Advancing Precision, Recall, F-score, and Jaccard Index: An Approach for Continuous, Ratio-scale Measurements

Katarzyna Krasnodębska[a], Wojciech Goch[b1], Johannes H. Uhl[c], Judith A. Verstegen[d], Martino Pesaresi[c*]

[a] katarzyna.krasnodebska@twarda.pan.pl, Institute of Geography and Spatial Organization, Polish Academy of Sciences, Twarda 51/55, 00-818 Warsaw, Poland

[b] wojciech.goch@wat.edu.pl, Cybernetics Faculty, Military University of Technology, gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland

[c] johannes.uhl@ec.europa.eu, martino.pesaresi@ec.europa.eu, European Commission, Joint Research Centre (JRC), Via E. Fermi, 2749, 21027 Ispra VA, Italy

[d] j.a.verstegen@uu.nl, Department of Human Geography and Spatial Planning, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands

* corresponding author: Martino Pesaresi, martino.pesaresi@ec.europa.eu

[1] Present address: goch@cs.cas.cz, Institute of Computer Science, Czech Academy of Sciences, Pod Vodárenskou věží 271/2, 182 00 Prague, Czech Republic

Abstract

Gridded data representing attribute estimates at ratio scale are increasingly common for modelling spatial-environmental variables, including class area estimates (e.g. built-up surface area), population abundance (e.g. number of inhabitants), or vegetation-related measurements (e.g. canopy height). The accuracy of model-based gridded data, including classifications of remotely-sensed data, is usually assessed with measures based on confusion matrices with site-specific class allocations. Yet, these measures can only be applied to categorical attributes, not to ratio-scale attributes. Here, we introduce an approach to extend commonly used agreement measures estimated from a confusion matrix (Precision, Recall, F-score and Jaccard index) to non-negative ratio-scale attributes. We test the proposed measures on a synthetic dataset, and in a realistic scenario using gridded data measuring built-up surface area. The proposed measures prove to be viable equivalents of their common categorical counterparts, suitable for evaluating the agreement of gridded data representing attribute estimates at ratio-scale.

Keywords

Agreement measures; accuracy assessment; Jaccard index; Precision; Recall; ratio scale measurements

1. Introduction

Ratio-scale measurements are an increasingly common form for representing land and vegetation attributes of geographic features. Ratio-scale attribute values are continuous estimates of magnitudes, absolute or relative, in units of equal size, with value of zero representing the absence of the geographic feature (Stevens, 1946). In our considerations we are focusing on gridded data, lower bounded with value of zero, and without an upper bound representing ratio scale measurements. The considered gridded data may represent attributes such as population density (e.g. population counts per grid cell, see Schiavina et al., 2023), absolute or relative class area estimates (e.g. built-up surface area, see Pesaresi et al., 2024), vegetation-related measurements (e.g. forest height or biomass density, see Matasci et al., 2018) or environmental measurements (e.g. ice thickness, see Copernicus Climate Change Service, 2018).

The usefulness of such datasets depends on the accuracy of the estimated attribute values, which is typically assessed as the agreement either between the product and reference data, or between the product and another product of the same (or a similar, coherent) attribute from a different source (Congalton, 2001; Foody, 2002). In case of categorical attributes (at nominal or ordinal scale), Recall (i.e., Producer's accuracy; fraction of correct classifications), Precision (i.e., User's accuracy; fraction of relevant correct classifications), their composite – the F-score, and the Jaccard index (relative overlap of two sets, also known as the Intersection over Union, IoU) are the most commonly used measures to assess the classifications of rare occurrences. They are appropriate for assessing the accuracy of categorical classifications of remotely sensed data, as they estimate the classification agreement within the domain of relevant classes and are not inflated by the correctly classified negative domain (Davis and Goadrich, 2006; Uhl and Leyk, 2022), which may be dominant in imbalanced binomial or multinomial distributions of the categorical classifications generated from the remotely sensed data (Congalton 2001). However, these measures can only be applied to categorical attributes, but not to continuous ones.

As such, we assert that variants of Recall, Precision, F-score and the Jaccard index are required that can be applied to the agreement assessment of gridded (or other) data representing estimates of attributes

at the ratio scale, while maintaining the same properties as their categorical counterparts: indifference to the absence of the geographic feature and relativity to the magnitude of the compared attributes. Therefore, we propose variants of these four measures (i.e., cont. Jaccard, cont. Precision, cont. Recall and cont. F-score), in which agreement is interpreted as closeness of the continuous (henceforth abbreviated as "cont.") attribute magnitude estimates and expressed as bounded, dimensionless measures. We illustrate the usefulness of the proposed measures using examples of gridded data representing building height estimates. We test the proposed and existing measures using a synthetic dataset with controlled disagreement as well as in a realistic scenario, comparing the agreement between two datasets representing the area of built-up surface within the extent of each grid cell.

2. Theory

Taking into account the increasing number of available datasets of continuous attribute estimates used in the field of remote sensing, there is an evident need to develop applicable agreement measures that are straightforward, intuitive and adjusted to the typically non-normal distribution of land use / land cover classes (Duveiller et al., 2016; Pontius and Millones, 2011; Riemann et al., 2010; Stehman and Foody, 2019). The agreement between continuous data is commonly addressed by measures of difference, such as mean deviation (MD), mean absolute deviation (MAD), root mean square deviation (RMSD), mean absolute percentage deviation (MAPD), as well as measures of association, namely Pearson correlation coefficient (r) and coefficient of determination (R^2) or Slope of the least squares line (Ji and Gallo, 2006; Pontius, 2022). However, each of these measures has properties which make them less suitable when used to assess the agreement of data representing estimates at the ratio scale: The commonly used MD, MAD and RMSD measures of disagreement are dimensional measures, dependent on the scale and unit of assessed products, while MAPD is unstable for comparing values near zero (Ji and Gallo, 2006). Correlation coefficient r and R^2 are measures of linear covariation between two datasets, insensitive to the systematic error in the estimated value, and as such are inappropriate when assessing estimates of the magnitude of an attribute. Several agreement measures were proposed, aiming to mitigate the limitations

of these commonly used measures: Among the most common ones are the Wilmotti and Mielke indices, which yield bounded, dimensionless and symmetric measures of agreement, but they are sensitive to the internal variance of compared datasets (Willmott et al., 2012; Willmott and Wicks, 1980). Moreover, the Concordance Correlation Coefficient (Lin, 1989) or the modified Mielke index (Duveiller et al., 2016) have been proposed, capturing in a single index the difference and the association between the data being compared. However, none of the aforementioned measures take into account 1) the closeness of the magnitude estimates being compared, or 2) the meaning of values equal to zero. The latter, on a ratio scale, implies the absence of the geographic feature for which the attribute is estimated. Given that geographic features may be sparsely distributed and not spatially exhaustive, the distribution of their estimated attribute values is non-normal, often peaking at zero (i.e., absence of the geographic feature) and with local maxima at the estimated attribute values (see ***Figure 1*** for an example). As a result, there is a clear need to develop agreement measures that are not distorted by the agreement encountered in the spatially dominating domain, where the geographic features of interest are absent, in particular when assessing gridded data on geographic features sparsely distributed across landscapes.

