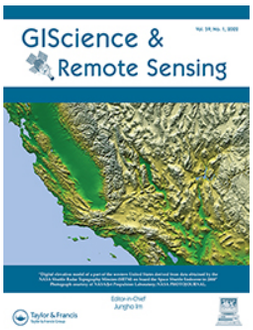

GIScience & Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgrs20

Assessing the relationship between morphology and mapping accuracy of built-up areas derived from global human settlement data

Johannes H. Uhl & Stefan Leyk

To cite this article: Johannes H. Uhl & Stefan Leyk (2022) Assessing the relationship between morphology and mapping accuracy of built-up areas derived from global human settlement data, GIScience & Remote Sensing, 59:1, 1722-1748, DOI: 10.1080/15481603.2022.2131192

To link to this article: https://doi.org/10.1080/15481603.2022.2131192

Published online: 12 Oct 2022.

GISCIENCE & REMOTE SENSING
2022, VOL. 59, NO. 1, 1722–1748
https://doi.org/10.1080/15481603.2022.2131192

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Assessing the relationship between morphology and mapping accuracy of built-up areas derived from global human settlement data

Johannes H. Uhl [a,b] and Stefan Leyk [b,c]

[a]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, Colorado, USA; [b]Institute of Behavioral Science, University of Colorado Boulder, Boulder, Colorado, USA; [c]Department of Geography, University of Colorado Boulder, Boulder, Colorado, USA

ABSTRACT

It is common knowledge that the level of landscape heterogeneity may affect the performance of remote sensing based land use/land cover classification. While this issue has been studied in depth for land cover data in general, the specific relationship between the mapping accuracy and morphological characteristics of built-up surfaces has not been analyzed in detail, an urgent need given the recent emergence of a variety of global, fine-resolution settlement datasets. Moreover, previous studies typically rely on aggregated, broad-scale landscape metrics to quantify the morphology of built-up areas, neglecting the fine-grained spatial variation and scale dependency of such metrics. Herein, we aim to fill this knowledge gap by assessing the associations between localized (focal) landscape metrics, derived from binary built-up surfaces and localized data accuracy estimates. We tested our approach for built-up surfaces from the Global Human Settlement Layer (GHSL) for Massachusetts (USA). Specifically, we examined the explanatory power of landscape metrics with respect to both commission and omission errors in the multi-temporal GHS-BUILT R2018A data product. We found that the Landscape Shape Index (LSI) calculated in focal windows exhibits, on average, the highest levels of correlation to focal accuracy measures. These relationships are scale-dependent, and become stronger with increasing level of spatial support. We found that thematic omission error, as measured by Recall, has the strongest relationship to measures of built-up surface morphology across different temporal epochs and spatial resolutions. The results of our regression analysis ($R^2 > 0.9$), estimating accuracy based on landscape metrics, confirmed these findings. Lastly, we tested the generalizability of our findings by regionally stratifying our regression models and applying them to a different version of the GHSL (i.e. the GHS-BUILT-S2) and a different study area. We observed varying levels of model transferability, indicating that the relationship between accuracy and landscape metrics may be sensor-specific, and is heavily localized for most accuracy metrics, but quite generalizable for the Recall measure. This indicates that there is a strong and generalizable association between morphological properties of built-up land and the degree to which it is "undermapped."

ARTICLE HISTORY
Received 19 May 2022
Accepted 27 September 2022

KEYWORDS
Global human settlement layer; spatially explicit accuracy assessment; landscape metrics; predictive uncertainty modeling; AdaBoost regression; domain shift

1. Introduction

In order to analyze the dynamics of human settlements on Earth, researchers typically rely on multi-temporal, remote-sensing-derived, gridded built-up surface datasets, such as the Global Human Settlement Layer (GHSL, Pesaresi et al. 2013), the Global Rural-Urban Mapping Project (GRUMP, Balk et al. 2005), the Global Artificial Impervious Area dataset (GAIA, Gong et al. 2020), or the World Settlement Footprint Evolution dataset (Marconcini et al. 2020a). In order to develop an unbiased understanding of the human settlement trends measured by these data, thorough knowledge of the uncertainty inherent in these multi-temporal datasets is crucial. The quantification of uncertainty in categorical spatial data is typically done by means of map comparison, i.e. the comparison to an independently compiled reference dataset of presumably higher accuracy (FGDC (Federal Geographic Data Committee) 1998), involving the creation of confusion matrices and the derivation of accuracy metrics (Fielding and Bell 1997). The accuracy assessment of remote-sensing-derived land cover/land use data is not straight-forward, for several reasons: (a) data accuracy is a spatially varying phenomenon, and accuracy estimates based on small samples, or aggregated to global or region-specific estimates, may ignore the fine-scale spatial non-stationarity of data accuracy (e.g. Strahler et al. 2006; Foody 2007; Wickham, Stehman, and

CONTACT Johannes H. Uhl johannes.uhl@colorado.edu

Homer 2018). (b) the accuracy metrics themselves may be biased, as they can be sensitive to sample size (e.g. Sim and Wright 2005; Bujang and Baharum 2017; Champagne et al. 2014) or class imbalance (see Rosenfield and Melley 1980; Wickham et al. 2010; Akosa 2017; Shao, Tang, and Liao 2019; Radoux, Waldner, and Bogaert 2020; Stehman and Wickham 2020). (c) The analytical unit at which an accuracy assessment is conducted, may affect the results (e.g. Pontius and Suedmeyer 2004; Pontius and Cheuk 2006; Stehman and Wickham 2011; Zhu et al. 2013; Ye, Pontius, and Rakshit 2018; Marconcini et al. 2020b), and (d) the appropriate choice of the sample size and distribution is critical to conduct an unbiased accuracy assessment (Congalton 1988; Hashemian, Abkar, and Fatemi 2004; Foody 2009; Stehman and Foody 2019). Lastly, the choice of the geographic unit, or assessment unit, for which accuracy metrics and the underlying confusion matrices are established, is crucial as well (e.g. Stehman 2009; Wardlow and Callahan 2014).

To account for the spatial variations in accuracy, researchers have started to use spatially explicit accuracy assessments, if reference data availability and computing resources permit (e.g. Löw et al. 2013; Khatami, Mountrakis, and Stehman 2017; Waldner et al. 2017; Mitchell, Downie, and Diesing 2018; Morales-Barquero et al. 2019; Uhl and Leyk 2022b) which are based on locally constrained confusion matrices (Foody 2007). Moreover, in order to account for the scarcity of reference data, their potentially resource-intensive creation, researchers have developed a wide range of methods for predictive accuracy modeling of geospatial data such as land cover data using a variety of techniques and explanatory variables (e.g. Steele, Winne, and Redmond 1998; Kyriakidis and Dungan 2001; Smith et al. 2003; Leyk and Zimmermann 2004, 2007; van Oort et al. 2004; Comber et al. 2012; Tsutsumida and Comber 2015; Zhang and Mei 2016; Wickham, Stehman, and Homer 2018; Mei et al. 2019; Ebrahimy et al. 2021; Cheng et al. 2021), while others have incorporated landscape metrics (LSMs) in land cover data accuracy assessments (Smith et al. 2002, 2003; Gu and Congalton 2020). Such studies typically focus on land cover data in general, and have not been applied to built-up surface data specifically.

In the specific case of built-up land datasets, accuracy assessments are often impeded by lack of reference data over large spatial extents (See et al. 2022), in particular for early points in time (Uhl and Leyk 2022a). Moreover, as it is well-known that the accuracy of remote-sensing-derived land use/land cover data products is related to structural landscape characteristics such as the level of landscape segregation or the patch size of urban land (Smith et al. 2002, 2003; Mück, Klotz, and Taubenböck 2017). In the same vein, Degen et al. (2018) show that the level of landscape heterogeneity affects the quantization of multispectral remote sensing data such as Landsat data. Previous research has shown that the accuracy of built-up surface layers varies regionally (Klotz et al. 2016; Liu et al. 2020), and across the rural-urban continuum (Leyk et al. 2018; Kaim et al. 2022), which is strongly related to morphological characteristics of landscapes in general (Vizzari 2011; Vizzari and Sigura 2013) and of settlements in particular (Cyriac and Firoz C 2022).

However, it has not been explicitly studied which morphological properties of settlements (as measured by landscape metrics) drive the accuracy at which they are mapped. Likewise, there is no literature that examines how individual uncertainty components (i.e. omission error, commission error) relate to morphological characteristics of built-up areas. This is the gap that this paper aims to fill. Knowledge of these relationships will help the users of settlement data (or of derived products such as fine-grained population data (e.g. Florczyk et al. 2019) to critically reflect on the data quality, and can guide data producers to improve data production pipelines, e.g. by using adaptive sampling and classification strategies based on the level of commission and omission errors expected in a region characterized by a specific built-up land morphology.

Herein, we make use of a multi-temporal reference dataset (i.e. the multi-temporal building footprint dataset for 33 U.S. counties (MTBF-33, Uhl and Leyk 2022a), enabling the creation of historical snapshots of built-up areas at fine spatial and temporal grain, for relatively large, contiguous regions. Using this reference dataset, we conducted a spatially exhaustive, localized accuracy assessment of the Global Human Settlement Layer (GHS-BUILT R2018A, Florczyk et al. 2019) in the state of Massachusetts (USA), for the epochs 1975 and 2014. Consistent to these multi-temporal, continuous surfaces of localized data accuracy estimates, we calculated focal landscape metrics

for a large sample of locations (N = 200,000 locations) to characterize the morphology of built-up areas. We used these data to (a) assess the association between localized data accuracy and landscape metrics at fine spatial grain, and over time, and (b) test the explanatory power of morphological characteristics of both the reference data and the GHSL with respect to data accuracy, using two different regression techniques. Finally, we tested the sensitivity of our results to the spatial support (i.e. the extent of the spatial sample used for focal/localized accuracy and landscape metrics computation) and to the assessment unit (i.e. the spatial resolution of the grid in which accuracy and landscape metrics are computed). Moreover, we analyzed the robustness of the relationships between landscape metrics and accuracy by means of domain adaptation (You et al. 2019) capabilities of our regression models to a different dataset (i.e. the GHS-BUILT-S2, Corbane et al. 2021) and to a study area outside of Massachusetts, as well as by model regionalization to the county-level within the state of Massachusetts to assess the spatial variation of these relationships.

This paper is structured as follows: In Section 2, we discuss the data and methods used, in Section 3, we present and discuss our results, and report our conclusions in Section 4.

2. Data and methods

In this section we introduce the used datasets and preprocessing steps undertaken (Section 2.1), as well as the methods used in the different parts of our analyses (Section 2.2).

2.1. Data and preprocessing

This study is based on gridded built-up surface layers from the GHSL project and on the multi-temporal building footprint dataset for 33 U.S. counties (MTBF-33, Uhl and Leyk 2022a).

2.1.1. Global human settlement layer (GHS-BUILT)

The GHS-BUILT R2018A dataset, which is derived from Landsat and Sentinel-2 data, maps built-up areas at a spatial resolution of 30 m, at a global extent, for the epoch (i.e. years) 1975, 1990, 2000, and 2014 (Florczyk et al. 2019, downloaded from https://data.jrc.ec.europa.eu/dataset/jrc-ghsl-10007). We used this data product, as the GHS-BUILT has been used in a range of scientific studies of different disciplines (Ehrlich et al. 2021) and has been input to the multi-temporal population datasets GHS-POP and the rural-urban classification datasets GHS-SMOD (Florczyk et al. 2019). Moreover, GHS-BUILT R2018A makes use of early Landsat 4 MSS data and thus, extends farther back in time than related datasets such as the WSF-evolution dataset, which dates back to 1985 (Marconcini et al. 2020a). The GHS-BUILT R2018A has been created using a sequential approach, extracting built-up areas in the most recent epoch (i.e. 2014) and spatially constraining built-up areas in prior epochs to the 2014 built-up mask. Note that Landsat 4 MSS data were upsampled to 30 m resolution to be integrated in this process (Corbane et al. 2019). For the two epochs 1975 and 2014, we extracted binary surfaces indicating built-up areas (1) and not built-up areas (0) (Figure 1a,b).

For the domain adaptation analysis, i.e. to test how regression models perform on data of a different distribution than the one they were trained on (Section 2.2.6), we employed the GHS-BUILT-S2 dataset, which provides estimates of built-up probability, in the range of 0–100, within a 10x10m grid. GHS-BUILT-S2 has been created from Sentinel-2 data acquired in 2018, using convolutional neural networks (Corbane et al. 2021, downloaded from https://ghsl.jrc.ec.europa.eu/ghs_bu_s2_2018.php). We used the data for a subset of the city of Charlotte, North Carolina. For our accuracy assessment, these continuous data needed to be converted into binary, presence-absence surfaces. To do so, we calculated the average built-up probability of the 10 m grid cells within 30x30m grid cells (aligned and consistent to the GHS-BUILT R2018A grid). We then applied a threshold of 50 to the built-up probabilities to generate binary built-up surface layers (Figure 1c), compatible to the GHS-BUILT R2018A data. This was done for two reasons: (a) the subsequent data processing requires binary, presence-absence surfaces, and (b) the original resolution of 10x10m is likely too fine-grained for direct calculation of landscape metrics. A target resolution of 30x30m generalizes the data such that meaningful landscape metrics can be derived (e.g. a contiguous patch of built-up surface should encompass the roads separating the actual buildings within that patch, and this may not be the case when using the

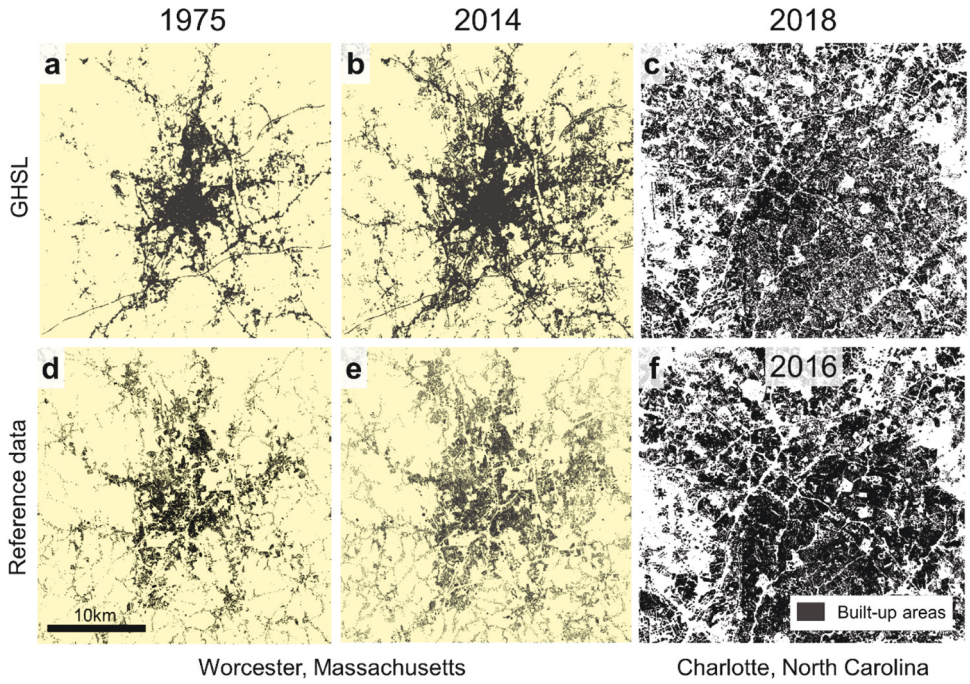

Figure 1. Samples of the input data used in this study. Built-up surfaces from the GHSL R2018A data (a) in 1975, and (b) in 2014, for the city of Worcester, Massachusetts, USA. Panel (c) shows the GHS-S2 built-up areas in 2018 for a subset of Mecklenburg County, North Carolina. The bottom row displays the reference data derived from the MTBF-33 dataset for Worcester (d) in 1975, and (e) in 2014. Panel (f) shows the MTBF-33 derived built-up areas for the North Carolina study area in 2016.

original resolution of 10x10m). Since the GHS-BUILT-S2 data stem from a different sensor, have a different resolution, processing strategy, encoding, accuracy, and, most importantly, exhibit unique configurations of built-up land patterns (and thus, unique combinations of landscape and accuracy metrics), this dataset and the derived landscape metrics represent different joint data distributions than the GHS-BUILT R2018A and the landscape metrics derived for Massachusetts.

