A FREE AND OPEN SOURCE TOOL TO ASSESS THE ACCURACY OF LAND COVER MAPS: IMPLEMENTATION AND APPLICATION TO LOMBARDY REGION (ITALY)

G. Bratic [1], M. A. Brovelli [1], M. E. Molinari [1], *

[1] Politecnico di Milano, Department of Civil and Environmental Engineering, Piazza Leonardo da Vinci 32, 20133 Milan, Italy - gorica.bratic@mail.polimi.it, (moniaelisa.molinari, maria.brovelli)@polimi.it

Commission IV, WG IV/4

KEY WORDS: Global Land Cover, Accuracy Assessment, Free and Open Source, Error Matrix, Accuracy Measures

ABSTRACT:

The availability of thematic maps has significantly increased over the last few years. Validation of these maps is a key factor in assessing their suitability for different applications. The evaluation of the accuracy of classified data is carried out through a comparison with a reference dataset and the generation of a confusion matrix from which many quality indexes can be derived. In this work, an ad hoc free and open source Python tool was implemented to automatically compute all the matrix confusion-derived accuracy indexes proposed by literature. The tool was integrated into GRASS GIS environment and successfully applied to evaluate the quality of three high-resolution global datasets (GlobeLand30, Global Urban Footprint, Global Human Settlement Layer Built-Up Grid) in the Lombardy Region area (Italy). In addition to the most commonly used accuracy measures, e.g. overall accuracy and Kappa, the tool allowed to compute and investigate less known indexes such as the Ground Truth and the Classification Success Index. The promising tool will be further extended with spatial autocorrelation analysis functions and made available to researcher and user community.

1. INTRODUCTION

Thanks to the continuous advance in remote sensing and mapping technologies, the availability of land use/land cover (LULC) maps has considerably grown over the last few years. These datasets provide valuable information in several fields related to environmental studies and land resource monitoring, and they are frequently released under open access licenses for research purposes. Obviously, being aware of the classification accuracy of LULC maps is a key factor to evaluate their suitability for the various applications where they are exploited. The accuracy assessment of digital remotely-sensed data started around 1975 and still represents an important research topic. Many recommendations and guidelines have been published over the years suggesting different approaches. First assessments were performed thanks to a simple visual checkup based on the “looking good” requirement. Afterward, the need for a reliable evaluation gave rise to the non-site-specific assessment approach (Meyer et al., 1975) which was performed by comparing the areal extent of land use classes for classified and ground truth datasets. While providing information about LULC correctness in terms of proportion of land use classes, this method was not able to extract any information about the location errors. To avoid this limitation the site-specific assessment, and particularly the error matrix technique (Congalton and Green, 1999), has spread. According to Lunetta and Lyon (2004), since the mid-1980s the error matrix (or confusion matrix) has been considered as “the standard descriptive reporting tool for accuracy assessment of remotely sensed data”. Derived from a comparison between a classified dataset and a reference one, this matrix represents the starting point from which to extract many useful indexes able to describe agreements and disagreements between the two considered datasets. These indexes span from the most commonly used *overall accuracy*, *user’s accuracy* and *producer’s accuracy* (Story and Congalton, 1986), to the more complex *Individual Classification Success* (Koukoulas and Blackburn, 2001) and *Ground Truth* (Türk, 1979) indexes.

Since it enables the comparison of two sources of spatial information, the error matrix computation represents a key tool for Geographical Information System (GIS) software, which are the most used tool for practical processing and analysis of spatial data. However, these software currently focus on the simple confusion matrices computation and they provide very few indexes, usually the most common used ones. As an example, ArcGIS (ESRI, 2018), probably the most widely known GIS proprietary software, allows the user to derive *omission* and *commission* errors and the *Kappa* index. Shifting the focus on the Free and Open Source Software (FOSS), which is the object of this study, the Accuracy Assessment Plugin of QGIS (QGIS Development Team, 2018) provides *user’s* and *producer’s accuracies* and *allocation* and *quantity disagreements* while the r.kappa module of GRASS GIS (GRASS Development Team, 2018) calculates *commission* and *omission errors*, *overall accuracy* and *kappa* statistics.

