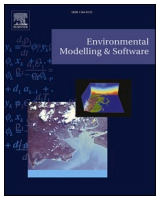
Environmental Modelling and Software 193 (2025) 106614

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



Advancing Precision, Recall, F-score, and Jaccard index: An approach for continuous, ratio-scale measurements

Katarzyna Krasnodębska [a], Wojciech Goch [b,1], Johannes H. Uhl [c], Judith A. Verstegen [d], Martino Pesaresi [c,*]

[a] Institute of Geography and Spatial Organization, Polish Academy of Sciences, Twarda 51/55, 00-818, Warsaw, Poland
[b] Cybernetics Faculty, Military University of Technology, Gen. Sylwestra Kaliskiego 2, 00-908, Warsaw, Poland
[c] European Commission, Joint Research Centre (JRC), Via E. Fermi 2749, Ispra, 21027, VA, Italy
[d] Department of Human Geography and Spatial Planning, Utrecht University, Princetonlaan 8a, 3584 CB, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Agreement measures
Accuracy assessment
cJaccard
cF-score
Intersection over union
Continuous Jaccard

ABSTRACT

Gridded data representing attribute estimates at the ratio scale are increasingly common for modelling spatial-environmental variables, including class area estimates (e.g. built-up surface area), population abundance, or vegetation-related measurements such as canopy height. The accuracy of 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 cannot be applied to attribute estimates at the ratio-scale. Here, we introduce an approach to extend commonly used agreement measures derived from a confusion matrix (i.e. Jaccard index, Precision, Recall and F-score) to non-negative, continuous ratio-scale attributes. The proposed measures, cJaccard, cPrecision, cRecall and cF-score, have been tested on synthetic datasets, and in a realistic scenario using gridded data measuring built-up surface area. They are viable equivalents to their binary counterparts, invariant to imbalanced data, and suitable for evaluating the agreement of various types of data representing ratio-scale attribute estimates.

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). Herein, we focus on gridded data representing ratio-scale measurements without upper bound and with a lower bound of zero. These 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 by measuring the agreement between the product and reference data, or between the product and another product measuring the same or a similar attribute from a different, independent source (Congalton, 2001; Foody, 2002). In case of categorical attributes at binary, nominal or ordinal scale, Precision (i.e. User's accuracy; fraction of relevant correct classifications), Recall (i.e. Producer's accuracy; fraction of 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 capture 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 cannot be applied to unbounded magnitude estimates at the

* Corresponding author.
E-mail addresses: katarzyna.krasnodebska@twarda.pan.pl (K. Krasnodębska), wojciech.goch@wat.edu.pl, goch@cs.cas.cz (W. Goch), johannes.uhl@ec.europa.eu (J.H. Uhl), j.a.verstegen@uu.nl (J.A. Verstegen), martino.pesaresi@ec.europa.eu (M. Pesaresi).
[1] Present address: Institute of Computer Science, Czech Academy of Sciences, Pod Vodárenskou věží 271/2, 182 00 Prague, Czech Republic.

https://doi.org/10.1016/j.envsoft.2025.106614
Received 9 December 2024; Received in revised form 3 July 2025; Accepted 5 July 2025
Available online 7 July 2025

ratio-scale.

As such, we assert that variants of Precision, Recall, F-score and the Jaccard index are needed to measure the agreement of gridded (or other) data representing attribute estimates at the ratio scale, while maintaining the same properties as their binary 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. continuous Precision, continuous Recall, continuous F-score and continuous Jaccard), in which agreement is interpreted as closeness of the continuous (henceforth abbreviated as 'cont.') attribute magnitude estimates. This agreement is expressed with bounded, dimensionless measures. We illustrate the usefulness of the proposed measures using examples of gridded data representing building height estimates. Moreover, we test the proposed and existing measures using a synthetic dataset with controlled disagreement, and in a realistic scenario comparing the agreement between gridded built-up surface estimates.

2. Theory

With an increasing number of available data products of continuous attribute estimates based on Earth observation data, 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). Moreover, often, accuracy of continuous data is quantified 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's correlation coefficient (r) and coefficient of determination (R^2) or slope of the least squares line (Ji and Gallo, 2006; Pontius, 2022). However, none of these measures inform about *agreement*. Each of these measures has properties which make them unsuitable when assessing the agreement of ratio scale estimates: The commonly used MD, MAD and RMSD measures of disagreement are dimensional measures, dependent on the scale and unit of assessed data, while MAPD is unstable for comparing values near zero (Ji and Gallo, 2006). Pearson's 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 Willmott 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: i) the agreement of the magnitude estimates being compared, or ii) the meaning of values equal to zero. The latter, on a ratio scale, implies the absence of the geographic feature (or, generically, of the *instance*) 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 Fig. 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. This is of particular interest when assessing gridded data on geographic features that are sparsely distributed across landscapes.

