NBER WORKING PAPER SERIES

GLOBAL HIGH-RESOLUTION ESTIMATES OF THE UNITED NATIONS HUMAN DEVELOPMENT INDEX USING SATELLITE IMAGERY AND MACHINE-LEARNING

Luke Sherman
Jonathan Proctor
Hannah Druckenmiller
Heriberto Tapia
Solomon M. Hsiang

Working Paper 31044
http://www.nber.org/papers/w31044

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2023

This work was supported by a grant from the Human Development Report Office of the United Nations Development Programme. We thank Pedro Conceicao and seminar participants at The Workshop in Environmental Economics and Data Science, the WIDER Development Conference (Bogota), IDinsight, and the American Geophysical Union Fall Meeting for their valuable feedback. We thank the Global Data Lab for sharing DHS cluster-level data on the International Wealth Index. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Global High-Resolution Estimates of the United Nations Human Development Index Using Satellite Imagery and Machine-learning
Luke Sherman, Jonathan Proctor, Hannah Druckenmiller, Heriberto Tapia, and Solomon M. Hsiang
NBER Working Paper No. 31044
March 2023
JEL No. C1,C8,I32,R1

ABSTRACT

The United Nations Human Development Index (HDI) is arguably the most widely used alternative to gross domestic product for measuring national development. This is in large part due to its multidimensional nature, as it incorporates not only income, but also education and health. However, the low country-level resolution of the global HDI data released by the Human Development Report Office of the United Nations Development Programme (N=191 countries) has limited its use at the local level. Recent efforts used labor-intensive survey data to produce HDI estimates for first-level administrative units (e.g., states/provinces). Here, we build on recent advances in machine learning and satellite imagery to develop the first global estimates of HDI for second-level administrative units (e.g., municipalities/counties, N = 61,591) and for a global 0.1 × 0.1 degree grid (N=806,361). To accomplish this we develop and validate a generalizable downscaling technique based on satellite imagery that allows for training and prediction with observations of arbitrary shape and size. This enables us to train a model using provincial administrative data and generate HDI estimates at the municipality and grid levels. Our results indicate that more than half of the global population was previously assigned to the incorrect HDI quintile within each country, due to aggregation bias resulting from lower resolution estimates. We also illustrate how these data can improve decision-making. We make these high resolution HDI estimates publicly available in the hope that they increase understanding of human wellbeing globally and improve the effectiveness of policies supporting sustainable development. We also make available the satellite features and software necessary to increase the spatial resolution of any other global-scale administrative data that is detectable via imagery.

Luke Sherman
University of California, Berkeley
lsherman@berkeley.edu

Jonathan Proctor
Harvard Center for the Environment
26 Oxford Street, 4th floor
Cambridge, MA 02138
jproctor1@fas.harvard.edu

Hannah Druckenmiller
Resources for the Future
1616 P St NW, Suite 600
Washington, DC 20036
hdruckenmiller@rff.org

Heriberto Tapia
Human Development Report
Office, United Nations
Development Programme
304E 45th Street, New York
FF-1288
New York, New 10017
heriberto.tapia@undp.org

Solomon M. Hsiang
Goldman School of Public Policy
University of California,
Berkeley 2607 Hearst Avenue
Berkeley, CA 94720-7320
and NBER
shsiang@berkeley.edu

Global HDI estimates and replication data/code are available at www.mosaiks.org/hdi

Introduction

The Human Development Index (HDI) has been widely used by policymakers and academics since the 1990s to summarize three key dimensions of well-being: the population's health, human capital, and standard of living (*1–4*). It was developed to be a more comprehensive measure of well-being than income or wealth alone (*2*, *3*, *5*) and is commonly used to categorize countries by their level of human development, which, in turn, can determine allocations of global resources, such as development assistance or the prices for international drugs (*6*). However, the Human Development Report Office of the United Nations Development Programme (HDRO/UNDP) releases official global estimates of HDI annually only at the highly aggregated national level (N=191), preventing the use of the indicator in applications that require sub-national information. Thus, while HDI is often considered a more meaningful metric of wellbeing than income, measures of income remain the dominant metric for evaluating development progress within countries in part because they are more readily available.

In an effort to address this, non-UN researchers (*7*) recently collated and processed extensive household survey data in order to produce the first HDI estimates for provinces and states, political units known as "first-level administrative units" or "ADM1 units". This extended previous such estimates (*8*) to the global scale. By constructing HDI estimates for ADM1 units (N=1,765), these efforts have substantially advanced our understanding of global development patterns, but these measures nonetheless remain too coarse for many modern policy applications where granular local information is needed, such as community-level aid targeting (*9*). Indeed, there tends to be substantial inequality in human development within ADM1 units (*10*). Furthermore, the reliance of all current HDI estimates on slow, infrequent, and costly global-scale ground-based data collection sharply limits the usability of HDI for most practical applications other than cross-national rankings.

Here, we produce the first global estimates of HDI at the level of municipalities and counties (N=61,591), known as "second-level administrative units" or "ADM2 units", and for a global $0.1^\circ \times 0.1^\circ$ (approximately 10km by 10km) grid. We construct these estimates by combining information from prior ADM1 estimates (*2*), described above, and global daytime and nighttime satellite imagery (*11*, *12*). To do this, we build on recent advances in machine-learning (*13*) to develop a general method that learns the relationship between satellite imagery and an outcome of interest (here, HDI) using imagery and measurements of the outcome from any set of political boundaries. We can then use that relationship to estimate the outcome for any other set of boundaries. Importantly, our method works for political units or grids of arbitrary shape and size, so models can be trained on coarse-resolution outcome measurements and make predictions at finer resolution, as detailed below. We apply this method to transform HDI measured for ADM1 units into finer resolution global estimates. Note that our fine-resolution global HDI estimates were produced in collaboration with researchers at the HDRO/UNDP and funded by the HDRO/UNDP, but data released with this paper should not be considered official United Nations indicators.

A general method for downscaling administrative data using satellite imagery and machine learning

The combination of satellite imagery and machine learning (SIML) is increasingly used to predict socioeconomic variables remotely at fine spatial resolution (*13–17*). The appeal of this approach is that it may enable information that is expensive to obtain through ground surveys to be estimated at low cost, thereby enabling more frequent data collection. In practice, SIML estimates generally do not replicate ground surveys exactly and researchers are actively studying the nature and impact of errors in SIML predictions (*18*, *19*); however, the quality of SIML estimates is now high enough that it can assist targeting of aid and program evaluation in remote communities where alternative sources of information are unavailable (*9*, *14*, *18*, *20*, *21*).

Currently, the ability of SIML systems to promote development is limited by the paucity of suitable observations for model training (*14*). This limitation, however, is partly due to the design of modern SIML methods, since large quantities of untapped administrative data are available, but existing systems are not designed to make use of them. To date, SIML approaches for predicting human outcomes require standardizing the structure of both the training labels and corresponding imagery so that the unit of analysis is a regular spatial structure, such as a square. For example, many systems use convolutional neural networks (CNNs)(*15*, *16*, *22*), which were originally developed to recognize "natural images" (e.g. photos taken from a hand-held camera) and tend to perform well on diverse computer vision tasks. These CNNs, however, typically require images to be a specific and constant size and shape, such as 224 x 224 x 3 pixels in the case of the commonly used ResNet-18 (*23*). This restriction has caused prior studies to rely on coarse approximations for linking irregularly shaped labels to corresponding imagery, for example, by interpolating or averaging polygon labels that overlap with the square image (*13*, *17*). Such procedures can introduce considerable error when administrative polygons are much larger or smaller than the chosen square size. This is particularly relevant for HDI, for which data is globally available only for relatively large political units, such as nations or provinces, which tend to be irregularly shaped and vary greatly in spatial extent. For example, the largest provincial polygon in our data is the Far Eastern Federal District of Russia, which is over 6 million km^2, and the smallest is Banjul of Gambia, which is 7 km^2. Developing a robust and widely applicable SIML system that can be trained on inputs that correspond with such diverse administrative structures requires an alternative strategy.

Training SIML systems using administrative data that correspond with irregular political units is a general challenge for many researchers. In an ideal setting, we would solve for a function that could directly map a single satellite image "tile," eg. 1km×1km, to the corresponding HDI for the same tile

$$HDI_{tile} = f(satellite\_image_{tile}) + \epsilon_{tile} \tag{1}$$

where ϵ is the component of HDI that is not measurable with imagery. In theory, Eq. 1 could be solved directly with many learning approaches, such as using a CNN or other techniques (*13*, *24*), but this is infeasible in practice because tile-level data on HDI (i.e. the left-hand side of Eq. 1, HDI_{tile}) does not exist. Instead, we can observe only aggregated estimates of HDI over politically-defined regions ($HDI_{country}$ or $HDI_{province}$) that correspond with large and irregular agglomerations of image tiles. For the SIML system described by $f(\cdot)$, this creates a mismatch between the spatial structure of inputs (satellite image tiles) and outputs (political administrative regions).

We solve this problem by converting image tiles into a generalizable set of descriptive features X_{tile}, such that $f(\cdot)$ can be structured as linear in these features

$$f(satellite\_image_{tile}) = \beta \cdot X_{tile}, \tag{2}$$

where β is a vector of weights. An obstacle to achieving this has been the notion that HDI may be a complex nonlinear function of information contained within the original image. However, if a suitable linearization of the image information can be achieved—specifically, if a basis for the imagery can be constructed such that outcomes of interest are well-represented by linear combinations of the basis vectors—then aggregate administrative measures of HDI will project onto corresponding aggregations of tile-level features with the same weights that would be recovered if the problem had been solved using only tile-level data. Thus, we aim to learn a model

$$HDI_{province} = \beta \cdot \underbrace{\left(\frac{1}{N} \sum_{tile \in province} X_{tile} \right)}_{\bar{X}_{province}} + \epsilon_{province} \tag{3}$$

and recover the same weights β that we would have recovered had we directly solved Eq. 1

using the linearization in Eq. 2. Note that $\bar{X}_{province}$ is simply the vector of average tile-level features within a province. The weights β can then be used to generate predictions for arbitrary aggregations of tiles. Specifically, we use these β to downscale HDI to the municipality level ($\beta \cdot \bar{X}_{municipality} = H\hat{D}I_{municipality}$) and the tile level ($\beta \cdot \bar{X}_{tile} = H\hat{D}I_{tile}$). The benefits of linearizing this problem have been understood in general terms, since linear models of basic scalar image properties (e.g. "greenness"(*25*) or nighttime luminosity (*26*)) have been widely used to downscale administrative-level data. However, to our knowledge, it has not been shown that such linearization is possible and skillful for the types of featurizations that capture complex spatial structures in imagery and enable modern high-performance SIML prediction.

To transform satellite imagery into descriptive features that exhibit high performance in linear models, we build on the recent development of Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS), an approach that achieves performance competitive with CNNs using an unsupervised image embedding combined with a linear ridge regression model (*13*). MOSAIKS features have been shown to be skillful at solving diverse prediction problems —such as forest cover, population, elevation, and house price— using only imagery as inputs and using only a single linear specification. This property makes MOSAIKS a particularly appealing approach for predicting HDI, which is constructed from multiple development indicators. Each MOSAIKS feature for a tile describes the similarity between the satellite image and a smaller patch of imagery, and is calculated as a nonlinear transformation of the image's pooled convolution with a random sub-image from the sample (i.e. random convolutional features (*27*)). Together, MOSAIKS features form a basis that can skillfully describe the rich structure contained within large imagery datasets through simple linear combinations of the features (*13*).

To compute local HDI via SIML, we transform a dataset of global Planet imagery ($\sim$4m resolution) into general-purpose MOSAIKS features (X) for 0.01° x 0.01° tiles ($\approx$ 1km x 1km; Figure 1A-D) (*11*). We supplement these MOSAIKS features with features that flexibly characterize the distribution of nighttime lights in each tile (Methods 2.2). We then learn a model that is linear in these features ($\beta \cdot X$) and use this linear model to estimate HDI at high resolution. Specifically, for both training and prediction, we assign image features to administrative polygons that contain the centroid of each tile and average them to the polygon-scale using population weights (Figure 1D, grid-scale predictions follow a similar procedure) (*28*). This results in one vector of image features for each province and municipality in the world. To learn the relationship between the image features and HDI, we train a ridge regression on province-level HDI labels and aggregated province-level image features (Figure 1E). We then predict municipality-level HDI using the municipality-level image features (Figure 1F), and we predict $0.1^{\circ} \times 0.1^{\circ}$ grid HDI using features for that grid. We tune the ridge hyper-parameter using 5-fold cross-validation and evaluate performance using a held-out test set (Methods 3.1). While we focus this analysis on the downscaling of HDI, this approach is generalizable to other types of administrative data associated with irregularly-shaped political units.

