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

Patrick Lehnert, Michael Niederberger and Uschi Backes-Gellner
Working Paper No. 165

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

Patrick Lehnert, Michael Niederberger and Uschi Backes-Gellner

March 2020

---


Discussion Papers are intended to make results of the Leading House research or its conferences and workshops promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the Leading House.

The Swiss Leading House on Economics of Education, Firm Behavior and Training Policies is a Research Program of the Swiss State Secretariat for Education, Research, and Innovation (SERI).

www.economics-of-education.ch
Proxying Economic Activity with Daytime Satellite Imagery: Filling Data Gaps Across Time and Space

Patrick Lehnert, Michael Niederberger, and Uschi Backes-Gellner
University of Zurich

March 5, 2020

First Draft. Preliminary Results.

Abstract

This paper develops a novel procedure for proxying economic activity across time periods and spatial units, for which other data is not available. In developing this proxy, we apply machine-learning techniques to a unique historical time series of daytime satellite imagery dating back to 1984. Compared to night lights intensity, a satellite-based proxy that economists commonly use, our proxy has the advantages of more precisely predicting economic activity over a longer time series and at smaller regional levels. We demonstrate the proxy’s usefulness for the example of Germany, where data on economic activity is otherwise unavailable, in particular for the regions belonging to the former German Democratic Republic. However, our procedure is generalizable to other regions and countries alike, and thus yields great potential for analyzing historical developments, evaluating local policy reforms, and controlling for economic activity at highly disaggregated regional levels in econometric applications.

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

* The authors would like to thank Simone Balestra, Dietmar Harhoff, Simon Janssen, Philip Jörg, Daniel Kükenbrink, Edward Lazear, Jens Mohrenweiser, and seminar participants at the University of Zurich for helpful comments. This study is partly funded by the Swiss State Secretariat for Education, Research, and Innovation (SERI) through its Leading House on the Economics of Education, Firm Behavior and Training Policies.
† Corresponding author. Address: Patrick Lehnert, University of Zurich, Plattenstrasse 14, CH-8032 Zurich, Switzerland. E-Mail: patrick.lehnert@business.uzh.ch
Note on Data Availability

In this paper, we produce data containing *surface groups* at various regional levels for Germany. We are happy to share the data we produce with interested researchers upon contacting the authors at patrick.lehnert@business.uzh.ch. Researchers using our data for their research projects must refer to our data and this working paper in their research.
1. Introduction

In this paper, we use daytime satellite imagery to develop a novel proxy for economic activity across time periods and highly disaggregated spatial units, for which other data is unreliable, inaccessible, or entirely inexistent. The proxy thus presents valuable information on economic activity over a uniquely long time series that other data do not cover. Furthermore, the use of daytime satellite imagery (as compared to night lights data) allows us to develop a better proxy for highly disaggregated regional units.

To study, for example, the economic impacts of natural disasters (e.g., Chen et al., 2017; Fu and Gregory, 2019), changes in land use regulations (e.g., de Janvry et al., 2015; Turner et al., 2014), and the impact of climate on economic activity (e.g., Brückner and Ciccone, 2011; Harari and La Ferrara, 2018), economists have started to use direct measures from satellite data in recent years. In addition, economists have increasingly used satellite data as indirect proxies for economic indicators such as Gross Domestic Product (GDP) or human development (e.g., Brüederle and Hodler, 2018; Ebener et al., 2005; Hodler and Raschky, 2014; Watmough et al., 2016). Satellite data have mainly been applied to developing countries, where reliable administrative or survey data is unavailable (e.g., Fetzer et al., 2018; Foster and Rosenzweig, 2003; König et al., 2017; Michalopoulos and Papaioannou, 2014), although existent data on economic activity for developed countries exhibit limited time series and levels of regional detail as well.

So far, night lights data is the most commonly used type of satellite data in economic research (e.g., Barsbai et al., 2017; Bazzi et al., 2016; Dingel et al., 2019; Lee, 2018; Michalopoulos and Papaioannou, 2013; Rohner et al., 2013). Simply put, this data captures the intensity of light sources on earth at night (see, e.g., Huang et al., 2014). This night lights intensity constitutes a valuable proxy for economic activity at the national level and at larger subnational levels such as federal states or metropolitan areas (e.g., Chen and Nordhaus, 2011; Elvidge et al., 1997; Henderson et al., 2012). Furthermore, the technological developments of the 21st century have improved both the accessibility of night lights data and the computational capabilities for processing this data (Donaldson and Storeygard, 2016). Consequently, night lights data has become an attractive data
source for economists in the last decade.

However, night lights are available only for a limited time series (from 1992) and, due to their spatial resolution (one kilometer at the equator), night lights are not reliable for highly disaggregated regional units such as municipalities or suburbs (e.g., Chen and Nordhaus, 2011; Kulkarni et al., 2011; Mellander et al., 2015). Similar problems appear for other data on economic activity (e.g., administrative data). The proxy we develop from daytime satellite imagery covers a longer time series (from 1984) and, due to the imagery’s detailed spatial resolution (30 meters), the proxy reliably represents economic activity at highly disaggregated regional units such as municipalities or even housing blocks.

One example for an application of our proxy is Lehnert et al.’s (2020) analysis of the effect of opening new tertiary education institutions in regions without reliable data on economic activity, such as the regions in East Germany (the former German Democratic Republic). For both East and West German regions, neither night lights nor other reliable data are available for a time series dating back far enough to test for regional economic developments before the openings of the education institutions and to investigate changes in economic activity resulting from these openings in 1985. In addition, night lights would not proxy economic activity at sufficiently detailed regional levels, because the policy effects of opening new education institutions are expected at rather small regional levels, such as municipalities. Furthermore, conventional administrative statistics on economic activity, even if they were available for a sufficiently long time series, do not contain this highly disaggregated regional level. Moreover, accessing regionally disaggregated databases of any country usually presents a challenge for non-resident researchers.

Therefore, analyses of policy reforms that are endogenously determined over time and space—such as the openings of new tertiary education institutions in Germany—can benefit immensely from using publicly available daytime satellite imagery as a proxy for economic activity over a long time series and at highly disaggregated regional units. We develop a novel procedure for computing such a proxy and show that, for Germany, this procedure produces results with a very high internal and external validity. In principle, our procedure is generalizable to any other region in the world, for which various types of
economic policy analyses might require data on economic activity.

Our paper comprises two parts that present our novel measure for proxying economic activity. In the first part, we demonstrate how we derive our measure from daytime satellite imagery by using machine-learning techniques. Furthermore, we assess the measure’s internal validity. In the second part, we analyze the external validity of our measure as a proxy for economic activity. Our measure reliably proxies regional economic activity both over a longer time series and at a more detailed regional level than other commonly used proxies.

In the first part of the paper, we explain how we derive our measure from Landsat satellite data, which has three advantages over night lights or other satellite data. First, the National Aeronautics and Space Administration (NASA) launched the first satellite of the Landsat program (Landsat-1) in 1972, making Landsat the earliest existing source of satellite data (e.g., Morain, 1998; Williams et al., 2006). Although Landsat data thus does exist from 1972, the data does not cover Germany until 1984. Therefore, we begin the time series of our analysis in 1984. Our measure goes further back than night lights data (which is available only from 1992) or other satellite data (Donaldson and Storeygard, 2016). Therefore, Landsat data allows researchers to conduct analyses over a longer time series, enabling them to study earlier events.

Second, Landsat satellites collect remotely sensed multispectral imagery of the earth, that is, they observe the energy that the earth reflects in different spectral bands (e.g., infrared or ultraviolet) (Donaldson and Storeygard, 2016; Williams et al., 2006). While Goldblatt et al. (2019) show that in Vietnamese regions the raw spectral data of Landsat-7 can help in slightly improving the prediction of economic activity as compared to a night lights-based prediction, our study shows that processing the spectral data of Landsat satellites with our machine-learning techniques better maps different types of land cover over time and space and thus provides a substantially improved proxy for economic activity.

---

1 For other countries, such as the U.S., Landsat data cover earlier years and thus potentially allow for an extended time series.

2 Although night lights has been recorded since the 1970s, a digital archive exists only from 1992, with earlier years available as a (barely processable) film archive (Donaldson and Storeygard, 2016; Elvidge et al., 2003).
Based on a survey of the literature that applies land cover classifications (e.g., Han et al., 2004; Pérez-Hoyos et al., 2012; Yu et al., 2011), we identify six types of land cover, which we refer to as six surface groups: built-up surfaces, grassy surfaces, forest-covered surfaces, surfaces with crop fields, surfaces without vegetation, and water surfaces. As some surface groups serve economic activity more than others (e.g., Keola et al., 2015; Sutton and Costanza, 2002; Yu et al., 2011), mapping surface groups can yield important information for predicting economic activity. For example, increases in built-up surfaces, which include agglomerations of cities or transportation networks, coincide with increases in economic activity (e.g., Davis et al., 2014; Holl, 2004). Landsat data thus provide a more nuanced proxy for economic activity for use in policy analyses than, for example, a single indicator of night lights intensity.

Third, Landsat data has a substantially higher resolution (30 meters) than night lights data resolution (one kilometer) (Donaldson and Storeygard, 2016). This higher resolution entails more precise information at a much more disaggregated regional level. Consequently, Landsat data makes possible economic analyses at even more detailed regional levels than night lights data and, in turn, allows assessments of the effect of policy changes in much smaller localities.

We compute the surface groups by applying a machine-learning algorithm in Google Earth Engine (GEE), a cloud-based platform that hosts an extensive catalog of satellite data and provides tools for processing and analyzing these data (see Gorelick et al., 2017). Through GEE, Landsat data (and many other geospatial data) have become available to the scientific community, with broadband Internet access as the sole requirement for using GEE (Wulder and Coops, 2014), that is, with no need for extensive data storage capacities, expensive software licenses, or administrative application procedures (e.g., restricted access for non-residents). However, due to the lack of comparable tools, Landsat data has so far received much less attention than night lights data in economic research. The few existing

---

3 Landsat-1, Landsat-2, and Landsat-3 data have a resolution of 79 meters and thus still a far higher resolution than night lights data (Williams et al., 2006).

