©Mokhamad Edliadi/CIFOR

CGIAR

RESEARCH PROGRAM ON Forests, Trees and Agroforestry

GEOSPATIAL DATA CURATION STEPS

Technical guidelines

(August 2021)

These guidelines and procedures has been developed consistent with the CGIAR Principles on the Management of Intellectual Assets and by the CGIAR's commitment to Open Access. This technical guide follows the guidelines and procedures in FTA's Geospatial Data Quality: Guidelines and Procedure, and FTA Metadata Guidelines and Procedures. This technical guide describes practical implementation in terms of objectives for curating and disseminating spatial data.

foreststreesagroforestry.org

SCOPE

Digital curation can be defined as maintaining and adding value to a trusted body of digital information for current and future use (Beagrie 2006). It is beyond archiving and preservation. Digital data curation is concerned with data management "for as long as it continues to be of scholarly, scientific, research and/or administrative interest, with the aim of supporting reproducibility of results, reuse of and adding value to that data, managing it from its point of creation until it is determined not to be useful, and ensuring its long-term accessibility and preservation, authenticity and integrity" (DCC n.d. in Sandifer n.d.).

Curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. This technical guide contains practical steps around curating geospatial data for dissemination purposes. It also contains information on how to disseminate geospatial data so that the data can be well accessed and/or used easily from the geoportal.

ROLES AND RESPONSIBILITIES

Project leaders and researchers

- Ensure that the dataset to be published does not have a copyright infringement.
- Ensure that the documentation of the dataset is complete from the producer dataset.
- Ensure privacy protections are implemented.

GIS specialists/producers

- Are responsible for the validation and quality of data that has been produced.
- Ensure completeness of metadata.

Data curators

- Ensure data are acquired properly.
- Implement geospatial data curation procedures properly.
- Ensure that metadata are filled in and follow the standards used within CGIAR System Office.

Data users

- Are responsible for data access that has been carried out including in terms of data distribution and terms of use.

TECHNICAL PROCEDURES OF GEOSPATIAL DATA CURATION

a. There are several things that need to be considered when acquiring the geospatial data:
   - The data are intended for publication and will be in the public domain.
   - There are no potential copyright infringements and the copyright owner has agreed to CIFOR-ICRAF's policy on data dissemination.
   - The dataset complies with the General Data Protection Regulation (GDPR) and other related policies and/or guidance on privacy and confidentiality protection.
   - All data documentation has been completed, including methods and procedures for data generation.

b. Check the quality of geospatial data in accordance with FTA's Geospatial Data Quality: Guidelines and Procedure. As for checking the quality of data, the steps for curating data can be followed (see Annex 1). Data with the curation status "does not pass" can still be acquired if the data are considered important enough to be archived.

c. Check the completeness of the metadata of the geospatial dataset to be published. Filling-out and validating procedures for metadata can be found in FTA Metadata Guidelines and Procedures.

d. Check the data format and consistency in the geographical projections so that the data can easily be used in various technological settings.

e. Data that are "pass" in quality, have metadata, and are formatted consistently are ready for archiving, cataloguing and disseminating in the geoportal.

TECHNICAL PROCEDURES FOR DATA DISSEMINATION AND PUBLICATION

a. Publish geospatial datasets for use by the public in user-accessible geoportals. All users can view and download directly from open data repositories and/or the Forest, Trees, and Agroforestry (FTA) geoportal. In order to maintain the publication of geospatial datasets in the FTA geoportal, the uniform resource locator (URL) address must be included in the metadata of geospatial datasets.

b. Geospatial datasets that will be disseminated can be delayed or cancelled for several reasons including:
   - There are risks identified when publishing the data either directly or indirectly to CIFOR-ICRAF. Data that has this risk can also be released with "restricted use".
   - Requests may come from researchers or project leaders not to publish or to delay publication. Geospatial datasets are still being acquired according to existing procedures and if there is an agreement, they can be published immediately.

c. Data citation is very important for any data that have been published and utilized by users; it can be used to measure the use and impact of research data and give credit for the recognition of those data. There are several basic elements required for a citation, including data generator, data title, distributor, date, version and persistent identifiers such as digital object identifier (DOI) or uniform resource identifier (URI).

REFERENCES

Beagrie N. 2006. Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation 1, 3–16. doi: 10.2218/ijdc.v1i1.2

Sandifer B. n.d. Good practice in research data management module 5: Deposit and long-term preservation. Newcastle University. Accessed 5 October 2021. https://slideplayer.com/slide/3353100/

Annex 1. Geospatial data curation steps

There are 4 criteria for curating geospatial data.