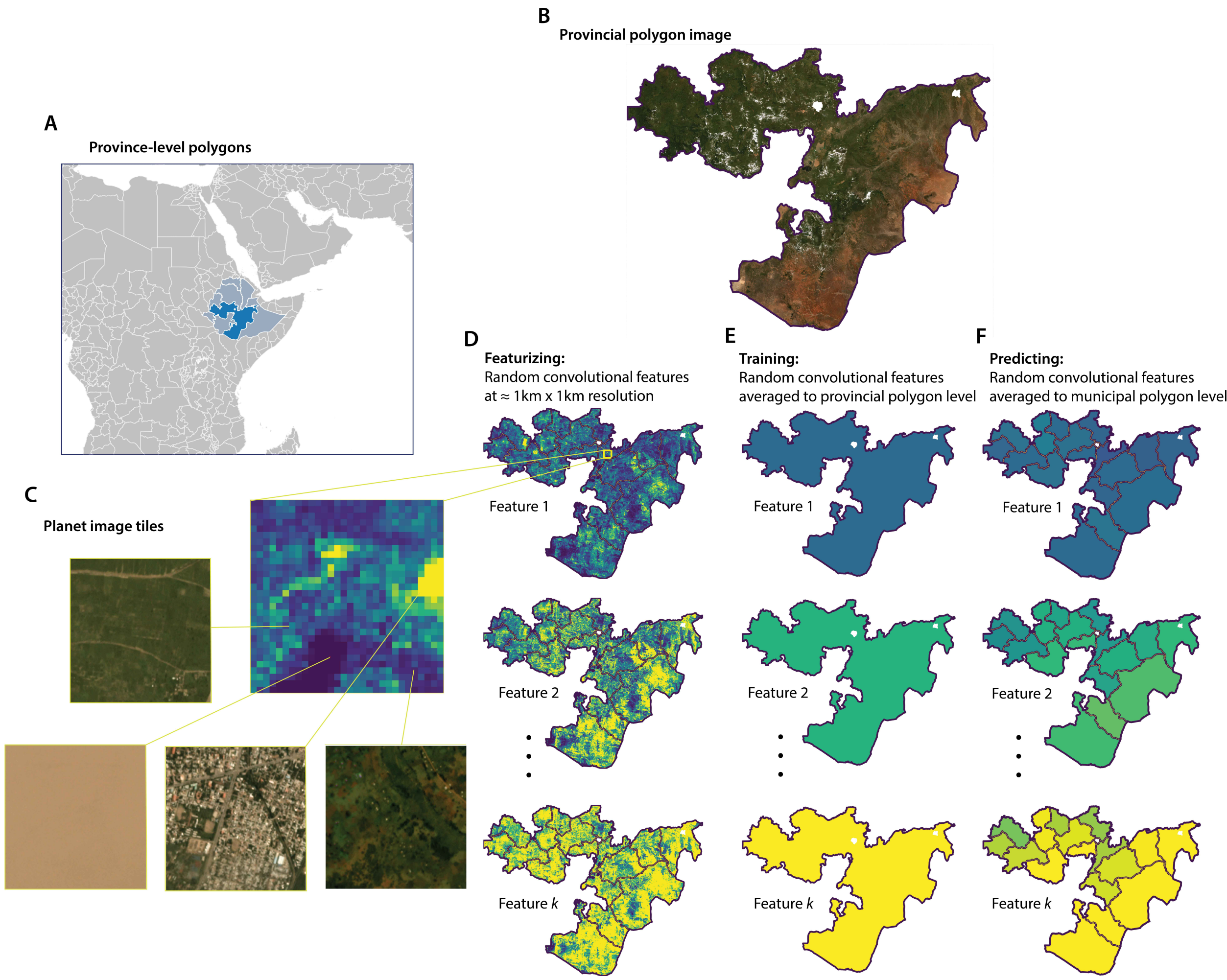

Figure 1: **MOSAIKS is used to transform satellite imagery for each administrative polygon into a vector of image features**. (**A**) The location of Oromia, an example province (ADM1 unit) within Ethiopia. (**B**) A composite of Planet imagery over Oromia in 2019. (**C**) A sample of 0.01° x 0.01° image tiles. (**D**) Three examples of MOSAIKS features over Oromia; each pixel shows the feature value for a single 0.01° x 0.01° image (X_{tile}). (**E**) The corresponding aggregation of these MOSAIKS features to the provincial polygon (ADM1) level for model training ($\bar{X}_{province}$). (**F**) Aggregation of these same MOSAIKS features to the municipal polygon (ADM2) level for fine-resolution prediction of HDI ($\bar{X}_{municipality}$).

Results

Our results have four sections. First we train and evaluate a global model for HDI at the province-level using aggregates of satellite features. Second, we implement multiple tests to validate that this model is skillful at downscaling data sets similar to our global HDI data, since direct global validation against HDI is impossible. Third, we generate the first high-resolution global HDI data using this procedure and we evaluate how these estimates compare to existing aggregated estimates. Last, we illustrate how these new high-resolution estimates could alter decision-making when targeting aid.

1. Predicting province-level HDI using satellite imagery

Using province-level administrative HDI data for training (as in Equation 3 and Figure 1), we find that predictions made using linear aggregates of MOSAIKS daytime and nighttime satellite image features, anchored to known country means, explain 96% of the variation in global provincial HDI values (Figure 2A, denoted "full variation performance"). Specifically, we train a model to predict provincial HDI deviations from the country mean and then add the known country mean onto the predicted provincial deviations. We take this mean-anchored approach because it reflects how SIML may be used to augment existing HDI data in practice. Evaluating the ability of this model to predict provincial HDI deviations from the country mean directly, we find that it explains 46% of the *within-country* variation in HDI (Figure 2B, Methods 3.2). This indicates that SIML provincial predictions of HDI add substantial fine-resolution information to existing national measures and supports our mean-anchoring approach. If we hide country-level data from the model and train on and predict provincial values directly (without anchoring estimates to country means), we find that the model explains 79% of the total (i.e. both across- and within-country) variation in provincial HDI (Table S1).

The greater difficulty predicting HDI variation within countries relative to across countries is due, in part, to the relatively smaller within-country variation available for model training and evaluation, as illustrated by the yellow and pink observations of provincial values for France and Ethiopia in Figure 2A and their demeaned values in Figure 2B. We note that models trained on provincial HDI deviations from the country-level HDI have higher performance predicting such deviations than models trained directly on the provincial values themselves (Table S1 col. 4). This results from model weights being optimized to explain the smaller within-country variation in the demeaned model, rather than the larger across-country deviations, leading to higher skill when predicting local-level variation. As we aim to predict local-level variation in HDI, in the following analysis we present results from models trained on deviations from the country mean and emphasize performance for the relatively more difficult task of predicting local (i.e. within-country and within-province) variation. Positive performance (i.e. $R^2 > 0$) explaining variation within the level of model training (evaluated below) indicates that model predictions are able to improve our understanding of the spatial distribution of human development.

Evaluation metrics for the provincial models above are calculated using a spatial-cross-validation procedure in which models are trained and evaluated on data from non-overlapping sets of countries. Similar performance is achieved in a held out test set from countries not previously used for model testing or training, indicating that the model was not over-fit to the training data (province-level full variation performance is $R^2 = 0.96$ and within-country performance is $R^2 = 0.39$, Table S2).

2. Validating downscaling of data below the province-level

We cannot directly evaluate the performance of municipality-level or grid-level HDI predictions worldwide because such highly-resolved estimates have not been previously constructed. Nonetheless, we test the performance of our downscaling technique using three alternative sources of similar data that allow predictions to be directly compared to "ground truth" at finer resolution than the province level. First, we directly compare our municipality-level HDI predictions in Mexico to a unique set of available survey-based estimates of HDI at the same resolution (*10*). Unfortunately, to the best of our knowledge, similar survey-based validation samples are not available outside of Mexico, preventing us from conducting analogous global-scale validation. Second, we train a model relating satellite imagery to the International Wealth Index (IWI) at the province level, and then construct downscaled predictions of IWI at the resolution of Demographic and Health Surveys (DHS) clusters where granular IWI measurements are available (*29*). The IWI is an alternative development indicator to HDI that omits measures of education and health. Third, we train a model to predict nighttime luminosity (NL), a common proxy for economic wellbeing (*30–34*), using MOSAIKS features constructed exclusively from daytime satellite imagery, and test whether our approach can downscale nighttime luminosity. Mirroring the structure of our HDI analysis, we train a model using only nighttime luminosity labels aggregated to the province-level, and then evaluate predictions of luminosity at the municipality-level. None of these three tests can directly validate the performance for downscaling HDI globally; however, all three tests taken together document the effectiveness of our downscaling strategy in general, using socioeconomic data that are similar to HDI.

When evaluating estimates made at finer resolution than that of the training data, we first find the optimal ridge hyperparameter using a spatial-cross-validation procedure, again splitting the data by country, and then re-train a new model using all of the available coarse resolution data before predicting at fine resolution. We generally mean-anchor downscaled estimates to the known provincial mean (Methods Section 3.4). This approach uses the satellite-based model to explain within-province variation, which is previously unknown, and the measured provincial values to explain the across-provincial variation, which is previously known, to produce the best possible estimates.

Comparison of downscaled HDI to municipality HDI measurements in Mexico As a direct evaluation of HDI downscaling performance, we compare municipal HDI predictions from the satellite-based MOSAIKS model trained on provincial HDI deviations from the country mean to municipality HDI derived from census-based calculations in Mexico in ref. [(*10*)]. Downscaled HDI predictions explain 40% of the municipal HDI variation in Mexico overall (Figure 2C, municipal predictions centered to the known provincial mean) and and 23% of the within-province variation (Figure 2D). These results indicate that our method for SIML-based downscaling substantially improves our understanding of the spatial distribution of HDI within Mexico; although, importantly, they are not a complete substitute for survey-based estimates when such data are available. However, since survey-based HDI estimates at the global scale do not exist, SIML-based estimates may be the only available option in many contexts.

Downscaling the International Wealth Index across DHS clusters We test the ability of MOSAIKS to downscale IWI internationally by training a model on province-level aggregates of IWI and then predicting IWI across DHS clusters. This is a more difficult task than predicting municipality-level values and an equally difficult task as predicting at the $0.1^\circ \times 0.1^\circ$ grid-level, since DHS clusters tend to be even finer resolution than municipalities and about the same size as the HDI grid (DHS cluster average area $\approx$ 180 km^2, municipality average area $\approx$ 2,000 km^2, HDI grid cell area $\approx$ 120 km^2). Models are trained on 863 provincial observations within 86 countries and evaluated at 51,996 DHS clusters (Table S1). Analogous to our approach with HDI, models are trained on province-level deviations from country-level means and predictions are re-centered to match the observed province-level mean, as these values are known and represented in the training data (see Methods Section 3.6). Downscaled IWI predictions explain 75% of the variation in IWI across all DHS clusters (Figure 2E) and 59% of the variation in IWI across DHS clusters within countries (i.e. of cluster deviations from the country mean, Figure 2F). Importantly, this approach is also able to predict variation in IWI within the provincial units that it was trained on,

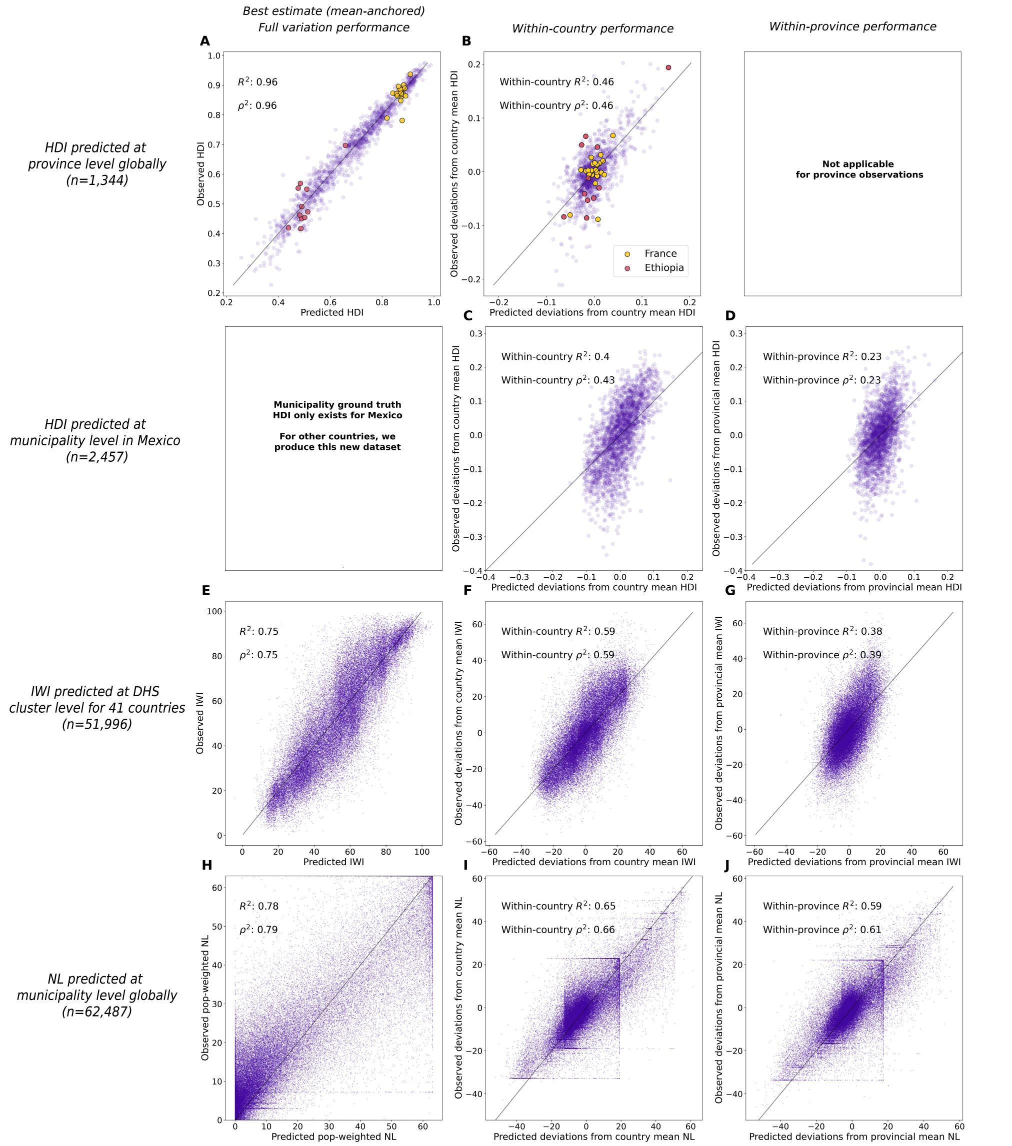

Figure 2: **MOSAIKS models perform well at regular and downscaled resolution.** (**A**) Observed and predicted HDI at the province level (country mean added to predicted deviations from country means). Note that the within-country variation is smaller than the across-country variation, as illustrated by France, in yellow, and Ethiopia, in pink. (**B**) Within-country observations and predictions of HDI. Provincial deviations from the country mean for France and Ethiopia are now centered at 0, and the model is evaluated on how well it can differentiate provinces that are relatively well and worse off within countries. (**C**) Predicted HDI at the municipality level in Mexico and census-derived data from Permanyer (2013) (*10*). (**D**) Predicted and census-derived HDI within Mexico's provincial units. (**E**) Observed and predicted IWI at the DHS cluster level (re-centered on province mean). (**F**) Observed and predicted IWI at the DHS cluster level within countries. (**G**) Observed and predicted IWI at the DHS cluster level within provinces. (**H**) Observed and predicted population weighted NL at the municipality level (country mean added back). (**I**) Observed and predicted NL at the municipality level within countries. (**J**) Observed and predicted NL at the municipality level within provinces.

explaining 38% of the variation in IWI cluster values within provinces (Figure 2G). This result demonstrates the ability of our downscaling approach to generate skillful global-scale predictions at resolutions higher than the training data, and also its ability generalize to measures other than HDI.

Downscaling nighttime luminosity globally using only daytime imagery To further evaluate our approach in a global test, we train a model on aggregate provincial nighttime luminosity and evaluate predictions at municipal resolution. Nighttime lights are not a direct measure of human welfare; however, they are generally correlated with income and other indicators of development (*32–34*) and are useful here because they allow us to design a validation test where true subnational values are known worldwide. We train and evaluate the model using provincial deviations from the country mean, and then predict municipal values as the predicted municipal deviations from the country mean plus the known country mean (Methods Section 3.7). Downscaled luminosity predictions capture 78% of municipal variation in nighttime lights globally (Figure 2H), 65% of the municipal variation within countries (Figure 2I), and, most importantly, 59% of the variation across municipalities within provinces (Figure 2J). These results further reinforce the ability of our approach to downscale global province-level data and underscore its generalizability to other non-HDI outcomes. Unlike the two downscaling experiments above, this experiment relies entirely on features generated using only daytime imagery (Figure S1).

