n=11). Scaled-up values are derived by extrapolating slum population counts from multiple sampled slums to the city level. Asterisks indicate estimates from this study. Box plots display

the median (central line), interquartile range (box), and minimum-maximum range (whiskers). (b) Comparison with official city-level statistics for the same six cities. (c) Comparison of country-level slum population shares between our estimates and World Bank data; each dot represents a Global South country. (d) Comparison of regional-level slum population shares between our estimates and UN-Habitat data.

Discussions

5.1 Robustness and external validation

Our approach builds on well-established deep learning techniques widely used for mapping poverty, infrastructure access, and progress toward the Sustainable Development Goals (Jean et al., 2016; Chi et al.,2022). We fine-tune the ResNet-34 model on 7-band satellite imagery to better capture the complex spatial-spectral patterns associated with deprived housing conditions…. The original RGB-pretrained ResNet-34 is optimized for general features (e.g., edges and textures), fine-tuning enables early and intermediate filters to learn context-specific cues, such as roof material, settlement density, and vegetation coverage ...

5.3 Spatial resolution and dataset selection

… The performance of any machine learning or deep-learning model relies heavily on the quality and suitability of its trained data (Meena et al., 2023). The DHS datasets provide accurate, representative, household-based survey data on demographic and socioeconomic conditions; our samples cover over 1 million households across 67,204 clusters in 53 countries of the Global South.

… To enrich the input feature space, we prioritize images with more spectral bands than standard RGB. Landsat imagery offers consistent temporal coverage, global reach, and multiple spectral bands at 30-meter resolution, making it well suited to our task (Wulder et al., 2022)

… Grided population data are also critical for measuring and mapping slum populations. Widely used global products, such as GPWv4.11 (CIESIN et al., 2018), GHS-POP (Pesaresi et al., 2016), and World Pop (Stevens et al., 2015), as well as regional/national products like HRSL (Smith et al., 2019), differ markedly in inputs, ancillary datasets, and dasymetric methods for redistributing population to grid cells. While regional products such as HRSL can achieve high accuracy at finer resolutions, their limited spatial coverage impedes large-scale, cross-country comparisons, especially in the Global South. Following the 'fitness-for-use' principle (Juran et al., 1979), we select GHS-POP for its distinctive advantages aligned with our objectives. Because it is grounded in human-settlement information, GHS-POP effectively refines population estimates along the urban gradient, which is critical for slum population

modeling. Moreover, the Global Human Settlement Layer (GHSL), which leverages Landsat imagery, aligns with the satellite images used in this study, reducing uncertainties introduced by heterogeneous sources.

Reference:

Breuer, J.H., Friesen, J., Taubenböck, H., Wurm, M. and Pelz, P.F. (2024). The unseen population: Do we underestimate slum dwellers in cities of the Global South? Habitat International, 148, p.103056.

Thomson, D.R., Gaughan, A.E., Stevens, F.R., Yetman, G., Elias, P. and Chen, R. (2021). Evaluating the accuracy of gridded population estimates in slums: A case study in Nigeria and Kenya. Urban Science, 5(2), p.48.

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

Chi, G., Fang, H., Chatterjee, S., Blumenstock, J.E. (2022). Microestimates of wealth for all low-and middle-income countries. Proceedings of the National Academy of Sciences 119, e2113658119.

5. Does the data used by the authors to verify the product have the underestimation issue as mentioned? If it does, how to ensure the validity of accuracy verification under such circumstances? Are there any other verification methods or supplementary data to enhance the reliability of the product?

Response: Thanks for your valuable comments. We agree that the reliability of validation data is critical. To assess our product as comprehensively as possible, we validate against multiple sources and spatial scales—from micro-level household surveys to macro-level regional and national statistics. Specifically, we draw on DHS household survey data, literature-reported estimates, official statistics, and datasets from international organizations (e.g., the World Bank, UN-Habitat), enabling comparisons at city, national, and regional levels. By reducing dependence on any single dataset, we strengthen the credibility of our accuracy assessment.

