Working Paper No. 165

Proxying Economic Activity with Daytime Satellite Imagery: Filling Data Gaps Across Time and Space

Patrick Lehnert, Michael Niederberger, Uschi Backes-Gellner and Eric Bettinger

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 165

Proxying Economic Activity with Daytime Satellite Imagery: Filling Data Gaps Across Time and Space

Patrick Lehnert, Michael Niederberger, Uschi Backes-Gellner and Eric Bettinger

September 2022 (previous versions: March and October 2020, July 2021)

Published as: "Proxying Economic Activity with Daytime Satellite Imagery: Filling Data Gaps Across Time and Space." *PNAS Nexus*, 2(2023)4. By Patrick Lehnert, Michael Niederberger, Uschi Backes-Gellner and Eric Bettinger.

DOI: https://doi.org/10.1093/pnasnexus/pgad099

Proxying economic activity with daytime satellite imagery: Filling data gaps across time and space*

Patrick Lehnert†a, Michael Niederberger[a,b], Uschi Backes-Gellner[a], and Eric Bettinger[c]

[a]Department of Business Administration, University of Zurich
[b]Department of Geography, University of Zurich
[c]Graduate School of Education, Stanford University

September 2022

*We thank participants of the World Congress of the Regional Science Association International in Marrakech, the Meeting of the Economics of Education Association in Zaragoza, the European Regional Science Association Congress in Bolzano, the Annual Conference of the Verein für Socialpolitik in Cologne, and participants of seminars at the University of Zurich, the Julius-Maximilians-Universität Würzburg, the Max Planck Institute for Innovation and Competition in Munich, and the ZEW Leibniz Center for European Economic Research in Mannheim. We thank Simone Balestra, Thomas Dohmen, Tor Eriksson, David Figlio, Dietmar Harhoff, Simon Janssen, Philip Jörg, Daniel Kükenbrink, Edward Lazear, Mark Long, Jens Mohrenweiser, Guido Neidhöfer, Harald Pfeifer, Natalie Reid, Michael E. Rose, Monika Schnitzer, Dinand Webbink, and Niels Westergård-Nielsen for helpful comments. This study was partly funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) through its "Leading House VPET-ECON: A Research Center on the Economics of Education, Firm Behavior and Training Policies". The code developed in this paper and the surface groups used for the analyses will be made public upon publication. Researchers interested in using our data can contact the authors at patrick.lehnert@business.uzh.ch.

†Corresponding author. Address: Patrick Lehnert, University of Zurich, Plattenstrasse 14, CH-8032 Zurich, Switzerland. E-Mail: patrick.lehnert@business.uzh.ch

Abstract

This paper develops a novel procedure for proxying economic activity with daytime satellite imagery across time periods and spatial units, for which reliable data on economic activity are otherwise not available. In developing this unique proxy, we apply machine-learning techniques to a historical time series of daytime satellite imagery dating back to 1984. Compared to satellite data on night light intensity, another common economic proxy, our proxy more precisely predicts economic activity at smaller regional levels and over longer time horizons. We demonstrate our measure's usefulness for the example of Germany, where East German data on economic activity are unavailable for detailed regional levels and historical time series. Our procedure is generalizable to any region in the world, and it has great potential for analyzing historical economic developments, evaluating local policy reforms, and controlling for economic activity at highly disaggregated regional levels in econometric applications.

Keywords: Daytime satellite imagery, Landsat, machine learning, economic activity, land cover

JEL Classification Numbers: E01, E23, O18, R11, R14

1 Introduction

The lack of credible data hampers our understanding of regional economic development, especially in historical contexts. Most countries lack data at the regional or even municipal levels, and the extant data either focus only on recent years or lack consistency across regions and/or time. To fill these data gaps across time and space, researchers have increasingly used satellite data on night light intensity as a proxy for economic activity (e.g., 1–4).

However, night light intensity data have significant weaknesses. They are available only for a limited time series (from 1992) and, due to their spatial resolution (one kilometer at the equator), they are not reliable for disaggregated regional units such as municipalities or suburbs (5–7). Administrative or survey data on economic activity encounter similar problems. They are typically not available for longer historical time series, not regionally disaggregated, or otherwise unreliable or unavailable to the research community. In a recent literature review, the lack of long time series and regional scalability have been identified as key weaknesses of former satellite-based metrics (8).

This paper solves these key weaknesses by offering an economic proxy from daytime satellite imagery with worldwide applicability based on a procedure that we developed in 2020 and that applies machine-learning techniques to Landsat imagery (9). The proxy presents valuable information on economic activity over a uniquely long time series (from 1984) at a level of regional disaggregation that is smaller (30-meter resolution) than any alternative.

Daytime satellite imagery from the Landsat program has so far received almost no attention in economics applications. The few existing applications rely on visual interpretation for identifying, for example, agricultural land use, (de)forestation, or urbanization (e.g., 10, 11). Developing new and more accurate proxies with Landsat data requires novel machine-learning techniques to adapt these data to economic settings. Tools such as the Google Earth Engine facilitate the processing and analysis of Landsat's large geographic datasets (12).

Landsat daytime satellite data have three advantages over other data sources. First, Landsat data have substantially higher disaggregation (30-meter resolution) than regional administrative or other satellite data such as night light intensity (one-kilometer resolution) (13). This higher resolution entails more precise information at a much more disaggregated regional level. Our economic proxy can characterize economic development at regional levels and even in much smaller localities such as municipalities or urban districts.

Second, NASA launched the first satellite of the Landsat program (Landsat-1) in 1972, making Landsat the earliest existing source of regionally highly disaggregated satellite data (14, 15). While Landsat did not reach its full potential until 1984, the long time horizon of the data allows researchers to construct longer historical data than other regional economic administrative data or other proxies based on satellite data such as night light intensity (which is available from 1992). In comparison to regional economic data, which might be available for some administrative locations, Landsat daytime data pre-date the break-up of the former Soviet Union, German Reunification, and other significant changes in regional or even local economic development.

Third, Landsat satellites collect multispectral imagery of the earth, that is, they observe the energy that the earth reflects in different spectral bands (e.g., infrared or ultraviolet). The geographic remote-sensing literature has successfully applied machine-learning techniques that exploit this multispectral information in the identification of different types of land cover from subsets of Landsat data (e.g., 16–20). We extend this literature by creating a procedure that combines all Landsat data available from 1984 to map six different types of land cover, which we refer to as *surface groups*: built-up surfaces, grassy surfaces, forest-covered surfaces, surfaces with crop fields, surfaces without vegetation, and water surfaces.

As some surface groups are more closely related to economic activity than others (21, 22), mapping surface groups yields important information on regional economic activity. For example, increases in built-up surfaces, which include agglomerations of cities or transportation networks, coincide with increases in economic activity (23, 24). Even holding built-up surfaces constant, the other surface groups provide greater predictability of local economic conditions. While previous research finds that the raw spectral values of Landsat-7 imagery can serve as a slightly better proxy than night light intensity in Vietnamese regions (25), we show that identifying the different surface groups through machine-learning techniques results in a substantially improved proxy for economic activity over time and space. Moreover, compared to a previous application that uses Landsat imagery to directly predict village asset wealth in Africa (26), our surface groups can function both as an indicator of land cover and as a more general proxy for economic activity with the potential for worldwide application. Which proxy to choose for empirical research depends on the concrete research question, with other proxies offering advantages through specialization in, for example, asset wealth (26) and our proxy offering advantages through painting an overall picture of regional (or even more subregional) economic activity.

Our procedure for detecting surface groups as a proxy for economic activity produces a metric with high internal and external validity. We lay the foundations for computing and validating the proxy using Germany as an example. In the context of the German Reunification, our proxy provides important, previously unavailable, yet reliable information on economic activity in East German regions. As such, the surface groups allow the examination of pre-reunification economic developments at highly disaggregated regional levels and—due to their independence of politically motivated economic statistics produced during the communist era—with very high validity. These analyses are otherwise impossible with other data. Our data and their applications easily extend to other settings and geographies throughout the world.

2 The value of surface groups as a proxy for economic activity

2.1 Features of surface groups

We use a supervised machine-learning algorithm to classify Landsat pixels into one of six surface groups. This classification procedure requires two external data sources. First, the raw imagery of Landsat satellites constitutes the input data to be classified. Before performing the classification, we pre-process this raw imagery to obtain pixel-based annual composites incorporating imagery from multiple Landsat satellites. Second, CORINE Land Cover (CLC) data (which are available only for the five reference years 1990, 2000, 2006, 2012, and 2018) serve as an external source of ground-truth information, that is, they indicate the true surface group for a subset of the input pixels. The training data for the classification algorithm consist of a stratified random sample of Landsat pixels matched to their true surface group from CLC data. The details of the classification procedure are outlined in Appendix A1.

Following prior literature utilizing land cover classifications (e.g., 27–31), we identify and map six different types of land cover—the surface groups—which are similar to previous work in a Chinese region (20). These groups include the following: 1) built-up surfaces feature buildings of non-natural materials such as concrete, metal, and glass (e.g., residential buildings, industrial plants, roads); 2) grassy surfaces are covered by green plants or groundcover with similar surface reflectance (e.g., natural grassland); 3) surfaces with crop fields include vegetation for agricultural purposes (e.g., grain fields); 4) forest-covered surfaces contain trees or other plants with similar surface reflectance (e.g., mixed forests); 5) surfaces without vegetation have (almost) no reflective vegetation or buildings (e.g., bare rock); and 6) water surfaces comprise any type of water surface (e.g., lakes). Our algorithm classifies these respective surfaces, which we then combine to form our proxy for economic activity.

The output of our procedure for detecting surface groups is a dataset containing the surface group of every Landsat pixel location in Germany annually from 1984 through 2020. One year comprises more than 630 million Landsat pixels, amounting to more than 23 billion pixel-year observations in the output data. Of these observations, 16.2% are classified as built-up; 20.9% as grass; 29.5% as crops; 25.6% as forest; 3.3% as no vegetation; and 3.8% as water. Only 0.6% of observations contain missing values due to, for example, cloud cover that is uninterrupted within a given year for single pixels in the Landsat data. For applications in research projects, researchers can aggregate this pixel-level information to the geographical unit matching their respective research objective (e.g., administrative regional units or ZIP code areas).

Fig. 1 illustrates the data sources we use and the output data we produce. As examples, the left column of Fig. 1 shows a large-scale area with the metropolitan region of *Nuremberg* (situated in mid-south Germany) in the center of the picture. The right column shows a small-scale area with the village of *Muhr-am-See* (*Muhr-at-the-lake*, situated about 30 miles south-west of *Nuremberg*) in the upper part of the picture and its accompanying lake (*Altmühlsee*) in the lower part of the picture (the area framed red in the left column). Fig. 1 *A*, which uses Landsat's visible spectral bands to approximate the perception of the human eye, shows the Landsat composite for 2018 (the input data). Fig. 1 *B* illustrates the six different types of land cover we identify from the CLC data (the ground-truth data). Fig. 1 *C* shows the surface group that our classification algorithm produces for every Landsat pixel location in 2018. As a reference, Fig. 1 *D* shows current high-resolution satellite images from Esri World Imagery (32).

2.2 Internal validity

To evaluate whether we achieve an accurate classification of Landsat pixels into the six surface groups (i.e., the measure's internal validity), we assess several indicators of prediction accuracy. In so doing, we follow the standard procedure in the remote-sensing literature that uses supervised machine learning to classify land cover (e.g., 16, 17, 33) and derive these indicators from five-fold cross-validation. This method draws five subsets from the input data and uses these subsets to perform five iterations of pixel classification (see appendix A1.5 for more details).

Using the classification output from the five-fold cross-validation, we calculate five

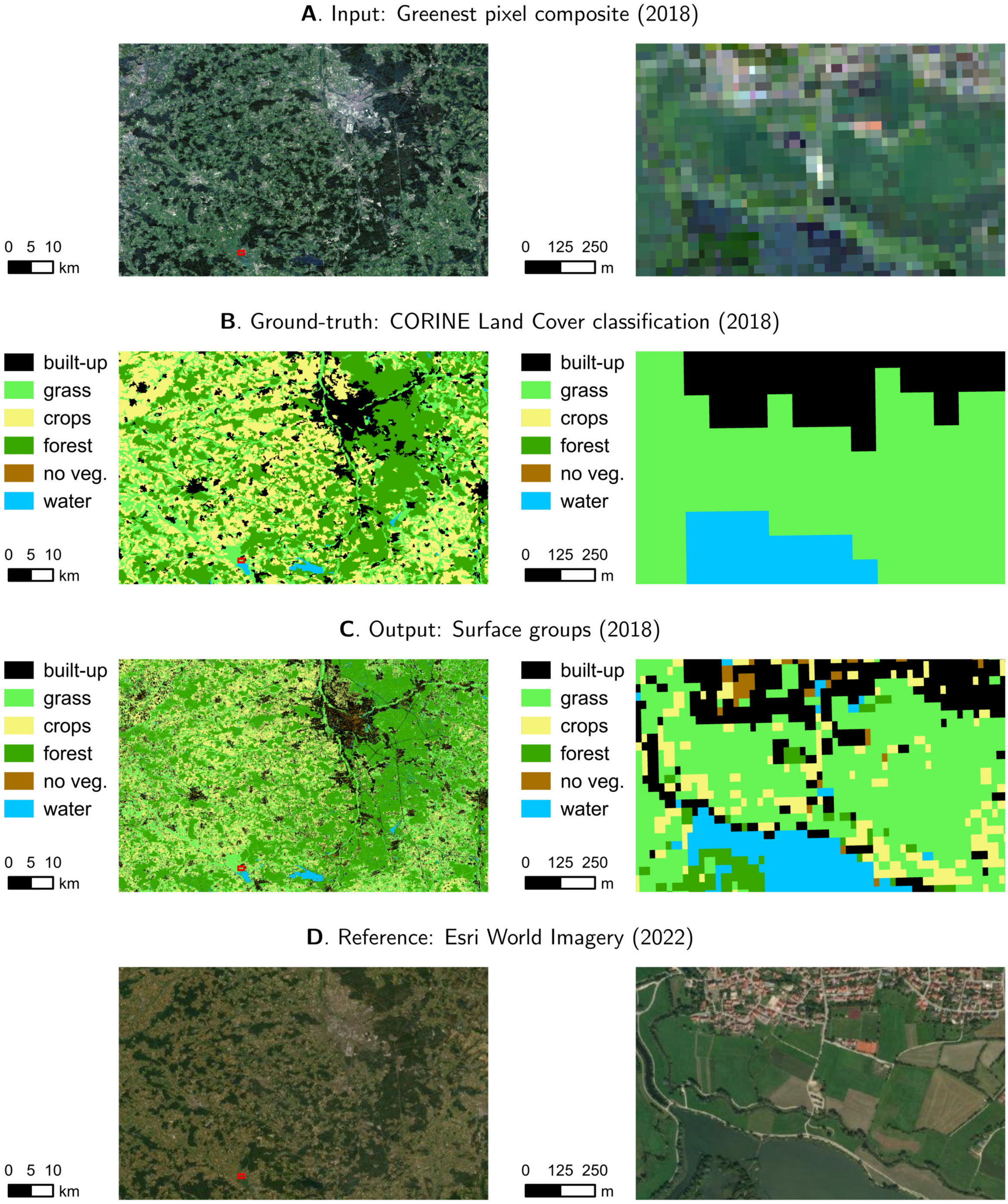

Fig. 1. Visual comparison of data sources. Pictures in the left column show the same approx. 78 $\times$ 49 square kilometers area with the metropolitan region of *Nuremberg* in the center. Pictures in the right column show the same approx. 1.3 $\times$ 0.8 square kilometers area with the village of *Muhr-am-See* in the upper part and its accompanying lake (*Altmühlsee*) in the lower part (area framed in red in the left column).

common indicators of prediction accuracy with respect to each surface group: overall accuracy, true-positive rate, true-negative rate, balanced accuracy, and user's accuracy. Overall accuracy denotes the percentage of pixels correctly classified, true-positive rate the percentage of pixels correctly classified as belonging to the respective surface group, true-negative rate the percentage of pixels correctly classified as not belonging to the respective surface group, balanced accuracy the average of true-positive rate and true-negative rate, and user's accuracy the percentage of pixels correctly classified as belonging to the respective surface group among all pixels belonging to the respective surface group.

Table 1 shows the five-fold cross-validation results with respect to each surface group. With 82.8%, overall accuracy for built-up surface areas is similar to that in other studies detecting built-up land with Landsat data (e.g., 16, 17). The other four indicators are also in line with other studies (e.g., 17, 33). Furthermore, we achieve very high overall accuracy for forest (89.5%), areas with no vegetation (87.0%), and water (90.9%).

The five-fold cross-validation results show that our output data constitute an internally valid measure of land cover. All indicators of prediction accuracy reinforce that our classification algorithm accurately identifies the six surface groups, suggesting that we adequately implemented the procedures from the remote-sensing literature. The high internal validity of the surface groups is a prerequisite for their external validity as a proxy for economic activity.

2.3 External validity

To evaluate the external validity of surface groups as a proxy for economic activity, we empirically analyze how much they explain of the variation in direct measures of regional economic activity (which are available for parts of our time series). We draw on two such external direct measures: First, from administrative statistics, we extract a regionally disaggregated direct measure of gross domestic product (GDP), the most commonly used economic indicator in the literature evaluating previous satellite-based proxies for economic activity (e.g., 5, 34). For Germany, this measure is available at the administrative county (*Kreis*) level from 2000. Second, we use a socioeconomic dataset that provides household income as a further indicator of economic activity with a very high level of regional detail (35). This indicator is available at the level of grid cells sized one square kilometer (and thus independent of administrative borders), but annually only from 2009. See Appendix A2.2 for more details on the two external validation data sources.

We analyze the external validity of our proxy by comparing the amount of variation in GDP that our proxy and night light intensity generate. We obtain this result from comparing Ordinary Least Squares (OLS) regressions of GDP on the surface groups with OLS regressions of GDP on night light intensity (see appendix A2.3 for more details on the methodology). Our preferred surface-groups specification, which additionally includes year and federal state fixed effects to cancel out any bias due to potential measurement error in the dependent or independent variables, explains 62.3% of the variation in GDP. Using night light intensity instead of surface groups in the same specification explains only 47.1% of this variation, that is, our proxy achieves 32.3% higher precision than previous data at the disaggregated regional level of counties.

The value of surface groups as a proxy for regional economic activity becomes even more obvious at the very small regional level of grid cells. In a similar OLS analysis, our preferred surface-groups specification explains a much larger percentage of the variation

Table 1. Five-fold cross-validation results

Surface group	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
built-up	0.828	0.606	0.877	0.741	0.514
grass	0.831	0.451	0.910	0.680	0.511
crops	0.832	0.381	0.932	0.657	0.563
forest	0.895	0.685	0.938	0.812	0.708
no veg.	0.870	0.756	0.886	0.821	0.490
water	0.909	0.672	0.958	0.815	0.765

Indicators calculated with respect to each surface group. Values indicate the average over all five iterations and all five reference years in the CLC data. See Appendix A1.5 for more details (including the results separately for every reference year).

in household income than the corresponding night-lights specification, with 67.5% vs. 30.7% (i.e., 119.9% higher precision).