Collectively, these three downscaling experiments demonstrate that our approach effectively combines coarse socioeconomic measurements with satellite data to produce skillful estimates of these measures at spatial resolutions finer than the aggregated province-level training data.

3. Global municipality-level and grid-level estimates of HDI

We use our model for subnational HDI (from Results Sections 1,2) to estimate HDI for 61,591 municipalities and 806,361 $0.1° \times 0.1°$ grid cells (Figure 3), the finest resolutions at which HDI has been estimated globally. Specifically, we use the model that was trained on provincial deviations from the country mean, using both daytime and nighttime satellite image features, to make predictions of the within-country distribution of HDI at the municipal and grid levels. We then center the mean of these estimates on the observed provincial means from Smits and Permanyer (*7*) (see Methods Section 3.8). We make these municipal and grid-level estimates of HDI publicly available for download at `mosaiks.org/hdi`.

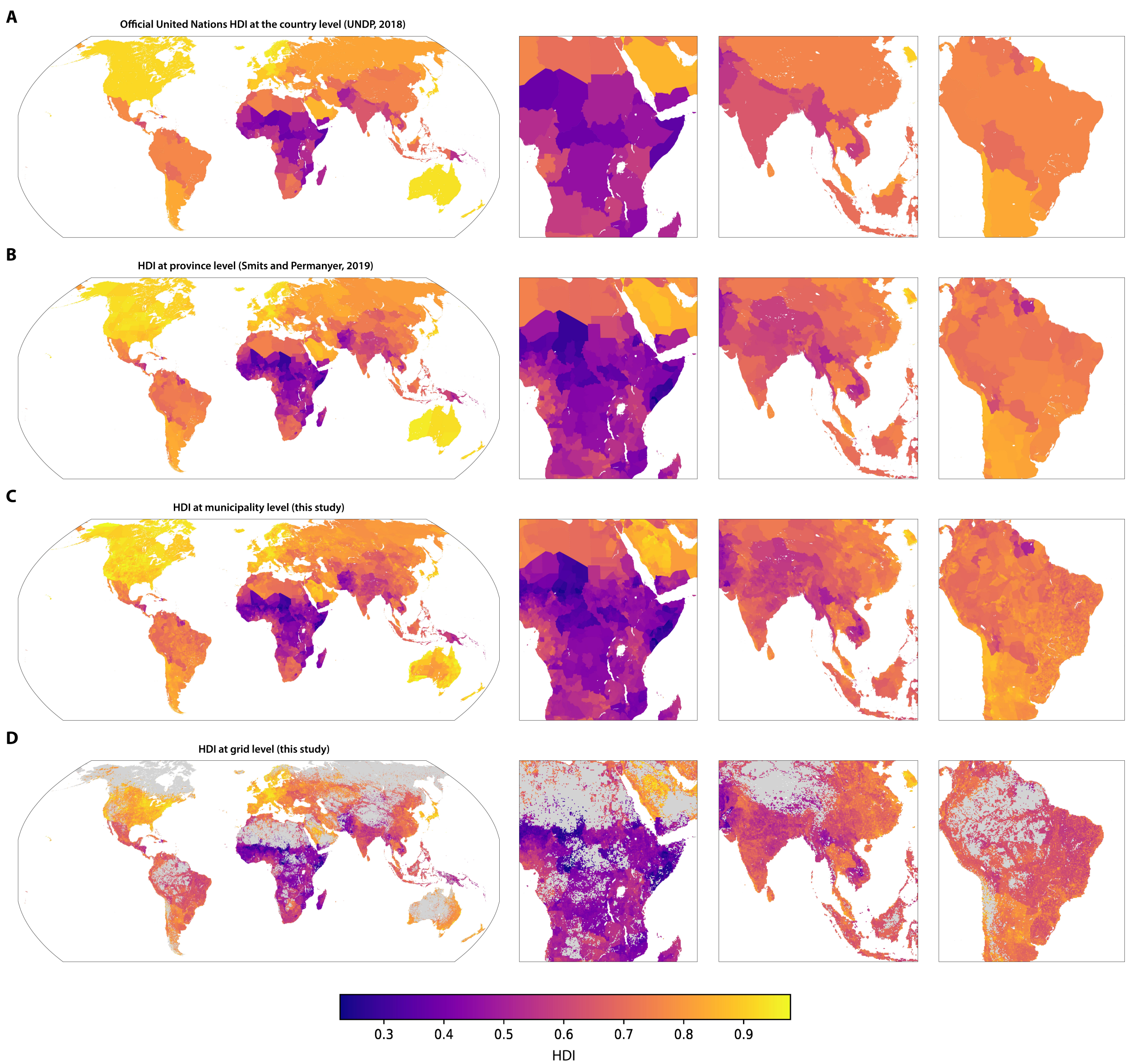

Figure 3: **Global HDI estimates at the municipal and grid levels.** (**A**) Official United Nations HDI at the country level (*35*) (**B**) HDI data at the province level from Smits and Permanyer (*7*). (**B**) Municipal level estimates of HDI produced here. (**C**) Grid level estimates of HDI at the 0.1° by 0.1° (approximately 10km by 10km) level produced here. Grey in the grid-level estimates indicates land area believed to be unsettled (*36*). All data shown are for the year 2018.

Our high-resolution estimates enable a substantially more detailed understanding of human development compared with national and provincial measures (Figure 3 A,B vs. C,D). Both municipal and grid-level estimates reveal within-province heterogeneity of HDI that was previously un-resolved (Figures S2, S3). The gridded HDI estimates tend to be relatively higher along major roadways, and especially at the intersection of roadways (Figure S4). Boarders, such as between Turkey, Georgia, Armenia, Azerbaijan, Iraq and Iran, are less apparent in the fine resolution estimates, indicating a greater continuity in human development across space than in the provincial maps. The wealthier city centers and poorer suburbs of capital cities such as Moscow in Russia and Antananarivo in Madagascar are also visible in the municipal and grid estimates, but obscured in the provincial estimates. The contribution of environmental features to human development is illustrated in eastern Pakistan and northwestern India, where human development is higher in the plains bordering the Indus River and its tributaries, and much lower in the neighboring deserts. Similarly, within Sonora and Sinaloa in Mexico, the coastal areas show relatively higher human development than the inland regions. This revealed local heterogeneity in HDI indicates that uniform assignment of HDI to populations based on their country or province of residence is inaccurate because it groups together populations with very different levels of human development.

Comparing the grid-level and municipal estimates, we see that while the grid-level estimates capture HDI variation within municipal polygons, they do not represent boundaries as sharply as the municipality-level estimates, which are mapped to boundaries in their construction. For example, grid-level estimates along coastlines can be extended incorrectly into the ocean, and small portions of the Dominican Republic that border Haiti are estimated to have a considerably lower HDI than they likely should due to the imperfect match between the administrative boundaries and the grid (Figure S4B). Municipal estimates capture these boundaries more precisely (Figure S4A).

We use our estimates to quantify the degree of aggregation bias that occurs when using only province-level estimates. Aggregation bias occurs in this setting because small units (i.e. grids or municipalities) are assigned the aggregate HDI of a larger unit (i.e. a province), but that assignment does not reflect the specific conditions within the smaller units. For example, a small urban region, where HDI tends to be high, embedded in a province that also contains large rural areas, where HDI tends to be lower, will be assigned a HDI level that is too low when coarse provincial measures are used. We quantify how frequently such mis-assignment occurs within countries by assigning populations to a quintile of HDI within their national HDI distribution based on provincial, municipal, or grid-level estimates. We then evaluate how frequently the province-level estimates agree with the more highly resolved estimates (Figure 4).

We find that a majority of the global population (53% using municipal estimates and 63% using grid estimates) is assigned to a different within-country HDI quintile compared to when using provincial estimates. For example, of the population measured to be in the bottom two HDI quintiles by the provincial estimates, 6.1% are measured to be in the top two HDI quintiles by the municipal estimates and 10.3% by the grid-estimates. Grid-level estimates tend to reveal larger amounts of aggregation bias in the provincial estimates due to their finer resolution. Based on our grid-level estimates, we estimate that 22.2% (18.4%) of the global population is one quintile lower (higher) than assigned using provincial estimates, and 5.4% (5.9%) are two quintiles lower (higher). Aggregation bias is especially concerning for communities with lower human development that live nearby communities with higher human development who, for example, might miss receiving assistance from development programs if only coarse HDI estimates were used to determine the allocation of aid. We calculate that over a hundred million people (1.3% of the global population) measured by the provincial estimates to be among the 40% most developed in their countries are measured by the grid-level estimates to actually be among the 20% least developed.

4. Illustrative application: targeting policy in Mexico

To explore how these new HDI measures could improve the efficiency of development policies, we conduct a simulation exercise for Mexico. We simulate how a geographically targeted policy based on provincial vs municipal HDI data (Figure 5A-B) might achieve different outcomes, noting that previous work has shown that more spatially granular targeting can produce meaningful welfare gains (*16*, *37–39*). We study the context of Mexico because we can access "ground truth" estimates of HDI (*10*), discussed earlier (recall Figure 2C-D), which can be used to evaluate the performance of the targeting exercise.

We consider a hypothetical scenario in which a program administrator has a fixed budget of aid to distribute to some portion of the population of Mexico. We suppose the administrator would like to target individuals with the lowest HDI. Following Chi et al. (*16*), we assume that the program will be geographically targeted and that all individuals within targeted regions will receive the same transfer, a practice used to reduce administrative costs (*37*, *40*). If the administrator has access to only province-level HDI measures, then eligibility is determined at the province-level. Alternatively, using our municipal-level estimates, the administrator is able to target the program at the municipality level. We evaluate how access to more granular HDI data improves the number of program recipients that are correctly targeted.

We evaluate the performance of the two targeting strategies using the census-derived

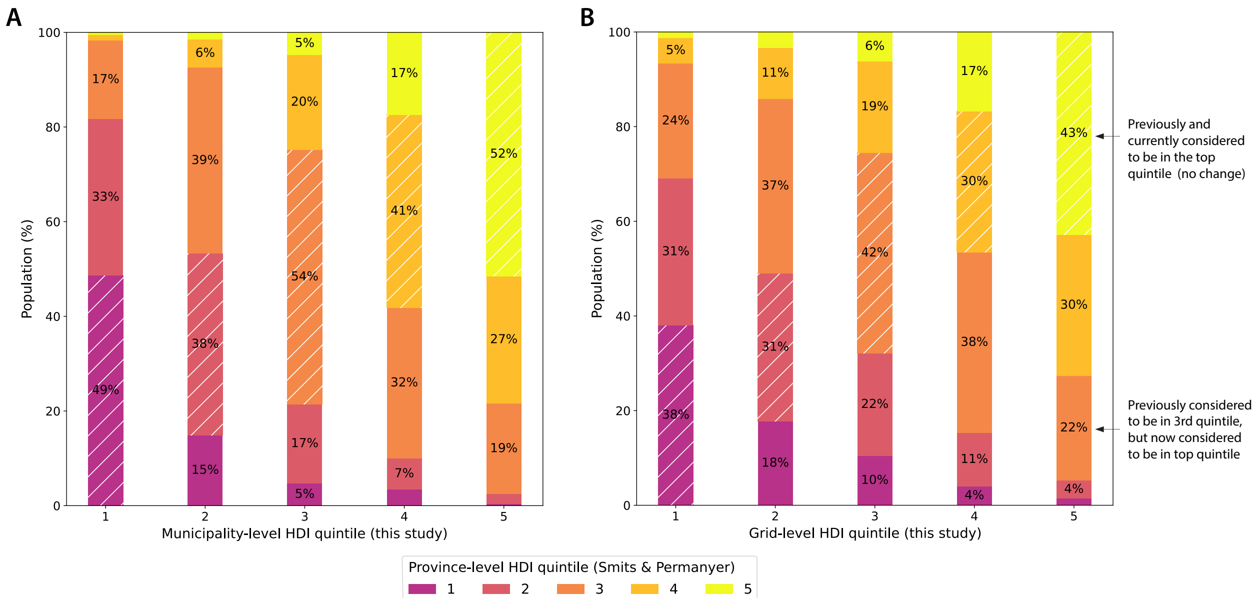

Figure 4: **Municipal and grid-level estimates of HDI assign more than half of the global population to a different within-country HDI quintile than provincial estimates.** (**A**) Shows the difference in estimated HDI quintile, within countries, using provincial vs. municipal data. (**B**) shows the same analysis using grid-level HDI estimates. Colors show the estimated HDI quintile using provincial data from (*7*), where yellow is high human development and purple is low human development. Bins along the x-axis show estimated HDI quintiles using the municipal and grid level data produced in this study. Hatch marks indicate no change in quintile assignment using municipal data. When provincial data do not allow for the creation of five distinct bins, the population is assigned first to the middle bin, followed by the neighboring bins. For example, if a country does not have any province-level data, the entire population is assumed to be in the middle quintile for that country. For this reason, a greater fraction of the global population is assigned to the middle quintile (32%) than to the outer quintiles when using the provincial data.

municipality-level HDI measurements (*10*) — which are not used to train our model — as the basis for ground truth. These estimates are municipality-level aggregates, so we simulate HDI for *individuals*–i.e. the targets of the policy–by imposing additional assumptions about the distribution of HDI across individuals within a municipality. We assume HDI within each municipality has a truncated normal distribution (bounded between 0 and 1) that is centered around the census-derived municipality mean (Figure S5). Because this distribution is not observable, we test the sensitivity of our results to different assumptions about the dispersion of this distribution.

The use of municipality-level data improves the number of program recipients correctly targeted. Supposing that the program director has funds to target 10% of the population and aims to provide assistance to the 10% of individuals with the lowest HDI, accuracy of program targeting increases by 7.9% percentage points (from 33.9% to 41.8%) when using municipal data compared with provincial data, assuming a within-municipality HDI standard deviation of 0.1 (Figure 5E). Use of the municipal data results in a much greater geographic dispersion of targeted municipalities (Figure 5C-D).

Targeting performance can also be evaluated using a receiver operating characteristic curve (ROC) curve (*9*). In this exercise, the aim is still to provide assistance to the 10% of individuals with the lowest HDI, but the fraction of the total population targeted is modified to examine how the true positive (individuals with low HDI correctly given aid) and false positive (individuals with HDI above the desired cutoff incorrectly given aid) rates change accordingly. Moving from left to right on the x-axis in Figure 5F implies a greater number of people receiving assistance. The area under the curve (AUC) shows the efficiency of the targeting, with higher values indicating that the HDI estimates help the program administrator better distinguish between individuals with low and high HDI. The AUC increases by 0.08 (+12% from 0.76 to 0.85) when municipal data are used instead of provincial data, indicating improved targeting performance across this range of program constraints.