Herein, we propose measures for the agreement assessment of gridded data representing attribute estimates by adapting the measures used for binary 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. Precision, Recall, F-score and Jaccard index) 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 interpreted in terms of frequency, a degree of membership or other arbitrary measure in a 0–1 range (e.g. Uhl et al., 2023). Such values are related to the occurrence of the class, rather than to a ratio-scale attribute estimate, such as e.g. vegetation height. To our knowledge, there is no generalizable adaptation of these well-known measures to data representing non-negative continuous estimate of magnitude, for which the construction of confusion matrices is not meaningful.

Herein, we present the mathematical foundation of the proposed measures (Section 3) and illustrate that the proposed measures are viable and easily interpretable for the evaluation of gridded and other data representing ratio-scale attributes, using a range of experiments (Sections 4 and 5).

3. Calculation

Accuracy assessments rely 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

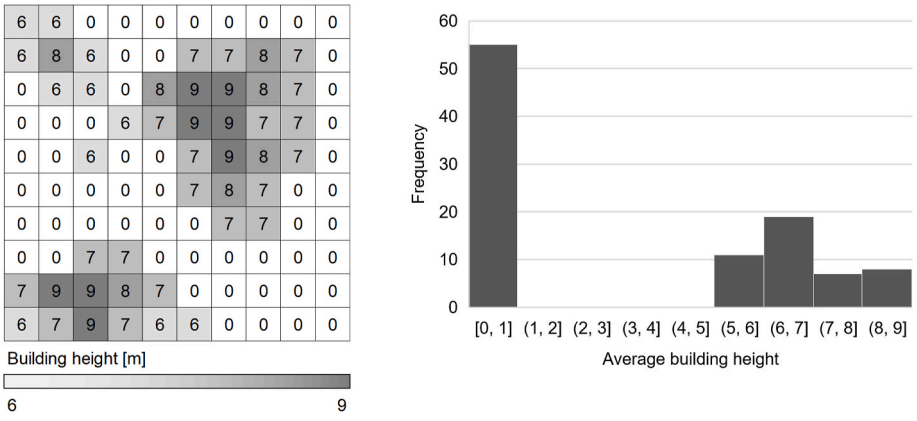


Fig. 1. Illustrating the concept of gridded ratio-scale measurements. Left: Gridded 10 × 10 dataset representing average building height per grid cell. Right: Histogram of the estimated values showing the dominating frequency of values equal to zero.

the 'model' can be any abstract (gridded or other) representation of the estimated geographic feature attribute to be evaluated referred to as the 'estimated attribute' henceforth. For binary variables, accuracy can be summarised based on a confusion matrix, where true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP) represent counts of cross-tabulated modelled and reference class elements. From these counts, a range of measures can be derived, including Precision, Recall, and the Jaccard index:

Precision, determining the fraction of modelled positive values that are correct, is defined as:

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq. (1)}$$

Recall, which in turn determines the fraction of reference presence that was classified as such, is defined as:

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq. (2)}$$

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

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

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 element-wise 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 element. We interpret their agreement as the ratio of their geometric intersection to their geometric union (Fig. 2). The intersection of two sets is interpreted as the area defined by the Minimum operator. The union of two sets is interpreted as the area defined by the Maximum operator.

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

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

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

The discussed measures are as follows:

$$cRecall = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} r_i} \quad \text{Eq. (4)}$$

$$cPrecision = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} m_i} \quad \text{Eq. (5)}$$

$$cJaccard = \frac{\sum_{i=1}^{N} min(m_i, r_i)}{\sum_{i=1}^{N} max\ (m_i, r_i)} \quad \text{Eq. (6)}$$

These measures can be interpreted similarly to their binary equivalents (see Fig. 2a). Specifically, cont. Precision can be understood as the rate of the total magnitude of the modelled attribute in agreement with the reference attribute; cont. Recall can be interpreted as the rate of the total magnitude of the reference attribute estimated by the model.