This work fits within this context and proposes the development of a new free and open source tool that can be easily integrated into GIS systems and enables users to automatically calculate all the statistics based on confusion matrix proposed by literature.
The potential of the tool has been tested and demonstrated in a case study related to the accuracy assessment of three different high resolution LULC maps (GlobeLand30, Global Urban Footprint, Global Human Settlement Layer Built-Up Grid) in the area of Lombardy Region (Northern Italy).

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-87-2018 | © Authors 2018. CC BY 4.0 License.

The remainder of the paper has four parts. Section 2 provides a description of the tool and lists all the confusion matrix-derived measures considered; section 3 focuses on the case study application by illustrating the main characteristics of the datasets and the adopted methodology for accuracy assessment analysis; section 4 discusses the obtained results. Finally, section 5 presents conclusions and future directions of the study.

2. TOOL DESCRIPTION

The computation of the indexes has been implemented as FOSS stand-alone tool taking advantage of Python programming language, which makes easier the integration within GIS environments (e.g. GRASS GIS or QGIS software packages) and their powerful spatial analysis functionalities. The tool, mainly based on *numpy* and *pandas* Python libraries, outputs a *csv* file with a set of accuracy measures derived from the confusion matrix provided as input by the users.

This study performed an accurate literature review to identify the existing quality measures that can be derived from a confusion matrix. *Overall accuracy* certainly represents the most common and simplest descriptive statistic and indicates the percentage of correctly classified samples. Since it provides a global evaluation of the dataset classification quality, this statistic is usually integrated with per-class accuracy indexes such as *user’s* and *producer’s accuracy*. The former identifies the probability of a reference sample unit being correctly classified while the latter identifies the probability that a sample unit classified on the map represents that category on the ground. Besides them, other less common but strictly correlated statistics can be mentioned. Fung and LeDrew (1988) proposed the *average of user’s and producer’s accuracies* while Nelson (1983) introduced the *combined user’s or producer’s accuracies*, used to dampen the inherent biases of the overall and average accuracies. *Hellden’s mean accuracy* is the harmonic mean of *user's* and *producer's accuracy* (Liu, 2007) and can be interpreted as a measure of overlapping between true and estimated classes. The *Short’s mean accuracy* (Labatut and Cherifi, 2011) represents instead the ratio of the estimated and true classes intersection to their union. Koukoulas and Blackburn (2001) proposed the *Individual Classification Success Index*, which reflects the classification effectiveness of a class as the average between *user’s* and *producer’s accuracy*. This index can be used to calculate the overall classification effectiveness by averaging its values for all categories (*Classification Success Index*) or some of them (*Group Classification Success Index*).

Despite the harsh criticism of many authors, i.e. Brennan and Prediger (1981), Stehman, (1997) and Foody (2008), also the use of Kappa global statistic still continues to be pervasive in matrix confusion-based accuracy assessment; for this reason all the Kappa-like statistics have been considered within the implemented tool. This group of statistics differs from overall accuracy since it takes into account the so-called “chance agreement” component that is calculated in different ways. Standard *Kappa* coefficient identifies the agreement that is expected when the raters are totally independent while *conditional Kappa* expresses the same “chance agreement” at per-classes level. The *weighted Kappa* (Cohen, 1968) is weighted according to the importance of the errors while *Tau* index (Ma and Redmond, 1995) calculates the “chance-agreement” based on prior probability of class membership. *Aickin’s alfa* (1990) assumes that the population of samples includes easy-to-classify and hard-to-classify items, out of which, only the latter is classified by chance. Finally, *Ground truth* index supposes that classifier includes an always correct component and a randomly correct one, which corresponds to the “chance-agreement”.

Margfit is a method proposed by Congalton and Green (1999) based on iterative proportional fitting with the aim to normalize the matrix. In this normalization process, differences in sample sizes used to generate the matrices are eliminated and, therefore, individual cell values within the matrix are directly comparable.

Finally, the last group of considered measures is proposed by Pontius and Millones (2011) that suggested the estimation of the *disagreement index* and its two components, i.e the *quantity* and *allocation disagreements*. These indexes identify the amount of difference between the reference map and a comparison map due to the less than perfect match in the proportions and the spatial allocation of the categories, respectively.

3. CASE STUDY: DATA AND IMPLEMENTATION

The case study selected for the testing of the implemented tool is the area corresponding to Lombardy Region (Northern Italy). In the following, information about considered datasets and data processing are provided.

