# LIMITATIONS OF NON-TRADITIONAL DATA: EVALUATING SATELLITE DATA AND OPENSTREETMAP FOR PREDICTING SPATIO-TEMPORAL VARIATION IN MATERIAL WEALTH

**Nathanael Schmidt-Ott***
Chair of International Economic Policy
University of Goettingen
n.schmidt.ott@gmail.com

**Krisztina Kis-Katos**
Chair of International Economic Policy
University of Goettingen
krisztina.kis-katos@uni-goettingen.de

## ABSTRACT

Granular and repeated measurements of material wealth are essential for understanding and improving economic livelihoods, but detailed survey data is scarce, especially in Africa — the continent most affected by poverty. Recent machine learning approaches successfully leverage non-traditional data sources such as satellite data to predict material wealth across locations but not its variation over time. This work systematically investigates the potential of publicly available satellite and OpenStreetMap data to predict levels and fluctuations in annual consumption expenditure and asset wealth at the village level. Models trained on panel data from five African countries between 2007 and 2021 explain up to 52% of spatial variation but struggle to explain temporal variation ($R^2$<0.04). This has important implications for retrospective policy evaluation, highlighting that in many contexts, non-traditional data sources cannot replace traditional survey data. Considering more nuanced input data constitutes a promising avenue for future research.

## 1 INTRODUCTION

In the global fight against poverty, it is vital to understand which policies and anti-poverty programs are most effective in raising living standards (Annan, 2018; Imbens, 2010). Impact evaluation of anti-poverty programs commonly relies on extensive and detailed household survey data. Ideally, this should be spatially and temporally detailed panel/longitudinal data as it will permit policy evaluations by allowing the establishment of accurate counterfactuals (Abadie & Cattaneo, 2018). However, detailed survey data is scarce — its collection is time and resource-intensive (Serajuddin et al., 2015).

In light of this evident "data-deprivation" (Serajuddin et al., 2015), economists and social scientists seek non-traditional data sources to approximate annual panel data on key economic variables at high regional resolution (Chen & Nordhaus, 2011; Henderson et al., 2012). For valid proxies of annual material wealth, non-traditional data sources must accurately differentiate wealth between locations (spatial prediction) and within locations over time (temporal prediction). Recent works highlight the use of high to medium resolution daytime satellite imagery (<1 m/pixel to 30 m/pixel) and mobile phone metadata to delineate wealth disparities among villages or neighbourhoods (Blumenstock et al., 2015; Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022). However, the accuracy of day-and-nighttime satellite images in predicting wealth dynamics within locations over time is still under scrutiny (Yeh et al., 2020; Wölk et al., 2023). It remains to be understood how this affects annual material wealth prediction, which is much needed for policy evaluation.

This paper systematically studies whether publicly available non-traditional data allow for predicting annual material wealth at the village or neighbourhood level (spatial units of 6.7X6.7km, called

*Corresponding author

clusters in this context).[1] In particular, we focus on non-traditional data sources that are available for more than two decades, including daytime satellite images (30 m/pixel), nightlights, vegetation and water indices and precipitation. Furthermore, we utilize land cover, settlement area and volunteered geographic information from OpenStreetMap (OSM). Using repeated observations from household panel survey data across five African countries (Table 1), we make three main contributions: (1) we introduce a prediction framework that combines the output of a spatial and a temporal model to obtain cluster-and year-specific predictions of material wealth; (2) for temporal prediction, we explore the use of time series information from vegetation and water indices and precipitation beyond day-and-nighttime satellite image features; and (3) we document that relatively high accuracy in annual wealth predictions ($R^2$=0.49) reflects the model's capability to predict material wealth across space ($R^2$=0.52), while at the same time being unable to predict wealth dynamics over time ($R^2$<0.04). Crucially, our findings underscore that the reliability of annual wealth approximations from non-traditional data sources is questionable, especially when not relying on proprietary data (e.g., high-resolution satellite images, mobile phone or social media usage data). This limits the scope of retrospective policy evaluations that include years before 2015, for which higher-resolution satellite imagery is not publicly available.