2.1.2. Gridded reference data

The reference dataset has been created from the MTBF-33 vector building footprint data (downloaded from https://doi.org/10.17632/w33vbvjtdy). MTBF-33 contains over 6 million building footprint vector geometries annotated with their construction year. For each county in the state of Massachusetts, we selected the MTBF-33 building footprints built-up by 1975, and 2014, respectively, and rasterized the vector data into the GHS-BUILT R2018A grid. To keep resampling uncertainty to a minimum, we first rasterized the vector polygons into a binary grid of 2x2m, and then down-sampled this grid to the target resolution of 30x30m, labeling all 30 m grid cells as "built-up" if they contain at least one 2 m building grid cell. A subset of these gridded surfaces is shown in Figure 1d,e). For the domain adaptation analysis, we carried out the same processes for the Mecklenburg County (i.e. the city of Charlotte, North Carolina) building footprints, but for the year 2016 only (Figure 1f).

2.2. Methods

Herein, we used the pre-processed GHS-BUILT layer as test data and applied the MTBF-33 (Section 2.1) as reference data. Our method consisted of the following steps: 1) Spatially explicit map comparison (Section 2.2.1) and calculation of localized accuracy estimates (Section 2.2.2), 2) the derivation of focal landscape metrics of built-up areas from both the reference and GHS-BUILT data (Section 2.2.3), 3) the correlation analysis of localized accuracy and landscape metrics (Section 2.2.4), 4) regression modeling (Section 2.2.5), and 5) assessing the sensitivity of these results to the spatial support, to the epoch, to the analytical unit, and to the study area (Section 2.2.6). This workflow is shown in Figure 2.

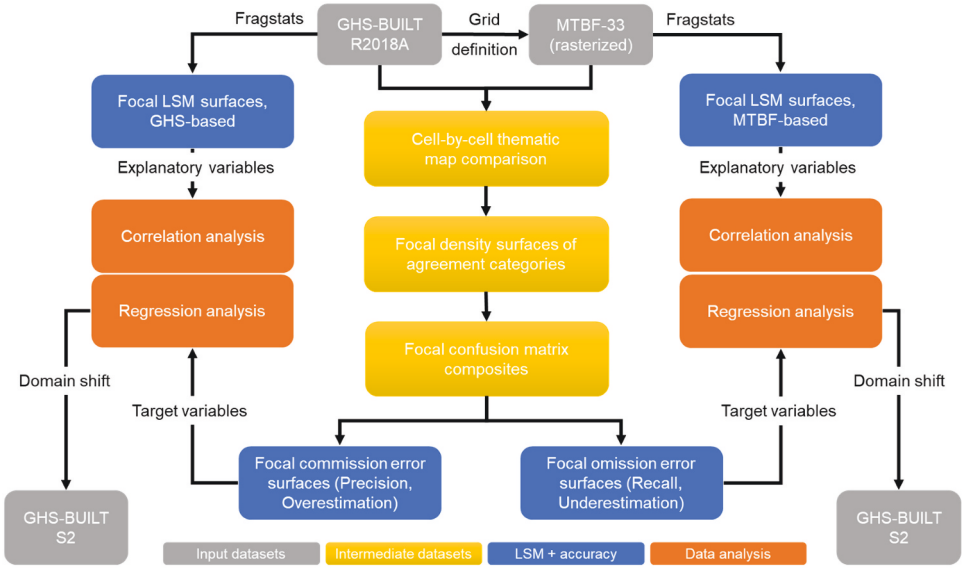

Figure 2. Processing workflow for this study.

2.2.1. Spatially explicit, exhaustive accuracy assessment

Based on the binary built-up presence/absence layers (Figure 1) we applied a method for efficient, spatially explicit accuracy assessment of categorical, gridded data, as proposed in Uhl and Leyk (2022b). This method first performs cell-by-cell map comparison and generates three gridded surfaces, each one containing a 1-hot encoding of one of the three relevant agreement classes (i.e. true positives, false positives, false negatives). Subsequently, the densities of each agreement class within focal windows of varying size (herein called the "spatial support") are calculated. Finally, these agreement class density surfaces are stacked cell-wise to a three-band focal confusion matrix composite, representing the localized (focal) confusion matrix at each location (i.e. grid cell). Moreover, we needed to account for potential effects of positional uncertainty in our data, that may cause misalignment between GHS-BUILT and reference data, and could severely bias the thematic accuracy estimates obtained at the "native" resolution of 30x30m (e.g. Congalton 2007; Gu and Congalton 2020). To mitigate such effects, we down-sampled the binary GHS-BUILT and reference grids to blocks of 3×3 pixels (i.e. corresponding to a resolution of 90x90m) and repeated the steps described above, for the 90x90m grids, as well as for both epochs (i.e. 1975 and 2014). Finally, we expected our focal accuracy estimates to be sensitive to the spatial support (Uhl and Leyk 2022b), and thus, we used focal windows of varying size s (1 km, 2.5 km, 5 km, and 10 km) to compute the agreement class density surfaces. Examples of the resulting confusion matrix composites for the different epochs, different levels of spatial support, and analytical units will be discussed in Section 3.4.

2.2.2. Focal accuracy measures

Based on the focal confusion matrix composites holding the densities of relevant confusion matrix elements TP (true positives), FP (false positives), and FN (false negatives) (see Section 2.2.1), we were able to efficiently calculate localized accuracy estimates at the grid-cell level. We calculated two thematic agreement metrics: Precision and Recall. Recall indicates the probability of a reference element being classified correctly, and is complementary to the omission error OE (error of exclusion), and the precision indicates the probability of a classified object being correct, and is complementary to the commission error CE (error of inclusion) (Story and Congalton 1986):

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

and

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Besides these thematic agreement measures, we used the absolute error (AE), a quantity agreement measure, which is independent from the spatial correspondence of class labels within the focal windows. AE is obtained as:

$$AE = (\mathrm{TP} + \mathrm{FP}) - (TP + FN) = FP - FN \quad (3)$$

Where TP + FP represents the built-up quantity reported in GHS-BUILT, and TP + FN represents the built-up quantity according to the reference data. These built-up quantities (i.e. the amount of 30x30m built-up grid cells per quadratic focal window of size ***s***x***s***, given in meters) can be converted into a measure of built-up surface density (in %, herein called "built-up density"), as follows

$$BUDENS_{REF,s}[\%] = 100 \cdot 30^2 \cdot \frac{(TP + FN)}{s} \quad (4)$$

Where s^2 is the area of the focal window (i.e. the spatial support in m^2). The GHSL-based built-up density is obtained as:

$$BUDENS_{GHSL,s}[\%] = 100 \cdot 30^2 \cdot \frac{(TP + FP)}{s} \quad (5)$$

Herein, we used these built-up density estimates to model the rural-urban continuum (Section 2.2.3). The aforementioned separation of thematic and quantity error has been proposed in a similar way by Pontius and Millones (2011), and allows to measure the agreement between test and reference data quantitatively, while ignoring the thematic agreement at the grid cell level. We used this strategy in regards to coarser-scale applications where the precise locations of built-up surfaces are irrelevant, for example when combining fine-scale built-up surface data with coarser population estimates. Examples of the resulting surfaces of thematic (i.e. precision and recall) and quantity agreement (i.e. AE) will be shown and discussed in Section 3.1. Here it is worth noting that the fourth category used in confusion matrices (i.e. the true negatives) was not calculated as none of the agreement or density metrics requires the true negatives. Generally, for binary comparisons such as built-up vs. not built-up grid cells, it is advised not to use accuracy metrics that include the true negatives, as they are the dominant class, in particular in sparsely built-up, rural areas, and will yield inflated values, e.g. overall accuracy (Rosenfield and Melley 1980; Stehman and Wickham 2020), as shown in the context of built-up land data in previous work (Uhl and Leyk 2022b).

2.2.3. Focal landscape metrics

We used the software FRAGSTATS v4.2 (McGarigal, Cushman, and Ene 2012) to calculate landscape metrics describing the shape and spatial structure of contiguous patches of built-up land within focal regions defined by the four support levels. To keep computational efforts manageable, we computed these metrics for a subset of N = 200,000 locations (i.e. grid cells) within Massachusetts (see Section 2.2.5, Appendix Figure 911). While previous work suggests that particularly the size of patches affects the classification accuracy (Smith et al. 2002, 2003; Klotz et al. 2016; Mück, Klotz, and Taubenböck 2017), we also assumed that certain shape and fragmentation characteristics may drive classification accuracy. Thus, we computed nine landscape level measures and seven patch-level measures, for both built-up areas from the GHSL and the reference data (Table 1). To characterize the distributions of all patch-level measures within the focal windows in Table 1, we calculated mean (MN), area-weighted mean (AM), median (MD), standard deviation (SD), coefficient of variation (CV), and range (RA), summing up to a total of 51 landscape metrics. As shown in Table 1, these commonly used metrics cover a wide range of the morphological, shape, and structure-related characteristics of built-up areas that are assumed to affect the classification accuracy of the GHSL in different ways. For our Charlotte study area, we calculated exhaustive surfaces of focal landscape metrics, using grid and support levels consistent to the focal accuracy surfaces. These focal landscape metrics were derived from both the reference data and the GHS built-up areas for the four levels of spatial support, as landscape metrics may be scale-sensitive (Lustig et al. 2015; Frazier 2022). These surfaces are discussed in Section 3.1. The surfaces of all 51 landscape metrics, derived from the reference data, for the four support levels are shown in Appendix Figure 1022.

2.2.4. Correlation analysis

Using the consistent gridded layers of spatially corresponding focal accuracy estimates and the focal landscape metrics we quantified the correlation between GHS-BUILT data accuracy and the morphological characteristics of the built-up areas. Specifically, we calculated Pearson's correlation coefficient between the landscape metrics and focal accuracy estimates (Section 3.1), as well as between landscape metrics and built-up density, since we assumed that built-up density may be a good proxy for GHS data accuracy,

Table 1. Landscape metrics used in this study include 9 landscape-level measures, and 7 patch-level measures. For each patch-level measure, six summary statistics were computed. Source: McGarigal (2015).

Metric type	Metric name	Short name	Measured characteristic
Landscape/class level	Aggregation Index	AI	Disaggregation
	Landscape Division Index	DIVISION	Segregation
	Landscape Shape Index	LSI	Shape complexity
	Largest Patch Index	LPI	Dominance, connectivity
	Number or Patches	NP	Segregation
	Percentage of Like Adjacencies	PLADJ	Contiguity
	Perimeter-area Fractal Dimension	PAFRAC	Shape complexity
	Edge Density	ED	Compactness, shape complexity, segregation
	Cohesion Index	COHESION	Connectivity
Patch level (MN, AM, MD, SD, CV, RA)	Contiguity Index	CONTIG	Contiguity
	Fractal Index	FRAC	Shape complexity
	Patch Area	AREA	Size
	Perimeter-Area Ratio	PARA	Shape complexity
	Radius of Gyration	GYRATE	Extension
	Related Circumscribing Circle	CIRCLE	Compactness
	Shape index	SHAPE	Shape complexity

as previous work has shown (Leyk et al. 2018; Uhl and Leyk 2022b).

2.2.5. Regression modeling

To further analyze the relationship between landscape metrics and accuracy, we assessed the explanatory power of landscape characteristics to estimate built-up land mapping accuracy across different levels of spatial support. We drew two subsamples of N = 100,000 from the initial sample through random selection, stratified by deciles of $BUDENS_{REF}$ (sample I) and $BUDENS_{GHSL}$ (sample II), respectively (Appendix Figure 1131b,113). This stratification ensured that the drawn samples are equally distributed across the rural-urban continuum and kept computational efforts feasible. For the locations in sample I, we computed focal landscape metrics based on the built-up patches in the reference data, and for the locations in sample II, we used the GHSL-derived built-up patches to compute landscape metrics (Section 2.2.3). As described above, we separately assessed thematic accuracy and quantity agreement. These components were further separated into omission and commission errors. While the reference data alone are independent from the test data (i.e. the GHSL), landscape metrics (LSMs) derived from the reference data (LSM_{REF}) neither contain any information on commission errors in the test data, nor do GHSL-based landscape metrics (LSM_{GHS}) allow for inferring on omission errors with respect to the reference data. For example, a road in an uninhabited region is mistakenly classified as built-up area (i.e. false positive). As the reference data indicates only not built-up grid cells in that region, the landscape metrics derived from the reference data will all be zero or not defined. Thus, the regression model ***precision = f(LSM$_{REF}$)*** will not be able to explain the low precision, since all covariates are zero. In other words, such a regression model could only work in regions where the spatial distributions of test and reference data are similar, but would fail in the areas that matter most, i.e. where extremely high commission error occur. Computing an R^2 of such a scenario would be unfair as we know a priori that the covariates have limited explanatory power.