3.1 Datasets

The implemented tool has been applied to evaluate the classification accuracy of three recently proposed high-resolution LULC maps, GlobeLand30 (hereafter GL30), Global Urban Footprint (hereafter GUF) and Global Human Settlement Layer Built-Up Grid (hereafter GHS).

The GL30 is a product of “Global Land Cover Mapping at Finer Resolution” project led by the National Geomatics Center of China (NGCC). It is a land cover dataset at 30m resolution available for the two timeline years of 2000 and 2010. It has been generated through the classification of multispectral images of Landsat Thematic Mapper (TM) and Enhanced TM plus (ETM+) satellites and China Environmental Disaster Alleviation Satellite (HJ-1). The dataset has been created by means of the pixel-object-knowledge-based (POK-based) classification approach (Chen et al., 2014). The classification system includes 12 land cover types, namely: *cultivated land*, *mixed forest*, *broadleaf forest*, *coniferous forest*, *grasslands*, *shrublands*, *wetlands*, *water*, *tundra*, *artificial surfaces*, *bare lands*, *permanent snow and ice*. The dataset used within this work is referred to 2010 and has been provided as GeoTiff dataset in WGS84 (World Geodetic System 1984) reference system and UTM (Universal Transverse Mercator) zone 32N projection (EPSG: 32632).

The GUF (Esch et al., 2017) is a global mask of built-up areas at a resolution of 0.4 arc second (about 12m at the equator). It has been generated taking advantage of the dedicated Urban Footprint Processor (UFP) implemented at the German Aerospace Center (DLR). The UFP has been applied on approximately 180.000 single TerraSAR-X/TanDEM-X image products for the reference year 2011. The dataset is provided as thematic raster map in GeoTiff format, WGS84 reference system (EPSG: 4326), and classification based on three values: 255 for built-up areas, 0 for non-built-up areas and 128 for missing data. According to DLR, a built-up area is defined as “a region featuring man-made building structures with a vertical component”.

The GHS is one of the products resulting from the Global Human Settlement Layer (GHSL) project carried out by the Joint Research Center (JRC) with the aim to provide global spatial information about the human presence on the planet over time. The dataset contains multitemporal information layers on built-up presence derived from Landsat image collections; it has been produced by means of Global Human Settlement Layer methodology (Pesaresi et al., 2016). The dataset selected for the present work is the GHS product related to 2014 characterized by a resolution of around 38m, available in GeoTiff format and Google Mercator Projection (EPSG:3857). The map is based on three values: 1 for non-built-up areas, 101 for built-up areas, and 0 for missing data. Similarly to GUF, GHS defines built-up areas as “the union of all the spatial units collected by the specific sensor and containing a building or part of it”.

A key point for the accuracy assessment of the three land cover maps is the comparison with a reference dataset characterized by a greater detail. This requirement is satisfied by DUSAF, a land cover database created in 2000–2001 for Lombardy Region (Credali et al., 2011). DUSAF consists of land cover vector maps at 1:10,000 scale referred to different time periods. This data is provided as Shapefile in WGS84 reference system, with the UTM zone 32N projection (EPSG: 32632). The adopted legend is structured in five hierarchical levels of detail. The first three levels comply with the Corine Land Cover nomenclature and the most general one consists of five classes: *artificial surfaces*, *agricultural areas*, *forest and semi natural areas*, *wetlands*, and *water bodies*. For the present work the DUSAF 4.0 related to 2012 has been selected.

3.2 Implementation

The data processing workflow carried out to perform the validation of each classified map (GL30, GUF, GHS) with respect to the reference (DUSAF) is presented in Figure 1. Since the data to be compared may be different in terms of reference system, format, resolution and thematic legend, some processing steps to harmonize them were required. More in detail, classified maps having a different reference system with respect to DUSAF were re-projected to WGS84/UTM32N while DUSAF was rasterized according to the resolution of the classified map considered for the comparison.

