Thematic quality assessment of land surface geospatial data based on confusion matrices: A matrix set for research on measures and procedures

Francisco J. Ariza-López¹ | José L. García-Balboa¹ | María V. Alba-Fernández² | José Rodríguez-Avi²

¹Departamento de Ingeniería Cartográfica Geodésica y Fotogrametría, Universidad de Jaén, Jaén, Spain
²Departamento de Estadística e Investigación Operativa, Universidad de Jaén, Jaén, Spain

Correspondence
José L. García-Balboa, Departamento de Ingeniería Cartográfica Geodésica y Fotogrametría, Universidad de Jaén, Campus Las Lagunillas, 23071 Jaén, Spain. Email: jlblalboa@ujaen.es

Abstract
The confusion matrix has long been adopted as the ‘de facto’ and ‘de jure’ standard method of reporting on the thematic accuracy assessment of any land surface geospatial dataset. This type of data supports decision-making in many different fields, so suitable quality is therefore essential in order to take the best decisions. Nevertheless, the creation and exploitation of the confusion matrix remains as an open topic with issues related to sampling design, quantitative indices derived from the matrix, statistical hypotheses that could be applied, etc. In connection with the latter, a confusion matrix dataset would be useful for a researcher in this matter. We have developed such a dataset retrieving confusion matrices from the literature, mainly research articles published in scientific journals included in WoS. We have collected almost 200 matrices in a database. This allows us to access the complete matrices and query different interesting properties of them and of the project where they were developed such as matrix size, sample size, location, year of data capture, labels of the classes, quality indices used, and extension and location of the project (where available).

Keywords
confusion matrix, geospatial data, land surface, quality assessment, thematic accuracy

Dataset Details
Identifier: https://doi.org/10.6084/m9.figshare.c.4768997
Creator: Francisco J. Ariza-López, María V. Alba-Fernández, José L. García-Balboa, José Rodríguez-Avi
Title: confusion matrix dataset from thematic quality assessments of land surface geospatial data
Publisher: Universidad de Jaén
Publication year: 2019
(Resource type): Dataset
(Version): 1.0

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. Geoscience Data Journal published by Royal Meteorological Society and John Wiley & Sons Ltd.
1 | INTRODUCTION

The so-called confusion matrix (also referred to as misclassification matrix or error matrix) is the usual structure by which the classification correctness of land surface geospatial data is represented (see Congalton and Green, 2009; Stehman and Foody, 2019, between many others). Classification correctness is one of the so-called data quality elements established by ISO 19157 (International Organization for Standardization, 2013) for thematic accuracy when describing geospatial data quality. Land surface geospatial data (spatial databases, topographic maps, thematic maps, classified images, remote sensing products, etc.) support decision-making in several fields such as climate change, crop forecasting, forest times, national defense and spatial planning. Therefore, suitable quality is essential in order to ensure that decisions based on it are technically the best.

The confusion matrix is also included in the list of standardized data quality measures of ISO 19157 (International Organization for Standardization, 2013). Different metrics can be derived from the matrix (see Congalton, 1991), but it is recommended to always report the raw confusion matrix so that the user of the data can derive any metric suitable for their needs (Salk et al., 2018). The structure of a confusion matrix is summarized below.

Suppose that \( k \) categories \( C_1, C_2, \ldots, C_k \) are given (i.e. land-cover categories, etc.) and \( n \) sample units were observed. All sample units were classified into categories through a certain classification method, and such classification is displayed in a contingency table called confusion matrix. The \((i, j)\) element, \( n_{ij} \), represents the number of sample units that actually belong to \( C_j \) and are classified into \( C_i \) for \( i, j = 1, \ldots, k \).

In this way, the columns and rows of the contingency table correspond, respectively, to reference (index \( j \)) and classified (index \( i \)) data (Table 1). In consequence, the elements in the diagonal are correctly classified items and the off-diagonal elements contain the number of confusions, namely the errors due to omissions and commissions.