Thus, we used precision and recall as response variables to be estimated based on the LSM_{GHS} and, LSM_{REF} respectively. Accordingly, we separated the absolute error (Equation 3) into overestimation (OE) and underestimation (UE) components as follows:

$$OE = \begin{cases} AE, AE > 0 \\ 0, AE \leq 0 \end{cases} \tag{6}$$

$$UE = \begin{cases} 0, AE > 0 \\ |AE|, AE \leq 0 \end{cases} \tag{7}$$

Thus, we established four models: (a) Estimating thematic commission error, i.e. the precision of GHSL given the reference data, based on the 51 GHSL-derived landscape metrics:

$$\begin{aligned} Precision_{GHSL \leftarrow REF} &= f(LSM_{GHSL}) \\ &= a_1 \cdot lsm_{GHSL,1} \quad + \ldots + a_{51} \\ &\quad \cdot lsm_{GHSL,51} \end{aligned} \tag{8}$$

(b) Estimating quantity commission error, i.e. the OE of GHSL given the reference data, based on the 51 GHSL-derived landscapemetrics:

$$OE_{GHSL\leftarrow REF} = f(LSM_{GHSL}) = a_1 \cdot lsm_{GHSL,1} + \ldots + a_{51} \cdot lsm_{GHSL,51} \quad (9)$$

(c) Estimating thematic omission error, i.e. the recall of GHSL given the reference data, based on the 51 reference-data derived landscape metrics:

$$Recall_{GHSL\leftarrow REF} = f(LSM_{REF}) = a_1 \cdot lsm_{REF,1} + \ldots + a_{51} \cdot lsm_{REF,51} \quad (10)$$

and (d) Estimating quantity omission error, i.e. the OE of GHSL given the reference data, based on the 51 reference-data derived landscape metrics:

$$UE_{GHSL\leftarrow REF} = f(LSM_{REF}) = a_1 \cdot lsm_{REF,1} + \ldots + a_{51} \cdot lsm_{REF,51} \quad (11)$$

These models were implemented as classical ordinary least squares (OLS) linear regression models as baseline models, and were compared to regression models using an AdaBoost regressor (Freund and Schapire 1997; Drucker 1997), which consists of an ensemble of shallow decision trees ("weak learners"). AdaBoost regression models have shown promising performance in other applications in the geosciences (e.g. Li et al. 2016; Belgiu and Drăguț 2016). This comparison was done in order to test whether complex (albeit black box) machine learning models such as the AdaBoost regressor are necessary to solve the given regression problem, or if classical, and more interpretable statistical models such as OLS are sufficient. All models were tested using these two techniques and separately for the four levels of spatial support, in order to assess cross-scale effects, yielding a total of 32 regression models (Section 3.5). For the AdaBoost regression, we performed hyperparameter tuning separately for each response variable and support level. The outcomes of this analysis will illuminate two questions: (a) What can landscape metrics derived from the reference data tell us about omission errors in built-up land reported in the GHS-BUILT ? (b) Can the GHS-BUILT itself be used to estimate its inherent uncertainty (i.e. commission errors)?

2.2.6. Sensitivity analyses

In our analytical setup, there are four components potentially affecting the performance of the regression models and the drawn conclusions. These components include:

(1) The spatial support of localized accuracy estimates and landscape metrics. To address that, we carried out all regression models based on the four levels of spatial support and compared their results.
(2) The analytical unit (i.e. the grid cell size). As mentioned before, positional uncertainty in our data may cause misalignment between the gridded GHS-BUILT and reference data, and could severely bias the thematic accuracy estimates obtained at the "native" resolution of 30x30m (Congalton 2007). Thus, we also computed the landscape metrics and localized accuracy estimates in coarser, 90x90m grids and carried out the regression analysis accordingly. This step is important because the effect of positional uncertainty on thematic accuracy estimates itself appears to depend on the landscape characteristics (Gu and Congalton 2020).
(3) The epoch or acquisition date. As GHS-BUILT R2018A is a multi-temporal data product (1975–2014) using multispectral data from various generations of the Landsat sensors as input, the relationship between classification accuracy and underlying landscape metrics may vary over time, as the properties and capabilities of the underlying sensors (Landsat MSS, TM, ETM+, OLI) have changed over time. Thus, we also computed the landscape metrics and localized accuracy estimates based on the 1975 GHSL epoch, and on a 1975 snapshot of the MTBF-33 reference data.
(4) The study area and data product. Landscape metrics may be very specific to the settlement patterns in Massachusetts. Moreover, the GHS-BUILT accuracy may be dependent of vegetation types, predominant roof material, and potentially ambiguous spectral responses between built-up and not built-up landscape features. Moreover, cloud cover frequency associated with a specific study area may also affect the GHS accuracy in that area. To account for that, we took two measures: First, we applied our regression models developed using the Massachusetts data to our Charlotte, North Carolina study area. For that study area,

which is also covered by the MTBF-33 reference database, localized accuracy estimates and landscape metrics were derived from the fine-resolution GHS-BUILT-S2 product (see Section 2.1.1). Second, we subdivided our Massachuetts study area into the 14 counties, and established individual regression models for each county. We then measured how well each county-level regression model estimates the focal accuracies in all other counties, in order to test the spatial stationarity of the relationships between landscape metrics and accuracy, Moreover, we applied this concept to the temporal domain, i.e. using regression models trained in 2014 applied to the 1975 epoch, and vice-versa.

In the subsequent analyses, we considered the 2014 epoch and the analytical unit of 30x30m as our baseline scenario.

3. Results and discussion

In this section, we first illustrate the created focal accuracy and landscape metrics surfaces (Section 3.1), and describe the results of the correlation analysis between focal accuracy estimates and landscape metrics (Section 3.2), and the regression-based accuracy models (Section 3.3). We then discuss the sensitivity of correlation and regression analysis to the GHS epoch (1975 and 2014) and to the analytical unit underlying the spatially explicit accuracy assessment and landscape metrics computation (Section 3.4). Then, we describe the results of the domain adaptation analysis, i.e. applying the regression models trained on GHS-BUILT R2018A in Massachusetts to GHS-BUILT-S2 data in North Carolina (Section 3.5). Lastly, we present the results of regression model regionalization to the county-level within Massachusetts to assess spatial variation of the target relationships (Section 3.6).

3.1. Surfaces of agreement metrics and landscape metrics

As discussed in Sections 2.2.1 to 2.2.3, we created a variety of different raster datasets used for the spatially explicit accuracy assessment and all further analyses. All surfaces were created using four levels of spatial support (i.e. focal windows of 1 km, 2.5 km, 5 km, and 10 km) These surfaces include the focal density of confusion matrix elements TP, FP, and FN (Figure 3a), the focal densities of built-up surface derived from both, reference and test data (Figure 3b), the focal accuracy surfaces including precision, recall, and absolute error (Figure 3c). The focal landscape metric surfaces are calculated in the same grid, using the same focal window sizes, derived from the reference data and from the GHSL (see Figures 3d and e for some examples).

3.2. Correlation analysis

As a first step, we systematically analyzed the correlation coefficients between each of the response variables, i.e. thematic commission error (precision), quantity commission error (overestimation), thematic omission error (recall), and quantity omission error (underestimation), and all 51 landscape metrics (i.e. 9 landscape-based measures, and 6 summary statistics for each of the 7 patch-based measures) used as explanatory variables. Figure 4a shows the correlation coefficients for these metrics (for patch-based measures, only the statistic with maximum average correlation across all 16 models is shown, see Figure 1243 for the full matrix, and Figure 1354 for a more detailed visualization of the relationships between landscape metrics and selected accuracy measures). The landscape metrics are sorted by their average correlation to all accuracy metrics at all support levels. We observe the following: In average, the Landscape shape index (LSI) exhibits the highest levels of correlation to the accuracy measures under test. Among the tested landscape metrics, highest levels of correlation were found for the reference-based landscape metrics (LSM_{REF}) and recall, and lowest for UE. In many cases, correlation increased with increasing spatial support. As shown in Figure 4a, contiguity, disaggregation, and connectivity measures of built-up land (AI, PLADJ, COHESION, and CONTIG) are highly correlated with recall, whereas quantity omission errors (i.e. UE) are highly correlated with measures of shape complexity (ED, LSI) but also with scatteredness (i.e. NP, also ED). The GHSL-based landscape metrics (LSM_{GHS}) highly correlated with commission error measures (i.e. precision and OE) are the Shape index (SHAPE), Fractal index (FRAC) and the GYRATE metric. They measure shape complexity and extension, and yield higher values e.g. where irregular road features are present,

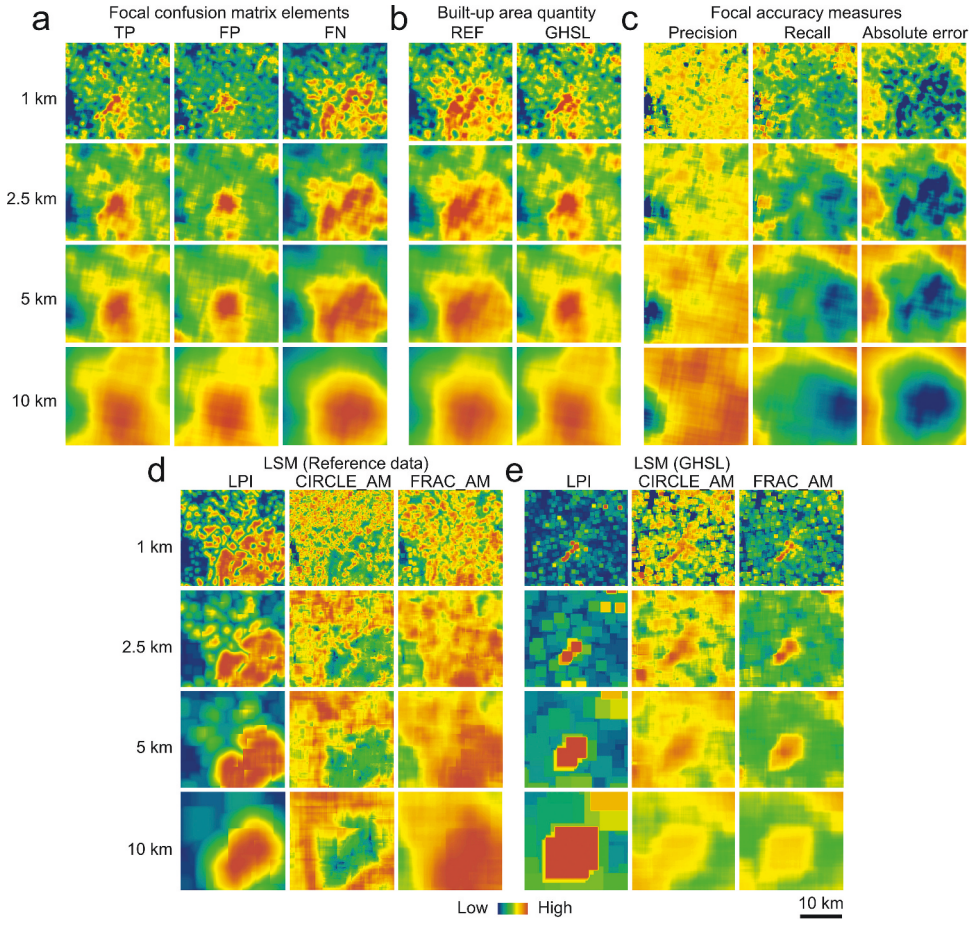

Figure 3. Continuous surfaces used in this study and the effect of spatial support: (a) density surfaces of grid cells for each thematic agreement category (i.e. true positives – TP, false positives – FP, and false negatives – FN), (b) derived measures of built-up quantity (measured in grid cells) derived from the reference data and the test data, (c) surfaces of focal, thematic and quantity agreement measures, (d) selected focal landscape metrics (LSMs) derived from (d) the reference data, and (e) the GHS-BUILT-S2, including the largest patch index (LPI), and the area-weighted mean of the Related Circumscribing Circle metric (CIRCLE_AM) and fractal index (FRAC_AM). The rows represent the different levels of spatial support (i.e. focal window size). All data are shown for a subset of Charlotte, North Carolina and based on the built-up areas shown in Figure 1c and f. A rank transform was applied to the continuous surfaces before color-coding.

which are often incorrectly classified as built-up area (i.e. representing commission errors). Thus, commission errors appear to be associated with the shape of the GHSL built-up areas, whereas omission errors are related to the contiguity and segregation of reference built-up areas.

Moreover, we visualized the 51 landscape metrics and the accuracy components in a two-dimensional space defined by their cross-support correlation trajectory with respect to quantity and thematic commission error measures (i.e. OE and precision, respectively), and quantity and thematic omission error measures (i.e. UE and recall, respectively) (Figure 4b). This way of visualizing the results shows highest correlations between landscape metrics and thematic omission error (i.e. recall) and most of these metrics also are highly correlated to thematic commission errors (i.e. precision). Conversely, some of the quantity agreement measures exhibited higher levels of correlation to OE, but low correlation to UE, indicating that structural properties of built-up areas determining the level of quantity overestimation are different from those that trigger underestimation.