In Methods, we describe our external comparisons. We have also added a discussion on the quality and limitations of the validation datasets, as follows:

Discussions

…

Another important consideration for accuracy verification is that external validation datasets may themselves contain uncertainties and limitations, stemming from inconsistent reporting standards, heterogeneous data collection methods, or

time lags in statistical updates. To address this, we minimize reliance on any single source by employing a diverse suite of validation data across multiple spatial scales. At the micro level, DHS household surveys provide standardized, representative indicators of slum households based on rigorous sampling. At the meso level, literature-reported estimates and city-level statistics, including both scaled-up values from individual slum areas and directly reported city totals, offer informative reference ranges. At the macro level, national and regional statistics from international organizations such as the World Bank and UN-Habitat provide cross-country benchmarks. Validating against multiple sources and scales strengthens robustness and reduces the risk of bias associated with any single dataset.

6. Currently, the product only includes data from 2018. Does the current method have transferability, and can it be applied to data processing in different years in the future to support temporal analysis of slum populations? If transferable, what conditions need to be met or what adjustments need to be made? This requires in-depth discussion.

Response: Thank you for highlighting the transferability of our approach. Our framework is transferable and supports dynamic, temporal analyses of slum populations. Although the current product focuses on 2018, this year was chosen because it offers the most complete, high-quality combination of ground-truth surveys and satellite datasets, providing a robust benchmark for methodological demonstration. The framework is particularly valuable in data-scarce regions, where satellite imagery can be used to generate spatial predictions of slum populations. Temporally, the model can be extended to future years by leveraging patterns learned from historical data in combination with updated satellite and nighttime light imagery. Nonetheless, we recommend validation against available ground-truth or survey data, as rapid urban development can significantly alter slum population distributions. We have added the following discussion in the revised manuscript, as follows:

Discussions

Our framework is transferable and supports dynamic, temporal analyses of slum populations. It enables forward-looking predictions by leveraging patterns learned from existing datasets in combination with updated Landsat imagery and nighttime lights data. Through data-driven learning, the framework extracts informative features from satellite imagery and captures the relationships between spectral characteristics and slum household proxies, facilitating generalization to new spatial and temporal contexts. To ensure reliable predictions, input satellite datasets must be

consistent in resolution and spatial extent, with preprocessing steps such as cloud masking carefully applied. Recalibration using newly released ground-truth survey data is recommended, as these labels provide an empirical benchmark to evaluate model performance and correct biases, particularly in rapidly urbanizing areas where slum populations may change significantly. By integrating spatial and temporal information, the framework provides a flexible and robust tool for rapid monitoring and spatiotemporal prediction in data-scarce and fast-changing contexts.

7. What are the specific application scenarios of the current product? For example, in urban planning, resource allocation, and social policy formulation, how can the data and results provided by this product be used to guide practical work?

Response: We agree that specifying potential application scenarios is critical for downstream use of this slum population product. Our product advances a human-cantered framework that prioritizes population living in inadequate housing rather than focusing solely on the physical delineation of slum settlements. Gridded slum population estimates provide a direct measure of how many people are affected by deficits in housing and infrastructure services, better reflecting the scale of human vulnerability and directing interventions toward densely populated areas where needs and risks are greatest. In line with your suggestions in comments 3 and 7, we have added a discussion of implications for urban planning, resource allocation, and social policy in Discussion, as follows:

Discussions

… Practically, this continuous representation enables flexible aggregation across multiple spatial scales (e.g., neighborhood, city, national) and provides a robust, data-driven foundation for human-centered urban planning, targeted resource allocation, and evidence-based policy design. For example, planners can optimize the spatial distribution of schools, clinics, and parks to improve accessibility and equity for vulnerable populations, thereby supporting inclusive urban development. These estimates also inform resource allocation and emergency response planning aligned with actual population needs. When combined with spatial information on environmental hazards, such as floods, heat islands, or air pollution, the gridded product becomes a critical tool for resilience planning and adaptive governance, enabling social protection and welfare programs tailored to those at greatest risk.