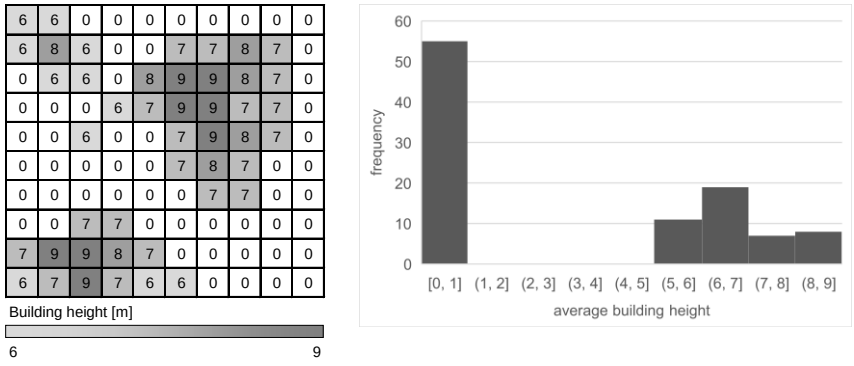

Figure 1. Illustrating the concept of gridded ratio-scale measurements. Left: Gridded 10×10 dataset representing average building height per grid cell. Right: Histogram of the estimated values showing the dominating frequency of values equal to zero.

Herein, we propose measures for the agreement assessment of gridded data representing attribute estimates by adapting the measures used for categorical classification of remotely-sensed data. As this is typically based on the evaluation of site-specific class allocations using confusion matrices (Foody, 2002),

the extension of well-known descriptive statistics for assessing classifications of rare occurrences (i.e., Jaccard, Precision, Recall and F-score) logically follows. Existing efforts for such extensions are based on confusion matrices build on the interpretation of grid cell values in terms of ambiguity (Binaghi et al., 1999), probability (Lewis and Brown, 2001) or fraction (Pontius and Cheuk, 2006) of class occurrence. Although these measures are based on continuous values, they are still related to class occurrence and not to a ratio-scale attribute of a given class, such as vegetation height. To our knowledge, there is no generalizable adaptation of these well-known measures to gridded data representing continuous estimates of ratio-scale attributes – not interpreted in terms of frequency, a degree of class membership or other arbitrary measure in a 0-1 range (e.g. Uhl et al., 2023), but as a representation of an actual unbounded, continuous estimate of magnitude, for which the construction of the confusion matrix is not meaningful.

Herein, we develop equivalents of Jaccard index, Precision, Recall and F-score measures for estimates of attributes at the ratio scale (Section 3); and illustrate that the proposed measures are viable and easily interpretable for the evaluation of gridded data representing attribute estimates of sparsely distributed geographic features, using a range of experiments (Sections 4 and 5).

3. Calculation

Accuracy assessment relies on comparing the attribute values under test against reference data of assumed higher accuracy (FGDC, 1998). Herein, we broadly refer to the data under test as ‘modelled data’, while the ‘model’ can be any abstract (gridded or other) representation of the estimated geographic feature attribute to be evaluated. For binary, nominal, or ordinal variables, accuracy can be summarized in a confusion matrix, *w*here true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP) represent counts resulting from cross-tabulation of the reference and the modelled dataset. Recall, determining the fraction of positive values that was classified as such, is defined as:

$$Recall = \frac{TP}{TP + FN} \qquad \textit{Eq. (1)}$$

Precision, which in turn determines the fraction of modelled positive values that are correct, is defined as:

$$Precision = \frac{TP}{TP + FP} \qquad \textit{Eq. (2)}$$

The Jaccard index is a measure of proximity between two sets given by the number of elements common to both sets (i.e., overlap) compared to the number of all elements present (i.e., sum). Applied specifically to the binary case, it measures the fraction of correctly assigned positives among all positives both in model and reference:

$$Jaccard = \frac{TP}{TP + FP + FN} \qquad \textit{Eq. (3)}$$

Extensions of agreement measures to continuous variables can be achieved in various ways, depending on the logic behind the derivation (e.g. the Tanimoto similarity index (Tanimoto, 1958), see ***Appendix A***). We propose a formulation of Precision, Recall, and the Jaccard index for ratio-scale attribute estimates using elementwise Minimum and Maximum operators. We propose to treat the reference and the modelled data separately, as sets of magnitudes of attribute estimates defined for each individual grid cell. We interpret their agreement as the ratio of their geometric overlap to their geometric union, with the intersection (union) of two sets interpreted as the area defined by the Minimum (Maximum) operator (***Figure 2***).

Let us define the modelled *N*-element dataset as *M* and the same-sized reference data set as *R*:

$$M = \{m_1, m_2, \dots, m_N\} \;\; where \;\; m_1, \dots, m_N \in \mathbb{R}_{\geq 0};$$

$$R = \{r_1, r_2, \dots, r_N\} \;\; where \;\; r_1, \dots, r_N \in \mathbb{R}_{\geq 0}$$

The discussed measures can then be determined as follows:

$$cont.Recall = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} r_i} \qquad \textit{Eq. (4)}$$

$$cont.Precision = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} m_i} \qquad \textit{Eq. (5)}$$

$$cont.Jaccard = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} max(m_i, r_i)} \qquad \textit{Eq. (6)}$$

These measures can be interpreted similarly to their binary equivalents (see ***Figure 2a***). Specifically, continuous Recall can be interpreted as the rate of the total magnitude of the reference attribute estimated by the model; cont. Precision can be understood as the rate of the total magnitude of the modelled attribute in agreement with the reference attribute. Combined, they yield information whether the model is predominantly overestimating the magnitude of the attribute (cont. Precision < cont. Recall) or underestimating (cont. Precision > cont. Recall). Cont. Jaccard can be viewed as a ratio of the magnitude of the given attribute correctly estimated by the model to the sum of correct estimates and all errors of omission and commission. Cont. Jaccard is in this sense equivalent to the Ružička similarity measure (Ružička, 1958), an abundance-based measure of relative difference (Ning et al., 2019), designed to assess the similarity of ecological communities.