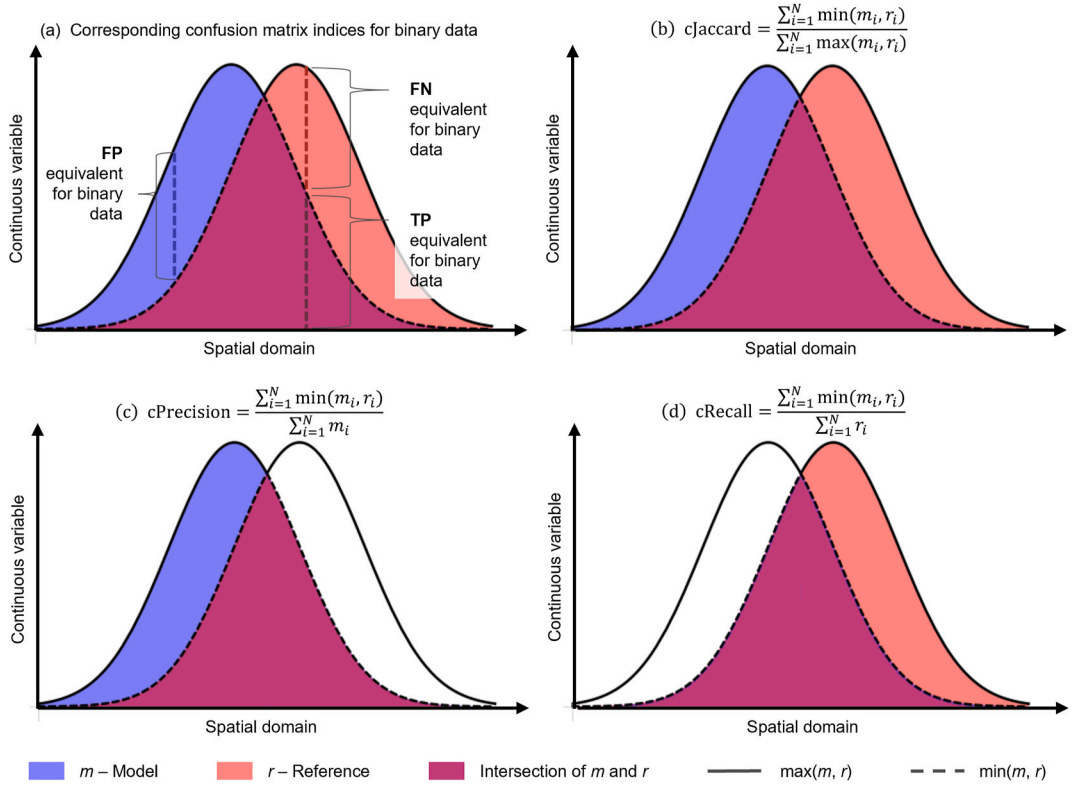


Fig. 2. Geometric representation of the concepts of intersection and union of continuous functions discretised to the spatial domain. a) Confusion matrix element equivalents on the ratio scale, b) cont. Jaccard data domain, c) cont. Precision data domain and d) cont. Recall data domain. The X-axes show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

Combined, they inform whether the model predominantly overestimates the magnitude of the attribute (cPrecision < cRecall) or underestimates (cPrecision > cRecall). 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. In the context of abundance-based measurements, cont. Jaccard is equivalent to the Ružička similarity measure of relative difference, designed to assess the similarity of ecological communities (Ružička, 1958).

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

$$F_\beta = (1+\beta^2)\frac{cRecall*cPrecision}{(\beta^2*cPrecision)+cRecall} \quad \text{Eq. (7)}$$

Where β factor defines model selection. 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:

$$F_0 = cPrecision;\ \ F_\infty = cRecall \quad \text{Eq. (8)}$$

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. F-score).

4. Materials and methods

We showcase the usefulness of the proposed measures using a small toy dataset (Section 4.1). Moreover, we assess their response using pairs of larger, synthetic datasets with different levels of disagreement (Section 4.2); and analyse their response to systematic variations in the continuous spatial variable using gridded built-up surface estimates (Section 4.3).

4.1. Showcase examples

In this experiment, we design three pairs of gridded reference and modelled toy datasets. These datasets represent the average height of buildings per grid cell, to establish a link to remote-sensing related applications. For each pair of datasets, we report the proposed agreement measures (cont. Precisions, cont. Recall, cont. F-score and cont. Jaccard). To demonstrate the added value of these measures, we compare them to commonly used measures of difference (Mean Error, ME, and Mean Absolute Error, MAE, corresponding to MD and MAD, respectively) and association (r and Slope) (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. We use a series of variables X from Table 8.1 from Pontius (2022) to populate grid cells in the reference dataset, and 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 (Fig. 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 with 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 (Fig. 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 (Fig. 3, example C).