The use of a confusion matrix is long-established in research studies, which include land classification processes. Nevertheless, there are different aspects related to the creation and exploitation of the matrix that remain as interesting topics, such as new tools (i.e. Bratic et al., 2018), sampling design (i.e. Stehman, 2000, 2014), indices derived from the matrix (i.e. Chen et al., 2010; Pontius and Millones, 2011; Salk et al. 2018; Stehman, 2013) and proposals for testing statistical hypotheses (i.e. Foody, 2004; García-Balboa et al., 2018). It would be useful for a researcher in this field to have available a wide set of land classifications in different locations for different applications, introducing different sampling designs or classification strategies, etc., which could be difficult or costly to develop by other means. Here, we propose, as an alternative, to retrieve from the literature a sufficient number of matrices so that a researcher could use those most suitable for a specific research goal.

Our proposal is to make available a database, which stores all this information. It covers not only the numeric data of each matrix but also complementary information related to the context of the research and to metadata of the spatial data involved. As an example, with this database a researcher can explore key issues such as:

- the size of the matrices (number of classes),
- the indices computed from the matrices,
- the sample size,
- type of geospatial data in both assessed datasets and control datasets,
- classifications used, and
- it also allows us to collect information for a meta-analysis in an easy way.

### TABLE 1 Structure of a confusion matrix with \( k \) categories

| Classified data | \( C_1 \) | \( C_2 \) | \( \ldots \) | \( C_k \) |
|-----------------|--------|--------|-----|--------|
| \( C_1 \)       | \( n_{11} \) | \( n_{12} \) | \( \ldots \) | \( n_{1k} \) |
| \( C_2 \)       | \( n_{21} \) | \( n_{22} \) | \( \ldots \) | \( n_{2k} \) |
| \( \ldots \)    | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( C_k \)       | \( n_{k1} \) | \( n_{k2} \) | \( \ldots \) | \( n_{kk} \) |

2 | DATA PRODUCTION METHODS

As a primary source of data, we have considered all confusion matrices included in publications that meet the following requirements:
It is a research article in a scientific journal.
- It is included in the WoS (Web of Science database).
- It belongs to the suitable categories of the WoS.

Therefore, the following search was performed in WoS:
- Topic: “confusion matrix” or “confusion matrices” or “error matrix” or “error matrices”. This search examines the following fields: title, abstract and keywords. The search terms are between quotes for an exact phrase.
- First filter: document type is article.
- Second filter: WoS categories are ‘remote sensing’ or ‘geosciences multidisciplinary’.
- Third filter: publication year is prior to year 2019.

More than four hundred articles (435) are obtained from the search, with an h-index of 57 and average citation of 27.81. The following figures summarize the results of the search:
- Figure 1 presents the number of articles published in each WoS (first 10 categories). As can be seen, more than the two selected categories are shown since journals are usually included in more than one category.
- Figure 2 includes the number of articles published in each journal (first 10 journals).
- Figure 3 shows the number of articles published each year, with a clear rising trend.

From the previous search, a list of research articles was obtained. Each one could contain, or not, one or more confusion matrices that could be added to a confusion matrix set. Therefore, the next goal is to revise each article to find out whether it contains a complete confusion matrix. The steps are as follows:
- Export the article list from WoS.
- Access the full text of each article. This step would require a subscription to every journal included in the search results.
- Read each article and localize every confusion matrix. Confusion matrices are usually included as tables throughout the text.
- Annotate the data included in each confusion matrix in order to include them in our confusion matrix database. These data include not only the numeric information but also the name of each of the classes. Also, when entering the matrix data into the database the criterion of putting the reference data in columns and classified data in rows was adopted.
- If available in the document, annotate any other complementary information that could be useful to understand the context in which the confusion matrix is applied. This information is related to location (name or coordinates), area size, field of application and indices computed. Also, metadata of both sets of data, controlled dataset (CDS) and reference dataset (RDS) are annotated: type of geospatial data, date, scale, resolution, accuracy, sampling method, etc.

3 | DATASET LOCATION AND FORMAT

All data are included in a database created following a relational model. This contains not only the numerical data of every confusion matrix found in the literature but also

![Figure 1](image-url) Figure 1 Number of articles in each WoS category from the search performed in WoS (only the first 10 categories are shown)
any other data collected. The entity–relationship diagram (ERD) is presented in Figure 4. There is a main table with the name ‘ConfusionMatrixSet’, which stores the numerical data of the matrix and some other complementary data (see Table 2).