The degree to which municipal HDI measures improve targeting performance relative to provincial measures depends on how dispersed individual HDI values are within municipalities, with lower (higher) assumed dispersion leading to larger (smaller) absolute — though similar proportional — gains (Figure 5E and Figure S5). Given the absence of ground-truth for individual-level HDI measures on which to evaluate performance, this simple exercise is not intended to produce numerically accurate results, but rather to demonstrate how access to more spatially granular data on human development might allow a program administrator to better direct resources towards those who need them most.

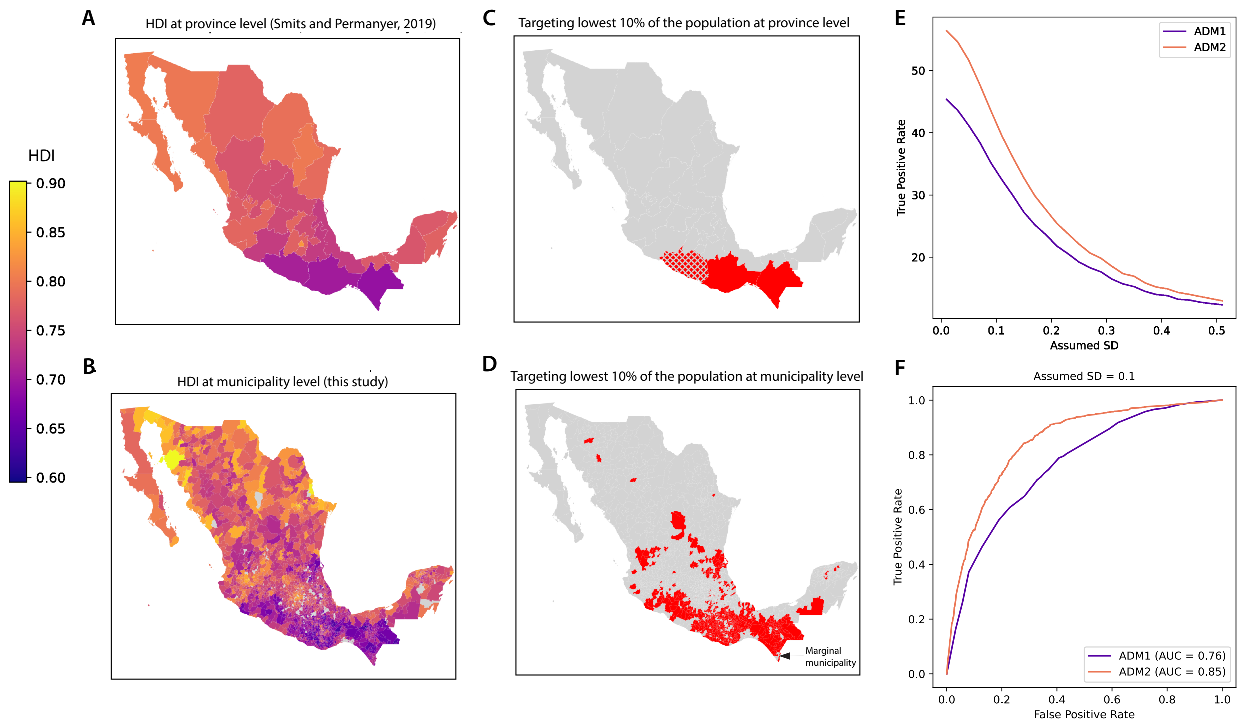

Figure 5: **Spatially granular HDI measures can improve decision-making.** (**A**) HDI at the provinice level of observation (*7*) (**B**) HDI estimates at the municipality level produced in this paper. (**C**) Lowest HDI provinces that would be targeted until 10% of the country's population is reached. (**D**) Lowest HDI municipalities that would be targeted until 10% of the country's population is reached. Hashing in (**C**) and (**D**) shows the marginal province and municipality that would be partially targeted. (**E**) Targeting accuracy (true positive rate) as a function of the assumed standard deviation of HDI within each municipality. (**F**) ROC curves illustrate the degree of improvement that comes with targeting at the municipality level relative to the province level (assumed SD of 0.1 within each municipality).

Discussion of model performance

Model performance for components of HDI One motivation for using MOSAIKS features to downscale HDI is their ability to predict a diversity of ground-based measures. This is particularly relevant for predicting HDI, since it is constructed from components that capture human health, education, and income. To consider which components of HDI are best captured by our estimates, we retrain models to predict each component of HDI separately. We find that MOSAIKS models explain 93%/7% of the full/within-country variation in provincial life expectancy, 93%/48% of mean years of schooling, 90%/22% of expected years of schooling, and 96%/46% of gross national income per capita (GNIpc) (Table S3). While the components of HDI do tend to be correlated (Table S4), these results indicate that instead of just capturing income, predictions of HDI using satellite imagery maintain the ability to capture multiple dimensions of human wellbeing. Predictions of HDI made from combinations of its individually predicted components perform nearly identically to the direct predictions of HDI used throughout this analysis.

Model performance across regions Analyzing the performance of MOSAIKS models across space, we find that performance tends to be the highest in low income regions, especially sub-Saharan Africa, where such measurements are likely to be of greatest value (Figure S6A,C). Specifically, we find that MOSAIKS models explain more variation of province-level HDI deviations from the country mean in areas of low human development (HDI < 0.6, $R^2 = 0.56$) than in areas of medium or high human development (HDI > 0.6, $R^2 = 0.29$, Figure S6A). We also see this pattern in predictions of DHS cluster-level IWI deviations from the province mean, with performance increasing monotonically from $R^2 = 0.24$ for countries with the highest HDI values to $R^2 = 0.48$ for countries with the lowest HDI values (Figure S6B,D). This improved performance may be due to increased variance of HDI and IWI values within these countries (Figure S6E-F), which provides more variation to exploit during model training. Alternatively, variations in well-being being may be relatively easier to see from satellite imagery in areas with lower human development. Relatedly, model performance for both HDI and IWI is higher in regions with higher inequality, highlighting the particular value of these estimates in regions of high inequality ($p < 0.01$ for both HDI and IWI; pearson's ρ is 0.30 for HDI and 0.48 for IWI comparing the within-country standard deviation of provincial values and the within-country predictive performance, measured by ρ^2).

Value from combining daytime and nighttime imagery The MOSAIKS system can use image features from multiple sensors simultaneously when training models, a property that is used throughout this analysis to predict HDI from both daytime and nighttime imagery. Analyzing the performance of MOSAIKS models based on the type of satellite imagery used, we find that daytime and nighttime imagery together explain 8% more variation in provincial HDI deviations from the country mean than does nighttime imagery alone, improving model fit by 21% (Table S1). This improved performance from using daytime and nighttime imagery together is especially strong in regions of low human development (HDI < 0.6) (Figure S6A), consistent with a previous finding that models using daytime imagery outperform models using nighttime imagery when predicting assets of the poorest populations in five African countries (*22*). Analyzing model performance for each component of HDI, we see that the improved performance predicting HDI using daytime and nighttime imagery stems from improved or comparable performance predicting each component of HDI (largest change in R^2 is 0.10 for mean years of schooling, smallest is no change in R^2 for life expectancy, Table S3).

Model performance training at the country-level To evaluate performance in an extremely data-limited setting, we re-train our model using only country-level data. Despite a low number of training observations (N = 86 to 170 across experiments) these models maintain 44% to 87% of the performance of our preferred models trained using provincial deviations from the country mean (N = 863 to 2,848) when evaluated on the relative ordering of predicted and observed values using ρ^2 in all experiments (Table S1). This indicates that our approach can achieve competitive predictive performance detecting locations with relatively higher and lower HDI even when trained on few and coarse observations of the variable of interest. Performance predicting the exact level of HDI is lower, especially when evaluated within-country, likely due to the large difference in the magnitude of HDI variation across countries versus within countries, discussed above (Figure 2A,B).

Conclusion

We produce and make freely available the first global-scale high-resolution estimates of HDI, enabling the use of broad-based measures of well-being for local decision-making and the prioritization of local policy actions. We achieve this by developing an approach for generating spatially granular measures of human well-being using SIML models trained on coarse and inconsistently structured administrative data. Since many forms of non-HDI data are also available only for administrative regions, we believe this generalizable method will increase the range of outcomes that can be used as labels in SIML models, as limited training data currently constrains the production and societal impact of SIML systems (*14*). To support researchers that may apply this approach to other outcomes, we also make freely available aggregations of features to the political boundaries used for training and prediction in this analysis.

Broadly, our approach of predicting outcomes at a fine resolution globally using only a small sample of coarse resolution labels differs considerably from what has been done in the poverty mapping literature to-date. For example, existing SIML literature using DHS asset measures has trained and evaluated at the same spatial scale, and has generally only considered observations from African countries, whereas we include all available data in our models (*15*, *22*, *33*).

Our strategy is motivated by the limited resolution of available training data for HDI, but our results do not exhibit major or obvious compromises in performance relative to existing alternatives that exploit high-resolution labels. The benchmark in this literature achieves $\rho^2 = 0.63$ to 0.67 predicting a wealth index when training and evaluating at the DHS cluster resolution (*15*, *41*). Though we do not train at the DHS cluster level, our performance predicting DHS cluster IWI is competitive with this previous analysis. We achieve lower performance when training on province values ($\rho^2 = 0.5$) and higher performance ($\rho^2 = 0.75$) when training on within-country anomalies and re-centering our estimates to the observed province averages (Table S1, col 1). Direct comparison, however, is complicated by the differences in training and evaluation methodologies.

Nonetheless, our study has several important limitations. First, there is a limit to how well every socioeconomic variable can be predicted using satellite imagery. While incorporating additional training data and imagery sources could further improve our performance, it is unlikely that any SIML system could predict 100% of the variation in HDI. One important property of the errors in our estimates, common in machine learning predictions generally and SIML predictions specifically (*13*, *18*, *19*), is that our predictions exhibit lower variance than the true values. For example, in Mexico where we predict 40% of municipal variation in HDI, the standard deviation of our satellite-derived estimates is approximately half that of census-derived values. As SIML measurements improve, survey and other traditional approaches to data collection will remain critical to informing the state of global human development and will continue to complement satellite models, which cannot be trained or evaluated without ground-truth measures.

Second, our model estimates and their evaluation are limited by the quality of HDI observations. For example, the province-level data on HDI that we use come from Smits and Permanyer (2019), whose estimates of HDI and its component indicators are also imperfect (*7*). Generally, such errors in training data reduce model performance, indicating that improved provincial measures could benefit our approach. Relatedly, errors in evaluation data tend to lead to overly conservative estimates of model performance, as SIML estimates may differ from ground data in part due to errors in the ground data itself (*14*).

A third limitation is that we focus on producing cross-sectional estimates. While we expect that our fine-resolution estimates of HDI will be useful in many ways, we also expect researchers, governments and non-government organizations to be additionally interested in changes to HDI over time. Evaluating the ability of our downscaling approach and other SIML systems to capture such changes is an important area for future investigation.

We emphasize that the approach described here can likely be used to predict a wide variety of labels for which country, province, and/or municipality level labels exist. To facilitate this use, we make the MOSAIKS features used in this analysis publicly available at the country, province, and municipality level via `https://mosaiks.org/hdi` (*11*). We offer these features aggregated to these administrative-unit levels using both area and population weights. Each of these files is relatively small, $\approx$ 3 GB or less, and thus relatively easy to process on a desktop computer. For comparison, the global set of features is $\approx$ 3 TB and the raw imagery is $\approx$ 30 TB. We hope that agencies and policymakers can leverage these published features, along with their own administrative datasets, to produce new downscaled estimates of socially-relevant outcomes. We believe that such spatially granular data will create new opportunities for achieving global development goals.

Methods

Throughout this analysis we use the term "province", the abbreviation "ADM1", and the subscript p to refer to first-level administrative regions; and "municipality", "ADM2" and the subscript m to refer to second-level administrative regions, though the terminology for these units varies by country. For example, "state" and "county" are the designations used for ADM1 and ADM2 units in the United States. We use the subscript c to denote observations at the country level.

1 Label data

National-level HDI data originate from the UNDP Human Development Data Center and are updated every year (*35*). HDRO/UNDP uses data from the World Bank, UNESCO, UNICEF, DHS, UN Stats, and other organizations to create these national-level indicators (*4*). We use data from 2018, as those were the most recently available data when we began our analysis.

Province-level data on HDI and its components come from the Global Data Lab (GDL) Sub-national HDI Database V4.0 (*7*, *42*). We omit 3% of the observations, which do not match with the associated GDL shapefile. The resulting province-level HDI dataset contains 1,707 provincial observations from 157 countries. Additionally, we include 22 country-level observations that do not have subnational province units (e.g., Qatar). Again, we use provincial HDI data from 2018, as those were the most recently available data when we began our analysis.

IWI data also come from GDL. These data are publicly available at the country and province levels and we use these data for 2018. GDL also provided us IWI data at the DHS cluster level, which are not publicly available. We use cluster-level IWI estimates from 2012 through 2019 in this analysis. We drop observations that do not overlap a parent province polygon and for which no imagery is available. This results in 51,996 DHS cluster observations from 41 countries.

We match all label data to time-constant satellite image features. The MOSAIKS daytime features are from 2019 imagery, and the nightlight features are from 2013 imagery. Because these features are not contemporaneous with all labels, our results present a conservative estimate of the ability of SIML to measure HDI globally. Given that HDI variation is substantially larger over space than time, however, we believe that perfectly contemporaneous measures would likely improve performance only modestly.

2 Creation of features

2.1 MOSAIKS features

We create daytime image features using Planet's Surface Reflectance Basemaps product from 2019, which has a pixel resolution of 4.77m x 4.77m at the equator. These quarterly mosaics are processed by Planet to minimize cloud cover, balance color across seasons, and remove seams from images (*11*, *12*). We use data from quarter 3 because it corresponds with less ice coverage in the northern hemisphere and less cloud cover in the tropics. We follow the methods described in Rolf et al. (2021) to generate a set of 4000 task agnostic daytime image features using Random Convolutional Features (RCF) (*11*, *13*). Two thousand of these features use a patch size of 4 x 4 x 3 pixels, and the other two thousand use a patch size of 6 x 6 x 3 pixels. The third dimension of the patch size refers to the number of color bands (i.e., red, green, and blue) that are available in Planet imagery. We selected these patch sizes because they maximized performance across three non-HDI prediction tasks: predicting nightlight intensity, road length, and forest cover at the global level. We tested patch sizes ranging from 3 x 3 x 3 to 10 x 10 x 3 and found that using a combination of two different patch sizes (with 2,000 patches each) outperformed using a single patch size (with 4,000 patches) across all three tasks.

We create features for all land tiles with available imagery on a global 0.01° x 0.01° equal-angle grid, amounting to $\approx$ 151 million feature vectors in total. Features become sparse above 60° latitude, due to a lack of available imagery. Figure 1C shows individual images spanning 0.01° x 0.01°, along with their corresponding MOSAIKS feature values.

