PNAS nexus

Supplementary Information for
Proxying economic activity with daytime satellite imagery: Filling data gaps across time and space

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

*To whom correspondence should be addressed: patrick.lehnert@business.uzh.ch (P.L.),
ebettinger@stanford.edu (E.B.)

This PDF file includes:

Supplementary texts S1 to S3
Figures S1 to S10
Tables S1 to S27
SI References

Structure of supplementary material

This supplementary material 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 paper and the underlying data.

In Section S1, we describe the procedure we develop for retrieving the surface groups measure from daytime satellite imagery and conduct the internal validity analysis. In Section S2, we perform several analyses to demonstrate the value of surface groups as a proxy for economic activity. In Section S3, we present our application of surface groups to the causal analysis comparing the effect of higher education institutions on regional innovation in East and West German regions.

S1 Computation of surface groups

S1.1 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., 1, 2). 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. S1 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., 6, 7), 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 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

[1] We use the Landsat Collections distributed by the U.S. Geological Survey (3) and directly accessible through GEE.

[2] The acronym "CORINE" stands for "coordination of information on the environment" (4).

[3] The CLC data are distributed by the EEA (5) and directly accessible through GEE.

the classification (i.e., the internal validity), we perform five-fold cross-validation.

S1.2 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 (8, 9). Throughout the history of Landsat, the various operating agencies have launched eight satellites, one of which (Landsat-6) failed to reach orbit (10, 11). As of 2022, Landsat-7, Landsat-8, and Landsat-9 remain active, with Landsat-9 having launched only in September 2021 (12, 13).[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 (16). 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 (17). Table S1 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 S1). 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 (18) 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 (19). 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 is scarce for Germany until the 1980s (11). 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 S1). 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 (20), we compute the surface groups annually. As Landsat satellites record a geographic location on earth multiple times per year (11), we have to use annual composites of these records. Unfortunately, pre-processed annual composites incorporating imagery from multiple Landsat

[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 (14) and the investigation of tree species diversity in the Alps (15).

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 (21). 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 (22).

For both the compositing and the actual pixel classification (see section S1.4), we follow studies from the remote-sensing literature (e.g., 1, 23) 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 (24, 25); Second, the Normalized Difference Water Index (NDWI) differentiates open water from other surfaces (26);[5] Third, the Normalized Difference Built-up Index (NDBI) differentiates built-up surfaces from other surfaces (28). Similar to prior work (1), we compute these three indices for Landsat data as follows:

$$NDVI_p = \frac{NIR_p - RED_p}{NIR_p + RED_p} \tag{S1}$$

$$NDWI_p = \frac{GREEN_p - NIR_p}{GREEN_p + NIR_p} \tag{S2}$$

$$NDBI_p = \frac{SWIR1_p - NIR_p}{SWIR1_p + NIR_p} \tag{S3}$$

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 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., 29). 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

[5] Another index exists under the name "NDWI", which was developed to identify liquid water inside plants (27). This other NDWI relies on different spectral bands than the NDWI we use.

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 S1) 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.

S1.3 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 (4, 30). 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) (31). Each database includes a map for the respective year with a pixel resolution of 100 meters, indicating land cover in a variety of classes (31, 32).

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 (31, 32). This imagery stems from various satellites, including Landsat satellites for CLC1990, CLC2000, and CLC2018 (31). The remote-sensing literature provides successful combinations of CLC and Landsat data in geospatial analyses (e.g., 33, 34).

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 S1.2). 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 35). 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) (36–39). Given that automated analyses of satellite data can detect only land cover and that determining land use requires manual interpretation (39), 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., 40, 41). However, as this literature does not provide an unambiguous assignment of CLC classes to larger groups with similar surface characteristics, we identify similarities in the spectral reflectance patterns of pixels in the CLC level 3 classes through visual inspection and 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 (7), best represents similar surface characteristics in Germany:

- Built-up surface (*builtup*): The surface group *builtup* contains surfaces with building of non-natural materials such as concrete, metal, and glass (e.g., residential buildings, industrial plants, roads). This surface group thus includes all artificial surfaces (CLC class 1) except for green urban areas (CLC class 141), which prior work (42) shows to have a lower resemblance with artificial surfaces than with vegetated surfaces.

- Grassy surfaces (*grass*): The surface group *grass* contains surfaces covered by grass or other plants with similar surface reflectance (e.g., natural grassland, city parks). This surface group thus includes pastures (CLC class 23) and natural grassland (CLC class 321), which have similar surface characteristics (6, 42). In addition, due to the similarities in surface reflectance that we detect in our trials, we add to the surface group *grass* the green urban areas (CLC class 141) that we exclude from the surface group *builtup*.

- Surfaces with crop fields (*crops*): The surface group *crops* contains surfaces with vegetation for agricultural purposes (e.g., hayfields, vineyards). This surface groups thus includes all agricultural areas (CLC class 2) except for pastures (CLC class 23), which belong to the surface groups *grass*.

- Forest-covered surfaces (*forest*): The surface groups *forest* contains surfaces covered by trees or other plants with similar surface reflectance (e.g., mixed forests, moors). This surface group thus includes all forests and semi-natural areas (CLC class 3) except for grassland (CLC class 321), which belongs to the surface group *grass*, and open spaces with little or no vegetation (CLC class 33), which differ in spectral reflectance from the remaining CLC classes in the surface group *forest* (42, 43)

- Surfaces without vegetation (*noveg*): The surface group *noveg* contains surfaces with (almost) no vegetation or buildings (e.g., bare rock, sand plains). This surface group thus includes open spaces with little or no vegetation (CLC class 33), which we exclude from the surface group *forest*.

- Water surfaces (*water*): The surface group *water* contains any type of water surface (e.g., rivers, lakes). This surface group thus includes wetlands (CLC class 4) and water bodies (CLC class 5), which we aggregate following prior work (44).

Table S2 provides a correspondence table of CLC classes for our algorithm. 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.

S1.4 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 (23), we perform pixel-based classification. For every pixel in our training sample, the machine-learning algorithm predicts the pixel's surface group from the spectral values and the added indices NDVI, NDWI, and NDBI. Compared to machine-learning algorithms that perform object-based classification, which additionally considers information from neighboring pixels, pixel-based classification requires less computational power (45, 46). Although the majority of studies in the remote-sensing literature suggest that machine-learning algorithms performing object-based classification better predict land cover than those performing pixel-based classification (e.g., 46), some studies find no significant performance difference (e.g., 47). In particular, one of these studies finds no significant difference using Landsat data (48). Therefore, given the spatial and temporal size of the data we analyze in this paper, we decided to use pixel-based classification through traditional machine-learnning algorithms such as RF or Support Vector Machines. Our assessments of external validity (see section "External validity" in the paper and section S2) confirm that choosing this machine-learning classification yields a valid proxy for economic activity.

However, in future research, object-based classification through deep-learning algorithms based on convolutional neural networks (CNNs) such as U-Net (49) or ResNet (50) has the potential to even better classify different types of land cover and thus offers and important avenue for future improvements. The remote-sensing literature has successfully applied CNNs to land-cover classification for specific geographic study areas and detected potential improvements in prediction accuracy (e.g., 51–53). While beyond the scope of our paper, extending these applications to a global scale has great potential for improving economic proxies for GDP or other socioeconomic indicators.[6]

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 algorithms when applied to land cover classification (e.g., 54, 56). 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 (23). Furthermore, RF requires less computational power (54). 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 (23).[8] Therefore, RF with ten decision trees best suits the purpose of our paper.

[6] We thank an anonymous reviewer for pointing us towards the potential benefits of CNNs in land-cover classification and for providing additional arguments for the discussion on potential improvements in this context and in the main text of the paper.

[7] For a description of the RF method's application for land cover classification, see, e.g., 54, and for a description of the method's application in economics, see, e.g., 55.

[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 (23).

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 (56, 57). Therefore, we choose a slightly larger number of pixels in the training data than in comparable applications from the remote-sensing literature (e.g., 2, 23) to achieve an accurate classification, but keep this number low enough to maintain a reasonable computational effort.

By restricting the pool of pixels from which we draw the stratified random sample we use as training data, we substantially reduce the influence of potentially imprecise ground-truth observations resulting from the difference in spatial resolution between Landsat (30 meters) and CLC data (100 meters). Due to the coarser CLC resolution, the CLC surface groups might not be accurate for Landsat pixels at the boundary of two CLC surface areas. While a CLC pixel might correctly belong to the *builtup* surface group, because more than half of the pixel's area contains built-up surfaces, not all Landsat pixels that fall within the CLC pixel are necessarily *builtup*. Therefore, we do not use Landsat pixels that fall within CLC pixels at the boundary of two CLC surface areas as training data, that is, the CLC pixel and all its neighboring pixels must belong to the same surface group. This restriction reduces the number of imprecise ground-truth observations and thus improves the quality of the training data, which correlates positively with the accuracy of the RF prediction when applied to land-cover classification (56, 58, 59).

A comparison of Figs. 1 *A* and 1 *B* (right column) in the paper illustrates the reason for the sampling restriction to inner CLC pixels. For example, we do not use the CLC pixels at the boundary of the water and grass surface areas in Fig. 1 *B*. At this boundary, some of the CLC water pixels contain parts of the vegetation at the lakeshore, and, vice versa, some of the CLC grass pixels contain parts of the lake. Therefore, the boundary CLC pixels are not representative for the true surface groups of the Landsat pixels that fall within these CLC pixels. Excluding the unrepresentative ground-truth in-formation reduces the risk of imprecise ground-truth information in the training data. The resulting benefit of this exclusion is that the algorithm can more accurately classify unrepresentative pixels (e.g., the forest and grass pixels that belong to the nature reserve (*Vogelinsel im Altmühlsee*) in the south-west of Fig. 1 *C*, as well as similar examples throughout Germany).