3.3. Regression analysis

Furthermore, we visualized the performance of the AdaBoost regression models generated for each of the four scenarios (Section 2.2.6), separately for

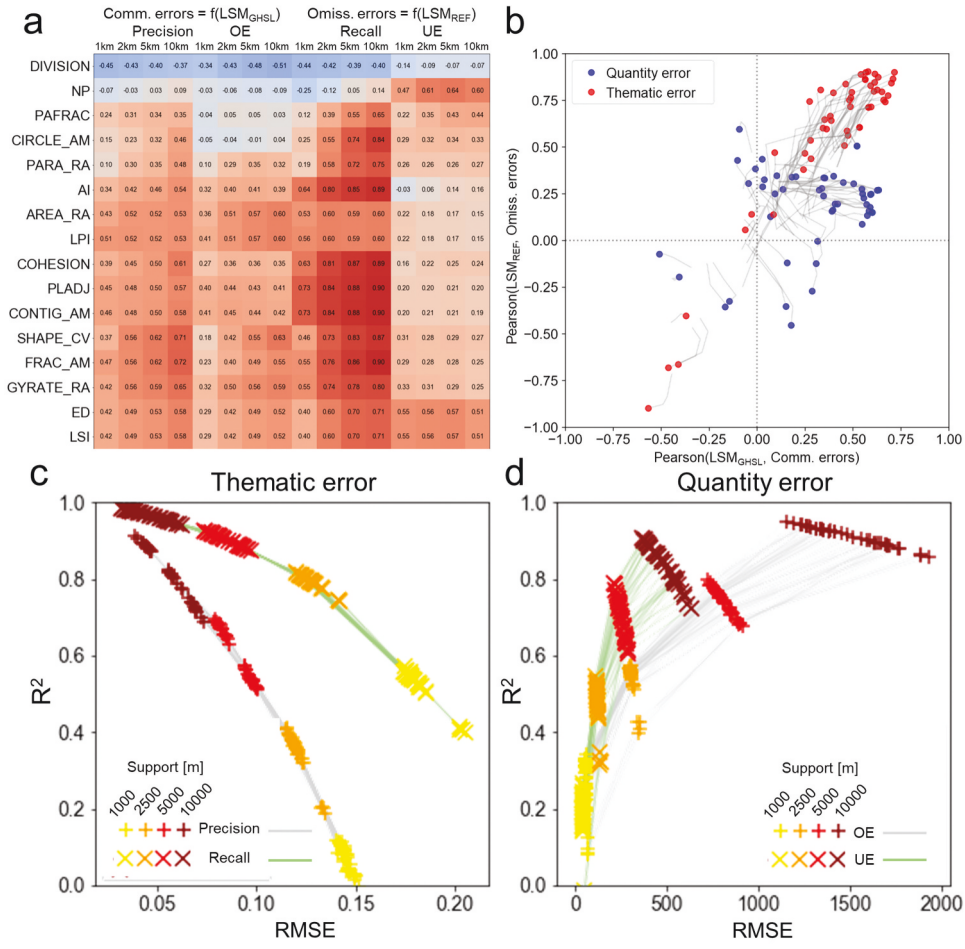

Figure 4. Estimating localized accuracy using regression analysis based on focal landscape metrics. (a) Pearson's correlation coefficients for the 16 most correlated landscape metrics, for each response variable and each level of spatial support. For patch-based metrics, only the summary statistic is shown that yields the highest overall correlation, see Appendix Figure 1133 for the full matrix; LSMs are sorted from top to bottom ascendingly by their average correlation across rows; (b) Correlation coefficients of the 51 landscape metrics in a bi-dimensional space of correlation to measures that characterize commission error (i.e. OE and precision; x-axis) and omission error (i.e. UE and recall; y-axis), color-coded by accuracy type (i.e. thematic or quantity agreement); correlation coefficients are shown for spatial support of 10 km, gray lines illustrate the cross-support trajectory for each LSM across the four levels of spatial support; (c) shows the AdaBoost model performance in bi-dimensional spaces of RMSE and R^2 for the two models estimating thematic errors, with each data point representing a different hyperparameter setting; (d) respective visualization for the two models estimating quantity errors. Lines connect the R^2-RMSE pairs for each hyperparameter combination across the levels of spatial support.

models estimating thematic accuracy (Figure 4c) and quantity agreement (Figure 4d). These visualizations show that for the optimal hyperparameters, all four models yielded R^2 values of >0.9, when using a spatial support of 10 km, indicating that all four accuracy components can be explained reliably based on the landscape metrics, within focal windows of 10 km x 10 km. Visualizing RMSE versus R^2 for each model revealed further that there are increasing levels of R^2 as spatial support increases. The opposite trends between Figure 4 c and d along the x-axis are due to the absolute nature of the OE and UE quantity error components, naturally increasing with increasing spatial support. As can be seen, the best-fitting models were achieved for estimating recall from LSM_{REF} and for estimating OE from LSM_{GHS}.

These observations imply that landscape metrics derived from the GHSL can reliably be employed as explanatory variables for commission errors in the absence of reference data, and the omission error component of the accuracy of built-up land data is highly affected by the level of spatial segregation and contiguity of built-up areas, confirming prior research results (e.g. Smith et al. 2002, 2003; Klotz et al. 2016; Mück, Klotz, and Taubenböck 2017).

Table 2. Regression analysis results for the modeling of GHS accuracy based on landscape metrics, using AdaBoost regression and Ordinary Least Squares.

		AdaBoost Regressor					OLS	
Landscape metrics source data	Response variable	Spatial support [m]	Max. Depth	Num. Estimators	R^2	RMSE	R^2	RMSE
GHSL	OE	1000	10	500	0.342	57.425	0.377	55.198
		2500	25	250	0.575	296.407	0.545	306.091
		5000	25	500	0.800	722.664	0.653	940.203
		10000	25	250	0.949	1151.146	0.751	2531.439
GHSL	Precision	1000	10	500	0.119	0.141	0.129	0.142
		2500	25	250	0.412	0.116	0.342	0.123
		5000	25	500	0.696	0.079	0.447	0.108
		10000	25	500	0.913	0.039	0.539	0.090
Reference data	UE	1000	10	500	0.268	37.740	0.264	37.796
		2500	25	500	0.550	107.468	0.442	120.762
		5000	25	250	0.791	209.211	0.525	315.497
		10000	25	500	0.909	358.610	0.511	838.903
Reference data	Recall	1000	10	250	0.573	0.174	0.561	0.177
		2500	25	500	0.819	0.119	0.781	0.133
		5000	25	500	0.928	0.073	0.857	0.104
		10000	25	500	0.985	0.032	0.903	0.080

Overall, recall appears to exhibit the strongest association to LSMs, and those models exhibit the highest explanatory power. It is also worth noting that while the machine-learning models (AdaBoost) consistently outperform the OLS models in most cases, OLS comes closest to the AdaBoost model performance for estimating recall for large spatial supports (Table 2).

3.4. Sensitivity to epoch and analytical unit

In the focal confusion matrix composites shown for the epochs 1975 and 2014, and for the analytical units of 30x30m and 90x90m (Figure 5), we observe interesting differences in the relative proportions of TP, FP, and FN instances (i.e. grid cells). For example, the RGB-encoding of these relative proportions yields green-yellow colors in the center of the map (i.e. the city of

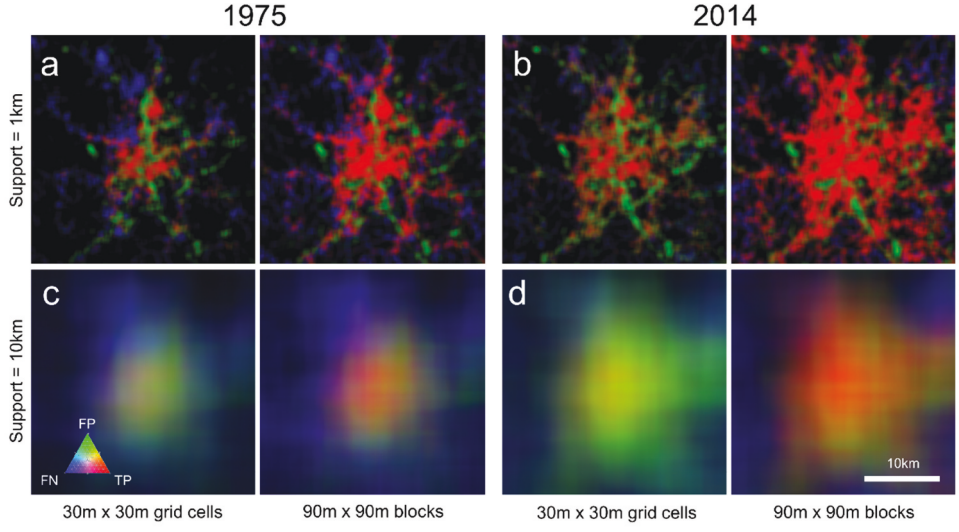

Figure 5. Focal confusion matrix composites for systematically varied parameters used in this study: Top row shows (a) Focal confusion matrix composites derived at a support level of 1x1km for analytical units of 30x30m, and 90x90m in 1975, and (b) in 2014, respectively. The bottom row shows focal confusion matrix composites derived at a support level of 10x10km for analytical units of 30x30m, and pixel blocks of 90x90m, in 1975, and (d) in 2014, respectively. Focal confusion matrix composites are RGB-encoded, i.e. the relative frequencies of the agreement categories are illustrated by the color tones. Specifically, true positives (TP) are represented by the red channel, false positives (FP) by the green channel, and false negatives (FN) by the blue channel. Thus, the colors provide a qualitative insight on the locally "dominating" agreement category and allows to visually detecting regions of high levels of agreement (red) or disagreement (blue for omission, green for commission errors). Data shown for the city of Worcester, Massachusetts, USA (cf. Figure 1).

Worcester, Massachusetts) for the 30 m scenario, and these areas turn red in the 90 m scenario, indicating higher proportions of grid cells switching from false positive to true positive when using a coarser analytical unit. This effect could be due to actual misalignments, which are mitigated by the 90 m aggregation, or could be caused by actual false positives (e.g. roads classified as built-up areas) nearby true positive grid cells. Moreover, in Figure 5 we observe a blue fringe around the city of Worcester in both 1975 scenarios, indicating higher levels of omissions in the GHS-BUILT epoch 1975 in peri-urban areas. These blue color tones are less pronounced in the 2014 scenarios, indicating a decrease of false negatives relative to the other categories (TP, FP).

These observations imply that classification accuracy varies considerably across GHSL epochs, and that the chosen analytical unit likely affects the magnitude of the resulting accuracy measures. How do these sensitivities affect the relationship between accuracy and landscape metrics, as measured by their correlation coefficients (Figure 4a, Appendix Figure 1133)? To shed light on this question, we visualized the correlation coefficients for all landscape and accuracy metrics for the four scenarios (i.e. using epochs 1975 and 2014, respectively, at an analytical unit of 30 m, and 90 m, respectively, see Appendix Figure 1355). While the overall trends seem consistent across these four scenarios, is the ranking of correlation coefficients, and thus the level of association between landscape and accuracy metrics consistent across scenarios? We transformed the correlation coefficients for each scenario in percentile-based ranks and visualized them in Q-Q plots (Figure 6).

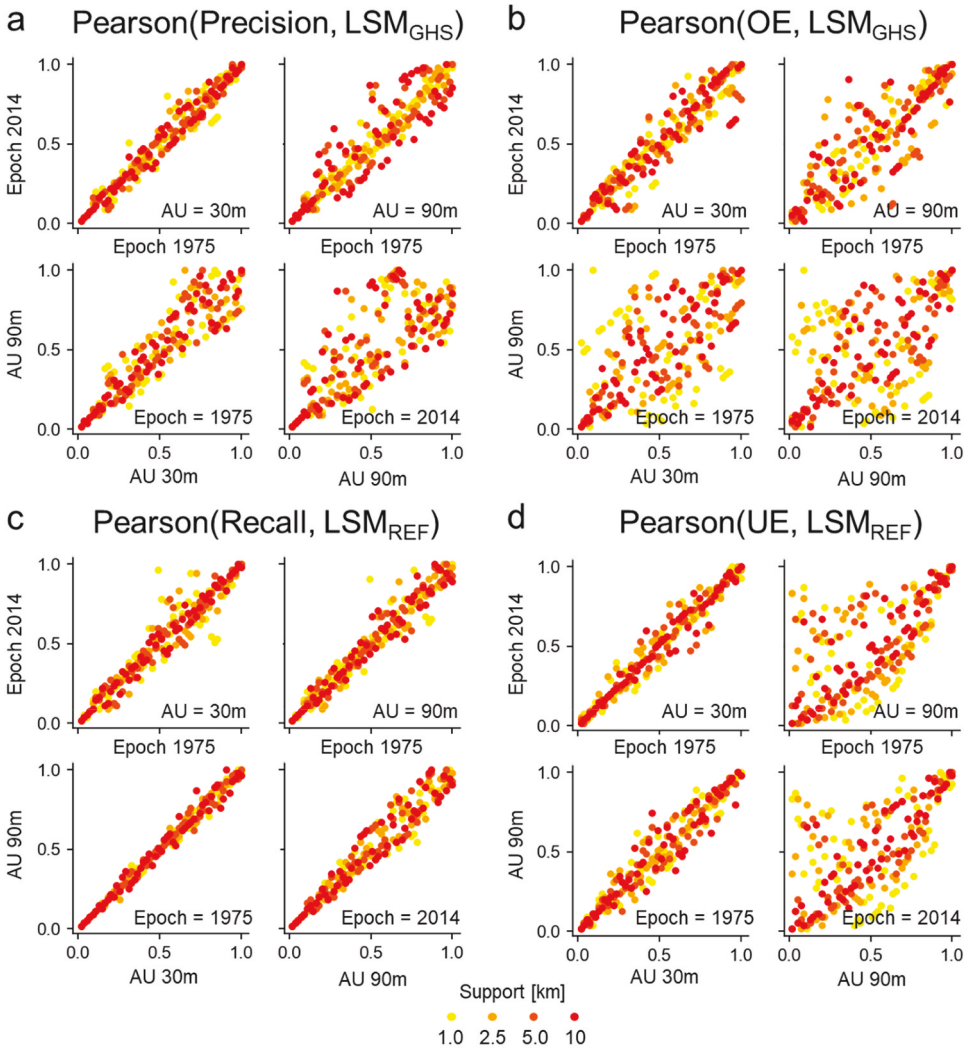

Figure 6. QQ-plots of correlation coefficients between accuracy metrics and LSMs, for the four scenarios i.e. different GHSL epochs (i.e. 1975 and 2014), and different analytical units (AU; i.e. 30 m, 90 m) used for the accuracy assessments.

The more spread the distributions in Figure 6 show, the more does either the epoch or the analytical unit (AU) affect the ranking of correlation coefficients. As can be seen in Figure 6c, the correlation coefficients between Recall and reference-data based landscape metrics experiences the least spread, with the points located nearby the main diagonal, indicating that the order of how strong the associations between specific landscape metrics and the Recall are, is largely independent from the GHSL epoch and from the chosen analytical unit. Conversely, the order of correlation coefficients between overestimation and GHSL-based landscape metrics is most affected by the epoch and analytical unit of the underlying data (Figure 6b).

The observed robustness of the correlation coefficients between individual landscape metrics and the recall in the GHS-BUILT built-up areas across epochs and analytical units is also reflected in the regression analyses carried out for the four scenarios (Table 3). The R^2 values of all regression models are relatively stable across the four scenarios. However, the coefficient of variation across the R^2 values of the OLS regression models that estimate recall using LSM_{REF} are considerably lower than for the other target variables. Importantly, the previously observed trend of increasing model fit with increasing spatial support (i.e. from 1 km toward 10 km) also persists when using the 1975 epoch or accuracy estimates obtained at 90 m analytical units. When looking at the average R^2 values across the models for each of the four scenarios (bottom row of Table 3), we observed, on average, lowest model fits for the 1975 GHSL epoch and using an analytical unit of 90 m. This drop in model fit is most pronounced when using GHS-based landscape metrics, indicating that the estimation of commission errors based on the GHS-BUILT alone is more difficult in 1975 than for the 2014 epoch.

3.5. Domain adaptation analysis

Finally, we investigated how our regression models perfom when deployed on data from a different distribution. This is called domain shift, and models that yield good results when performing a domain shift, are capable of domain adaptation (You et al. 2019). To do so, we applied the models trained on the GHS-BUILT R2018A sample collected in Massachusetts, to a region in Charlotte, North Carolina, where focal accuracy and GHS-based landscape metrics were obtained from the GHS-BUILT-S2 product (see Sections 2.2.1 and 2.2.6). We visually compared three accuracy surfaces: (a) the calculated accuracy surfaces based on map comparison between GHS-BUILT-S2 and the MTBF-33 reference data, (b) the accuracy surfaces as estimated by the regression model trained on Massachusetts data (i.e. domain shift), and (c) the accuracy surfaces as estimated by a regression model trained on 80% of the data based on GHS-BUILT-S2 in the Charlotte study area (i.e. no domain shift). These surfaces are shown in Figure 7, for all target variables, support levels, and for the two regression techniques. As can be seen, the modeled accuracy surfaces using

Table 3. Regression results across the four spatial support levels for GHSL epochs 1975 and 2014, and for analytical units of 30 m and 90 m.