4 In addition, the U.S. Geologic Survey (USGS), which has been running the Landsat program since the 1990s, made the Landsat archives publicly available free of charge only in 2008 (Morain, 1998; Williams et al., 2006; Woodcock et al., 2008; Wulder and Coops, 2014).
applications rely mainly on visual interpretation for identifying, for example, agricultural land use, (de)forestation, or urbanization (e.g., Burchfield et al., 2006; Deng et al., 2008; Foster and Rosenzweig, 2003; Mertens et al., 2002). In our approach, we draw from the geographic 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., Dewan and Yamaguchi, 2009; Goldblatt et al., 2018; Liu et al., 2018). For our economic application, we extend this literature by developing a procedure that combines all Landsat data available from 1984 to map six different surface groups. This procedure can be used or adapted by other researchers for their own applications and research purposes.

In the second part of the paper, we assess the value of the surface groups for economic research using three reliable and regionally detailed indicators of economic activity as validation data. First, since 2000 the German Federal Statistical Office (GFSO) has provided administrative data on GDP, the indicator of economic activity most frequently used in the literature evaluating the usefulness of satellite-based measures for economic research (e.g., Chen and Nordhaus, 2011; Doll et al., 2006; Henderson et al., 2012), at the level of small administrative regional units (counties) since the 2000s. We can use this data to test the external validity of our proxy. This proxy then allows us to determine economic activity for time series and regional units for which no (reliable) administrative or other data are available. Second, data on household income from Leibniz Institute for Economic Research (RWI) and Micromarketing-Systeme und Consult GmbH (microm) (2019) make possible an additional analysis of the validity of surface groups as a proxy for economic activity at a regional micro-level representing areas as small as housing blocks, and thus also provide a very high level of regional detail. Third, to assess the value of the surface groups for economic research analyzing, for example, urbanization or agriculture, we use regionally detailed administrative data on land cover from the GFSO. In sum, this rich set of validation data allows us to extensively investigate the value of the surface groups for economic research. In so doing, we also contribute to the literature on the validation of new satellite-based measures of economic activity.
To test the external validity of our proxy, we regress the indicators of regional economic activity from the validation data on the surface groups using Ordinary Least Squares (OLS). To assess the quality of our proxy, we investigate potential biases in the OLS predictions by examining the distribution of the regression residuals. In so doing, we analyze both the overall distribution of the residuals and their distribution over time and space. To get a better grasp of the value of the surface groups for economic research, we also compare the surface groups-based prediction to the night lights-based prediction in our analysis of the prediction bias.

Our analyses show that the combination of surface groups within a region constitutes a valid proxy for economic activity. At the levels of both small administrative units and regional micro-levels, the regional combination of surface groups predicts regional economic activity more accurately (i.e., with a smaller prediction error and a smaller bias) than, for example, night lights intensity. The surface groups that we identify from Landsat data thus constitute a valuable measure for various applications in economic research, particularly for assessing policy changes affecting small regional units and for retrospectively analyzing long time series.

The paper proceeds as follows. Section 2 presents our computations of the surface groups in GEE. We outline the specifications of the machine-learning algorithm, describe the data it uses, and evaluate the algorithm’s performance. Section 3 assesses the validity of the surface groups as a novel six-dimensional proxy for economic activity and as separate proxies for different types of land cover. Section 4 concludes.

2. Detection of Surface Groups

2.1. 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., Goldblatt et al., 2016; Liu et al., 2018; Schneider, 2012). 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 this 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 Figure 1 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. As we choose those observations that best differentiate between vegetated and unvegetated areas for this composite data, we refer to it as “greenest” pixel

---

5 For more information on GEE, see Gorelick et al. (2017) or GEE’s documentation website (https://developers.google.com/earth-engine/, last retrieved on November 11, 2019)
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)\(^6\) data as an external source of ground-truth information. This data, which comes from a pan-European project commissioned by the European Environment Agency (EEA), maps land cover in European countries for five reference years (1990, 2000, 2006, 2012, and 2018). Based on a survey of the literature that applies land cover classifications (e.g., Gallego and Bamps, 2008; Waser and Schwarz, 2006; Yu et al., 2011), we obtain from the CLC data six different types of land cover, which we refer to as six \textit{surface groups}: built-up surfaces (\textit{builtup}), grassy surfaces (\textit{grass}), forest-covered surfaces (\textit{forest}), surfaces with crop fields (\textit{crops}), surfaces without vegetation (\textit{noveg}), and water surfaces (\textit{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 this 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 contains the surface group of every Landsat pixel location annually for the period 1984 through 2018. To assess

\footnote{The acronym “CORINE” stands for “coordination of information on the environment” (Büttner et al., 2002).}
Figure 2: Visual comparison of data sources


Source: Authors’ illustrations using Google satellite imagery, Landsat data, and CLC data.
Notes: Pictures a–d show the same approx. 84 × 56 square kilometers area with the metropolitan region of Nuremberg in the center. Pictures e–h show the same approx. 1.3 × 0.9 square kilometers area with the village of Muhr am See in the north and the Altmühlsee (lake) in the south (area framed in red in pictures a–d).
the accuracy of the classification (i.e., the internal validity), we perform five-fold cross-validation. We find that our procedure accurately classifies pixels into the surface groups.

Figure 2 illustrates the data sources we use and the output data we produce. Figures 2a–d show a large-scale sample area with the metropolitan region of Nuremberg in the center and Figures 2e–h show a small-scale sample area with the village of Muhr am See in the north and the Altmühlsee (lake) in the south (the area framed in red in figures 2a–d). As a reference, Figures 2a and 2e show current high-resolution satellite images from GEE. Figures 2b and 2f, which use Landsat’s visible spectral bands to approximate the perception of the human eye, shows the greenest pixel composite of Landsat imagery for 2018 (the input data). Figures 2c and 2g illustrates the six different types of land cover we identify from the CLC data for the same year (the ground-truth data). Figures 2d and 2h show the surface group that our classification algorithm produces for every Landsat pixel location in 2018 (the output data). For better comprehensibility of our procedure for detecting surface groups, we will refer to Figure 2 in this section as appropriate.

2.2. Greenest Pixel Composite of Landsat Data as Input Data

Satellite data from the Landsat program functions 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 (Lauer et al., 1997; Wulder et al., 2012, 2008, 2016). Throughout the history of Landsat, the various operating agencies have launched eight satellites, one of which (Landsat-6) failed to reach orbit (Morain, 1998; Williams et al., 2006; Wulder et al., 2016). As of 2020, Landsat-7 and Landsat-8 remain active, with the launch of Landsat-9 scheduled for December 2020 (McCorkel et al., 2018; Wulder et al., 2016).

For descriptions of the Landsat program’s legal and political history, including the switch from public to private operators and back again, see, e.g., Goward et al. (2006); Green (2006); Lauer et al. (1997); Morain (1998); Williams et al. (2006); Wulder et al. (2012). The remote-sensing literature and related disciplines have applied Landsat data for numerous purposes. For example, Lyzenga (1981) assesses water conditions in the Bahamas; Yacobi et al. (1995) examine plankton species in Lake Kinneret; Sobrino et al. (2004) measure surface temperature in Spain; and Torresani et al. (2019) investigate tree species diversity in the Alps.
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 (Markham et al., 2004). 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 (Markham and Helder, 2012). Table 1 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.9

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 1). 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 (Rumerman, 1999) 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 (Morain, 1998). 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 that time, imagery in the Landsat archives is scarce for Germany until the 1980s (Wulder et al., 2016). 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

---

9 Table 1 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 (Loveland and Dwyer, 2012; Wulder et al., 2016). 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 (Roy et al., 2014).
# Table 1: Technical specifications of Landsat sensors

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Multispectral Scanner (MSS)</th>
<th>Thematic Mapper (TM)</th>
<th>Enhanced Thematic Mapper Plus (ETM+)</th>
<th>Operational Land Imager (OLI) / Thermal Infrared Sensor (TIRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial resolution</td>
<td>79 meters</td>
<td>30 meters</td>
<td>30 meters</td>
<td>30 meters</td>
</tr>
<tr>
<td></td>
<td>Landsat-3 (1978–1983)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band name</td>
<td>Wavelength (in μm)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ultra blue</td>
<td></td>
<td>0.43–0.45</td>
<td>0.43–0.45</td>
<td></td>
</tr>
<tr>
<td>Visible blue (BLUE)</td>
<td></td>
<td>0.45–0.52</td>
<td>0.45–0.52</td>
<td>0.45–0.51</td>
</tr>
<tr>
<td>Visible green (GREEN)</td>
<td>0.50–0.60</td>
<td>0.52–0.60</td>
<td>0.52–0.60</td>
<td>0.53–0.59</td>
</tr>
<tr>
<td>Visible red (RED)</td>
<td>0.60–0.70</td>
<td>0.63–0.69</td>
<td>0.63–0.69</td>
<td>0.64–0.67</td>
</tr>
<tr>
<td>Short-wave infrared 1 (SWIR1)</td>
<td>1.55–1.75</td>
<td>1.55–1.75</td>
<td>1.57–1.65</td>
<td>1.57–1.65</td>
</tr>
<tr>
<td>Short-wave infrared 2 (SWIR2)</td>
<td>2.08–2.35</td>
<td>2.08–2.35</td>
<td>2.11–2.29</td>
<td></td>
</tr>
<tr>
<td>Near-infrared 1 (NIR)</td>
<td>0.70–0.80</td>
<td>0.76–0.90</td>
<td>0.77–0.90</td>
<td>0.85–0.88</td>
</tr>
<tr>
<td>Near-infrared 2</td>
<td>0.80–1.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermal infrared 1</td>
<td>10.40–12.50 (120-meter resolution)</td>
<td>10.40–12.50 (60-meter resolution)</td>
<td>10.60–11.19 (100-meter resolution)</td>
<td></td>
</tr>
<tr>
<td>Thermal infrared 2</td>
<td>11.50–12.51 (100-meter resolution)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panchromatic</td>
<td>0.52–0.90 (15-meter resolution)</td>
<td>0.50–0.68 (15-meter resolution)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cirrus</td>
<td></td>
<td></td>
<td></td>
<td>1.36–1.38</td>
</tr>
</tbody>
</table>

Source: Authors’ representation based on Goldblatt et al. (2016); Loveland and Dwyer (2012); Markham et al. (2004); Roy et al. (2014); Williams et al. (2006); Wulder et al. (2016).