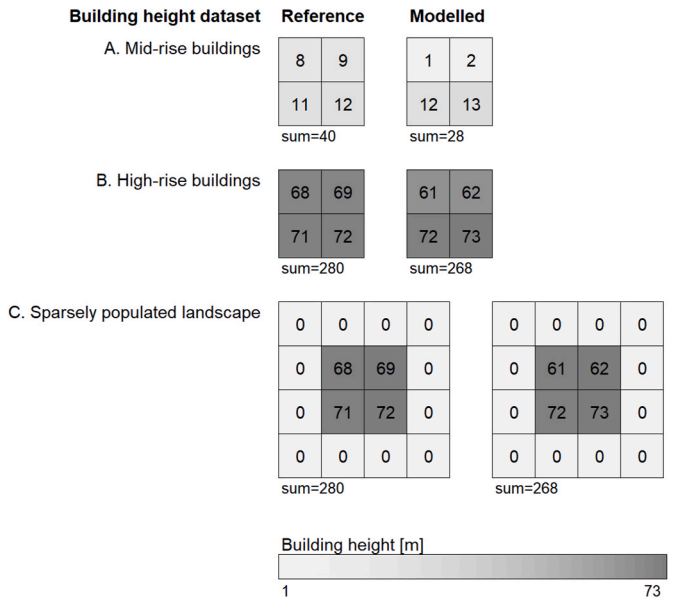


Fig. 3. Toy data representing average building heights: A. of mid-rise buildings, with low estimated values; B. of high-rise buildings, with higher values than in example B but with the same levels of association and error between the reference and modelled datasets; C. the same values as in example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

4.2. Experiment using synthetic data

In the experiment using synthetic data we use an artificial landscape, representing estimates of crop canopy height (Fig. 4a, left), generated with Gaussian window functions, adding noise using a two-dimensional Gaussian filter. Here, we use a grid of 1000 × 1000 cells with continuous values between zero and two (approximating corn canopy height, see Appendix B for details).

We consider the produced synthetic landscape as the reference dataset and generate corresponding modelled datasets by adding varying levels of absolute bias, relative bias, and random noise to the reference data (Fig. 4). To inject absolute bias, we add to each grid cell a value between −0.5 and 0.5 in a series of runs; to inject relative bias, we multiply each grid cell with value from 0.5 to 1.5; and to inject random noise, we add to each grid cell a value drawn from a normal distribution with a mean of zero and a standard deviation (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 proposed 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 data (Section 5.1).

4.3. Realistic experiment with controlled error

To illustrate the utility of the proposed measures in a realistic context, we use gridded data representing the built-up surface density for a region in Sao José dos Campos (Brazil) at 10 m resolution, derived from building footprint data (Copernicus Emergency Mapping Service, 2 (2022). In this experiment, we assess the response of the proposed measures to controlled injection of commission and omission errors of varying magnitudes.

The gridded built-up surface density dataset represents our reference data (Fig. 5a). By modifying a copy of this dataset, we produce a corresponding modelled dataset, injecting omission errors, by setting a selection of grid cells with low values to 0, and commission errors, by replacing selected grid cells with values of zero by positive, low values