S1.5 Accuracy assessment of output data

To assess the prediction accuracy of our classification in the output data, we follow prior work (23) 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

[9] For descriptions and discussions of this method, see, e.g., 60, 61

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 the *Internal Validity* section in the paper). Complementing the five-fold cross-validation results for the entire sample in Table 1 of the paper, Tables S3 through S8 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, 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 S2.[10]

S1.6 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 (64, 65). 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

[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 (62)

[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 63).

[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 63).

(zone E) (64).[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 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.

[14] The Köppen-Geiger climate classification data (from 64) are publicly available at https://doi.org/10.6084/m9.figshare.6396959.

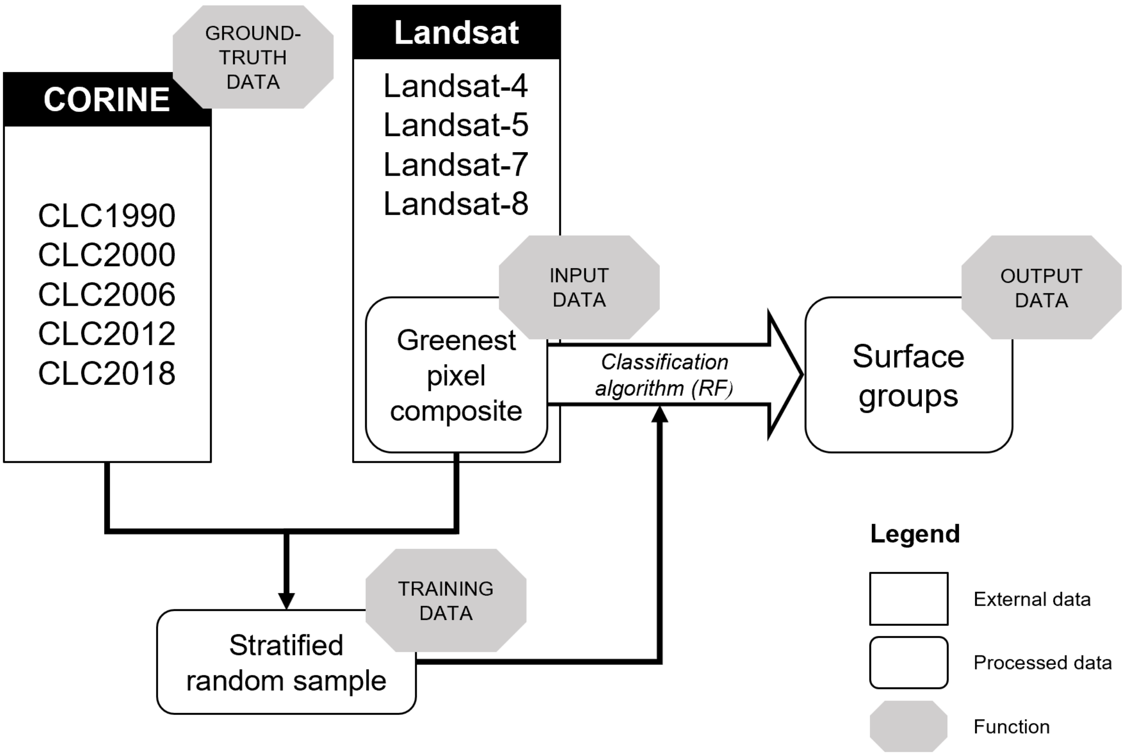

Fig. S1. Overview of procedure for detecting surface groups.

Table S1. 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 (10, 11, 16, 23, 66, 67). 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 (11, 66). 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 (67).

Table S2. CLC classes and assignment for algorithm

CLC class						Our
Level 1		Level 2		Level 3		algorithm
1	Artificial surfaces	11	Urban fabric	111	Continuous urban fabric	*builtup*
				112	Discontinuous urban fabric	*builtup*
		12	Industrial, commercial, and transport units	121	Industrial or commercial units and public facilities	*builtup*
				122	Road and rail networks and associated land	*builtup*
				123	Port areas	*builtup*
				124	Airports	*builtup*
		13	Mine, dump, and construction sites	131	Mineral extraction sites	*builtup*
				132	Dump sites	*builtup*
				133	Construction sites	*builtup*
		14	Artificial, non-agricultural vegetated areas	141	Green urban areas	*grass*
				142	Sport and leisure facilities	*builtup*
2	Agricultural areas	21	Arable land	211	Non-irrigated arable land	*crops*
				212	Permanently irrigated arable land	*crops*
				213	Rice fields	*crops*
		22	Permanent crops	221	Vineyards	*crops*
				222	Fruit tree and berry plantations	*crops*
				223	Olive groves	*crops*
		23	Pastures	231	Pastures, meadows, and other permanent grasslands under agricultural use	*grass*
		24	Heterogeneous agricultural areas	241	Annual crops associated with permanent crops	*crops*
				242	Complex cultivation patterns	*crops*
				243	Land principally occupied by agriculture, with significant areas of natural vegetation	*crops*
				244	Agro-forestry areas	*crops*
3	Forest and semi-natural areas	31	Forests	311	Broad-leaved forest	*forest*
				312	Coniferous forest	*forest*
				313	Mixed forest	*forest*
		32	Shrubs and/or herbaceous vegetation associations	321	Natural grassland	*grass*
				322	Moors and heathland	*forest*
				323	Sclerophyllous vegetation	*forest*
				324	Transitional woodland/shrub	*forest*
		33	Open spaces with little or no vegetation	331	Beaches, dunes, and sand plains	*noveg*
				332	Bare rock	*noveg*
				333	Sparsely vegetated areas	*noveg*
				334	Burnt areas	*noveg*
				335	Glaciers and perpetual snow	*noveg*
4	Wetlands	41	Inland wetlands	411	Inland marshes	*water*
				412	Peatbogs	*water*
		42	Coastal wetlands	421	Coastal salt marshes	*water*
				422	Salines	*water*
				423	Intertidal flats	*water*
5	Water bodies	51	Inland waters	511	Water courses	*water*
				512	Water bodies	*water*
		52	Marine waters	521	Coastal lagoons	*water*
				522	Estuaries	*water*
				523	Sea and ocean	*water*

Authors' illustration based on a prior illustration (68, p. 27). CLC classes listed as in official CLC nomenclature (32).

Table S3. 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 of the paper.

Table S4. Five-fold cross-validation results with respect to grassy surfaces (surface group *grass*)

Year	Overall	True-positive	True-negative	Balanced	User's
Year	accuracy	rate	rate	accuracy	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 of the paper.

Table S5. 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 of the paper.

Table S6. 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 of the paper.

Table S7. 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 of the paper.

Table S8. 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 of the paper.

S2 External validity analyses

S2.1 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., 69, 70). The administrative GDP measure is available at the county (*Kreis*) level[15] from 2000. Second, we use the socioeconomic dataset RWI-GEO-GRID (71) 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), assess the surface groups' performance in relation to data on built-up land cover from the Global Human Settlement Layer (GHSL), and compare the predictive value of surface groups to a prior approach in Africa (72).

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.

S2.2 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 "Regionaldaten-

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

bank 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 (70), we use real (i.e., deflated)[19] GDP measures in euros as a validation measure for our analyses. We denote real GDP as GDP.

Second, we use RWI-GEO-GRID (71), 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 73). 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 (73). The grid cells in this dataset follow the system of the European Reference Grid distributed by the European Soil Data Centre (ESDAC)[20] (73). 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 (73). 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 S2.3), 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 (74). 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 (69, 75, 76). The technological developments of the 21st century have improved both the accessibility of night lights data and the computational capabilities for processing these data (20). Consequently, night lights

[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=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).

data have become an attractive data source for economists in the last decade. Similar to prior work (70), 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 three other data sources in Section S2.4. First, we use VIIRS night lights data, which prior research has confirmed to be a valid proxy for economic activity (77, 78), 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 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 (79), 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 the GHSL data, which are provided by the European Commission and contain, among other things, information on built-up land cover in five-year intervals. These pixel-level data have a 100-meter resolution and are based on a classification of daytime satellite imagery (including Landsat). From these data, we extract two measures of built-up land cover—one that indicates absolute built-up surface in square meters and one that indicates absolute built-up volume in cubic meters—and take their regional averages (denoted as $GHSL_{surface}$ and $GHSL_{volume}$).[24] The information on built-up volume thus allows us to assess the potential role of building height in proxying economic activity. While the GHSL data start already in 1975, they are available only in five-year intervals.

Third, we use an index for village-level asset wealth from prior work in Africa (72). The authors (72) 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 $\widehat{AWI}$).[25]

To assess the value of the surface groups we derive from Landsat data as a proxy for

[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 (70), we use the average of those observations.

[23] We use the annual VIIRS night lights composites version 2 (79), available from the Colorado School of Mines at https://eogdata.mines.edu/nighttime_light/annual/v20/ (accessed October 27, 2021).

[24] For surface, we use the 100-meter resolution GHS-BUILT-S R2022A data (80), which are publicly available from the European Commission at https://ghsl.jrc.ec.europe.eu/download.php?ds=bu (accessed December 7, 2022). For volume, we use the 100-meter resolution GHS-BUILT-V R2022A data (81), which are publicly available from the European Commission at https://ghsl.jrc.ec.europa.eu/download.php?ds=builtV (accessed December 7, 2022). For more information on the concept and methodology underlying the GHSL data, see the GHSL data package (82).

[25] Both the original and the predicted asset wealth index are available as a supplement to 72.

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*.[26] 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 $\%cloud$).[27]

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.

S2.3 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:

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t} + \beta_2 C_{i,t} + \nu_{i,t} \tag{S4}$$

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.[28]

[26] 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).

[27] Other reasons for missing values could be implausible spectral values or inexistence of imagery (see section S1.2). However, cloud cover is the most likely reason.