Note: Spectral bands used for detecting surface groups highlighted gray.
thermal infrared spectral bands, because their technical specifications change over time and differ from the remaining bands (e.g., coarser spatial resolution, different number of bands; see table 1). 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 lights data that economists commonly use (Donaldson and Storeygard, 2016), we compute the surface groups annually. As Landsat satellites record a geographic location on earth once every 16 days (Loveland and Dwyer, 2012), one satellite ideally gives us 23 images of a pixel location per year. During years with two active satellites (1984–1993 and 1999–present), we can thus use a maximum of 46 images per pixel and year. Unfortunately, pre-processed annual composites incorporating imagery from multiple Landsat satellites do not exist, requiring us to produce such composites from the available images and use these composites as input data for our algorithm.

We produce pixel-based annual composites of Landsat images. Among all available observations of a given pixel within a year, we choose the one pixel that best serves the purpose of detecting surface groups. This pixel-based compositing procedure (as compared to scene-based compositing) prevents a loss of information due to, for example, cloud-covered pixels and enables the researcher to choose those pixels best suitable for a specific application—in our case the detection of surface groups (Griffiths et al., 2013). 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 (Trianni et al., 2015).

For both the compositing and the actual pixel classification (see section 2.4), we follow studies from the remote-sensing literature (e.g., Goldblatt et al., 2018, 2016; Liu et al., 2018) and add three indices to the data: First, the Normalized Difference Vegetation Index (NDVI), which dates to the work of Rouse Jr et al. (1973), differentiates vegetated from unvegetated surfaces and is one of the most frequently used indices in remote-sensing.

---

10 Technically, the Landsat data consists of images (also called scenes) with a size of approximately 185 × 185 square kilometers (Wulder et al., 2016). One image therefore comprises several million pixels sized 30 × 30 square meters at the equator.
applications (Townshend and Justice, 1986; Xue and Su, 2017); Second, the Normalized Difference Water Index (NDWI) differentiates open water from other surfaces (McFeeters, 1996, 2013); \(^{11}\) Third, the Normalized Difference Built-up Index (NDBI) differentiates built-up surfaces from other surfaces (Zha et al., 2003). Similar to Liu et al. (2018), we compute these three indices for Landsat data as follows:

\[
NDVI_p = \frac{NIR_p - RED_p}{NIR_p + RED_p}
\]  

(1)

\[
NDWI_p = \frac{GREEN_p - NIR_p}{GREEN_p + NIR_p}
\]  

(2)

\[
NDBI_p = \frac{SWIR1_p - NIR_p}{SWIR1_p + NIR_p}
\]  

(3)

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. 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., Zhu and Woodcock, 2012). 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 lights as a proxy for GDP—we expect economic activity to concentrate in urban or industrial areas. Therefore, a clear distinction between built-up surfaces and other surfaces will improve our proxy for economic activity.

Our procedure of compositing Landsat data provides us with a greenest pixel composite that we can use as input data for the machine-learning algorithm. This composite covers the geographical area of Germany and consists of one observation per pixel for every year.

\(^{11}\) Gao (1996) developed another index under the name “NDWI” to identify liquid water inside plants. The computation of Gao’s (1996) NDWI relies on different spectral bands than McFeeters’ (1996).
since 1984. The variables in the dataset are the pixel’s values in the six spectral bands we use in this paper (see table 1) 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 contains missing values. Figures 2b and 2f visualize the greenest pixel composite with the visible spectral bands BLUE, GREEN, and RED for 2018.

2.3. 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 the CLC data in 1985, with the goal of creating a standardized database on land cover to support policymakers in environmental affairs (Büttner et al., 2002; EEA, 2017). 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) (Büttner and Kosztra, 2017). Each database includes a map for the respective year with a pixel resolution of 100 meters, indicating land cover in a variety of classes (Büttner and Kosztra, 2017; Kosztra et al., 2019).

Although the medium underlying the classification changed over the years from hard-copies to computer-assisted technologies, classification still relies mainly on visual interpretation of satellite imagery (Büttner and Kosztra, 2017; EEA, 2007; Kosztra et al., 2019). This imagery stems from various satellites, including Landsat satellites for CLC1990, CLC2000, and CLC2018 (Büttner and Kosztra, 2017). The remote-sensing literature provides successful combinations of CLC data and Landsat data in geospatial analyses (e.g., Matejicek and Kopackova, 2010; Pekkarinen et al., 2009; Sánchez et al., 2008; Stathopoulou et al., 2007).

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 2.2). 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 lights data with a resolution of one kilometer in Goldblatt et al., 2018). 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 provides 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 constitutes 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 (i.e., the land’s directly observable terrestrial features) and land use (i.e., the land’s socioeconomic purpose) (Cihlar and Jansen, 2001; Comber et al., 2008; Feranec et al., 2007; Fisher et al., 2005). Given that automated analyses of satellite data can detect only land cover, and that determining land use requires manual interpretation (Fisher et al., 2005), 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., Balzter et al., 2015; Han et al., 2004; Yu et al., 2011). However, as this literature does not provide an unambiguous assignment of CLC classes to larger groups with similar surface characteristics, we perform repeated trials of our classification procedure with varying assignments of CLC level 3 classes to larger groups. These trials yield the result that a classification consisting of six surface groups, which correspond to Yu et al.’s (2011) six types of surfaces identified from subsets of Landsat data for the Daqing region in China, best represent similar surface characteristics in Germany:

- Built-up surfaces (builtup): surfaces with buildings of non-natural materials such
as concrete, metal, and glass (e.g., residential buildings, industrial plants, roads) (following, e.g., Neumann et al., 2007).

- Grassy surfaces (grass): surfaces covered by grass or other plants with similar surface reflectance (e.g., natural grassland) (following, e.g., Neumann et al., 2007; Waser and Schwarz, 2006).

- Forest-covered surfaces (forest): surfaces covered by trees or other plants with similar surface reflectance (e.g., mixed forests) (following, e.g., Neumann et al., 2007; Pérez-Hoyos et al., 2012).

- Surfaces with crop fields (crops): surfaces with vegetation for agricultural purposes (e.g., hayfields).

- Surfaces without vegetation (noveg): surfaces with (almost) no vegetation or buildings (e.g., bare rock).

- Water surfaces (water): any type of water surface (e.g., lakes) (following, e.g., Gallego and Bamps, 2008).

These six surface groups into which the classification algorithm divides the input data constitute the basis for our proxy for economic activity. Figures 2c and 2g visualize the ground-truth surface groups that we obtain from the CLC2018 data.

### 2.4. 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, forest, crops, 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 Random Forest (RF) algorithm with ten decision trees.

Following Goldblatt et al. (2016), 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 object-based classification, which additionally considers information from neighboring pixels, pixel-based classification requires considerably less computational power (Myint
et al., 2011; Whiteside et al., 2011). Although the majority of studies in the remote-
sensing literature suggest that object-based classification performs better than pixel-based
classification (e.g., Jebur et al., 2014; Whiteside et al., 2011), some studies find no significant
performance difference (e.g., Berhane et al., 2018; Duro et al., 2012), and in particular,
Dingle Robertson and King (2011) find no significant difference using Landsat data.\textsuperscript{12}
Therefore, given the spatial and temporal size of the data we analyze in this paper, pixel-
based classification is the preferable choice. Our assessments in Section 3 confirm that
choosing this computationally less intensive classification yields a valid proxy for economic
activity.

To classify the pre-processed Landsat data, we use the RF algorithm with ten decision
trees.\textsuperscript{13} Several studies in the remote-sensing literature find that RF outperforms other
algorithms when applied to land cover classification (e.g., Gislason et al., 2006; Ok et al.,
2012; Rodriguez-Galiano et al., 2012). For example, Goldblatt et al. (2016), who predict
built-up land cover in India with Landsat-7 and Landsat-8 data, assess the performance of
three different algorithms that the remote-sensing literature commonly uses (Classification
and Regression Tree, Support Vector Machines, and RF) and find that RF performs best.
Furthermore, the RF algorithm requires less computational power (Gislason et al., 2006;
Pal, 2005). As to the number of decision trees, Goldblatt et al. (2016) conclude that
performance increases with the number of trees, although after ten trees the increase is
negligibly small relative to the increase in computational power required.\textsuperscript{14} Therefore, RF
with ten decision trees best suits our purpose.

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 the CLC data covers (1990, 2000, 2006, 2012,
and 2018), we randomly choose 1,000 pixels of each of the six surface groups. Generally,

\textsuperscript{12} See, e.g., Goldblatt et al. (2016) for a short summary of the literature on the advantages and
disadvantages of pixel-based vs. object-based classification, and see, e.g., Ma et al. (2017) for an
extensive review.

\textsuperscript{13} For a general description of the RF method, see, e.g., Breiman (2001). For a description of the method’s
application for land cover classification, see, e.g., Gislason et al. (2006), and for a description of the
method’s application in economics, see, e.g., Athey and Imbens (2019).