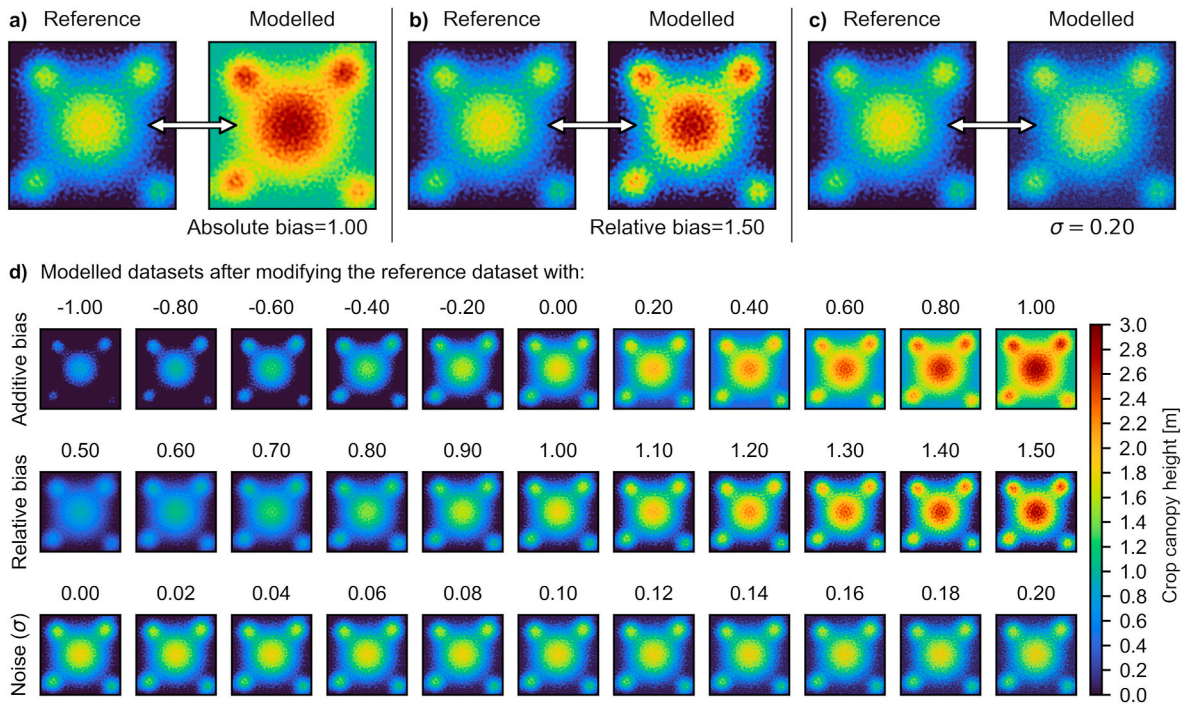


Fig. 4. Pairs of synthetic reference and modelled data generated for three types of disagreement: (a) absolute bias, (b) relative bias and (c) random noise. Panel (d) shows modelled datasets generated in a series of runs.

(Fig. 5b). The resulting pair of data represents a realistic case of a model performance: High built-up surface densities dominate the area of overlap, while lower density values are found in areas of omission and commission. We consider this our baseline scenario.

We binarize the baseline datasets using a cutoff value of 0 and compute the confusion matrix elements: FN, FP and TP. We then define three spatial domains based on the binary classification: an omission area defined by FN grid cells, a commission area defined by FP grid cells and an overlap area defined by TP grid cells (Fig. 5c).

We systematically increase the continuous grid cell densities in reference and modelled data within the three domains in a permutation-based approach: For example, we increase the reference density in the omission area by factor 2 (Fig. 5d), and at the same time, we increase the modelled density in the commission area by factor 2 (Fig. 5e). Each combination of these two states (i.e. using (i) the assigned baseline values, or (ii) the increased values, which are the baseline values multiplied by 2) yields a different scenario of density levels in reference and modelled data in the three domains. These 16 possible scenarios are shown in Fig. 5e and in Appendix Figure C.1.

For each of these 16 scenarios we compute the proposed 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 (Section 5.3).

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 agreement measures 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 observe an overestimation of only 1 m in the cells in the bottom row, with estimates of 12 and 13 m instead of 11 and 12 m.

Secondly, the proposed measures underline the importance of the relativity of commission and omission errors to the magnitude of the estimated attribute. While the datasets in examples A and B have by design the same level of association and error, their agreement strongly differs (Table 1, ***examples A and B)***. The value of cont. Jaccard estimated in example B (cJaccard = 0.94) is substantially higher than in example A (cJaccard = 0.62). Moreover, the cont. Recall shows a difference of 30 percentage points in estimated magnitude allocation between these two examples (cRecall = 0.65 in example A, cRecall = 0.95 in example B). This observation has, for example, implications when choosing a model over another: In this example, the accuracy values for high-rise building are larger than the corresponding accuracy values for mid-rise buildings because the agreement computed by the minimum function is larger for high-rise buildings than for 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. for sparsely distributed geographic features), the proposed measures (cont. Jaccard, cont. Precision, cont. Recall and cont. F-score) capture the agreement in the presence of reference and modelled estimates of the attribute magnitudes.

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

Results of the experiment using synthetic data (see Section 4.2) show that the proposed cont. Jaccard, cont. Precision, cont. Recall and cont. F-score are sensitive to both bias and noise, as desired (Fig. 6). The cont. Precision and cont. Recall measures are indicators of over- and under-estimation of the estimated magnitude, whereas the proposed cont. Jaccard is a symmetric measure of closeness between two magnitude estimates. These measures are not inflated by the agreement on the absence of the geographic feature, in contrast to the other tested