1. SYSTEM CO-ORDINATES AND GEOGRAPHICAL PROJECTION

System co-ordinates and projection system information can be found through the properties of the spatial data. There are common co-ordinate and projection systems: Universal Transverse Mercator (UTM); Mercator Projection System; and World Geodetic System 1984 (WGS84).

Inspection procedures:

a. Check the existence of co-ordinate system information and/or system projections of geospatial data.

b. Check the type of the co-ordinate system and/or projection using spatial data.

Curation status is **pass** if:

- A co-ordinate system and/or projections exist on the spatial data according to the location of the data.
- Information related to data position is marked with clear boundaries of north, south, west, and east co-ordinate points.

Curation status is **does not** pass and requires improvement if:

- Data do not have a definition of a co-ordinate system or a projection and so the quality of these data needs to be improved by adding the appropriate co-ordinate system and/or projection information. Documentation is needed to record all improvement processes.

Note: The CGIAR FTA geospatial data quality standard (2021) stipulates that the mandatory co-ordinate system used in archival data and publications is WGS84, so all well-curated data must use this co-ordinate system.

2. ATTRIBUTE TABLE

The attribute table is an important item that contains information from each feature in the spatial data.

Inspection procedures:

- Check the clarity of the attribute column headings.
- Check the clarity of the contents in the attribute table.
- Check the completeness of the contents of the attribute table.
- Check the repetition of information in the attribute table.
- Check the consistency of the data attribute contents.

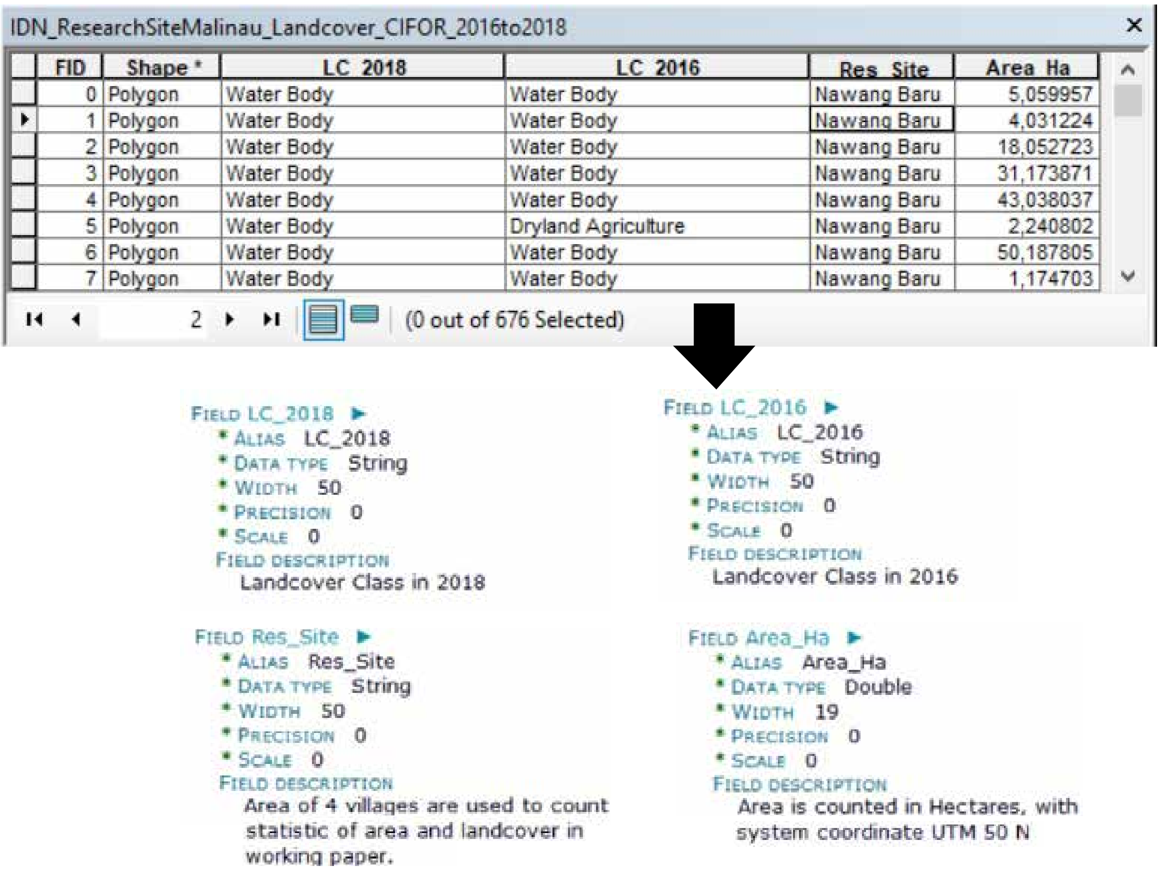
Some spatial attribute conditions:

a. Column headings

Curation status is **pass** if:

- Column headings are written clearly, easily understood and in accordance with the title of the data.
- Column headings are written using abbreviations and accompanied by explanatory notes that explain the abbreviated column headings and are in accordance with the type of data.

Notes can be written as metadata embedded in the spatial data or in the separate notes as in Figure A1.



IDN_ResearchSiteMalinau_Landcover_CIFOR_2016to2018

FID	Shape *	LC_2018	LC_2016	Res_Site	Area_Ha
0	Polygon	Water Body	Water Body	Nawang Baru	5,059957
1	Polygon	Water Body	Water Body	Nawang Baru	4,031224
2	Polygon	Water Body	Water Body	Nawang Baru	18,052723
3	Polygon	Water Body	Water Body	Nawang Baru	31,173871
4	Polygon	Water Body	Water Body	Nawang Baru	43,038037
5	Polygon	Water Body	Dryland Agriculture	Nawang Baru	2,240802
6	Polygon	Water Body	Water Body	Nawang Baru	50,187805
7	Polygon	Water Body	Water Body	Nawang Baru	1,174703

2 (0 out of 676 Selected)

FIELD LC_2018
- ALIAS LC_2018
- DATA TYPE String
- WIDTH 50
- PRECISION 0
- SCALE 0

FIELD DESCRIPTION
Landcover Class in 2018

FIELD LC_2016
- ALIAS LC_2016
- DATA TYPE String
- WIDTH 50
- PRECISION 0
- SCALE 0

FIELD DESCRIPTION
Landcover Class in 2016

FIELD Res_Site
- ALIAS Res_Site
- DATA TYPE String
- WIDTH 50
- PRECISION 0
- SCALE 0

FIELD DESCRIPTION
Area of 4 villages are used to count statistic of area and landcover in working paper.

FIELD Area_Ha
- ALIAS Area_Ha
- DATA TYPE Double
- WIDTH 19
- PRECISION 0
- SCALE 0

FIELD DESCRIPTION
Area is counted in Hectares, with system coordinate UTM 50 N

Figure A1. The curation status of this attribute table as regards column headings is pass.

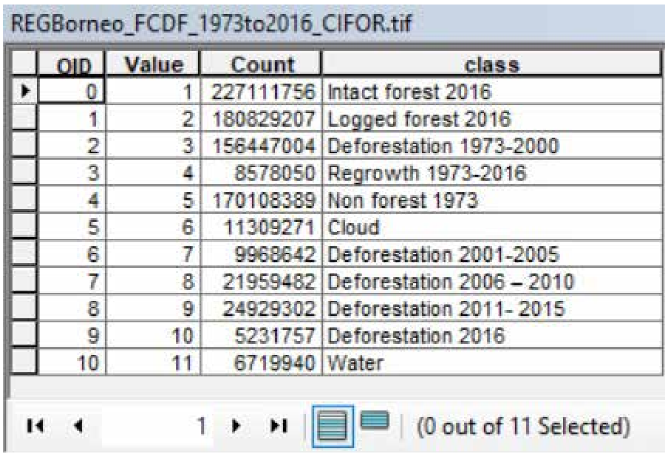
Curation status is **does not pass** and requires improvement if:

- Column headings are written in an unclear manner, without any explanatory notes. Then, the attribute table in terms of column headings is not appropriate.
- Column headings are written in such a way that they do not match/do not support the type of data.

b. Attribute contents

Curation status is **pass** if:

- The contents of all rows, which represent features, are easy to understand and match the column headings and type of data and are fully filled out, so the attribute table in terms of content is appropriate/clear, consistent and complete. Figure A2 shows a complete attribute example.
- The contents of the attribute table are only values and are accompanied by notes explaining the meaning of these values to ensure the information is complete and clear. Figure A3 depicts a sample of such data.