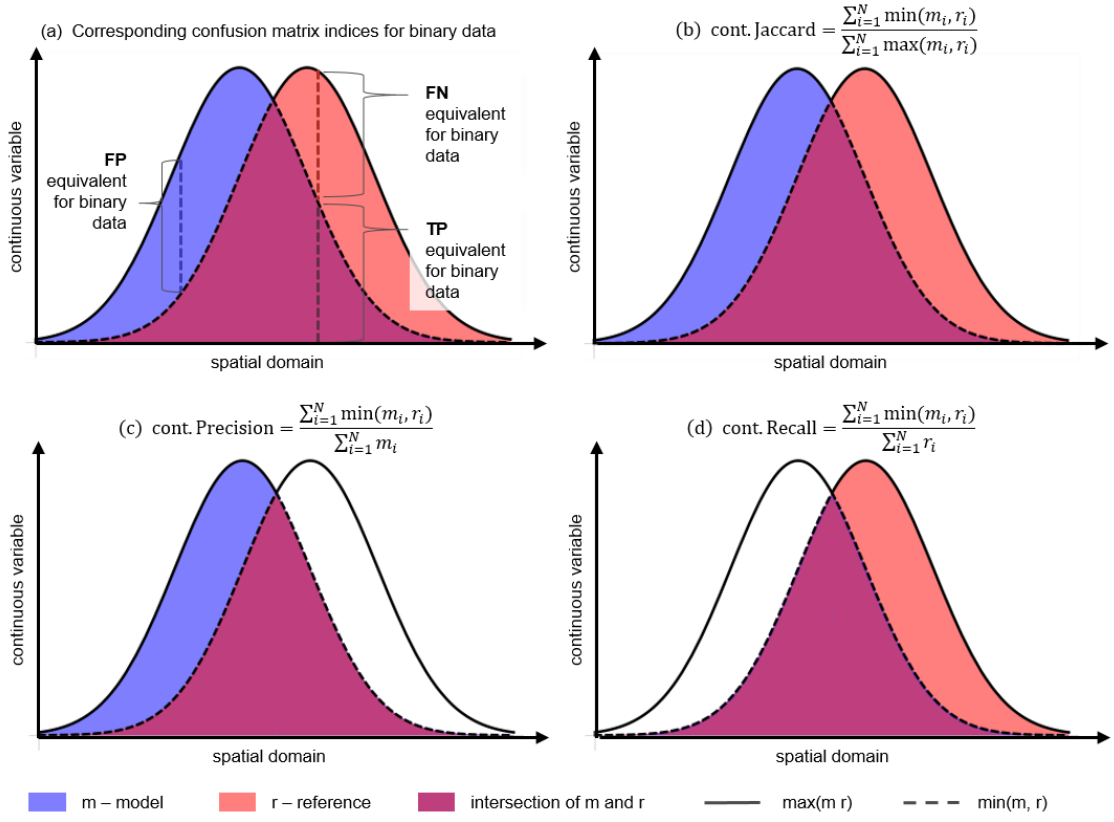

Figure 2. Geometric representation of the concepts of intersection and union of continuous functions discretised to the spatial domain. a) Confusion matrix indices equivalents on the ratio scale, b) cont. Jaccard data domain, c) cont. Precision data domain and d) cont. Recall data domain. The X axes show the discrete spatial domain in a symbolic, one-dimensional representation. The Y axes show the attribute magnitude estimate at the ratio scale.

Cont. Recall and cont. Precision measures can be combined into a single F_β score:

$$F_\beta = (1+\beta^2)\frac{cont.Recall * cont.Precision}{(\beta^2 * cont.Precision) + cont.Recall} \quad \textit{Eq. (7)}$$

Where smaller β values favour models characterized by smaller cont. Recall, whereas larger β values favour models with smaller cont. Precision. In extreme cases, F_β simplifies to:

$$F_0 = cont.Precision; \;\; F_\infty = cont.Recall \quad \textit{Eq. (8)}$$

The $\beta = 1$ yields the harmonic mean of Precision and Recall, also known as Dice coefficient (Dice, 1945), which we test in this study (cont. F1-score).

4. Materials and methods

We showcase the usefulness of our proposed continuous measures using a set of small toy data (Section ***4.1***). Moreover, we investigate their behaviour using pairs of larger, synthetic datasets with different levels of disagreement (Section ***0***); and assess their response to variations in the continuous spatial variable using actual data on built-up surface area (Section ***4.3***).

4.1. Showcase examples

We design three pairs of gridded datasets, reference and modelled, illustrating the usefulness of the proposed measures for assessing the agreement of gridded data representing ratio-scale attribute estimates. In the examples below, we are comparing gridded toy datasets interpreted as the average height of buildings per grid cell, in order to link these toy data to remote-sensing related applications. For each pair of datasets, we report our proposed measures of agreement (cont. Jaccard, cont. Precisions, cont. Recall and cont. F1-score). To demonstrate the added value of the proposed measures for estimating the agreement of ratio-scale measurements, we compare them with measures proven to be effective in estimating the difference (Mean Error, ME, and Mean Absolute Error, MAE, corresponding to MD and MAD, respectively) and association (r and Slope) of interval-scale measurements, showing units of all quantitative phenomena (Pontius, 2022). To do so, we build upon a theoretical experiment by Pontius (2022) and use original data underlying this experiment, to generate reference and modelled data values.

The first pair of gridded data is represented by grids of dimension 2×2. From (Pontius, 2022) to populate grid cells in the reference dataset, and we construct the modelled dataset by adding to each grid cell the deviation reported for series E from the same table, herein representing medium-rise building heights (***Figure 3, example A***). To highlight the utility of our proposed measures in assessing the (relative) closeness of magnitude estimates, we construct a second pair of gridded datasets to have higher numerical values, but the same level of association and the same absolute errors as the previous pair. This is done by adding a constant value of 60 to each grid cell in the datasets of example A, both reference and modelled, herein representing the average building height estimates of high-rise buildings (***Figure 3, example B***). Finally, to illustrate the performance of the measures in a sparsely populated landscape (i.e., higher levels of absence of the geographic feature, i.e., building), we generate the reference and modelled datasets by extending datasets B into 4×4 grids, by padding the edges with values of zero, indicating absence (***Figure 3, example C***).