\textsuperscript{14} For example, while the overall accuracy of Goldblatt et al.’s (2016) prediction increases by approx. two
percentage points when increasing the number of trees from three to ten, the overall accuracy increases
only by approx. one more percentage point when increasing the number of trees from ten to 100.
the number of pixels in the training data correlates positively with prediction accuracy but negatively with computational effort (Millard and Richardson, 2015; Rodriguez-Galiano et al., 2012). Therefore, we choose a slightly larger number of pixels in the training data than in comparable applications from the remote-sensing literature (e.g., Goldblatt et al., 2016; Schneider, 2012) to achieve an accurate classification, but keep the number of pixels low enough to maintain a reasonable computational effort. Furthermore, to account for the lower spatial resolution of the CLC data, we do not use Landsat pixels that fall within CLC pixels at the boundary of two CLC surface areas.

2.5. Accuracy Assessment of Output Data

To assess the prediction accuracy of our classification in the output data, we follow Goldblatt et al. (2016) and perform five-fold cross-validation 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 and year, we now draw only 250. Thus each of the five subsets consists of 7,500 pixels, that is, 250 per surface group and year. For the cross-validation to be valid, the subsets must not overlap. In other words, one pixel can belong to only one subset.

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

As indicators of prediction accuracy, for every iteration and for each of the six surface groups separately, we calculate overall accuracy (OA), true-positive rate (TPR), true-negative rate (TNR), balanced accuracy (BA), and user’s accuracy (UA) (see, e.g.,

15 For descriptions and discussions of this method see, e.g., Arlot and Celisse (2010); Rodríguez et al. (2010); Wong (2015).
Goldblatt et al., 2018, for the formulae to calculate these indicators). OA denotes the percentage of pixels correctly classified. $TPR^{16}$ denotes the percentage of pixels correctly classified as belonging to the respective surface group. $TNR$ indicates the percentage of pixels correctly classified as not belonging to the respective surface group. $BA$ indicates the average of $TPR$ and $TNR$. $UA$ indicates the percentage of pixels correctly classified as belonging to the respective surface group among all pixels classified as belonging to the respective surface group.

Table 2: Five-fold cross-validation results

<table>
<thead>
<tr>
<th>Surface group</th>
<th>Overall accuracy (OA)</th>
<th>True-positive rate (TPR)</th>
<th>True-negative rate (TNR)</th>
<th>Balanced accuracy (BA)</th>
<th>User’s accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>builtup</td>
<td>0.852</td>
<td>0.625</td>
<td>0.898</td>
<td>0.762</td>
<td>0.553</td>
</tr>
<tr>
<td>grass</td>
<td>0.835</td>
<td>0.546</td>
<td>0.893</td>
<td>0.719</td>
<td>0.505</td>
</tr>
<tr>
<td>forest</td>
<td>0.897</td>
<td>0.734</td>
<td>0.930</td>
<td>0.832</td>
<td>0.684</td>
</tr>
<tr>
<td>crops</td>
<td>0.849</td>
<td>0.501</td>
<td>0.919</td>
<td>0.710</td>
<td>0.562</td>
</tr>
<tr>
<td>noveg</td>
<td>0.907</td>
<td>0.697</td>
<td>0.949</td>
<td>0.823</td>
<td>0.733</td>
</tr>
<tr>
<td>water</td>
<td>0.912</td>
<td>0.653</td>
<td>0.964</td>
<td>0.808</td>
<td>0.783</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: Indicators calculated with respect to each surface group. Values indicate the average over all five iterations and all five CLC reference years.

Table 2 shows the results of the five-fold cross-validation with respect to each of the six surface groups. With 85.2 percent, our OA score for builtup is similar to that in other studies detecting built-up land with Landsat data (e.g., Dewan and Yamaguchi, 2009; Goldblatt et al., 2018, 2016). The other indicators—$TPR$, $TNR$, $BA$, and $UA$—are also in line with other studies (e.g., Goldblatt et al., 2018, 2016). Furthermore, we achieve very high OA scores for forest (89.7 percent), noveg (90.7 percent) and water (91.2 percent), and good OA scores for grass (83.5 percent) and crops (84.9 percent). Appendix A shows the five-fold cross-validation results separately for every CLC year.

The five-fold cross-validation results show that the output data we produce constitute

16 $TPR$ is also known as producer’s accuracy in the remote-sensing literature.
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 are highly suitable for analyzing whether the surface groups are an externally valid proxy for economic activity in Section 3.

Finally, the output of our procedure for detecting surface groups is a dataset containing the surface group of every Landsat pixel location in Germany annually from 1984 through 2018. One year comprises more than 630 million Landsat pixels, amounting to more than 22 billion pixel-year observations in the output data. Of these observations, 15.0 percent are classified as *builtup*; 20.8 percent as *grass*; 24.9 percent as *forest*; 29.1 percent as *crops*; 4.1 percent as *noveg*; and 3.8 percent as *water*. Only 2.3 percent of observations contain missing values due to, for example, cloud cover. We can aggregate the pixel-level information to any geographical unit (e.g., administrative regional units or grid cells) within GEE by uploading a polygon shapefile (i.e., a file containing geospatial information on regional borders) and calculating the number of pixels per surface group within a predefined area.

### 3. Validation of Surface Groups for Applications in Economics

#### 3.1. Overview

After computing our measure of surface groups and assessing the measure’s prediction accuracy (i.e., internal validity) in Section 2, we investigate the measure’s external validity as a novel six-dimensional proxy for economic activity in this section. The purpose of the surface groups is to proxy 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

---

17 For reference, Germany comprises 355,888 square kilometers excluding water areas (calculated from geodata provided by the Federal Agency for Cartography and Geodesy (BKG)).
analyze the quality of a surface groups-based prediction of economic activity.

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 satellite-based proxies for economic activity (e.g., Chen and Nordhaus, 2019; Henderson et al., 2012; Sutton et al., 2007). The administrative GDP measure is available at the county (Kreis) level\(^{18}\) since 2000. Second, the socioeconomic dataset RWI-GEO-GRID (RWI and microm, 2019) provides household income as a further indicator of economic activity with a very high level of regional detail. The household income measure is available at the level of grid cells sized one square kilometer (and thus independent of administrative borders) for 2005 and since 2009.

To evaluate the surface groups-based prediction of economic activity, we perform 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 bias in the prediction of economic activity. Throughout this evaluation, we compare the surface groups-based prediction of economic activity to the night lights-based prediction. The night lights-based prediction thus serves as a benchmark for assessing the quality of the surface-groups based prediction.

In a further analysis, we investigate the quality of surface groups as proxies for their corresponding types of land cover as indicated in administrative statistics. This analysis complements the assessment of internal validity in Section 2.5 and evaluates whether the surface groups are a useful measure for studies examining different types of land cover (e.g., urbanization, deforestation). The administrative measures of land cover are available at the municipality (Gemeinde) level\(^{19}\) from 2008 through 2015. We perform separate OLS regressions for each surface group on the corresponding type of administrative land cover and examine the regression residuals.

By using external validation data that are available for limited time series, the analyses

\(^{18}\) As of January 1, 2017, Germany comprised 401 counties.

\(^{19}\) As of January 1, 2017, Germany comprised 11,266 municipalities.
in this section provide insight into the quality of the surface groups as a measure for applications in economic research. Section 3.2 describes the external data we use for these analyses in more detail. Section 3.3 presents the analysis of surface groups as a novel six-dimensional proxy for economic activity and Section 3.4 provides the analysis of surface groups as separate proxies for land cover as indicated in administrative statistics.

3.2. Validation Data

To obtain economic indicators at detailed regional levels, we draw on two data sources. First, we use administrative regional data. We access this data via “Regionaldatenbank Deutschland”; a database belonging to the GFSO’s data portal, GENESIS. This database comprises a variety of regional statistics from the GFSO and the statistical offices of the federal states (Bundesländer), with varying time series and levels of regional disaggregation. We extract two measures for our analyses from GENESIS, GDP (available from 2000), and land cover (available from 2008 through 2015). Second, we use RWI-GEO-GRID (RWI and microm, 2019), a grid-level dataset containing socioeconomic indicators collected from a variety of public and private sources (for a more detailed description of this dataset, see Breidenbach and Eilers, 2018). From this dataset, we extract a measure of household income (available for 2005 and from 2009) that allows us to analyze economic activity at a regional level even more detailed than the administrative county level.

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 2016. Following Henderson et al. (2012), we use real (i.e., deflated) GDP measured in euros as a validation measure for our analyses. We denote real GDP as \(GDP\).

Household income in RWI-GEO-GRID is available at the level of grid cells sized one square kilometer, an extremely high level of regional detail. The grid cells in this dataset

---

20 https://www.regionalstatistik.de/genesis/online/ (last retrieved on August 21, 2019).
21 The acronym “GENESIS” stands for “Gemeinsames Nemes Statistisches Informations-System”. See https://www.statistikportal.de/de/datenbanken (last retrieved on August 31, 2019).
22 Germany is divided into 16 federal states.
23 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 (last retrieved on November 6, 2019).
follow the system of the European Reference Grid distributed by the European Soil Data Centre (ESDAC)\(^{24}\) (Breidenbach and Eilers, 2018). 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 (Breidenbach and Eilers, 2018). 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 approximately 360,000 grid cells, more than 150,000 of which contain positive values of household income. We denote real household income as \(HHI\).

In addition, to compare the quality of the prediction that uses surface groups to the prediction that uses night lights, we use night lights data from the U.S. Air Force Defense Meteorological Satellite Program Operational Linescan System (DMSP OLS), available from 1992 through 2013. Like Landsat data, the night lights data is accessible through GEE. Similar to Henderson et al. (2012), we use the pre-processed version of this data (i.e., the version corrected for, for example, clouds or unusual lighting such as forest fires). This version contains one observation per pixel and year, indicating the intensity of lights on earth at night.\(^{25}\) The intensity variable ranges between 0 and 63. To achieve regional correspondence with the administrative GDP data and RWI-GEO-GRID, we calculate the average night lights intensity at the county and at the grid level (denoted as \(NL\)).