## 2 Data and Methods

**Survey Data.** We obtain a panel on ground truth wealth labels from the World Bank's Living Standards Measurement Study. The panel data encompasses surveys conducted between 2007 and 2021 in five African countries: Ethiopia, Malawi, Nigeria, Tanzania, and Uganda. Depending on survey specifics, households are revisited between one and six times. Households are grouped into clusters, the smallest geographical areas at which geolocations are available. We aggregate the wealth indicators to the cluster level by taking the average over all households within each cluster. If households retire from the survey, we remove them entirely from our sample. This prevents household composition within each cluster from driving temporal variation in the data. However, the resulting attrition introduces bias to the cluster estimates, because at baseline they are on average wealthier than non-attrition households by about 0.86$ per capita per day. Our data includes 17,445 households across 2,128 clusters and 6,401 cluster-year observations (Table 1).[2]

We use daily per capita consumption expenditures, expressed in 2017 international dollars, as the primary indicator of material wealth. Additionally, we derive a relative asset wealth index using PCA on the pooled sample including all clusters over all years. The asset index ranks households based on their asset ownership relative to one another (Filmer & Pritchett, 2001). In line with other studies (Jean et al., 2016; Yeh et al., 2020), we find a moderate correlation between the asset index and log consumption expenditures at the household (r=0.53) and at the cluster level (r=0.53).

| | Consumption | | | | Asset wealth | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | $\sigma_B$ | $\sigma_W$ | $\sigma_O$ | Mean | $\sigma_B$ | $\sigma_W$ | $\sigma_O$ | N | Clusters | Years |
| Ethiopia | 2.47 | 2.10 | 0.96 | 1.48 | -0.74 | 1.34 | 0.19 | 0.82 | 1187 | 431 | 11, 13, 15 |
| Malawi | 4.13 | 5.37 | 1.13 | 3.88 | -0.31 | 0.86 | 0.13 | 0.62 | 366 | 183 | 10, 13 |
| Nigeria | 4.61 | 4.49 | 2.33 | 3.27 | 0.37 | 1.31 | 0.22 | 0.80 | 1167 | 419 | 10, 12, 15, 18 |
| Tanzania | 4.42 | 5.38 | 1.81 | 3.68 | 0.56 | 1.68 | 0.20 | 1.07 | 1951 | 781 | 08, 10, 12, 14, 19, 20 |
| Uganda | 4.13 | 7.47 | 1.87 | 3.60 | -0.30 | 1.57 | 0.25 | 0.71 | 1730 | 314 | 09, 10, 11, 13, 15, 18, 19 |
| Pooled | 4.00 | 5.27 | 1.79 | 3.38 | 0.00 | 1.71 | 0.22 | 1.00 | 6401 | 2128 | 08–20 (not 16, 17) |

Table 1: **Summary statistics of ground truth labels.** $\sigma_B$, $\sigma_W$, $\sigma_O$ indicate the standard deviation between clusters, within clusters, and overall. N is the total number of cluster-year observations in respective countries. Years indicates the years in which data was collected. Each cluster is visited at least twice. Consumption expenditure is measured in 2017 international dollars. Asset wealth are the first principal component scores over a set of household assets.

[1] In line with Yeh et al. (2020), we define a cluster as a grid cell measuring 6,720m in width and height.

[2] Disclosed GPS coordinates of the clusters' locations are randomly displaced by up to 2 km in urban areas and up to 5 km in rural areas. We recenter the geolocations to the closest populated place on OSM. We consider the following populated places: villages, neighbourhoods, quarters, city blocks, suburbs, towns, cities, residential areas, buildings, isolated dwellings, and hamlets.