REGBorneo_FCDF_1973to2016_CIFOR.tif

OID	Value	Count	class
0	1	227111756	Intact forest 2016
1	2	180829207	Logged forest 2016
2	3	156447004	Deforestation 1973-2000
3	4	8578050	Regrowth 1973-2016
4	5	170108389	Non forest 1973
5	6	11309271	Cloud
6	7	9968642	Deforestation 2001-2005
7	8	21959482	Deforestation 2006 – 2010
8	9	24929302	Deforestation 2011- 2015
9	10	5231757	Deforestation 2016
10	11	6719940	Water

Figure A2. The curation status of this attribute table as regards attribute contents is pass.

Curation status is **does not pass** and requires improvement if:

- The contents of all rows are easy to understand but do not match the column headings and/or types of data, so the attribute table in terms of content is inconsistent.
- If the contents of the attribute table are found to contain:
  a. no content or a blank cell in one line or more than one line
  b. incomplete words with cut-off or missing characters
  c. missing information in the form of dots ("....")
  d. only a number (value)
  and are not accompanied by a note explaining the intent and condition, then in terms of the contents of the attribute table it is declared incomplete, and the information is not clear.

3. TOPOLOGY ERRORS

Topology errors can occur in the process of checking vector data. This process aims to see the condition of features in spatial data and is often done using topology tools in ArcGIS.

Inspection procedures:

- Check the overall condition of vector data.
- Check for overlap.
- Check for empty areas between vector data (holes/gaps).

Curation status is **pass** if:

- The results of the topology error check show that there are no overlaps or gaps, so the vector data are appropriate and complete.
- The results of the topology error check are overlapping, but there are documents that explain the reasons for this condition, so these vector data are appropriate and complete. Some data that are generally allowed to overlap are as follows:
  a. Overlapping on concessions
     - Overlapping can occur due to the unfinished licensing process.

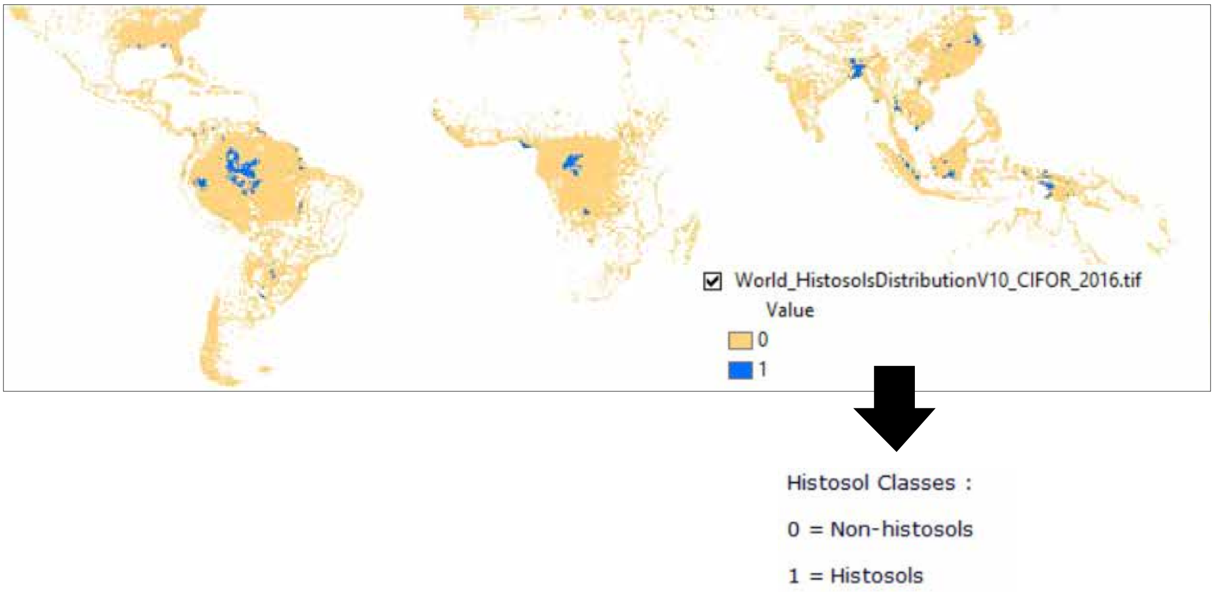


Figure A3. The content of the attributes is pass because it contains code values and explanations about the meaning of values.