[28] The household income data are also available for 2005, but we exclude this year to consistently examine

To assess whether the combination of surface groups is a valid proxy for economic activity, we follow prior work (70) 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 and are thus not collinear by construction. However, an increase in one of the surface groups is associated with a decrease in at least one other surface group. Still, as our purpose is to achieve the best possible prediction of economic activity and not to identify the exact association between each surface group and economic activity, we can include all six surface groups in the regressions despite a certain degree of collinearity. In an assessment of night light intensity as a country-level proxy for GDP (70), 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.[29]

County-level analysis of GDP. The results of the county-level analysis with real GDP as the dependent variable in Table S9 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), 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* in the paper show that the statistical distribution of the residuals from the OLS regressions with control variables (columns 2 and 4 of table S9) 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

patterns in the temporal distribution of the regression residuals by maintaining a data structure of consecutive years.

[29] 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 S2.2) 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 S25 and S26).

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. S3 illustrates the temporal and spatial distribution of the residuals from the regressions in column 2 of Table S9. For reference, Fig. S2 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. S3 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. S4 proceeds similarly for DMSP OLS night light intensity, illustrating the residuals from the regression in column 4 of Table S9.

Figs. S3 and S4 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. S3. For DMSP OLS night light intensity, this pattern appears for 339 counties (85.0%), leading to the four maps in Fig. S4 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. S3 and S4 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 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. S3.

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. S5 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. S5 *A*) and federal states (fig. S5 *B*). In addition to the original data points obtained from the regressions, Fig. S5 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 S10 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* in the paper 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. S6 and S7 show the spatial and temporal distribution of the prediction error in household income for surface groups and DMSP OLS night light intensity, respectively. 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 S1.5), 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.[30] To obtain more valid results, we apply outlier corrections in the applications of surface groups in this work, that is, in the comparison of GDP developments across German counties (fig. 3 in the paper) and in the analysis of the impacts higher education institutions in East and West Germany (section *Essential improvements in social science research through surface groups data* in the paper). For details on the outlier removal procedure, see Sections S2.5 and S3.

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 historical development (83). However, for small regional units and early time series, surface groups constitute a valuable and more accurate proxy for economic activity.

S2.4 Additional validation analyses

We present four 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 assess the surface groups' performance in relation to GHSL data that provide other metrics for built-up land cover. Third, we analyze within-region heterogeneity in predicted GDP to demonstrate that surface groups

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

enable the isolation of subregional changes in economic activity. Fourth, a comparison to prior work in Africa (72) 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. S4 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 S11 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 S12 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.

Comparison to GHSL land-cover metrics. To evaluate how the surface groups perform compared to other land-cover metrics in proxying economic activity, we use the GHSL data.[31] These data (a) provide an alternative measure of built-up surface and (b) go beyond our surface groups by offering information on building volume (i.e., they add the dimension of building height). As the extension of built-up land associated with economic activity can occur both horizontally and vertically, this additional information can be an important determinant of economic activity.

As in our comparisons to night light intensity, we regress the economic validation indicators (GDP and household income) on the natural logarithms of either one of the two GHSL-based measures ($GHSL_{surface}$ and $GHSL_{volumne}$), year FE, and federal state FE according to Eq. S4, and compare them to a similar specification with the surface groups metrics as independent variables instead of the GHSL-based metrics. The observation years are those available in all datasets (2000, 2005, 2010, and 2015 for the GDP analysis; 2005,[32] 2010, and 2015 for the household income analysis).

[31] We thank an anonymous reviewer for suggesting to link the surface groups proxy and the GHSL data.

[32] Note that we do not use the 2005 data on household income in our original comparison to night light intensity. The reason is that we analyze consecutive annual data in this comparison to consistently examine patterns in the temporal distribution of the regression residuals (see section S2.3), and the household income data are not available from 2006 through 2008. As the comparison to GHSL data is feasible only in five-year intervals, we can use the 2005 data on household income in this new comparison.

Table S13 presents the county-level estimation results for GDP. At this level, our surface groups measure and the GHSL built-up surfaces measure perform equally well in predicting GDP (explaining 64.4% and 64.3% of the variation in GDP in columns 1 and 2, respectively), suggesting that other measures of built-up land cover at the surface can match the performance of our proxy. The GHSL built-up volume measure outperforms our surface groups measure at the county level (explaining 83.1% of the variation in GDP in column 3). This result suggests that adding the dimension of building height can substantially improve economic proxies.

The results of the grid-level analysis of household income in Table S14 suggest that our surface groups perform better in proxying economic activity at this very small regional level compared to both GHSL measures. The surface groups explain 69.4% of the variation in grid-level household income (column 1), whereas GHSL built-up surface explains 55.1% (column 2) and GHSL built-up volume 57.8% of this variation (column 3). Thus the combination of different types of land cover appears to play a more important in role in proxying grid-level economic activity than building height.

We conclude from these analyses that building volume as indicated in the GHSL data can play an important role in proxying economic activity but comes at the cost of losing temporal information. As GHSL data are available only in five-year intervals, they have the disadvantage of not providing sufficient information for answering research questions that address short term changes in economic activity and thus benefit from annual data. As an example, the application of surface groups to studying immediate economic effects of higher education institutions after the fall of the Iron Curtain (section *Essential improvements of social science research through surface groups data* in the paper) requires annual data, and other potential applications to studying policy interventions with outcomes expected in the short term both in Europe and across the world have the same requirement. In contrast, the advantage of the GHSL data is that they may more precisely proxy economic activity, particularly in areas where heights matter, if less precise temporal information is acceptable. Moreover, our proxy offers additional information on other types of land cover potentially related to economic activity (e.g., cropland) and can thus provide more precision in countries with, for example, a large agricultural sector.

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. S4 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. We conduct these analyses at the level of municipalities, the smallest administrative regional unit in Germany.[33] Although GDP data do not exist at the municipality level, we

[33] 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. We use the territorial status of 2017, because it corresponds to the territorial status of the data on higher education institutions from prior work (84) that we use for the social science application we present in the paper and in Section S3.

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 S9, 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. S4 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 S17 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. S8 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 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. S8 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. S9, which illustrates for one county (*Wunsiedel*) as an example the two indicators plotted in Fig. S8. 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. S9 *A6* and *B6*.

Figs. S9 *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. S9 *A3* and *A4* illustrate the surface groups-based prediction of standardized municipality-level GDP for these two years, and Figs. S9 *B3* and *B4* the DMSP OLS night light intensity-based prediction. Figs. S9 *A5* and *B5* indicate the difference between Figs. S9 *A3* and *A4* and that between Figs. S9 *B3* and *B4*, respectively, that is, the changes in the GDP predictions between 2000 and 2013. Fig. S9 *A6* then shows the municipality-county difference in the change in predicted standardized GDP using surface groups as predictor and Fig. S9 *B6* shows this difference using DMSP OLS night light intensity (i.e., the same indicators for which Fig. S8 plots the distribution).

Two properties become noticeable. First, the colors in Figs. S9 *A3* through *A6* are much more intense than in Figs. S9 *B3* through *B6*. This higher intensity is in line with Fig. S8, 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. S9 *A6*), a pattern that is not visible when using DMSP OLS night light intensity (Fig. S9 *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. S4 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 (70). 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 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 S15 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 S9, 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.[34]

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 (72). While both approaches provide indicators for economic conditions, the approaches differ in the type of economic conditions they proxy. The prior approach (72) 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 (72) 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 S1.6. Choosing these four countries ensures the fairest possible comparison, because for them the prior approach (72) yields an above-average prediction quality (according to R^2 reported in Fig. 2 of 72). The prior work (72) 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 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 (72) DHS-based asset wealth index (see table S18 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

[34] 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 S16). 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.

corresponding percentage the prior approach (72) explains.

The results of this comparison show that our approach also contributes to explaining the variation in the prior work's (72) 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 (72) (corresponds to red R^2 in Fig. 2a of 72).[35] Our surface groups proxy thus explains a significant percentage of the index, although lower than the prior approach (72). Despite our metric not being designed to identify asset wealth like the prior metric (72), our approach performs 81.1% as well as the prior metric (72) in predicting asset wealth.

While the prior approach in Africa (72) 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 72, 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.

S2.5 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. S4. To establish the external validity of such a single-variable proxy, for both GDP and household income we estimate Eq. S4 using only one randomly selected quarter of the sample (the training sample). With the OLS coefficients obtained from the training-sample estimation, we predict GDP ($\widehat{ln(GDP)}$) and household income ($\widehat{ln(HHI)}$) for one randomly selected half of the sample (the left-out sample). We do not use the remaining quarter of the sample to avoid too strong similarity between the training and left-out samples due to spatial proximity of the training and left-out regions. Randomization takes place at the region level so that all observations from one region end up in the same sample.

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 S20 and S21 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 variation in economic activity than the night lights-based proxy (63.2% vs. 50.6% for GDP and 67.6% vs. 30.9% for household income). This finding corroborates the findings

[35] 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 (72) (corresponds to black R^2 in Fig. 2a of 72). Conducting the analyses at the administrative district level (see table S19 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 72), 71.4% vs. 90.9% when separating by country and weighting (corresponds to black weighted R^2 in Fig. 2b of 72), 50.9% vs. 63.2% when pooling and not weighting (corresponds to red unweighted R^2 in Fig. 2b of 72), and 48.7% vs. 78.8% when separating and not weighting (corresponds to black unweighted R^2 in Fig. 2b of 72).

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 S22 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 S2.2). This estimation underlies the time series plots of predicted GDP in Fig. 3 in the paper. In producing Fig. 3, we follow our recommendation in Section S2.3 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.

S2.6 Combination of surface groups and GHSL data

While this Section S2 has shown how the surface groups can function as one proxy for economic activity, combining them with other metrics can further improve our understanding of regional economic activity. As an example, in Tables S23 and S24 we combine our six surface groups with GHSL built-up volume for GDP and household income, respectively. The regressions in these tables correspond to the model in Eq. S4 but use both surface groups and GHSL built-up volume as independent variables. We find that the combination of the two data sources outperforms the separate specifications in proxying economic activity, explaining 86.8% of the variation in county-level GDP and 73.1% of the variation in grid-level household income.