**Dynamic Features**. Dynamic features include day-and-nighttime satellite images, NDVI, NDWI and precipitation and are available annually. We obtain median annual RGB satellite images from all available cloud-free Landsat 5, 7, and 8 images, mean annual NDVI and NDWI from cloud-free Modis satellite images, harmonised nighttime satellite images (nightlights) from Chen et al. (2021), and precipitation data from Funk et al. (2015). Similar to Chi et al. (2022), we extract vector representations from RGB daytime satellite images using a ResNet18 pre-trained on ImageNet. We do not finetune the network. We remove the network's last layer to obtain a 512-dimensional vector representation of each satellite image and apply PCA, reducing the dimensionality of the image representations to 25. For NDVI, NDWI, and nightlights, we extract the cluster grid cell means and standard deviations. We sum precipitation data over the year at cluster centroids.

**Static Features.** Static features include land cover from 2020 (Zanaga et al., 2021), human settlement areas from 2019 (Marconcini et al., 2021), and OSM features as of March 2023. Static features do not vary over time. For land cover, we calculate the area covered by different classes such as cropland or built-up areas. From OSM, we compute the road network length and the number of amenities within clusters.[3] We approximate cluster remoteness by calculating the distance from clusters to the nearest paved/primary road and to the amenities. A caveat of OSM data is its incompleteness (Herfort et al., 2023), potentially introducing noise into the analysis. However, excluding OSM data reduced model performance.

**Methods.** The ideal model explains variation in material wealth between clusters (spatial variation) and within clusters over time (temporal variation). To achieve this, we split the model into a spatial and a temporal component. The combined model aggregates the output of both model components to obtain annual material wealth predictions. This approach has two advantages over directly predicting annual material wealth. First, the spatial and temporal models are specialised to their tasks. Second, this allows for explicitly assessing whether the combined predictions reflect spatial and temporal variation in material wealth.

We disentangle the material wealth indicator into a between and within component. $\bar{w}_c$ — the between component — is the mean material wealth over all time periods $t$ of cluster $c$. $\widetilde{w}_{ct} = w_{ct} - \bar{w}_c$ — the within component — denotes the annual deviation of clusters' material wealth from its mean over time. The target variable for the spatial model is $\bar{w}_c$. The target variable for the temporal model is $\widetilde{w}_{ct}$. The spatial model takes static features and dynamic features averaged over time as input. The temporal model takes exclusively dynamic features as inputs. Consistent with the target variable of the temporal model, we demean all input features for the temporal model. For both models, our main algorithm is a Random Forests (RF) algorithm.[4] Additionally, as a more flexible model, we estimate a gradient boosted regression tree (GBM).[5] In order to allow for a better comparison to the literature, we also train a ResNet18 model on multi-spectral (MS) day-time satellite images in an end-to-end fashion as suggested by Yeh et al. (2020).[6] We train the ResNet18 for day-time satellite images only, because Yeh et al. (2020) find that including nightlights reduces temporal model performance.

[3] Amenities include bars, cafes, markets, schools, universities, libraries, fuel stations, pharmacies, hospitals, and clinics.

[4] We set the maximum number of predictors at each split to the square root of the total number of predictors, the number of estimators to 3000, and the minimum number of observations in each leaf to five for both models.

[5] We tune GBM hyperparameters using nested 5-fold cross-validation, performing a grid search over the maximum tree depth (2, 4, 6, 8, 10) and the number of estimators (250, 500, 750) to maximise $R^2$.