The value of surface groups in comparison to night light intensity as a proxy for economic activity thus substantially increases with the degree of regional disaggregation. This finding is supported by an additional analysis on the prediction of county-level GDP by county-size category (see appendix A2.3). On average, the surface groups explain a larger percentage of the variation in GDP for smaller counties than for larger counties.

Fig. 2 underscores and visualizes these findings. It plots the statistical distribution of the OLS regression residuals, which are smaller when the measure is a better proxy for economic activity. The plots show that, for both GDP and household income, this distribution is smoother and narrower for surface groups (figs. 2 *A* and *C*) than for night light intensity (figs. 2 *B* and *D*). For household income, the residual distribution of the night-lights specification even exhibits a plateau—instead of a real peak—around the value zero, whereas the surface groups show a very clear peak and a narrow residual distribution.

Furthermore, we conduct four additional validation analyses in the supplementary material (text S2). First, we find that surface groups are a temporally and spatially less biased proxy for economic activity than night light intensity. This feature is important for the surface groups to serve as a valid proxy for comparisons of economic activity over time and between regions. Temporal bias would occur if the OLS residual is constant for a given region throughout all observation years, and spatial bias would occur if this residual is equal for clusters of regions. Surface groups yield a considerably smaller temporal bias that outweighs their somewhat larger spatial bias in comparison to night light intensity. Second, in line with their smaller bias, surface groups offer more information on within-region changes in economic activity than night light intensity through higher within-region heterogeneity. That is, our surface groups allow for a more precise determination of which subregional units drive the change in a region's economic activity by isolating the change in each subregional unit. Third, surface groups outperform also newer night light intensity data with higher spatial resolution in proxying economic activity. Fourth, we validate surface groups as a proxy for economic conditions in developing countries by comparing their predictive power to that of a prior metric of village asset wealth in Africa (26), a similar but more specialized outcome variable. This analysis shows that the validity of surface groups is not restricted to developed European countries such as Germany, but that surface groups can also provide valuable insights on economic conditions in developing countries across the world.

2.4 Surface groups economic proxy

As having one single proxy may be desirable when economic activity is the dependent variable in an analysis, we compute predicted county-level GDP using our OLS model. To assess the external validity of this single-variable proxy, we use one half of the sample to train the coefficients showing the predictive power of our surface groups proxy, and then for the second half of our sample compute predicted GDP. Corroborating the results of the first analysis of external validity, GDP predicted using surface groups explains 63.4% of the variation in actual GDP in the second half of the sample, whereas GDP predicted using night light intensity explains only 48.8% of this variation (i.e., 29.9% higher precision). The corresponding values for household income are 67.4% using surface groups vs. 30.8% using night light intensity (i.e., 118.8% higher precision). However,

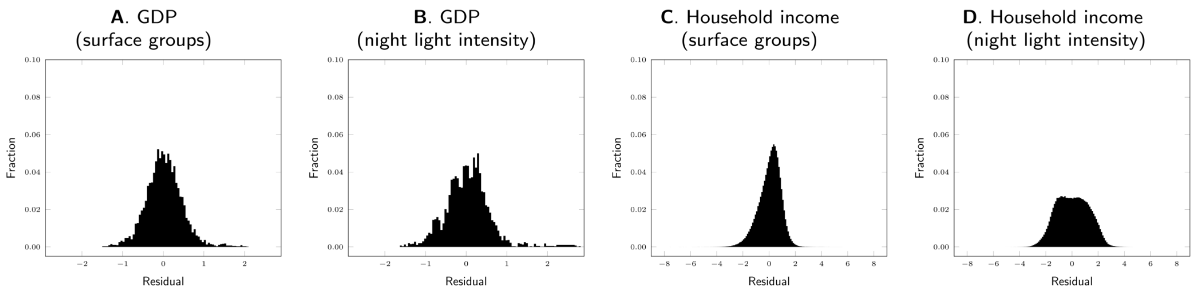

Fig. 2. Statistical distribution of OLS regression residuals. See Appendix A2.3 for details on the regression specifications. Bin width of histograms is 0.05 in panels A and B and 0.1 in panels C and D.

when using the proxy as an independent variable, we recommend using the full set of proxy variables to minimize the noise and measurement error that might come from the predictive process.

Finally, Fig. 3 demonstrates the usefulness of our economic proxy. The curves marked with triangles show the extant data for regional economic activity in four regions of Germany—including areas in both East Germany (*Rostock*, *Börde*) and West Germany (*Groß-Gerau*, *Passau*). The thicker curves without triangles show our single-variable proxy for economic activity (with OLS coefficients trained on the entire sample) for its available years. Starting in 1984, the improved coverage achieved through surface groups almost doubles the number of available years compared to administrative data (which start in 2000 for Germany). Compared to other proxies such as night light intensity (which starts in 1992), surface groups are the only proxy pre-dating the German Reunification. To better visualize trends over time, Fig. 3 plots the three-year moving average of administrative GDP and the surface groups proxy. While the variation between years in the surface groups proxy is larger than in the administrative metric, all curves exhibit identical trends over time. The longer time series and the differences in the developments of the regions over time (e.g., GDP in *Börde* falls below that in *Rostock* after reunification) emphasize the proxy's potential for enabling previously impossible analyses.

3 Conclusion and discussion

As Fig. 3 demonstrates, the proxy we create from daytime satellite imagery is a strong proxy across time periods and across highly disaggregated regional levels, for which other data are unreliable, inaccessible, or entirely inexistent. Moreover, in this particular example, the proxy provides valuable, previously unavailable information on economic activity for East German regions before the fall of the iron curtain.

More generally, our procedure has worldwide relevance. While we apply our procedure to Germany and establish its validity for this country, the procedure is transferable to any region or country in the world (as we demonstrate in appendices A1.6 and A2.4). Our analyses for Germany exemplify that our machine-learning approach using daytime satellite imagery can predict both disaggregated and potentially missing or erroneous economic activity data (e.g., GDP at highly disaggregated levels within a country). However, the methodology and the data it provides for countries across the world can be extended globally to additional contexts where specific economic and developmental markers are needed. Our insight is to demonstrate that our methodology can be helpful for many economic and social science applications where varying degrees of disaggregation are required and where missing or incorrect data are prevalent. Surface groups thus constitute a valuable resource for analyzing historical developments, evaluating local policy reforms, and controlling for economic activity in econometric applications within a country. Although a country's history or industry structure affects the economic importance of different types of land cover (36), the principle that land cover, which the surface groups reflect, relates to economic activity applies to any country in the world. Therefore, surface groups have a potential for economic research that investigates small regions within the same country or within a homogeneous group of countries.

The Landsat daytime satellite data are available for extremely small regional units such as municipalities or urban districts, thus providing new opportunities for urban and regional economic researchers to understand differences in even small regional variation

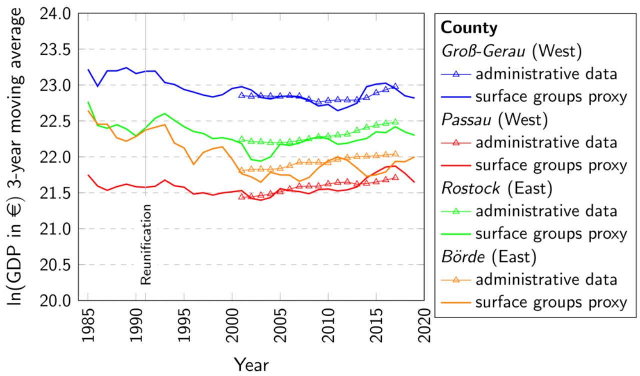

Fig. 3. Time series of GDP measures in four counties. Plots show three-year moving averages. Curves marked by triangles show the natural logarithm of GDP in administrative data for all years for which county-level administrative GDP data are available. Thick curves without triangles show the surface groups proxy for GDP (predicted from OLS estimates, see appendix A2.5 for details). *Groß-Gerau* is situated in mid-west Germany, *Passau* in south Germany (at the border to Austria), *Rostock* in north Germany (at the Baltic Sea), and *Börde* in mid-north Germany.

in economic development. The surface groups we derive from these data thus contribute to analyses of the regional impacts of local policy reforms by providing information on economic activity at very detailed regional levels, for which other data sources are entirely unavailable for the necessary observation period, unreliable, less precise, or inaccessible for non-residents of the respective country. With these particular features, the surface groups complement other satellite-based measures for economic activity such as night light intensity.

The use of satellite data is a significant advancement in measuring regional economic activity and over time will generate new opportunities to strengthen our understanding of local economic conditions. For example, as the methods for utilizing daytime satellite data advance, researchers will eventually be able to analyze high-resolution satellite data with image recognition procedures to discern more about the nature of built-up surfaces. Such analyses could, for example, identify stores or industrial buildings, evaluate neighborhood housing quality, or determine when buildings have been renovated. However, retrieving more sophisticated metrics on economic activity requires satellite data with an even finer spatial resolution than Landsat data, such as the Advanced Spaceborne Thermal Emissions and Reflection Radiometer (ASTER) or the Sentinel mission. These or other satellite data offer promising venues for future research, for which this paper lays first methodological foundations. However, the ASTER and Sentinel data cover substantially shorter time series than Landsat and are thus not as valuable for historical analyses.

Appendix

This appendix provides the technical details of retrieving our surface groups measure as a proxy for economic activity. It presents all procedures and analyses referred to in the main text and the underlying data.

In Section A1, we describe the procedure we develop for retrieving the surface groups measure from daytime satellite imagery and conduct the internal validity analysis. In Section A2, we perform several analyses to demonstrate the value of surface groups as a proxy for economic activity.

A1 Computation of surface groups

A1.1 Overview

This section describes our procedure for detecting surface groups as a novel proxy for economic activity at detailed regional levels. In developing this procedure, we follow the remote-sensing literature, which has successfully applied machine-learning techniques to identifying, for example, built-up land cover from subsets of Landsat data (e.g., 18, 37). Our procedure adds to this literature by combining data from four Landsat satellites to produce a time series of data on different types of land cover starting in 1984. To produce these data, we use GEE as a platform and apply supervised machine-learning techniques with the objective of classifying the annual type of land cover of every Landsat pixel location in Germany. We proceed in four steps that Fig. A1 illustrates.

First, we prepare the Landsat data to retrieve the input data for the classification algorithm. We combine the data of four Landsat satellites (Landsat-4, Landsat-5, Landsat-7, and Landsat-8) to produce composite data containing the qualitatively best observation per pixel location and year.[1] As we choose those observations that best differentiate between vegetated and unvegetated areas for this composite, we refer to it as "greenest" pixel composite. This greenest pixel composite constitutes the input data that we pass on to the classification algorithm.

Second, to be able to classify the observations in the greenest pixel composite, we add CORINE Land Cover (CLC)[2] data as an external source of ground-truth information. These data, which come from a pan-European project commissioned by the European Environment Agency (EEA),[3] map land cover in European countries for five reference years (1990, 2000, 2006, 2012, 2018). Based on a survey of the literature that applies land cover classifications (e.g., 20, 31), we obtain from the CLC data the six different types of land cover that we refer to as *surface groups*: built-up surfaces (*builtup*), grassy surfaces (*grass*), surfaces with crop fields (*crops*), forest-covered surfaces (*forest*), surfaces without vegetation (*noveg*), and water surfaces (*water*). The classification algorithm requires this ground-truth information on surface groups to be able to recognize patterns in the input data and link these patterns to the different surface groups. For example, the spectral values of an input pixel showing a grassy surface differ from those of an input pixel showing a built-up surface. The CLC data provide the classification algorithm with the true surface group for a subset of the input pixels. By using external ground-truth

[1] We use the Landsat Collections distributed by the U.S. Geological Survey (38) and directly accessible through GEE.

[2] The acronym "CORINE" stands for "coordination of information on the environment" (39).

[3] The CLC data are distributed by the EEA (40) and directly accessible through GEE.

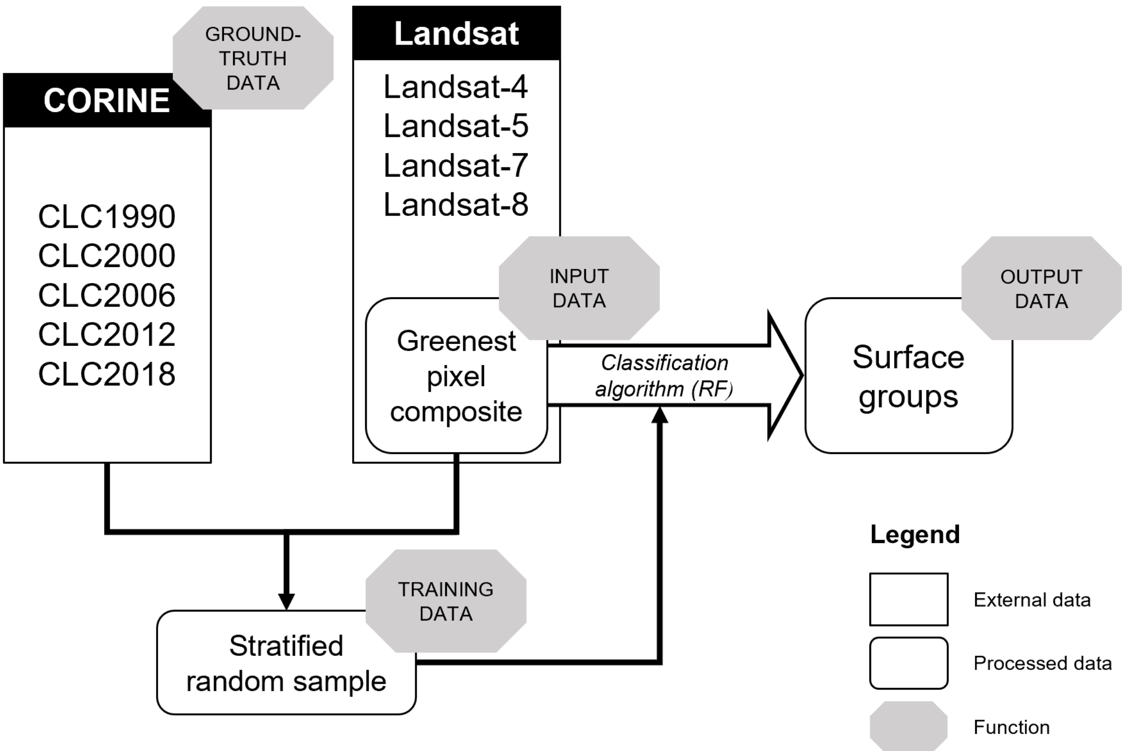

Fig. A1. Overview of procedure for detecting surface groups.

data, we overcome the resource-intensive necessity of visually interpreting (i.e., manually classifying) input pixels to retrieve ground-truth information.

Third, we produce the training data for the classification algorithm. To obtain these training data, we draw a stratified random sample of pixels from the greenest pixel composite and match the CLC ground-truth information on surface groups to the pixels in this sample. We then use the training data to train the classification algorithm, which is a Random Forest (RF) algorithm with ten decision trees. After training the algorithm, it classifies every observation in the greenest pixel composite into one of the six surface groups.

Fourth, the classification result is the output data that contain the surface group of every Landsat pixel location annually from 1984 through 2020. To assess the accuracy of the classification (i.e., the internal validity), we perform five-fold cross-validation.

A1.2 Greenest pixel composite of Landsat data as input data

Satellite data from the Landsat program serve as input data for the machine-learning procedure for detecting surface groups. Since 1972, Landsat satellites have continuously recorded remotely sensed imagery of the earth, providing a unique basis for various applications in mapping and monitoring land cover (41, 42). Throughout the history of Landsat, the various operating agencies have launched eight satellites, one of which (Landsat-6) failed to reach orbit (15, 43). As of 2022, Landsat-7, Landsat-8, and Landsat-9 remain active, with Landsat-9 having launched only in September 2021 (44, 45).[4]

We gather the input data for our algorithm to detect surface groups from the spectral information that Landsat satellites capture. Every Landsat satellite carries sensors that remotely measure the spectral reflectance of the earth's surface (49). The improving technical specifications of these sensors from one satellite generation to the next entail an increase in the number of spectral bands that each satellite captures (51). Table A1 provides the technical specifications of the different sensors that Landsat satellites carry, including their spectral resolution, years of operation, and wavelengths of the spectral bands that the sensors capture.

We use information in the six spectral bands that the sensors on Landsat-4, Landsat-5, Landsat-7, and Landsat-8 have in common (highlighted gray in table A1). These bands contain the surface reflectance in the visible blue (BLUE), visible green (GREEN), visible red (RED), short-wave infrared (SWIR1 and SWIR2), and near-infrared (NIR) ranges of the electromagnetic spectrum. Consequently, we begin our observation period with the 1982 launch of Landsat-4. However, due to a series of technical failures throughout the lifetime of Landsat-4 (52) and the resulting scarcity of Landsat-4 imagery for Germany, the effective start of our observation period is 1984 (although we include Landsat-4 imagery in later years whenever available).

We exclude imagery from the pre-Landsat-4 period and information in the thermal infrared spectral bands for the following reasons. We exclude pre-Landsat-4 satellites because they differ substantially from their successors in captured wavelength and in spatial resolution (14). Therefore, when combining all sensors, we cannot achieve a consistent pixel classification, which is a prerequisite for a valid economic measure. Moreover, due to technological and organizational constraints at the time, imagery in the Landsat archives

[4] The remote-sensing literature and related disciplines have applied Landsat data for numerous purposes, for example, the assessment of water conditions in the Bahamas (46) and the investigation of tree species diversity in the Alps (47).

Table A1. Technical specifications of Landsat sensors

Sensor	Multispectral Scanner (MSS)	Thematic Mapper (TM)	Enhanced Thematic Mapper Plus (ETM+)	Operational Land Imager (OLI) / Thermal Infrared Sensor (TIRS)
Spatial resolution	79 meters	30 meters	30 meters	30 meters
Satellites (Operating Years)	Landsat-1 (1972–1978) Landsat-2 (1975–1982) Landsat-3 (1978–1983) Landsat-4 (1982–1993) Landsat-5 (1984–1995)	Landsat-4 (1982–1993) Landsat-5 (1984–2014)	Landsat-7 (1999–present)	Landsat-8 (2013–present)
Band name	Wavelength (in µm)			
Ultra blue				0.43–0.45
Visible blue (BLUE)		0.45–0.52	0.45–0.52	0.45–0.51
Visible green (GREEN)	0.50–0.60	0.52–0.60	0.52–0.60	0.53–0.59
Visible red (RED)	0.60–0.70	0.63–0.69	0.63–0.69	0.64–0.67
Short-wave infrared 1 (SWIR1)		1.55–1.75	1.55–1.75	1.57–1.65
Short-wave infrared 2 (SWIR2)		2.08–2.35	2.08–2.35	2.11–2.29
Near-infrared 1 (NIR)	0.70–0.80	0.76–0.90	0.77–0.90	0.85–0.88
Near-infrared 2	0.80–1.10			
Thermal infrared 1		10.40–12.50 (120-meter resolution)	10.40–12.50 (60-meter resolution)	10.60–11.19 (100-meter resolution)
Thermal infrared 2				11.50–12.51 (100-meter resolution)
Panchromatic			0.52–0.90 (15-meter resolution)	0.50–0.68 (15-meter resolution)
Cirrus				1.36–1.38

Authors' representation based on previous representations (15, 17, 43, 48–50). Spectral bands used for detecting surface groups highlighted gray. The table excludes technical details that are beyond the scope of this paper. For example, the MSS sensor of Landsat-5 was decommissioned in 1995, but the MSS archives only contain data until the sensor became unable to relay data in 1992 (43, 48). OLI and TIRS, the sensors that Landsat-8 carries, are two separate sensors, with the TIRS capturing the two thermal infrared bands and OLI the remaining ones (50).

is scarce for Germany until the 1980s (43). This scarcity of imagery makes the detection of surface groups unfeasible for the pre-Landsat-4 period, regardless of the sensors the satellites carried. Furthermore, we do not use the thermal infrared spectral bands because their technical specifications change over time and differ from the remaining bands (e.g., coarser spatial resolution, different numbers of bands, see table A1). In addition, the bands' specifications notwithstanding, temperatures in Germany vary over the seasons so that thermal information would be of little help for detecting surface groups.