For evaluating the surface groups as proxies for their corresponding types of land cover, we derive such corresponding measures of land cover from administrative land use statistics. These statistics report how much of a region’s area serves a specific land use purpose. The categories of land use are aggregated versions of the categories in the official real estate register that indicates land use within administrative regions (see Working Committee of the Surveying Authorities of the Laender of the Federal Republic of Germany


\(^{25}\) For a few observation years, two satellites collected night lights intensity. Consequently, the night lights data contains two observations per pixel for these years. Following Henderson et al. (2012), we use the average of those observations for our calculations.

26
As with the CLC classes (see section 2.3), the administrative categories indicate either land use alone or a mixture of land use and land cover. Unfortunately, at their level of aggregation, the administrative data does not allow us to identify all subcategories representing a given surface group in the real estate register. Consequently, the administrative measures we obtain constitute lower bounds of their actual values. Therefore, to obtain corresponding measures of land cover, we sum up those categories that unambiguously indicate land uses belonging to the six surface groups. We thus retrieve six administrative measures of land cover, denoted as $\text{builtup}_{\text{adm}}$, $\text{grass}_{\text{adm}}$, $\text{forest}_{\text{adm}}$, $\text{crops}_{\text{adm}}$, $\text{noveg}_{\text{adm}}$, and $\text{water}_{\text{adm}}$ (measured in square kilometers).

The administrative data indicate land use at the municipality ($\text{Gemeinde}$) level, the smallest administrative regional unit in Germany, for the time period 2008–2015. As Germany experienced a number of territorial reforms over the last two decades and the territorial status underlying the administrative land use information varies between years, we need to update this information to a common territorial status. Because the territorial reforms mainly involved municipal mergers, updating land use primarily entails adding the information from formerly separated municipalities that merged, thereby preventing a loss of measurement precision. To determine each municipality’s territorial status as of January 1, 2017, we follow Egger et al. (2017) and use the BKG’s historical archive of polygon shapefiles, data files that contain geospatial information on municipal borders for the relevant time period. For municipalities belonging to the federal state of Lower Saxony, inconsistencies in the region identifier prevent us from reliably assigning the administrative data to the correct geographic location. Therefore, we drop these municipalities from

---

26 For example, the higher level of aggregation in our data does not allow us to distinguish public and historical buildings from other types of non-built-up land cover.

27 Since 2016, the land use categories follow a definition different from that in AdV (1991). The new definition no longer allows us to consistently identify the six types of land cover corresponding to the surface groups.

28 The territorial reforms did not affect the county level, at which the administrative data provides GDP.

29 The number of municipalities diminished from 12,485 on January 1, 2008 to 11,266 on January 1, 2017.

30 In the rare cases of municipal separations, we divide the original land use information relative to the areas of the new municipalities.

31 We chose the territorial status of January 1, 2017, because it was the most current one when we started work on this paper.

32 Available from http://www.geodatenzentrum.de/download/archiv/vektor/vg250_ebenen/ (last retrieved on April 17, 2019).
Apart from this exclusion, we are able to calculate $\text{builtup}_{\text{adm}}$, $\text{grass}_{\text{adm}}$, $\text{forest}_{\text{adm}}$, $\text{crops}_{\text{adm}}$, $\text{noveg}_{\text{adm}}$, and $\text{water}_{\text{adm}}$ annually for all municipalities.

To assess the predictive value of surface groups we derive from Landsat data, we need to aggregate the pixel-level surface group information to the different regional units of the validation data. We perform this aggregation directly in GEE by uploading the polygon shapefile indicating the regional borders of the validation data (county borders, municipality borders, and grid cell borders), and then counting the number of pixels in each surface group per regional unit and year. This procedure generates, at the respective regional level, six variables indicating the number of pixels per surface group: $\text{builtup}_{\text{sat}}$, $\text{grass}_{\text{sat}}$, $\text{forest}_{\text{sat}}$, $\text{crops}_{\text{sat}}$, $\text{noveg}_{\text{sat}}$, and $\text{water}_{\text{sat}}$. 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.34

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 and as separate proxies for land cover. We argue that if the quality of the surface groups-based prediction is high in the years that the validation data cover, we can safely conclude that 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 1984–2018). 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, 1984–2004 for household income, and 1984–2007 for land cover).

33 Apart from a reduction in sample size, we do not expect that dropping municipalities in the federal state of Lower Saxony affects our results, because the remaining parts of Germany still exhibit much variation in land cover.

34 Other reasons for missing values could be implausible values in the spectral information or inexistence of imagery (see section 2.2). However, cloud cover is the most likely reason.
3.3. Validation of Surface Groups as a Proxy for Economic Activity

To assess the validity of surface groups as a proxy for economic activity and to compare their prediction to compare them to night lights—which are a widely accepted proxy in economic research—we perform OLS regressions of the following form:

\[ Y_{i,t} = \beta_0 + \beta_1 X_{i,t} + \beta_2 C_{i,t} + \nu_{i,t} \]  

(4)

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 \( \ln(builtup_{sat} + 1) \), \( \ln(grass_{sat} + 1) \), \( \ln(crops_{sat} + 1) \), \( \ln(noveg_{sat} + 1) \), and \( \ln(water_{sat} + 1) \), but excluding \( \ln(forest_{sat} + 1) \) as a reference because the six surface groups are collinear)\(^{35}\) or \( \ln(NL + 1) \). \( C \) represents a vector of control variables and \( \nu \) 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. Furthermore, to consistently examine patterns in the temporal distribution of the regression residuals, we maintain a data structure of consecutive years in the household income analysis by excluding the 2005 observations. The years of observation are thus 2000 through 2013 for the GDP analysis and 2009 through 2013 for the household income analysis.

To assess whether the combination of surface groups validly proxies economic activity, we follow Henderson et al. (2012) by using the natural logarithms of the dependent variables and the independent variables. As the variables in \( X \) contain values of zero, we add the value one to the variables before taking their natural logarithms. In their assessment of night lights as a country-level proxy for GDP, Henderson et al. (2012) argue

\(^{35}\) We choose the surface group \( forest \) as the reference group because it is the largest of the six. Our results are robust to choosing a different reference group except for the surface group \( builtup \) (available upon request). This result is in line with our expectation that \( builtup \) is the most important of the six for predicting economic activity.
that night lights 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. 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 the Landsat or the night lights 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.\footnote{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 3.2) 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 (available upon request).}

The results of the county-level analysis with real GDP as the dependent variable in Table 3 show that surface groups explain more of the variation in GDP than night lights. In the specifications without control variables, surface groups explain 42.5 percent of the variation in GDP (column 1), whereas night lights explain only 22.9 percent of this variation (column 3). Including the control variables does not affect this pattern, with surface groups explaining 60.5 percent (column 2) and night lights explaining 49.0 percent of the variation in GDP (column 4). As the specifications including the control variables explain a larger percentage of the variation in GDP for both surface groups and night lights, controlling for potential measurement error improves the prediction but neither affects the predictive properties of surface groups nor those of night lights. 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.

Figure 3 shows that the statistical distribution of the residuals from the OLS regressions including the control variables (columns 2 and 4 of table 3) looks smoother and narrower...
Table 3: OLS prediction of GDP using surface groups and using night lights (county level, 2000–2013)

<table>
<thead>
<tr>
<th>Dep. var.: ln(GDP)</th>
<th>Surface groups</th>
<th>Night lights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>ln(builtup$_{sat} + 1$)</td>
<td>1.659***</td>
<td>1.297***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>ln(grass$_{sat} + 1$)</td>
<td>-0.093***</td>
<td>-0.223***</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>ln(forest$_{sat} + 1$)</td>
<td>(ref.)</td>
<td>(ref.)</td>
</tr>
<tr>
<td>ln(crops$_{sat} + 1$)</td>
<td>-0.339***</td>
<td>-0.230***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>ln(noveg$_{sat} + 1$)</td>
<td>-0.454***</td>
<td>-0.250***</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>ln(water$_{sat} + 1$)</td>
<td>-0.191***</td>
<td>-0.083***</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>ln(NL + 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year FE</td>
<td>No</td>
<td>Yes***</td>
</tr>
<tr>
<td>Federal state FE</td>
<td>No</td>
<td>Yes***</td>
</tr>
<tr>
<td>N</td>
<td>5,586</td>
<td>5,586</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>0.425</td>
<td>0.605</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, GENESIS data, and BKG data.
Notes: Robust standard errors in parentheses. All models include intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Figure 3: Statistical distribution of GDP residuals for surface groups and for night lights.

(a) Surface groups

(b) Night lights

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, GENESIS data, and BKG data.

Note: Residuals stem from the regressions in Table 3 (OLS specifications including year FE and federal state FE). Figure shows histograms with a bin width of 0.05.

for surface groups than for night lights. This finding is in line with the surface groups explaining more of the variation in GDP than night lights, as indicated by the adjusted $R^2$ of the regressions. Moreover, for both surface groups and night lights, 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 night lights.

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, Figure C1 illustrates the temporal and spatial distribution of the residuals from the regression in column 2 of Table 3. (For reference, appendix B 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, Figure C1 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, Figure C2 proceeds similarly for the night lights data, illustrating the
residuals from the regression in column 4 of Table 3.

Figures C1 and C2 in Appendix C 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 204 counties (51.1 percent), and, for the remaining regions, the color varies over time in Figure C1. For night lights, this pattern appears for 347 counties (87.0 percent), leading to the four maps in Figure C2 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 lights.

The distribution of the residuals across regions in Figures C1 and C2 suggests a somewhat larger spatial bias in prediction error for surface groups than for night lights. If such a bias existed, clusters of similarly colored regions would appear. For surface groups, 1,014 observations (18.2 percent) have the same color as all their geographically neighboring observations, whereas for night lights, this pattern shows for only 538 observations (9.6 percent). However, for both surface groups and night lights, 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 17 counties (4.3 percent) have the same color as all their neighboring observations and, simultaneously, the same color throughout all observation years. For surface groups, this pattern appears for 26 counties (6.5 percent). This finding reflects in the small clusters of counties with the same color not showing up in consecutive years in Figure C1.

