RESEARCH ARTICLES

ECONOMICS

Combining satellite imagery and machine learning to predict poverty

Neal Jean,[1,2*] Marshall Burke,[3,4,5*†] Michael Xie,[1] W. Matthew Davis,[4] David B. Lobell,[3,4] Stefano Ermon[1]

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.

Accurate measurements of the economic characteristics of populations critically influence both research and policy. Such measurements shape decisions by individual governments about how to allocate scarce resources and provide the foundation for global efforts to understand and track progress toward improving human livelihoods. Although the quantity and quality of economic data available in developing countries have improved in recent years, data on key measures of economic development are still lacking for much of the developing world (*1*). This data gap is hampering efforts to identify and understand variation in these outcomes and to target intervention effectively to areas of greatest need (*2*, *3*).

Data gaps on the African continent are particularly constraining. According to World Bank data, during the years 2000 to 2010, 39 of 59 African countries conducted fewer than two surveys from which nationally representative poverty measures could be constructed. Of these countries, 14 conducted no such surveys during this period (*4*) (Fig. 1A), and most of the data from conducted surveys are not in the public domain. Coverage is similarly limited for the Demographic and Health Surveys (DHS), the primary source for population-level health statistics in most developing countries as well as for internationally comparable data on household assets—a common measure of wealth (Fig. 1B). For the same 11-year period, 20 of the 59 countries had no DHS asset-based surveys taken, and an additional 19 had only one. These shortcomings have prompted calls for a "data revolution" to sharply scale up data collection efforts within Africa and elsewhere (*1*). But closing these data gaps with more frequent household surveys is likely to be both prohibitively costly—perhaps costing hundreds of billions of U.S. dollars to measure every target of the United Nations Sustainable Development Goals in every country over a 15-year period (*5*)—and institutionally difficult, as some governments see little benefit in having their lackluster performance documented (*2*, *6*).

Given the difficulties of scaling up traditional data collection efforts, an alternative path to measuring these outcomes might use novel sources of passively collected data, such as data from social media, mobile phone networks, or satellites. A popular recent approach leverages satellite images of luminosity at night ("nightlights") to estimate economic activity (*7*–*10*). While this particular technique has shown promise in improving existing country-level economic production statistics (*7*, *10*), it appears less capable of distinguishing differences in economic activity in areas with populations living near and below the international poverty line ($1.90 per capita per day). In these impoverished areas, luminosity levels are generally also very low and show little variation (Fig. 1, C to F, and fig. S1), making nightlights potentially less useful for studying and tracking the livelihoods of the very poor. Other recent approaches using mobile phone data to estimate poverty (*11*, *12*) show promise, but could be difficult to scale across countries given their reliance on disparate proprietary data sets.

Here we demonstrate a novel machine learning approach for extracting socioeconomic data from high-resolution daytime satellite imagery. We then validate this approach in five African countries for which recent georeferenced local-level data on economic outcomes are available. In contrast to existing methods, ours can produce fine-grained poverty and wealth estimates using only data available in the public domain.

Transfer learning

High-resolution satellite imagery is increasingly available at the global scale and contains an abundance of information about landscape features that could be correlated with economic activity. Unfortunately, such data are highly unstructured and thus challenging to extract meaningful insights from at scale, even with intensive manual analysis. Recent applications of deep learning techniques to large-scale image data sets have led to marked improvements in fundamental computer vision tasks such as object detection and classification, but these techniques are generally most effective in supervised learning regimes where labeled training data are abundant (*13*). In our setting, however, labeled data are scarce. Even in the instances where detailed household surveys do exist (Fig. 1, A and B), individual surveys typically only contain information for hundreds of locations, yielding data sets many orders of magnitude smaller than those typically used in deep learning applications. Thus, although deep learning models such as convolutional neural networks could in principle be trained to directly estimate economic outcomes from satellite imagery, the scarcity of training data on these outcomes makes the application of these techniques challenging.