The primary key (‘Id’) of the table ‘ConfusionMatrixSet’ is used as a foreign key in different supplementary tables (see Figure 4), which complement the information about each confusion matrix. A different table is included for different aspects that could be of interest to a researcher.

Additional tables are as follows:

- Table ‘CDS_Metadata’. This includes metadata related to the assessed data to which the confusion matrix refers. Metadata are derived/extracted from the original document. The table contains several fields: type of geospatial data, date (year), scale, resolution (m) and accuracy (m). One tuple is included for each confusion matrix (if a document includes several confusion matrices a tuple is included for each one). If the value of a numerical field
unknown, a value of ‘9999’ is set. The type of geospatial
data is a text field with the most concise description pos-
sible, such as satellite image, SPOT imagery, crop map,
CORINE Land Cover and lidar data. If it is unknown, a
value of ‘NAV’ (not available value) is set.

- Table ‘RDS_Metadata’. These metadata describe the con-
trol/reference data used to assess the data and are derived
from the original document. The same fields as the meta-
data of the assessed data are included (data source type,
date, scale/resolution, accuracy). An additional field is
added for the sampling method (e.g. random sampling) and
sample unit (e.g. pixel).

Complementary tables are as follows:

- Table ‘Field’. We have assigned an application domain to
each confusion matrix. It represents the main topic related
to the research, that is land cover, land uses and agricul-
ture. If it is unknown, a ‘NAV’ code is set.

- Table ‘Classes’. The name of each of the classes of each
confusion matrix is included, as named in the original
document. The table contains one field in order to include
the name of the class. One tuple is included for each class
of each confusion matrix.

- Table ‘Indices’. The name of each of the indices computed
from each confusion matrix is included, as presented in
the original document. They have not been standardized
so as not to introduce interpretations. The table contains
one field to include the name of the index. One tuple is in-
cluded for each index of each confusion matrix. The value
of the indices is not included because they can be com-
puted directly from the confusion matrix.

- Table ‘Citations’. A complete reference to the original doc-
ument (journal article, book, chapter, etc.) in APA style is
included for each confusion matrix. The table contains one
field to include a reference. One tuple is included for each
confusion matrix.

The whole dataset is offered to the scientific community
in an open format, therefore to be managed with free open-
source software. Additionally, a complementary version is
offered using widely used proprietary formats. A collection

**FIGURE 4** Entity–relationship diagram of the dataset
TABLE 2 Fields included in the table ‘ConfusionMatrixSet’

| Name                | Data type (size)            | Description                                                                 |
|---------------------|----------------------------|-----------------------------------------------------------------------------|
| Id                  | AutoNumber (long integer)   | Primary key field, which starts from 101.                                    |
| CM_Data             | Long Text (1 Gbyte)         | Matrix data. Cell values are separated by commas. Rows are separated by brackets. For example, the confusion matrix published in Park, Kyriakidis and Hong (2016, table 3) is stored in the database as follows: \[[445,84,30,5,38],[22,51,17,1,9],[11,11,32,0,9],[2,2,1,47,0],[9,6,11,0,63]\] |
| CM_ExcelSheet       | Short Text (255)            | Excel sheet. An independent Excel file contains the data of each confusion matrix, each one in an independent sheet. This field contains the name of the Excel sheet. |
| CM_MatrixSize       | Number (long integer)       | Matrix size. Since the matrix is squared, it equals the number of rows or columns. It is also the number of categories \(k\). |
| CM_SampleSize       | Number (long integer)       | Size of the sample used to build the matrix. Equal to the sum of the cell values in the matrix. |
| Area                | Number (double)             | Surface area (in km\(^2\)) of the land described by the confusion matrix.   |
| Coord_yLL           | Number (double)             | Minimum bounding rectangle. Minimum latitude (decimal degrees).              |
| Coord_xLL           | Number (double)             | Minimum bounding rectangle. Minimum longitude (decimal degrees).             |
| Coord_yUR           | Number (double)             | Minimum bounding rectangle. Maximum latitude (decimal degrees).              |
| Coord_xUR           | Number (double)             | Minimum bounding rectangle. Maximum longitude (decimal degrees).             |
| Place               | Short Text (255)            | Value of the field ‘name’ of one or more records in the GeoNames database, separated by commas. Two or more records can refer to adjacent areas or hierarchical administrative classifications. |
| Observations        | Long Text (1 Gbyte)         | Any useful comment, for example about the hypothesis of the research and errors detected in the data. |

is available at Figshare (https://doi.org/10.6084/m9.figshare.c.4768997). This collection contains the following files:

- File CM_database.sqlite. It includes the database in SQLite format.
- File CM_database.accdb. It includes the database in Microsoft Access format. Although it is not an open format, it is offered because it is widely used.
- File CM_data.xlsx. A Microsoft Excel workbook, which contains a sheet for every confusion matrix included in the database. This file is only included for an alternative access to the numeric data of a confusion matrix. It is the same data included in the field ‘CM_Data’ in Table 2.
- File CM_data.ods. Contains the same data as the Microsoft Excel file but in the ODS open format.

4 | DATASET USE

Classification correctness assessment throughout confusion matrices is an open and interesting topic. The dataset of published confusion matrices about land surface geospatial data constitutes an important resource in the advance of research related to geospatial thematic data quality. A researcher can select from the dataset those cases that are suitable for the
specific research topic. Different filtering criteria can be considered. For example, it could be interesting to locate cases attending to:

- matrix size,
- sample size,
- location,

**FIGURE 5** Frequencies of the matrix size (number of categories) in the confusion matrix dataset

**FIGURE 6** Frequencies of the indices computed for matrices in the dataset
• year of data capture,
• quality indices used and
• extension and location of the project (where available).

As an example, a researcher can be interested on the use of the kappa coefficient (κ) when assessing Sentinel 2 images. Performing the corresponding query in the tables ‘CDS_metadata’ and ‘Indices’, it can be found that κ is applied in the matrices 210-213, published in Wessel et al. (2018).

Also, the whole dataset can be taken into account to perform analyses that could clarify how the scientific community has used the confusion matrix. As an example, in Liu et al. (2007) the number of categories was explored through a frequency histogram. They found that the majority of matrices have between 3 and 7 categories. It is easy to obtain a similar histogram with our dataset (Figure 5). We found that the most frequent matrix size is 5 categories and that the majority range between 2 and 7 categories.

A similar analysis could be performed with any other aspect of interest, like the quality indices derived from the matrices. Widely adopted indices for thematic accuracy controls upon confusion matrices are the overall accuracy (OA) and κ (see Congalton, 1991 or Congalton and Green, 2009), but more indices can be found in the literature (global or category-related). In Liu et al. (2007), 20 global indices and 14 category-related indices are compiled, and more recently, Morales-Barquero et al. (2019) review popular measures of accuracy in the context of natural resources and Stehman and Foody (2019) look at temporal trends in accuracy reporting. Our database can help a lot in studies of this type. So, in Figure 6 it can be seen that in the dataset the more frequent indices are OA and κ (global), producer's accuracy and user's accuracy. Any other proposals are minoritary.

5 | CONCLUSIONS

The use of the confusion matrix remains as an open topic in geospatial data. There are many aspects to take into account when applying this tool and when assessing thematic accuracy: sample size, sampling unit, sampling design, quantitative indices derived from the matrix, statistical hypotheses that could be applied, etc. We have therefore developed a dataset that could be useful for any researcher interested on the matter. We have created a database with almost 200 confusion matrices published in research articles included in WoS database. Any researcher can perform queries to retrieve useful data. The dataset is offered in both open and proprietary formats at Figshare. Further works can expand the database, not only in the quantity of matrices, but also in exploring other literature sources. Reports from recognized institutions and research articles not included in WoS but in other databases can also be of interest.

ACKNOWLEDGEMENTS

The research presented in this paper was funded by Grant CMT2015-68276-R from the Spanish Ministry of Economy and Competitiveness.

OPEN PRACTICES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.6084/m9.figsh are.c.4768997 Learn more about the Open Practices badges from the Center for OpenScience: https://osf.io/tvyxz/ wiki.