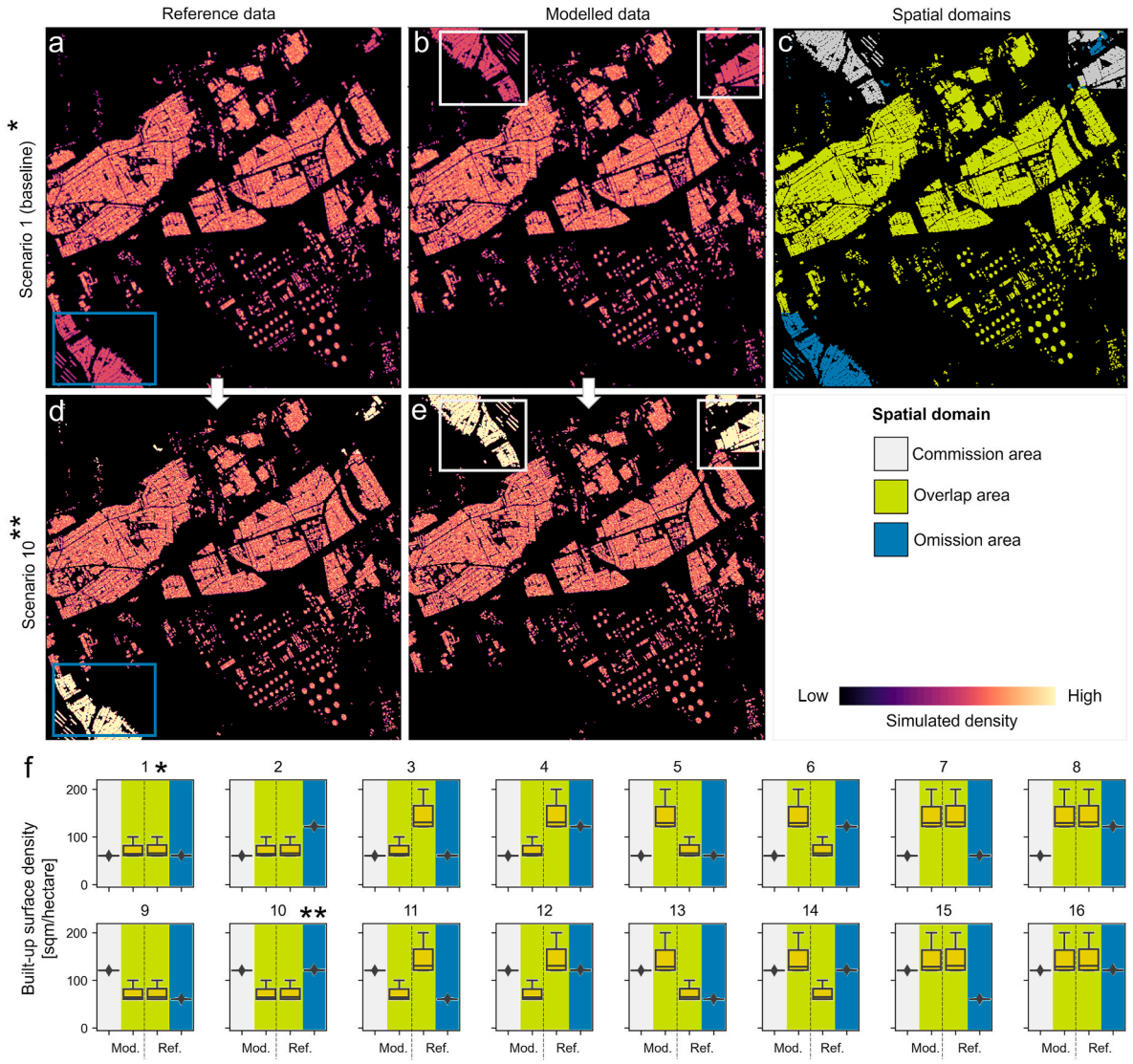


Fig. 5. Illustrating the input data for a realistic experiment with controlled error using gridded built-up surface density data. (a) Baseline reference data, (b) baseline modelled data, (c) derived overlap, commission and omission areas. Panel (d) shows the reference data from (a) with increased densities in the omission area (highlighted by turquoise boxes), while panel (e) shows the modelled data with increased densities in the commission area (highlighted by light-grey boxes). Panel (e) shows the density distributions of modelled and reference data in the three domains. Asterisks link the gridded data shown in (a), (b), (d) and (e) to their corresponding scenarios in (f).

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	cJaccard	cPrecision	cRecall	cF-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

measures, where the imposed lower bound at value zero lowers the estimated (absolute) measure values (Fig. 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 reflects the magnitude of the estimated attribute mapped in both datasets. This is in contrast to measures reporting averaged error (MAE, ME, RMSE) or the correlation coefficient r, reporting about association, which is indifferent to bias.

5.3. Response of the proposed measures to variations 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 (Fig. 7a), while the binary agreement remains constant (Jaccard = 0.81, Precision = 0.91, Recall = 0.87).