We overcome this challenge through a multistep "transfer learning" (*14*) approach (see supplementary materials section 1), whereby a noisy but easily obtained proxy for poverty is used to train a deep learning model (*15*). The model is then used to estimate either average household expenditures or average household wealth at the "cluster" level (roughly equivalent to villages in rural areas or wards in urban areas), the lowest level of geographic aggregation for which latitude and longitude data are available in the public-domain surveys that we use (see supplementary materials 1.4). Household expenditures, where available, are the standard basis from which national poverty statistics are calculated in poor countries, and we use expenditure data from the World Bank's Living Standards Measurement Study (LSMS) surveys. To measure wealth, we use an asset index drawn from the DHS, computed as the first principal component of survey responses to multiple questions about asset ownership. Although the asset index cannot be used directly to construct benchmark measures of poverty, asset-based measures are thought to better capture households' longer-run economic status (*16*, *17*), with the added advantage that many of the enumerated assets are directly observable to the surveyor and therefore are measured with relatively little error.

To estimate these outcomes, our transfer learning pipeline involves three main steps. First, we start with a convolutional neural network (CNN) model that has been pretrained on ImageNet, a large image classification data set that consists of labeled images

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [3]Department of Earth System Science, Stanford University, Stanford, CA, USA. [4]Center on Food Security and the Environment, Stanford University, Stanford, CA, USA. [5]National Bureau of Economic Research, Boston, MA, USA.
*These authors contributed equally to this work. †Corresponding author. Email: mburke@stanford.edu

from 1000 different categories (*18*). In learning to classify each image correctly (e.g., "hamster" versus "weasel"), the model learns to identify low-level image features such as edges and corners that are common to many vision tasks (*19*).

Next, we build on the knowledge gained from this image classification task and fine-tune the CNN on a new task, training it to predict the nighttime light intensities corresponding to input daytime satellite imagery. Here we use the word "predict" to mean estimation of some property that is not directly observed, rather than its common meaning of inferring something about the future. Nightlights are a noisy but globally consistent—and globally available—proxy for economic activity. In this second step, the model learns to "summarize" the high-dimensional input daytime satellite images as a lower-dimensional set of image features that are predictive of the variation in nightlights (see Fig. 2). The trained CNN can be treated as a feature extractor that has learned a nonlinear mapping from each input image to a concise feature vector representation (supplementary materials 1.1). Both daytime imagery (drawn here from the Google Static Maps API) and nightlights (*20*) are available at relatively high resolutions for the entire global land surface, providing a very large labeled training data set.

Finally, we use mean cluster-level values from the survey data along with the corresponding image features extracted from daytime imagery by the CNN to train ridge regression models that can estimate cluster-level expenditures or assets. Regularization in the ridge model guards against overfitting, a potential challenge given the high dimensionality of the extracted features and the relatively small survey data sets. Intuitively, we expect that some subset of the features that explain variation in nightlights is also predictive of economic outcomes.

How might a model partially trained on an imperfect proxy for economic well-being—in this case, the nightlights used in the second training step above—improve upon the direct use of this proxy as an estimator of well-being? Although nightlights display little variation at lower expenditure levels (Fig. 1, C to F), the survey data indicate that other features visible in daytime satellite imagery, such as roofing material and distance to urban areas, vary roughly linearly with expenditure (fig. S2) and thus better capture variation among poorer clusters. Because both nightlights and these features show variation at higher income levels, training on nightlights can help the CNN learn to extract features like these that more capably capture variation across the entire consumption distribution.

Nightlights also have difficulty distinguishing between poor, densely populated areas and wealthy, sparsely populated areas, an added motivation for not using nightlights to estimate per capita consumption. Our approach does not depend on nightlights being able to make this distinction, and instead uses nightlights only as intermediate labels to learn image features that are correlated with economic well-being. The final step of our analysis, in which we train a model to directly estimate local per capita outcomes from daytime image features, does not rely on nightlights.

Visualization of the extracted image features suggests that the model learns to identify some livelihood-relevant characteristics of the landscape (Fig. 2). The model is clearly able to discern semantically meaningful features such as urban areas, roads, bodies of water, and agricultural areas, even though there is no direct supervision—that is, the model is told neither to look for such features, nor that they could be correlated with economic outcomes of interest. It learns on its own that these features are useful for estimating nighttime light intensities. This is in contrast to existing efforts to extract features from satellite imagery, which have relied heavily on human-annotated data (*21*).