			RMSE per analytical unit and epoch				R^2 per analytical unit and epoch				
LSM source	Spatial support [m]	Accuracy measure	30 m, 2014	30 m, 1975	90 m, 2014	90 m, 1975	30 m, 2014	30 m, 1975	90 m, 2014	90 m, 1975	R^2 Coefficient of variation
GHS	1000	OE	90.999	93.864	9.241	11.622	0.298	0.245	0.191	0.114	0.320
GHS	2500	OE	372.065	390.618	37.702	49.065	0.421	0.362	0.247	0.203	0.284
GHS	5000	OE	1174.696	1214.789	122.756	139.083	0.464	0.427	0.201	0.279	0.313
GHS	10000	OE	833.944	2117.633	190.991	264.154	0.972	0.821	0.802	0.699	0.119
Reference	1000	UE	37.796	41.537	9.402	9.626	0.264	0.188	0.271	0.240	0.136
Reference	2500	UE	120.762	138.817	30.432	25.837	0.442	0.277	0.496	0.471	0.203
Reference	5000	UE	315.497	362.587	83.073	67.511	0.525	0.372	0.589	0.533	0.159
Reference	10000	UE	838.903	895.912	227.732	179.768	0.511	0.443	0.670	0.570	0.152
GHS	1000	Precision	0.128	0.122	0.163	0.196	0.363	0.383	0.298	0.258	0.153
GHS	2500	Precision	0.119	0.115	0.159	0.202	0.515	0.525	0.386	0.303	0.214
GHS	5000	Precision	0.099	0.099	0.134	0.189	0.591	0.580	0.461	0.282	0.259
GHS	10000	Precision	0.063	0.069	0.088	0.129	0.776	0.735	0.659	0.464	0.182
Reference	1000	Recall	0.177	0.184	0.203	0.237	0.561	0.526	0.480	0.391	0.130
Reference	2500	Recall	0.133	0.146	0.152	0.185	0.781	0.734	0.745	0.655	0.063
Reference	5000	Recall	0.104	0.116	0.119	0.147	0.857	0.822	0.841	0.770	0.040
Reference	10000	Recall	0.080	0.089	0.087	0.107	0.903	0.880	0.900	0.860	0.020
Average							0.578	0.520	0.515	0.443	

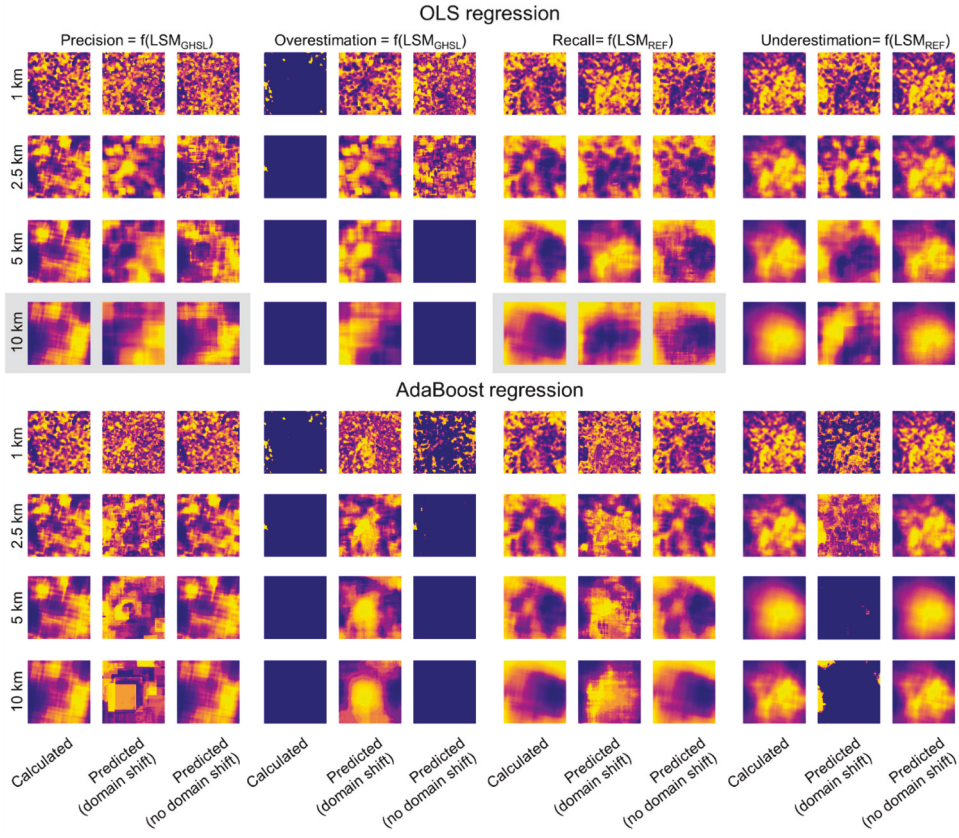

Figure 7. Results of the domain adaptation tests for OLS and AdaBoost regression. Best domain adaptation results are achieved for estimating focal precision and recall using an OLS regression model at 10 km spatial support (highlighted in gray). Values are rank-transformed; high values shown in yellow.

domain shift differ, in many cases, considerably from the calculated surfaces. In some cases (e.g. OLS-based recall modeling at a spatial support of 1 km and 5 km) the resulting surfaces are even inverted, indicating that the underlying relationships between specific landscape metrics and data accuracy may be inverted between the Landsat-based GHS-BUILT R2018A and the GHS-BUILT-S2 product, in the analyzed study area. Importantly, the OLS-based regression models perform the domain shift better at a spatial support of 10 km for most target variables, in particular for the models estimating precision and recall measures (R^2 of 0.42, and 0.46, respectively, highlighted in gray in Figure 7). Poor performance for the overestimation models is due to predominant built-up quantity underestimation in our Charlotte study area, and the resulting sparsity of focal regions where quantity overestimation occurs, impede the successful estimation.

Moreover, we observed that at a spatial support of 10 km, OLS-based models appear to outperform AdaBoost regression models (e.g. three out of four OLS models show – visually – acceptable domain shift results at a support level of 10 km, whereas this is not the case for any of the target variables using AdaBoost regression). This is in contrast to the better model fits of AdaBoost compared to OLS in Table 2, and indicates that the AdaBoost models may be overfitted to the Massachusetts study area, whereas the OLS-based models, despite exhibiting lower levels of model fit in the Massachusetts study area, appear to be more generalizable to other study areas, when the spatial support is large enough.

These results indicate that the morphological landscape characteristics that drive the presence or absence of thematic omission errors are largely identical for the GHS-BUILT R2018A and the GHS-BUILT-S2 product, given that sufficient spatial context is provided.

While the presented analysis focused on the state of Massachusetts, we calculated focal landscape metric surfaces based on the MTBF-33 reference

data for all 33 counties covered by MTBF-33. In previous work, we showed that there are strong associations between GHSL data accuracy and the density of built-up surface within a given spatial unit (Uhl and Leyk 2022b). Thus, we calculated the correlation coefficients between each landscape metric and built-up density for each of the 33 counties, and for three levels of spatial support (i.e. 1 km, 2.5 km, and 5 km, see Appendix Figure 1466). As can be seen, across the three levels of support, most landscape metrics exhibit high positive of negative correlation with built-up density, and these correlations are very consistent across the 33 counties, out of which 19 are located outside of the state of Massachusetts.

3.6. Regional and temporal model generalization

We spatially stratified our data samples by county (see county boundaries in Appendix Figure 2), and established individual regression models for each county, and then calculated the R^2 values of each county-level model when estimating the accuracy in all other counties. The results are a set of cross-tabulated R^2 values for each of the four regression models (Figure 8). Note that we only did this for using OLS regression and for a spatial support of 10 km, as these models showed the best performance in the previously discussed domain adaptation analysis. As can be seen in Figure 8, R^2 values

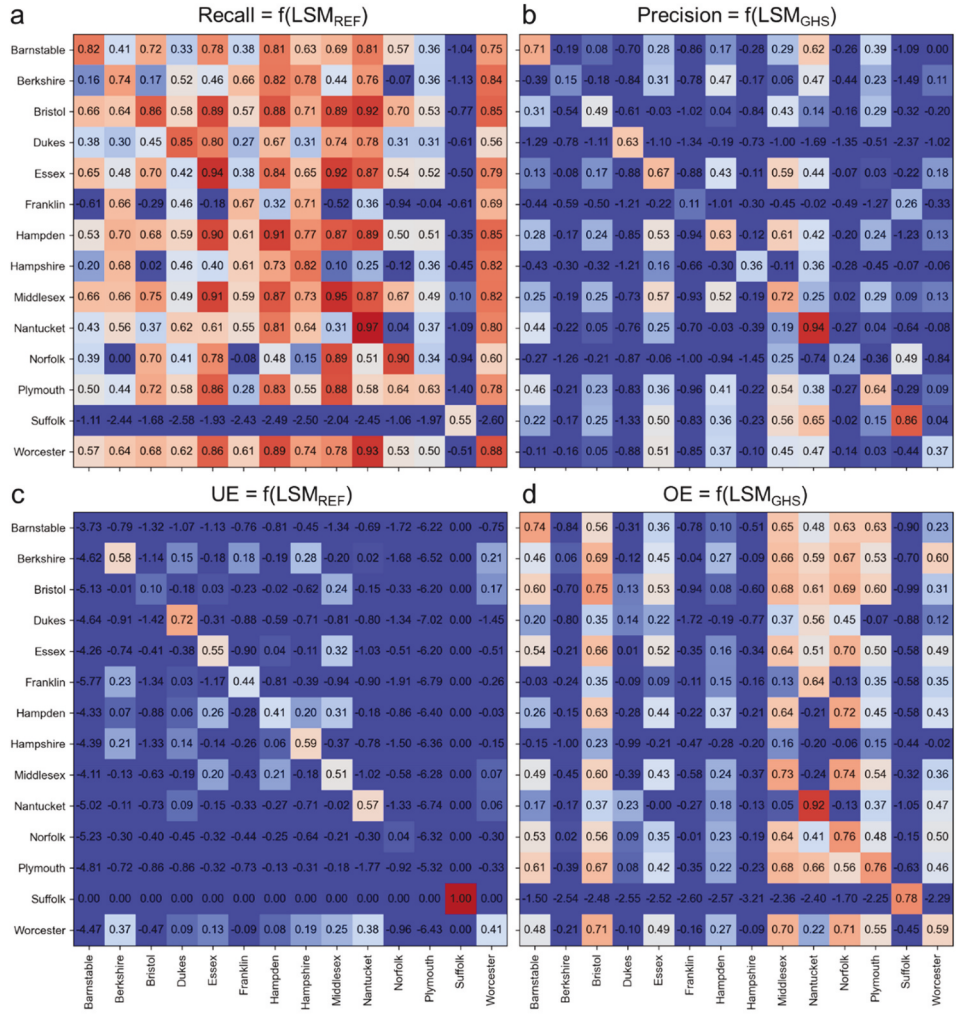

Figure 8. Regression model regionalization. Shown are matrices of R^2 values from OLS regression models estimating accuracy components from landscape metrics, trained on county A (x-axis) and then used for inference in county B (y-axis), for the following accuracy metrics: (a) Recall, (b) Precision, (c) Underestimation component (UE), and (d) Overestimation component (OE). High R^2 values in cells other than on the main diagonal indicate high regression model generalizability between counties.

Table 4. Domain adaptation over time (OLS models, spatial support of 10 km).

train/infer	1975	2014	train/infer	1975	2014
Precision = f(LSM_{GHS})			Recall = f(LSM_{REF})		
1975	0.259	0.296	1975	0.912	0.923
2014	0.101	0.455	2014	0.890	0.942
OE = f(LSM_{GHS})			UE = f(LSM_{REF})		
1975	0.584	0.625	1975	0.208	0.341
2014	0.493	0.672	2014	−0.174	0.347

are generally highest on the main diagonal (i.e. model trained and employed in the same county). These R^2 values are largely in agreement with the R^2 values for the state-level OLS regression models at a spatial support of 10 km (cf. Table 2). Interestingly, off-diagonal R^2 values are low for most accuracy metrics, and for most counties, except for Recall (Figure 8a) and, to some extent, for OE (Figure 8d). This indicates that the relationship between landscape metrics and Recall (Figure 8a) exhibits highest levels of spatial stationarity, i.e. the morphological properties of built-up areas contribute to the level of thematic omission error in similar ways across our study area. A clear exception is Suffolk County, where the city of Boston is located: the relationship between LSMs and recall is not found in any other Massachusetts county. Conversely, the relationship found in Worcester county seems to be most generalizable, possibly because this county contains balanced proportions of urban and rural regions. Thus, while the relationship between accuracy metrics and landscape metrics are highly localized, the recall metric takes an idiosyncratic position, as observed in Figure 8, and in line with the previously discussed findings (e.g. Table 2, Figures 4 and 6), underlining once more the strong and generalizable association between morphological properties of built-up land and the degree to which it is "undermapped."

Finally, we also tested the generalizability of the accuracy-LSM relationship in the temporal domain (i.e. setting up regression models using the 2014 epoch, and estimating accuracy in the 1975 epoch, and vice-versa). We observed similar trends, i.e. higher levels of generalizability over time for the Recall measure, and low levels for the other accuracy metrics (Table 4).

4. Conclusions

In this article, we conducted a detailed assessment of the relationships between morphological characteristics of built-up surfaces (measured by means of landscape metrics), and the data accuracy of built-up areas reported in the gridded, multi-temporal GHS-BUILT R2018A dataset. We identified varying associations between accuracy measures and morphological characteristics of built-up areas, and relatively high explanatory power in the accuracy models, in particular when estimating omission errors from landscape metrics. These findings are useful to determine areas where omission errors are expected to be high, and could be incorporated into classifier training procedures, in order to improve future settlement layers. Moreover, some of the presented regression models could be applied to existing built-up land data, to identify regions where commission errors are expected to be high, in the absence of reference data, and could inform the sampling design of future accuracy assessments.

While the tree-based AdaBoost regressor outperformed the OLS regression models in the "baseline scenario" (i.e. for the epoch 2014, using the full analytical resolution of 30x30m grid cells), our domain adaptation analysis revealed that these AdaBoost models likely overfitted to the Massachusetts study area, as they performed poorly in the "unseen" Charlotte study area. This important insight highlights the importance of domain shift/domain adaptation analyses when evaluating machine learning models. Moreover, the poor performance in our domain adaptation analysis indicates that the relationships between morphology and accuracy of built-up land are highly regional, and not generalizable, except for the Recall metric which exhibits higher levels of generalizability, across regions, and remains largely unaffected by the choice of the underlying analytical unit.