These results indicate that combining proxies can improve the prediction of economic activity. Studies that do not require specific information available in only one dataset (such as the consecutive annual time series of the surface or the height dimension of the GHSL data) can thus benefit from combining data sources. Further examining how economic proxies can be combined for analyzing regional economic activity thus provides great opportunities for future research.

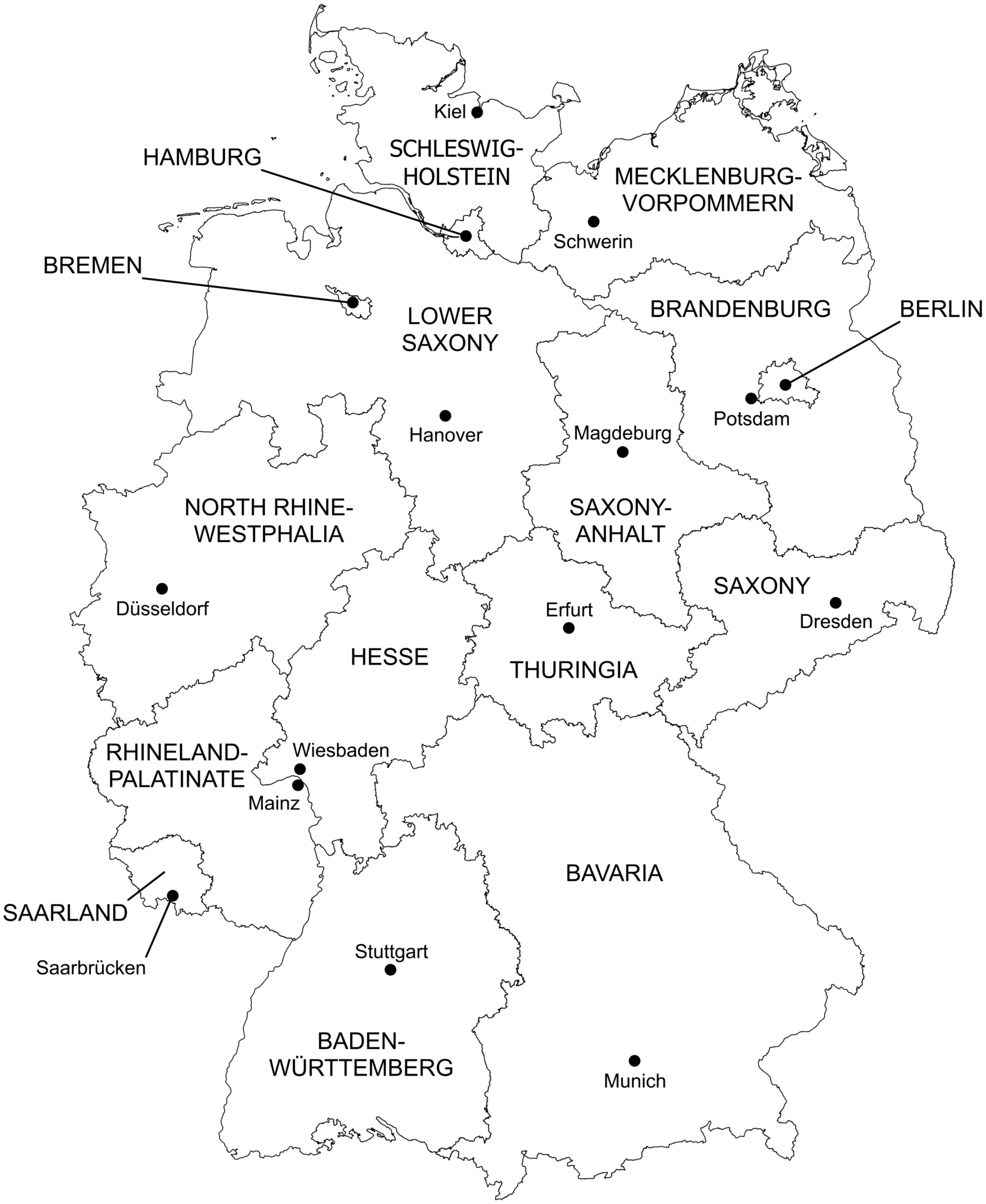

Fig. S2. Reference map of German federal states and their capitals.

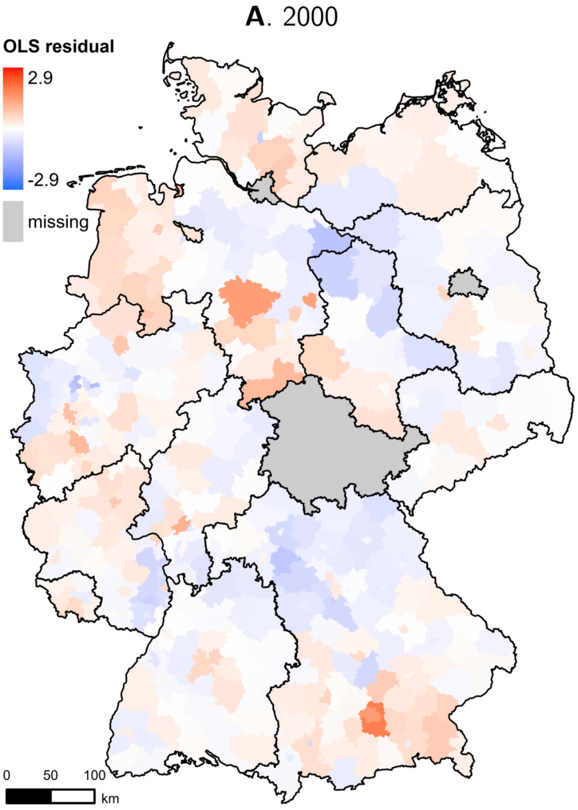

Fig. S3. Spatial and temporal distribution of GDP residuals for surface groups. Maps illustrate residuals from the regression in column 2 of Table S9.

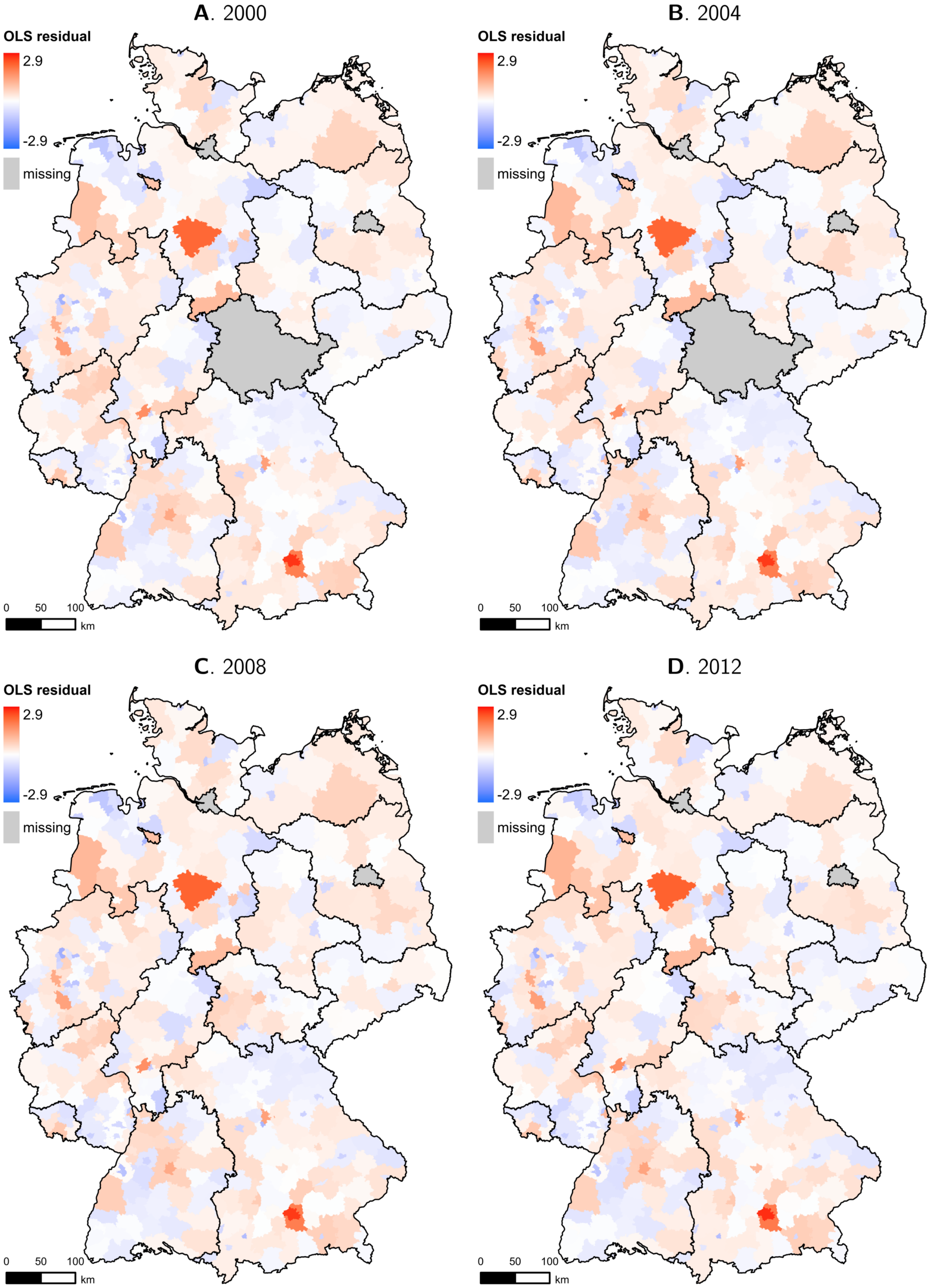

Fig. S4. Spatial and temporal distribution of GDP residuals for DMSP OLS night light intensity. Maps illustrate residuals from the regression in column 4 of Table S9.