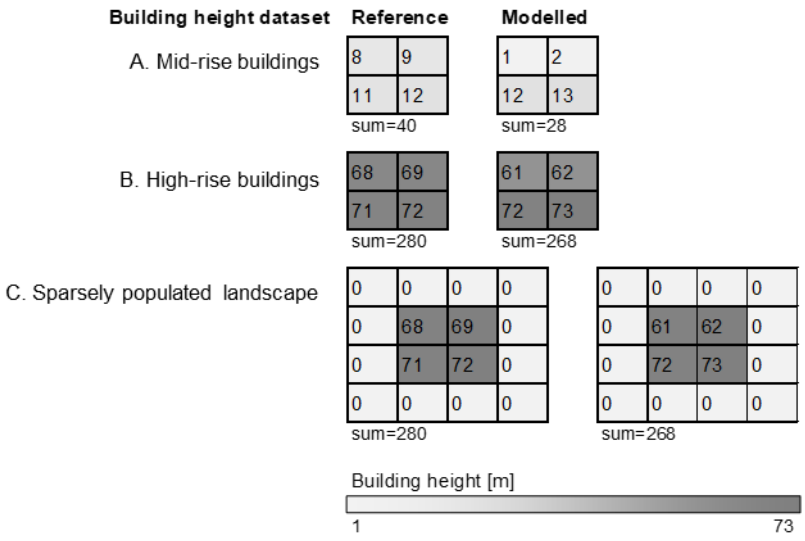

Figure 3. Toy data to illustrate the utility of the proposed measures, representing average building height: A. of mid-rise buildings, with low estimated values; B. of high-rise buildings, with higher values than in example B but with the same levels of association and error between the reference and modelled datasets; C. the same values as in example B, but embedded into a sparsely populated landscape. Data in A. adopted from (Pontius, 2022).

4.2. Synthetic experiment

In the synthetic experiment we use an artificial landscape, resembling estimates of crop canopy height (***Figure 4a, left***), generated with five shifted Gaussian windows stacked together, with a noise added to the data using a two-dimensional Gaussian filter. In this experiment, we use a grid of 1000×1000 cells, with continuous cell values between zero and two (approximating corn canopy height). Details on synthetic data preparation can be found in ***Appendix B***.

To estimate the agreement with the proposed cont. Recall, cont. Precision, cont. Jaccard and cont. F1-score, we consider the produced synthetic landscape as the reference dataset and generate corresponding (synthetic) modelled datasets by adding to the reference data: absolute bias and relative bias, referring to systematic disagreement, and random noise. To test the influence of the magnitude of bias and noise on our measures, we add increasing bias (absolute and relative) and random noise with increasing standard deviation σ, to the reference dataset (***Figure 4)***. For absolute bias, we add to each grid cell a value between -0.5 and 0.5 in a series of runs; for relative bias, we multiply each grid cell with value from 0.5 to 1.5; and for random noise, we add to each grid cell a value drawn from a normal distribution with a mean of zero and a σ varying from 0 to 0.2. For modelling absolute bias, we replace the negative values in the modelled dataset with zeros, representing absence of the estimated attribute. We compare our continuous agreement measures with existing measures, i.e., correlation coefficient ***r***, and the difference measures MAE, ME and Root Mean Squared Error (RMSE), computed for the same pairs of datasets (***Section 5.1***).

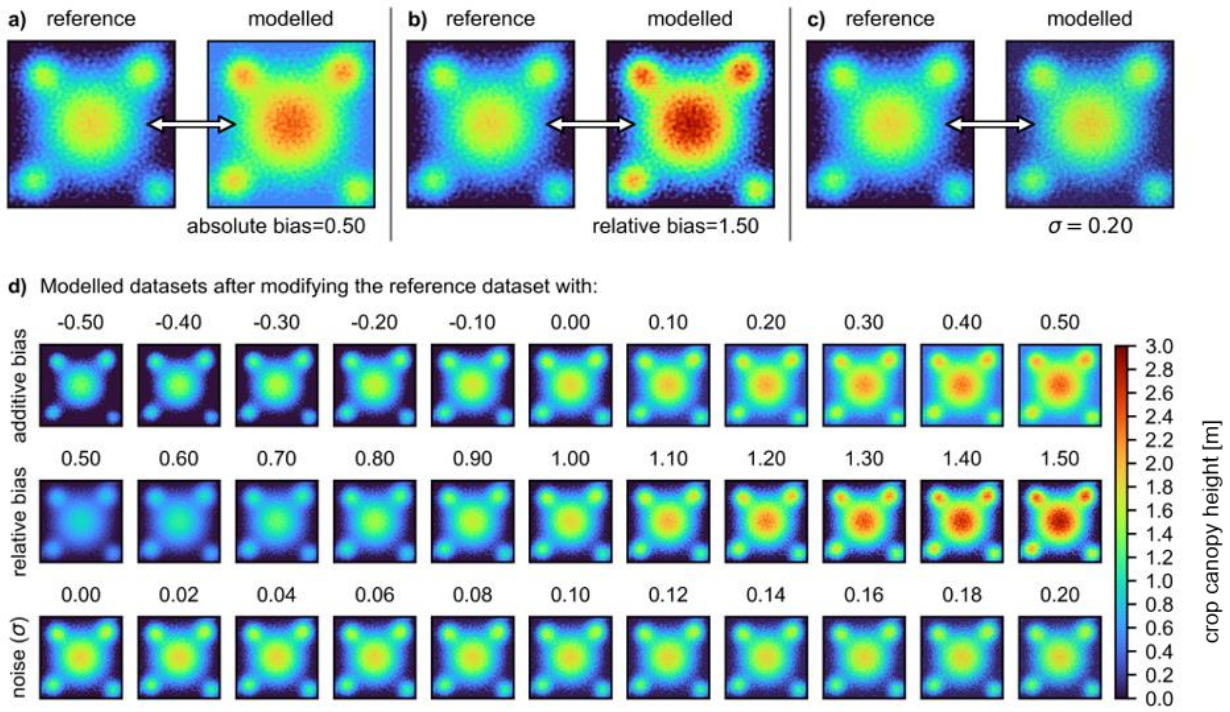

Figure 4. Synthetic reference datasets (left grid in panels a, b and c) compared with modelled datasets generated for three types of disagreement: absolute bias (panel a), relative bias (b) and random noise (c). Panel d shows modelled datasets generated in a series of runs. The colour scale for each dataset is the same.