- Field conditions do overlap due to different uses, for example, mining concessions and plantations in overlapping areas.

b. Overlapping territorial boundaries

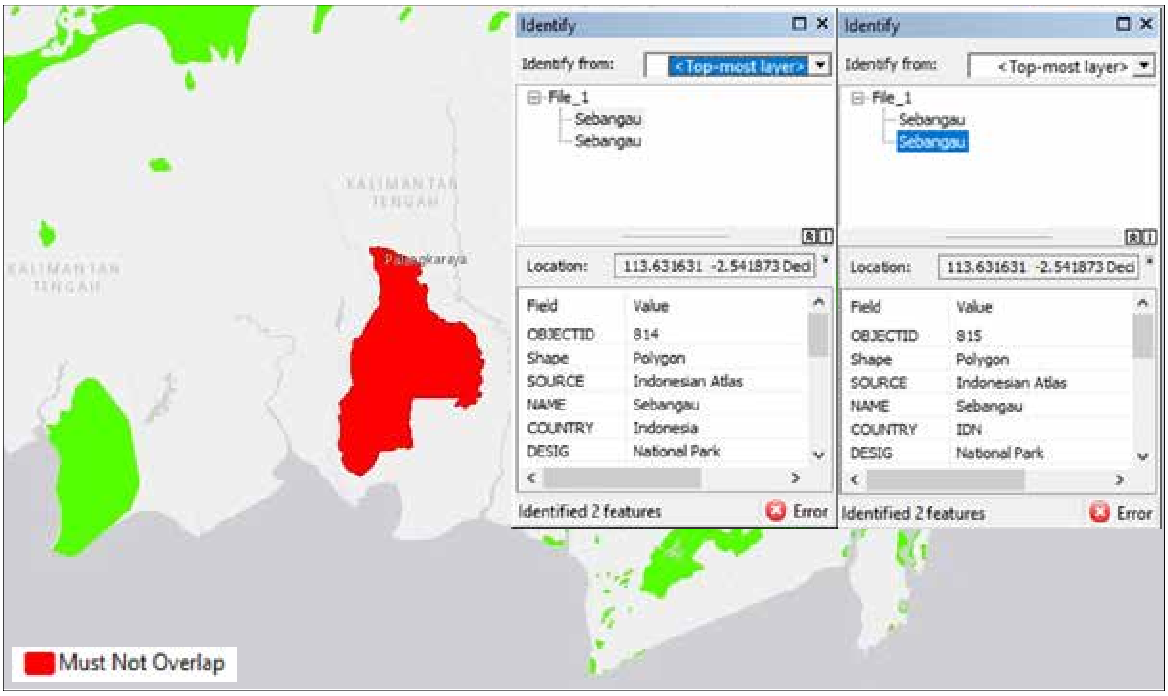
- This can happen because the territorial boundaries are still not definitive or there is no agreement yet.

c. Overlapping of other data accompanied by documentation containing history/reasons for the overlap

- The results of the check for topology errors have gaps and are accompanied by notes explaining the reasons for these conditions; therefore, these vector data are appropriate and complete.

Curation status is **does not pass** and requires improvement if:

- The results of topology errors are overlapping and not accompanied by a note explaining the reason for the condition, so these vector data are not appropriate. Figure A4 provides an example of data with an overlapping error topology, where two polygons have the same information.
- The results of the topology error contain gaps between other vector data and are not accompanied by notes explaining the reason for the condition. Then, these vector data are not appropriate. Figure A5 shows an example of vector data with overlaps (red) and gaps (black) within the area.



Identify

Identify from: <Top-most layer>

File_1
Sebangau
Sebangau

Location: 113.631631 -2.541873 Deci

Field	Value
OBJECTID	814
Shape	Polygon
SOURCE	Indonesian Atlas
NAME	Sebangau
COUNTRY	Indonesia
DESIG	National Park