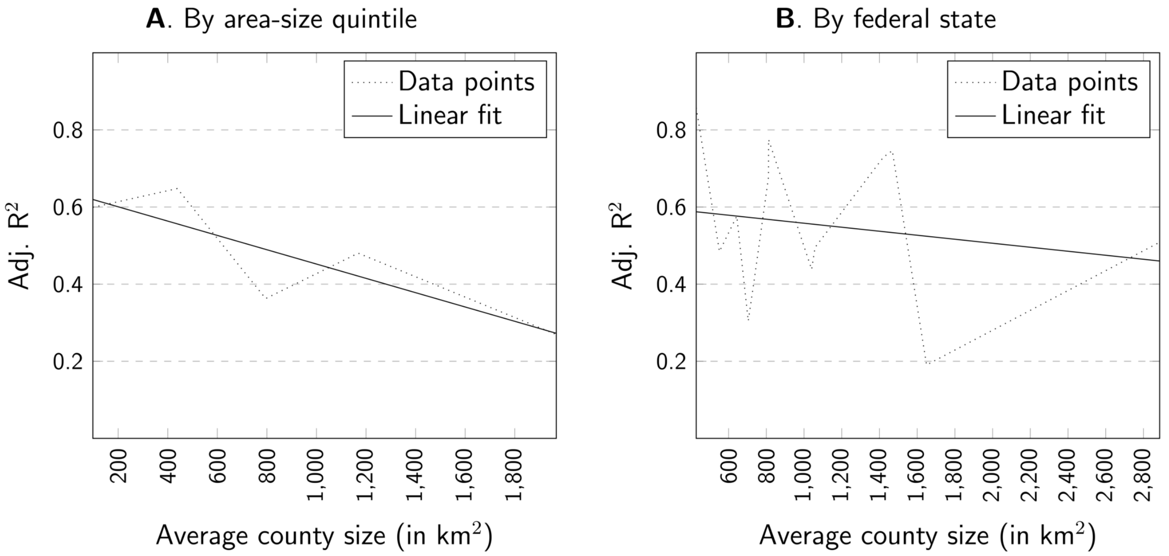

Fig. S5. Adj. R^2 by county-size group. Values stem from separate regressions of surface groups on GDP corresponding to the specification in column 2 of Table S9

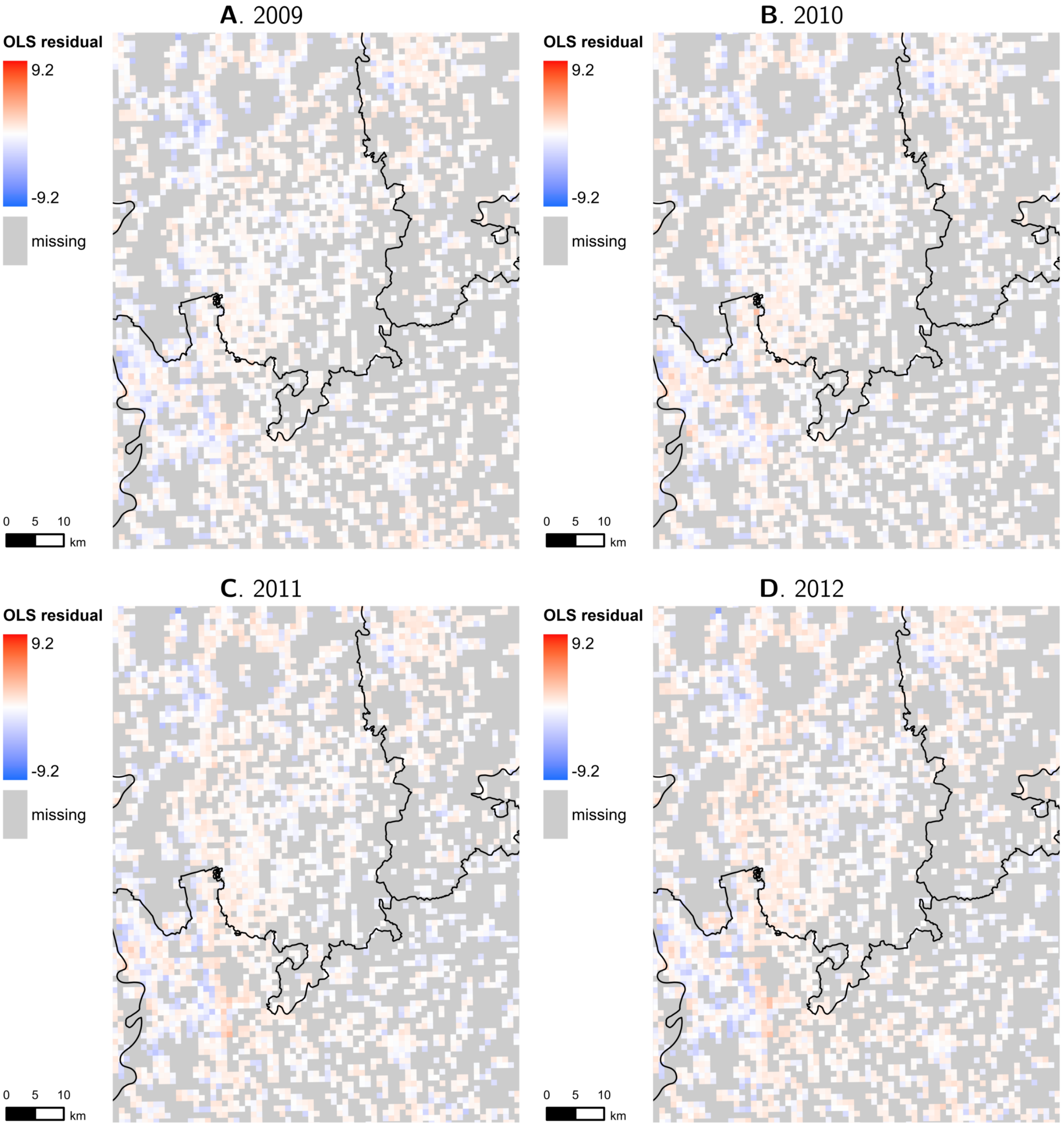

Fig. S6. Spatial and temporal distribution of household income residuals for surface groups. Maps illustrate residuals from the regression in column 2 of Table S10. Maps show an area at the borders of the four federal states *Rhineland-Palatinate*, *Hesse*, *Baden-Württemberg*, and *Bavaria*.

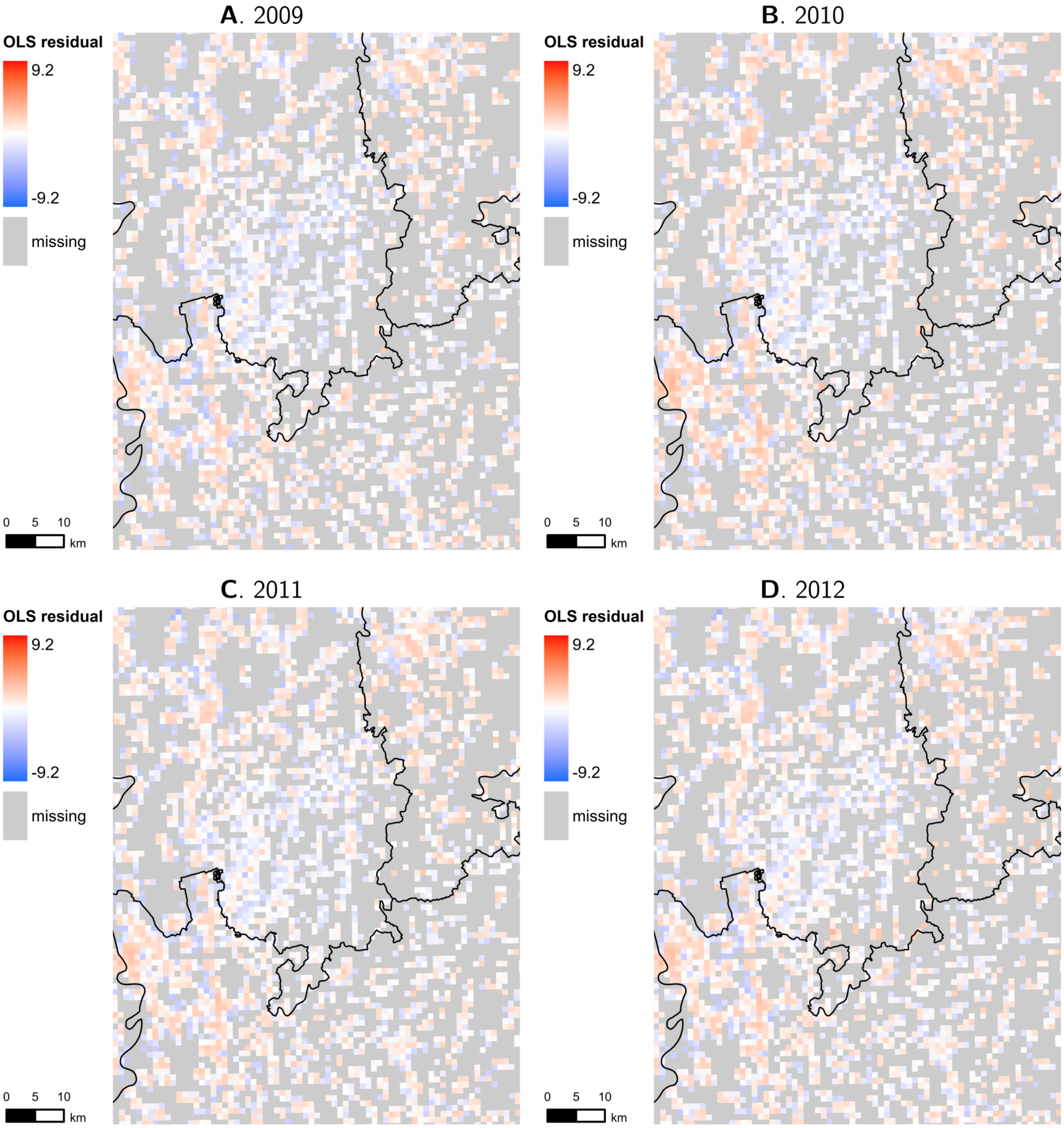

Fig. S7. Spatial and temporal distribution of household income residuals for DMSP OLS night light intensity. Maps illustrate residuals from the regression in column 4 of Table S10. Maps show an area at the borders of the four federal states *Rhineland-Palatinate*, *Hesse*, *Baden-Württemberg*, and *Bavaria*.