4.3. Realistic experiment with controlled error

To test the proposed continuous measures in a realistic scenario and to assess their response to variations in the continuous spatial variable, we estimate the agreement of two datasets representing the built-up surface area for a study area in Sao José dos Campos, Brazil. We generate a reference dataset by rasterizing building vector footprints (Copernicus Emergency Mapping Service, 2022) to 10 m resolution, and generate a modelled dataset by ingesting both commission and omission errors into the reference dataset.

We define four spatial domains, derived by binarizing each dataset using a cutoff value of 0 and computing the confusion matrix elements TP, FP and FN, considering the first dataset as the reference and the second dataset as the modelled data (***Figure 5a***). We then draw from two overlapping, skewed normal distributions (***Figure 5b***), and replace the grid cell values in each domain, to create a baseline scenario:

High values for both, reference and model grid cells in TP domain (TP_{ref} and TP_{model}, respectively), and low values for the reference grid cells in the FN domain (FN_{ref}), as well as for the modelled grid cells in the FP domain (FP_{model}) (***Figure 5c, d***). This baseline scenario represents the case when grid cells of high reference values (henceforth called densities, as they reflect a density, i.e., built-up surface per grid cell area) are correctly detected, falsely omitted grid cells contain low densities in the reference data, and false positive grid cells are assigned with a low density as well. We then systematically modify the continuous grid cell densities within these domains by multiplying them with a factor of 2. Specifically, there are two states that the densities in the four domains TP_{ref}, TP_{model}, FN_{ref}, and FP_{model} can take: 1) the assigned baseline values, or 2) the increased values, which are the baseline values multiplied by 2. All possible combinations of the two states in the four domains yield $4^2 = 16$ different scenarios, each of them representing different levels of densities in the four domains. This allows to simulate scenarios of different levels of bias in the modelled data. For example, we simulate a scenario where the reference and modelled densities in the TP_{ref} and TP_{model} domains are similarly high, and the modelled densities in the FP_{model} domain are high as well – simulating a model output that heavily overestimates the density in the FP_{model} domain, for which we expect a low cont. Precision (***Figure 5e***). For each of the 16 scenarios (***Figure 5f***) we generate the proposed continuous agreement measures, and compare them to existing, commonly used measures: Mean Absolute Percentage Error (MAPE), its weighted alternative, wMAPE (Kolassa and Schütz, 2007), MAE, ME and r, and observe their response (***Section 5.3***). See ***Appendix Figure C.1*** for maps corresponding to the 16 scenarios. For all these scenarios, the categorical agreement (when binarizing the data with a cutoff value of 0) remains constant, and thus, this experiment will illustrate the responsiveness of the proposed variables to density variations that would not be captured by a simple categorical agreement assessment of binarized data.

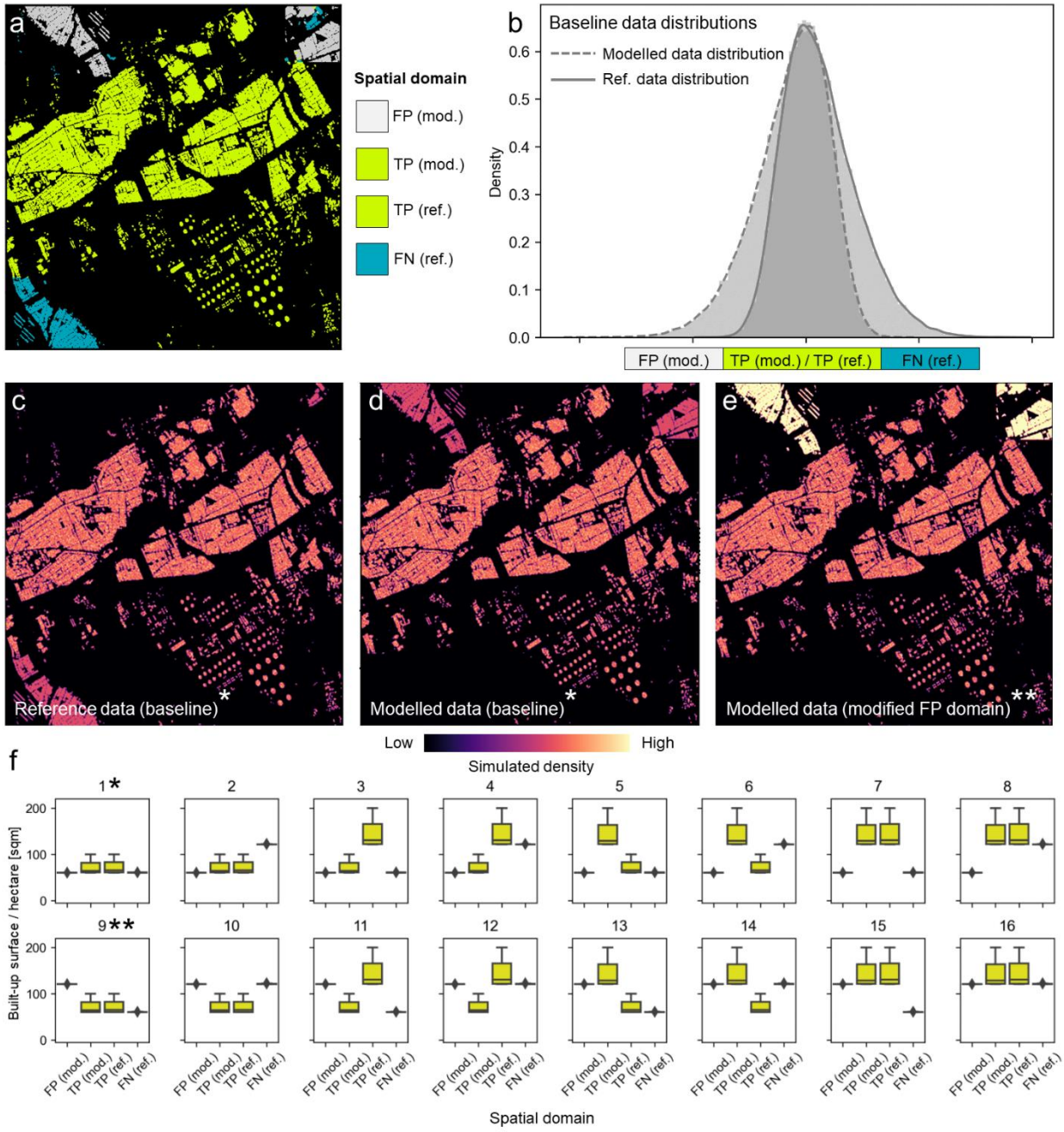

Figure 5. Illustrating the input data for a realistic experiment with controlled error. (a) Categorical agreement map, (b) skewed, overlapping probability density functions from which data are drawn to populate grid cells in the TP_{model}, TP_{ref}, FP_{model}, and FN_{ref} domains for the baseline scenario, resulting baseline scenario for the (c) reference data, and (d) for the modelled data; (e) example of systematically modified modelled data with increased densities in the FP_{model} domain, and (f) showing the density distributions within the four spatial domains for the 16 numbered scenarios, where the densities in each combination of domains are either kept original or multiplied by factor 2. Asterisks in (f) denote the correspondence to the maps in (c) – (e).