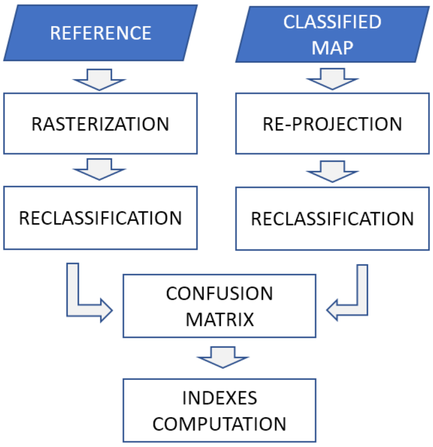

Figure 1. Data processing workflow

The third step of the processing aims to provide a common set of land cover classes between the maps under comparison. To that goal a binary reclassification (non-built-up, built-up) was performed to DUSAF for comparing with GUF or GHS. More in detail, the DUSAF classes corresponding to *continuous and discontinuous urban fabric* and *industrial*, *commercial*, *public and private units* were reclassified as built-up and all the others as non-built-up. For the comparison between GL30 and DUSAF, two different methods were considered. In the first case the classes of DUSAF have been reclassified according to GL30 thematic legend based on eleven classes (as proposed in Table 1). Vice versa, in the second case the GL30 has been reclassified to the five first-level classes of DUSAF as shown in Table 2. After these processing steps, the datasets are comparable, and the confusion matrix and the accuracy measures were computed.

The implementation of the processing workflow was performed by taking advantage of the re-projection, rasterization and reclassification modules provided by GRASS GIS, which also has a specific module to compute the error matrix between two raster maps. The tool for indexes calculation was integrated as a GRASS GIS script in such a way to automate the procedure.

DUSAF classes	GLOBELAND30 classes
Agricultural Areas	Cultivated land
Mixed Forest	Mixed forest
Broad-leaved forest, Recent afforestation	Broadleaf forest
Coniferous forest	Coniferous forest
Natural Grassland	Grasslands
Moors and heathland, Transitional woodland/shrub	Shrublands
Wetlands	Wetlands
Artificial Areas	Artificial surfaces
Beaches, dunes and sand planes, Bare Rock, Sparsely vegetated areas	Bare lands
Water Bodies	Water
Glaciers and perpetual snow	Permanent snow and ice

Table 1. Rules adopted to reclassify DUSAF according to GlobeLand30 classes

DUSAF classes	GLOBELAND30 classes
Artificial surfaces	Artificial surfaces
Agricultural areas	Cultivated land
Forest and semi natural areas	Broadleaf forest, Coniferous forest, Mixed forest, Grasslands, Shrublands, Bare lands, Permanent snow and ice
Wetlands	Wetlands
Water bodies	Water

Table 2. Rules adopted to reclassify GlobeLand30 according to DUSAF first-level classes

4. RESULTS

Table 3 shows the global accuracy measures obtained for the three evaluated land cover datasets. All data ranges from 0% (no agreement) to 100% (perfect agreement); only for disagreement parameters a value of 0% indicates perfect agreement and a value of 100% identifies no agreement.

As explained in the previous section, the accuracy assessment of GL30 has been performed taking into account two different classification methods based on 11 classes (hereafter GL30-11) and 5 classes (hereafter GL30-5), respectively.

Regarding the GL30-11 analysis, results show a value of *overall accuracy* equal to 73.4%. Since the present work has not the aim to evaluate the dataset according to a target application, the judgement of the goodness of this result is a challenging task. Literature provides very different thresholds to define the acceptable value of *overall accuracy*, e.g Pringle et al. (2009) define satisfying an overall accuracy of 70% while Anderson et al. (1976) require a value of at least 85%. The analysis of the other statistics, such as the K-like indexes, suggests that the classification quality is not high; in fact, according to Landis and Koch (1977) a value of *Kappa* coefficient equal to 64.4% suggests a moderate agreement between GL30-11 and DUSAF. The same behaviour is highlighted by the *Classification Success Index*, which is equal to 61.2%, much lower with respect to the optimal threshold value proposed by Koukoulas & Blackburn (2001). The value of *Group Classification Success Index*, which has been obtained by excluding *mixed forest*, *shrubland*, *bareland* and *permanent ice and snow* classes from the computation, is equal to 77.5% and suggests that some of the removed land cover classes are the main responsible of the less than optimal classification. Finally, the disagreements measures show that the most of incoherence between GL30-11 and DUSAF is mostly due to allocation component (19.6%) with respect to the quantitative one (7%).