As with the night light intensity data that economists commonly use (13), we compute the surface groups annually. As Landsat satellites record a geographic location on earth multiple times per year (43), we have to use annual composites of these records. Unfortunately, pre-processed annual composites incorporating imagery from multiple Landsat satellites do not exist, requiring us to produce such composites from the available images and use these composites as input data for our algorithm.

We produce pixel-based annual composites of Landsat images. Among all available observations of a given pixel within a year, we choose the one pixel that best serves the purpose of detecting surface groups. This pixel-based compositing procedure (as compared to scene-based compositing) prevents a loss of information due to, for example, cloud-covered pixels and enables the researcher to choose those pixels best suitable for a specific application—in our case, the detection of surface groups (53). Given the long time span that we analyze, the production of annual composites also entails less computational effort than other approaches such as data stacking (54).

For both the compositing and the actual pixel classification (see section A1.4), we follow studies from the remote-sensing literature (e.g., 17, 18) and add three indices to the data: First, the Normalized Difference Vegetation Index (NDVI) differentiates vegetated from unvegetated surfaces and is one of the most frequently used indices in the remote-sensing literature (55, 56); Second, the Normalized Difference Water Index (NDWI) differentiates open water from other surfaces (57);[5] Third, the Normalized Difference Built-up Index (NDBI) differentiates built-up surfaces from other surfaces (59). Similar to prior work (18), we compute these three indices for Landsat data as follows:

$$NDVI_p = \frac{NIR_p - RED_p}{NIR_p + RED_p} \tag{A1}$$

$$NDWI_p = \frac{GREEN_p - NIR_p}{GREEN_p + NIR_p} \tag{A2}$$

$$NDBI_p = \frac{SWIR1_p - NIR_p}{SWIR1_p + NIR_p} \tag{A3}$$

with p denoting pixels as the unit of observation.

For the compositing of Landsat images, we proceed in three steps. First, we collect all images available within a given calendar year for Germany, our study region. We restrict the pool of images to those taken between March and November, that is, we exclude the meteorological winter months in the northern hemisphere. We do so because the potential snow cover and the absence of large parts of the vegetation during winter might confuse the machine-learning algorithm. Second, we drop pixels showing clouds or cloud shadow and pixels with implausible values in one of the spectral bands. Clouds

[5]Another index exists under the name "NDWI", which was developed to identify liquid water inside plants (58). This other NDWI relies on different spectral bands than the NDWI we use.

obscure the actual surface we want to observe, and cloud shadow distorts a pixel's actual reflectance, whereas a pixel with clear vision does not (e.g., 60). Implausible values, such as a negative reflectance in one of the spectral bands, might result from erroneous data transmission. Third, among the remaining pixels we choose the best one available. In so doing, we emphasize the distinction of built-up land from other surfaces, because—as with the logic underlying the use of night light intensity as a proxy for gross domestic product (GDP)—we expect economic activity to concentrate in urban or industrial areas. Therefore, a clear distinction between built-up surfaces and other surfaces will improve our proxy for economic activity.

Our procedure of compositing Landsat data provides us with a greenest pixel composite that we can use as input data for the machine-learning algorithm. This composite covers the geographical area of Germany and consists of one observation per pixel for every year since 1984. The variables in the dataset are the pixel's values in the six spectral bands we use in this paper (see table A1) and the added indices NDVI, NDWI, and NDBI. If the compositing procedure cannot identify a valid observation for a pixel location within a calendar year (e.g., if all available pixels show clouds), the data contain missing values. Fig. 1 *A* in the paper visualizes the greenest pixel composite with the visible spectral bands BLUE, GREEN, and RED for 2018.

A1.3 CLC data as ground-truth data

To retrieve ground-truth information for a subset of the greenest pixel composite, we use CLC data. The European Commission began the CORINE program that produces these data in 1985, with the goal of creating a standardized database on land cover to support policymakers in environmental affairs (39, 61). Since then, five phases of the program have produced CLC databases for the five reference years 1990, 2000, 2006, 2012, and 2018 (hereafter denoted as CLC1990, CLC2000, CLC2006, CLC2012, and CLC2018) (62). Each database includes a map for the respective year with a pixel resolution of 100 meters, indicating land cover in a variety of classes (62, 63).

Although the medium underlying the classification changed over the years from hardcopies to computer-assisted technologies, classification still relies mainly on visual interpretation of satellite imagery by professional experts (62, 63). This imagery stems from various satellites, including Landsat satellites for CLC1990, CLC2000, and CLC2018 (62). The remote-sensing literature provides successful combinations of CLC and Landsat data in geospatial analyses (e.g., 19, 64).

To train our machine-learning algorithm, we exploit the CLC data as a source of ground-truth information for three reasons. First, the earliest of the CLC data's five reference years (1990) still falls within the operating time of Landsat-4 (1982–1993), the oldest Landsat satellite we use in our computations (see section A1.2). This time overlap improves the prediction of surface groups by providing a better temporal fit of ground-truth data and input data. Second, although with 100 meters the spatial resolution of CLC pixels is lower than that of Landsat pixels, CLC pixels still have a much higher resolution than other external ground-truth data used in the remote-sensing literature (e.g., night light intensity data with a resolution of one kilometer in 33). This high resolution improves the prediction of surface groups by providing a better spatial fit of ground-truth data and input data. Third, the CLC data provide a detailed classification of surfaces, allowing us to distinguish between various types of surfaces, such as built-up land, forests, or water. In sum, the CLC data constitute an excellent external source of ground-truth information for the purpose of detecting surface groups.

The CLC classification consists of five larger groups (level 1), which are further subdivided into 15 subgroups (level 2) and 44 detailed groups (level 3). However, even at levels 1 and 2, this classification simultaneously indicates types of land cover (the land's directly observable terrestrial features) and land use (the land's socioeconomic purpose) (65–68). Given that automated analyses of satellite data can detect only land cover and that determining land use requires manual interpretation (68), we cannot directly apply this classification for the training of our algorithm.

To obtain a classification of land cover types that we can use to train our algorithm, we aggregate the CLC level 3 classes to larger groups with similar surface characteristics. We base this aggregation on a survey of the literature that uses CLC data or Landsat data for classifying land cover (e.g., 27, 28). However, as this literature does not provide an unambiguous assignment of CLC classes to larger groups with similar surface characteristics, we perform repeated trials of our classification procedure with varying assignments of CLC level 3 classes to larger groups. These trials yield the result that a classification consisting of six surface groups, which correspond to the six types of surfaces identified from subsets of Landsat data for the Daqing region in China (20), best represents similar surface characteristics in Germany: *builtup* (following, e.g., 29), *grass* (following, e.g., 29, 31), *crops* (following, e.g., 29, 30), *noveg*, and *water* (following, e.g. 69). These six surface groups into which the classification algorithm divides the input data constitute the basis for our proxy for economic activity. Fig. 1 *B* in the paper visualizes the ground-truth surface groups that we obtain from the CLC2018 data.

A1.4 Training data and classification algorithm

We apply a machine-learning algorithm that classifies the input data of the greenest pixel composite into the six surface groups *builtup*, *grass*, *crops*, *forest*, *noveg*, and *water*. From the input data, we draw a stratified random sample of pixels to train the algorithm and retrieve the corresponding ground-truth information from CLC data. The classifier we use is a RF algorithm with ten decision trees.

Following prior work (17), we perform pixel-based classification. For every pixel in our training sample, the machine-learning algorithm predicts the pixel's surface groups from the spectral values and the added indices NDVI, NDWI, and NDBI. Compared to object-based classification, which additionally considers information from neighboring pixels, pixel-based classification requires considerably less computational power (70, 71). Although the majority of studies in the remote-sensing literature suggest that object-based classification performs better than pixel-based classification (e.g., 71), some studies find no significant performance difference (e.g., 72), and in particular, one of these studies finds no significant difference using Landsat data (73).[6] Therefore, given the spatial and temporal size of the data we analyze in this paper, pixel-based classification is the preferable choice. Our assessments of external validity confirm that choosing this classification yields a valid proxy for economic activity.

To classify the pre-processed Landsat data, we use the RF algorithm with ten decision trees.[7] Several studies in the remote-sensing literature find that RF outperforms other

[6] See, e.g., 74 for a review of the literature on the advantages and disadvantages of pixel-based vs. object-based classification.

[7] For a description of the RF method's application for land cover classification, see, e.g., 75, and for a description of the method's application in economics, see, e.g., 76.

algorithms when applied to land cover classification (e.g., 75, 77). For example, an assessment of the performance of three different algorithms that the remote-sensing literature commonly uses (Classification and Regression Tree, Support Vector Machines, and RF) reveals that RF performs best in predicting built-up land cover in India with Landsat-7 and Landsat-8 data (17). Furthermore, RF requires less computational power (75). As to the number of decision trees, performance increases with the number of trees, although after ten trees the increase is negligibly small relative to the increase in computational power required (17).[8] Therefore, RF with ten decision trees best suits the purpose of our paper.

We draw a stratified random sample of a total of 30,000 pixels to serve as training data for the classification algorithm. For every year in the CLC data (1990, 2000, 2006, 2012, 2018), we randomly choose 1,000 pixels of each surface group. Generally, the number of pixels in the training data correlates positively with prediction accuracy but negatively with computational effort (77, 78). Therefore, we choose a slightly larger number of pixels in the training data than in comparable applications from the remote-sensing literature (e.g., 17, 37) to achieve an accurate classification, but keep this number low enough to maintain a reasonable computational effort. Furthermore, to account for the lower spatial resolution of the CLC data, we do not use Landsat pixels that fall within CLC pixels at the boundary of two CLC surface areas.

A1.5 Accuracy assessment of output data

To assess the prediction accuracy of our classification in the output data, we follow prior work (17) and perform five-fold cross-validation[9] by drawing five subsets from the greenest pixel composite. In drawing the subsets, we apply the same stratification criteria as for the training dataset, with the only difference being that instead of 1,000 pixels per surface group, we now draw only 250. Thus each of the five subsets consists of 7,500 pixels, that is, 250 per surface group and year. For the cross-validation to be valid, the subsets must not overlap. In other words, one pixel can belong to only one subset.

Next, imitating our procedure for generating the output data, we use the five subsets to perform five iterations of pixel classification. During each iteration, we use four of the subsets as a training set. Consequently, every iteration leaves out a different subset, and the training set of four subsets includes precisely the same number of pixels as the training set we actually use for the computations. We train the classification algorithm with the four-subset training set, then classify the left-out subset.

As indicators of prediction accuracy, for every iteration and for each of the six surface groups separately, we calculate overall accuracy, true-positive rate, true-negative rate, balanced accuracy, and user's accuracy (see section 2.2). Complementing the five-fold cross-validation results for the entire sample in Table 1, Tables A2 through A7 show the results separately for every CLC year.

The five-fold cross-validation results show that our output data constitute an internally valid measure of land cover. All indicators of prediction accuracy reveal that our classification algorithm accurately identifies the six surface groups, suggesting that we adequately implemented the procedures from the remote-sensing literature. Therefore,

[8] For example, while the prediction's overall accuracy increases by about two percentage points when increasing the number of trees from three to ten, it increases only by about one more percentage point when increasing the number of trees from ten to 100 (17).

[9] For descriptions and discussions of this method, see, e.g., 79, 80

Table A2. Five-fold cross-validation results with respect to built-up surfaces (surface group *builtup*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.827	0.664	0.859	0.761	0.481
2000	0.838	0.610	0.886	0.748	0.530
2006	0.838	0.593	0.897	0.745	0.585
2012	0.844	0.606	0.887	0.747	0.493
2018	0.795	0.558	0.853	0.705	0.482
Average	0.828	0.606	0.877	0.741	0.514

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

Table A3. Five-fold cross-validation results with respect to grassy surfaces (surface group*grass*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.839	0.468	0.912	0.690	0.514
2000	0.826	0.513	0.891	0.702	0.495
2006	0.823	0.517	0.888	0.703	0.496
2012	0.852	0.428	0.937	0.682	0.575
2018	0.813	0.327	0.921	0.624	0.477
Average	0.831	0.451	0.910	0.680	0.511

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

Table A4. Five-fold cross-validation results with respect to surfaces with crop fields (surface group *crops*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.816	0.461	0.889	0.675	0.458
2000	0.847	0.416	0.937	0.677	0.583
2006	0.828	0.396	0.931	0.664	0.581
2012	0.852	0.348	0.955	0.652	0.611
2018	0.817	0.281	0.950	0.615	0.583
Average	0.832	0.381	0.932	0.657	0.563

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

Table A5. Five-fold cross-validation results with respect to forest-covered surfaces (surface group *forest*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.892	0.509	0.969	0.739	0.771
2000	0.899	0.725	0.936	0.830	0.701
2006	0.886	0.756	0.914	0.835	0.648
2012	0.901	0.711	0.940	0.825	0.713
2018	0.895	0.726	0.933	0.830	0.709
Average	0.895	0.685	0.938	0.812	0.708

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

Table A6. Five-fold cross-validation results with respect to surfaces without vegetation (surface group *noveg*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.890	0.732	0.921	0.827	0.652
2000	0.891	0.754	0.913	0.834	0.585
2006	0.894	0.644	0.918	0.781	0.434
2012	0.850	0.847	0.850	0.849	0.543
2018	0.827	0.801	0.829	0.815	0.236
Average	0.870	0.756	0.886	0.821	0.490

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

Table A7. Five-fold cross-validation results with respect to water surfaces (surface group *water*)

Year Year	Overall accuracy	True-positive rate	True-negative rate	Balanced accuracy	User's accuracy
1990	0.908	0.683	0.952	0.817	0.740
2000	0.902	0.623	0.959	0.791	0.759
2006	0.915	0.693	0.961	0.827	0.787
2012	0.915	0.692	0.959	0.826	0.768
2018	0.905	0.667	0.957	0.812	0.772
Average	0.909	0.672	0.958	0.815	0.765

The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values as indicated in Table 1.

the output data of our algorithm is highly suitable for analyzing whether the surface groups are an externally valid proxy for economic activity in Section A2.[10]

A1.6 Transfer to all countries across the world

Producing our surface groups proxy depends on two external datasets—Landsat imagery (to retrieve the greenest pixel composite as input data) and CLC data (ground-truth data). While Landsat data are available for the entire world,[11] consistent ground-truth data are not. As such, we use two different strategies—one covering the European countries included in the CLC data (CLC countries)[12] and another covering the rest of the world (non-CLC countries)—to retrieve ground-truth data.

Our procedure for detecting surface groups is straightforwardly transferable to CLC countries (i.e., most European countries). For these countries, CLC data include comprehensive and consistent ground-truth information. Therefore, producing the surface groups for any given CLC country works exactly as for our German example. As the data do not cover 1990 (the first of the five CLC reference years) for a few CLC countries, we make one adjustment to the training-sample construction for these countries.[13] In the stratified random sample to serve as training data, we randomly draw 1,250 instead of 1,000 pixels per surface group and year. Consequently, as for CLC countries that cover all five reference years, the training data comprise a total of 30,000 pixels (thus a size identical to that for CLC countries with ground-truth data for all five reference years).

We address the challenge in producing our proxy for non-CLC countries—the selection of adequate ground-truth data from which to draw the training sample—through a selection rule based on the Köppen-Geiger climate classification system (83, 84). At the highest level of aggregation, this system differentiates between five climate zones of the world: tropical (zone A), arid (zone B), temperate (zone C), cold (zone D), and polar (zone E) (83).[14] When classifying surface groups for non-CLC countries, we calculate which percentage of a country's area falls within each of the five climate zones. We then draw a random sample of 30,000 pixels (same size as in the procedure for CLC countries) from all available CLC data (i.e., from all countries participating in CORINE) and stratify the pixel selection by climate zone, that is, for each climate zone the percentage of pixels in the training sample belonging to that climate zone corresponds to the target country's percentage of pixels belonging to that climate zone. For example, if 30 percent of a country's area belong to climate zone C and the remaining 70 percent to climate zone D, the training sample will consist of 9,000 pixels from climate zone C and 21,000

[10] Additional analyses on the correlation between surface groups and administrative measures of land cover also reveal that surface groups validly indicate their corresponding type of land cover in administrative statistics.

[11] For example, Landsat-7 covers any region between the 81.8° north and south latitudes, thus not covering uninhabited places such as Antarctica and the far northern part of Greenland (81)

[12] Countries included in the CLC data are Albania, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Kosovo, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Montenegro, The Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, and the United Kingdom (see 82).

[13] CLC countries without data for the reference year 1990 are Albania, Bosnia and Herzegovina, Cyprus, Finland, Iceland, Kosovo, North Macedonia, Norway, Sweden, Switzerland, and the United Kingdom (see 82).

[14] The Köppen-Geiger climate classification data (from 83) are publicly available at https://doi.org/10.6084/m9.figshare.6396959.

pixels from climate zone D. All other stratification criteria for CLC countries (e.g., same number of pixels per surface group and CLC year) also apply for non-CLC countries.

As none of the CLC countries features the tropical climate zone A, we assign the percentage of a non-CLC target country's area in climate zone A (if any) to CLC pixels in the temperate climate zone C. We do so, because climate zone C is most similar to climate zone A in terms of vegetation (the main selection criterion in constructing our greenest pixel composite from Landsat data as input data). As we restrict the pool of Landsat images for constructing the greenest pixel composite and, consequently, the training data to those images taken between March and November (thus excluding the meteorological winter months in the northern hemisphere), climate zones A and C also have similar temperature levels during the period we consider.

As in the procedure for CLC countries, we exclude Landsat images taken during meteorological winter months in constructing the greenest pixel composite for non-CLC countries. For non-CLC countries in the northern hemisphere, we exclude images taken between December and February (similar to the exclusion for CLC countries), while we exclude images taken between June and August for non-CLC countries in the southern hemisphere. For countries within the Tropic of Cancer and the Tropic of Capricorn we do not exclude any images, because temperatures (and thus vegetation) in these countries stay almost constant over the seasons.

The procedure for producing surface groups for non-CLC countries also offers the flexibility of classifying Landsat pixels only for subregions of a country, with all steps of our classification procedure (i.e., draw of training sample, training of algorithm, and classification of pixels in the Landsat greenest pixel composite) taking place for each subregion separately. Such a separation of subregions can be useful for large countries with differences in vegetation and climate across subregions. For example, splitting the U.S. by states could improve the classification output because the states differ substantially in terms of vegetation and climate. Moreover, the average area size of a U.S. state roughly equals that of a CLC country, so that through splitting the U.S. into states the proportion of training data size and size of the greenest pixel composite would stay constant, thus potentially improving the classification output further. The same reasoning applies to other large countries such as Australia, Canada, or China.