We create polygon-level feature vectors by averaging values across the feature tiles associated with each polygon. Each administrative polygon is represented by a single vector of 4000 daytime image features. We assign each feature tile to the administrative polygon that contains its centroid. For small municipal and DHS polygons that do not contain any tile centroids, we represent the polygon by the nearest feature tile. When averaging, we weight by population using data from the Gridded Population of the World V4 (GPW) (*28*). The GPW data product has global coverage at the 30 arcsecond (0.008°) resolution.

2.2 Nighttime light features

We create non-linear NL features from Defense Meteorological Satellite Program (DMSP) stable lights data (*30*). Specifically, we use an annual composite from the year 2013, the last year that composites are available for which sources of light contamination have been removed. The data has global coverage at 30 arcsecond (0.008°) resolution. We use DMSP data because it is widely used in the literature (*32–34*) and has a more uniform luminosity distribution than the more recent data from the Visible Infrared Imaging Radiometer Suite. DMSP luminosity values range from 0 to 63. We do not directly create random convolutional features from these NL data because the pixels are so large that each tile would not contain meaningful spatial structure. There are on average 1.5 NL pixels per tile.

Instead, we create features that flexibly characterize the distribution of the NL data using indicator variables that represent whether the luminosity value of each NL pixel falls into each of 20 bins. The 20 bins are comprised of a single bin at zero and 19 equally-spaced bins from 0 to 63. Due to the coarse resolution of the NL data, this basis captures similar information to what random convolutional features would capture if implemented, but is computationally simpler to implement. Analogous to aggregating the daytime imagery features to the polygon level, we calculate the population-weighted average NL luminosity value in each of the 20 bins for a given polygon. These polygon-level NL features denote the fraction of each polygon's population that is covered by nighttime light values represented by each of the 20 bins. This approach allows NL to associate non-linearly with the outcome variables in our linear models.

3 Analysis

3.1 General model specification

All models are trained at either the country or province level and use either the 4000 MOSAIKS daytime imagery features, the 20 NL features, or both. We train models using a five-fold cross-validation procedure with basic ridge hyper-parameter tuning. Data are split by country during cross-validation to account for spatial autocorrelation and to ensure that the model is predicting provincial outcomes when no observations from within the same country have been observed. We apply a clipping procedure that restricts model predictions to the minimum and maximum value observed in the training data. Hyper-parameter tuning is done with this clipping procedure. We allow for a different hyper-parameter between the MOSAIKS and NL feature sets, though this has only a minor impact on our results.

The general linearized model, representing Eq. 2, that we implement is

$$Y = \beta_0 + \boldsymbol{\beta}_1 \, \mathbf{X}_{MOSAIKS} + \boldsymbol{\beta}_2 \, \mathbf{X}_{NL} + \epsilon \tag{4}$$

Where Y is used to refer to the HDI, IWI, or NL labels interchangeably. We use this same model but predict each outcome separately. $\mathbf{X}_{MOSAIKS}$ is the matrix of daytime MOSAIKS features and $\mathbf{X}_{NL}$ is the matrix of nightlight features. We learn β_0, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ using ridge regression, following Rolf et al. (*13*). When predicting the NL outcome, we always exclude the $\mathbf{X}_{NL}$ feature matrix. For each outcome, we report performance for models trained at the country, province, and within-country levels (Table S1).

3.2 Performance metrics

For each model specification, we report two metrics, both of which are used in the literature. The *coefficient of determination* (R^2), used to evaluate related models by Chi *et al.* (*16*) and others, describes the accuracy of the raw model predictions and is a direct measure of model skill. The *square of the correlation coefficient* (ρ^2), used by Jean *et al.* (*22*) and others, scores performance after allowing model predictions to be linearly re-scaled before they are compared to observed values. We calculate both of these metrics when evaluating the full variation in labels and when decomposing the variation in labels into components that are visible within-countries or within-provinces (in contrast to between-countries and between-provinces).

Full variation performance The "full variation" performance metrics describe how well we estimate subnational HDI globally when we use all information available to us. To calculate full variation performance, we calculate ρ^2 and R^2 on the predicted and observed values of subnational HDI directly. This evaluates the ability of model predictions to capture the total variation in the observed values – i.e. variation across countries, across provinces within countries, and across municipalities or DHS clusters within provinces. Because most of the variation in HDI and other outcomes is between countries (Figure 2A-B), a large portion of the model's full variation performance comes directly from the mean-anchoring procedure (when it is used). Thus, the full variation performance metrics do not precisely evaluate the model's ability to predict local variation in isolation, and so they are not our preferred evaluation metric for understanding model performance within countries. Instead, we focus our analysis on within-country and, when applicable, within-province performance.

Within-country performance The "within-country" performance measures the amount of variation in the provincial deviations from the country mean that can be explained by the model. This metric evaluates the ability of the model to explain local variation in the outcome by removing large-scale variation in the outcome across countries in the demeaning step before predictions and observations are compared. To calculate within-country performance, we calculate ρ^2 and R^2 after demeaning predictions and observed values at the country level (i.e., after subtracting the predicted and observed country average value from each predicted and observed data point, respectively).

Within-province performance The "within-province" performance metric evaluates the ability of the model to explain hyper-local variation in the outcome, such as which DHS clusters within each province have higher or lower IWI. It does this by removing all between-province variation in the predicted and observed values before they are compared. To calculate within-province performance, we calculate ρ^2 and R^2 after demeaning predictions and observed values at the province level.

3.3 Model evaluation

Evaluation of HDI at the same provincial resolution as model training When reporting model performance at the same resolution as training (Figure 2 top row, Table S1 upper section, and Table S3), we evaluate predictions from the validation folds of the five-fold cross-validation procedure. This enables more observations to be used when evaluating model performance. We also evaluate models on a held-out test set of countries that were not included when tuning the HDI model. Before analysis, we set aside 20% of the provincial HDI data to be used as a final evaluation test set by randomly sampling 35 countries and their respective provinces. Evaluation on this test set was conducted after all hyper-parameter tuning and analysis decisions were made. We find that performance is not meaningfully different in the validation and tests sets, which indicates that the models evaluated on the validation folds did not over-fit to the data (Table S2).

Evaluation of HDI, IWI and NL at finer resolution than model training In the downscaling experiments, we evaluate performance using fine-resolution municipal or DHS cluster observations that were not used for model training or tuning. After tuning the model using cross-validation we retrain the model using the optimal hyper-parameters on all the ADM1 or ADM0 observations before predicting at downscaled resolution.

3.4 Mean anchoring

In our primary within-country model, we anchor our estimates to country or province-level means depending on the experiment. Estimates from models trained on provincial or national observations in "levels" (Equations 7,8) are never mean-anchored.

Anchoring to country means The procedure for anchoring to the country mean is illustrated in the top row of Figure 2, where we evaluate performance at the same resolution as model training. Our within-country model is trained to predict within-country anomalies, so in order to predict HDI in "levels" (Figure 2A), we add back the known country average

HDI (Equation 6). This procedure enables us to calculate full variation performance (Figure 2A) but has no impact on the reported within-country performance (Figure 2B).

Anchoring to provincial means In the downscaling application, our goal is to produce the best possible estimates at fine resolution. Thus, when producing downscaled estimates of IWI and HDI, we anchor our predictions to the observed provincial value of the outcome (Equation 12). This re-centering procedure impacts the full variation performance and the within-country performance but does not impact the within-province performance (Table S1). Note that we anchor municipal NL estimates to country means, rather than provincial means, because we find that this improves the estimates (Methods Section 3.7).

3.5 HDI model training

Within-country model training Because our focus is explaining subnational variation in HDI, we specifically train our primary model to predict within-country deviations of HDI. To do this, we first demean subnational observations by country and then train a model to use imagery to predict these residualized deviations. Specifically, we transform observed ADM1 HDI for province p (HDI_p^{ADM1}) into the deviation of this value from the country mean HDI ($\widetilde{HDI}_p^{ADM1}$). We then solve a ridge regression to predict $\widetilde{HDI}_p^{ADM1}$ based only on provincial daytime ($\widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1}$) and nightlight ($\widetilde{\mathrm{X}}_{NL,p}^{ADM1}$) features that have been similarly residualized relative to the country mean values for these variables. We learn the model

$$\widetilde{HDI}_p^{ADM1} = \beta_0 + \beta_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} + \beta_2 \, \widetilde{\mathrm{X}}_{NL,p}^{ADM1} + \epsilon_p \tag{5a}$$

where :

$$\widetilde{HDI}_p^{ADM1} = HDI_p^{ADM1} - \sum_{p \in c} \frac{HDI_p^{ADM1}}{N_c} \tag{5b}$$

$$\widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} = \mathrm{X}_{MOSAIKS,p}^{ADM1} - \sum_{p \in c} \frac{\mathrm{X}_{MOSAIKS,p}^{ADM1}}{N_c} \tag{5c}$$

$$\widetilde{\mathrm{X}}_{NL,p}^{ADM1} = \mathrm{X}_{NL,p}^{ADM1} - \sum_{p \in c} \frac{\mathrm{X}_{NL,p}^{ADM1}}{N_c}. \tag{5d}$$

Here, N_c is the number of provinces in country c. Note that we restrict predictions from this demeaned model to be between the observed minimum and maximum HDI deviations from the country mean.

Anchoring to country means via re-centering To evaluate full variation performance using the within-country model (Table S1, col. 1-2) we need HDI predictions in "levels" rather than predicted deviations from the country mean. To construct predicted HDI values in "levels" we anchor estimate to country means, since they are observed and used in the estimation procedure. Practically, this means we add the country mean HDI, which was subtracted from the observations before model training, back onto the predicted deviations:

$$\widehat{HDI}_p^{ADM1} = \widehat{\widetilde{HDI}}_p^{ADM1} + \sum_{p \in c} \frac{HDI_p^{ADM1}}{N_c} \tag{6}$$

Note that it is not necessary to implement this procedure when evaluating within-country performance.

Province and country model training In Table S1, we additionally report performance for models trained on province and country-level data directly. Unlike the within-country model, these models are trained on values in "levels" instead of deviations from the country mean. In these experiments, we learn the models:
Province model:

$$HDI_p^{ADM1} = \beta_0 + \beta_1 X_{MOSAIKS,p}^{ADM1} + \beta_2 X_{NL,p}^{ADM1} + \epsilon_p \tag{7}$$

Country model:

$$HDI_c^{ADM0} = \beta_0 + \beta_1 \ X_{MOSAIKS,c}^{ADM0} + \beta_2 \ X_{NL,c}^{ADM0} + \epsilon_c \tag{8}$$

We do not apply a mean-anchoring procedure with these models as their predictions are already in "levels" rather than predicted deviations. Note that 22 of the 179 total countries do not have subnational data (e.g., Qatar) and that these 22 country-only observations are included in both province and country models.

3.6 Downscaling validation with IWI

Labels IWI is similar to the wealth index reported in DHS surveys, except that it was created to be comparable across countries (*29*). IWI data are available both for provincial polygons, which we use for training, and for DHS clusters, which we use for evaluation. For each survey cluster, DHS provides coordinate points associated with the cluster centroid. To protect privacy, the actual GPS coordinates of the center of each cluster are randomly displaced by up to 2km for urban clusters and up to 5km for rural clusters, with a random 1% of rural cluster coordinates displaced by up to 10km. According to DHS, the displaced coordinate is guaranteed to fall within the same DHS-provided administrative boundaries as the true cluster centroid. To map these point observations to administrative polygons, we spatially buffer urban cluster coordinates using a 2km radius and rural cluster coordinates using a 10km radius. We then clip these buffers to the finest DHS-provided administrative boundaries that are available.

Training We train within-country, province level, and country level IWI models following the structure of models for HDI (Methods Section 3.5). Provincial IWI observations are denoted IWI_p^{ADM1}.

The within-country IWI model, our preferred model specification, takes the same form as Equation 5a:

$$\widetilde{IWI}_p^{ADM1} = \beta_0 + \beta_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} + \beta_2 \, \widetilde{\mathrm{X}}_{NL,p}^{ADM1} + \epsilon_p \tag{9a}$$

where :

$$\widetilde{IWI}_p^{ADM1} = IWI_p^{ADM1} - \sum_{p \in c} \frac{IWI_p^{ADM1}}{N_c} \tag{9b}$$

Note that $\widetilde{\mathrm{X}}_{MOSAIKS,i}^{ADM1}$ and $\widetilde{\mathrm{X}}_{NL,i}^{ADM1}$ are the same feature matrices defined in Equation 5c and 5d but with a different number of observations due to differing availability of outcome data.

Prediction We evaluate the IWI model performance at a finer resolution than it was trained. We use the trained provincial model (Equation 9a) to produce predictions of IWI at the DHS cluster level and compare those predictions to the cluster-level IWI measurements from the GDL, which were not used for model training. We calculate DHS cluster-level features in the same way as for the other administrative polygons.

To make predictions of IWI deviations from the country mean at the DHS cluster level using the within-country model trained on provincial deviations from the country mean, we multiply model weights with the demeaned DHS cluster-level satellite features:

$$\widetilde{\widehat{IWI}}_d^{DHS} = \hat{\beta}_0 + \hat{\beta}_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,d}^{DHS} + \hat{\beta}_2 \, \widetilde{\mathrm{X}}_{NL,d}^{DHS} \tag{10}$$

where d indexes DHS cluster and $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are estimated in Equation 9a. Tildes denote that these predictions are predicted deviations from the country mean. In our within-country IWI model, we demean DHS cluster-level satellite image features by the same country average feature values as in the training procedure:

$$\widetilde{\mathrm{X}}_{MOSAIKS,d}^{DHS} = \mathrm{X}_{MOSAIKS,d}^{DHS} - \sum_{p \in c} \frac{\mathrm{X}_{MOSAIKS,p}^{ADM1}}{N_c} \tag{11a}$$

$$\widetilde{\mathrm{X}}_{NL,d}^{DHS} = \mathrm{X}_{NL,d}^{DHS} - \sum_{p \in c} \frac{\mathrm{X}_{NL,p}^{ADM1}}{N_c} \tag{11b}$$

where we note that these averages are constructed by averaging province-level features, but have similar values that averages of nationally-representative sets of cluster-level features would have.

Anchoring to provincial means via re-centering To construct estimates of cluster-level IWI in levels ($\widehat{IWI}_d^{DHS}$), we anchor predicted cluster-level deviations from the country mean ($\widetilde{\widehat{IWI}}_d^{DHS}$) to the known provincial value (IWI_p^{ADM1}) using a provincial level adjustment:

$$\widehat{IWI}_d^{DHS} = \widetilde{\widehat{IWI}}_d^{DHS} + \underbrace{IWI_p^{ADM1} - \sum_{d \in p} \frac{\widetilde{\widehat{IWI}}_d^{DHS}}{N_p}}_{\text{centers DHS clusters to known provincial values}} \tag{12}$$

Here, N_p denotes the number of DHS clusters contained by ADM1 polygon p, and IWI_p^{ADM1} denotes the observed ADM1-level value for polygon p. This anchors the mean of our DHS cluster-level predictions within each provincial polygon to the respective known province value used in training.