[6] We closely follow the training procedure outlined by Yeh et al. (2020): for spatial prediction, we use a ResNet18, pre-trained on ImageNet. We modify the first convolutional layer to accommodate MS satellite images, adopting the same-scaled weight initialisation scheme (Yeh et al., 2020). For temporal prediction, we train a ResNet18 from scratch, initialising weights randomly using Kaiming weight initialisation (He et al., 2015). We stack the image of a given year and the respective mean over time on top of each other, creating an image of size *224×224×12* for a cluster-year observation. The ResNet18 models are trained with the Adam optimiser (Kingma & Ba, 2014) and a mean squared-error loss function in batches of 64 images. We perform a grid search over the learning rate (1e-2, 1e-3, 1e-4) and $L_2$ weight regularisation (1e-2, 1e-3). The learning rate is decayed by a factor of 0.96 after every epoch. We randomly flip images horizontally and vertically with a probability of 50%. We train models for 150 epochs and use the model with the highest $R^2$ on the validation set for model comparison. We tune hyperparameters on a spatially and temporally stratified 15% validation sample within the training sample of each CV-iteration.

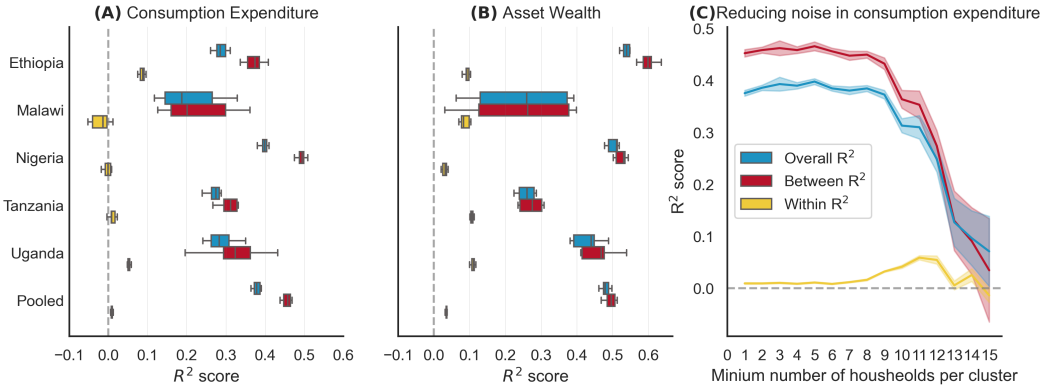


Figure 1: **Temporal, spatial and overall performance.** *Panel A*: the model consistently explains a relatively high share of the variation ($R^2$) between clusters (red) in all countries, but fails to explain temporal variation (yellow) in consumption expenditure. *Panel B*: the model explains slightly higher shares of spatial and temporal variation in asset wealth as compared to consumption expenditure. *Panel C*: reducing noise in consumption expenditure by removing clusters with few households from the sample does not improve model performance substantially, but leads to a sharp deterioration in model performance after the sample is severely reduced.

We use a spatially and temporally stratified version of 5-fold cross-validation (CV) to estimate out-of-sample performance, addressing temporal and spatial autocorrelation (Rolf, 2023; Chi et al., 2022). In every iteration, we ensure that spatially overlapping clusters or repeated cluster observations belong only to the training or the test fold. We use the coefficient of determination ($R^2$) to quantify model performance.

## 3 Results

Consistent with past work (Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022; Engstrom et al., 2022), we find that the spatial model explains a relatively high share of the variation in consumption expenditure ($R^2$=0.46) and asset wealth ($R^2$=0.52). However, we find that the temporal model explains virtually none of the temporal variation in consumption expenditure ($R^2$<0.01) and only marginal variation in asset wealth ($R^2$<0.04). Compared to RF, the more complex GBM icreases prediction accuracy for the spatial model, not for the temporal model. The ResNet18 yields significantly lower performance for all metrics, most likely because of comparatively few samples for a deep learning algorithm (Table 2). During hyperparameter tuning of the ResNet18 model and the GBM, we find that as soon as training performance of the temporal model increases, the validation performance deteriorates, indicating that for temporal prediction the models pick up noise rather than signal in the features.