A2 External validity analyses

A2.1 Overview

In this section, we investigate the surface groups measure's external validity as a proxy for economic activity. The purpose of the measure is to approximate economic activity over a long time series and at small regional levels. To examine whether the surface groups fulfill this purpose, we require external data on economic activity at small regional levels. With such external data, we can empirically analyze the quality of a surface groups-based prediction of economic activity.

In our main validation analyses, we draw on two external sources of validation data to analyze the surface groups-based prediction of economic activity. First, from administrative statistics, we extract a regionally disaggregated direct measure of GDP, the most commonly used indicator of economic activity in the literature evaluating previous satellite-based proxies for economic activity (e.g., 5, 34). The administrative GDP measure is available at the county (*Kreis*) level[15] from 2000. Second, we use the socioeconomic dataset RWI-GEO-GRID (35) that provides household income as a further indicator of economic activity with a very high level of regional detail. This indicator is available at the level of grid cells sized one square kilometer (and thus independent of administrative borders), but annually only from 2009.

To evaluate the surface groups-based prediction of economic activity, we perform Ordinary Least Squares (OLS) regressions of the two indicators of economic activity (GDP and household income) on the surface groups. These regressions allow us to determine how much of the variation in economic activity the surface groups explain. Furthermore, we analyze the distribution of the regression residuals to assess potential biases in the prediction of economic activity. Throughout this evaluation, we compare the surface groups-based prediction of economic activity to the prediction based on night light intensity data from the U.S. Air Force Defense Meteorological Satellite Program Operational Linescan System (DMSP OLS). This commonly used night lights-based prediction thus serves as a benchmark for assessing the quality of our daytime-based prediction using surface groups.

In additional validation analyses, we examine further predictive properties of surface groups. We investigate within-region predictive power, evaluate surface groups against newer night light intensity data with higher spatial resolution from the Visible Infrared Imaging Radiometer Suite (VIIRS), and compare the predictive value of surface groups to a prior approach in Africa (26).

By using external validation data that are available for limited time series, the analyses in this section provide insight into the quality of the surface groups as a measure for applications in economic research. After describing the external data we use for these analyses in more detail, we present the analysis of surface groups as a novel six-dimensional proxy for economic activity. Finally, we show how the six surface groups can be combined into a single-variable proxy.

A2.2 Validation data

To obtain economic indicators at detailed regional levels, we draw on two data sources. First, we use administrative regional data. We access these data via the "Regionaldatenbank Deutschland",[16] a database belonging to the German Federal Statistical Office's (GFSO) data portal, GENESIS.[17] This database comprises a variety of regional statistics from the GFSO and the statistical offices of the 16 federal states (*Bundesländer*), with varying time series and levels of regional disaggregation. GDP information in the administrative statistics is available at the county level, the next lower administrative regional unit after the federal states, from 2000 through 2018.[18] Following prior work (34), we use real (i.e., deflated)[19] GDP measures in euros as a validation measure for our analyses.

[15] As of 2020, Germany comprised 401 counties.

[16] https://www.regionalstatistik.de/genesis/online/ (accessed July 19, 2021).

[17] The acronym "GENESIS" stands for "Gemeinsames Neues Statistisches Informations-System". See https://www.statistikportal.de/de/datenbanken (accessed July 19, 2021).

[18] We use data table 82111-01-05-4 "Bruttoinlandsprodukt/Bruttowertschöpfung nach Wirtschaftsbereichen – Jahressumme – regionale Tiefe: Kreise und krfr. Städte" available at https://www.regionalstatistik.de/genesis/online?operation=previous&levelindex=1&step=1&titel=Tabellenaufbau&levelid=1626691580813&acceptscookies=false#abreadcrumb (accessed June 29, 2021).

[19] We deflate to 2000 prices according to the consumer price index provided by the GFSO. See https://www-genesis.destatis.de/genesis/online?sequenz=tabelleErgebnis&selectionname=

We denote real GDP as GDP.

Second, we use RWI-GEO-GRID (35), a grid-level dataset containing socioeconomic indicators collected from a variety of public and private sources, but annually available only from 2009 through 2016 (for a more detailed description of this dataset, see 85). From this dataset, we extract a measure of household income that allows us to analyze economic activity at a regional level even more detailed than the administrative county level. This measure is available at the level of grid cells sized one square kilometer, an extremely high level of regional detail, and indicates the total purchasing power of all households living in a grid cell (85). The grid cells in this dataset follow the system of the European Reference Grid distributed by the European Soil Data Centre (ESDAC)[20] (85). To evaluate the quality of the surface groups-based prediction at this very detailed regional level, we use real household income measured in euros at the grid level as a further indicator of economic activity. For data protection, the dataset contains missing or zero values for grid cells with a population below five inhabitants or households (85). However, we expect economic activity and thus household income in these grid cells to be negligibly small, so that our analysis excludes grid cells essentially without economic activity. Altogether, Germany comprises 381,425 grid cells, between 146,382 and 148,509 of which (depending on the year) contain positive values of household income within our observation period. We denote real household income as HHI.

To compare the quality of the prediction that uses surface groups to the prediction that uses night light intensity in our main validation analyses (section A2.3), we use DMSP OLS night lights data, available from 1992 through 2013.[21] Simply put, these data capture the intensity of light sources on earth at night (86). This night light intensity constitutes a valuable proxy for economic activity at the national level and at larger subnational levels such as federal states or metropolitan areas (4, 5, 87). The technological developments of the 21st century have improved both the accessibility of night lights data and the computational capabilities for processing these data (13). Consequently, night lights data have become an attractive data source for economists in the last decade. Similar to prior work (34), we use the pre-processed version of the DMSP OLS data (i.e., the version corrected for, e.g., clouds or unusual lighting such as forest fires). This version contains one observation per pixel and year, indicating the intensity of light sources on earth at night.[22] The intensity variable is a digital number ranging between 0 and 63. To achieve regional correspondence with the administrative GDP data and RWI-GEO-GRID, we calculate the average DMSP OLS night light intensity at the county and at the grid level (denoted as $NL_{DMSPOLS}$).

Furthermore, we use two other data sources in Section A2.4. First, we use VIIRS night lights data as an alternative to the DMSP OLS benchmark.[23] While VIIRS data offer a higher spatial resolution than DMSP OLS data (500 meters vs. one kilometer at the

61111-0001&startjahr=1991#abreadcrumb (accessed November 4, 2021).

[20] Available from https://esdac.jrc.ec.europa.eu/content/european-reference-grids (accessed August 13, 2019).

[21] We use the Version 4 DMSP-OLS Nighttime Lights Time Series distributed by the National Oceanic and Atmospheric Administration's (NOAA) National Geophysical Data Center, available at https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html#AVSLCFC (accessed October 25, 2021).

[22] For a few observation years, two satellites collected night light intensity. Consequently, the night lights data contain two observations per pixel for these years. Following prior work (34), we use the average of those observations.

[23] We use the annual VIIRS night lights composites version 2 (88), available from the Colorado School of Mines at https://eogdata.mines.edu/nighttime_light/annual/v20/ (accessed October 27, 2021).

equator), their available time series is substantially shorter (2012–2020 vs. 1992–2013). As the 2012 and 2013 VIIRS composites differ from later years by not being built from stray-light corrected data (88), we do not use these two years in our analyses to have a consistent benchmark. Like DMSP OLS data, VIIRS data contain one observation per pixel and year. We denote the regional average of the VIIRS night light intensity variable, which indicates radiance measured in nano Watts per square centimeter per steradian, as NL_{VIIRS}.

Second, we use an index for village-level asset wealth from prior work in Africa (26). The authors (26) use African Demographic and Health Survey (DHS) data to construct this index, including measures for quality of living (e.g., if households have running water). They then train a neural network to directly predict the index from a combination of Landsat and DMSP OLS night light intensity data. We use both their original DHS-based asset wealth index as an outcome to validate the surface groups against (denoted as AWI) and their predicted asset wealth index as benchmark (denoted as $\widehat{AWI}$).[24]

To assess the value of the surface groups we derive from Landsat data as a proxy for economic activity, we aggregate the pixel-level surface groups information to the different regional units of the validation data. We do so by counting the number of pixels in each surface group per regional unit and year, thus generating, at the respective regional level, six variables indicating the number of pixels per surface group: *builtup*, *grass*, *crops*, $forest$, *noveg*, and *water*.[25] Moreover, to improve the evaluation by accounting for potential measurement error in the number of pixels per surface group, we calculate a region's percentage of pixels with values missing because of, for example, cloud cover as an indicator of potential measurement error (denoted as $\%cloud$).[26]

In sum, this set of validation data allows us to perform a precise validation analysis of surface groups as a novel six-dimensional proxy for economic activity. We argue that if the quality of the surface groups-based prediction is high in the years that the validation data cover, this quality is high for earlier periods as well because we consistently measure the surface groups over time (i.e., for the entire period from 1984–2020). Put differently, we have no reason to believe that our results on the validity of surface groups as a proxy for economic activity would change if the validation data were already available from 1984. Therefore, we assume that the conclusions we draw from the validation analysis also hold for earlier periods for which validation data are not available (1984–1999 for GDP and 1984–2008 for household income) and, consequently, that the surface groups proxy economic activity equally well from 1984 through 2020.

[24]Both the original and the predicted asset wealth index are available as a supplement to 26.

[25]For better efficiency, we perform the aggregation tasks of surface groups (and that of any other regionally aggregated variables in our analyses such as night light intensity) to the different regional units using Esri's ArcPy package. However, these tasks can be achieved using freeware such as PyQGIS with similar results. The polygon shapefiles indicating the regional borders of the validation data in our analyses are available from the German Federal Agency for Cartography and Geodesy at https://daten.gdz.bkg.bund.de/produkte/vg/vg250_ebenen_0101/ (accessed November 3, 2021; administrative regional borders in Germany), from ESDAC at https://esdac.jrc.ec.europa.eu/content/european-reference-grids (accessed August 13, 2019; grid-cell borders for EU25 countries), and from the Database of Global Administrative Areas (GADM) at https://gadm.org/download_country.html (accessed November 22, 2021; administrative borders of African and other countries).

[26]Other reasons for missing values could be implausible spectral values or inexistence of imagery (see section A1.2). However, cloud cover is the most likely reason.

A2.3 Validation of surface groups as a proxy for economic activity

To assess the external validity of surface groups as a proxy for economic activity and to compare them to night light intensity—which has become a widely accepted proxy in economic research—we perform OLS regressions of the following form:

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t} + \beta_2 C_{i,t} + \nu_{i,t} \tag{A4}$$

with i denoting the regional unit of observation (i.e., counties for the GDP analysis and grid cells for the household income analysis), t denoting the year of observation, and Y denoting the dependent variable $ln(GDP)$ or $ln(HHI)$. X denotes the independent variables, that is, the vector of surface groups (including $ln(builtup+1)$, $ln(grass+1)$, $ln(crops+1)$, $ln(forest+1)$, $ln(noveg+1)$, and $ln(water+1)$) or $ln(NL_{DMSPOLS}+1)$. C represents a vector of control variables and ν constitutes the error term.

To compare the surface groups-based prediction to the night lights-based prediction, we restrict the observation periods to those years for which all variables entering the equation are available. The years of observation are thus 2000 through 2013 for the GDP analysis and 2009 through 2013 for the household income analysis.[27]

To assess whether the combination of surface groups is a valid proxy for economic activity, we follow prior work (34) by using the natural logarithms of the dependent variables and the independent variables. We add the value one to the variables in X before taking their natural logarithms, because they contain values of zero. As the variables in X do not represent percentage points, they do not add up to 100. Thus including all six surface groups in the regressions does not lead to multicollinearity. In an assessment of night light intensity as a country-level proxy for GDP (34), the authors argue that night light intensity might be more sensitive to a growth in GDP than to a decline in it, because technology and other factors constantly change over time. The same logic applies to surface groups. For example, while a growth in GDP and the construction of new buildings might occur simultaneously, a decline in GDP might involve a stagnation of construction activities or an abandonment of buildings rather than a remotely sensible reduction in built-up surfaces. Therefore, surface groups might also be more sensitive to a growth in GDP than to a decline in it.

The vector C comprises two control variables that cancel out any bias due to potential measurement error in the dependent or independent variables. First, year fixed effects (FE) account for potential quality differences between years in Landsat or DMSP OLS data. Such differences might occur due to, for example, the technological performance of satellites or weather conditions. Second, federal state FE control for potential differences in administrative data collected by the statistical offices of the federal states.[28]

County-level analysis of GDP. The results of the county-level analysis with real GDP as the dependent variable in Table A8 show that surface groups explain more of the variation in GDP than DMSP OLS night light intensity. In the specifications without control variables, surface groups explain 43.9% of the variation in GDP (column 1),

[27] The household income data are also available for 2005, but we exclude this year to consistently examine patterns in the temporal distribution of the regression residuals by maintaining a data structure of consecutive years.

[28] As we compare the surface groups-based prediction to the night lights-based prediction, we do not include the percentage of cloud cover (see section A2.2) as a control variable for potential measurement error in the number of pixels per surface group. The results do not change when we include this control variable in the prediction using surface groups (see tables A9 and A11).

Table A8. OLS prediction of GDP using surface groups and using DMSP OLS night light intensity (county level, 2000–2013)

	Surface groups		DMSP OLS night light intensity	
Dep. var.: $ln(GDP)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.625***	1.368***		
	(0.029)	(0.035)		
$ln(grass + 1)$	-0.050***	-0.132***		
	(0.015)	(0.013)		
$ln(crops + 1)$	-0.354***	-0.269***		
	(0.012)	(0.012)		
$ln(forest + 1)$	-0.095***	-0.162***		
	(0.011)	(0.011)		
$ln(noveg + 1)$	-0.408***	-0.246***		
	(0.016)	(0.015)		
$ln(water + 1)$	-0.153***	0.002		
	(0.017)	(0.015)		
$ln(NL_{DMSPOLS} + 1)$			0.532***	0.432***
			(0.015)	(0.017)
Year FE	No	Yes***	No	Yes***
Federal state FE	No	Yes***	No	Yes***
N	5,402	5,402	5,402	5,402
Adj. R^2	0.439	0.623	0.230	0.471

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9. OLS prediction of GDP using surface groups (county level, 2000–2013)

Dep. var.: $ln(GDP)$	(1)	(2)
$ln(builtup + 1)$	1.642***	1.360***
	(0.029)	(0.035)
$ln(grass + 1)$	-0.030**	-0.116***
	(0.015)	(0.014)
$ln(crops + 1)$	-0.357***	-0.282***
	(0.012)	(0.012)
$ln(forest + 1)$	-0.104***	-0.172***
	(0.012)	(0.012)
$ln(noveg + 1)$	-0.407***	-0.241***
	(0.016)	(0.015)
$ln(water + 1)$	-0.151***	0.002
	(0.017)	(0.015)
Year FE	No	Yes***
Federal state FE	No	Yes***
$\%cloud$	-2.327**	-4.247***
	(0.960)	(0.923)
N	5,402	5,402
Adj. R^2	0.439	0.624

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

whereas night light intensity explains only 23.0% of this variation (column 3). Including the control variables does not affect this pattern, with surface groups explaining 62.3% (column 2) and night light intensity explaining 47.1% of the variation in GDP (column 4). As the specifications with control variables explain a larger percentage of the variation in GDP for both surface groups and night light intensity, controlling for potential measurement error improves the prediction but neither affects the predictive properties of surface groups nor those of DMSP OLS night light intensity. At the disaggregated regional level of counties, the combination of surface groups and control variables thus explains a significant percentage of the variation in GDP.

Figs. 2 *A* and *B* show that the statistical distribution of the residuals from the OLS regressions with control variables (columns 2 and 4 of table A8) looks smoother and narrower for surface groups than for DMSP OLS night light intensity. This finding is in line with surface groups explaining more of the variation in GDP than night light intensity, as indicated by the adjusted R^2 of the regressions. Moreover, for both surface groups and night light intensity, the residuals are normally distributed, although the distribution has more pronounced local maxima in the night lights specification. Surface groups thus proxy GDP more precisely than DMSP OLS night light intensity.

Furthermore, using surface groups to compare GDP over time and between regions requires that the prediction error be neither temporally nor spatially biased. Temporal bias would occur if the prediction error is constant for a given region throughout all observation years, and spatial bias would occur if the prediction error is equal for clusters of regions. To assess the existence of such biases, Fig. A3 illustrates the temporal and spatial distribution of the residuals from the regressions in column 2 of Table A8. For reference, Fig. A2 provides a map indicating the names of the federal states and the locations of their capitals. In four-year intervals evenly spread over our observation period, Fig. A3 shows the estimated residuals for all counties in the respective year, that is, the degree to which GDP is overestimated (blue counties) or underestimated (red counties). For comparison, Fig. A4 proceeds similarly for DMSP OLS night light intensity, illustrating the residuals from the regression in column 4 of Table A8.

Figs. A3 and A4 suggest that the surface groups-based prediction yields a considerably smaller temporal bias than the night lights-based prediction. If a temporal bias in prediction error existed, the color of a given region would stay the same over the entire observation period. For surface groups, such a pattern exists for 179 counties (44.9%), and, for the remaining regions, the color varies over time in Fig. A3. For DMSP OLS night light intensity, this pattern appears for 339 counties (85.0%), leading to the four maps in Fig. A4 hardly differing in color. Therefore, although we cannot definitely rule out the existence of a temporal bias for some regions when proxying GDP with surface groups, this temporal bias is far less severe than that of proxying GDP with night light intensity.