Results

Our transfer learning model is strongly predictive of both average household consumption expenditure and asset wealth as measured at the cluster level across multiple African countries. Cross-validated predictions based on models trained separately for each country explain 37 to 55% of the variation in average household consumption

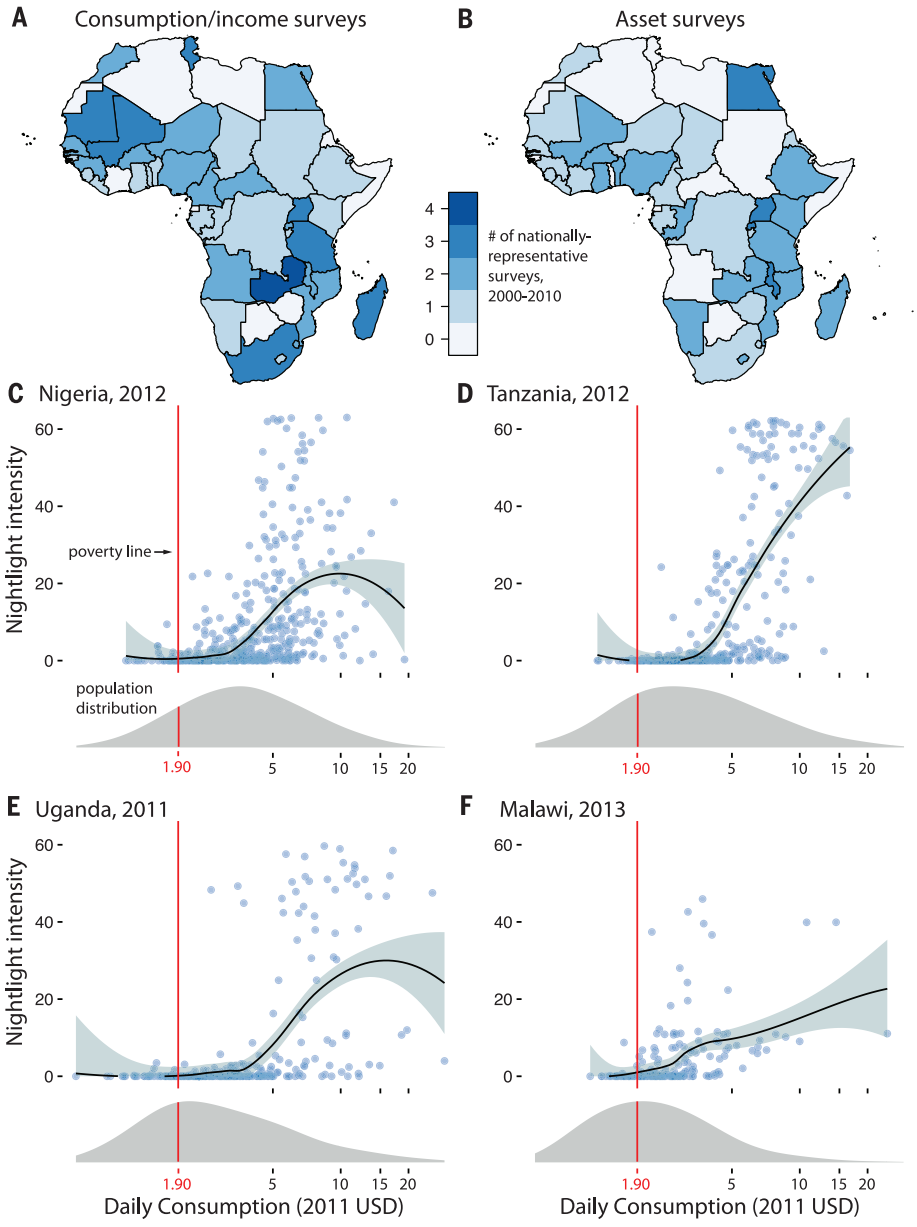

Fig. 1. Poverty data gaps. (**A**) Number of nationally representative consumption surveys occurring in each African country between 2000 and 2010. (**B**) Same as (A), for DHS surveys measuring assets. (**C** to **F**) Relationship between per capita consumption expenditure (measured in U.S. dollars) and nightlight intensity at the cluster level for four African countries, based on household surveys. Nationally representative share of households at each point in the consumption distribution is shown beneath each panel in gray. Vertical red lines show the official international extreme poverty line ($1.90 per person per day), and black lines are fits to the data with corresponding 95% confidence intervals in light blue.