Identified 2 features Error

Identify

Identify from: <Top-most layer>

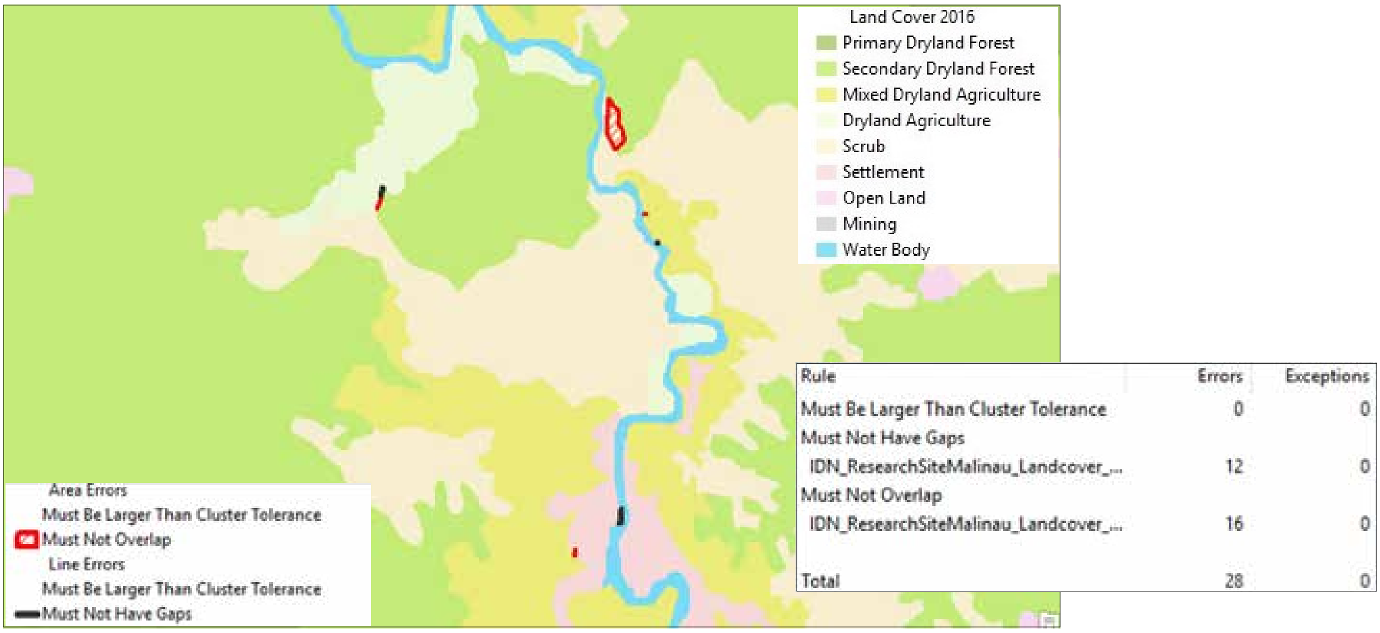
File_1
Sebangau
Sebangau

Location: 113.631631 -2.541873 Deci

Field	Value
OBJECTID	815
Shape	Polygon
SOURCE	Indonesian Atlas
NAME	Sebangau
COUNTRY	IDN
DESIG	National Park

Identified 2 features Error

Figure A4. Example of data with an overlapping error topology.



Rule	Errors	Exceptions
Must Be Larger Than Cluster Tolerance	0	0
Must Not Have Gaps		
IDN_ResearchSiteMalinau_Landcover_...	12	0
Must Not Overlap		
IDN_ResearchSiteMalinau_Landcover_...	16	0
Total	28	0

Figure A5. An example of vector data with overlaps (red) and gaps (black) within the area.

4. STATISTICAL INFORMATION

Statistical information refers to statistical values stored as geospatial data and can usually be accessed through attribute data.

Inspection procedures:

- Check the consistency of the total area/length/ circumference of each feature, including those of each data classification class.
- Check reports related to spatial data such as working papers or scientific publications.

Curation status is **pass** if:

- Statistical information of the spatial data and the report is the same, so that the spatial data in terms of statistics are appropriate and consistent. An example of spatial data consistent with data from a published document is shown in Figure A6.

Curation status is **does not pass** and requires improvement if:

- Statistical information of the spatial data and reports is different, and there is no document about this difference; so, the spatial data are not appropriate and consistent, and therefore need improvement. An example of spatial data with statistical information that contrasts with published information is shown in Figure A7.