Index	GL30-5	GL30-11	GHS	GUF
Overall accuracy	86.2	73.4	94.2	95.5
Average user’s accuracy	78.3	62.3	86.6	90.2
Average producer's accuracy	73.2	60.1	84.5	87.7
User’s combined accuracy	82.2	67.8	90.4	92.9
Producer’s combined accuracy	79.7	66.7	89.3	91.6
Classification Success Index	75.8	61.2	85.6	88.9
Group Classification Success Index	72.2	77.5	-	-
Kappa coefficient	78.1	64.4	71.1	77.8
Conditional Kappa	73.9	59.2	71.2	77.9
Weighted Kappa	77.7	37.0	-	-
Tau	86.2	70.7	94.2	95.5
Alpha	78.1	68.7	83.5	87.4
Margfit	87.7	67.0	90.3	92.7
Disagreement	13.8	26.6	5.8	4.5
Allocation disagreement	9.8	19.6	5.1	3.6
Quantity disagreement	4.0	7.0	0.7	0.9

Table 3. Case studies results: global accuracy measures

Figure 2 reports the cumulative values of the computed individual class accuracy measures; here, each index value is ranging from 0 (no agreement) to 1 (perfect agreement). The proposed statistics clearly identify *shrubland*, *mixed forest*, *wetlands* and *grasslands* as the classes with a significantly lower level of accuracy.

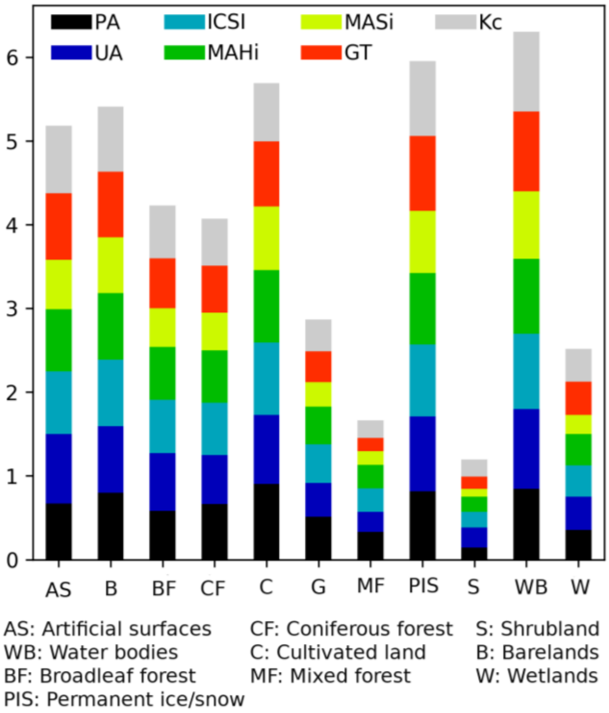

Figure 2. GL30-11 analysis results: values of per-class indexes (producer’s accuracy *PA*; user’s accuracy *UA*; Individual Classification Success Index *ICSI*; Hellden’s mean accuracy *MAHi*; Short’s mean accuracy *MASi*; Ground Truth *GT*; Conditional Kappa *Kc*)

The case study related to GL30-5 merged in a single land cover class (*forests and seminatural areas*) most of the categories characterized by a low classification accuracy in GL30-11 analysis. The adoption of a less detailed classification legend leads to an increasing of the quality of global indexes (Table 3); namely, the *overall accuracy* (86.2%), the K-like indexes, and the *classification success index* are significantly higher with respect the previous case and exceed or are very close to the thresholds suggested by literature as satisfying values. Information extracted from disagreement measures confirms that the allocation component is the primary cause of incoherence between classes. The cumulative values of individual class accuracy measures (Figure 3) show that *wetland* is the only poorly classified land cover class.

Regarding the accuracy assessment of GUF and GHS datasets, very similar results have been obtained. All the indexes undoubtedly identify a very good classification quality since they are almost always higher than 75% and some measures reach the 95%, e.g. the *overall accuracy*. As observed in the previous case studies, the low values of disagreement are mostly related to the allocation component.