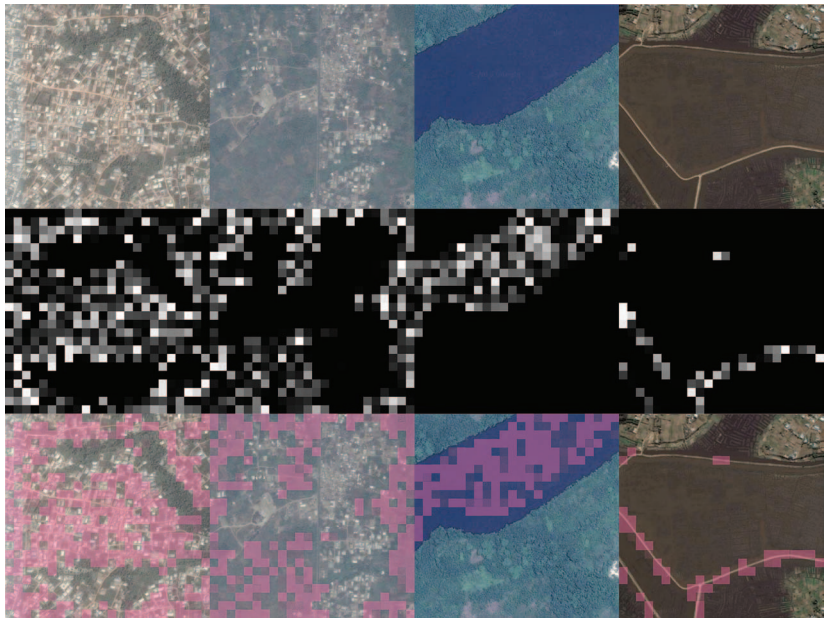

Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

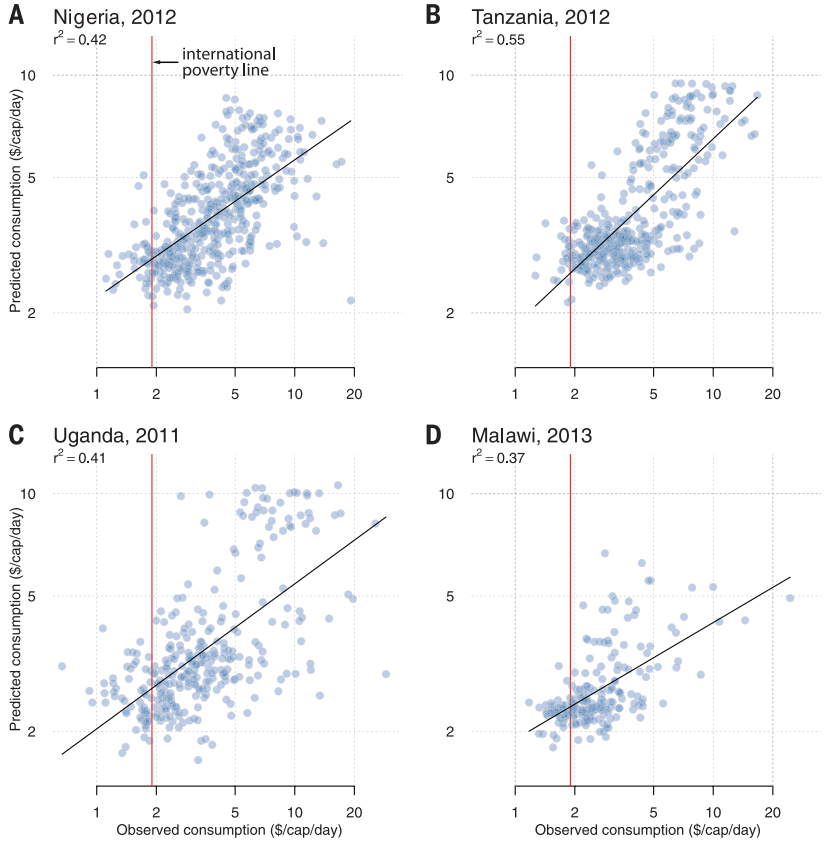

Fig. 3. Predicted cluster-level consumption from transfer learning approach (*y* axis) compared to survey-measured consumption (*x* axis). Results are shown for Nigeria (**A**), Tanzania (**B**), Uganda (**C**), and Malawi (**D**). Predictions and reported r^2 values in each panel are from fivefold cross-validation. Black line is the best fit line, and red line is international poverty line of $1.90 per person per day. Both axes are shown in logarithmic scale. Countries are ordered by population size.

across four countries for which recent survey data are available (Fig. 3), and 55 to 75% of the variation in average household asset wealth across five countries with recent survey data (fig. S3). Models trained on pooled consumption or asset observations across all countries (hereafter "pooled model") perform similarly, with cross-validated predictions explaining 44 to 59% of the overall variation in these outcomes (fig. S4).