3.7 Downscaling validation using nighttime lights as labels

In our analysis of the downscaling perfomance of our approach, we design an experiment in which NL are used as *labels* and are *not used as features* (Figure S1). This experiment is useful because it is the only validation experiment where the groundtruth data are available globally and at municipal resolution. Thus, this experiment allows us to evaluate predictions at a downscaled resolution for the entire globe using a procedure that mirrors how we will generate downscaled HDI estimates (such global high resolution labels do not exist for our other outcomes). We do not expect NL predictions to be perfect proxies for HDI data in this regard, but if NL can be downscaled successfully, it provides support for the *procedure* we use to downscale HDI.

Labels We use population estimates from GPW and fine resolution NL data from DMSP to create a population-weighted average NL luminosity at the province level. NL observations are population-weighted to mirror the construction of HDI, which is also population-weighted. We construct municipality-level NL observations using a municipal (ADM2) shapefile from geoBoundaries (*43*), which links municipalities to provincial "parent" polygons.

We exclude Ireland from the geoBoundaries ADM2 dataset because Irish municipalities (ADM2 units) are so small that they alone represent 45% of the global municipality observations. Thus, they would be over-represented in global performance metrics relative to their size if not removed. Still, we find similar performance to the global results when evaluating downscaled NL for Ireland. Full variation $R^2 = 0.68$ and within-province $R^2 = 0.38$ when predicting NL for Ireland's very small municipality units. The vertical and horizontal streaking patterns in the scatter plots in Figure 2 are caused by three other countries that also have very spatially dense municipalities, though not to the same degree as Ireland. This, in turn, leads them to have very similar true and predicted values, which is compounded by the bunching of nightlight values at the maximum and minimum of the sensor range. The countries are Great Britain ($\approx$ 9,000 units), Spain ($\approx$ 8,000 units), and Brazil ($\approx$ 5,000 units).

Training We train a model using only MOSAIKS features constructed from daytime imagery to predict NL:

$$NL = \beta_0 + \boldsymbol{\beta}_1 \, \mathbf{X}_{MOSAIKS} + \epsilon \tag{13}$$

This model structure is broadly the same training procedure described in Methods Section 3.5 and in Equation 4; however, we do not include NL features when predicting average NL luminosity. NL is also now a vector of scalar NL observations rather than a matrix of features.

Prediction To generate municipal predictions, indexed by m, from the within-country model, we first create municipal predictions of NL deviations from the country mean. We demean $X^{ADM2}_{MOSAIKS,m}$ by country by subtracting the country mean feature values and then multiplying the resulting demeaned features by the estimated model weights. This corresponds to what is done when evaluating downscaled IWI performance in Section 3.6 and shown in Equation 11a.

Anchoring to country means via re-centering When converting the predicted municipal NL deviations from the country mean ($\widehat{\widetilde{NL_m^{ADM2}}}$) into predicted municipal NL values in levels ($\widehat{NL_m^{ADM2}}$), we anchor values to the known country mean:

$$\widehat{NL_m^{ADM2}} = \widehat{\widetilde{NL_m^{ADM2}}} + \sum_{p \in c} \frac{NL_p^{ADM1}}{N_c} \tag{14}$$

Note that we anchor fine resolution NL predictions to the known country mean rather than the provincial mean (following Equation 6 rather than Equation 12) because we find that this substantially improves full variation performance. Most of the variation in nightlight luminosity occurs within countries, rather than between countries, which is considerably different from what we observe for HDI and IWI. Importantly, the choice to use a different re-centering procedure for NL does not impact the downscaled within-province performance (Figure 2J), which we believe provides the most important evaluation of downscaling performance. After re-centering, 20% of downscaled NL predictions are outside of range of valid values for the DMSP nightlight raster. We winsorize these values to the limits of the allowable range prior to evaluating performance.

3.8 Producing downscaled estimates of HDI

Municipality-level HDI We follow the downscaling approach for IWI described in Section 3.6, but adjusted for HDI data, to produce municipality (ADM2) level estimates of HDI. We use the within-country model specified in Equation 5a to estimate HDI at the municipality level, using a municipality (ADM2) shapefile from geoBoundaries (*43*). We anchor municipality estimates by centering predicted deviations on the observed province-level HDI value, identical to the procedure for downscaled IWI predictions (Equation 12). We do not release HDI estimates for municipalities that cannot be linked to a parent province with a province-level HDI estimate from Smits and Permanyer (*7*) because there is not a known provincial value to anchor on.

Grid-level HDI To produce $0.1° \times 0.1°$ estimates of HDI, we similarly use the within-country model specified in Equation 5a. However, we make predictions at the native resolution of the MOSAIKS features ($0.01° \times 0.01°$). This results in gridded estimates of HDI at approximately 1km^2 resolution. We mask out locations where humans are not believed to be settled based on the Global Human Settlement Layer (*36*) (keeping areas with population > 0) and then mean-anchor our tile estimates such that the population-weighted average of the grid tiles within each province matches known provincial HDI values. Population weights are taken from GPW. Finally, we down-sample predictions to a $0.1° \times 0.1°$ grid to reduce noise and more closely match the spatial scale of the DHS clusters used in our finest-resolution downscaling validation experiment.

3.9 Downscaling validation with HDI data in Mexico

We compare our municipal HDI estimates with census-derived municipal estimates for HDI, which are available in Mexico for the year 2010 (*10*). The census-derived data have a different mean than the 2018 HDI data that we use elsewhere in this analysis; though, this has no influence on the within-country or within-province evaluation metrics (Figure 2 C,D) because predictions and observations are demeaned at the country and province level, respectively, before the metrics are calculated (Methods 3.2).

Code availability Replication code is available at `github.com/lukesherman/hdi_downscaling_mosaiks`.

Data availability All data used in this analysis, other than the DHS cluster-level IWI data from the Global Data Lab, is from free, publicly available sources. Details on how to access data for replication can be found at `github.com/lukesherman/hdi_downscaling_mosaiks`. HDI estimates are available at `mosaiks.org/hdi`.

Funding This work was supported by a grant from the Human Development Report Office of the United Nations Development Programme.

Acknowledgements We thank Pedro Conceição and seminar participants at The Workshop in Environmental Economics and Data Science, the WIDER Development Conference (Bogotá), IDinsight, and the American Geophysical Union Fall Meeting for their valuable feedback. We thank the Global Data Lab for sharing DHS cluster-level data on the International Wealth Index.

References

1. P. Conceição, M. Kovacevic, T. Mukhopadhyay, *Human Development: A Perspective on Metrics*, pp. 83–115, ISBN: 9780128190579.

2. I. Permanyer, J. Smits, *Population and Development Review*, ISSN: 17284457.

3. J. Klugman, F. Rodríguez, H. J. Choi, The HDI 2010: New controversies, old critiques. *Journal of Economic Inequality* **9**, 249–288, ISSN: 15691721 (2011).

4. *Technical notes - Human Development Report 2018*, 2018.

5. "Human Development Report 1990: Concept and Measurement of Human Development", tech. rep. (United Nations Development Programme, New York, 1990).

6. H. Wolff, H. Chong, M. Auffhammer, Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index. *Economic Journal* **121**, 843–870, ISSN: 00130133 (2011).

7. J. Smits, I. Permanyer, The subnational human development database. *Scientific Data* **6**, 1–15, ISSN: 20524463, (`https://doi.org/10.1038/sdata.2019.38`) (Mar. 2019).

8. M. Kummu, M. Taka, J. H. Guillaume, Gridded global datasets for gross domestic product and Human Development Index over 1990–2015. *Scientific Data* **5**, 1–15 (2018).

9. I. S. Smythe, J. E. Blumenstock, Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences* **119**, e2120025119 (2022).

10. I. Permanyer, Using Census Data to Explore the Spatial Distribution of Human Development. *World Development* **46**, 1–13, ISSN: 0305-750X (June 2013).

11. Carleton, Chong, Druckenmiller, Noda, Proctor, Rolf and Hsiang, *Multi-Task Observation Using Satellite Imagery and Kitchen Sinks (MOSAIKS) API*, `https://mosaiks.org`, version 1.0, 2022.

12. Planet Team, *Planet Application Program Interface: In Space for Life on Earth*, Planet, 2017, (`https://api.planet.com`).

13. E. Rolf *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 2021 12:1* **12**, 1–11, ISSN: 2041-1723 (July 2021).

14. M. Burke, A. Driscoll, D. B. Lobell, S. Ermon, *Using satellite imagery to understand and promote sustainable development*, Mar. 2021, arXiv: 2010.06988.

15. C. Yeh *et al.*, Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* **11**, 1–11, ISSN: 20411723 (Dec. 2020).

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

17. A. Khachiyan *et al.*, Geographic microtargeting of social assistance with high-resolution poverty maps. *American Economic Review: Insights* **4**, 491–506 (2022).

18. N. Ratledge, G. Cadamuro, B. de la Cuesta, M. Stigler, M. Burke, Using machine learning to assess the livelihood impact of electricity access. *Nature* **611**, 491–495 (2022).

19. J. Proctor, T. Carleton, S. Sum, "Parameter Recovery Using Remotely Sensed Variables", tech. rep. (National Bureau of Economic Research, 2023).

20. E. Aiken, S. Bellue, D. Karlan, C. Udry, J. E. Blumenstock, Machine learning and phone data can improve targeting of humanitarian aid. *Nature* **603**, 864–870 (2022).

21. L. Y. Huang, S. M. Hsiang, M. Gonzalez-Navarro, "Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs", tech. rep. (National Bureau of Economic Research, 2021).

22. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794, ISSN: 10959203 (Aug. 2016).

23. Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21 (2021).

24. N. Jean *et al.*,
Tile2vec: Unsupervised representation learning for spatially distributed data.
Proceedings of the AAAI Conference on Artificial Intelligence **33**, 3967–3974 (2019).

25. S. Mohanasundaram, K. Kasiviswanathan, C. Purnanjali, I. P. Santikayasa, S. Singh,
Downscaling Global Gridded Crop Yield Data Products and Crop Water Productivity Mapping Using Remote Sensing Derived Variables in the South Asia.
International Journal of Plant Production, 1–16 (2022).

26. P. Rayner, M. Raupach, M. Paget, P. Peylin, E. Koffi, A new global gridded data set of CO2 emissions from fossil fuel combustion: Methodology and evaluation.
Journal of Geophysical Research: Atmospheres **115** (2010).

27. A. Rahimi, B. Recht, Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning.
Advances in neural information processing systems **21** (2008).

28. U. Center for International Earth Science Information Network (CIESIN) Columbia,
Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 10, Palisades, NY, 2018, (`https://doi.org/10.7927/H49C6VHW`).

29. J. Smits, R. Steendijk, The International Wealth Index (IWI).
Social Indicators Research 2014 122:1 **122**, 65–85, ISSN: 1573-0921 (July 2014).

30. *Version 4 DMSP-OLS Nighttime Lights Time Series*,
(`https://eogdata.mines.edu/products/dmsp/%5C#v4%5C_dmsp%5C_download`).

31. X. Chen, W. D. Nordhaus, Using luminosity data as a proxy for economic statistics.
Proceedings of the National Academy of Sciences **108**, 8589–8594, ISSN: 0027-8424 (May 2011).

32. J. V. Henderson, A. Storeygard, D. N. Weil,
Measuring Economic Growth from Outer Space.
American Economic Review **102**, 994–1028 (Apr. 2012).

33. A. Bruederle, R. Hodler,
Nighttime lights as a proxy for human development at the local level.
PLOS ONE **13**, 1–22 (Sept. 2018).

34. R. Bluhm, G. C. McCord, What Can We Learn from Nighttime Lights for Small Geographies? Measurement Errors and Heterogeneous Elasticities.
Remote Sensing **14**, ISSN: 2072-4292 (2022).

35. *UNDP Human Development Data Center*, 2018 data,
(`http://hdr.undp.org/en/data`).

36. *Global Human Settlement Layer - Population (GHS POP E2020)*,
(https://ghsl.jrc.ec.europa.eu/download.php?ds=pop).

37. C. Elbers, T. Fujii, P. Lanjouw, B. Özler, W. Yin,
Poverty alleviation through geographic targeting: How much does disaggregation help?
Journal of Development Economics **83**, 198–213 (2007).

38. M. Ravallion,
Poverty alleviation through regional targeting: A case study for Indonesia.
The Economics of Rural Organization: Theory, Practice and Policy, 373–77 (1993).

39. D. P. Coady, The welfare returns to finer targeting: The case of the PROGRESA program in Mexico. *International tax and public finance* **13**, 217–239 (2006).

40. R. Hanna, B. A. Olken, Universal basic incomes versus targeted transfers:
Anti-poverty programs in developing countries.
Journal of Economic Perspectives **32**, 201–26 (2018).

41. C. Yeh *et al.*, SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. arXiv: 2111.04724 (Nov. 2021).

42. *Subnational-HDI Database V4.0*,
(https://globaldatalab.org/shdi/download_files/).

43. D. Runfola *et al.*,
geoBoundaries: A global database of political administrative boundaries.
PLOS ONE **15**, e0231866, ISSN: 1932-6203 (Apr. 2020).