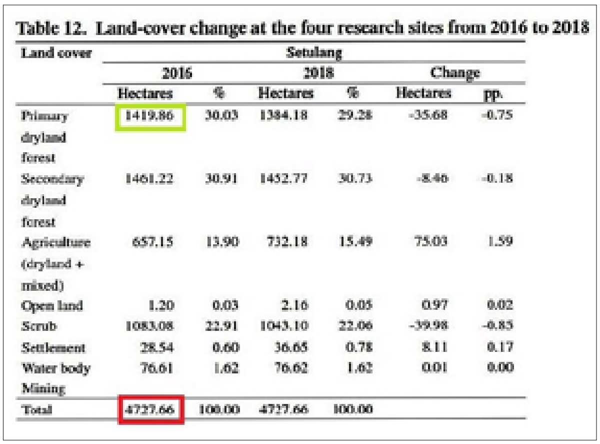
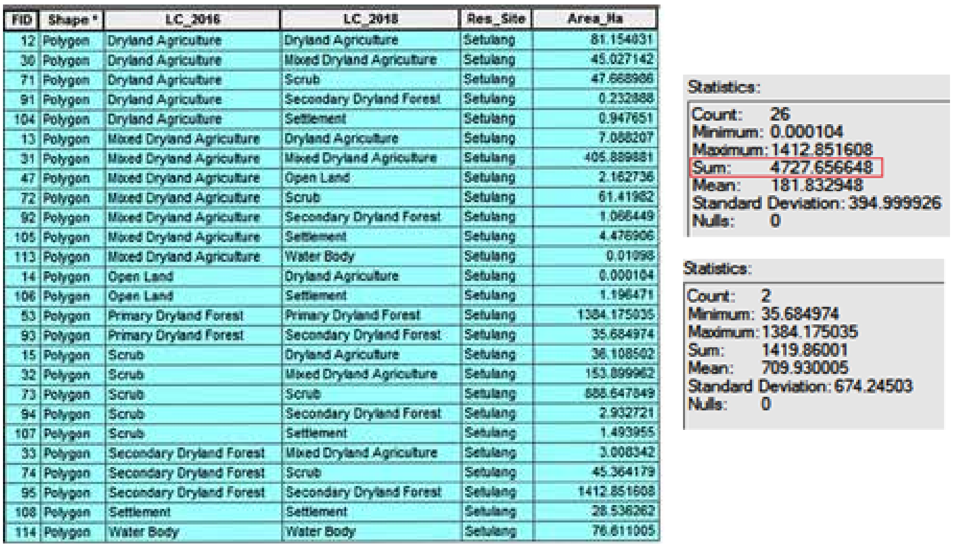


Table 12. Land-cover change at the four research sites from 2016 to 2018

Land cover	Setulang					
	2016		2018		Change	
	Hectares	%	Hectares	%	Hectares	pp.
Primary dryland forest	1419.86	30.03	1384.18	29.28	-35.68	-0.75
Secondary dryland forest	1461.22	30.91	1452.77	30.73	-8.46	-0.18
Agriculture (dryland + mixed)	657.15	13.90	732.18	15.49	75.03	1.59
Open land	1.20	0.03	2.16	0.05	0.97	0.02
Scrub	1083.08	22.91	1043.10	22.06	-39.98	-0.85
Settlement	28.54	0.60	36.65	0.78	8.11	0.17
Water body	76.61	1.62	76.62	1.62	0.01	0.00
Mining						
Total	4727.66	100.00	4727.66	100.00		



FID	Shape *	LC_2016	LC_2018	Res_Site	Area_Ha
12	Polygon	Dryland Agriculture	Dryland Agriculture	Setulang	81.154031
30	Polygon	Dryland Agriculture	Mixed Dryland Agriculture	Setulang	45.027142
71	Polygon	Dryland Agriculture	Scrub	Setulang	47.668986
91	Polygon	Dryland Agriculture	Secondary Dryland Forest	Setulang	0.232888
104	Polygon	Dryland Agriculture	Settlement	Setulang	0.947651
13	Polygon	Mixed Dryland Agriculture	Dryland Agriculture	Setulang	7.088207
31	Polygon	Mixed Dryland Agriculture	Mixed Dryland Agriculture	Setulang	405.889881
47	Polygon	Mixed Dryland Agriculture	Open Land	Setulang	2.162736
72	Polygon	Mixed Dryland Agriculture	Scrub	Setulang	61.41982
92	Polygon	Mixed Dryland Agriculture	Secondary Dryland Forest	Setulang	1.066449
105	Polygon	Mixed Dryland Agriculture	Settlement	Setulang	4.476906
113	Polygon	Mixed Dryland Agriculture	Water Body	Setulang	0.01098
14	Polygon	Open Land	Dryland Agriculture	Setulang	0.000104
106	Polygon	Open Land	Settlement	Setulang	1.196471
53	Polygon	Primary Dryland Forest	Primary Dryland Forest	Setulang	1384.175035
93	Polygon	Primary Dryland Forest	Secondary Dryland Forest	Setulang	35.684974
15	Polygon	Scrub	Dryland Agriculture	Setulang	36.108502
32	Polygon	Scrub	Mixed Dryland Agriculture	Setulang	153.899962
73	Polygon	Scrub	Scrub	Setulang	888.647849
94	Polygon	Scrub	Secondary Dryland Forest	Setulang	2.932721
107	Polygon	Scrub	Settlement	Setulang	1.493955
33	Polygon	Secondary Dryland Forest	Mixed Dryland Agriculture	Setulang	3.008342
74	Polygon	Secondary Dryland Forest	Scrub	Setulang	45.364179
95	Polygon	Secondary Dryland Forest	Secondary Dryland Forest	Setulang	1412.851608
108	Polygon	Settlement	Settlement	Setulang	28.536262
114	Polygon	Water Body	Water Body	Setulang	76.611005