This high overall predictive power is achieved despite a lack of temporal labels for the daytime imagery (i.e., the exact date of each image is unknown), as well as imperfect knowledge of the location of the clusters, as up to 10 km of random noise was added to cluster coordinates by the data collection agencies to protect the privacy of survey respondents. Predictive power for assets is nearly uniformly higher than for consumption, perhaps reflecting the larger sample sizes available in the asset surveys; that the asset index is thought to serve as a better proxy for households' longer-run economic status (*16*, *17*) (which could be better correlated with landscape features that change slowly over time); and/or the possibility that certain assets in the index (such as roof type) are directly identified in extracted features (see supplementary materials 2.1). We investigate these potential explanations by constructing our own asset index from variables available in the Uganda LSMS and comparing predictive performance for that index relative to performance for consumption measured in the same survey. We find that differences in the outcome being measured, rather than differences in survey design or direct identification of key assets in daytime imagery, likely explain these performance differences (see supplementary materials 2.1 and fig. S5). Finally, asset-estimation performance of our model in Rwanda surpasses performance in a recent study using cell phone data to estimate identical outcomes (*11*) (cluster-level r^2 = 0.62 in that study, and r^2 = 0.75 in our study; r^2 is the coefficient of determination), again with the added advantage that our predictions can be constructed entirely from publicly available data, obviating the need to obtain and evaluate proprietary data sets when scaling across countries.

To test whether our transfer learning model improves upon the direct use of nightlights to estimate livelihoods, we ran 100 trials of 10-fold cross-validation separately for each country and for the pooled model, each time comparing the predictive power of our transfer learning model to that of nightlights alone. To understand relative performance on different subsets of the consumption distribution, trials were run separately with the sample of clusters restricted to those whose average consumption fell below each quintile of the consumption distribution. The same procedure was repeated for assets.

Despite being trained partially on nightlights, our model is on average substantially more predictive of variation in consumption and assets than nightlights alone. For expenditures, our model outperforms nightlights at nearly all points in the consumption distribution, for both the pooled model and for countries run independently (Fig. 4A and fig. S6). In the pooled setting, for

clusters below the international poverty line, our model outperforms nightlights in 81.3% of trials, with an average increase in r^2 of 0.04. For clusters below two times the poverty line, our model outperforms nightlights in 98.5% of trials, with an average increase in r^2 of 0.10, an 81.2% increase in explanatory power. For clusters below three times the poverty line, our model outperforms nightlights in 99.5% of trials, with an average increase in r^2 of 0.12, corresponding to a 54.2% increase in explanatory power. Results for individual countries are similar, with the predictive power of our model outperforming nightlights for all countries at nearly all parts of the consumption distribution (fig. S6). Our model's relative performance against nightlights is even better for assets than for consumption (Fig. 4B), particularly for clusters with low average asset levels. Using more information in nightlights beyond mean luminosity leads to some improvement in nightlights performance, but this improved use of nightlights is still outperformed by our model (see supplementary materials 2.2 and fig S7).

We also study whether our approach improves upon other simpler approaches to extracting information from daytime imagery and predicting economic outcomes using available survey data. We find that our CNN feature extractor far outperforms common general-purpose image features such as color histograms and histograms of oriented gradients (see supplementary materials 2.3 and fig. S8). Our approach also performs as well as or better than an intuitive approach of using data from past surveys to predict outcomes in more recent surveys (see supplementary materials 2.4 and table S2).

To further quantify the statistical significance of our results, we perform an experiment in which we randomly reassign daytime imagery to survey locations and retrain the model on these incorrect images (see supplementary materials 1.7). We repeat this experiment 1000 times within each country and for the pooled model, then compare the predictive power when daytime images were assigned to their correct locations (as in Fig. 3) to the distribution of r^2 values obtained from the 1000 placebo trials. As shown in Fig. 4, C and D, the r^2 values obtained using "correct" daytime imagery are much higher than any of the r^2 values obtained from the reshuffled images, for both consumption and assets, indicating that our model's level of predictive performance is unlikely to have arisen by chance.