The distribution of the residuals across regions in Figs. A3 and A4 suggests a somewhat larger spatial bias in prediction error for surface groups than for DMSP OLS night light intensity. If such as bias existed, clusters of similarly colored regions would appear. For surface groups, 992 observations (18.4%) have the same color as all their geographically neighboring observations, whereas for night light intensity, this pattern shows for only 565 observations (10.5%). However, for both surface groups and night light intensity, the clusters appear randomly distributed across the country rather than concentrated in specific parts (e.g., clusters not only in rural areas, clusters not only in the north). Therefore, the spatial distribution of the prediction error appears random but yields a larger

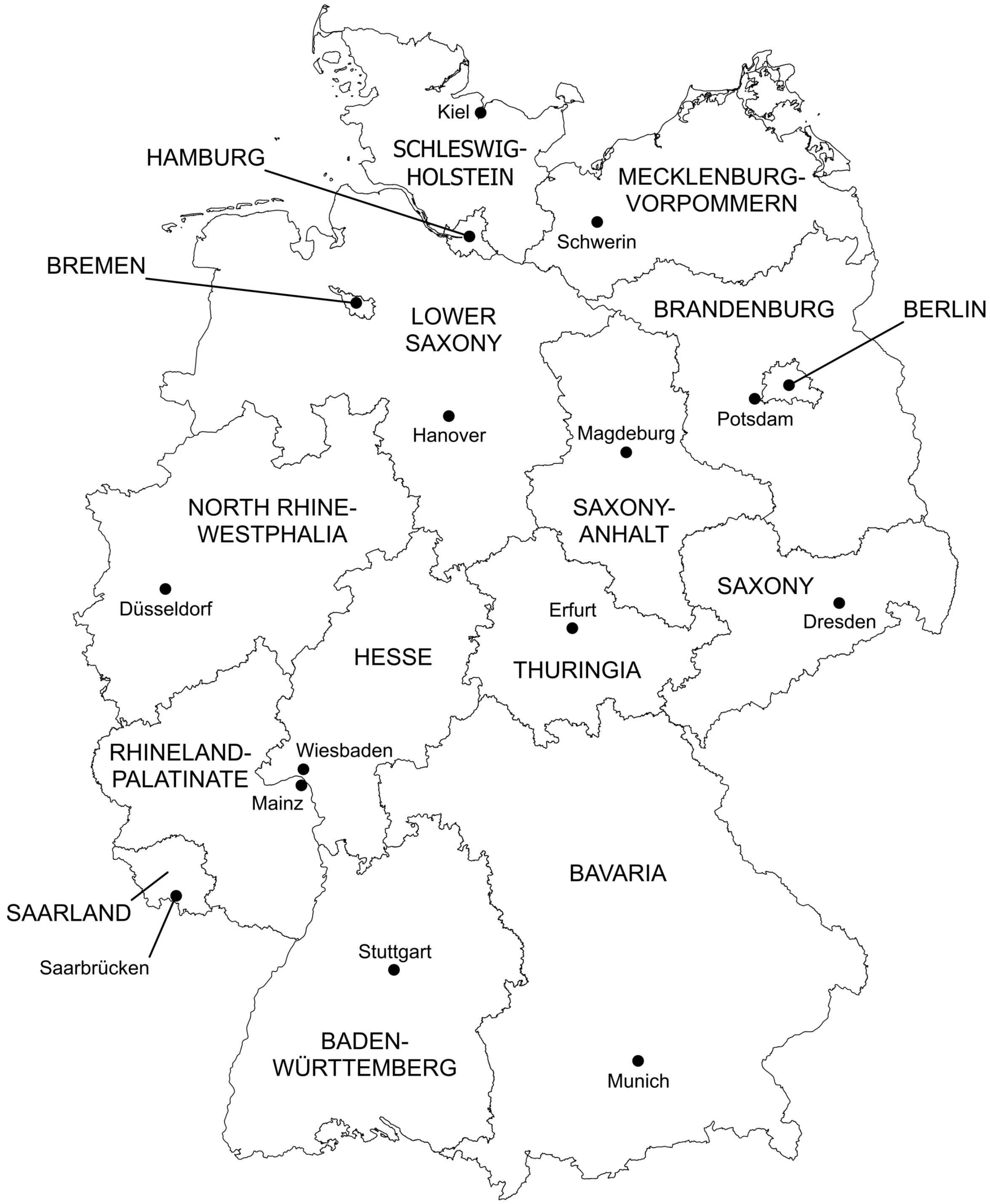

Fig. A2. Reference map of German federal states and their capitals.

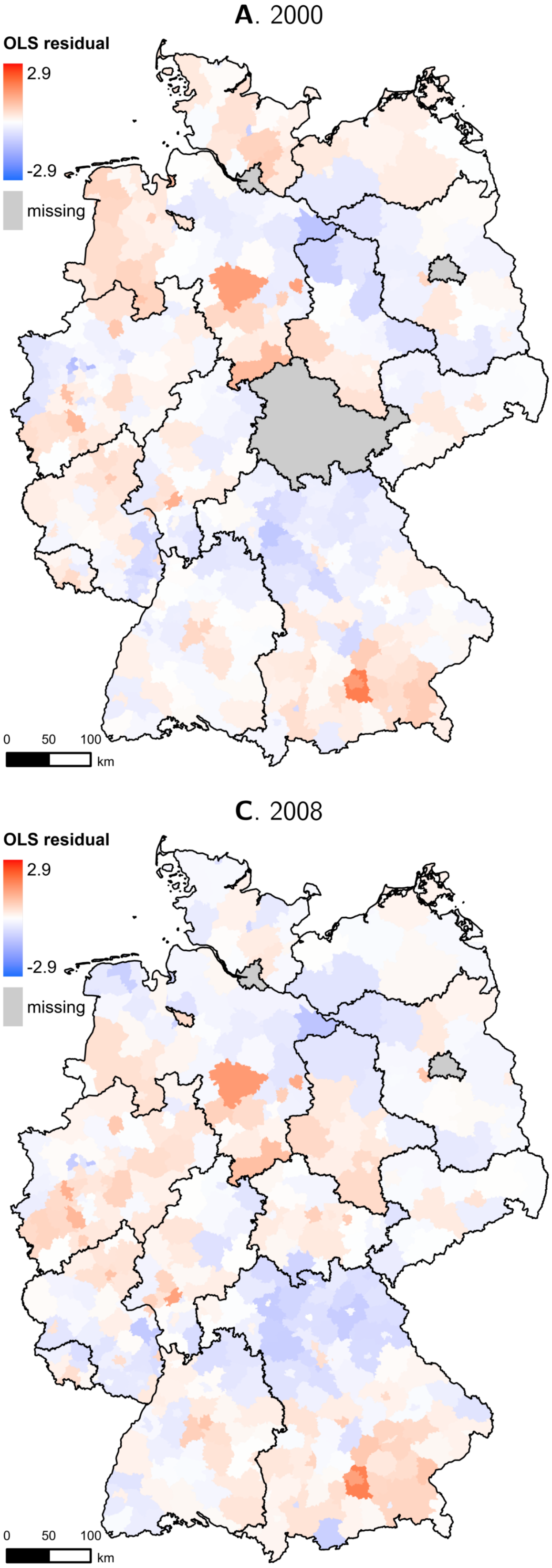

Fig. A3. Spatial and temporal distribution of GDP residuals for surface groups. Maps illustrate residuals from the regression in column 2 of Table A8.

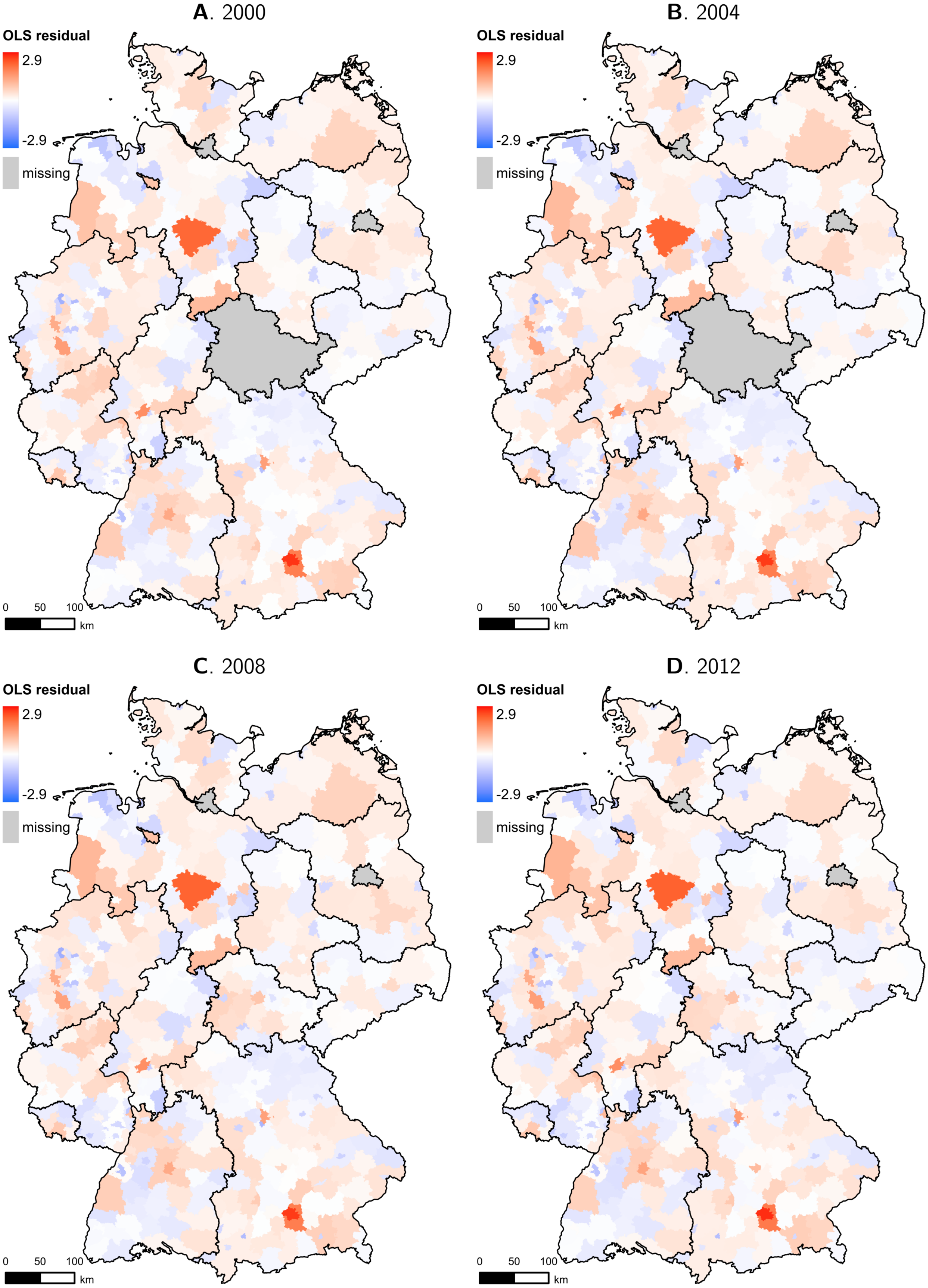

Fig. A4. Spatial and temporal distribution of GDP residuals for DMSP OLS night light intensity. Maps illustrate residuals from the regression in column 4 of Table A8.

bias for surface groups.

Combining the indicators of temporal and spatial bias shows that the smaller temporal bias of the surface groups-based prediction outweighs the prediction's larger spatial bias as compared to the night lights-based prediction. For surface groups, only 11 counties (2.8%) have the same color as all their neighboring observations and, simultaneously, the same color throughout all observation years. For DMSP OLS night light intensity, this pattern appears for 26 counties (6.5%). This finding reflects in the small clusters of similarly colored counties not showing up in consecutive years in Fig. A3.

In addition, to show that the value of surface groups as a proxy for economic activity increases with the degree of regional disaggregation, we estimate our OLS model separately by county-size groups. Fig. A5 plots average county size within a group against the adjusted R^2 obtained from the separate regressions. As county-size groups, we use quintiles of the county-size distribution (fig. A5 *A*) and federal states (fig. A5 *B*). In addition to the original data points obtained from the regressions, Fig. A5 also plots the linear fitted values to visualize the trend in the data. For both county-size groups, the plots show a declining trend, that is, the percentage of the variation in GDP explained by surface groups declines with an increase in county size. Put differently, the smaller the county size the better the proxy. This finding emphasizes the potential of surface groups as a valuable measure for analyses at detailed regional levels.

In essence, the county-level analysis of the surface groups-based prediction of GDP yields the finding that surface groups are a highly suitable proxy for GDP. They explain a significant percentage of the variation in GDP. Moreover, in comparison to the DMSP OLS night lights-based prediction, the surface groups-based prediction shows a smaller bias in the regression residuals. Therefore, surface groups provide a useful alternative for proxying GDP at disaggregated regional levels such as German counties.

Grid-level analysis of household income. In the grid-level analysis of surface groups as a proxy for household income, we find the same patterns as in the county-level analysis of surface groups as a proxy for GDP. Table A10 presents the estimation results for this grid-level analysis. At this very detailed regional level, the surface groups-based prediction explains a much larger percentage of the variation in household income than the DMSP OLS night lights-based predictions (63.6% vs. 27.2% in the specifications without control variables in columns 1 and 3, and 67.5% vs. 30.7% in the specifications with control variables in columns 2 and 4). In comparison to the GDP analysis, the control variables (year FE and federal state FE) improve the prediction only slightly in the household income analysis, probably because the number of observation years is smaller and because the dependent variable is not collected within administrative borders.

Figs. 2 *C* and *D* confirm the findings of the regressions. The statistical distribution of the prediction error for household income is much narrower (although slightly left-skewed) for surface groups than for night light intensity. The distribution of the prediction error for night light intensity is slightly right-skewed and, instead of a peak at the value zero, the distribution exhibits a plateau around this value. Therefore, surface groups proxy household income at the grid level much more precisely than DMSP OLS night light intensity.

Furthermore, the assessment of the temporal and spatial distribution of the prediction error in the household income analysis yields results similar to those in the GDP analysis. Figs. A6 and A7 show the spatial and temporal distribution of the prediction error in household income for surface groups and DMSP OLS night light intensity, respectively.

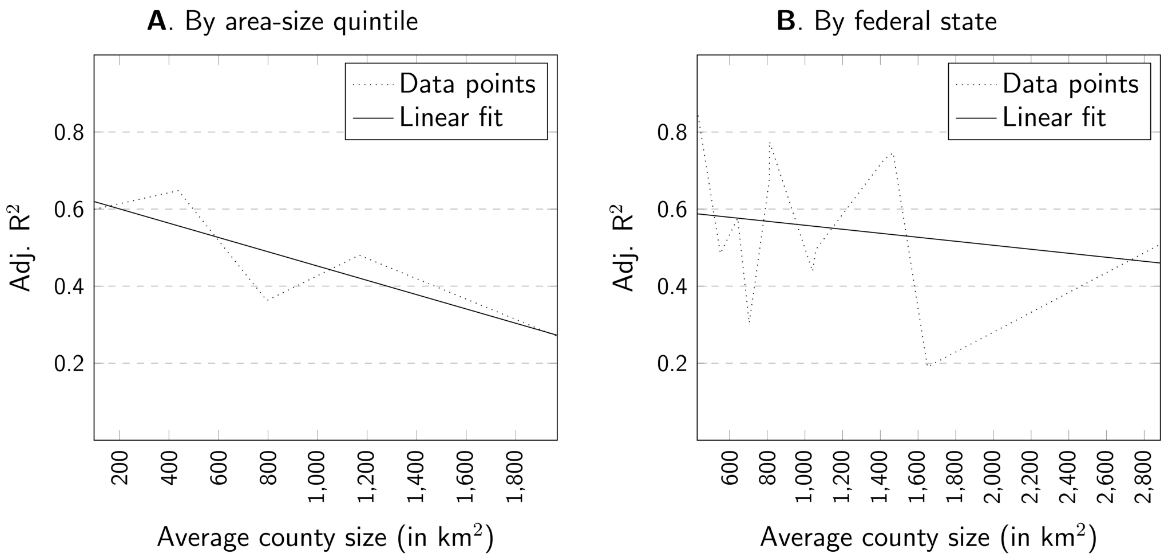

Fig. A5. Adj. R^2 by county-size group. Values stem from separate regressions of surface groups on GDP corresponding to the specification in column 2 of Table A8

Table A10. OLS prediction of household income using surface groups and using DMSP OLS night light intensity (grid level, 2009–2013)

	Surface groups		DMSP OLS night light intensity	
Dep. var.: $ln(HHI)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.449***	1.412***		
	(0.002)	(0.002)		
$ln(grass + 1)$	-0.090***	-0.126***		
	(0.002)	(0.002)		
$ln(crops + 1)$	-0.422***	-0.371***		
	(0.002)	(0.002)		
$ln(forest + 1)$	-0.053***	-0.066***		
	(0.001)	(0.001)		
$ln(noveg + 1)$	-0.200***	-0.173***		
	(0.001)	(0.001)		
$ln(water + 1)$	-0.268***	-0.211***		
	(0.001)	(0.001)		
$ln(NL_{DMSPOLS} + 1)$			0.936***	0.953***
			(0.002)	(0.002)
Year FE	No	Yes***	No	Yes***
Federal state FE	No	Yes***	No	Yes***
N	737,626	737,626	737,626	737,626
Adj. R^2	0.636	0.675	0.272	0.307

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11. OLS prediction of household income using surface groups (grid level, 2009–2013)

Dep. var.: $ln(HHI)$	(1)	(2)
$ln(builtup + 1)$	1.462***	1.426***
	(0.002)	(0.002)
$ln(grass + 1)$	-0.0832***	-0.118***
	(0.002)	(0.002)
$ln(crops + 1)$	-0.413***	-0.360***
	(0.001)	(0.001)
$ln(forest + 1)$	-0.044***	-0.057***
	(0.001)	(0.001)
$ln(noveg + 1)$	-0.200***	-0.173***
	(0.001)	(0.001)
$ln(water + 1)$	-0.270***	-0.214***
	(0.001)	(0.001)
Year FE	No	Yes***
Federal state FE	No	Yes***
$\%cloud$	0.804***	0.904***
	(0.052)	(0.053)
N	737,626	737,626
Adj. R^2	0.637	0.675

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

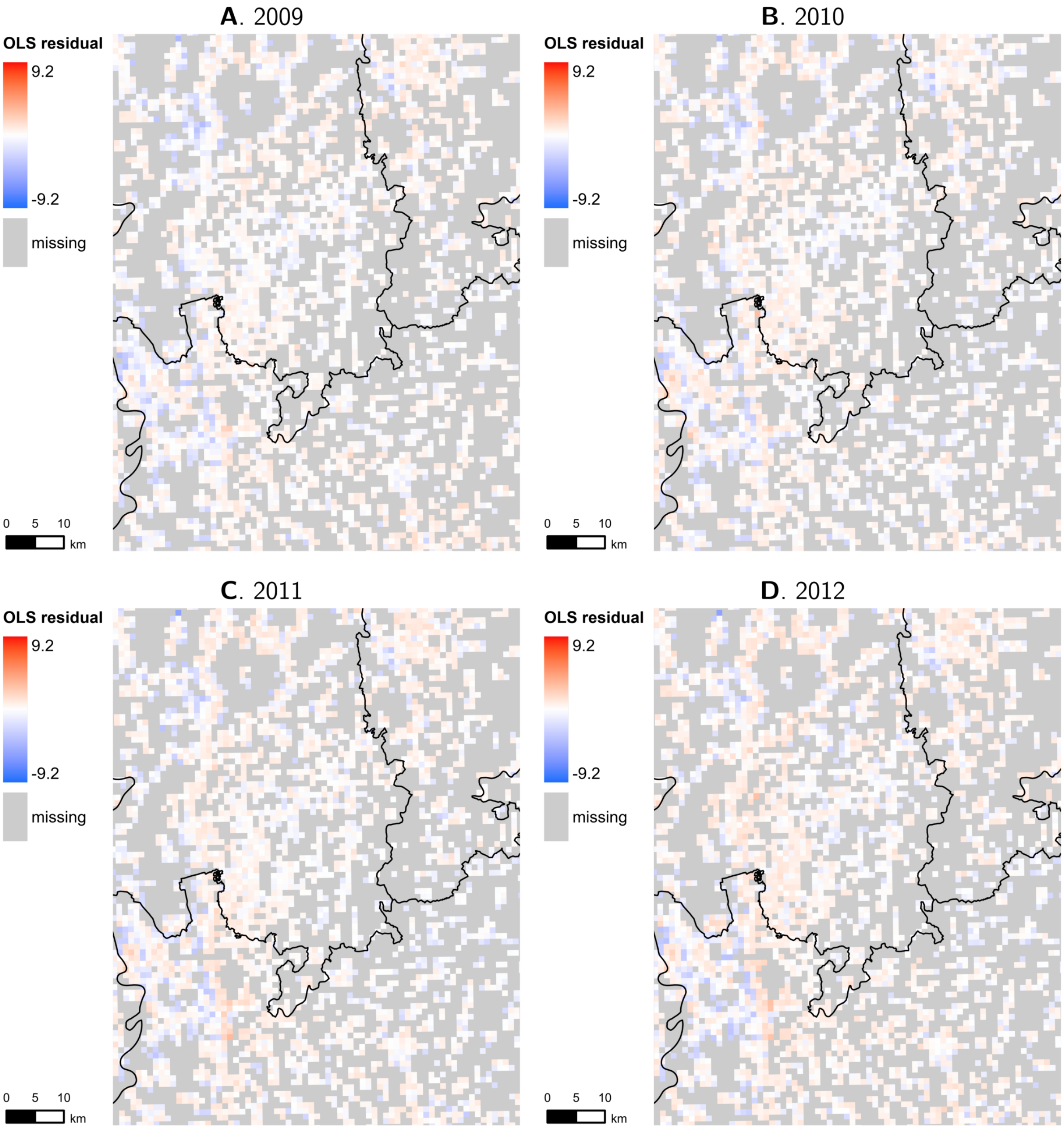

Fig. A6. Spatial and temporal distribution of household income residuals for surface groups. Maps illustrate residuals from the regression in column 2 of Table A10. Maps show an area at the borders of the four federal states *Rhineland-Palatinate*, *Hesse*, *Baden-Württemberg*, and *Bavaria*.