5. Results and Discussion

5.1. Utility of the proposed measures

Using a set of three simple examples (see Section 4.1), we illustrate how the proposed measures are suitable for assessing the agreement of gridded data representing estimates of attributes at the ratio scale. First, the proposed measures of agreement complement error measures reporting underestimation (***Table 1***, ***example A***, ME = -3 m, MAE = 4 m). The cont. Recall informs that 65% of the estimated building height was allocated in the modelled dataset while cont. Precision informs that up to 93% of the modelled building height was allocated correctly. Indeed, we can only observe an overestimation in the bottom-right cell, with an estimate of 13 m instead of 12 m.

Secondly, the proposed measures underline the importance of the relativity of commission and omission errors to the magnitude of the estimated attribute. Despite the datasets in examples A and B have by design exactly the same level of association and the same errors, their agreement strongly differs (***Table 1***, ***examples A and B)***. The value of cont. Jaccard estimated in example B (cont. Jaccard = 0.94) is substantially higher than in example A (cont. Jaccard = 0.62). Moreover, the cont. Recall shows a difference of 30 percentage points in estimated magnitude allocation between these two examples (cont. Recall = 0.65 in example A, cont. Recall = 0.95 in example B). This observation has implications for the choice of model for different applications, among others. In this example, a model that worked reasonably well for estimating the height of high-rise buildings would perform poorly for height estimation of mid-rise buildings.

Third, and importantly, the proposed measures are invariant to the absence of the geographic feature, represented in ratio scale with grid cells of value zero, both in the reference and in the modelled datasets (***Table 1***, ***examples B and C)***. This is not the case for measures of error (ME, MAE) and association (r and Slope), which show stronger agreement due to the correct detection of the absence of the geographic feature of interest. In cases where this is irrelevant (e.g., non-sparsely distributed geographic features), the

proposed measures (cont. Jaccard, cont. Precision, cont. Recall and cont. F-score) capture the actual agreement between the reference and the modelled estimates of the attribute magnitudes.

Table 1. Measures of error (ME, MAE), association (r, Slope) and the proposed continuous agreement measures computed for pairs of gridded datasets in examples A, B and C.

Building height dataset	ME	MAE	r	Slope	cont. Jaccard	cont. Precision	cont. Recall	Cont. F1-score
A. Mid-rise buildings	-3,00	4,00	0,97	0.28	0,62	0,93	0,65	0,77
B. High-rise buildings	-3,00	4,00	0,97	0.28	0,94	0,99	0,95	0,97
C. Sparsely populated landscape	-1,00	1,00	0,998	1,04	0,94	0,99	0,95	0,97

5.2. Impact of bias and noise on the proposed agreement measures

Results of the synthetic experiment (see Section 4.2) show that the proposed cont. Jaccard, cont. Precision, cont. Recall and cont. F1-score measures are sensitive to both bias and noise, as desired (***Figure 6).*** The proposed cont. Precision and cont. Recall measures are indicators of over- and underestimation of the estimated magnitude, whereas the proposed cont. Jaccard is a symmetric measure of closeness between two magnitude estimates. The proposed measures are not inflated by the agreement on the absence of the geographic feature, in contrast to the other tested measures, where the imposed lower bound at value zero lowers the estimated (absolute) measure values (***Figure 6d, negative bias***). As measures of relative difference, the proposed measures respond differently in the case of positive or negative bias in the modelled dataset (dictating varying magnitudes of the estimated attribute, see ***Appendix D*** for details). The agreement is expressed on an interpretable, bounded scale from 0 to 1 and estimated in relation to the magnitude of the estimated attribute mapped in both datasets, as opposed to the measures reporting averaged error (MAE, ME, RMSE) or correlation coefficient r, reporting about

association, indifferent to bias.

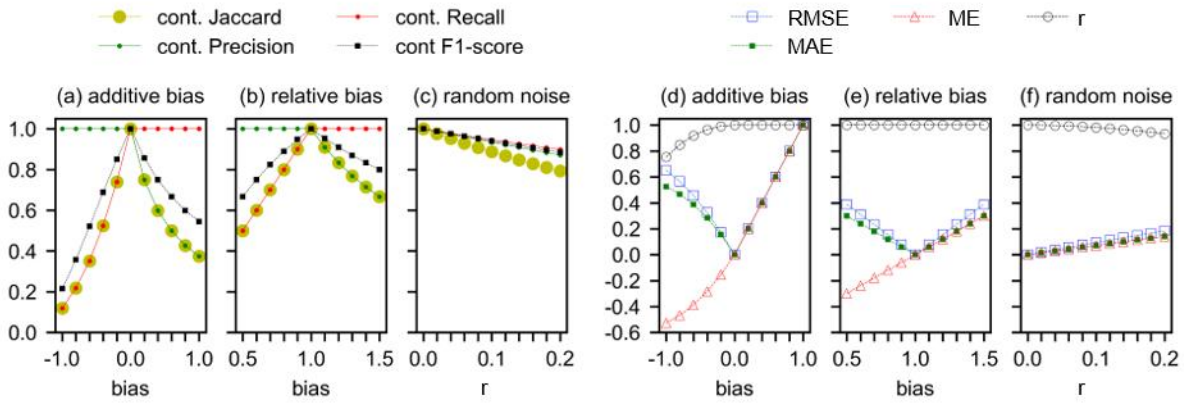

Figure 6. Agreement measures computed for the synthetic landscapes. Columns a, b and c show the introduced continuous agreement measures: cont. Jaccard, cont. Precision, cont. Recall and cont. F1-score. Columns d, e and f show agreement measures typically used for accuracy assessment of continuous data: r and error measures MAE, ME and RMSE.