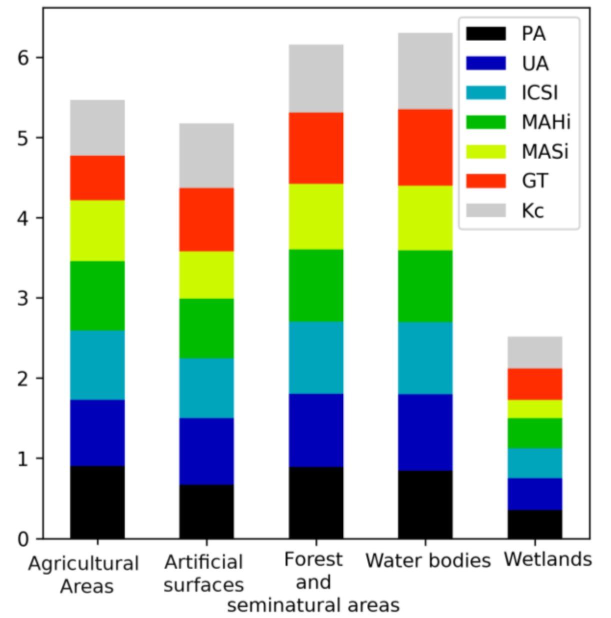

Figure 3. GL30-5 analysis results: per-class indexes (producer’s accuracy *PA*; user’s accuracy *UA*; Individual Classification Success Index *ICSI*; Hellden’s mean accuracy *MAHi*; Short’s mean accuracy *MASi*; Ground Truth *GT*; Conditional Kappa *Kc*)

5. CONCLUSIONS

According to the “good practices” suggested by many authors, confusion matrix and its derived statistics represent the standard approach for the accuracy assessment of the classification of remotely-sensed data. Although many GIS software packages provide modules for confusion matrix computation, a comprehensive tool enabling the calculation of all the accuracy indexes proposed by literature is not yet available.

In the present work, a detailed investigation about confusion matrix-derived indexes was performed and a Python FOSS module was implemented to facilitate their automatic computation. The tool was successfully applied to evaluate the classification accuracy of three high-resolution LULC datasets, i.e. GL30, GHS, and GUF on the area of Lombardy Region (Italy). The assessment was performed through a comparison between each dataset and a reference map, i.e DUSAF. The analysis of results suggests a very satisfactory accuracy of GUF and GHS built-up datasets. Regarding the GL30 dataset, a high overall accuracy is obtained by considering the classification system based on the first-level classes of DUSAF nomenclature. Instead, the accuracy decreases if a more detailed thematic legend is considered, especially for the classes related to vegetation (*shrubland*, *grassland*, *mixed forests*). Obviously, the correct matching between two detailed classification systems represents a challenging task, thus errors due to interpretation of ambiguous classes may have been introduced during the processing phase.

Some improvements to the tool implemented in this work are planned in the future. Investigations are ongoing to extend the tool with additional functions able to detect any patterns of error in discrepancies between the LULC products. To this purpose, different spatial autocorrelation measures, e.g. *Moran’s I* (Moran, 1950), *Geary’s C* (Geary, 1954), *join counts* (Moran, 1948), *Getis-Ord G* (Getis and Ord, 1992), have been explored to analyse the geographical distribution of errors in the classified maps. Currently the work is focused on the *join counts* statistic, which is not easy to handle when there is a number of categories larger than 2 or 3. *Join counts* tests have been performed on land cover maps involving just two classes (binary classification). Further experiments are planned in the short term to apply the *joint counts* or other autocorrelation tests to enable multivariate classification error pattern detection.

Finally, future development of the work will consist also in the integration of the tool as an Addons for GRASS GIS and a Plugin for QGIS, with the purpose of creating a user-friendly Graphical User Interface and widen its usage among users, professionals and researchers.

ACKNOWLEDGEMENTS

Authors would like to thank National Geomatics Center of China (NGCC), German Aerospace Center (DLR), Joint Research Center (JRC) and Lombardy Region for kindly providing land cover datasets for free. This work was partially supported by URBAN-GEO BIG DATA, a Project of National Interest (PRIN) funded by the Italian Ministry of Education, University and Research (MIUR) – id. 20159CNLW8.

REFERENCES

Aickin, M., 1990. Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model and Its Relation to Cohen's Kappa. *Biometrics*, 46(2), pp. 293-302.

Anderson, J. R., Hardy, E. E., Roach, J. T. and Witmer, R. E., 1976. A land Use and Land Cover Classification System for Use with Remote Sensor Data. Geological Survey Professional Paper 964.