Notably, both correlations and model fits increased with the level of spatial support, indicating that the choice of an appropriate level of spatial support is crucial when creating and analyzing localized accuracy estimates and local landscape metrics. This effect is somewhat expected, and can be attributed to the general case of the Modifiable Areal Unit Problem (MAUP; Openshaw 1984). These trends across different spatial support levels underline the importance of scale-

related considerations in geospatial analyses. However, which level of spatial support is appropriate for a specific purpose needs to be decided for each individual case, taking into account the tradeoff between model robustness (which increases with increasing support level in this study) on the one hand, and loss of spatial granularity on the other hand.

At this point it is important to mention that despite the domain adaptation analysis presented in Section 3.5, further work using a larger set of study areas is required to formalize general guidelines on the effects of landscape characteristics and GHS-BUILT data accuracy. In particular, the temporal gap of two years between the GHS-BUILT-S2 (from 2018) and the MTBF-33 data (from 2016) may introduce a small bias into our domain adaptation analysis. However, as the Charlotte study area is located in the inner part of the city, rather than in a peri-urban area, it has not experienced substantial urban growth between these two years, and thus, we believe that this bias is of minor nature.

Importantly, in this work, we used landscape metrics derived from the test and from the reference data. Thus, the created regression models are of limited use for predictive uncertainty modeling, as the reference data required to generate the explanatory variables (i.e. the landscape metrics) could also be used to perform the accuracy assessment by map comparison rather than using the predictive model. In future work, we will also test the use of completely independent explanatory variables (e.g. land cover data, census data) for the purpose of predictive uncertainty modeling. An important limitation here is that the spatial support of such predictive models needs to be large enough (e.g. 10x10km), as we observed rather weak associations at lower levels of spatial support (e.g. 1x1km). However, having accuracy surfaces based on a support level of 10x10km is still an improvement over simple global accuracy estimates neglecting spatial accuracy variations, as still many studies do.

In future work, we will also focus on the application of the described framework to different built-up surface/settlement data products and we will analyze in detail the sensitivity of landscape metrics to spatial support, taking into account potential bias introduced by the scale sensitivity of the landscape metrics themselves (see Lustig et al. 2015). While the relationships between landscape characteristics and data accuracy have been studied in the case of land cover data in general (Smith et al. 2002, 2003), and, in the case of built-up land data (Klotz et al. 2016; Mück, Klotz, and Taubenböck 2017), this work demonstrated at unprecedented depth, that the accuracy of remote-sensing derived built-up land data products such as the GHS-BUILT is affected by the morphology of the built-up area patterns, but differently for commission and omission error components. Concluding, this work contributes to a better understanding of the spatial structure and variation of the uncertainty inherent in data products such as the GHS-BUILT R2018A, and ultimately, to a more informed and reflected use of such data products.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development [2P2CHD066613-06]; NSF [1924670,2121976]; Open Philantropy Project [1433]. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

ORCID

Johannes H. Uhl http://orcid.org/0000-0002-4861-5915

Data availability

Focal landscape metrics and focal accuracy estimates computed for the epochs 1975 and 2014, for analytical units of 30x30m and 90x90m, as well as for the four levels of spatial support are available at https://doi.org/10.6084/m9.figshare.19785877. Python code for spatially explicit accuracy assessments of binary, gridded datasets is available at https://github.com/johannesuhl/local_accuracy.

References

Akosa, J. 2017. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. In Proceedings of the SAS Global Forum, Orlando, FL, USA: 2–5.

Balk, D., F. Pozzi, G. Yetman, U. Deichmann, and A. Nelson. 2005. The Distribution of People and the Dimension of Place: Methodologies to Improve the Global Estimation of

Urban Extents. In International Society for Photogrammetry and Remote Sensing, proceedings of the urban remote sensing conference, Tempe, AZ, USA: 1–14.

Belgiu, M., and L. Drăguţ. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

Bujang, M. A., and N. Baharum. 2017. "Guidelines of the Minimum Sample Size Requirements for Kappa Agreement Test." *Epidemiology, Biostatistics and Public Health* 14 (2): 1–10.

Champagne, C., H. McNairn, B. Daneshfar, and J. Shang. 2014. "A Bootstrap Method for Assessing Classification Accuracy and Confidence for Agricultural Land Use Mapping in Canada." *International Journal of Applied Earth Observation and Geoinformation* 29: 44–52. doi:10.1016/j.jag.2013.12.016.

Cheng, K. S., J. Y. Ling, T. W. Lin, Y. T. Liu, Y. C. Shen, and Y. Kono. 2021. "Quantifying Uncertainty in Land-Use/Land-Cover Classification Accuracy: A Stochastic Simulation Approach." *Frontiers in Environmental Science* 9: 46. doi:10.3389/fenvs.2021.628214.

Comber, A., P. Fisher, C. Brunsdon, and A. Khmag. 2012. "Spatial Analysis of Remote Sensing Image Classification Accuracy." *Remote Sensing of Environment* 127: 237–246. doi:10.1016/j.rse.2012.09.005.

Congalton, R. G. 1988. "A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data." *Photogrammetric Engineering and Remote Sensing* 54 (5): 593–600.

Congalton, R. G. 2007. Thematic and Positional Accuracy Assessment of Digital Remotely Sensed Data. In: R. E. McRoberts, G. A. Reams, P. C. Van Deusen, and W. H. McWilliams, eds. Proceedings of the seventh annual forest inventory and analysis symposium; October 3-6, 2005; Portland, ME. Gen. Tech. Rep. WO-77. Washington, DC: US Department of Agriculture, Forest Service: 149–154.

Corbane, C., M. Pesaresi, T. Kemper, P. Politis, A. J. Florczyk, V. Syrris, P. Soille, F. Sabo, and P. Soille. 2019. "Automated Global Delineation of Human Settlements from 40 Years of Landsat Satellite Data Archives." *Big Earth Data* 3 (2): 140–169. doi:10.1080/20964471.2019.1625528.

Corbane, C., V. Syrris, F. Sabo, P. Politis, M. Melchiorri, M. Pesaresi, P. Soille, and T. Kemper. 2021. "Convolutional Neural Networks for Global Human Settlements Mapping from Sentinel-2 Satellite Imagery." *Neural Computing & Applications* 33 (12): 6697–6720. doi:10.1007/s00521-020-05449-7.

Cyriac, S., and M. Firoz C. 2022. "Dichotomous Classification and Implications in Spatial Planning: A Case of the Rural-Urban Continuum Settlements of Kerala, India." *Land Use Policy* 114: 105992. doi:10.1016/j.landusepol.2022.105992.

Degen, A., C. Corbane, M. Pesaresi, and T. Kemper. 2018. *A Statistical Analysis of the Relationship between Landscape Heterogeneity and the Quantization of Remote Sensing Data, EUR 29340 EN, JRC111818*. Luxembourg: Publications Office of the European Union.

Drucker, H. 1997. "Improving Regressors Using Boosting Techniques." *Icml* 97: 107–115.

Ebrahimy, H., B. Mirbagheri, A. A. Matkan, and M. Azadbakht. 2021. "Per-pixel Land Cover Accuracy Prediction: A Random forest-based Method with Limited Reference Sample Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 172: 17–27. doi:10.1016/j.isprsjprs.2020.11.024.

Ehrlich, D., S. Freire, M. Melchiorri, and T. Kemper. 2021. "Open and Consistent Geospatial Data on Population Density, Built-Up and Settlements to Analyse Human Presence, Societal Impact and Sustainability: A Review of GHSL Applications." *Sustainability* 13 (14): 7851. doi:10.3390/su13147851.

FGDC (Federal Geographic Data Committee). 1998. *Geospatial Positioning Accuracy Standards - Part 3: National Standard for Spatial Data Accuracy*. Washington, DC: Federal Geographic Data Committee.

Fielding, A. H., and J. F. Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation presence/absence Models." *Environmental Conservation* 24 (1): 38–49. doi:10.1017/S0376892997000088.

Florczyk, A. J., C. Corbane, D. Ehrlich, S. Freire, T. Kemper, L. Maffenini, M. Melchiorri, et al. 2019. *GHSL Data Package 2019, EUR 29788 EN*. Luxembourg: Publications Office of the European Union.

Foody, G. M. 2007. "Local Characterization of Thematic Classification Accuracy through Spatially Constrained Confusion Matrices." *International Journal of Remote Sensing* 26 (6): 1217–1228. doi:10.1080/01431160512331326521.

Foody, G. M. 2009. "Sample Size Determination for Image Classification Accuracy Assessment and Comparison." *International Journal of Remote Sensing* 30 (20): 5273–5291. doi:10.1080/01431160903130937.

Frazier, A. E. 2022. "Scope and Its Role in Advancing a Science of Scaling in Landscape Ecology." *Landscape Ecology* 22: 1–7.

Freund, Y., and R. E. Schapire. 1997. "A decision-theoretic Generalization of on-line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1): 119–139. doi:10.1006/jcss.1997.1504.

Gong, P., X. Li, J. Wang, Y. Bai, B. Chen, T. Hu, X. Liu, et al. 2020. "Annual Maps of Global Artificial Impervious Area (GAIA) between 1985 and 2018." *Remote Sensing of Environment* 236: 111510. doi:10.1016/j.rse.2019.111510.

Gu, J., and R. G. Congalton. 2020. "Analysis of the Impact of Positional Accuracy When Using a Single Pixel for Thematic Accuracy Assessment." *Remote Sensing* 12 (24): 4093. doi:10.3390/rs12244093.

Hashemian, M. S., A. A. Abkar, and S. B. Fatemi. 2004. "Study of Sampling Methods for Accuracy Assessment of Classified Remotely Sensed Data." In International congress for photogrammetry and remote sensing, Istanbul, Turkey: 1682–1750.

Kaim, D., E. Ziółkowska, S. R. Grădinaru, and R. Pazúr. 2022. "Assessing the Suitability of urban-oriented Land Cover Products for Mapping Rural Settlements." *International Journal of Geographical Information Science* 1–15. doi:10.1080/13658816.2022.2075877.

Khatami, R., G. Mountrakis, and S. V. Stehman. 2017. "Mapping per-pixel Predicted Accuracy of Classified Remote Sensing Images." *Remote Sensing of Environment* 191: 156–167. doi:10.1016/j.rse.2017.01.025.

Klotz, M., T. Kemper, C. Geiß, T. Esch, and H. Taubenböck. 2016. "How Good Is the Map? A multi-scale cross-comparison Framework for Global Settlement Layers: Evidence from Central Europe." *Remote Sensing of Environment* 178: 191–212. doi:10.1016/j.rse.2016.03.001.

Kyriakidis, P. C., and J. L. Dungan. 2001. "A Geostatistical Approach for Mapping Thematic Classification Accuracy and Evaluating the Impact of Inaccurate Spatial Data on Ecological Model Predictions." *Environmental and Ecological Statistics* 8 (4): 311–330. doi:10.1023/A:1012778302005.

Leyk, S., J. H. Uhl, D. Balk, and B. Jones. 2018. "Assessing the Accuracy of multi-temporal built-up Land Layers across rural-urban Trajectories in the United States." *Remote Sensing of Environment* 204: 898–917. doi:10.1016/j.rse.2017.08.035.

Leyk, S., and N. E. Zimmermann. 2004. A Predictive Uncertainty Model for field-based Survey Maps Using Generalized Linear Models. International Conference on Geographic Information Science, Adelphi, Maryland, USA: 191–205.

Leyk, S., and N. E. Zimmermann. 2007. "Improving Land Change Detection Based on Uncertain Survey Maps Using Fuzzy Sets." *Landscape Ecology* 22 (2): 257–272. doi:10.1007/s10980-006-9021-2.

Li, M., L. Ma, T. Blaschke, L. Cheng, and D. Tiede. 2016. "A Systematic Comparison of Different object-based Classification Techniques Using High Spatial Resolution Imagery in Agricultural Environments." *International Journal of Applied Earth Observation and Geoinformation* 49: 87–98. doi:10.1016/j.jag.2016.01.011.

Liu, F., S. Wang, Y. Xu, Q. Ying, F. Yang, Y. Qin, and S. Fu. 2020. "Accuracy Assessment of Global Human Settlement Layer (GHSL) built-up Products over China." *Plos one* 15 (5): e0233164. doi:10.1371/journal.pone.0233164.

Löw, F., U. Michel, S. Dech, and C. Conrad. 2013. "Impact of Feature Selection on the Accuracy and Spatial Uncertainty of per-field Crop Classification Using Support Vector Machines." *ISPRS Journal of Photogrammetry and Remote Sensing* 85: 102–119. doi:10.1016/j.isprsjprs.2013.08.007.

Lustig, A., D. B. Stouffer, M. Roigé, and S. P. Worner. 2015. "Towards More Predictable and Consistent Landscape Metrics across Spatial Scales." *Ecological Indicators* 57: 11–21. doi:10.1016/j.ecolind.2015.03.042.

Marconcini, M., N. Gorelick, A. Metz-Marconcini, and T. Esch. 2020a. Accurately Monitoring Urbanization at Global scale–the World Settlement Footprint. In IOP Conference Series: Earth and Environmental Science, Florence, Italy (Vol. 509, No. 1, p. 012036); IOP Publishing.

Marconcini, M., A. Metz-Marconcini, S. Üreyen, D. Palacios-Lopez, W. Hanke, F. Bachofer, … M. Paganini. 2020b. "Outlining Where Humans Live, the World Settlement Footprint 2015." *Scientific Data* 7 (1): 1–14. doi:10.1038/s41597-020-00580-5.

McGarigal, K. 2015. FRAGSTATS Help. Accessible online at: https://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf. Accessed on 04 October 2020

McGarigal, K., S. Cushman, and E. Ene. 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Accessible online at: http://www.umass.edu/landeco/research/fragstats/fragstats.html.

Mei, Y., J. Zhang, W. Zhang, and F. Liu. 2019. "A Composite Method for Predicting Local Accuracies in Remotely Sensed Land-Cover Change Using Largely Non-Collocated Sample Data." *Remote Sensing* 11 (23): 2818. doi:10.3390/rs11232818.

Mitchell, P. J., A. L. Downie, and M. Diesing. 2018. "How Good Is My Map? A Tool for semi-automated Thematic Mapping and Spatially Explicit Confidence Assessment." *Environmental Modelling and Software* 108: 111–122. doi:10.1016/j.envsoft.2018.07.014.

Morales-Barquero, L., M. B. Lyons, S. R. Phinn, and C. M. Roelfsema. 2019. "Trends in Remote Sensing Accuracy Assessment Approaches in the Context of Natural Resources." *Remote Sensing* 11 (19): 2305. doi:10.3390/rs11192305.