Supplementary Data

Supplementary tables

HDI trained at:	**Features**	**Predicted at province level** (n=1,363) *Full variation performance* ρ^2 (1)	R^2 (2)	*Within-country performance* ρ^2 (3)	R^2 (4)
Within-country (n=1,344)	**MOSAIKS+NL**	0.96	0.96	0.46	0.46
	MOSAIKS	0.95	0.95	0.35	0.34
	NL	0.95	0.95	0.39	0.38
Province level (n=1,363)	**MOSAIKS+NL**	0.79	0.79	0.36	0.08
	MOSAIKS	0.72	0.72	0.23	< 0
	NL	0.58	0.58	0.38	< 0
Country level (n=144)	**MOSAIKS+NL**	0.71	0.69	0.4	< 0
	MOSAIKS	0.6	0.58	0.15	< 0
	NL	0.59	0.5	0.38	< 0

HDI trained at:	**Features**	**Predicted at municipality level in Mexico** (n=2,457) *Within-country performance* ρ^2 (3)	R^2 (4)	*Within-province performance* ρ^2 (5)	R^2 (6)
Within-country (n=1,344)	**MOSAIKS+NL**	0.43	0.4	0.23	0.23
	MOSAIKS	0.48	0.44	0.3	0.29
	NL	0.5	0.43	0.32	0.27
Province level (n=1,363)	**MOSAIKS+NL**	0.07	< 0	0.05	< 0
	MOSAIKS	0.12	< 0	0.11	< 0
	NL	0.2	< 0	0.19	< 0
Country level (n=144)	**MOSAIKS+NL**	0.19	< 0	0.15	< 0
	MOSAIKS	0.11	< 0	0.08	< 0
	NL	0.21	< 0	0.19	< 0

IWI trained at:	**Features**	**Predicted at DHS cluster level** (n=51,996) *Full variation performance* ρ^2 (1)	R^2 (2)	*Within-country performance* ρ^2 (3)	R^2 (4)	*Within-province performance* ρ^2 (5)	R^2 (6)
Within-country (n=863)	**MOSAIKS+NL**	0.75	0.75	0.59	0.59	0.39	0.38
	MOSAIKS	0.69	0.68	0.48	0.47	0.22	0.2
	NL	0.76	0.76	0.59	0.59	0.39	0.39
Province level (n=864)	**MOSAIKS+NL**	0.5	0.38	0.3	0.19	0.19	< 0
	MOSAIKS	0.37	0.31	0.14	< 0	0.08	< 0
	NL	0.4	< 0	0.3	0.02	0.27	< 0
Country level (n=86)	**MOSAIKS+NL**	0.41	< 0	0.3	< 0	0.25	< 0
	MOSAIKS	0.27	0.09	0.12	< 0	0.08	< 0
	NL	0.42	< 0	0.33	< 0	0.29	< 0

NL trained at	**Features**	**Predicted at municipality level** (n=62,487) *Full variation performance* ρ^2 (1)	R^2 (2)	*Within-country performance* ρ^2 (3)	R^2 (4)	*Within-province performance* ρ^2 (5)	R^2 (6)
Within-country (n=2,848)	**MOSAIKS**	0.79	0.78	0.66	0.65	0.61	0.59
Province level (n=2,848)	**MOSAIKS**	0.76	0.76	0.63	0.61	0.58	0.55
Country level (n=170)	**MOSAIKS**	0.6	0.58	0.44	0.38	0.4	0.3

Table S1: Performance for models trained to predict HDI, IWI, and population-weighted nightlight luminosity (NL). We show performance evaluated at the province level for HDI and evaluate downscaled performance for HDI in Mexico, IWI, and NL. Performance scatters from the within-country models are shown in Figure 2.

		Predicted at province level (n=366)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
HDI trained at:	***Features***	(1)	(2)	(3)	(4)
Within-country	**MOSAIKS+NL**	0.96	0.96	0.41	0.39
	MOSAIKS	0.96	0.96	0.26	0.25
	NL	0.96	0.96	0.37	0.37
Province level	**MOSAIKS+NL**	0.79	0.79	0.38	< 0
	MOSAIKS	0.79	0.77	0.24	< 0
	NL	0.65	0.63	0.35	< 0
Country level	**MOSAIKS+NL**	0.73	0.7	0.36	< 0
	MOSAIKS	0.71	0.67	0.11	< 0
	NL	0.65	0.62	0.34	< 0

Table S2: This is similar to the upper portion of Table S1 except that here we have evaluated on a 35 country ($\approx$ 20%) test set that was not used during model tuning. There is only a slight decline in within-country performance as compared to Table S1.

Life expectancy trained at:	Features	Predicted at province level (n=1,363) Full variation performance ρ^2 (1)	Full variation performance R^2 (2)	Within-country performance ρ^2 (3)	Within-country performance R^2 (4)
Within-country	**MOSAIKS+NL**	0.93	0.93	0.07	0.07
	MOSAIKS	0.93	0.93	0.02	0.02
	NL	0.93	0.93	0.07	0.07
Province level	**MOSAIKS+NL**	0.68	0.68	0.03	< 0
	MOSAIKS	0.66	0.66	0.02	< 0
	NL	0.42	0.42	0.07	< 0
Country level	**MOSAIKS+NL**	0.54	0.47	0.07	< 0
	MOSAIKS	0.49	0.46	0.02	< 0
	NL	0.44	0.36	0.08	< 0

Mean years schooling trained at:	Features	Predicted at province level (n=1,363) Full variation performance ρ^2 (1)	Full variation performance R^2 (2)	Within-country performance ρ^2 (3)	Within-country performance R^2 (4)
Within-country	**MOSAIKS+NL**	0.93	0.93	0.48	0.48
	MOSAIKS	0.91	0.91	0.37	0.37
	NL	0.92	0.92	0.39	0.38
Province level	**MOSAIKS+NL**	0.69	0.69	0.33	0.22
	MOSAIKS	0.63	0.62	0.23	< 0
	NL	0.52	0.52	0.37	0.16
Country level	**MOSAIKS+NL**	0.65	0.63	0.32	< 0
	MOSAIKS	0.54	0.51	0.1	< 0
	NL	0.53	0.52	0.36	< 0

Expected years schooling trained at:	Features	Predicted at province level (n=1,363) Full variation performance ρ^2 (1)	Full variation performance R^2 (2)	Within-country performance ρ^2 (3)	Within-country performance R^2 (4)
Within-country	**MOSAIKS+NL**	0.9	0.9	0.23	0.22
	MOSAIKS	0.89	0.89	0.19	0.19
	NL	0.89	0.89	0.13	0.13
Province level	**MOSAIKS+NL**	0.58	0.58	0.17	< 0
	MOSAIKS	0.56	0.56	0.16	< 0
	NL	0.42	0.42	0.14	< 0
Country level	**MOSAIKS+NL**	0.52	0.46	0.15	< 0
	MOSAIKS	0.49	0.46	0.12	< 0
	NL	0.42	0.37	0.14	< 0

GNIpc trained at:	Features	Predicted at province level (n=1,363) Full variation performance ρ^2 (1)	Full variation performance R^2 (2)	Within-country performance ρ^2 (3)	Within-country performance R^2 (4)
Within-country	**MOSAIKS+NL**	0.96	0.96	0.46	0.46
	MOSAIKS	0.95	0.95	0.31	0.31
	NL	0.96	0.96	0.45	0.45
Province level	**MOSAIKS+NL**	0.71	0.71	0.38	0.03
	MOSAIKS	0.62	0.61	0.18	< 0
	NL	0.56	0.56	0.43	< 0
Country level	**MOSAIKS+NL**	0.54	< 0	0.17	0.08
	MOSAIKS	0.42	< 0	0.06	0.04
	NL	0.4	< 0	0.16	0.09

Table S3: This is similar to the top section of Table S1 except that here we show performance for each HDI component evaluated at the province level.

	HDI	Life expectancy	Mean years schooling	Expected years schooling
HDI				
Life expectancy	0.79			
Mean years schooling	0.84	0.54		
Expected years schooling	0.83	0.6	0.62	
GNIpc	0.63	0.44	0.51	0.46
Within-country	**HDI**	**Life expectancy**	**Mean years schooling**	**Expected years schooling**
HDI				
Life expectancy	0.32			
Mean years schooling	0.83	0.14		
Expected years schooling	0.65	0.11	0.46	
GNIpc	0.2	0.04	0.11	0.1

Table S4: Individual components of HDI tend to be correlated. We report the squared Pearson's correlation coefficient (ρ^2) between HDI and its components at the province level. We also report the squared correlation coefficients after demeaning provincial observations by country. This ρ^2 metric used here is intended to be comparable to the metrics reported in Tables S1 and S3. Notably, within-country correlation between HDI and GNIpc is low, yet we are still able to predict those separate outcomes with considerable skill using MOSAIKS.

Supplementary figures

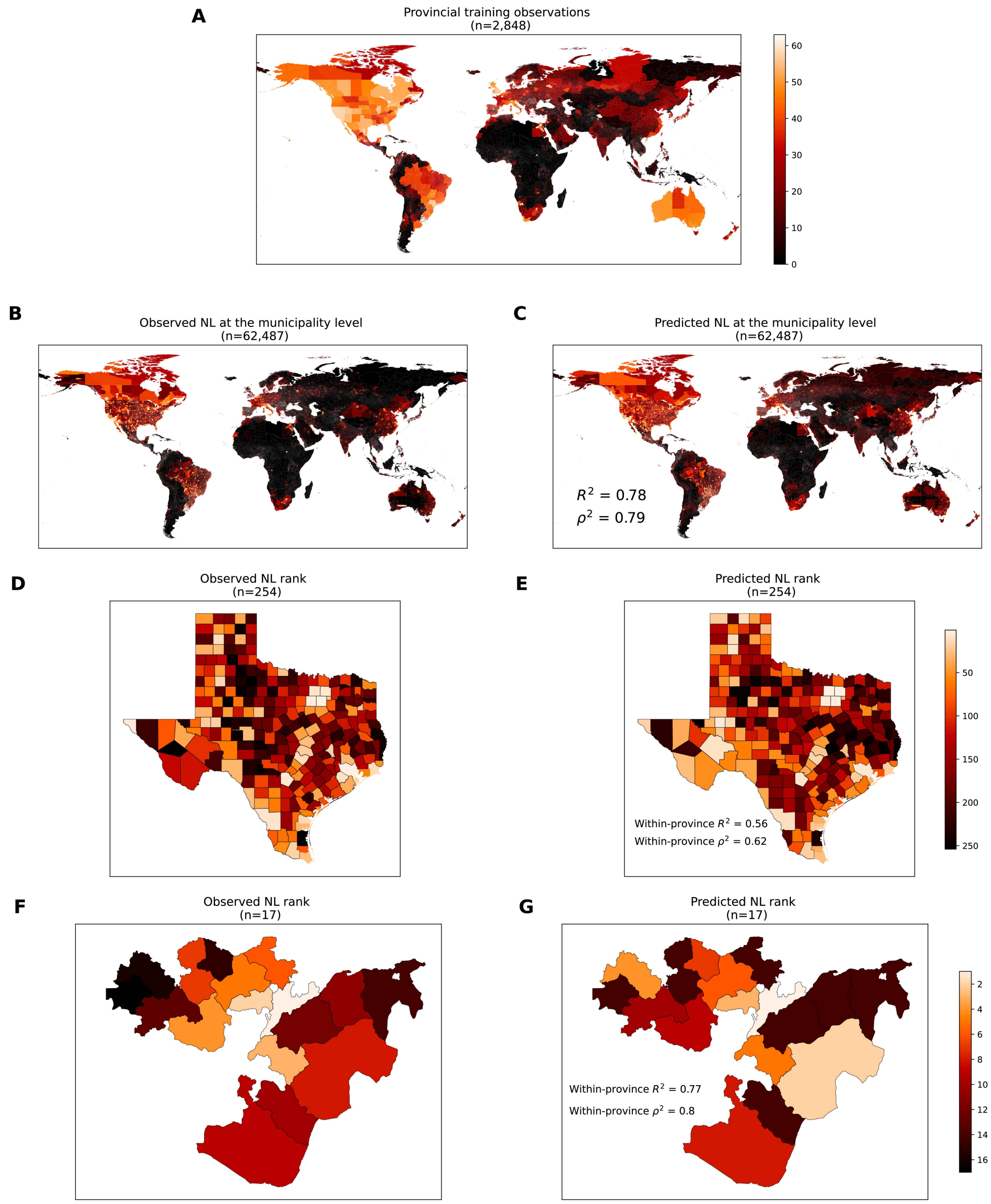

Figure S1: **A MOSAIKS model trained at the province level can effectively predict NL at the municipality level.** These maps show population-weighted NL luminosity that has been predicted using MOSAIKS. (**A**) Population-weighted NL averaged up to the provincial polygon. These are the data used to train the model. (**B**) True population-weighted NL at the municipality level. (**C**) Predicted population-weighted NL at the municipality level (country mean added back). (**D**) Municipalities ranked by luminosity within Texas, a single province in the United States. (**E**) Predicted nightlight luminosity rank within Texas. (**F**) Municipalities ranked by luminosity within Oromia, a single province in Ethiopia. (**G**) Predicted nightlight luminosity rank within Oromia. Panels D-G illustrate the downscaling efficacy of MOSAIKS. Each of these polygons (Texas and Oromia) represent a single training observation. All predictions come from a within-country model (country mean added back). Note that panels A-C use the same colorbar. See Table S1 for detailed performance metrics.

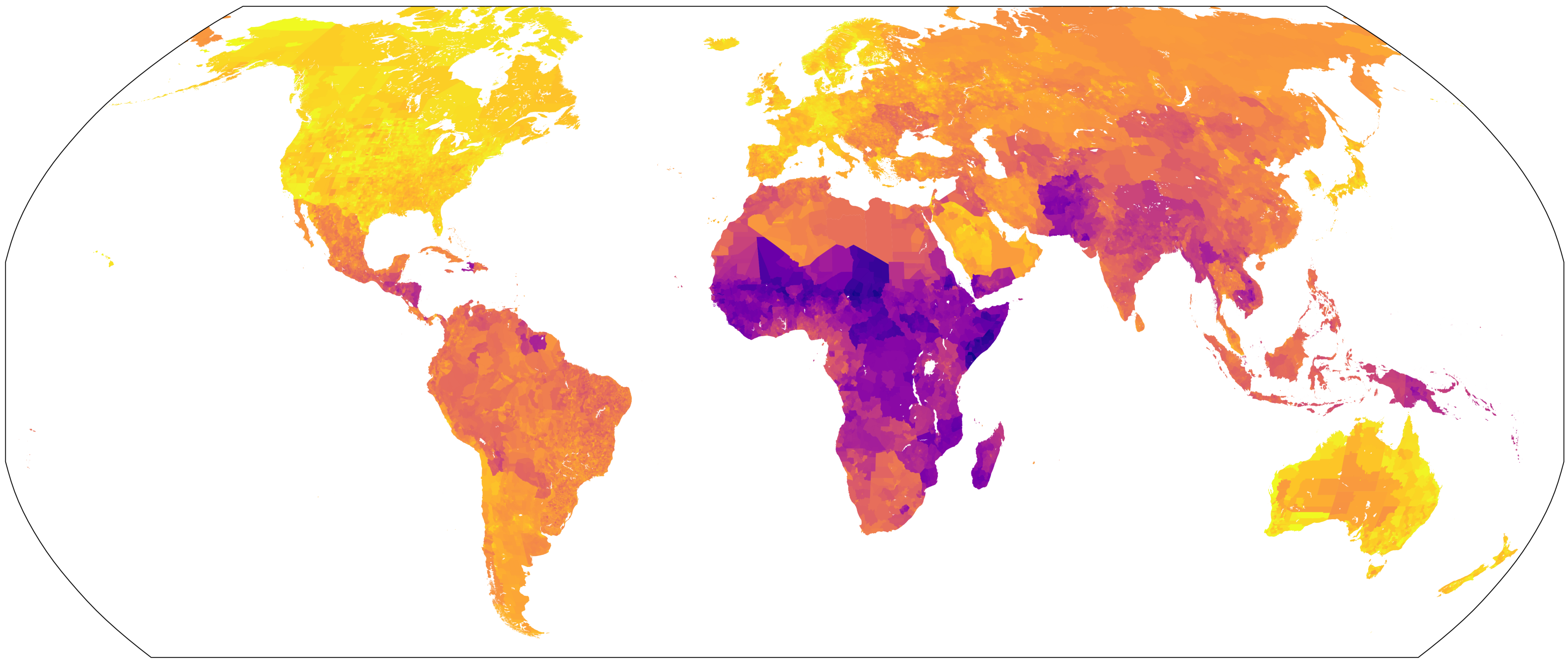

Figure S2: **Full page map of HDI estimates at the municipal level.** This is the same data as shown in Figure 3C.