The scenario with high densities and high agreement in the overlap area yields the highest values for our continuous agreement measures (Fig. 7a, scenario 7). Lowest values are found for scenarios 12 and 14, where the disagreement in the overlap area is high, and in addition to

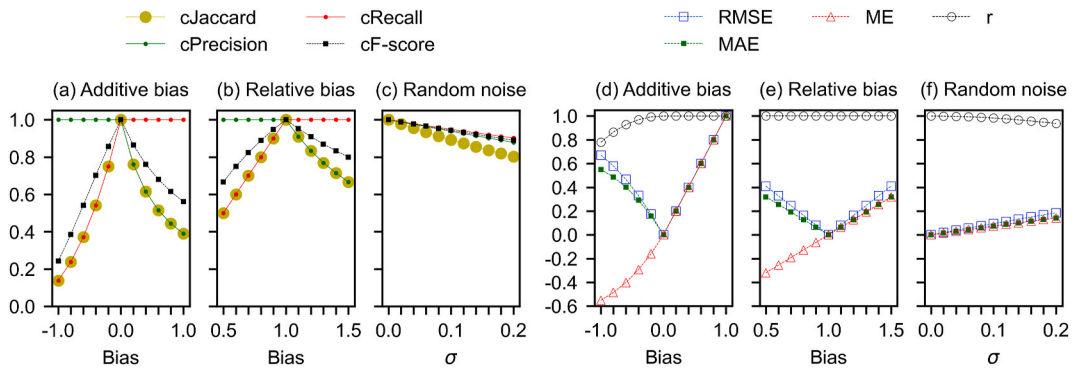


Fig. 6. Agreement measures computed for different pairs of synthetic landscapes. Shown is the response of the proposed agreement measures (cJaccard, cPrecision, cRecall and cF-score) for landscape pairs reflecting (a) additive bias, (b) relative bias, and (c) random noise. Panels (d) to (f) show agreement measures typically used for accuracy assessment of continuous data: r and error measures MAE, ME and RMSE for the same sets of landscapes.

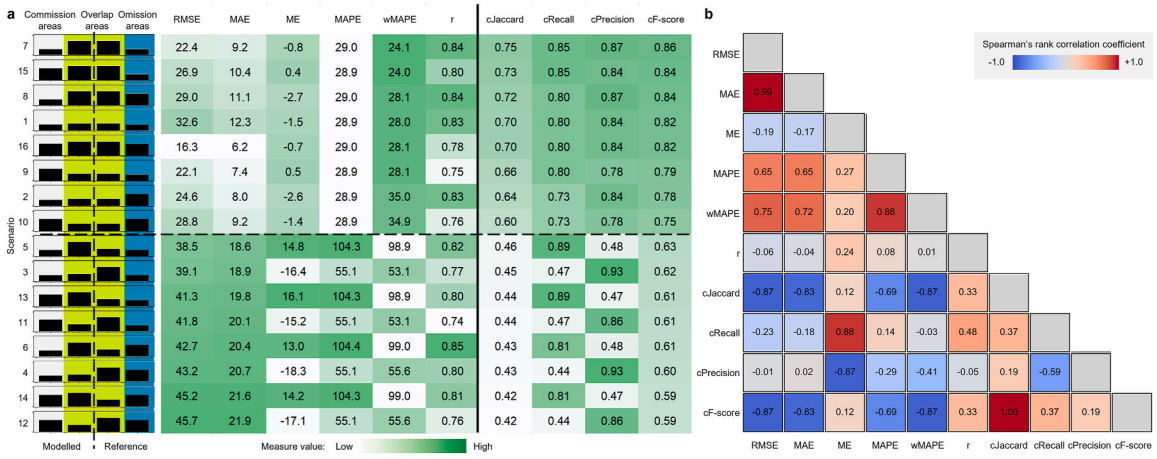


Fig. 7. Comparison of agreement measures for 16 scenarios of systematically modified data in the overlap, commission and omission domains. (a) Values of commonly used (left part) and proposed (right part) measures for each scenario, sorted by the cont. Jaccard index. Numbering of the scenarios corresponds to Fig. 5f. The dashed line separates scenarios by their level of agreement in the overlap domain (above dashed line = agreement in overlap domain, below dashed line = disagreement in overlap domain). Bar plots to the left indicate the average density values for each scenario within each domain (colours correspond to Fig. 5c). Panel (b) shows the cross-correlation matrix of the ten measures, based on the data shown in (a). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

that, the densities in the commission and omission areas are high as well.