Mück, M., M. Klotz, and H. Taubenböck. 2017. Validation of the DLR Global Urban Footprint in Rural Areas: A Case Study for Burkina Faso. In 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates (pp. 1-4); IEEE.

Openshaw, S. 1984. "The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography."

Pesaresi, M., G. Huadong, X. Blaes, D. Ehrlich, S. Ferri, L. Gueguen, M. Halkia, et al. 2013. "A Global Human Settlement Layer from Optical HR/VHR RS Data: Concept and First Results." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6 (5): 2102–2131. doi:10.1109/JSTARS.2013.2271445.

Pontius, R. G., and M. L. Cheuk. 2006. "A Generalized Cross-tabulation Matrix to Compare Soft-classified Maps at Multiple Resolutions." *International Journal of Geographical Information Science* 20 (1): 1–30. doi:10.1080/13658810500391024.

Pontius, R. G., and M. Millones. 2011. "Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment." *International Journal of Remote Sensing* 32 (15): 4407–4429. doi:10.1080/01431161.2011.552923.

Pontius, R. G., and B. Suedmeyer. 2004. "Components of Agreement between Categorical Maps at Multiple Resolutions." In *Remote Sensing and GIS Accuracy Assessment*, edited by Lunetta, Ross S., Lyon, John G., 233–251. Boca Raton, Florida: CRC Press.

Radoux, J., F. Waldner, and P. Bogaert. 2020. "How Response Designs and Class Proportions Affect the Accuracy of Validation Data." *Remote Sensing* 12 (2): 257. doi:10.3390/rs12020257.

Rosenfield, G., and M. Melley. 1980. "Applications of Statistics to Thematic Mapping." *Photogrammetric Engineering and Remote Sensing* 46: 1287–1294.

See, L., I. Georgieva, M. Duerauer, T. Kemper, C. Corbane, L. Maffenini, J. Gallego, et al. 2022. "A Crowdsourced

Global Data Set for Validating built-up Surface Layers." *Scientific Data* 9 (1): 1–14. doi:10.1038/s41597-021-01105-4.

Shao, G., L. Tang, and J. Liao. 2019. "Overselling Overall Map Accuracy Misinforms about Research Reliability." *Landscape Ecology* 34 (11): 2487–2492. doi:10.1007/s10980-019-00916-6.

Sim, J., and C. C. Wright. 2005. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements." *Physical Therapy* 85 (3): 257–268. doi:10.1093/ptj/85.3.257.

Smith, J. H., S. V. Stehman, J. D. Wickham, and L. Yang. 2003. "Effects of Landscape Characteristics on land-cover Class Accuracy." *Remote Sensing of Environment* 84 (3): 342–349. doi:10.1016/S0034-4257(02)00126-8.

Smith, J. H., J. D. Wickham, S. V. Stehman, and L. Yang. 2002. "Impacts of Patch Size and land-cover Heterogeneity on Thematic Image Classification Accuracy." *Photogrammetric Engineering and Remote Sensing* 68: 65–70.

Steele, B. M., J. C. Winne, and R. L. Redmond. 1998. "Estimation and Mapping of Misclassification Probabilities for Thematic Land Cover Maps." *Remote Sensing of Environment* 66 (2): 192–202. doi:10.1016/S0034-4257(98)00061-3.

Stehman, S. V. 2009. "Sampling Designs for Accuracy Assessment of Land Cover." *International Journal of Remote Sensing* 30 (20): 5243–5272. doi:10.1080/01431160903131000.

Stehman, S. V., and G. M. Foody. 2019. "Key Issues in Rigorous Accuracy Assessment of Land Cover Products." *Remote Sensing of Environment* 231: 111199. doi:10.1016/j.rse.2019.05.018.

Stehman, S. V., and J. D. Wickham. 2011. "Pixels, Blocks of Pixels, and Polygons: Choosing a Spatial Unit for Thematic Accuracy Assessment." *Remote Sensing of Environment* 115 (12): 3044–3055. doi:10.1016/j.rse.2011.06.007.

Stehman, S. V., and J. Wickham. 2020. "A Guide for Evaluating and Reporting Map Data Quality: Affirming Shao Et Al."." *Overselling Overall Map Accuracy Misinforms about Research Reliability". Landscape Ecology* 35 (6): 1263–1267.

Story, M., and R. G. Congalton. 1986. "Accuracy Assessment - a Users Perspective." *Photogrammetric Engineering and Remote Sensing* 52 (3): 397–399.

Strahler, A. H., L. Boschetti, G. M. Foody, M. A. Friedl, M. C. Hansen, M. Herold, P. Mayaux, J. T. Morisette, S. V. Stehman, and C. E. Woodcock. 2006. "Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps." *European Communities, Luxembourg* 51 (4): 1–60.

Tsutsumida, N., and A. J. Comber. 2015. "Measures of spatio-temporal Accuracy for Time Series Land Cover Data." *International Journal of Applied Earth Observation and Geoinformation* 41: 46–55. doi:10.1016/j.jag.2015.04.018.

Uhl, J. H., and S. Leyk. 2022a. "MTBF-33: A multi-temporal Building Footprint Dataset for 33 Counties in the United States (1900-2015)." *Data in Brief* 43: 108369. doi:10.1016/j.dib.2022.108369.

Uhl, J. H., and S. Leyk. 2022b. "A scale-sensitive Framework for the Spatially Explicit Accuracy Assessment of Binary built-up Surface Layers." *Remote Sensing of Environment* 279: 113117. doi:10.1016/j.rse.2022.113117.

van Oort, P. A., A. K. Bregt, S. de Bruin, A. J. de Wit, and A. Stein. 2004. "Spatial Variability in Classification Accuracy of Agricultural Crops in the Dutch National land-cover Database." *International Journal of Geographical Information Science* 18 (6): 611–626. doi:10.1080/13658810410001701969.

Vizzari, M. 2011. Spatio-temporal Analysis Using urban-rural Gradient Modelling and Landscape Metrics. In International Conference on Computational Science and Its Applications (pp. 103-118). Berlin, Heidelberg; Springer.

Vizzari, M., and M. Sigura. 2013. "Urban-rural Gradient Detection Using Multivariate Spatial Analysis and Landscape Metrics." *Journal of Agricultural Engineering* 44 (s2). doi:10.4081/jae.2013.333.

Waldner, F., M. C. Hansen, P. V. Potapov, F. Löw, T. Newby, S. Ferreira, P. Defourny, and K. P. Vadrevu. 2017. "National-scale Cropland Mapping Based on spectral-temporal Features and Outdated Land Cover Information." *PloS one* 12 (8): e0181911. doi:10.1371/journal.pone.0181911.

Wardlow, B. D., and K. Callahan. 2014. "A multi-scale Accuracy Assessment of the MODIS Irrigated Agriculture data-set (Mirad) for the State of Nebraska, USA." *GIScience and Remote Sensing* 51 (5): 575–592. doi:10.1080/15481603.2014.952546.

Wickham, J. D., S. V. Stehman, J. A. Fry, J. H. Smith, and C. G. Homer. 2010. "Thematic Accuracy of the NLCD 2001 Land Cover for the Conterminous United States." *Remote Sensing of Environment* 114 (6): 1286–1296. doi:10.1016/j.rse.2010.01.018.

Wickham, J., S. V. Stehman, and C. G. Homer. 2018. "Spatial Patterns of the United States National Land Cover Dataset (NLCD) land-cover Change Thematic Accuracy (2001–2011)." *International Journal of Remote Sensing* 39 (6): 1729–1743. doi:10.1080/01431161.2017.1410298.

Ye, S., R. G. Pontius, and R. Rakshit. 2018. "A Review of Accuracy Assessment for object-based Image Analysis: From per-pixel to per-polygon Approaches." *ISPRS Journal of Photogrammetry and Remote Sensing* 141: 137–147. doi:10.1016/j.isprsjprs.2018.04.002.

You, K., M. Long, Z. Cao, J. Wang, and M. I. Jordan. 2019. Universal Domain Adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA: 2720–2729.

Zhang, J., and Y. Mei. 2016. "Integrating Logistic Regression and Geostatistics for user-oriented and uncertainty-informed Accuracy Characterization in remotely-sensed Land Cover Change Information." *ISPRS International Journal of Geo-Information* 5 (7): 113. doi:10.3390/ijgi5070113.

Zhu, L., P. Xiao, X. Feng, Z. Wang, and L. And Jiang. 2013. "Multi-scale Accuracy Assessment of Land Cover Datasets Based on histo-variograms." *Journal of Remote Sensing* 17 (6): 1–8.

Appendix

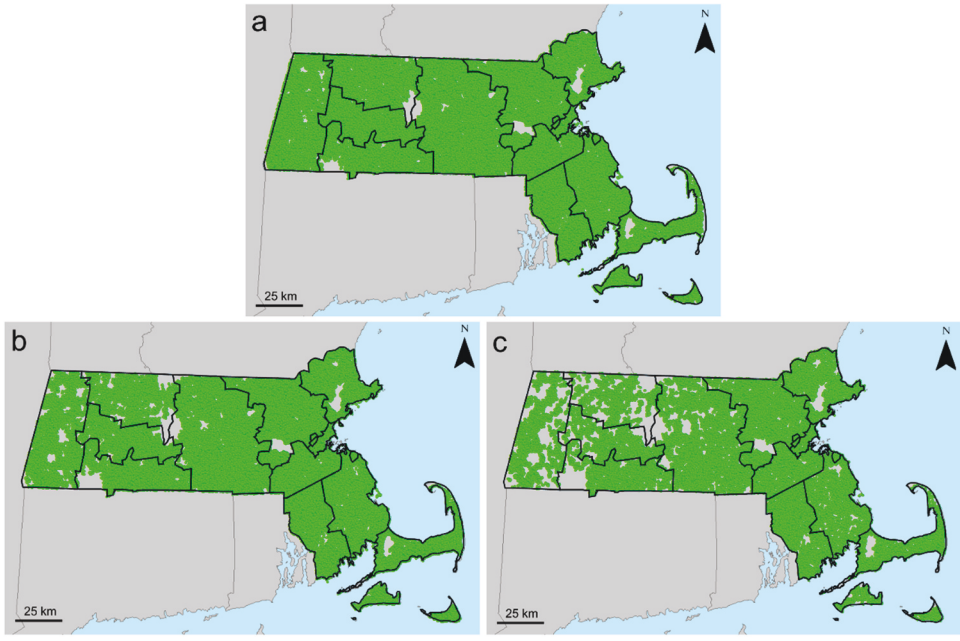

Figure A1. Sampling locations (a) (N = 200,000) in the state of Massachusetts at which focal landscape metrics were computed, and the subsamples at which regression analysis was carried out for (b) reference data based landscape metrics, and (c) GHSL-based landscape metrics, of sample size N = 100,000 each. Black lines represent the boundaries of the 14 counties in Massachusetts.

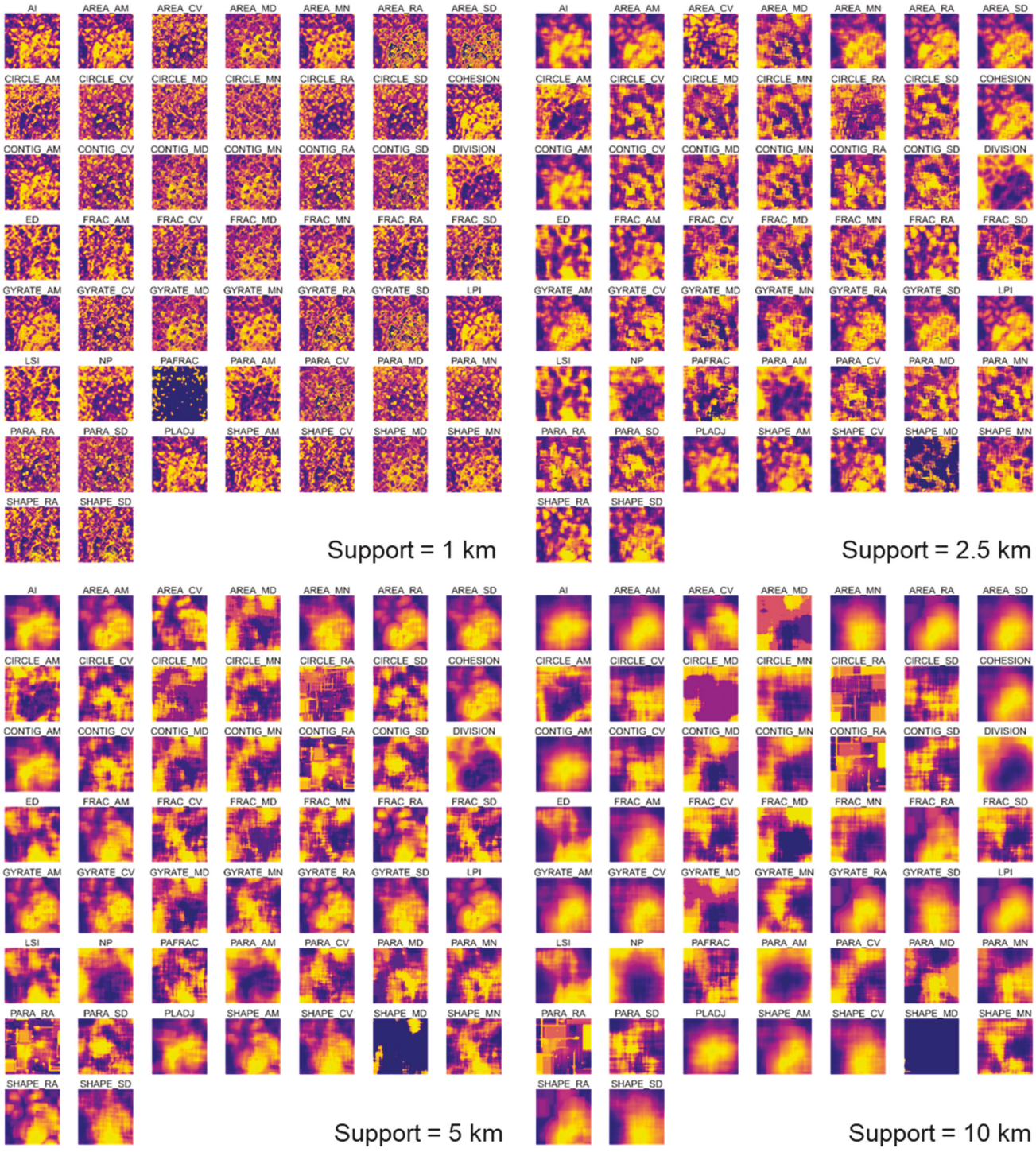

Figure A2. Exhaustive focal landscape metric surfaces at various levels of spatial support, shown for the city of Charlotte, North Carolina. Values are rank-transformed; high values shown in yellow.