5.3. Responsiveness of the proposed measures to variation in data distributions

For each of the 16 different scenarios in our controlled error experiment (see Section 4.3), the variations in the continuous gridded datasets are reflected in the values of the continuous agreement measures (***Figure 7a***), while the categorical agreement remains constant (cat. Jaccard = 0.81, cat. Precision = 0.91, cat. Recall = 0.87). As expected, the baseline scenario with increased densities in the TP_{model} and TP_{ref} domains (i.e., very high average densities in the TP_{model} and TP_{ref}, low densities in the FP_{model} and FN_{ref} domains (***Figure 7a,*** scenario 7) yields the highest values for our continuous agreement measures. Lowest values are found for scenario 12, where the disagreement between reference and modelled datasets in the grid cells of overlap (i.e., TP_{model} and TP_{ref} domains) is high, and in addition to that, the densities in the FP_{model} and FN_{ref} domains are high as well. Cont. Precision and cont. Recall are responsive to the density variations, in particular within the TP_{model} and TP_{ref} domains, and to a lesser degree, within the FP_{model} and FN_{ref} domains, respectively. This is due to the imbalanced spatial support between the TP_{model}, TP_{ref}, FP_{model}, and FN_{ref} domains (Figure 5a; number of grid cells: N_{TP} = 78 272, N_{FP} = 7 245, N_{FN} = 11 005). As an example, we compare scenarios 7 and 8: the density increase in the FN_{ref} cells (i.e., mimicking an omission error) causes a drop in the cont. Recall by only 0.05 (from 0.85 to 0.80),

while this effect is much stronger when we decrease the densities in the TP_{model} domain (scenario 7 versus scenario 3), where the induced omission error causes the cont. Recall to drop by 0.38 (from 0.85 to 0.47). This observation has as important implication: the probabilistic interpretation of the cont. Precision and cont. Recall allows to infer on the total proportion of misallocated modelled and reference densities. E.g., in scenario 3 (cont. Recall=0.47, cont. Precision=0.93), we observe a case of underestimation, where only 47% of reference densities are captured by the model, but 93% of the modelled densities are allocated correctly.

The proposed cont. Jaccard measure is bounded to 0-1 scale and mostly correlates with RMSE and MAE (***Figure 7***), which confirms its general usefulness. Pearson's correlation coefficient *r* is barely responsive to the systematic disagreement injected in the data. The MAPE, and to a lesser degree, wMAPE, yield almost identical results for all scenarios where the agreement in the TP domains is similar for the modelled and reference distributions (i.e., scenarios above the dashed line, ***Figure 7a***), while cont. Jaccard ranges from 0.60 to 0.75 between these scenarios. Moreover, the ME as a signed measure oscillates with increasing densities in the FP and FN domain, respectively, but does not allow for an individual disentanglement of omission and commission errors, as opposed to the cont. Precision and cont. Recall measures (see ***Section 3***). The latter observations indicate that the 'classical' measures are very useful for quantifying *individual components* of the differences between the data compared, the proposed measures summarize some of the classical measures, while they disentangle others, and thus, represent valuable alternative measures of agreement. Looking at cross-correlations between measures (***Figure 7b***), we observe a wide range of variability. Notably, ME correlates highly (positively and negatively) with cont. Recall, and cont. Precision, respectively, further highlighting the capability of these measures to disentangle overall, signed measures such as the ME into components of omission and commission errors, respectively, in an interpretable manner.

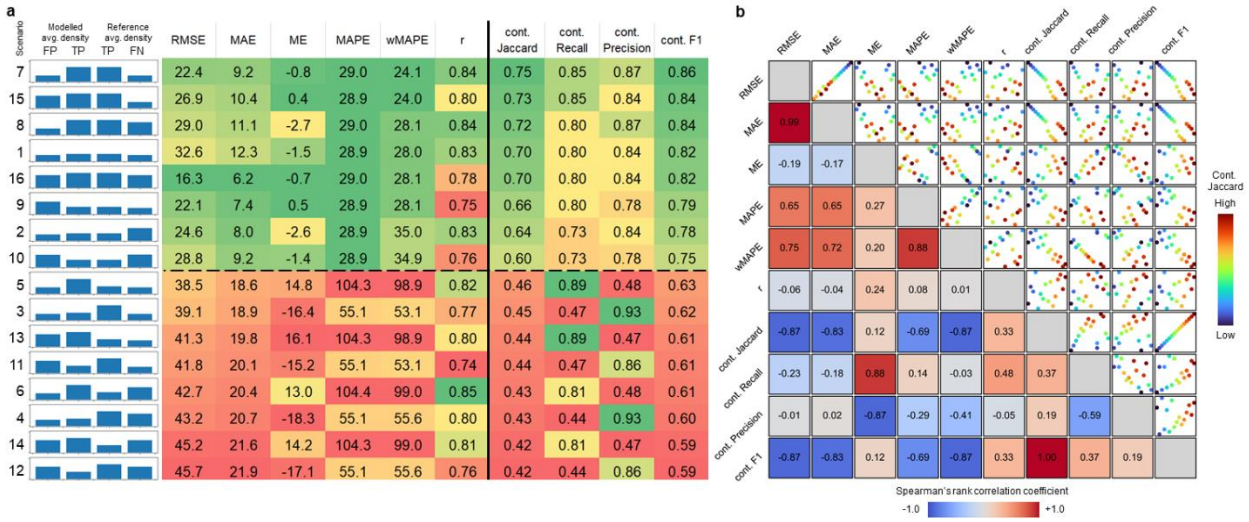

Figure 7. Comparison of agreement measures for 16 scenarios of systematically modified data in the TP_{ref}, TP_{model}, FP_{model}, and FN_{ref} domains. (a) Values of commonly used (left part) and proposed (right part) measures for each scenario, sorted by the continuous Jaccard index. Numbering of the scenarios corresponds to Fig. 3f. The colors in panel (a) correspond to the values of each measure, , with green colors indicating higher levels of agreement or association, or lower levels of error. The dashed line separates scenarios by their level of agreement in the TP domains (above dashed line = agreement in TP_{ref} and TP_{model} domains, below dashed line = disagreement in TP_{ref} and TP_{model} domains). Panel (b) shows the cross-correlation matrix of the ten measures, and their pair-wise scatterplots, based on the 16 scenarios, color-coded by the continuous Jaccard index. For this qualitative assessment, data in scatterplots are rank-transformed, and correlation is measured using Spearman's rank correlation coefficient.

6. Conclusions

In this study, we presented four measures for assessing the accuracy of gridded data representing estimates of attributes at the ratio scale: cont. Jaccard, cont. Precision, cont. Recall, and cont. F1-score. These measures were applied and tested in a range of experiments. We establish that due to its robustness and interpretability, the cont. Jaccard measure, an extension of the widely recognized IoU agreement measure, is a practical tool for comparing gridded datasets of attributes at the ratio scale, which include absolute or relative estimates (e.g. canopy height or built-up surface density), but can also be applied to any non-spatial, ratio-scale data. Additionally, we illustrated that cont. Precision and cont. Recall offer the capability to disentangle commission and omission errors in the total proportion of misallocated magnitudes, a property that has been largely overlooked in the evaluation of data representing continuous estimates.