Brennan, R. L. and Prediger, D. J., 1981. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41, pp. 687-699.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS J. Photogram. Remote Sens*., 103, pp. 7-27.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), pp. 213-220.

Congalton, R.G. and Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data Principles and Practices, Lewis Publishers, Boca Raton, FL, USA.

Credali, M., Fasolini, D., Minnella, L., Pedrazzini, L., Peggion, M., Pezzoli, S., 2011. Tools for territorial knowledge and government. In *Land Cover Changes in Lombardy over the Last 50 Years*, Fasolini, D., Pezzoli, S., Sale, V.M., Cesca, M., Coffani, S., Brenna, S., Eds., ERSAF-Lombardy Region, Milan, Italy.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E., 2017. Breaking new ground in mapping human settlements from space – The Global Urban Footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134, pp. 30-42.

ESRI, 2018. ArcGIS Desktop. http://desktop.arcgis.com/en/arcmap (26 March 2018).

Foody, G. M., 2008. Harshness in image classification accuracy assessment. *International Journal of Remote Sensing,* 29(11), pp. 3137-3158.

Fung, T., LeDrew, E., 1988. The Determination of Optimal Threshold Levels for Change Detection Using Various Accuracy Indices. *Photogrammetric Engineering and Remote Sensing,* 54(10), pp. 1449-1454.

Geary, R.C., 1954. The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3), 115-145.

Getis, A. and Ord, K. J., 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographycal Analiysis,* 24, pp. 189-206.

GRASS Development Team, 2018. Geographic Resources Analysis Support System (GRASS) Software, Version 7.4. Open Source Geospatial Foundation. http://grass.osgeo.org (26 March 2018).

Koukoulas, S., Blackburn, G. A., 2001. Introducing New Indices for Accuracy Evaluation of Classified Images Representing Semi-Natural Woodland Environments. *Photogrammetric Engineering & Remote Sensing*, 67(4), pp.499-510.

Labatut, V. and Cherifi, H., 2012. Accuracy Measures for the Comparison of Classifiers. Proceedings of the 5th International Conference on Information Technology, Chania Crete, 7-9 July 2014, pp. 1-5.

Landis, R. J. and Koch, G. G., 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33, pp. 159-174.

Liu, C., Frazier, P. and Kumar, L., 2007. Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment,* 107, pp. 606-616.

Lunetta, R.S., Lyon, J.G., 2004. Remote Sensing and GIS Accuracy Assessment, CRC Press, Boca Raton, FL, p. 326.

Ma, Z. and Redmond, R. L., 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering & Remote Sensing,* 61(4), pp. 435-439.

Meyer, M., Brass, J., Gerbig, B. and Batson, F., 1975. ETRS Data Applications to Surface Resource Surveys of Potential Coal Production Lands in Southeast Montana, IARSL Final Research Report, University of Minnesota, USA.

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37, pp. 17-3.

Moran, P., 1948. The interpretation of statistical maps. *Journal of the Royal Society of London Series B*, 10, pp.243-251.

Nelson, R. F., 1983. Detecting Forest Canopy Change Due To Insect Activity Using Landsat MSS. *Photogrammetric Engineering and Remote Sensing*, 49(9), pp. 1303-1314.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Carneiro, F.S.M., Halkia, S., Julea, A.M., Kemper, T., Soille, P., Syrris, V., 2016. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. Publications Office of the European Union, EUR 27741 EN.

Pontius, R. G. and Millones, M., 2011. Death to Kappa: birth of quality disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing,* 32(15), pp. 4407-4429.

Pringle, M. J., Schmidt, M. and Muir, J. S., 2009. Geostatical interpolation of SLC-off Landsat ETM+ images. *ISPRS Journal of Photogrammetry and Remote Sensing,* 64, pp. 654-664.

QGIS Development Team, 2018. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.org (26 March 2018).

Stehman, S. V., 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), pp. 77-89.

Story, M., Congalton, R.G., 1986. Accuracy Assessment: A User’s Perspective. *Photogrammetric Engineering and Remote Sensing*, 52, pp. 397-399.

Türk, G., 1979. Gt index: A measure of the success of prediction. *Remote Sensing of Environment*, 8(1), pp.65–75.