Finally, capitalizing on our survey-based measures of consumption and assets in multiple countries, we study the extent to which a model trained using data and satellite image features from one country can estimate livelihoods in other countries. Examining whether a particular model generalizes across borders is useful for understanding whether accurate predictions can be made from imagery alone in areas with no survey data—an important practical concern given the paucity of existing survey data in many African countries (see Fig. 1)—as well as for gaining insight about commonalities in the determinants of livelihoods across countries.

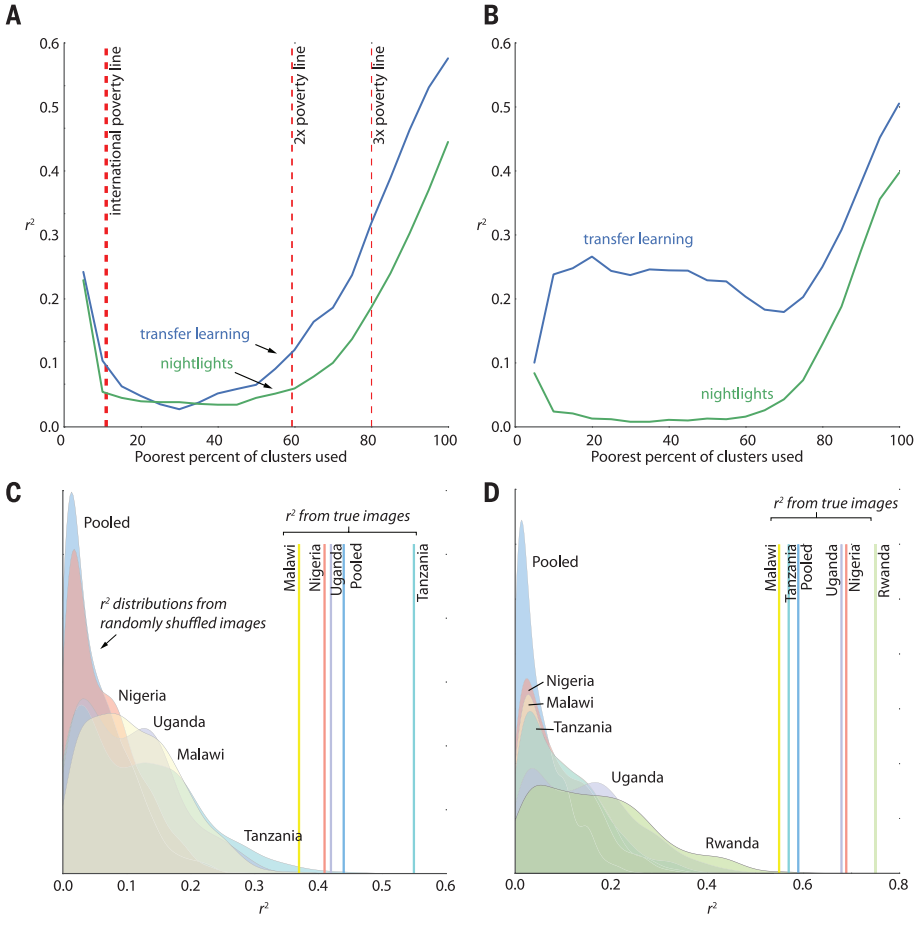

Fig. 4. Evaluation of model performance. (**A**) Performance of transfer learning model relative to nightlights for estimating consumption, using pooled observations across the four LSMS countries. Trials were run separately for increasing percentages of the available clusters (e.g., *x*-axis value of 40 indicates that all clusters below 40th percentile in consumption were included). Vertical red lines indicate various multiples of the international poverty line. Image features reduced to 100 dimensions using principal component analysis. (**B**) Same as (A), but for assets. (**C**) Comparison of r^2 of models trained on correctly assigned images in each country (vertical lines) to the distribution of r^2 values obtained from trials in which the model was trained on randomly shuffled images (1000 trials per country). (**D**) Same as (C), but for assets. Cross-validated r^2 values are reported in all panels.