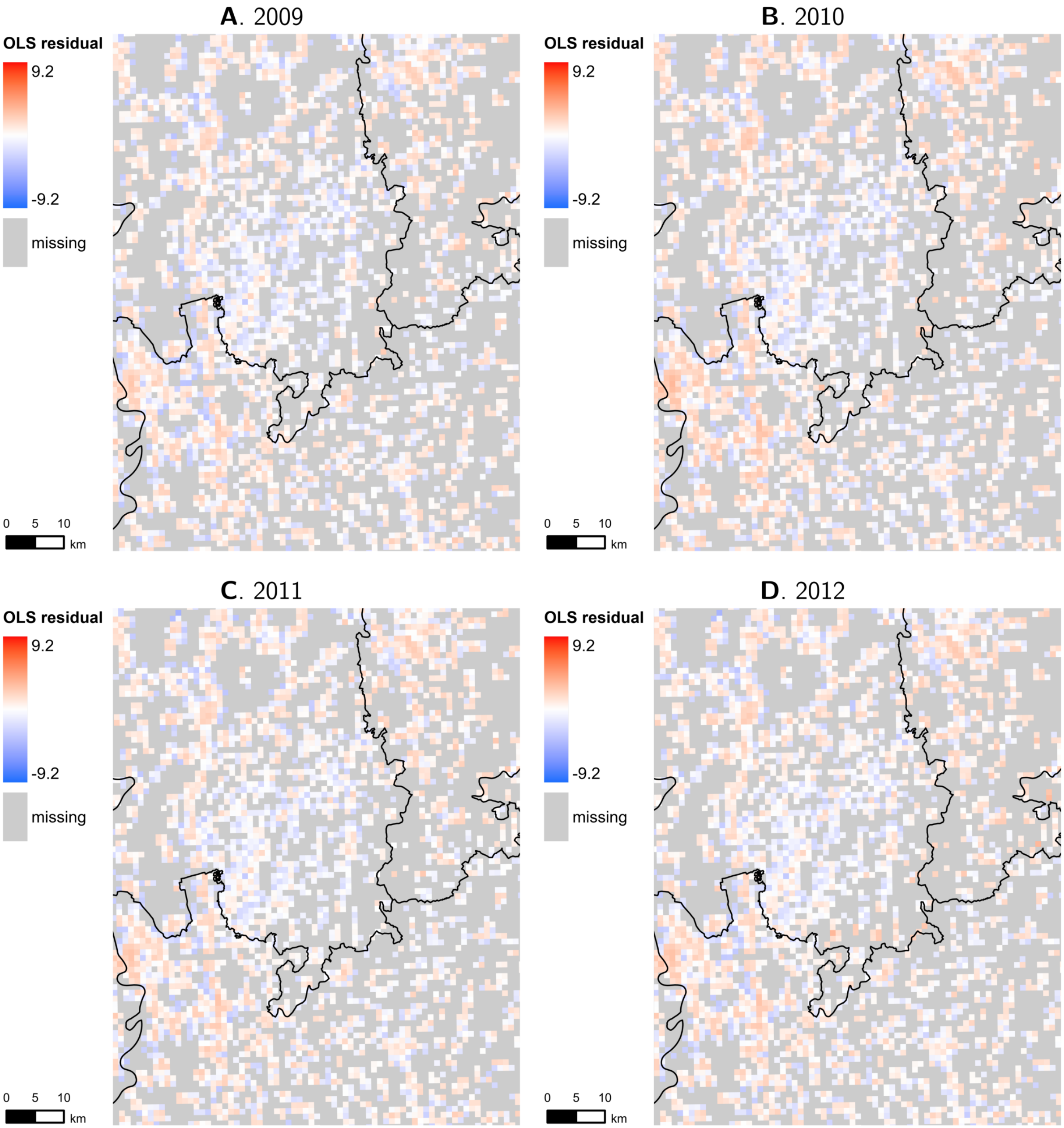

Fig. A7. Spatial and temporal distribution of household income residuals for DMSP OLS night light intensity. Maps illustrate residuals from the regression in column 4 of Table A10. Maps show an area at the borders of the four federal states *Rhineland-Palatinate*, *Hesse*, *Baden-Württemberg*, and *Bavaria*.

For a better illustration of the very small grid cells, the map shows an area at the borders of four federal states, with the metropolitan region of *Ludwigshafen-am-Rhein/Mannheim* in the south-west and the rural *Odenwald* region in the east. The gray cells are those with missing values (i.e., uninhabited or only sparsely inhabited areas).

Again, the smaller temporal bias in the surface groups-based prediction in comparison to the night lights-based prediction outweighs the larger spatial bias. For surface groups 90,054 grid cells (59.5%) have the same color throughout all observation years, whereas this number amounts to 131,704 grid cells (87.0%) for DMSP OLS night light intensity. Moreover, the spatial bias of the surface groups-based prediction is only slightly larger than the spatial bias of the night lights-based prediction, with 167,095 observations (22.7%) for surface groups and 126,703 observations (17.2%) for night light intensity having the same color as all their geographical neighbors. Combining the two types of biases shows that for surface groups, 8,166 grid cells (5.4%) have the same color as their neighbors and, simultaneously, the same color throughout all observation years. For DMSP OLS night light intensity, this pattern applies to 15,058 grid cells (9.9%). Therefore, the smaller temporal bias of surface groups again outweighs their slightly larger spatial bias.

Summary. To summarize our main analyses of the surface groups' external validity, we show that at the county level (GDP) and at the grid level (household income) surface groups can serve as a valid proxy for economic activity. At both levels, the surface groups predict a significant percentage of the variation in economic activity, and this prediction is more precise (i.e., less biased) for surface groups than for DMSP OLS night light intensity. Furthermore, the comparative advantage of surface groups as a proxy for economic activity becomes more pronounced in the grid-level analysis than in the county-level analysis. This finding, in combination with the GDP analysis by county-size group, suggests that surface groups are particularly useful for applications that investigate very small regional units. Although we derive these findings from external validation data with limited time series, we argue that, due to the high and temporally stable internal validity of the surface groups measure (see section A1.5), surface groups can also function as a valid proxy for economic activity for earlier years.

To ensure the surface groups' validity across all years in economic or other applications, we recommend (a) including the number of cloud-covered pixels as a control variable and (b) checking the data for outlier observations and remove those from empirical analyses for particular years and regions. Such outliers can occur in few regions in years with scarce Landsat imagery (particularly in the 1980s). For these years, our greenest pixel composite features higher percentages of cloud-covered pixels, pixels showing cloud shadow, or otherwise invalid pixels. As the filters we apply in constructing the greenest pixel composite cannot detect some of these pixels, our algorithm potentially produces an erroneous classification for these pixels.[29] To obtain more valid results, we apply an outlier correction in the application of surface groups to the comparison of GDP developments across German counties (fig. 3). For details on the outlier removal procedure, see Section A2.5.

While surface groups offer substantial advantages in proxying economic activity at disaggregated levels, night light intensity might still be the more appropriate proxy for cross-country studies or other larger regions. The reason is that land use characteristics might have heterogeneous meanings for a country's economy, depending on the country's

[29] Visual inspections of the classification show that most of these undetected invalid pixels are classified as *builtup*.

historical development (36). However, for small regional units and early time series, surface groups constitute a valuable and more accurate proxy for economic activity.

A2.4 Additional validation analyses

We present three additional analyses on the surface groups' external validity. First, we use VIIRS night light intensity data as a benchmark to show that surface groups offer higher precision in predicting economic activity than night light intensity data with higher spatial resolution than DMSP OLS data. Second, we analyze within-region heterogeneity in predicted GDP to demonstrate that surface groups enable the isolation of subregional changes in economic activity. Third, a comparison to prior work in Africa (26) suggests that surface groups can function as a proxy for economic conditions also in developing countries.

VIIRS night light intensity as benchmark. To analyze whether surface groups outperform night light intensity data with higher spatial resolution than DMSP OLS data in proxying economic activity, we reestimate the OLS model specified in Eq. A4 both at the county level (with GDP as outcome) and at the grid level (with household income as outcome) with VIIRS night light intensity as a benchmark. The observation periods of this analysis start in 2014 (first consistent year in the VIIRS data). They end in 2018 for the county-level analysis (last year in the GDP data) and in 2016 for the grid-level analysis (last year in the household income data).

Table A12 presents the county-level analysis that compares surface groups and VIIRS night light intensity as proxies for GDP. Our surface groups proxy achieves 142.2% of the VIIRS precision in predicting GDP, thus offering a much higher precision. While VIIRS night light intensity explains only 46.9% of the variation in GDP in the specification with control variables (column 4), surface groups explain 66.7% of this variation (column 2). Therefore, at the county level our surface groups proxy outperforms even night light intensity data with a higher spatial resolution than DMSP OLS data.

The grid-level analysis of household income in Table A13 supports the county-level finding that surface groups outperform VIIRS night light intensity in predicting regional economic activity. With 51.8%, VIIRS night light intensity explains a lower percentage of the variation in household income than surface groups with 70.0% (columns 2 and 4). While VIIRS night light intensity thus appears to perform better in proxying household income than DMSP OLS night light intensity, our surface groups proxy outperforms both sources of night light intensity data.

Within-region predictive power. To analyze the surface groups' predictive power of within-region changes in economic activity, we (a) conduct analyses at a higher level of disaggregation to contrast the usefulness of disaggregated vs. aggregated metrics and (b) reestimate our model specified in Eq. A4 with region unit (i.e., county) FE. The results show that (a) surface groups are more useful than night light intensity in disentangling which subregional units contribute to regional changes in economic activity, while (b) in more aggregated settings (i.e., settings that do not consider subregional variation) region unit and year FE alone explain almost all of the variation in economic activity with neither surface groups nor night light intensity adding any significant value.

Our analyses at a higher level of disaggregation illustrate that surface groups contribute to a better understanding of within-county changes in regional economic activity.

Table A12. OLS prediction of GDP using surface groups and using VIIRS night light intensity (county level, 2014–2018)

	Surface groups		VIIRS night light intensity	
Dep. var.: $ln(GDP)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.419***	1.249***		
	(0.040)	(0.049)		
$ln(grass + 1)$	-0.054**	-0.151***		
	(0.026)	(0.024)		
$ln(crops + 1)$	-0.312***	-0.233***		
	(0.018)	(0.019)		
$ln(forest + 1)$	-0.165***	-0.205***		
	(0.018)	(0.019)		
$ln(noveg + 1)$	0.028	0.043		
	(0.033)	(0.030)		
$ln(water + 1)$	-0.239***	-0.043*		
	(0.024)	(0.022)		
$ln(NL_{VIIRS} + 1)$			0.482***	0.382***
			(0.025)	(0.026)
Year FE	No	Yes***	No	Yes
Federal state FE	No	Yes***	No	Yes***
N	1,995	1,995	1,995	1,995
Adj. R^2	0.499	0.667	0.213	0.469

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13. OLS prediction of household income using surface groups and using VIIRS night light intensity (grid level, 2014–2016)

	Surface groups		VIIRS night light intensity	
Dep. var.: $ln(HHI)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.297***	1.275***		
	(0.002)	(0.002)		
$ln(grass + 1)$	-0.123***	-0.160***		
	(0.002)	(0.002)		
$ln(crops + 1)$	-0.356***	-0.324***		
	(0.002)	(0.002)		
$ln(forest + 1)$	-0.076***	-0.074***		
	(0.002)	(0.002)		
$ln(noveg + 1)$	-0.061***	0.058***		
	(0.002)	(0.002)		
$ln(water + 1)$	-0.222***	-0.184***		
	(0.002)	(0.001)		
$ln(NL_{VIIRS} + 1)$			1.394***	1.377***
			(0.003)	(0.003)
Year FE	No	Yes***	No	Yes***
Federal state FE	No	Yes***	No	Yes***
N	446,524	446,524	446,524	446,524
Adj. R^2	0.671	0.700	0.497	0.518

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We conduct these analyses at the level of municipalities, the smallest administrative regional unit in Germany.[30] Although GDP data do not exist at the municipality level, we can use the surface groups to derive a prediction of GDP at this level. We then compare the municipality-level change over time in this GDP prediction to the county-level change in the administrative GDP measure. If the change in GDP is similar at both geographic levels, the municipality-level prediction of GDP does not add any informative value to the county-level measure. However, if the change in municipality-level predicted GDP differs from the change in county-level GDP, the new municipality-level prediction can be informative about within-county heterogeneity in economic development, thus allowing assessments of which municipalities drive county-level economic activity (i.e., how economic activity develops heterogeneously within a county). To investigate which proxy offers more insight into within-county heterogeneity, we also compare the surface groups-based and the DMSP OLS night light intensity-based municipality-level GDP predictions.

We proceed in two steps to analyze municipality-level GDP. First, we predict GDP at the municipality level. Because both the continuous independent variables and the dependent variable are natural logarithms of their original values in the county-level prediction in Table A8, the estimation coefficients are not directly transferable to the municipality level. Therefore, we standardize these county-level variables to have a mean of 0 and a standard deviation of 1, then estimate the OLS model specified in Eq. A4 using the standardized variables. As the standardization does not affect the variables' distributional properties except for the mean and the standard deviation, the OLS result in Table A14 has the same properties (adjusted R^2, F-value, coefficients' t-values) as the original unstandardized result. Assuming that the distributional properties of the variables in the model are identical at the county level and at the municipality level, we can use the coefficients from the county-level estimation with standardized variables to predict standardized GDP at the municipality level. We produce one prediction of standardized municipality-level GDP using surface groups as predictor and one using DMSP OLS night light intensity.

Second, we construct an indicator for the difference between the municipality-level change in predicted GDP and the county-level change in administrative GDP. To obtain the municipality-level change in predicted GDP, for each municipality and for both surface groups and DMSP OLS night light intensity we calculate the difference between the prediction of standardized GDP in 2013 (the last year in the DMSP OLS night light intensity data) and that in 2000 (the first year in the administrative GDP data). To obtain the county-level change in standardized administrative GDP, we proceed similarly at the county level by calculating the difference in administrative GDP between 2013 and 2000. As final indicators, we then calculate for both surface groups and DMSP OLS night light intensity the difference between the municipality-level change in the prediction of standardized GDP and the county-level change in standardized administrative GDP. These indicators measure at the municipality-level whether and to what extent the municipality-level change in GDP over time deviates from the county-level change in GDP over time.

Fig. A8 plots the distribution of the two indicators. The figure shows that DMSP OLS night light intensity yields a lower degree of additional information at the municipality

[30] We use the territorial status of 2017, because it was the most recent one when we started work on this paper. As of January 1, 2017, Germany comprised 11,266 municipalities (i.e., on average 28.1 municipalities per county), with one municipality belonging to only one county.

Table A14. OLS prediction of GDP using surface groups and using DMSP OLS night light intensity with standardized variables (county level, 2000–2013)

	Surface groups		DMSP OLS night light intensity	
Dep. var.: standardized $ln(GDP)$	(1)	(2)	(3)	(4)
standardized $ln(builtup + 1)$	1.975***	1.642***		
	(0.035)	(0.041)		
standardized $ln(grass + 1)$	-0.109***	-0.285***		
	(0.032)	(0.028)		
standardized $ln(crops + 1)$	-0.771***	-0.585***		
	(0.026)	(0.025)		
standardized $ln(forest + 1)$	-0.224***	-0.381***		
	(0.025)	(0.027)		
standardized $ln(noveg + 1)$	-0.782***	-0.471***		
	(0.032)	(0.029)		
standardized $ln(water + 1)$	-0.296***	0.003		
	(0.033)	(0.030)		
standardized $ln(NL_{DMSPOLS} + 1)$			0.486***	0.395***
			(0.014)	(0.015)
Year FE	No	Yes***	No	Yes***
Federal state FE	No	Yes***	No	Yes***
N	5,402	5,402	5,402	5,402
Adj. R^2	0.439	0.623	0.230	0.471

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

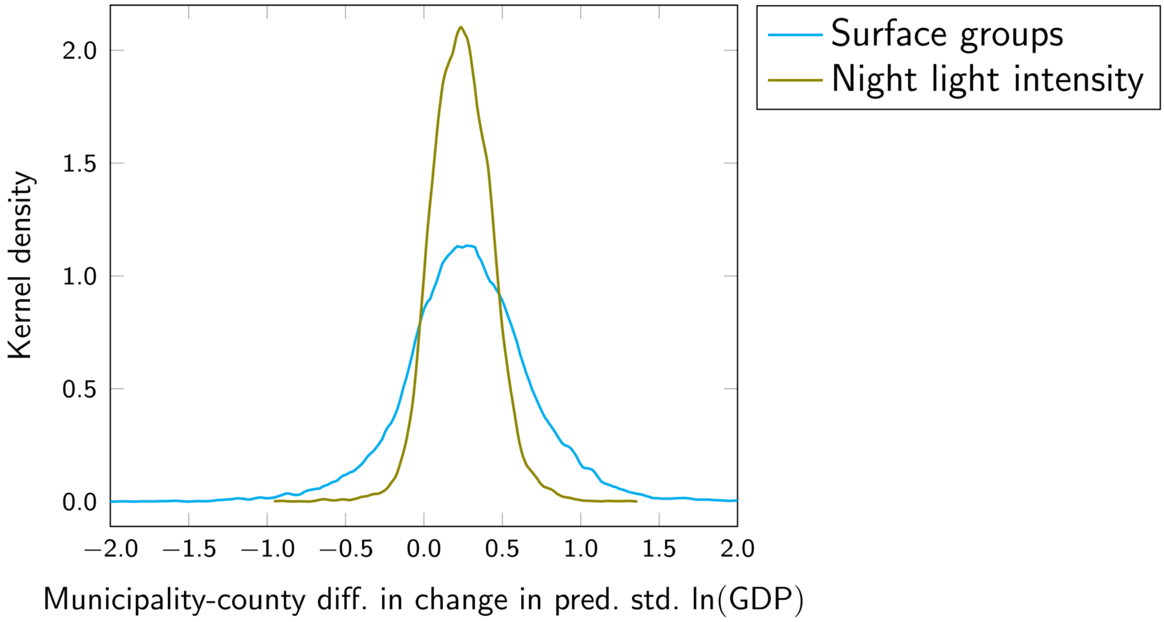

Fig. A8. Distribution of municipality-county difference in the change in predicted standardized $ln(GDP)$ between 2000 and 2013 for surface groups-based and DMSP OLS night light intensity-based prediction. Figure shows univariate kernel density estimates at 300 points using the Epanechnikov kernel function with a kernel half-width of 0.025.

level in comparison to the county level, that is, surface groups have higher within-region predictive power for geographies below the county level than DMSP OLS night light intensity. Fig. A8 reveals this relationship through the stronger concentration towards its mean in the indicator for DMSP OLS night light intensity compared to the larger variation in the indicator for surface groups. Therefore, surface groups offer more additional information at the municipality level. The change in the municipality-level prediction of standardized GDP using surface groups thus yields substantially more information on within-county heterogeneity in GDP change in comparison to DMSP OLS night light intensity.