To replicate Henderson et al.’s (2012) cross-country analysis at the level of counties, we estimate Equation 4 using FE regression in a separate specification. The reason that

---

37 The county FE will cancel out the federal state FE in these regressions.
Henderson et al. (2012) include FE at the country level (their regional unit of observation) is to control for differences in night lights 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, such as our analysis of German regions, such differences do not occur. Furthermore, our objective is to proxy differences in economic activity both within and between regions. Therefore, we expect neither surface groups nor night lights to achieve equally compelling FE results as in Henderson et al. (2012).

We report the FE results in Table E1 in Appendix E. As expected and in line with the findings of Goldblatt et al. (2019), the adjusted within-$R^2$ is smaller than in Henderson et al.’s (2012) country-level analysis. In the FE prediction, surface groups and night lights perform almost equally in predicting within-county variation in GDP, with surface groups explaining 29.0 percent of this variation and night lights explaining 29.8 percent.

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. Surface groups explain a significant percentage of the variation in GDP. Moreover, in comparison to the 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.

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 4 presents the estimation results for this grid-level analysis. At this very detailed regional level, the surface groups-based predictions explain a much larger percentage of the variation in household income than the night lights-based predictions (64.4 percent vs. 28.1 percent in the specifications without control variables in columns 1 and 3, and 68.2 percent vs. 31.8 percent 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 boundaries.

Table 4: OLS prediction of household income using surface groups and using night lights (grid level, 2009–2013)

<table>
<thead>
<tr>
<th>Dep. var.: ln(HHI)</th>
<th>Surface groups</th>
<th>Night lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(builtup\textsubscript{sat} + 1)</td>
<td>1.529***</td>
<td>1.509***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>ln(grass\textsubscript{sat} + 1)</td>
<td>-0.064***</td>
<td>-0.105***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>ln(forest\textsubscript{sat} + 1)</td>
<td>(ref.)</td>
<td>(ref.)</td>
</tr>
<tr>
<td>ln(crops\textsubscript{sat} + 1)</td>
<td>-0.406***</td>
<td>-0.346***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>ln(noveg\textsubscript{sat} + 1)</td>
<td>-0.250***</td>
<td>-0.219***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>ln(water\textsubscript{sat} + 1)</td>
<td>-0.290***</td>
<td>-0.242***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>ln(NL + 1)</td>
<td></td>
<td>0.976***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.005***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.002)</td>
</tr>
<tr>
<td>Year FE</td>
<td>No</td>
<td>Yes***</td>
</tr>
<tr>
<td>Federal state FE</td>
<td>No</td>
<td>Yes***</td>
</tr>
</tbody>
</table>

| N                  | 737,822      | 737,822     |
| Adj. R\textsuperscript{2} | 0.644        | 0.682       |

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, RWI and microm (2019) data, and ESDAC data.
Notes: Robust standard errors in parentheses. All models include intercept. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).

Figure 4 confirms the findings of the regressions. The figure shows that the statistical distribution of the prediction error for household income is much narrower (although slightly left-skewed) for surface groups than for night lights. The distribution of the prediction error for night lights 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...
Figure 4: Statistical distribution of household income residuals for surface groups and for night lights

(a) Surface groups

(b) Night lights

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, RWI and microm (2019) data, and ESDAC data.

Note: Residuals stem from the regressions in Table 4 (OLS specifications including year FE and federal state FE). Figure shows histograms with a bin width of 0.1.

household income at the grid level much more precisely than night lights.

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. Figures D1 and D2 in Appendix D show the spatial and temporal distribution of the prediction error in household income for surface groups and night lights, respectively. For a better illustration of the very small grid cells, the map shows an area at the borders of four federal states—Rhineland-Palatinate, Hesse, Baden-Württemberg, and Bavaria—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, 87,819 grid cells (58.0 percent) have the same color throughout all observation years, whereas this number amounts to 131,025 grid cells (86.5 percent) for night lights. 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 239,733 observations (32.5 percent)
for surface groups and 223,424 observations (30.3 percent) for night lights having the same color as all their geographical neighbors. Combining the two types of biases shows that for surface groups, 22,326 grid cells (14.7 percent) have the same color as their neighbors and, simultaneously, the same color throughout all observation years. For night lights, this pattern applies to 35,130 grid cells (23.2 percent). Therefore, the smaller temporal bias of surface groups again outweighs their slightly larger spatial bias.

In Table E2 in Appendix E, we replicate Henderson et al.’s (2012) country-level analysis at the grid level using FE regression. Both the surface groups-based and the night lights-based prediction explain only a small percentage of the within-grid cell variation in household income. However, as in the GDP analysis, the FE specification does not fit our application of proxying household income for small regional units within the same country.

In sum, our analyses at the county level (GDP) and at the grid level (household income) suggest that 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 night lights. 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, suggesting 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 2.5), surface groups can also function as a valid proxy for economic activity for earlier years. Nonetheless, for cross-country studies or other larger regions, night lights might still be more appropriate. The reason is that land use characteristics might have heterogeneous meanings for a country’s economy, depending on the country’s historical development (Henderson et al., 2018). However, for small regional units and early time series, surface groups constitute a valuable and more accurate proxy for economic activity.

For 1984 and 1985, the percentage of cloud-covered pixels in the Landsat data in Germany is higher than in subsequent years. When applying the surface groups to time series containing these two years, including the number of cloud-covered pixels as a control variable might improve the proxy substantially.
3.4. Validation of Surface Groups as Proxies for Land Cover

After analyzing the quality of surface groups as a proxy for economic activity, we investigate their usefulness for economic research analyzing actual land cover (e.g., urbanization). In so doing, we use external data from administrative statistics to examine the quality of surface groups as proxies for their corresponding types of land cover. This analysis allows us to draw conclusions on the external validity of surface groups as proxies for administrative land cover, and the analysis also complements the assessment of internal validity in Section 2.5.

To assess the validity of surface groups as proxies for their corresponding types of land cover, we perform separate OLS regressions for each of the six surface groups. The dependent variables in these regressions are the natural logarithms of land cover as indicated in administrative data: $\ln(builtup_{adm} + 1)$, $\ln(grass_{adm} + 1)$, $\ln(forest_{adm} + 1)$, $\ln(crops_{adm} + 1)$, $\ln(noveg_{adm} + 1)$, or $\ln(water_{adm} + 1)$. The main independent variables are the natural logarithms of the corresponding surface groups we retrieve from satellite data: $\ln(builtup_{sat} + 1)$, $\ln(grass_{sat} + 1)$, $\ln(forest_{sat} + 1)$, $\ln(crops_{sat} + 1)$, $\ln(noveg_{sat} + 1)$, or $\ln(water_{sat} + 1)$. To control for potential measurement error in the dependent and independent variables, we again include year FE and federal state FE (see section 3.3).

As a comparison to night lights data makes no sense for assessing land cover, we also include the percentage of missing pixels (due to cloud cover) to capture potential measurement error in the number of pixels per surface group (see section 3.2). The unit of observation in these regressions is the municipality, because the administrative land cover data is available at this level. The observation period is 2008–2015.

Table 5 shows the results of the regressions for each surface group. For all six surface groups, an increase in the satellite-based measure is significantly associated with an increase in the administrative measure, whether we include the control variables or not. As adjusted $R^2$ indicates, with 80.2 percent builtup is the surface group that explains most of the variation in the corresponding administrative measure (column 2), closely followed by crops, with 79.7 percent (column 8), and forest, with 77.3 percent (column 6). The surface groups grass (44.2 percent), noveg (53.7 percent), and water (59.3 percent)
Table 5: OLS prediction of administrative land cover using surface groups (municipality level, 2008–2015)

<table>
<thead>
<tr>
<th></th>
<th>$\ln(\text{builtup}_{\text{adm}} + 1)$</th>
<th>$\ln(\text{grass}_{\text{adm}} + 1)$</th>
<th>$\ln(\text{forest}_{\text{adm}} + 1)$</th>
<th>$\ln(\text{crops}_{\text{adm}} + 1)$</th>
<th>$\ln(\text{noveg}_{\text{adm}} + 1)$</th>
<th>$\ln(\text{water}_{\text{adm}} + 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td>$\ln(\text{builtup}_{\text{sat}} + 1)$</td>
<td>0.874***</td>
<td>0.883***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln(\text{grass}_{\text{sat}} + 1)$</td>
<td></td>
<td>0.812***</td>
<td>0.816***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.007)</td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln(\text{forest}_{\text{sat}} + 1)$</td>
<td></td>
<td></td>
<td>1.126***</td>
<td>1.126***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln(\text{crops}_{\text{sat}} + 1)$</td>
<td></td>
<td></td>
<td></td>
<td>0.834***</td>
<td>0.812***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.003)</td>
<td>(0.004)</td>
<td></td>
</tr>
<tr>
<td>$\ln(\text{noveg}_{\text{sat}} + 1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.254***</td>
<td>0.954***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.003)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>$\ln(\text{water}_{\text{sat}} + 1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.927***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.004)</td>
</tr>
<tr>
<td>Cloud cover</td>
<td>No</td>
<td>Yes***</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Year FE</td>
<td>No</td>
<td>Yes***</td>
<td>No</td>
<td>Yes***</td>
<td>No</td>
<td>Yes***</td>
</tr>
<tr>
<td>Federal state FE</td>
<td>No</td>
<td>Yes***</td>
<td>No</td>
<td>Yes***</td>
<td>No</td>
<td>Yes***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>80,543</td>
<td>80,543</td>
<td>80,543</td>
<td>80,543</td>
<td>80,543</td>
<td>80,543</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>0.755</td>
<td>0.802</td>
<td>0.130</td>
<td>0.442</td>
<td>0.748</td>
<td>0.773</td>
</tr>
</tbody>
</table>

Source: Authors' calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Notes: Robust standard errors in parentheses. All models include intercept. * $p<0.10$, ** $p<0.05$, *** $p<0.01$. 
explain less of the variation in the corresponding administrative measure. Although one reason for their lower predictive value may be that our classification algorithm detects these surface groups with a lower accuracy, the results of the five-fold cross-validation in Section 2.5 suggest otherwise. Therefore, more likely is that the administrative measures do not indicate the six types of land cover with equal accuracy, due to the level of aggregation in the data (see section 3.2). In other words, the lower predictive values likely result from the differing aggregations of land cover types in the administrative data.