Cont. Precision and cont. Recall measures are responsive to the density variations, in particular within the overlap area, and to a lesser degree, within the commission and omission areas. This is due to the imbalanced spatial support between the overlap, commission and omission domains (Fig. 5a; number of grid cells: overlap area = 78 272, commission area = 7 245, omission area = 11 005). As an example, we compare scenarios 7 and 8: the density increase in the omission area 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 overlap areas (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 (cRecall = 0.47, cPrecision = 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, bounded to the range [0, 1], mostly correlates with RMSE and MAE (Fig. 7a). 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 overlap areas is similar for the modelled and reference densities (i.e. scenarios above the dashed line, Fig. 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 commission and omission areas, 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 tested measures of difference are very useful for quantifying *individual components* of the differences between the data compared. Specifically, the proposed cont. Jaccard summarises RMSE, MAPE, and MAE, and cont. Precision and cont. Recall disentangle them. Thus, the proposed measures represent valuable alternative measures of agreement. Looking at cross-correlations between measures (Fig. 7b), we observe a wide range of variability. Notably, ME correlates highly (positively and negatively) with cont. Recall, and cont. Precision, respectively. This highlights the capability of these measures to decompose overall, signed measures such as the ME into distinct components of omission and commission, providing an interpretable measure of these error components.

6. Conclusions

In this study, we presented four measures for assessing the agreement

of gridded (and other) data representing continuous estimates of attributes at the ratio scale: continuous Jaccard, continuous Precision, continuous Recall, and continuous F-score. These measures were applied and tested in a range of experiments. We establish that due to its robustness and interpretability, the continuous Jaccard measure, an extension of the widely recognized IoU agreement measure, is a practical method for comparing datasets of attributes at the ratio scale, which include absolute or relative estimates (e.g. canopy height or built-up surface density). Additionally, we illustrated that continuous Precision and continuous 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 of attribute values.

The proposed measures are easily interpretable due to the similarity to their well-known binary 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 make them suitable measures for estimating the accuracy of 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 equip researchers and analysts with a toolset to evaluate spatial-environmental data consisting of increasingly common continuous, rather than binary measurements. Importantly, the usefulness of these measures extends far beyond the context of gridded, spatial, and environmental data, as they can be applied to any ratio-scale variable, i.e. non-negative continuous variables with an absolute zero. These measures complement the suite of existing agreement measures, and ultimately, aim to contribute to increase uncertainty awareness among practitioners (Goch et al., 2023), enabling informed interpretation of geospatial data and beyond.

CRediT authorship contribution statement

Katarzyna Krasnodębska: Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Wojciech Goch:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Johannes H. Uhl:** Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Judith A. Verstegen:** Writing – review & editing, Supervision, Methodology. **Martino Pesaresi:** Writing – review & editing, Supervision, Conceptualization.

Code availability

- Name of software: continuous-jaccard
- Developers: Katarzyna Krasnodębska, Johannes H. Uhl and Martino Pesaresi
- Contact: katarzyna.krasnodebska@twarda.pan.pl and JRC-GHSL-DATA@ec.europa.eu
- Date first available: January 14, 2025
- Software required: None
- Program language: Python and R
- Source code at: https://code.europa.eu/jrc-ghsl/continuous-jaccard
- Documentation: Brief description of code stored in the repository can be found at https://code.europa.eu/jrc-ghsl/continuous-jaccard/-/blob/main/README.md
- No data for local installation and use of software is required.

Data availability statement

Data to reproduce experiments are available with the code at https://code.europa.eu/jrc-ghsl/continuous-jaccard/-/tree/main/publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the reviewer for their thoughtful comments and suggestions, which improved the clarity and validity of this article. Their guidance, including references to relevant literature, were invaluable in enhancing the quality of this work.

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 (JRC). Part of the work was carried out at the Centre for Complex Systems Studies (CCSS) at Utrecht University, supported by the Swaantje Mondt travel fund.

Appendices

Apendix to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106614.

Data availability

All datasets and codes used in this study are available in the repository at https://code.europa.eu/jrc-ghsl/continuous-jaccard, as described in the Code Availability Statement and Data Availability Statement.

References

Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern Recognit. Lett. 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. Rem. Sens. Environ. 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4.

Goch, K., Verstegen, J.A., Uhl, J.H., 2023. Influence of uncertainty in reference data on the validation of multi-temporal settlement layers. In: Proceedings of the 2023 Conference on Big Data from Space. Presented at the BIDS 2023. Publications Office of the European Union, Luxembourg, Vienna, pp. 269–272. https://doi.org/10.2760/46796.

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: Int. J. Appl. Forecast. 40–43.

Lewis, H.G., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. Int. J. Rem. Sens. 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. Rem. Sens. Environ. 209, 90–106. https://doi.org/10.1016/j.rse.2017.12.020.

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. Int. J. 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. Int. J. Geogr. Inf. Sci. 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. Int. J. Rem. Sens. 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. Rem. Sens. Environ. 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). Biol. Bratisl. 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. Rem. Sens. Environ. 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. Int. J. Climatol. 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. Phys. Geogr. 1, 59–73. https://doi.org/10.1080/02723646.1980.10642189.