The higher degree of additional municipality-level information obtainable from surface groups also becomes obvious in Fig. A9, which illustrates for one county (*Wunsiedel*) as an example the two indicators plotted in Fig. A8. In essence, surface groups detect much more variation in economic activity in this county's municipalities, represented by the higher intensity of colors in Figs. A9 *A6* and *B6*.

Figs. A9 *A1* and *A2* show the surface groups classification underlying the GDP prediction for 2000 and 2013, and Figs. *B1* and *B2* the corresponding raw DMSP OLS night light intensity. Figs. A9 *A3* and *A4* illustrate the surface groups-based prediction of standardized municipality-level GDP for these two years, and Figs. A9 *B3* and *B4* the DMSP OLS night light intensity-based prediction. Figs. A9 *A5* and *B5* indicate the difference between Figs. A9 *A3* and *A4* and that between Figs. A9 *B3* and *B4*, respectively, that is, the changes in the GDP predictions between 2000 and 2013. Fig. A9 *A6* then shows the municipality-county difference in the change in predicted standardized GDP using surface groups as predictor and Fig. A9 *B6* shows this difference using DMSP OLS night light intensity (i.e., the same indicators for which Fig. A8 plots the distribution).

Two properties become noticeable. First, the colors in Figs. A9 *A3* through *A6* are much more intense than in Figs. A9 *B3* through *B6*. This higher intensity is in line with Fig. A8, confirming that surface groups offer substantially more information on within-county heterogeneity by detecting variation in economic activity at the municipality level. Second, the municipalities at the south-western border of the county exhibit a substantially lower growth in GDP than the county when using surface groups for prediction (blue-colored municipalities in Fig. A9 *A6*), a pattern that is not visible when using DMSP OLS night light intensity (Fig. A9 *B6*). These municipalities differ from the other municipalities by being unincorporated areas, that is, typically uninhabited areas (e.g., forests) belonging to the county but without their own municipal governments. Therefore, that these uninhabited municipalities exhibit a substantially lower growth in GDP is a logical consequence of their characteristics. The surface groups detect these characteristics, whereas DMSP OLS night light intensity does not.

At the more aggregated county level, reestimation of our model specified in Eq. A4 with region unit FE corresponds to, for example, a cross-country analysis of DMSP OLS night light intensity as a predictor for economic activity in prior work (34). The reason that this prior work includes region-level FE (in this case countries) is to control for differences in night light intensity resulting from cultural or economic differences. Such differences can affect the country-wide use of night lights because of, for example, the relative importance of daytime activities in comparison to nighttime activities or the level of technological advancement for producing electricity. However, for within-country applications analyzing small subnational regions—the type of application that we develop our proxy for—such differences are less likely to create heterogeneity over time.

The FE estimations show that county and year FE explain almost all of the variation

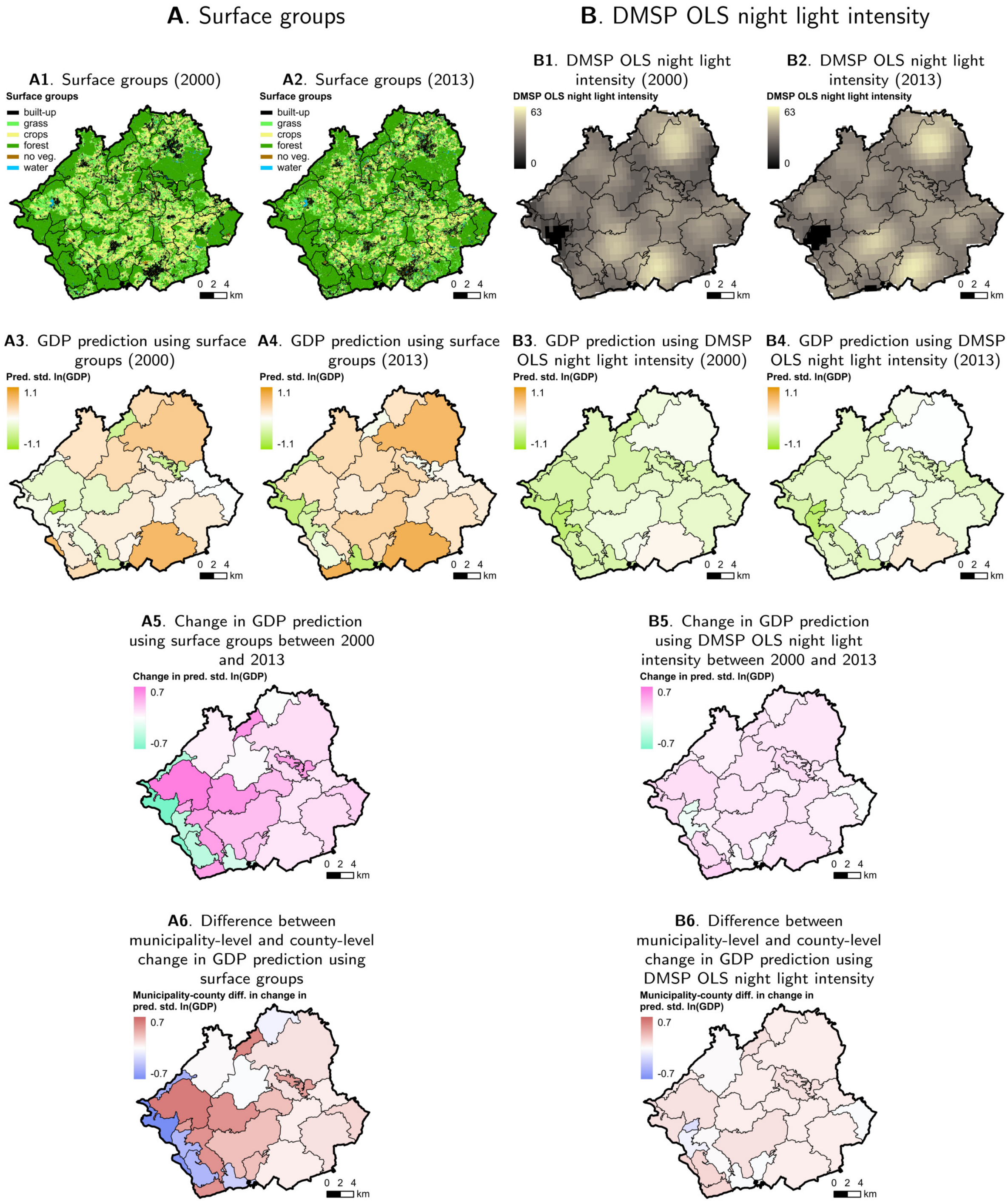

Fig. A9. Surface groups, DMSP OLS night light intensity, predictions of standardized $ln(GDP)$, changes in predicted standardized $ln(GDP)$, and municipality-county differences in the changes in predicted standardized $ln(GDP)$. Maps show the county of *Wunsiedel* (situated in south-east Germany at the border to the Czech Republic).

Table A15. FE prediction of GDP using surface groups and using DMSP OLS night light intensity (county level, 2000–2013)

	County FE covariates			County FE through within-estimator		
Dep. var.: $ln(GDP)$	No proxy (1)	Surface groups (2)	DMSP OLS night light intensity (3)	No proxy (4)	Surface groups (5)	DMSP OLS night light intensity (6)
$ln(builtup + 1)$		0.023^{***}			0.023^{***}	
		(0.006)			(0.007)	
$ln(grass + 1)$		-0.002			-0.002	
		(0.006)			(0.006)	
$ln(crops + 1)$		-0.021^{***}			-0.021^{***}	
		(0.005)			(0.005)	
$ln(forest + 1)$		0.007			0.007	
		(0.005)			(0.007)	
$ln(noveg + 1)$		-0.012^{***}			-0.012^{***}	
		(0.003)			(0.003)	
$ln(water + 1)$		0.001			0.001	
		(0.003)			(0.004)	
$ln(NL_{DMSPOLS} + 1)$			0.076^{***}			0.076^{***}
			(0.010)			(0.010)
Year FE	Yes***	Yes***	Yes***	Yes***	Yes***	Yes***
N	5,402	5,402	5,402	5,402	5,402	5,402
Adj. R^2	0.996	0.996	0.996			
Adj. within-R^2				0.295	0.301	0.307

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16. FE prediction of household income using surface groups and using DMSP OLS night light intensity (grid level, 2009–2013)

	Grid cell FE covariates			Grid cell FE through within-estimator		
Dep. var.: $ln(HHI)$	No proxy (1)	Surface groups (2)	DMSP OLS night light intensity (3)	No proxy (4)	Surface groups (5)	DMSP OLS night light intensity (6)
$ln(builtup + 1)$		0.003***			0.003***	
		(0.001)			(0.001)	
$ln(grass + 1)$		0.000			0.000	
		(0.000)			(0.000)	
$ln(crops + 1)$		0.004***			0.004***	
		(0.000)			(0.000)	
$ln(forest + 1)$		0.002***			0.002***	
		(0.000)			(0.000)	
$ln(noveg + 1)$		-0.001***			-0.001***	
		(0.000)			(0.000)	
$ln(water + 1)$		-0.001***			-0.001***	
		(0.000)			(0.000)	
$ln(NL_{DMSPOLS} + 1)$			-0.000			-0.000
			(0.001)			(0.001)
Year FE	Yes***	Yes***	Yes***	Yes***	Yes***	Yes***
N	737,626	737,626	737,626	737,626	737,626	737,626
Adj. R^2	0.997	0.997	0.997			
Adj. within-R^2				0.044	0.044	0.044

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in economic activity. That is, neither surface groups nor DMSP OLS night light intensity have enough within-county variation over time to significantly contribute to explaining within-region changes. Table A15 shows the results of three different FE models that illustrate this finding: The first model includes only county and year FE without any of the two proxies; the second includes the surface groups in addition to county and year FE; and the third includes DMSP OLS night light intensity in addition to county and year FE. The models thus correspond to the OLS regressions in Table A8, with the difference of containing county instead of federal state FE. We estimate all three models using two different estimation methods, one including the county FE as covariates to obtain an estimate of the overall variance explained by the models and one considering the county FE by subtracting the county-level mean of the dependent variable to obtain an estimate of the within-county variation explained by the model. Both estimation methods show that the inclusion of any proxy leads only to a negligibly small increase in (adjusted) R^2, with the county and year FE explaining 99.6% of the overall variation in GDP.[31]

Validation of surface groups for developing countries. To investigate whether surface groups can serve as a proxy for economic activity in developing countries, we compare our approach to a prior approach for African countries (26). While both approaches provide indicators for economic conditions, the approaches differ in the type of economic conditions they proxy. The prior approach (26) uses African DHS data to construct an index for village-level asset wealth, including measures for quality of living (e.g., if households have running water), and then trains a neural network to directly predict this index from a combination of Landsat and DMSP OLS night light intensity data. In contrast, our approach intends to proxy regional economic activity as indicated in administrative statistics, and thus represents primarily industrial economic activity rather than asset wealth of villages. Moreover, by classifying Landsat pixels into the six surface groups before using them to predict economic activity, our approach offers a direct measure for land cover with a potential for applications in regional science studies. The prior work (26) thus demonstrates that satellite data can be used to predict a particular developmental characteristic (village asset wealth), while our approach demonstrates that satellite data can be trained to predict both disaggregated and potentially missing or erroneous economic activity data (e.g., GDP at disaggregated levels within a county).

To make the comparison, we produce our surface groups proxy for four African countries—Guinea, Togo, Uganda, and Zimbabwe—using the procedure we outline in Section A1.6. Choosing these four countries ensures the fairest possible comparison, because for them the prior approach (26) yields an above-average prediction quality (according to R^2 reported in Fig. 2 of 26). The prior work (26) provides both its village-level asset wealth index and its prediction of this index for the years available in the underlying DHS data—2012 for Guinea; 2013 for Togo; 2009, 2011, and 2014 for Uganda; and 2010 and 2015 for Zimbabwe. The locations of villages are indicated by the coordinates of their geographic centers. Similar to the prior approach, we consider the area within a radius of 6.72 kilometers of a village's center for predicting the village's asset wealth

[31] Conducting the FE analysis for household income at the grid level yields similar results, with the grid-cell and year FE explaining 99.7% of the overall variation in household income and neither including surface groups nor including night light intensity increases adjusted R^2 (table A16). However, the grid-level analysis can draw on only five observation years (2009–2013) and thus much fewer years than the county-level analysis (14 years, 2000–2013). For such short time series, FE estimation in general is an inappropriate econometric method. Therefore, we do not further interpret these results.

Table A17. OLS prediction of asset wealth index in African countries using surface groups (village level)

Dep. var.: AWI	Guinea (1)	Togo (2)	Uganda (3)	Zimbabwe (4)
$ln(builtup + 1)$	0.430***	0.559***	0.774***	0.668***
	(0.066)	(0.051)	(0.041)	(0.034)
$ln(grass + 1)$	0.321***	-0.038	-0.050	-0.248***
	(0.093)	(0.083)	(0.043)	(0.093)
$ln(crops + 1)$	-0.452***	-0.458***	-0.559***	-0.403***
	(0.115)	(0.089)	(0.052)	(0.073)
$ln(forest + 1)$	-0.298***	0.030	0.007	-0.182***
	(0.076)	(0.054)	(0.028)	(0.060)
$ln(noveg + 1)$	-0.039	-0.042	-0.265***	0.014
	(0.054)	(0.040)	(0.021)	(0.063)
$ln(water + 1)$	0.236***	-0.067***	0.021	0.184***
	(0.059)	(0.024)	(0.017)	(0.032)
Year FE	n/a	n/a	Yes***	Yes***
$\%cloud$	0.466	-0.081	-0.365	0.198
	(0.627)	(0.379)	(0.268)	(0.644)
N	300	300	778	793
R^2	0.624	0.663	0.533	0.457

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. AWI denotes the DHS-based asset wealth index from prior work (26). Available years are 2012 for Guinea, 2013 for Togo, 2009, 2011, and 2014 for Uganda, and 2010 and 2015 for Zimbabwe.

Table A18. OLS prediction of asset wealth index in African countries using surface groups (district level)

Dep. var.: AWI	Guinea (1)	Togo (2)	Uganda (3)	Zimbabwe (4)
$ln(builtup+1)$	0.226	0.738**	0.424***	0.242*
	(0.211)	(0.249)	(0.076)	(0.123)
$ln(grass+1)$	0.751**	-0.300	0.106**	-0.485**
	(0.320)	(0.307)	(0.046)	(0.224)
$ln(crops+1)$	-0.766**	-0.611*	-0.462***	-0.125
	(0.371)	(0.290)	(0.074)	(0.205)
$ln(forest+1)$	-1.034***	0.239	0.018	-0.085
	(0.265)	(0.197)	(0.029)	(0.128)
$ln(noveg+1)$	-0.230	-0.016	-0.201***	0.118
	(0.198)	(0.140)	(0.030)	(0.116)
$ln(water+1)$	0.934***	-0.043	-0.034***	0.151*
	(0.259)	(0.150)	(0.010)	(0.081)
Year FE	n/a	n/a	Yes***	Yes***
$\%cloud$	-19.351***	-1.979	-8.864**	-16.622
	(5.772)	(8.518)	(4.218)	(16.197)
N	34	21	397	120
R^2	0.657	0.675	0.389	0.227

Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. AWI denotes the DHS-based asset wealth index from prior work (26). Available years are 2012 for Guinea, 2013 for Togo, 2009, 2011, and 2014 for Uganda, and 2010 and 2015 for Zimbabwe.

with surface groups. For each of the four countries separately, we run an OLS regression of the surface groups, the percentage of cloud cover, and year FE (if applicable) on the prior work's (26) DHS-based asset wealth index (see table A17 for the regression results). The predictions derived from these regressions allow us to calculate the percentage of the variation in the asset wealth index our approach explains and to compare it to the corresponding percentage the prior approach (26) explains.

The results of this comparison show that our approach also contributes to explaining the variation in the prior work's (26) DHS-based asset wealth index. Pooling over all villages in the four countries, our approach explains 59.7% of the variation in the asset wealth index, compared to 73.6% with the prior approach (26) (corresponds to red R^2 in Fig. 2a of 26).[32] Our surface groups proxy thus explains a significant percentage of the index, although lower than the prior approach (26). Despite our metric not being designed to identify asset wealth like the prior metric (26), our approach performs 81.1% as well as the prior metric (26) in predicting asset wealth.

While the prior approach in Africa (26) is designed to optimally predict the asset wealth index this work constructs from DHS data, our approach focuses on predicting a much broader proxy for regional economic activity. Both approaches explain substantial variation in the outcome variables they respectively predict. Each approach has comparative advantages and disadvantages depending on the research question (e.g., advantage for focused, village-level analyses in developing countries with the prior approach of 26, advantage for broader regional-level analyses in developed countries with our new approach). Satellite data can provide insight, predictability, and accuracy to various developmental indicators when trained specifically toward predicting the outcome in context.

A2.5 Surface groups economic proxy

The six surface groups can be combined into a single-variable proxy by computing a predicted indicator of economic activity using our OLS model specified in Eq. A4. To establish the external validity of such a single-variable proxy, for both GDP and household income we estimate Eq. A4 using only one randomly selected half of the sample (the training sample). With the OLS coefficients obtained from the training-sample estimation, we predict GDP and household income for the second half of the sample (the left-out sample), that is, we predict $\widehat{ln(GDP)}$ and $\widehat{ln(HHI)}$. To assess whether this predicted single-variable proxy is as valid as the original proxy, we then re-estimate the OLS model using only the left-out sample and using the single-variable proxy as independent variable instead of the original proxy. Again, we proceed similarly for DMSP OLS night light intensity to have a benchmark comparison.