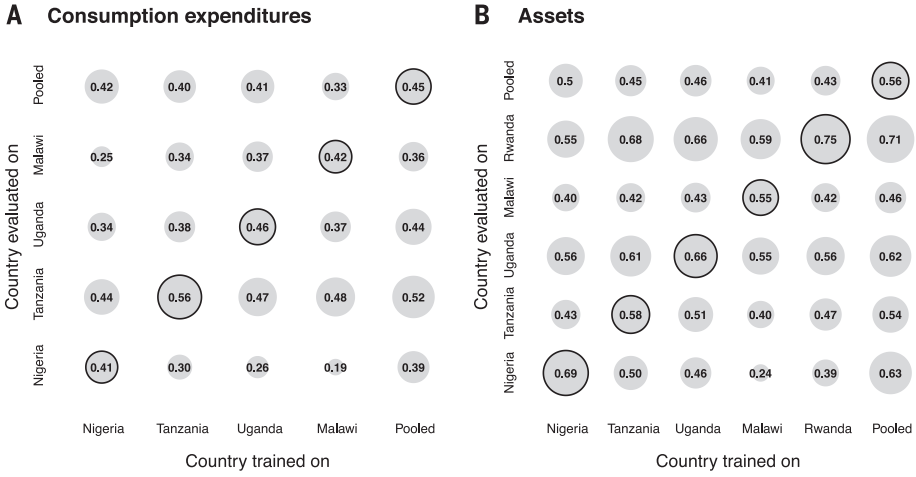

Fig. 5. Cross-border model generalization. (**A**) Cross-validated r^2 values for consumption predictions for models trained in one country and applied in other countries. Countries on *x* axis indicate where model was trained, countries on *y* axis where model was evaluated. Reported r^2 values are averaged over 100 folds (10 trials, 10 folds each). (**B**) Same as in (A), but for assets.

We find that for both consumption and assets, models trained in-country uniformly outperform models trained out-of-country (Fig. 5), as would be expected. But we also find that models appear to "travel well" across borders, with out-of-country predictions often approaching the accuracy of in-country predictions. Pooled models trained on all four consumption surveys or all five asset surveys very nearly approach the predictive power of in-country models in almost all countries for both outcomes. These results indicate that, at least for our sample of countries, common determinants of livelihoods are revealed in imagery, and these commonalities can be leveraged to estimate consumption and asset outcomes with reasonable accuracy in countries where survey outcomes are unobserved.

Discussion

Our approach demonstrates that existing high-resolution daytime satellite imagery can be used to make fairly accurate predictions about the spatial distribution of economic well-being across five African countries. Our model performs well despite inexact data on both the timing of the daytime imagery and the location of clusters in the training data, and more precise data in either of these dimensions are likely to further improve model performance.

Notably, we show that our model's predictive power declines only modestly when a model trained in one of our sample countries is used to estimate consumption or assets in another country. Despite differences in economic and political institutions across countries, model-derived features appear to identify fundamental commonalities in the determinants of livelihoods across settings, suggesting that our approach could be used to fill in the large data gaps resulting from poor survey coverage in many African countries. In contrast to other recent approaches that rely on proprietary commercial data sets, our method uses only publicly available data and so is straightforward and nearly costless to scale across countries.

Although our model outperforms other sources of passively collected data (e.g., cellphone data, nightlights) in estimating economic well-being at the cluster level, we are currently unable to assess its ability to discern differences within clusters, as public-domain survey data assign identical coordinates to all households in a given cluster to preserve respondent privacy. In principle, our model can make predictions at any resolution for which daytime satellite imagery is available, though predictions on finer scales would likely be noisier. New sources of ground truth data, whether from more disaggregated surveys or novel crowdsourced channels, could enable evaluation of our model at the household level. Combining our extracted features with other passively collected data, in locations where such data are available, could also increase both household- and cluster-level predictive power.

Given the limited availability of high-resolution time series of daytime imagery, we also have not yet been able to evaluate the ability of our transfer learning approach to predict changes in economic well-being over time at particular locations. Such predictions would be very helpful to both researchers and policy-makers and should be enabled in the near future as increasing amounts of high-resolution satellite imagery become available (*22*).

Our transfer learning strategy of using a plentiful but noisy proxy shows how powerful machine learning tools, which typically thrive in data-rich settings, can be productively employed even when data on key outcomes of interest are scarce. Our approach could have broad application across many scientific domains and may be immediately useful for inexpensively producing granular data on other socioeconomic outcomes of interest to the international community, such as the large set of indicators proposed for the United Nations Sustainable Development Goals (*5*).