Appendix F depicts the statistical distribution of the residuals by surface group, stemming from the regressions in columns 2, 4, 6, 8, 10, and 12 of Table 5. In line with the assessment of adjusted $R^2$, the distribution is much narrower and smoother for builtup, forest, and crops than for the other three groups. Thus differences exist in the predictive values of the surface groups.

Appendix G shows the temporal and spatial distribution of the residuals by surface group. Confirming the previous finding, the intensity of the colors (i.e., the degree of prediction error) is far lower for builtup, forest, and crops than for the other three groups. Regarding temporal bias, the residual has the same direction throughout all observation years in 3,754 municipalities (36.7 percent) for builtup, 7,530 municipalities (73.7 percent) for grass, 4,101 municipalities (40.1 percent) for forest, 5,575 municipalities (54.5 percent) for crops, 5,614 municipalities (54.9 percent) for noveg, and 5,211 municipalities (51.0 percent) for water. Regarding spatial bias, we find the same direction of the residual in all neighboring regions for 24,823 observations (30.8 percent) for builtup, 17,285 observations (21.5 percent) for grass, 27,405 observations (34.0 percent) for forest, 28,247 observations (35.1 percent) for crops, 19,063 observations (23.7 percent) for noveg, and 18,854 observations (23.4 percent) for water. The combined bias persists in 390 municipalities (3.8 percent) for builtup, 965 municipalities (9.4 percent) for grass, 527 municipalities (5.2 percent) for forest, 926 municipalities (9.1 percent) for crops, 512 municipalities (5.0 percent) for noveg, and 468 municipalities (4.6 percent) for water.

In sum, we argue that the six surface groups—builtup, grass, forest, crops, noveg, and water—are valid predictors of the corresponding types of land cover in administrative
statistics and thus useful measures for economic research that investigates actual land cover. Based on the explained percentage of the variation in the administrative measure and the combined bias indicator, the surface group $builtup$ performs best in proxying its administrative counterpart. The prediction error and bias indicators we observe in our analysis for some of the remaining surface groups likely originates in the differing aggregations of land cover types caused by the data structure of the administrative statistics.

4. Conclusion

In this paper, we develop a novel procedure for proxying economic activity across time periods and highly disaggregated regional levels, for which other data is unreliable, inaccessible, or entirely inexistent. We develop this proxy by applying machine-learning techniques to daytime satellite imagery that dates back to 1984. Compared to night lights intensity, a commonly used satellite-based proxy for economic activity, our proxy has the advantages of more precisely predicting economic activity over a longer time series and at more detailed regional levels. We demonstrate the proxy’s usefulness for Germany, where data on economic activity is otherwise unavailable for the time periods and spatial units we analyze. In this particular example, the proxy provides valuable information on economic activity for the regions belonging to the former German Democratic Republic. However, our procedure is generalizable to any region or country, and thus constitutes a valuable resource for analyzing historical developments, evaluating local policy reforms, and controlling for economic activity at highly disaggregated regional levels in econometric applications.

To construct our proxy, we identify six surface groups—different types of surfaces covering earth—from Landsat satellite data. We detect these groups using machine-learning techniques on the GEE platform. We show that the regional combination of these groups provides a valid proxy for economic activity, even for small regional units such as municipalities or even housing blocks. Furthermore, we compare the predictive value of
surface groups to the benchmark value of night lights, the most commonly used proxy for GDP in economic research (Donaldson and Storeygard, 2016), and find that, for small regional units in Germany, surface groups explain a higher percentage of the variation in economic activity than night lights. Moreover, surface groups can function as a direct indicator of different types of land cover.

Therefore, the six surface groups that we construct constitute a useful measure for applications in economic research such as evidence-based policy analyses. Surface groups are available from 1984, thus providing a uniquely long time series of data. This long time series can help improve identification in analyses in the context of historical events, such as the fall of the iron curtain. Furthermore, surface groups can proxy economic activity at very detailed regional levels. Surface groups thus contribute to analyses of the regional impacts of local policy reforms by providing information on economic activity at very detailed regional levels, such as municipalities. For example, the openings of new tertiary education institutions in Germany affect the economy at the local level, and analyzing the effects of such openings requires information on economic activity at very detailed regional levels, for which other data sources are entirely unavailable for the necessary observation period, unreliable, less precise, or inaccessible for non-residents.

In developing a procedure for detecting surface groups from Landsat data, we present a novel approach to using this data for economic research. While we apply our procedure to Germany and establish its validity for this country, the procedure is transferable to other countries. Although a country’s history or industry structure affects the economic importance of different types of land cover (Henderson et al., 2018), the principle that different types of land cover, which the surface groups reflect, relate to economic activity applies to any country in the world. Therefore, surface groups have a potential for economic research that investigates small regions within the same country or within a homogeneous group of countries. Retrieving the surface groups for other countries requires the availability of suitable training data for the machine-learning algorithm, and potentially further country-specific computational adjustments.

Future research could try to exploit data from other satellite programs, such as the
Advanced Spaceborne Thermal Emissions and Reflection Radiometer (ASTER) (e.g., Yamaguchi et al., 1998) or the Sentinel mission (e.g., Berger et al., 2012), for applications in economics. For the years that they cover, these data could provide more accurate classifications of land cover and thus potentially even more precise predictions of economic activity at very small regional levels. However, these data cover substantially shorter time series than Landsat data.
References


### Appendices

#### Appendix A. Five-Fold Cross-Validation Results by Surface Group

Table A1: Five-fold cross-validation results with respect to built-up surfaces (surface group *builtup*).

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy $(OA)$</th>
<th>True-positive rate $(TPR)$</th>
<th>True-negative rate $(TNR)$</th>
<th>Balanced accuracy $(BA)$</th>
<th>User’s accuracy $(UA)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.840</td>
<td>0.662</td>
<td>0.875</td>
<td>0.768</td>
<td>0.516</td>
</tr>
<tr>
<td>2000</td>
<td>0.856</td>
<td>0.610</td>
<td>0.905</td>
<td>0.758</td>
<td>0.563</td>
</tr>
<tr>
<td>2006</td>
<td>0.867</td>
<td>0.636</td>
<td>0.913</td>
<td>0.774</td>
<td>0.593</td>
</tr>
<tr>
<td>2012</td>
<td>0.840</td>
<td>0.609</td>
<td>0.886</td>
<td>0.748</td>
<td>0.517</td>
</tr>
<tr>
<td>2018</td>
<td>0.859</td>
<td>0.610</td>
<td>0.909</td>
<td>0.760</td>
<td>0.575</td>
</tr>
<tr>
<td>Average</td>
<td>0.852</td>
<td>0.625</td>
<td>0.898</td>
<td>0.762</td>
<td>0.553</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.

Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Table A2: Five-fold cross-validation results with respect to grassy surfaces (surface group \textit{grass})

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy (\textit{OA})</th>
<th>True-positive rate (\textit{TPR})</th>
<th>True-negative rate (\textit{TNR})</th>
<th>Balanced accuracy (\textit{BA})</th>
<th>User’s accuracy (\textit{UA})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.833</td>
<td>0.477</td>
<td>0.905</td>
<td>0.691</td>
<td>0.501</td>
</tr>
<tr>
<td>2000</td>
<td>0.839</td>
<td>0.646</td>
<td>0.877</td>
<td>0.762</td>
<td>0.513</td>
</tr>
<tr>
<td>2006</td>
<td>0.835</td>
<td>0.590</td>
<td>0.884</td>
<td>0.737</td>
<td>0.504</td>
</tr>
<tr>
<td>2012</td>
<td>0.843</td>
<td>0.491</td>
<td>0.913</td>
<td>0.702</td>
<td>0.530</td>
</tr>
<tr>
<td>2018</td>
<td>0.825</td>
<td>0.524</td>
<td>0.885</td>
<td>0.705</td>
<td>0.478</td>
</tr>
<tr>
<td>Average</td>
<td>0.835</td>
<td>0.546</td>
<td>0.893</td>
<td>0.719</td>
<td>0.505</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Table A3: Five-fold cross-validation results with respect to forest-covered surfaces (surface group *forest*).