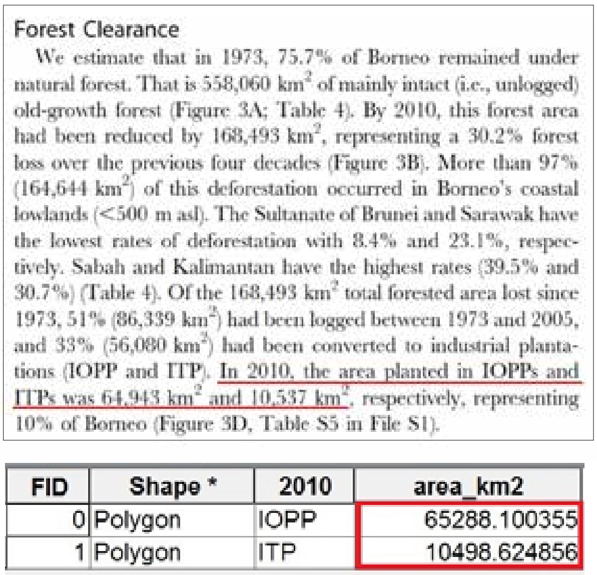
Statistics:

```
Count:     26
Minimum:   0.000104
Maximum:   1412.851608
Sum:       4727.656648
Mean:      181.832948
Standard Deviation: 394.999926
Nulls:     0
```

Statistics:

```
Count:     2
Minimum:   35.684974
Maximum:   1384.175035
Sum:       1419.86001
Mean:      709.930005
Standard Deviation: 674.24503
Nulls:     0
```

Figure A6. **An example of spatial data consistent with data from a published document.**



Forest Clearance

We estimate that in 1973, 75.7% of Borneo remained under natural forest. That is 558,060 km^2 of mainly intact (i.e., unlogged) old-growth forest (Figure 3A; Table 4). By 2010, this forest area had been reduced by 168,493 km^2, representing a 30.2% forest loss over the previous four decades (Figure 3B). More than 97% (164,644 km^2) of this deforestation occurred in Borneo's coastal lowlands (<500 m asl). The Sultanate of Brunei and Sarawak have the lowest rates of deforestation with 8.4% and 23.1%, respectively. Sabah and Kalimantan have the highest rates (39.5% and 30.7%) (Table 4). Of the 168,493 km^2 total forested area lost since 1973, 51% (86,339 km^2) had been logged between 1973 and 2005, and 33% (56,080 km^2) had been converted to industrial plantations (IOPP and ITP). In 2010, the area planted in IOPPs and ITPs was 64,943 km^2 and 10,537 km^2, respectively, representing 10% of Borneo (Figure 3D, Table S5 in File S1).

FID	Shape *	2010	area_km2
0	Polygon	IOPP	65288.100355
1	Polygon	ITP	10498.624856

Figure A7. **An example of spatial data with statistical information that contrasts with published information.**

The CGIAR Research Program on Forests, Trees and Agroforestry (FTA). FTA is the world's largest research for development program to enhance the role of forests, trees and agroforestry in sustainable development and food security and to address climate change. CIFOR leads FTA in partnership with ICRAF, the Alliance of Bioversity International and CIAT, CATIE, CIRAD, INBAR and TBI.

FTA's work is supported by the CGIAR Trust Fund: cgiar.org/funders/