| | Consumption | | | Asset Wealth | | |
|---|---|---|---|---|---|---|
| | (A) Between | (B) Within | (C) Overall | (D) Between | (E) Within | (F) Overall |
| RF | 0.456(0.012) | 0.009(0.002) | 0.379(0.009) | 0.495(0.014) | 0.035(0.002) | 0.481(0.011) |
| GBM | 0.457(0.015) | -0.009(0.005) | 0.378(0.013) | 0.521(0.014) | 0.022(0.005) | 0.494(0.010) |
| ResNet18 | 0.329 | -0.006 | 0.189 | 0.455 | 0.007 | 0.359 |
| RF Dir | - | - | 0.381(0.010) | - | - | 0.491(0.009) |
| GBM Dir | - | - | 0.378(0.012) | - | - | 0.489(0.006) |
| N | 2128 | 6401 | 6401 | 2128 | 6401 | 6401 |

Table 2: **Model performance for different algorithms.** Columns present $R^2$ values for different algorithms. RF: Random Forests, GBM: Gradient Boosted Regression Tree. ResNet18 is trained on MS satellite images only. *Dir* indicates that the algorithm is trained to predict annual material wealth levels directly. Reported $R^2$ values are based on 5-fold spatial CV. Models are trained and evaluated on the same data splits. Performance metrics are averaged over 50 folds (10 runs, 5 folds each)—standard errors in parentheses. For ResNet18, spatial CV is not repeated to limit computations.

For asset wealth, we achieve slightly higher performance when training and evaluating the temporal model on individual countries ($R^2$=0.04–0.11). We also find slight heterogeneity in temporal prediction accuracy with respect to countries for consumption expenditures. The combined prediction accuracy reflects the model's capability to predict material wealth across space, not over time. Thus, any temporal variation in predicted annual material wealth levels is likely to be random (Panels A and B in Figure 1).

Reducing noise in the ground truth labels by excluding clusters with few households does not increase model performance. Especially for the temporal model, small changes in the wealth indicators over time might be obscured by noise in survey measurements (Yeh et al., 2020). We observe an increase in within $R^2$, alongside a substantial decrease in between and overall $R^2$ only after excluding clusters with less than eight households from the data (Panel C in Figure 1). However, we attribute these changes in performance to a significant shift in the sample composition, almost entirely removing clusters from Tanzania and Uganda.

Restricting the sample to specific time periods with higher image quality, separating rural and urban clusters, or requiring larger time differences between cluster-year observations does not improve predictions of the temporal model for either wealth indicator. The quality of day-and-nighttime images increases after 2013, when Landsat 8 daytime satellite images and VIIRS-DNB nighttime satellite images become available. However, restricting the sample to the respective time period does not affect temporal model performance (column B in Table 3). In urban areas, the temporal model performs worse as compared to rural areas (columns C, D in Table 3). Finally, annual changes in material wealth might be too small or too noisy for the temporal model to capture them. Nevertheless, ensuring at least three years between observations of the clusters does not increase model performance significantly vis-a-vis the full sample (column E in Table 3).

| | (A) Full | (B) Post 2013 | (C) Rural | (D) Urban | (E) Min 3 years |
|---|---|---|---|---|---|
| $R^2$ Consumption | 0.005(0.003) | 0.003(0.002) | 0.017(0.001) | -0.045(0.005) | 0.021(0.002) |
| $R^2$ Asset wealth | 0.035(0.002) | 0.042(0.005) | 0.044(0.004) | 0.009(0.005) | 0.070(0.005) |
| N | 6401 | 2932 | 4417 | 1984 | 4011 |

Table 3: **Temporal model performance for different subsamples of the data.** Columns present $R^2$ values when restricting the sample based on different criteria. Reported $R^2$ values are averaged over 10 repeated runs of 5-fold spatial CV — standard deviations in parentheses. (A) All cluster-year observations, (B) only cluster-year observations post 2013, (C) and (D) only rural or urban clusters, (E) ensures at least 3 years between repeated observations.