ORCID

Francisco J. Ariza-López https://orcid.org/0000-0002-5204-3630
José L. García-Balboa https://orcid.org/0000-0002-3109-5888
María V. Alba-Fernández https://orcid.org/0000-0002-4747-740X
José Rodríguez-Avi https://orcid.org/0000-0002-1673-9876

REFERENCES

Bratic, G., Brovelli, M.A. & Molinari, M.E. (2018) A free and open source tool to assess the accuracy of land cover maps: Implementation and application to lombardy region (Italy). International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, XLII-3, 87–92. https://doi.org/10.5194/isprs-archives-XLII-3-87-2018

Chen, J., Zhu, X., Imura, H. & Chen, X. (2010) Consistency of accuracy assessment indices for soft classification: Simulation analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 65, 156–164. https://doi.org/10.1016/j.isprsjprs.2009.10.003

Congalton, R.G. (1991) A review of Assessing the Accuracy of Classifications of Remotely Sensed Data. Remote Sensing of Environment, 37, 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Congalton, R.G. & Green, K. (2009) Assessing the accuracy of remote sensed data. Principles and practices, 2nd edn. CRC Press: Boca Raton, FL, USA. https://doi.org/10.1201/9781420055139

Foody, G. (2004) Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. Photogrammetric Engineering & Remote Sensing, 70, 627–633. https://doi.org/10.14358/PERS.70.5.627

García-Balboa, J.L., Alba-Fernández, M.V., Ariza-López, F.J. & Rodríguez-Avi, J. (2018) Analysis of Thematic Similarity Using Confusion Matrices. ISPRS International Journal of Geo-Information, 7, 233. https://doi.org/10.3390/ijgi7060233

International Organization for Standardization (2013) ISO 19157:2013 - Geographic information – Data quality. International Organization for Standardization: Geneva, Switzerland. https://www.iso.org/standard/32575.html
Liu, C., Frazier, P. & Kumar, L. (2007) Comparative assessment of the measurement of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606–616. https://doi.org/10.1016/j.rse.2006.10.010

Morales-Barquero, L., Lyons, M.B., Phinn, S.R. & Roelfsema, C.M. (2019) Trends in remote sensing accuracy assessment approaches in the context of natural resources. *Remote Sensing*, 11, 2305. https://doi.org/10.3390/rs11192305

Park, N.-W., Kyriakidis, P.C. & Hong, S.-Y. (2016) Spatial Estimation of Classification Accuracy Using Indicator Kriging with an Image-Derived Ambiguity Index. *Remote Sensing*, 8, 320. https://doi.org/10.3390/rs8040320

Pontius, R.G. Jr & Millones, M. (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15), 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Salk, C., Fritz, S., See, L., Dresel, C. & McCallum, I. (2018) An Exploration of Some Pitfalls of Thematic Map Assessment Using the New Map Tools Resource. *Remote Sensing*, 10, 376. https://doi.org/10.3390/rs10030376

Stehman, S.V. (2000) Practical implications of Design-Based Sampling Inference for thematic Map Accuracy Assessment. *Remote Sensing of the Environment*, 72, 35–45. https://doi.org/10.1016/S0034-4257(99)00090-5

Stehman, S.V. (2013) Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, 132, 202–211. https://doi.org/10.1016/j.rse.2013.01.016

Stehman, S.V. (2014) Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing*, 35, 4923–4939. https://doi.org/10.1080/01431161.2014.930207

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

Wessel, M., Brandmeier, M. & Tiede, D. (2018) Evaluation of Different Machine Learning Algorithms for Scalable Classification of Tree Types and Tree Species Based on Sentinel-2 Data. *Remote Sensing*, 10, 1419. https://doi.org/10.3390/rs10091419

**How to cite this article:** Ariza-López FJ, García-Balboa JL, Alba-Fernández MV, Rodríguez-Avi J. Thematic quality assessment of land surface geospatial data based on confusion matrices: A matrix set for research on measures and procedures. *Geosci Data J*. 2021;00:1–9. https://doi.org/10.1002/gdj3.116