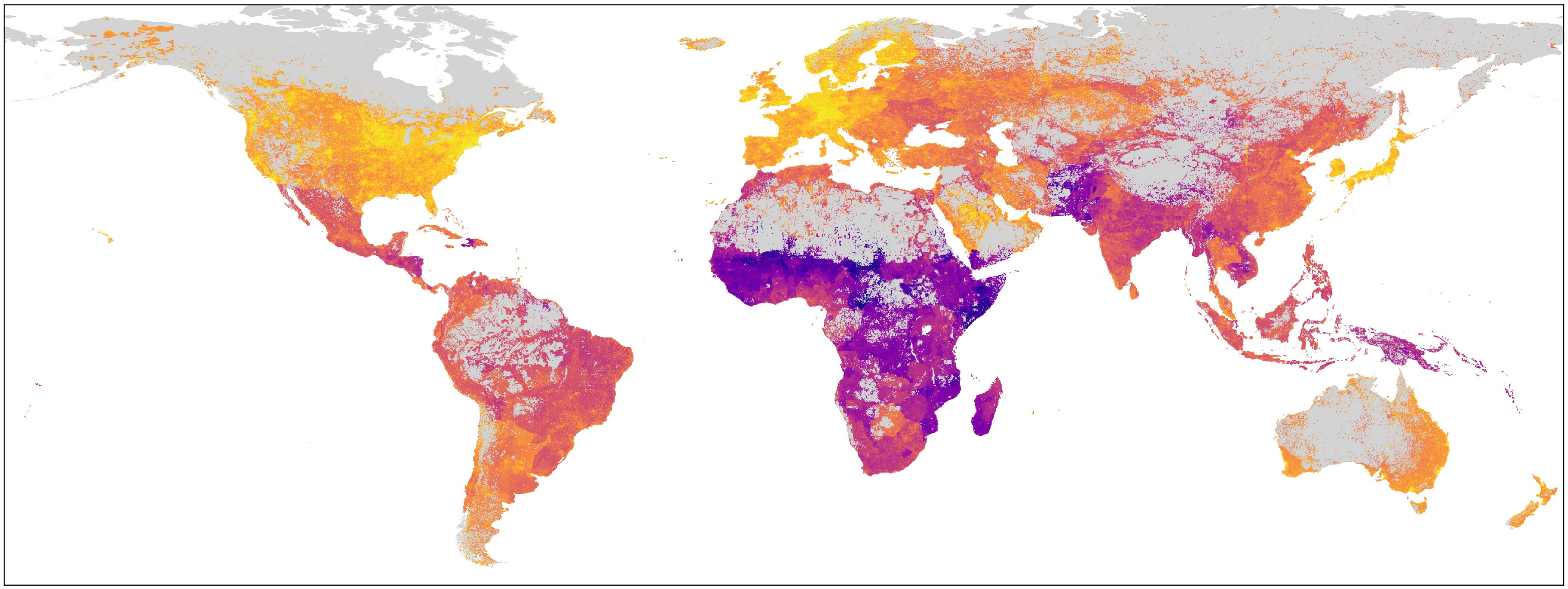

Figure S3: **Full page map of HDI estimates at the grid level.** This is the same data as shown in Figure 3D.

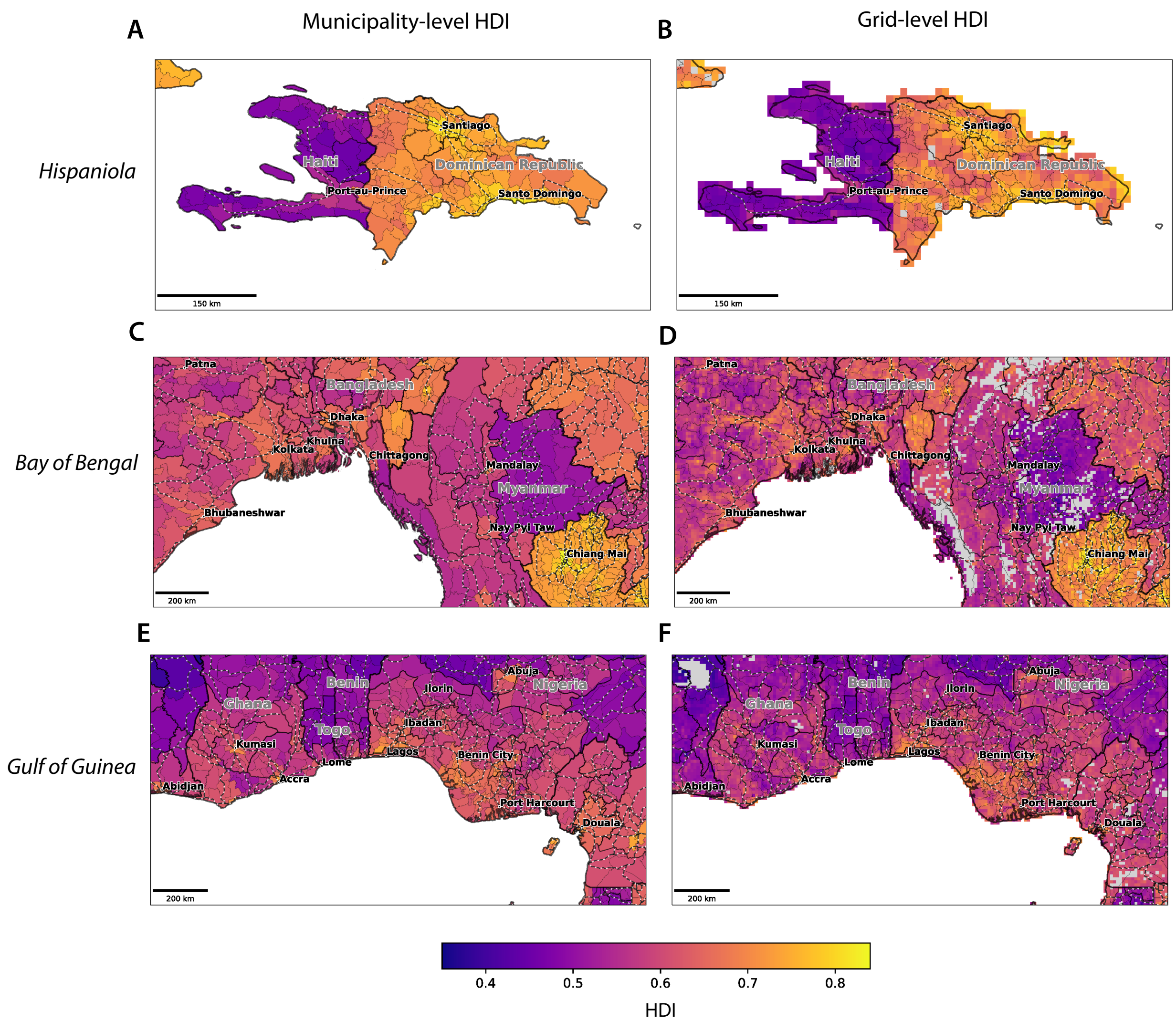

Figure S4: **Regional maps of HDI estimates at the municipal and grid levels.** (**A-B**) HDI estimates on Hispaniola (**C-D**) HDI estimates around the Bay of Bengal (**E-F**) HDI estimates around the Gulf of Guinea. All panels show country, province, and municipality borders as solid lines. Dashed lines show major roadways. Grey in the grid-level estimates indicates land area believed to be unsettled (*36*).

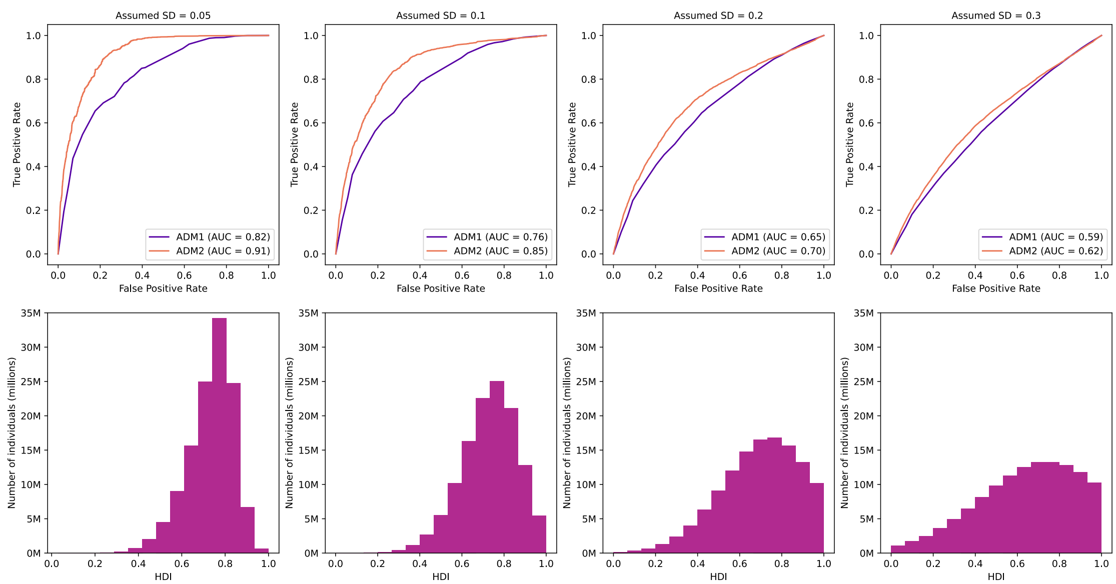

Figure S5: **The improvement in geographic targeting efficacy from using municipal values (ADM2) depends on the assumed variability of individual-level HDI within municipalities**. ROC curves as in Figure 5F for different assumed standard deviations (SD) of individual-level HDI within municipalities. Using municipal instead of provincial HDI estimates increases the AUC by 0.09 (+11% from 0.82 to 0.91) when the within-municipality HDI standard deviation is assumed to be 0.05 and by 0.05 (+8% from 0.65 to 0.7) when it is assumed to be 0.2. Histograms show the distribution of simulated individual-level HDI for each assumed SD, using a truncated normal distribution centered on the ADM2 values calculated by Permanyer (*10*).

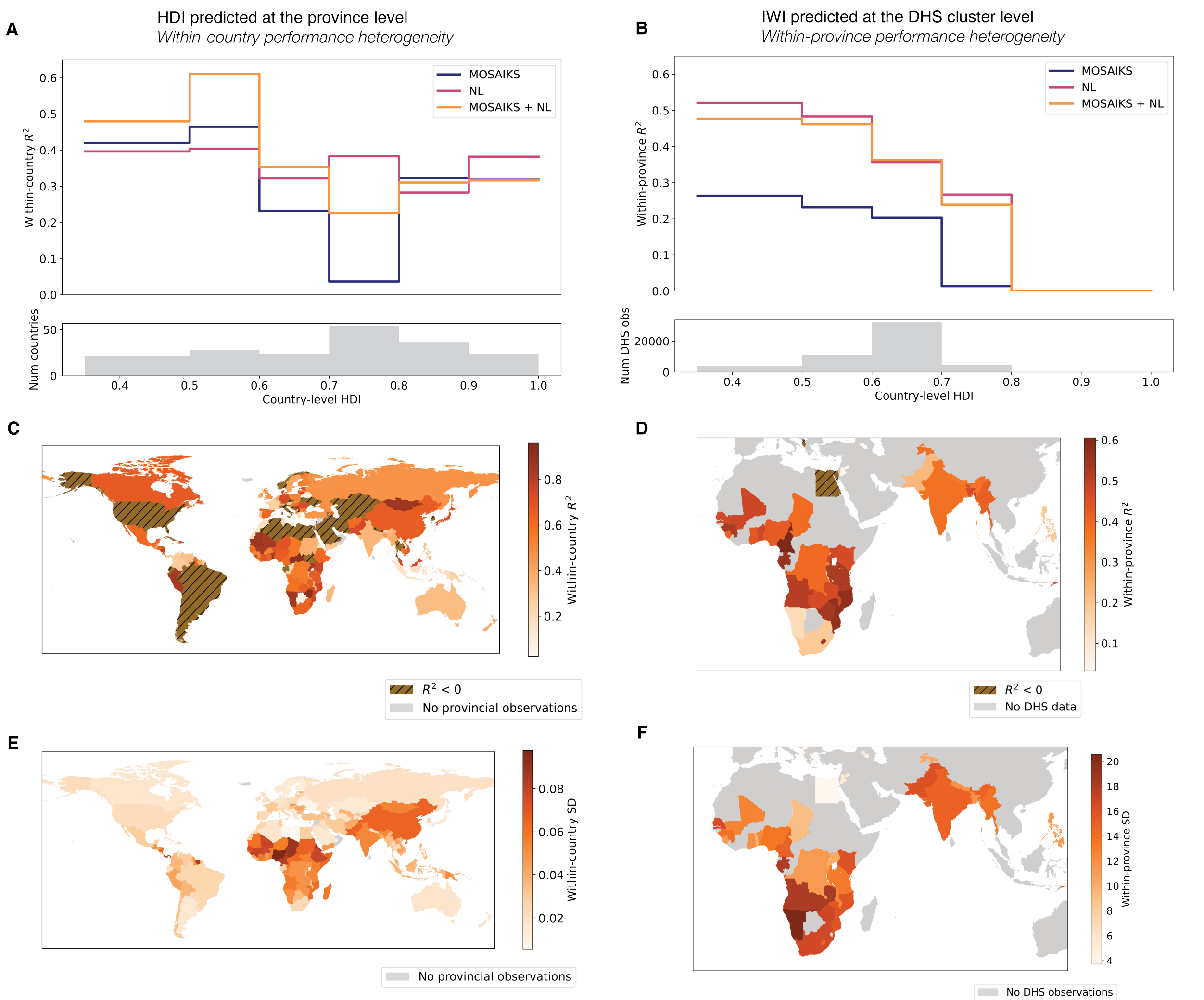

Figure S6: **Heterogeneity of HDI and IWI predictions.** On the left we show heterogeneity in HDI performance evaluated at the province level. On the right, we show heterogeneity in IWI performance evaluated at the DHS cluster level. (**A**) HDI performance as a function of parent country HDI. (**B**) IWI performance as a function of parent country HDI.(**C**) Mapped performance of HDI within-countries (within-country MOSAIKS + NL model). (**D**) Mapped performance of IWI within-provinces (within-country MOSAIKS + NL model). (**E**) Standard deviation of provincial HDI by country (**F**) Standard deviation of DHS cluster-level IWI within-provinces by country.