REFERENCES AND NOTES

1. United Nations, "A World That Counts: Mobilising the Data Revolution for Sustainable Development" (2014).
2. S. Devarajan, *Rev. Income Wealth* **59**, S9–S15 (2013).
3. M. Jerven, *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It* (Cornell Univ. Press, 2013).
4. World Bank, PovcalNet online poverty analysis tool, http://iresearch.worldbank.org/povcalnet/ (2015).
5. M. Jerven, "Benefits and costs of the data for development targets for the Post-2015 Development Agenda," Data for Development Assessment Paper Working Paper, September (Copenhagen Consensus Center, Copenhagen, 2014).
6. J. Sandefur, A. Glassman, *J. Dev. Stud.* **51**, 116–132 (2015).
7. J. V. Henderson, A. Storeygard, D. N. Weil, *Am. Econ. Rev.* **102**, 994–1028 (2012).
8. X. Chen, W. D. Nordhaus, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8589–8594 (2011).
9. S. Michalopoulos, E. Papaioannou, *Q. J. Econ.* **129**, 151–213 (2013).
10. M. Pinkovskiy, X. Sala-i-Martin, *Q. J. Econ.* **131**, 579–631 (2016).
11. J. Blumenstock, G. Cadamuro, R. On, *Science* **350**, 1073–1076 (2015).
12. L. Hong, E. Frias-Martinez, V. Frias-Martinez, "Topic models to infer socioeconomic maps," AAAI Conference on Artificial Intelligence (2016).
13. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436–444 (2015).
14. S. J. Pan, Q. Yang, *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
15. M. Xie, N. Jean, M. Burke, D. Lobell, S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," AAAI Conference on Artificial Intelligence (2016).
16. D. Filmer, L. H. Pritchett, *Demography* **38**, 115–132 (2001).
17. D. E. Sahn, D. Stifel, *Rev. Income Wealth* **49**, 463–489 (2003).
18. O. Russakovsky *et al.*, *Int. J. Comput. Vis.* **115**, 211–252 (2014).
19. A. Krizhevsky, I. Sutskever, G. E. Hinton, *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
20. National Geophysical Data Center, Version 4 DMSP-OLS Nighttime Lights Time Series (2010).
21. V. Mnih, G. E. Hinton, in *11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5 to 11 September 2010* (Springer, 2010), pp. 210–223.
22. E. Hand, *Science* **348**, 172–177 (2015).

ACKNOWLEDGMENTS

We gratefully acknowledge support from NVIDIA Corporation through an NVIDIA Academic Hardware Grant, from Stanford's Global Development and Poverty Initiative, and from the AidData Project at the College of William & Mary. N.J. acknowledges support from the National Defense Science and Engineering Graduate Fellowship Program. S.E. is partially supported by NSF grant 1522054 through subcontract 72954-10597. We declare no conflicts of interest. All data and code needed to replicate these results are available at http://purl.stanford.edu/cz134jc5378.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6301/790/suppl/DC1
Materials and Methods
Figs. S1 to S22
Tables S1 to S3
References (*23*–*27*)

30 March 2016; accepted 6 July 2016
10.1126/science.aaf7894

STATISTICAL PHYSICS

Quantum thermalization through entanglement in an isolated many-body system

Adam M. Kaufman, M. Eric Tai, Alexander Lukin, Matthew Rispoli, Robert Schittko, Philipp M. Preiss, Markus Greiner*

Statistical mechanics relies on the maximization of entropy in a system at thermal equilibrium. However, an isolated quantum many-body system initialized in a pure state remains pure during Schrödinger evolution, and in this sense it has static, zero entropy. We experimentally studied the emergence of statistical mechanics in a quantum state and observed the fundamental role of quantum entanglement in facilitating this emergence. Microscopy of an evolving quantum system indicates that the full quantum state remains pure, whereas thermalization occurs on a local scale. We directly measured entanglement entropy, which assumes the role of the thermal entropy in thermalization. The entanglement creates local entropy that validates the use of statistical physics for local observables. Our measurements are consistent with the eigenstate thermalization hypothesis.

When an isolated quantum system is perturbed—for instance, owing to a sudden change in the Hamiltonian (a so-called quench)—the ensuing dynamics are determined by an eigenstate distribution that is induced by the quench (*1*). At any given time, the evolving quantum state will have amplitudes that depend on the eigenstates populated by the quench and the energy eigenvalues of the Hamiltonian. In many cases, however,

Department of Physics, Harvard University, Cambridge, MA 02138, USA.
***Corresponding author. Email: greiner@physics.harvard.edu**