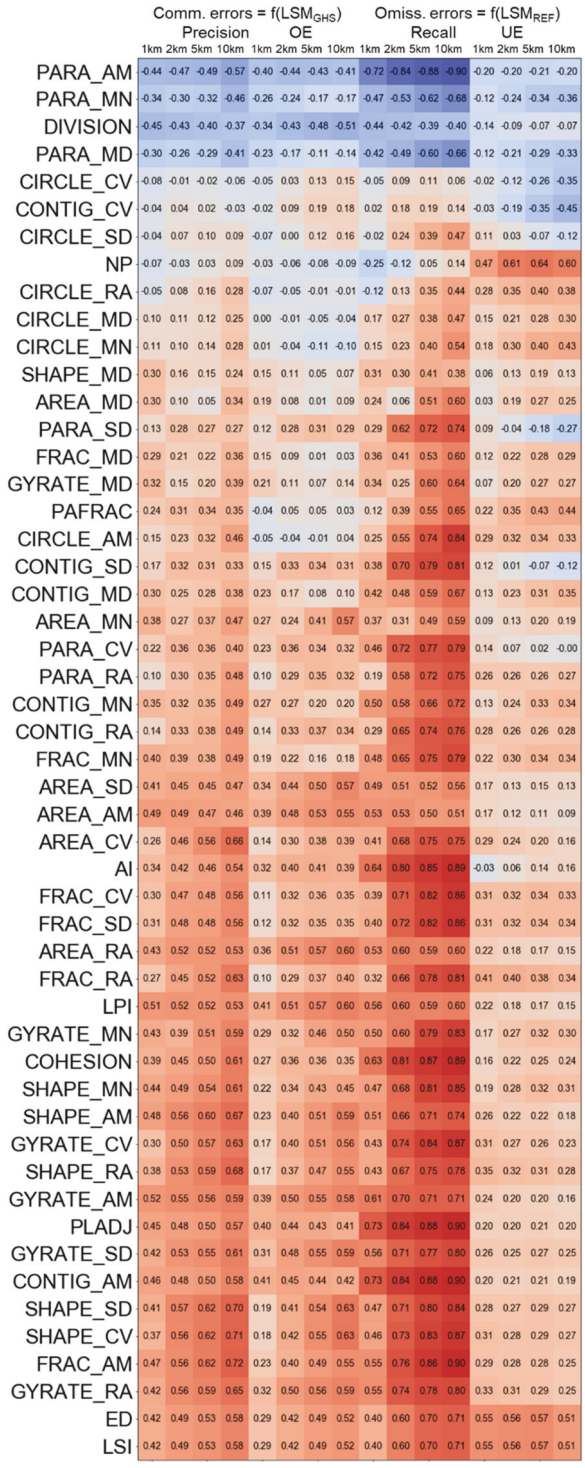

Figure A3. Pearson's correlation coefficients between all 51 landscape metrics and accuracy components.

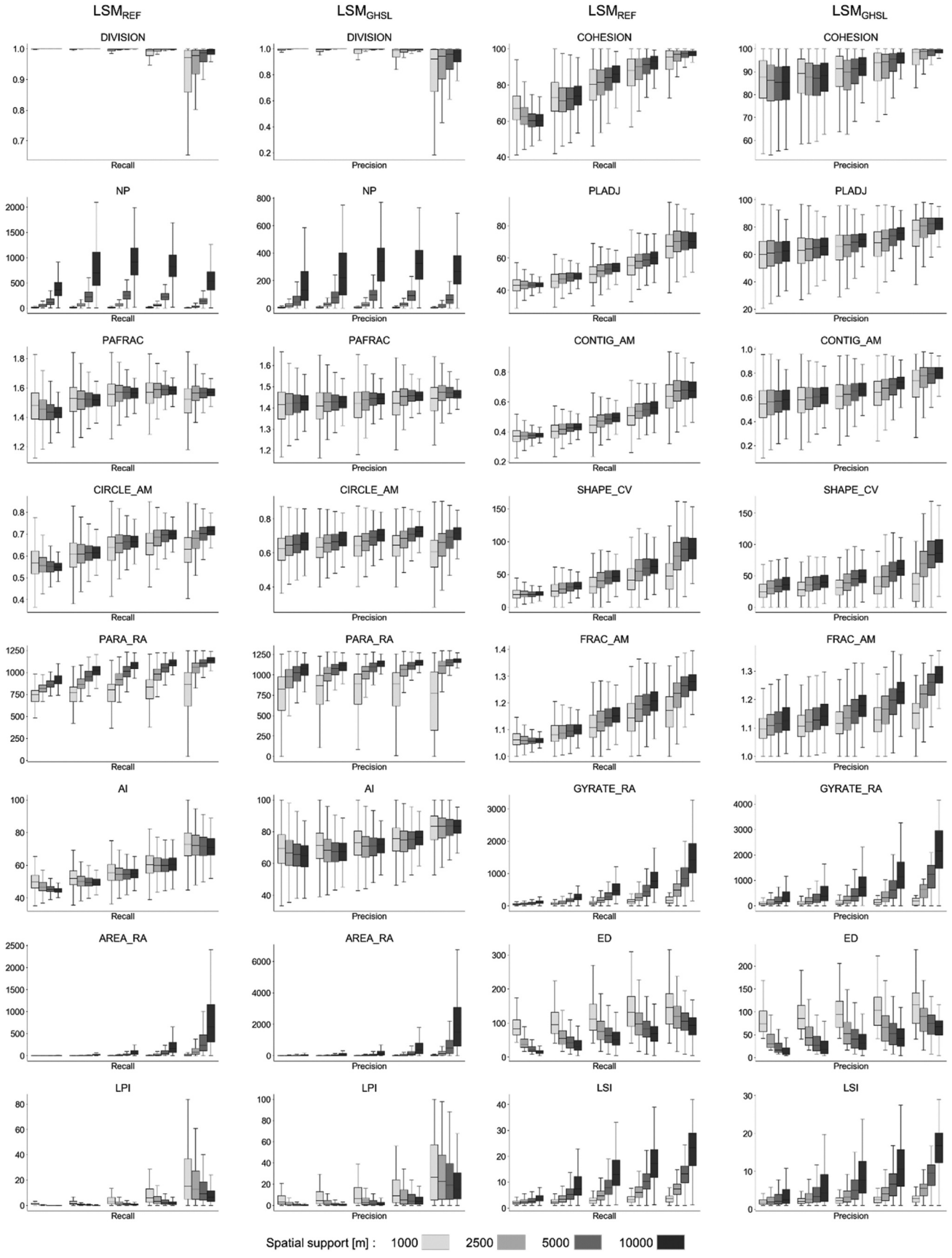

Figure A4. Distributions of landscape metrics derived from the reference data (LSM_{REF}), and from the GHSL (LSM_{GHS}), within strata defined by quintiles of the response variables recall and precision, respectively. LSMs in the upper left exhibit least, in the lower right highest average correlation to the response variable across all support levels.

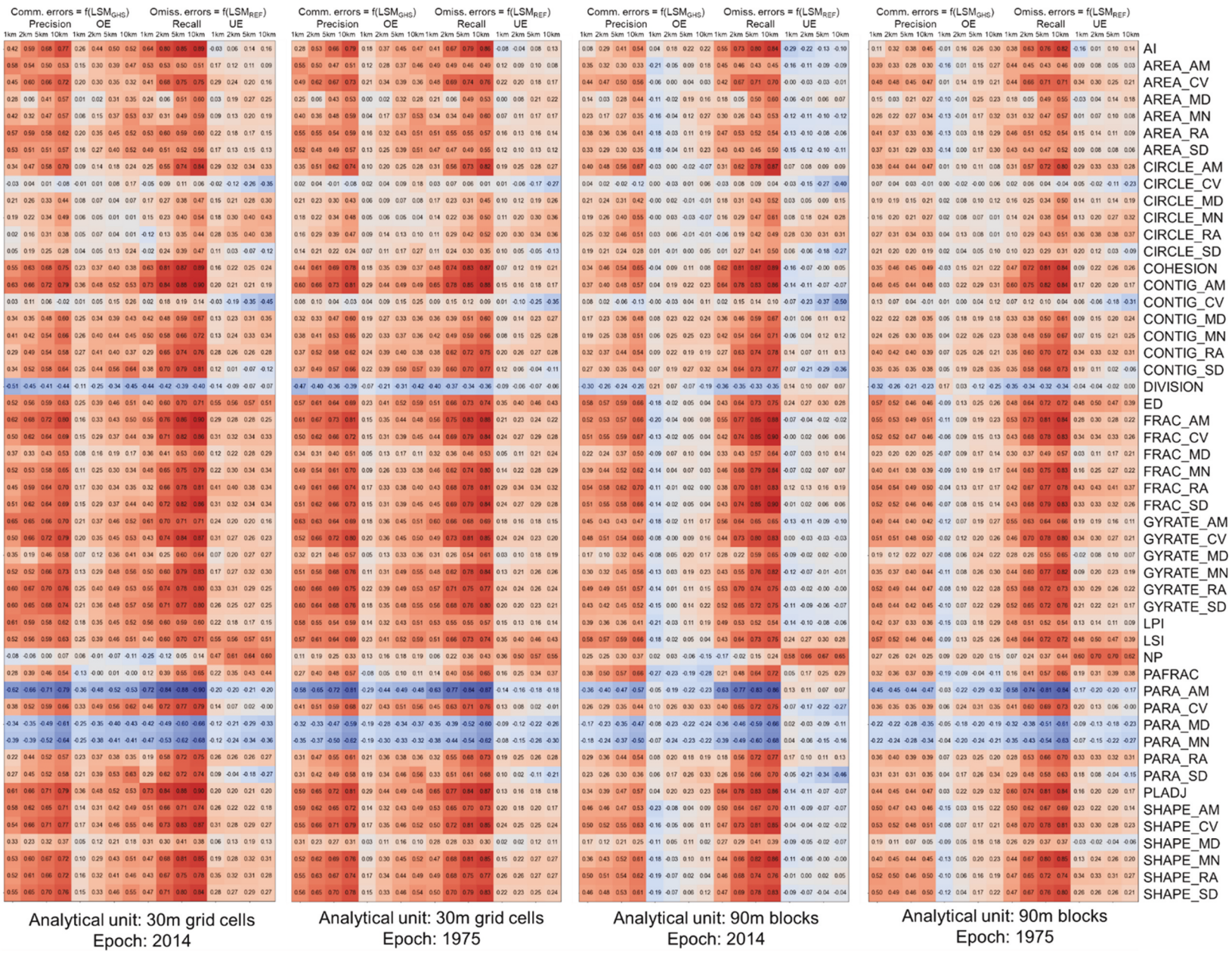

Figure A5. Sensitivity analysis: Correlation coefficients between the 51 landscape metrics and the accuracy estimates, over time and for different analytical units.

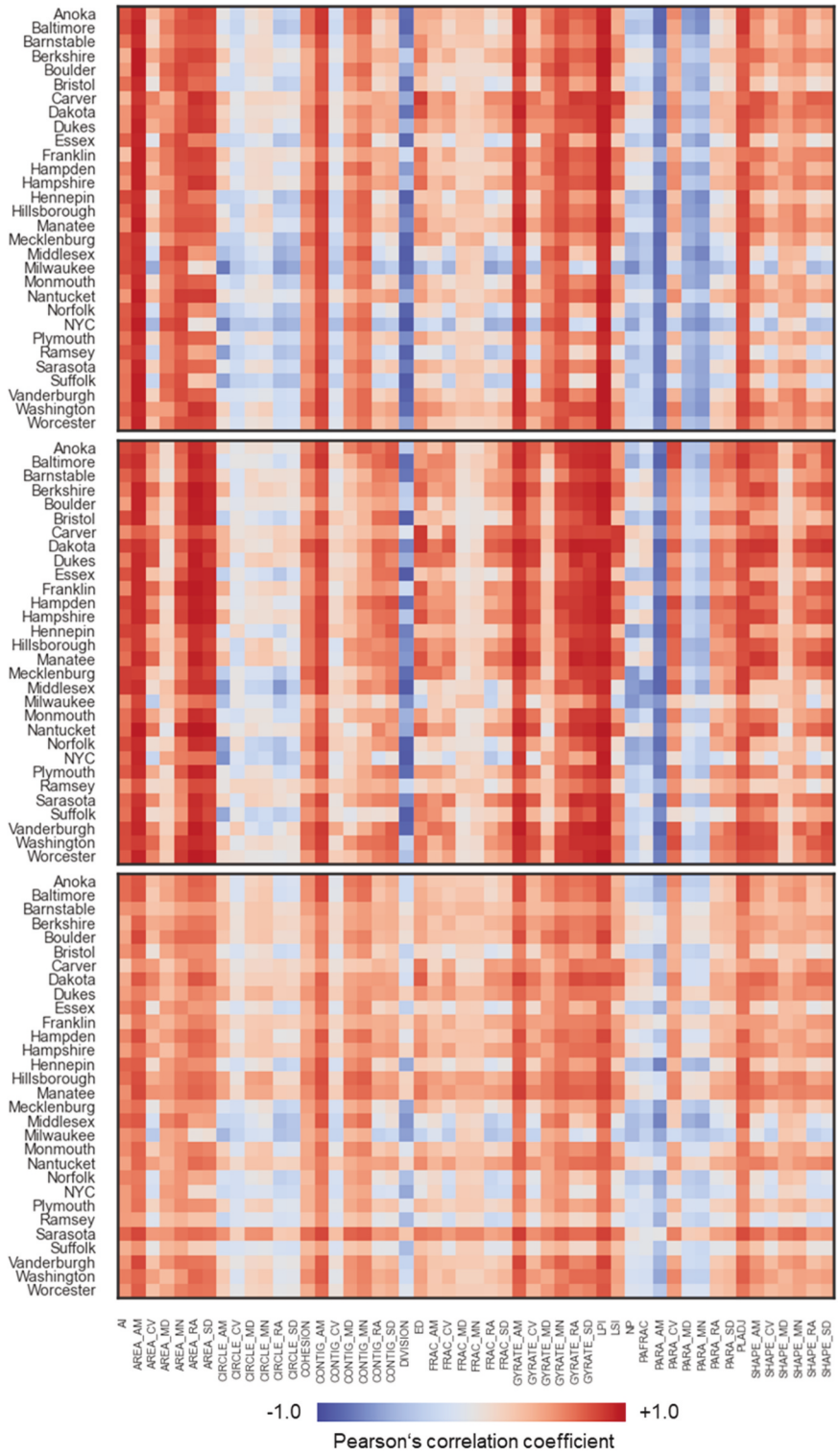

Figure A6. Correlation of the 51 landscape metrics and built-up surface density in 30 U.S. counties, for 1 km (top), 2.5 km (middle), and 5 km spatial support (bottom).