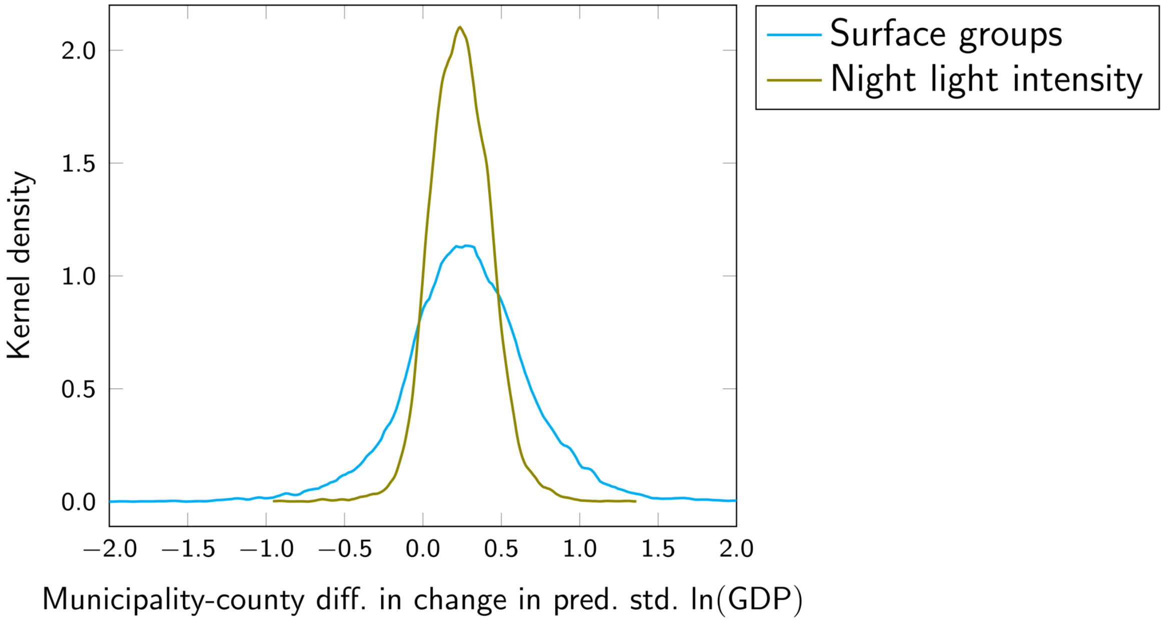

Fig. S8. Distribution of municipality-county difference in the change in predicted standardized $ln(GDP)$ between 2000 and 2013 for surface groups-based and DMSP OLS night light intensity-based prediction. Figure shows univariate kernel density estimates at 300 points using the Epanechnikov kernel function with a kernel half-width of 0.025.

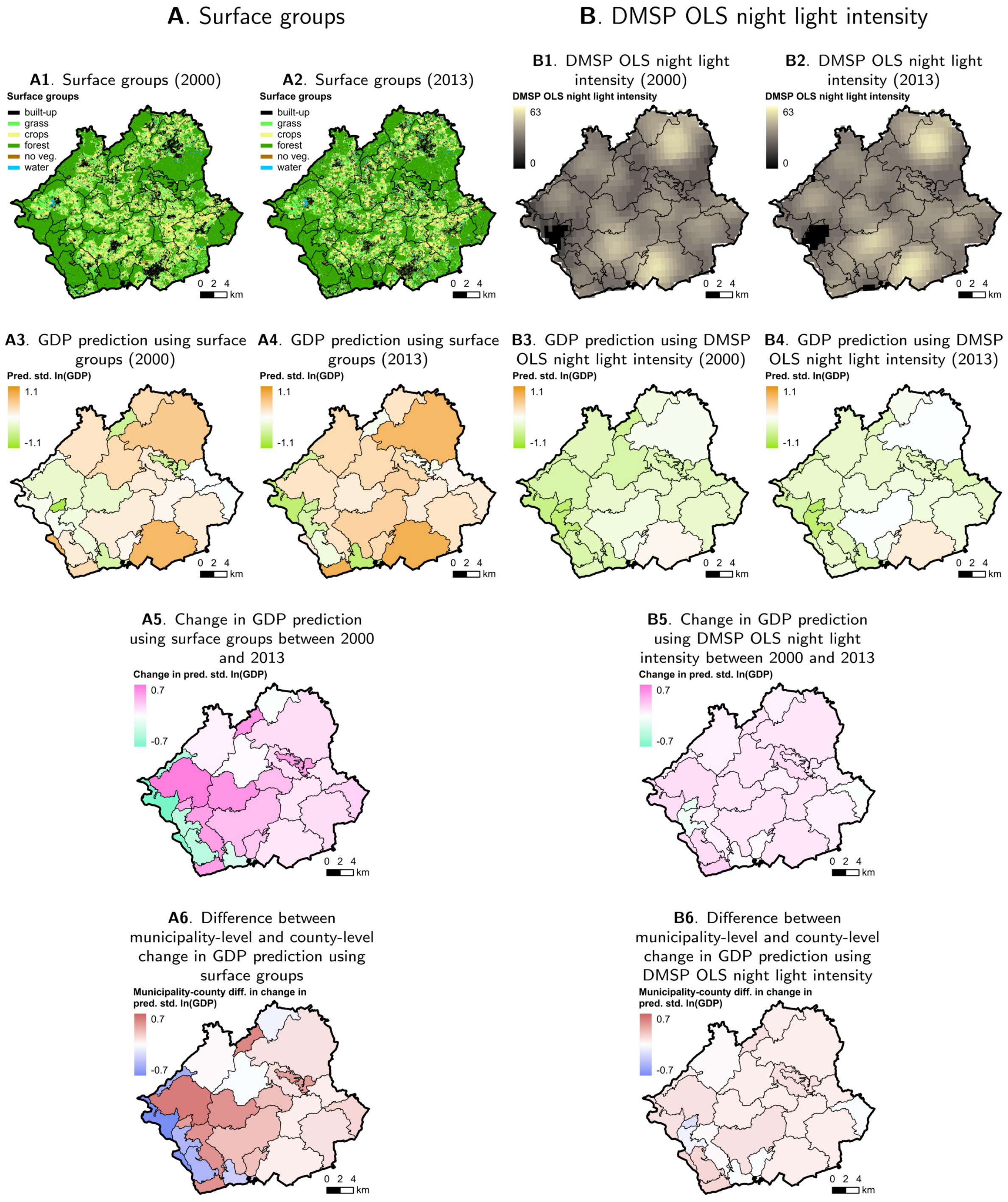
A. Surface groups
B. DMSP OLS night light intensity
A1. Surface groups (2000)
Surface groups
built-up
grass
crops
forest
no veg.
water
A2. Surface groups (2013)
Surface groups
built-up
grass
crops
forest
no veg.
water
B1. DMSP OLS night light intensity (2000)
DMSP OLS night light intensity
63
0
B2. DMSP OLS night light intensity (2013)
DMSP OLS night light intensity
63
0
A3. GDP prediction using surface groups (2000)
Pred. std. ln(GDP)
1.1
-1.1
A4. GDP prediction using surface groups (2013)
Pred. std. ln(GDP)
1.1
-1.1
B3. GDP prediction using DMSP OLS night light intensity (2000)
Pred. std. ln(GDP)
1.1
-1.1
B4. GDP prediction using DMSP OLS night light intensity (2013)
Pred. std. ln(GDP)
1.1
-1.1
A5. Change in GDP prediction using surface groups between 2000 and 2013
Change in pred. std. ln(GDP)
0.7
-0.7
B5. Change in GDP prediction using DMSP OLS night light intensity between 2000 and 2013
Change in pred. std. ln(GDP)
0.7
-0.7
A6. Difference between municipality-level and county-level change in GDP prediction using surface groups
Municipality-county diff. in change in pred. std. ln(GDP)
0.7
-0.7
B6. Difference between municipality-level and county-level change in GDP prediction using DMSP OLS night light intensity
Municipality-county diff. in change in pred. std. ln(GDP)
0.7
-0.7
0 2 4
km

Fig. S9. 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 S9. 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 S10. 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 S11. 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 S12. 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S13. OLS prediction of GDP using surface groups, using GHSL built-up surface, and using GHSL built-up volume (county level, 2000–2015 in 5-year intervals)

Dep. var.: $ln(GDP)$	Surface groups (1)	GHSL built-up surface (2)	GHSL built-up volume (3)
$ln(builtup + 1)$	1.296*** (0.059)		
$ln(grass + 1)$	-0.119*** (0.022)		
$ln(crops + 1)$	-0.258*** (0.020)		
$ln(forest + 1)$	-0.176*** (0.021)		
$ln(noveg + 1)$	-0.096*** (0.029)		
$ln(water + 1)$	-0.049*** (0.028)		
$ln(GHSL_{surface} + 1)$		0.803*** (0.030)	
$ln(GHSL_{volume} + 1)$			1.003*** (0.021)
Year FE	Yes***	Yes***	Yes***
Federal state FE	Yes***	Yes***	Yes***
N	1,550	1,550	1,550
Adj. R^2	0.644	0.643	0.831