## 4 Discussion

Our main findings suggest that publicly available medium-resolution satellite and OSM data have a limited capacity to approximate annual material wealth at the cluster level. While these data sources allow for explaining a considerable share of spatial variation, they fall short of explaining temporal variation in material wealth. A central real world application of this finding is that, when used without more sophisticated and proprietary data (e.g., reflecting mobile phone or internet usage), these non-traditional data sources cannot shed light on which policies are most effective in raising living standards because they do not allow for approximating the evolution of material wealth over time. Nevertheless this work confirms that satellite and OSM data produce detailed, static maps of material wealth, which can aid efficient targeting of anti-poverty programs (Aiken et al., 2022; Smythe & Blumenstock, 2022).

Overall, these findings underscore the need for more nuanced models to predict temporal changes in material wealth, enabling rigorous policy evaluations in the region. Repeatedly counting objects on high resolution satellite images, such as vehicles or roofing materials and comparing them over time is a promising avenue for future research to improve temporal predictions. If successful, such an approach would also increase model interpretability, as predictions could be linked to the specific object counts. Moreover, exploring data related to human activity, such as mobile phone meta data (Blumenstock et al., 2015) or social media data (Fatehkia et al., 2020) for spatial prediction, offer alternative avenues for temporal prediction.

## REFERENCES


Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E Blumenstock. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870, 2022.

Kofi Annan. Data can help to end malnutrition across africa. *Nature*, 555(7697):7–8, 2018.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

Xi Chen and William D Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

Zuoqi Chen, Bailang Yu, Chengshu Yang, Yuyu Zhou, Shenjun Yao, Xingjian Qian, Congxiao Wang, Bin Wu, and Jianping Wu. An extended time series (2000-2018) of global npp-viirs-like nighttime light data from a cross-sensor calibration. *Earth System Science Data*, 13(3):889–906, 2021.

Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E Blumenstock. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119 (3):1–11, 2022.

Ryan Engstrom, Jonathan Hersh, and David Newhouse. Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, 36(2): 382–412, 2022.

Masoomali Fatehkia, Isabelle Tingzon, Ardie Orden, Stephanie Sy, Vedran Sekara, Manuel Garcia-Herranz, and Ingmar Weber. Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):2–15, 2020.

Deon Filmer and Lant H Pritchett. Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115–132, 2001.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data*, 2(1):1–21, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring economic growth from outer space. *American economic review*, 102(2):994–1028, 2012.

Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap. *Nature Communications*, 14(1):3985, 2023.

Guido W Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature*, 48(2):399–423, 2010.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Mattia Marconcini, Annekatrin Metz-Marconcini, Thomas Esch, and Noel Gorelick. Understanding current trends in global urbanisation-the world settlement footprint suite. *GI_Forum*, 9(1):33–38, 2021.

Esther Rolf. Evaluation challenges for geospatial ml, 2023.

Umar Serajuddin, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen. Data deprivation: another deprivation to end. *World Bank Policy Research Working Paper* No. 7252, World Bank, Washington, DC., 2015.

Isabella S Smythe and Joshua E Blumenstock. Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, 119(32): e2120025119, 2022.

Fabian Wölk, Tingting Yuan, Krisztina Kis-Katos, and Xiaoming Fu. A temporal–spatial analysis on the socioeconomic development of rural villages in thailand and vietnam based on satellite image data. *Computer Communications*, 203:146–162, 2023.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

D. Zanaga, R. Van De Kerchoveand, W. De Keersmaecker, N. Souverijns, C. Brockmann, R. Quast, J. Wevers, A. Grosu, A. Paccini, S. Vergnaud, O. Cartus, M. Santoro, S. Fritz, M. Herold, L. Li, N.E. Tsendbazar, F. Ramoino, and O. Arino. Esa worldcover 10 m 2020 v100, 2021. Available at `https://zenodo.org/records/5571936`, accessed on April 16, 2024.