The proposed measures are easily interpretable due to the similarity to their well-known categorical counterparts, universal, and straightforward to comprehend in terms of their underlying assumptions. They exhibit two particular properties: indifference to the absence of the geographic feature of the estimated attribute and relativity to the magnitude of the compared values. These properties, combined with their considerable recognisability make them suitable measures for estimating the accuracy of gridded datasets representing unevenly distributed and dispersed attributes at the ratio scale, such as area estimates of human settlements (Pesaresi et al., 2024). The proposed measures provide a tool for researchers and analysts to assess the agreement between remote-sensing derived data represented by increasingly common continuous, rather than categorical measurements, contributing to increase uncertainty awareness in geospatial data and beyond.

7. Acknowledgements

This research was funded in part by National Science Centre, Poland, grant number 2021/43/O/HS4/02700 and the institutional work program 2024 of the European Commission, Joint Research Centre.

8. Author contributions:

Katarzyna Krasnodębska: Methodology, Software, Visualization, Writing - original draft;

Wojciech Goch: Methodology, Formal analysis, Writing - original draft;

Johannes H. Uhl: Software, Methodology, Visualization, Writing - original draft;

Judith A. Verstegen: Methodology, Writing - Review & Editing, Supervision;

Martino Pesaresi: Conceptualization, Writing - Review & Editing, Supervision.

9. Code availability

Python and R code to calculate the proposed measures, as well as data to reproduce experimental results, will be made available upon acceptance.

10. References

Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern recognition letters 20, 935–948.

Congalton, R.G., 2001. Accuracy assessment and validation of remotely sensed and other spatial information. Int. J. Wildland Fire 10, 321. https://doi.org/10.1071/WF01031

Copernicus Climate Change Service, 2018. Sea ice thickness monthly gridded data for the Arctic from 2002 to present derived from satellite observations. https://doi.org/10.24381/CDS.6679A99A

Copernicus Emergency Mapping Service, 2022. EMSN132: Copernicus Exposure Mapping (GHSL) reference data. https://emergency.copernicus.eu/mapping/list-of-components/EMSN132

Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning - ICML '06. Presented at the ICML '06, ACM Press, Pittsburgh, Pennsylvania, pp. 233–240. https://doi.org/10.1145/1143844.1143874

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302. https://doi.org/10.2307/1932409

Duveiller, G., Fasbender, D., Meroni, M., 2016. Revisiting the concept of a symmetric index of agreement for continuous datasets. Sci Rep 6, 19401. https://doi.org/10.1038/srep19401

FGDC, 1998. Geospatial positioning accuracy standards - Part 3: National standard for spatial data accuracy.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4
Ji, L., Gallo, K., 2006. An Agreement Coefficient for Image Comparison. PE&RS 72, 823–833.
Kolassa, S., Schütz, W., 2007. Advantages of the MAD/Mean Ratio over the MAPE. Foresight: The International Journal of Applied Forecasting 40–43.
Lewis, H.G., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. International Journal of Remote Sensing 22, 3223–3235. https://doi.org/10.1080/01431160152558332
Lin, L.I.-K., 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45, 255. https://doi.org/10.2307/2532051
Matasci, G., Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., Zald, H.S.J., 2018. Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. Remote Sensing of Environment 209, 90–106. https://doi.org/10.1016/j.rse.2017.12.020
Ning, D., Deng, Y., Tiedje, J.M., Zhou, J., 2019. A general framework for quantitatively assessing ecological stochasticity. Proc. Natl. Acad. Sci. U.S.A. 116, 16892–16898. https://doi.org/10.1073/pnas.1904623116
Pesaresi, M., Schiavina, M., Politis, P., Freire, S., Krasnodębska, K., Uhl, J.H., Carioli, A., Corbane, C., Dijkstra, L., Florio, P., Friedrich, H.K., Gao, J., Leyk, S., Lu, L., Maffenini, L., Mari-Rivero, I., Melchiorri, M., Syrris, V., Van Den Hoek, J., Kemper, T., 2024. Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data. International Journal of Digital Earth 17, 2390454. https://doi.org/10.1080/17538947.2024.2390454
Pontius, R.G., 2022. Metrics that make a difference: how to analyze change and error, Advances in geographic information science. Springer, Cham.
Pontius, R.G., Cheuk, M.L., 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. International Journal of Geographical Information Science 20, 1–30. https://doi.org/10.1080/13658810500391024
Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32, 4407–4429. https://doi.org/10.1080/01431161.2011.552923
Riemann, R., Wilson, B.T., Lister, A., Parks, S., 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. Remote Sensing of Environment 114, 2337–2352. https://doi.org/10.1016/j.rse.2010.05.010
Ružička, M., 1958. Anwendung mathematisch-statistischer methoden in der geobotanik (synthetische bearbeitung von aufnahmen). Biologia, Bratislava 13, 647–661.
Schiavina, M., Freire, S., MacManus, K., 2023. GHS-POP R2023A - GHS population grid multitemporal (1975-2030). https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE
Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment 231, 111199. https://doi.org/10.1016/j.rse.2019.05.018
Stevens, S.S., 1946. On the Theory of Scales of Measurement. Science 103, 677–680. https://doi.org/10.1126/science.103.2684.677
Tanimoto, T.T., 1958. Elementary mathematical theory of classification and prediction (Internal IBM Technical Report). IBM.
Uhl, J.H., Leyk, S., 2022. A scale-sensitive framework for the spatially explicit accuracy assessment of binary built-up surface layers. RSE 279, 113117. https://doi.org/10.1016/j.rse.2022.113117
Uhl, J.H., Royé, D., Burghardt, K., Aldrey Vázquez, J.A., Borobio Sanchiz, M., Leyk, S., 2023. HISDAC-ES: historical settlement data compilation for Spain (1900–2020). Earth Syst. Sci. Data 15, 4713–4747. https://doi.org/10.5194/essd-15-4713-2023
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Intl Journal of Climatology 32, 2088–2094. https://doi.org/10.1002/joc.2419

Willmott, C.J., Wicks, D.E., 1980. An Empirical Method for the Spatial Interpolation of Monthly Precipitation within California. Physical Geography 1, 59–73. https://doi.org/10.1080/02723646.1980.10642189