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

Table S14. OLS prediction of household income using surface groups, using GHSL built-up surface, and using GHSL built-up volume (grid level, 2005–2015 in 5-year intervals)

		GHSL	GHSL
	Surface	built-up	built-up
	groups	surface	volume
Dep. var.: $ln(HHI)$	(1)	(2)	(3)
$ln(builtup + 1)$	1.363*** (0.003)		
$ln(grass + 1)$	-0.164*** (0.002)		
$ln(crops + 1)$	-0.342*** (0.002)		
$ln(forest + 1)$	-0.066*** (0.002)		
$ln(noveg + 1)$	-0.111*** (0.002)		
$ln(water + 1)$	-0.227*** (0.002)		
$ln(GHSL_{surface} + 1)$		0.468*** (0.001)	
$ln(GHSL_{volume} + 1)$			0.416*** (0.001)
Year FE	Yes***	Yes***	Yes***
Federal state FE	Yes***	Yes***	Yes***
N	438,601	438,601	438,601
Adj. R^2	0.694	0.551	0.578

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

Table S15. 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		
	No proxy	Surface groups	DMSP OLS night light intensity	No proxy	Surface groups	DMSP OLS night light intensity
Dep. var.: $ln(GDP)$	(1)	(2)	(3)	(4)	(5)	(6)
$ln(builtup + 1)$		0.023^{***} (0.006)			0.023^{***} (0.007)	
$ln(grass + 1)$		-0.002 (0.006)			-0.002 (0.006)	
$ln(crops + 1)$		-0.021^{***} (0.005)			-0.021^{***} (0.005)	
$ln(forest + 1)$		0.007 (0.005)			0.007 (0.007)	
$ln(noveg + 1)$		-0.012^{***} (0.003)			-0.012^{***} (0.003)	
$ln(water + 1)$		0.001 (0.003)			0.001 (0.004)	
$ln(NL_{DMSPOLS} + 1)$			0.076^{***} (0.010)			0.076^{***} (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 S16. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S17. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S18. OLS prediction of asset wealth index in African countries using surface groups (village level)

	Guinea	Togo	Uganda	Zimbabwe
Dep. var.: AWI	(1)	(2)	(3)	(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 (72). Available years are 2012 for Guinea, 2013 for Togo, 2009, 2011, and 2014 for Uganda, and 2010 and 2015 for Zimbabwe.

Table S19. OLS prediction of asset wealth index in African countries using surface groups (district level)

	Guinea	Togo	Uganda	Zimbabwe
Dep. var.: AWI	(1)	(2)	(3)	(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 (72). Available years are 2012 for Guinea, 2013 for Togo, 2009, 2011, and 2014 for Uganda, and 2010 and 2015 for Zimbabwe.

Table S20. 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)
$ln(builtup+1)$	1.368^{***}			
	(0.083)			
$ln(grass+1)$	-0.172^{***}			
	(0.025)			
$ln(crops+1)$	-0.221^{***}			
	(0.025)			
$ln(forest+1)$	-0.084^{***}			
	(0.021)			
$ln(noveg+1)$	-0.212^{***}			
	(0.027)			
$ln(water+1)$	-0.087^{***}			
	(0.031)			
$ln(NL_{DMSPOLS}+1)$			0.296^{***}	
			(0.036)	
$\widehat{ln(GDP)}$ from (1)		1.067^{***}		
		(0.032)		
$\widehat{ln(GDP)}$ from (3)				1.727^{***}
				(0.073)
Year FE	Yes***	Yes	Yes	Yes
Federal state FE	Yes***	Yes***	Yes***	Yes***
N	1,324	2,764	1,324	2,764
Adj. R^2	0.643	0.632	0.475	0.506

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

Table S21. 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)
$ln(builtup + 1)$	1.415***			
	(0.005)			
$ln(grass + 1)$	-0.128***			
	(0.003)			
$ln(crops + 1)$	-0.381***			
	(0.003)			
$ln(forest + 1)$	-0.066***			
	(0.002)			
$ln(noveg + 1)$	-0.179***			
	(0.003)			
$ln(water + 1)$	-0.214***			
	(0.003)			
$ln(NL_{DMSPOLS} + 1)$			0.943***	
			(0.005)	
$ln(\widehat{HHI})$ from (1)		0.999***		
		(0.001)		
$ln(\widehat{HHI})$ from (3)				1.016***
				(0.004)
Year FE	Yes***	Yes	Yes***	Yes
Federal state FE	Yes***	Yes***	Yes***	Yes***
N	184,323	368,088	184,323	368,088
Adj. R^2	0.672	0.676	0.308	0.309

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

Table S22. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S23. OLS prediction of GDP combining surface groups and GHSL built-up volume (county level, 2000–2015 in 5-year intervals)

Dep. var.: $ln(GDP)$	(1)
$ln(builtup + 1)$	-0.058
	(0.041)
$ln(grass + 1)$	0.056***
	(0.014)
$ln(crops + 1)$	-0.135***
	(0.013)
$ln(forest + 1)$	-0.013
	(0.012)
$ln(noveg + 1)$	0.025
	(0.017)
$ln(water + 1)$	0.003
	(0.018)
$ln(GHSL_{volumne} + 1)$	1.103***
	(0.024)
Year FE	Yes***
Federal state FE	Yes***
$\%cloud$	-0.677
	(1.028)
N	1,550
Adj. R^2	0.868

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

Table S24. OLS prediction of GDP combining surface groups and GHSL built-up volume (grid level, 2005–2015 in 5-year intervals)

Dep. var.: $ln(HHI)$	(1)
$ln(builtup + 1)$	1.015***
	(0.003)
$ln(grass + 1)$	-0.131***
	(0.002)
$ln(crops + 1)$	-0.301***
	(0.002)
$ln(forest + 1)$	-0.035***
	(0.001)
$ln(noveg + 1)$	-0.133***
	(0.002)
$ln(water + 1)$	-0.170***
	(0.002)
$ln(GHSL_{volumne} + 1)$	0.174***
	(0.002)
Year FE	Yes***
Federal state FE	Yes***
$\%cloud$	0.007
	(0.051)
N	438,601
Adj. R^2	0.731

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

Table S25. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S26. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S3 Example for application of surface groups in social science research

In studying causal effects of higher education institutions in less developed East Germany compared to developed West Germany (section *Essential improvements in social science research through surface groups data* in the paper), we use our surface groups proxy because it allows us to compare economic conditions in East and West German regions before reunification. In addition, we use a dataset containing information on the locations and opening years of University of Applied Sciences (UAS) campuses in Germany from prior work (84),[36] which collects this information primarily through extensive online research. The original dataset extends back in time until 1980 and indicates the exact locations, opening years, and study fields of all public UAS campuses in Germany. Moreover, it contains annual municipality-level innovation outcomes based on patenting activities.

For our analysis, we use a municipality-level excerpt from the prior work's (84) dataset. For each municipality, this excerpt indicates whether in a given year a municipality is located within a 25-kilometer travel-distance radius of a UAS campus that offers study fields in science, technology, engineering, and mathematics (STEM). This definition of UAS campus areas and the restriction to STEM fields follow previous research on the innovation effects of UAS campus openings in Switzerland (88–90).[37] As innovation outcomes, the excerpt includes two indicators of regional innovation—patent quantity and patent quality. Patent quantity indicates the number of priority patent applications per municipality and year and patent quality indicates the average number of forward citations three years after a patent's publication per municipality and year. Both indicators are constructed as in prior work on Swiss UASs (88) and constitute well-established indicators for regional innovation (e.g., 91, 92). In the data excerpt we use in our analysis, the indicators are retrieved from the European Patent Office's Worldwide Patent Statistical Database (October 2019 version), which has complete information on patenting activities from 1980 for West Germany and from 1991 for East Germany (84).[38] This data excerpt thus allows us to study causal effects of UASs immediately after the fall of the Iron Curtain (for more details on the dataset and on UASs, see 84).

Descriptive analyses show that in 1991—that is, shortly after the fall of the Iron Curtain—East German regions lag far behing West German ones in both patent quantity and patent quality. Fig. S10 shows the differences between East and West German municipalities from 1991 through 2015 in both outcomes.[39] For both patent quantity (Fig. S10 *A*) and patent quality (Fig. S10 *B*), the difference is positive throughout the entire observation period, suggesting that East German municipalities had lower levels than West German ones immediately after reunification and never reach the West German levels. However, while the East-West gap increased in the first five to ten years after reunification, those gaps have been closing since 2005. Based on this starting point, we can use our surface groups proxy to examine whether the UASs helped close the gap between

[36] For developed countries, previous literature has shown the positive effects of higher education institutions in general (e.g., 85, 86) and of UASs in particular (e.g., 87, 88).

[37] The 25-kilometer travel-distance radius is based on commuting behavior and the restriction to STEM fields is done because patenting outcomes represent technological innovation rather than, for example, social innovation (88–90).

[38] We thank Dietmar Harhoff from the Max Planck Institute for Innovation and Competition in Munich for providing the patent data for this analysis.

[39] Due to its historical situation, we exclude the city of Berlin from this analysis.

less developed East Germany and developed West Germany. That is, we analyze whether UASs in East Germany yield different effects than UASs in West Germany and could thus bring East German regions with a UAS closer to their West German counterparts.

For this analysis, we divide the post-reunification period (beginning in 1991) into three-year periods p (i.e., $p = 1993$ denotes the first period from 1991 through 1993, $p = 1996$ the second period from 1994 through 1996, etc.).[40] We use the first three-year period as the baseline period (representing the level of innovation immediately after reunification) to compare the subsequent periods to this baseline. For every three-year period, we calculate the means of the two outcome variables at the municipality level.