Tables A19 and A20 present the estimation results for GDP and household income, respectively. In the specifications using the single-variable proxy as independent variable (columns 2 and 4), the surface groups-based proxy explains a higher percentage of the

[32] Calculating this indicator separately for each of the four countries and then averaging it, our approach explains 56.9% of the variation in the asset wealth index, compared to 78.8% with the prior approach (26) (corresponds to black R^2 in Fig. 2a of 26). Conducting the analyses at the administrative district level (see table A18 for the OLS regression results) yields indicators of 69.4% vs. 81.8% when pooling over all districts in the four countries and weighting by the number of villages (corresponds to red weighted R^2 in Fig. 2b of 26), 71.4% vs. 90.9% when separating by country and weighting (corresponds to black weighted R^2 in Fig. 2b of 26), 50.9% vs. 63.2% when pooling and not weighting (corresponds to red unweighted R^2 in Fig. 2b of 26), and 48.7% vs. 78.8% when separating and not weighting (corresponds to black unweighted R^2 in Fig. 2b of 26).

Table A19. OLS prediction of single-variable proxy for GDP using surface groups and using DMSP OLS night light intensity (county level, 2000–2013)

	Surface groups		DMSP OLS night light intensity	
	Training sample	Left-out sample	Training sample	Left-out sample
Dep. var.: $ln(GDP)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.371***			
	(0.051)			
$ln(grass + 1)$	-0.136***			
	(0.019)			
$ln(crops + 1)$	-0.260***			
	(0.017)			
$ln(forest + 1)$	-0.159***			
	(0.016)			
$ln(noveg + 1)$	-0.261***			
	(0.022)			
$ln(water + 1)$	-0.001			
	(0.022)			
$ln(NL_{DMSPOLS} + 1)$			0.434***	
			(0.024)	
$\widehat{ln(GDP)}$ from (1)		1.014***		
		(0.033)		
$\widehat{ln(GDP)}$ from (3)				0.995***
				(0.054)
Year FE	Yes***	Yes*	Yes**	Yes***
Federal state FE	Yes***	Yes***	Yes***	Yes***
N	2,687	2,715	2,687	2,715
Adj. R^2	0.610	0.634	0.457	0.488

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A20. OLS prediction of single-variable proxy for household income using surface groups and using DMSP OLS night light intensity (grid level, 2009–2013)

	Surface groups		DMSP OLS night light intensity	
	Training sample	Left-out sample	Training sample	Left-out sample
Dep. var.: $ln(HHI)$	(1)	(2)	(3)	(4)
$ln(builtup + 1)$	1.414***			
	(0.003)			
$ln(grass + 1)$	-0.126***			
	(0.002)			
$ln(crops + 1)$	-0.371***			
	(0.002)			
$ln(forest + 1)$	-0.066***			
	(0.002)			
$ln(noveg + 1)$	-0.172***			
	(0.002)			
$ln(water + 1)$	-0.212***			
	(0.002)			
$ln(NL_{DMSPOLS} + 1)$			0.955***	
			(0.003)	
$\widehat{ln(HHI)}$ from (1)		0.996***		
		(0.001)		
$\widehat{ln(HHI)}$ from (3)				0.996***
				(0.003)
Year FE	Yes***	Yes**	Yes***	Yes***
Federal state FE	Yes***	Yes**	Yes***	Yes***
N	367,668	369,958	367,668	369,958
Adj. R^2	0.676	0.674	0.307	0.308

Robust standard errors in parentheses. All models include intercept.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

variation in economic activity than the night lights-based proxy (63.4% vs. 48.8% for GDP and 67.4% vs. 30.8% for household income). This finding corroborates the findings of the county-level analysis of GDP and of the grid-level analysis of household income. Therefore, the surface groups can provide a valid single-variable proxy of economic activity, which might be desirable when economic activity is the dependent variable in an analysis.

Finally, Table A21 shows the results of an OLS estimation that uses all available GDP data (2000–2018) to train the single-variable surface groups-based economic proxy. Moreover, to improve the quality of the prediction, this estimation also includes the regional percentage of pixels with cloud cover as a further indicator of potential measurement error (see section A2.2). This estimation underlies the time series plots of predicted GDP in Fig. 3. In producing Fig. 3, we follow our recommendation in Section A2.3 and remove outlier observations. More specifically, we consider a county-year observation an outlier if the number of *builtup* pixels in that year is more than twice as large as the median number of *builtup* pixels among all observations from the same county or if more than ten percent of the observation's pixels are covered by clouds.

Table A21. OLS prediction of GDP using surface groups (county level, 2000–2018)

Dep. var.: $ln(GDP)$	(1)
$ln(builtup + 1)$	1.307***
	(0.029)
$ln(grass + 1)$	-0.114***
	(0.012)
$ln(crops + 1)$	-0.259***
	(0.010)
$ln(forest + 1)$	-0.187***
	(0.010)
$ln(noveg + 1)$	-0.185***
	(0.013)
$ln(water + 1)$	-0.006
	(0.013)
Year FE	Yes***
Federal state FE	Yes***
$\%cloud$	-5.029***
	(0.792)
N	7,397
Adj. R^2	0.630

Robust standard errors in parentheses. Model includes intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

[1] J. I. Dingel, A. Miscio, and D. R. Davis. Cities, lights, and skills in developing economies. *Journal of Urban Economics*, 125:103174, 2021.

[2] R. Hodler and P. A. Raschky. Regional favoritism. *The Quarterly Journal of Economics*, 129(2):995–1033, 2014.

[3] S. Michalopoulos and E. Papaioannou. Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, 81(1):113–152, 2013.

[4] M. Pinkovskiy and X. Sala-i-Martin. Lights, camera ... income! Illuminating the national accounts–household surveys debate. *The Quarterly Journal of Economics*, 131(2):579–631, 2016.

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

[6] R. Kulkarni, K. Haynes, R. Stough, and J. Riggle. Light based growth indicator (LGBI): Exploratory analysis of developing a proxy for local economic growth based on night lights. *Regional Science Policy & Practice*, 3(2):101–113, 2011.

[7] C. Mellander, J. Lobo, K. Stolarick, and Z. Matheson. Night-time light data: A good proxy measure for economic activity? *PLoS One*, 10(10):e0139779, 2015.

[8] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(1219):eabe8628, 2021.

[9] P. Lehnert. *Higher education institutions and their impact on employment and innovation: Regional identification and empirical analyses.* Dissertation, University of Zurich, 2020.

[10] M. Burchfield, H. G. Overman, D. Puga, and M. A. Turner. Causes of sprawl: A portrait from space. *The Quarterly Journal of Economics*, 121(2):587–633, 2006.

[11] A. D. Foster and M. R. Rosenzweig. Economic growth and the rise of forests. *The Quarterly Journal of Economics*, 118(2):601–637, 2003.

[12] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

[13] D. Donaldson and A. Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198, 2016.

[14] S. A. Morain. A brief history of remote sensing applications, with emphasis on Landsat. In D. Liverman, E. F. Moran, R. R. Rindfuss, and P. C. Stern, editors, *People and pixels: Linking remote sensing and social science*, pages 28–50. The National Academies Press, Washington, DC, 1998.

[15] D. L. Williams, S. Goward, and T. Arvidson. Landsat: Yesterday, today, and tomorrow. *Photogrammetric Engineering & Remote Sensing*, 72(10):1171–1178, 2006.

[16] A. M. Dewan and Y. Yamaguchi. Land use and land cover in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Applied Geography*, 29(3):390–401, 2009.

[17] R. Goldblatt, W. You, G. Hanson, and A. K. Khandelwal. Detecting the boundaries of urban areas in India: A dataset for pixel-based image classification in Google Earth Engine. *Remote Sensing*, 8:634, 2016.

[18] X. Liu, G. Hu, Y. Chen, X. Li, X. Xu, S. Li, and F. Pei. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. *Remote Sensing of Environment*, 209:227–239, 2018.

[19] A. Pekkarinen, L. Reithmaier, and P. Strobl. Pan-European forest/non-forest mapping with Landsat ETM+ and CORINE Land Cover 2000 data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(2):171–183, 2009.

[20] W. Yu, S. Zang, C. Wu, W. Liu, and X. Na. Analyzing and modeling land use land cover change (LUCC) in the Daqing City, China. *Applied Geography*, 31(2):600–608, 2011.

[21] S. Keola, M. Andersson, and O. Hall. Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. *World Development*, 66:322–334, 2015.

[22] P. C. Sutton and R. Costanza. Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics*, 41(3):509–527, 2002.

[23] M. A. Davis, J. D. M. Fisher, and T. M. Whited. Macroeconomic implications of agglomeration. *Econometrica*, 82(2):731–764, 2014.

[24] A. Holl. Manufacturing location and impacts of road transport infrastructure: Empirical evidence from Spain. *Regional Science and Urban Economics*, 34(3):341–363, 2004.

[25] R. Goldblatt, K. Heilmann, and Y. Vaizman. Can medium-resolution satellite imagery measure economic activity at small geographies? Evidence from Landsat in Vietnam. *The World Bank Economic Review*, 34(3):635–653, 2020.

[26] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11:2583, 2020.

[27] H. Balzter, B. Cole, C. Thiel, and C. Schmullius. Mapping CORINE land cover from Sentinel-1A SAR and SRTM digital elevation model data using random forests. *Remote Sensing*, 7(11):14876–14898, 2015.

[28] K.-S. Han, J.-L. Champeaux, and J.-L. Roujean. A land cover classification product over France at 1 km resolution using SPOT4/VEGETATION data. *Remote Sensing of Environment*, 92(1):52–66, 2004.

[29] K. Neumann, M. Herold, A. Hartley, and C. Schmullius. Comparative assessment of CORINE2000 and GLC2000: Spatial analysis of land cover data for Europe. *International Journal of Applied Earth Observation and Geoinformation*, 9(4):425–437, 2007.

[30] A. Pérez-Hoyos, F. J. García-Haro, and J. San-Miguel-Ayanz. A methodology to generate a synergetic land-cover map by fusion of different land-cover products. *International Journal of Applied Earth Observation and Geoinformation*, 19:72–87, 2012.

[31] L. T. Waser and M. Schwarz. Comparison of large-area land cover products with national forest inventories and CORINE land cover in the European Alps. *International Journal of Applied Earth Observation and Geoinformation*, 8(3):196–207, 2006.

[32] Esri, Maxar, Earthstar Geographics, USDA FSA, USGS, Aerogrid, IGN, IGP, GIS User Community. World Imagery (updated January 14, 2022. Dataset, Esri, Redlands, CA, 2009.

[33] R. Goldblatt, M. F. Stuhlmacher, B. Tellman, N. Clinton, G. Hanson, M. Georgescu, C. Wang, F. Serrano-Candela, A. K. Khandelwal, W.-H. Cheng, and R. C. Balling Jr. Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sensing of Environment*, 205:253–275, 2018.

[34] J. V. Henderson, A. Storeygard, and D. N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012.

[35] Leibniz Institute for Economic Research (RWI) and Micromarketing-Systeme and Consult GmbH (microm). RWI-GEO-GRID: Socio-economic data on grid level – scientific use file (wave 8). version: 1. Dataset, RWI, Essen, 2019.

[36] J. V. Henderson, T. Squires, A. Storeygard, and D. Weil. The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133(1):357–406, 2018.

[37] A. Schneider. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. *Remote Sensing of Environment*, 124:689–704, 2012.

[38] U.S. Geological Survey (USGS). Landsat collections. Fact sheet 2018-3049, USGS, Sioux Falls, 2018.

[39] G. Büttner, J. Feranec, and G. Jaffrain. Corine land cover update 2002: Technical guidelines. Technical report 89, European Environment Agency, Copenhagen, 2002.

[40] European Environment Agency (EEA). Copernicus land monitoring service: Corine land cover. Dataset, European Union, 2021.

[41] M. A. Wulder, J. C. White, S. N. Goward, J. G. Masek, J. R. Irons, M. Herold, W. B. Cohen, T. R. Loveland, and C. E. Woodcock. Landsat continuity: Issues and opportunities for land cover monitoring. *Remote Sensing of Environment*, 112(3):955–969, 2008.

[42] M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, and C. E. Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, 122:2–10, 2012.

[43] M. A. Wulder, J. C. White, T. R. Loveland, C. E. Woodcock, A. S. Belward, W. B. Cohen, E. A. Fosnight, J. Shaw, J. G. Masek, and D. P. Roy. The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, 185:271–283, 2016.

[44] K. Lulla, M. D. Nellis, B. Rundquist, P. K. Srivastava, and S. Szabo. Mission to earth: LANDSAT 9 will continue to view the world. *Geocarto International*, 36(20):2261–2263, 2021.

[45] J. G. Masek, M. A. Wulder, B. Markham, J. McCorkel, C. J. Crawford, J. Storey, and D. T. Jenstrom. Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, 248:111968, 2020.

[46] D. R. Lyzenga. Remote sensing of bottom reflectance and water attenuation parameters in shallow water using aircraft and Landsat data. *International Journal of Remote Sensing*, 2(1):71–82, 1981.

[47] M. Torresani, D. Rocchini, R. Sonnenschein, M. Zebisch, M. Marcantonio, C. Ricotta, and G. Tonon. Estimating tree species diversity from space in an alpine conifer forest: The Rao's Q diversity index meets the spectral variation hypothesis. *Ecological Informatics*, 52:26–34, 2019.

[48] T. R. Loveland and J. L. Dwyer. Landsat: Building a strong future. *Remote Sensing of Environment*, 122:22–29, 2012.

[49] B. L. Markham, J. C. Storey, D. L. Williams, and J. R. Irons. Landsat sensor performance: History and current status. *IEEE Transactions on Geoscience and Remote Sensing*, 42(12):2691–2694, 2004.

[50] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy, T. A. Scambos, C. B. Schaaf, J. R. Schott, Y. Sheng, E. F. Vermote, A. S. Belward, R. Bindschadler, W. B. Cohen, F. Gao, J. D. Hipple, P. Hostert, J. Huntington, C. O. Justice, A. Kilic, V. Kovalskyy, Z. P. Lee, L. Lymburner, J. G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R. H. Wynne, and Z. Zhu. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

[51] B. L. Markham and D. L. Helder. Forty-year calibrated record of earth-reflected radiance from Landsat: A review. *Remote Sensing of Environment*, 122:30–40, 2012.

[52] J. A. Rumerman. *NASA historical data books (SP-4012) volume VI: NASA space applications, aeronautics and space research and technology, tracking and data acquisition/Support operations, commercial programs, and resources, 1979–1988.* NASA History Division, Washington, DC, 1999. Updated October 15, 2010.

[53] P. Griffiths, S. van der Linden, T. Kuemmerle, and P. Hostert. A pixel-based Landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5):2088–2101, 2013.

[54] G. Trianni, G. Lisini, E. Angiuli, E. A. Moreno, P. Dondi, A. Gaggia, and P. Gamba. Scaling up to national/regional urban extent mapping using Landsat data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3710–3719, 2015.

[55] J. W. Rouse Jr, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Progress Report RSC 1978-1, Texas A&M University Remote Sensing Center, College Station, 1973.

[56] J. Xue and B. Su. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017, 2017.

[57] S. K. McFeeters. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432, 1996.

[58] B.-C. Gao. NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996.

[59] Y. Zha, J. Gao, and S. Ni. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3):583–594, 2003.

[60] Z. Zhu and C. E. Woodcock. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, 118:83–94, 2012.

[61] European Environment Agency (EEA). Copernicus Land Service – Pan-European component: CORINE Land Cover. EEA, Copenhagen, 2017.

[62] G. Büttner and B. Kosztra. CLC2018 technical guidelines. Technical Report Service Contract No 3436/R0-Copernicus/EEA.56665, Environment Agency Austria, Vienna, 2017.

[63] B. Kosztra, G. Büttner, G. Hazeu, and S. Arnold. Updated CLC illustrated nomenclature guidelines. Technical Report Service Contract No 3436/R0-Copernicus/EEA.57441 Task 3, D3.1 – Part 1, Environment Agency Austria, Vienna, 2019.

[64] L. Matejicek and V. Kopackova. Changes in croplands as a result of large scale mining and the associated impact on food security studied using time-series Landsat images. *Remote Sensing*, 2(6):1463–1480, 2010.

[65] J. Cihlar and L. J. M. Jansen. From land cover to land use: A methodology for efficient land use mapping over large areas. *The Professional Geographer*, 53(2):275–289, 2001.

[66] A. J. Comber, R. Wadsworth, and P. Fisher. Using semantics to clarify the conceptual confusion between land cover and land use: The example of 'forest'. *Journal of Land Use Science*, 3(2):185–198, 2008.

[67] J. Feranec, G. Hazeu, S. Christensen, and G. Jaffrain. Corine land cover change detection in Europe (case studies of the Netherlands and Slovakia). *Land Use Policy*, 24(1):234–247, 2007.

[68] P. Fisher, A. J. Comber, and R. Wadsworth. Land use and land cover: Contradiction or complement. In Peter Fisher and David J. Unwin, editors, *Re-presenting GIS*, pages 85–98. John Wiley & Sons Ltd, West Sussex, 2005.

[69] J. Gallego and C. Bamps. Using CORINE land cover and the point survey LUCAS for area estimation. *International Journal of Applied Earth Observation and Geoinformation*, 10(4):467–475, 2008.

[70] S. W. Myint, P. Gober, A. Brazel, S. Grossmann-Clarke, and Q. Weng. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing of Environment*, 115(5):1145–1161, 2011.

[71] T. G. Whiteside, G. S. Boggs, and S. W. Maier. Comparing object-based and pixel-based classifications for mapping savannas. *International Journal of Applied Earth Observations and Geoinformation*, 13(6):884–893, 2011.

[72] D. C. Duro, S. E. Franklin, and M. G. Dubé. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118:259–272, 2012.

[73] L. Dingle Robertson and D. J. King. Comparison of pixel- and object-based classification in land cover change mapping. *International Journal of Remote Sensing*, 32(6):1505–1529, 2011.

[74] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu. A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:277–293, 2017.

[75] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random Forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[76] S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

[77] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.

[78] K. Millard and M. Richardson. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, 7(7):8489–8515, 2015.

[79] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[80] T.-T. Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015.

[81] R. Bindschadler. Landsat coverage of the earth at high latitudes. *Photogrammetric Engineering & Remote Sensing*, 69(12):1333–1339, 2003.

[82] Copernicus Land Monitoring Service. Clc seamless data coverage. V2020_v20u1, Copernicus Land Monitoring Service, 2020.

[83] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, 5:180214, 2018.

[84] W. Köppen. Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet. *Meteorologische Zeitschrift*, 1(21):215–226, 1884.

[85] P. Breidenbach and L. Eilers. RWI-GEO-GRID: Socio-economic data on grid level. *Journal of Economics and Statistics*, 238(6):609–616, 2018.

[86] Q. Huang, X. Yang, B. Gao, Y. Yang, and Y. Zhao. Application of DMSP/OLS nighttime light images: A meta-analysis and a systematic literature review. *Remote Sensing*, 6(8):6844–6866, 2014.

[87] C. D. Elvidge, K. E. Baugh, E. A. Kihn, H. W. Kroehl, E. R. Davis, and C. W. Davis. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6):1373–1379, 1997.

[88] C. D. Elvidge, M. Zhizhin, T. Ghosh, F.-C. Hsu, and J. Taneja. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021.