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy (OA)</th>
<th>True-positive rate (TPR)</th>
<th>True-negative rate (TNR)</th>
<th>Balanced accuracy (BA)</th>
<th>User’s accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.894</td>
<td>0.534</td>
<td>0.966</td>
<td>0.750</td>
<td>0.763</td>
</tr>
<tr>
<td>2000</td>
<td>0.903</td>
<td>0.790</td>
<td>0.926</td>
<td>0.858</td>
<td>0.680</td>
</tr>
<tr>
<td>2006</td>
<td>0.892</td>
<td>0.806</td>
<td>0.909</td>
<td>0.857</td>
<td>0.639</td>
</tr>
<tr>
<td>2012</td>
<td>0.896</td>
<td>0.767</td>
<td>0.922</td>
<td>0.845</td>
<td>0.665</td>
</tr>
<tr>
<td>2018</td>
<td>0.899</td>
<td>0.774</td>
<td>0.924</td>
<td>0.849</td>
<td>0.672</td>
</tr>
<tr>
<td>Average</td>
<td>0.897</td>
<td>0.734</td>
<td>0.930</td>
<td>0.832</td>
<td>0.684</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Table A4: Five-fold cross-validation results with respect to surfaces with crop fields (surface group *crops*)

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy (OA)</th>
<th>True-positive rate (TPR)</th>
<th>True-negative rate (TNR)</th>
<th>Balanced accuracy (BA)</th>
<th>User’s accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.808</td>
<td>0.517</td>
<td>0.867</td>
<td>0.692</td>
<td>0.441</td>
</tr>
<tr>
<td>2000</td>
<td>0.850</td>
<td>0.482</td>
<td>0.924</td>
<td>0.703</td>
<td>0.559</td>
</tr>
<tr>
<td>2006</td>
<td>0.864</td>
<td>0.492</td>
<td>0.938</td>
<td>0.715</td>
<td>0.613</td>
</tr>
<tr>
<td>2012</td>
<td>0.860</td>
<td>0.468</td>
<td>0.939</td>
<td>0.703</td>
<td>0.604</td>
</tr>
<tr>
<td>2018</td>
<td>0.862</td>
<td>0.545</td>
<td>0.925</td>
<td>0.735</td>
<td>0.594</td>
</tr>
<tr>
<td>Average</td>
<td>0.849</td>
<td>0.501</td>
<td>0.919</td>
<td>0.710</td>
<td>0.562</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Table A5: Five-fold cross-validation results with respect to surfaces without vegetation (surface group *noveg*)

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy (OA)</th>
<th>True-positive rate (TPR)</th>
<th>True-negative rate (TNR)</th>
<th>Balanced accuracy (BA)</th>
<th>User’s accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.893</td>
<td>0.682</td>
<td>0.934</td>
<td>0.808</td>
<td>0.668</td>
</tr>
<tr>
<td>2000</td>
<td>0.909</td>
<td>0.651</td>
<td>0.960</td>
<td>0.806</td>
<td>0.766</td>
</tr>
<tr>
<td>2006</td>
<td>0.907</td>
<td>0.630</td>
<td>0.963</td>
<td>0.796</td>
<td>0.774</td>
</tr>
<tr>
<td>2012</td>
<td>0.906</td>
<td>0.760</td>
<td>0.935</td>
<td>0.848</td>
<td>0.702</td>
</tr>
<tr>
<td>2018</td>
<td>0.920</td>
<td>0.761</td>
<td>0.951</td>
<td>0.856</td>
<td>0.758</td>
</tr>
<tr>
<td>Average</td>
<td>0.907</td>
<td>0.697</td>
<td>0.949</td>
<td>0.823</td>
<td>0.733</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Table A6: Five-fold cross-validation results with respect to water surfaces (surface group *water*)

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall accuracy (OA)</th>
<th>True-positive rate (TPR)</th>
<th>True-negative rate (TNR)</th>
<th>Balanced accuracy (BA)</th>
<th>User’s accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.910</td>
<td>0.665</td>
<td>0.959</td>
<td>0.812</td>
<td>0.762</td>
</tr>
<tr>
<td>2000</td>
<td>0.906</td>
<td>0.610</td>
<td>0.966</td>
<td>0.788</td>
<td>0.780</td>
</tr>
<tr>
<td>2006</td>
<td>0.907</td>
<td>0.658</td>
<td>0.956</td>
<td>0.807</td>
<td>0.750</td>
</tr>
<tr>
<td>2012</td>
<td>0.915</td>
<td>0.689</td>
<td>0.961</td>
<td>0.825</td>
<td>0.780</td>
</tr>
<tr>
<td>2018</td>
<td>0.921</td>
<td>0.644</td>
<td>0.976</td>
<td>0.810</td>
<td>0.844</td>
</tr>
<tr>
<td>Average</td>
<td>0.912</td>
<td>0.653</td>
<td>0.964</td>
<td>0.808</td>
<td>0.783</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data and CLC data.
Notes: The yearly values indicate the average over all five iterations within the respective year. Average indicates the average over the yearly values.
Appendix B. Reference Map of German Federal States and Their Capitals

Figure B1: Reference map of German federal states and their capitals

Source: Authors’ illustration with BKG data.
Appendix C. Spatial and Temporal Distribution of GDP Residuals Using Surface Groups and Using Night Lights

Figure C1: Spatial and temporal distribution of GDP residuals for surface groups

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Note: Maps illustrate residuals from the regression in column 2 of Table 3.
Figure C2: Spatial and temporal distribution of GDP residuals for night lights

(a) 2000

(b) 2004

(c) 2008

(d) 2012

Source: Authors’ calculations based on DMSP OLS data, GENESIS data, and BKG data.
Note: Maps illustrate residuals from the regression in column 4 of Table 3.
Appendix D. Spatial and Temporal Distribution of Household Income Residuals Using Surface Groups and Using Night Lights

Figure D1: Spatial and temporal distribution of household income residuals for surface groups

Source: Authors’ calculations based on Landsat data, CORINE data, RWI and microm (2019) data, and ESDAC data.

Note: Maps illustrate residuals from the regression in column 2 of Table 4. Maps show an area at the borders of the four federal states Rhineland-Palatinate, Hesse, Baden-Württemberg, and Bavaria.
Figure D2: Spatial and temporal distribution of household income residuals for night lights

(a) 2009

(b) 2010

(c) 2011

(d) 2012

Source: Authors’ calculations based on DMSP OLS data, RWI and microm (2019) data, and ESDAC data.
Note: Maps illustrate residuals from the regression in column 5 of Table 4. Maps show an area at the borders of the four federal states Rhineland-Palatinate, Hesse, Baden-Württemberg, and Bavaria.
### Appendix E. FE Prediction of Economic Activity Using Surface Groups and Using Night Lights

**Table E1:** FE prediction of GDP using surface groups and using night lights (county level, 2000–2013)

<table>
<thead>
<tr>
<th>Dep. var.: ( \ln(GDP) )</th>
<th>Surface groups</th>
<th>Night lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ln(\text{builtup}_{sat} + 1) )</td>
<td>0.015*</td>
<td>(ref.)</td>
</tr>
<tr>
<td>(0.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(\text{grass}_{sat} + 1) )</td>
<td>-0.006</td>
<td>(ref.)</td>
</tr>
<tr>
<td>(0.006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(\text{forest}_{sat} + 1) )</td>
<td>(ref.)</td>
<td>(ref.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(\text{crops}_{sat} + 1) )</td>
<td>-0.016***</td>
<td>(0.005)</td>
</tr>
<tr>
<td>(0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(\text{noveg}_{sat} + 1) )</td>
<td>-0.012***</td>
<td>(0.003)</td>
</tr>
<tr>
<td>(0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(\text{water}_{sat} + 1) )</td>
<td>0.005</td>
<td>(0.004)</td>
</tr>
<tr>
<td>(0.004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln(NL + 1) )</td>
<td>0.071***</td>
<td>(0.010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes***</td>
<td>Yes***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>5,586</td>
<td>5,586</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj. within-( R^2 )</td>
<td>0.290</td>
<td>0.298</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, GENESIS data, and BKG data.

Notes: Robust standard errors in parentheses. All models include intercept. *\( p < 0.10 \), **\( p < 0.05 \), ***\( p < 0.01 \).
Table E2: FE prediction of household income using surface groups and using night lights (grid level, 2009–2013)

<table>
<thead>
<tr>
<th>Dep. var.: ( \ln(HHI) )</th>
<th>Surface groups</th>
<th>Night lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ln(builtup_{sat} + 1) )</td>
<td>0.002***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>( \ln(grass_{sat} + 1) )</td>
<td>0.001*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>( \ln(forest_{sat} + 1) )</td>
<td>(ref.)</td>
<td></td>
</tr>
<tr>
<td>( \ln(crops_{sat} + 1) )</td>
<td>0.004***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>( \ln(noveg_{sat} + 1) )</td>
<td>-0.001***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>( \ln(water_{sat} + 1) )</td>
<td>-0.001***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>( \ln(NL + 1) )</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
</tr>
</tbody>
</table>

Year FE | Yes*** | Yes***
---|---|---
\( N \) | 737,822 | 737,822
Adj. within-\( R^2 \) | 0.044 | 0.044

Source: Authors’ calculations based on Landsat data, CORINE data, DMSP OLS data, RWI and microm (2019) data, and ESDAC data.

Notes: Robust standard errors in parentheses. All models include intercept. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).
Appendix F. Statistical Distribution of Land Cover Residuals by Surface Group

Figure F1: Statistical distribution of land cover residuals by surface group

(a) builtup
(b) grass
(c) forest
(d) crops
(e) noveg
(f) water

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data. Note: Residuals stem from the regressions in Table 5 (specifications including cloud cover, year FE, and federal state FE). Figure shows histograms with a bin width of 0.1.
Appendix G. Spatial and Temporal Distribution of Land Cover Residuals by Surface Group

Figure G1: Spatial and temporal distribution of land cover residuals for surface group \textit{builtup}

(a) 2008
(b) 2010
(c) 2012
(d) 2014

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data. Note: Maps illustrate residual that stems from the regression in column 2 of Table 5.
Figure G2: Spatial and temporal distribution of land cover residuals for surface group *grass*

(a) 2008  
(b) 2010  
(c) 2012  
(d) 2014

Source: Authors' calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.  
Note: Maps illustrate residual that stems from the regression in column 4 of Table 5.
Figure G3: Spatial and temporal distribution of land cover residuals for surface group forest

(a) 2008  (b) 2010  (c) 2012  (d) 2014

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Note: Maps illustrate residual that stems from the regression in column 6 of Table 5.
Figure G4: Spatial and temporal distribution of land cover residuals for surface group crops

(a) 2008  (b) 2010

(c) 2012  (d) 2014

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Note: Maps illustrate residual that stems from the regression in column 8 of Table 5.
Figure G5: Spatial and temporal distribution of land cover residuals for surface group *novelg*

(a) 2008  
(b) 2010  
(c) 2012  
(d) 2014

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Note: Maps illustrate residual that stems from the regression in column 10 of Table 5.
Figure G6: Spatial and temporal distribution of land cover residuals for surface group water.

Source: Authors’ calculations based on Landsat data, CORINE data, GENESIS data, and BKG data.
Note: Maps illustrate residual that stems from the regression in column 12 of Table 5.