To ensure a comparison of regions with similar levels of economic activity before reunification, we use the surface groups data as the only reliable proxy for regional economic development in East Germany before the fall of the Iron Curtain. More specifically, we perform propensity-score matching to compare similar regions affected by a UAS campus in East and West Germany, that is, regions that—other than being located in different parts of the country—have similar pre-reunification characteristics. To do so, we focus on municipalities with a UAS campus area in the first year of a three-year period p. To ensure similarity in pre-reunification economic activity, we match East German municipalities and West German ones based on their average pre-reunification growth in the six surface groups that proxy the pre-reunification trend in economic activity.[41] Thus we compare municipalities in East Germany with a UAS campus (denoted as *East*) to similar municipalities in West Germany. We conduct separate analyses for the seven three-year periods following the baseline period.[42] As outcome variables, we use the differences in patent quantity and patent quality between the observed three-year period and the baseline period, denoted as $PQUAN^{diff}$ and $PQUAL^{diff}$, respectively.

Our results of the propensity-score matching analysis in Table S27 show that the increase in patent quantity is significantly smaller in East German UAS regions than in West German ones until 2008, that is, even 17 years after reunification. The same type of educational policy thus has very different effects in a developed country as compared to a less developed, former communist country. Our surface groups proxy allows us to perform these causal analyses that otherwise would have been impossible or less reliable. Our detailed analyses also show that the effect on patent quality is roughly identical in similar East and West German regions. This finding again supports the importance of reliable data on economic activity at sufficiently disaggregated regional levels, such as the surface groups proxy we develop in our paper.

[40] As graduates are one important channel of knowledge transfer from higher education institutions to the private sector (e.g., 89, 93), we choose the minimum number of years a student needs for graduating from a UAS to determine period length for this analysis, thus following previous studies on UASs in Switzerland (e.g., 88, 89).

[41] Again, to achieve more valid results we follow our recommendation in Section S2.3 and remove outlier observations, that is, municipality-year observations with a number of *builtup* pixels more than twice as large as the median number of *builtup* pixels among all observations from the same municipality or with more than ten percent cloud cover.

[42] These are the three-year periods 1994–1996, 1997–1999, 2000–2002, 2003–2005, 2006–2008, 2009–2011, and 2012–2014. We do not consider the 2015–2017 period, because the underlying patent data are complete only until 2018 and we need to end our observation period at least three years earlier to ensure correct representation of the three-year citation window used to construct that patent quality indicator.

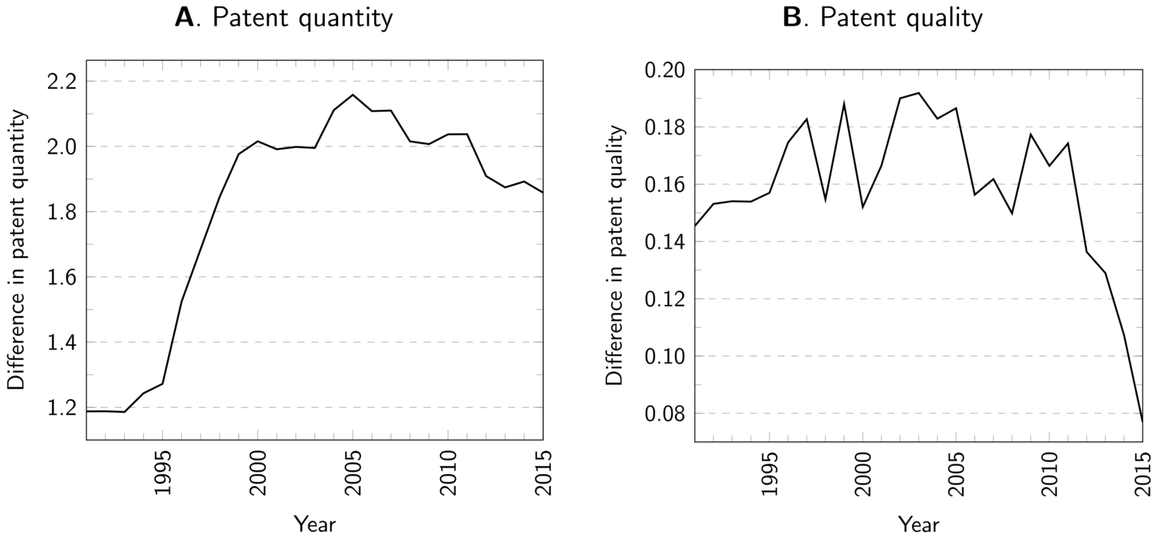

Fig. S10. Differences in patent quantity and patent quality between East and West German municipalites. Differences calculated as West – East.

Table S27. Propensity-score matching results on patent quantity and patent quality, comparing UAS municipalities in East and West Germany

			$PQUAN^{diff}$	$PQUAL^{diff}$
	p	N	(1)	(2)
ATT ($East = 1$)	1996	2,810	-0.542**	0.022
			(0.256)	(0.030)
	1999	3,333	-1.045***	0.016
			(0.327)	(0.031)
	2002	3,333	-1.382**	0.067
			(0.552)	(0.044)
	2005	3,365	-1.197***	-0.013
			(0.454)	(0.028)
	2008	3,365	-1.091**	0.052*
			(0.503)	(0.031)
	2011	3,544	-1.082	0.040
			(0.839)	(0.031)
	2014	3,556	-0.287	0.059*
			(0.443)	(0.028)

Table shows average treatment effects on the treated (ATT). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

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

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

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical image computing and computer-assisted intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, Munich, 2015. Springer.

[50] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[51] W. Li, R. Dong, H. Fu, J. Wang, L. Yu, and P. Gong. Integrating Google Earth imagery with Landsat data to improve 30-m resolution land cover mapping. *Remote Sensing of Environment*, 237:11563, 2020.

[52] F. H. Wagner, A. Sanchez, Y. Tarabalka, R. G. Lotte, M. P. Ferreira, M. P M. Aidar, E. Gloor, O. L. Phillips, and L. E. O. C. Aragão. Using the u-net convolutional neural network to map forest types and disturbance in the atlantic rainforest with very high resolution images. *Remote Sensing in Ecology and Conservation*, 5(4):360–375, 2019.

[53] M. Wang, X. Zhang, X. Niu, F. Wang, and X. Zhang. Scene classification of high-resolution remotely sensed image based on ResNet. *Journal of Geovisualization and Spatial Analysis*, 3:16, 2019.

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

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

[56] 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.

[57] 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.

[58] A. Mellor, S. Boukir, A. Haywood, and S. Jones. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:155–168, 2015.

[59] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112(5):2272–2283, 2008.

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

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

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

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

[64] 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.

[65] 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.

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

[67] 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.

[68] Y. Heymann, C. Steenmans, G. Croisille, and M. Bossard. CORINE land cover: Technical guide. Technical Report EUR 12585 EN, Office for Official Publications of the European Communities, Luxembourg, 1994.

[69] 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.

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

[71] 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.

[72] 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.

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

[74] 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.

[75] 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.

[76] 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.

[77] R. C. M. Beyer, E. Chhabra, V. Galdo, and M. Rama. Measuring districts' monthly economic activity from outer space. Policy Research Working Paper No. 8523, World Bank Group, 2018.

[78] M. Zhao, W. Cheng, C. Zhou, M. Li, N. Wang, and Q. Liu. GDP spatialization and economic differences in south China based on NPP-VIIRS nighttime light imagery. *Remote Sensing*, 9(7):673, 2017.

[79] 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.

[80] M. Pesaresi and P. Politis. GHS-BUILT-S R2022A – GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975–2030). Dataset, European Commission, Joint Research Centre, Brussels, 2022.

[81] M. Pesaresi and P. Politis. GHS-BUILT-V R2022A – GHS built-up volume grids derived from joint assessment of Sentinel2, Landsat, and global DEM data, for 1975–2030 (5yrs interval). Dataset, European Commission, Joint Research Centre, Brussels, 2022.

[82] M. Schiavina, M. Melchiorri, M. Pesaresi, P. Politis, S. Freire, L. Maffenini, P. Florio, D. Ehrlich, K. Goch, P. Tommasi, and T. Kemper. *GHSL data package 2022.* Publications Office of the European Union, Luxembourg, 2022.

[83] 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.

[84] P. Lehnert, C. Pfister, D. Harhoff, and U. Backes-Gellner. Innovation effects and knowledge complementarities in a diverse research landscape. Leading House "Economics of Education" Working Paper No. 164, Swiss Leading House VPET-ECON, Zurich, 2022.

[85] R. Cowan and N. Zinovyeva. University effects on regional innovation. *Research Policy*, 42(3):788–800, 2013.

[86] O. Toivanen and L. Väänänen. Education and invention. *The Review of Economics and Statistics*, 98(2):382–396, 2016.

[87] M. Fritsch and R. Aamoucke. Fields of knowledge in higher education institutions, and innovative start-ups: An empirical investigation. *Papers in Regional Science*, 96(S1):S1–S27, 2017.

[88] C. Pfister, M. Koomen, D. Harhoff, and U. Backes-Gellner. Regional innovation effects of applied research institutions. *Research Policy*, 50(4):104197, 2021.

[89] P. Lehnert, C. Pfister, and U. Backes-Gellner. Employment of R&D personnel after an educational supply shock: Effects of the introduction of Universities of Applied Sciences in Switzerland. *Labour Economics*, 66:101883, 2020.

[90] T. Schlegel, C. Pfister, D. Harhoff, and U. Backes-Gellner. Innovation effects of universities of applied sciences: An assessment of regional heterogeneity. *The Journal of Technology Transfer*, 47:63–118, 2022.

[91] D. Harhoff, F. Narin, F. M. Scherer, and K. Vopel. Citation frequency and the value of patent inventions. *The Review of Economics and Statistics*, 81(3):511–515, 1999.

[92] M. Squicciarini, H. Dernis, and C. Criscuolo. Measuring patent quality: Indicators of technological and economic value. OECD Science, Technology and Industry Working Papers 2013/03, Organisation for Economic Co-operation and Development, Paris, 2013.

[93] M. Andrews. How do institutions of higher education affect local invention? Evidence from the establishment uf U.S. colleges. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3072565, 2020.