©Mokhamad Edliadi/CIFOR

CGIAR

RESEARCH PROGRAM ON
Forests, Trees and Agroforestry

GEOSPATIAL DATA QUALITY

Guidelines and Procedures

(July 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, and explain how to implement geospatial data quality management practices. This document provides guidelines and procedures on managing the quality of geospatial data – from applying data quality evaluation procedures to developing geospatial data quality documentation.

SCOPE

In this document, the term 'geospatial data' is used to refer to data that provides information on the location of features, attributes, and often temporal information as well. Different types of geospatial data can be produced for different purposes and through different processes.

Types of geospatial data considered in these guidelines include:

a. Vector data: data consisting of points, lines and polygons;
b. Raster data: data consisting of pixels, with each pixel having an associated value.

Geospatial datasets are collected from many sources and channels with different methodologies and techniques. Throughout this document, geospatial data quality management can refer to the validation of errors relating to potential discrepancies arising from:

Different types of data: the many spatial data formats need to be evaluated to ensure appropriate and consistent data values;

- Unintended changes and modifications to geospatial datasets, such as deletion of features;
- Errors during data transfer.

ROLES AND RESPONSIBILITIES

Project leaders

- Set the culture of practice based on guidance, and delegate specific responsibilities as appropriate;
- Are accountable for ensuring that, should they leave their organization during the life of the project, the data stays with the project.

Researchers

- Maintain records of geospatial data and ensure these records and the geospatial data and materials are securely stored;
- Make the geospatial data available to other researchers via open or negotiated access as appropriate and in accordance with the requirements of research funding bodies and CIFOR's Research Data Management Policy;
- Ensure that, where projects span several institutions, an agreement is developed at the outset covering the ownership and storage of geospatial data and primary materials within each institution in accordance with CGIAR policies and guidelines;
- Ensure that adequate back-up, archival and monitoring strategies are in place to prevent the loss of geospatial data and primary materials, and associated delays in completing research.

GIS specialists

- As subject specialists, must understand the technical framework underlying data quality evaluation, improvement and management activities in their respective area;
- Ensure overall geospatial data validity, integrity and conformity of information gathered by their portfolio;
- Oversee and implement geospatial data quality procedures.

Data users

- Responsible and accountable for all data access made through their user accounts and the subsequent use and distribution of the data.

GEOSPATIAL DATA QUALITY ASSESSMENT

To ensure quality is achieved to the highest standard, geospatial data should be assessed and evaluated based on the selected standard. These guidelines adopt the standard from the International Organization for Standardization (ISO) for assessing geospatial data quality.

There are two levels of assessment:

- The first level of assessment is performed by the GIS specialist as the producer of geospatial data. This level of assessment is based on data quality checks based on given data specifications (Annex 2).
- The second level of data quality assessment is performed by geospatial data consumers (researchers and data users) where feedback is taken from consumers for improving data quality.

Suggested steps for geospatial data quality assessment at the first level (producer) are as follows:

a. Select and review geospatial data quality elements and sub-elements (Annex 1) – Geospatial data producers should select element(s) to use to determine how well the dataset meets requirements;
b. Determine geospatial data quality scope – data producers should determine the scope of data evaluation;
c. Choose the data quality assessment method(s) for data evaluation (Annex 3);
d. Specify conformance quality level;
e. Assess conformance to product specification (Annex 2);
f. Report quality evaluation results.

Suggested steps for geospatial data quality assessment at the second level (user) are as follows:

a. Select data quality elements and sub-elements – geospatial data users should select element(s) to use to determine how well the dataset meets requirements;
b. Define data quality scope – users should set the scope of quality evaluation;
c. Select data quality measures – users should choose data quality measures to be evaluated;
d. Specify data quality conformance level – using user requirements, data users should select the data quality conformance level – data may be accepted even though one or more data quality element fails to meet requirements;
e. Provide feedback to data producers on geospatial data quality improvement needs so data can be analysed / rectified based on feedback.

Geospatial data quality assessment results should provide:

- The conformance or non-conformance status of the geospatial data to data quality elements and sub-elements;
- Interpretations of results to draw conclusions about data quality;
- A dataset of acceptable quality even when one or more data quality elements fail to meet requirements. Documentation is mandatory to record such acceptance.

GEOSPATIAL DATA QUALITY DOCUMENTATION

It is mandatory for geospatial data producers to document data quality in the dataset's file and metadata. An additional report may be created to provide a detailed statement on data quality.

Documentation should have appropriate content, including:

- Reference to the geospatial data quality assessment procedures;
- Methods applied for geospatial data assessment;
- Comparisons and interpretations of geospatial data quality assessment results;
- Any interventions or improvements to the geospatial dataset (if applicable) – another important step during data quality documentation is to ensure quality report status is kept up to date to include any new dataset creations, updates, modifications and/or deletions.

©Mokhamad Edliadi/CIFOR

Annex 1. Geospatial data quality elements and sub-elements

Element	Sub-element	Description
Data and model completeness All features should be measured, with the number of missing features kept to a minimum. The level of completeness can be gauged by the extent to which the completeness of objects (entities) in the real world are represented through the model in the spatial dataset.	Object entity completeness	Explains the extent to which a real-world object (e.g., a river or road) can be represented by a geospatial dataset
	Attribute data completeness	Explains the extent to which information on a model from the real world (in the form of a data attribute) is complete (An example of an incomplete attribute for river data, for instance, is a missing river name)
	Formal completeness	Explains the formal structure the spatial data should have, data format and metainformation for instance
Logical consistency Logical consistency in a dataset can be explained as the absence of conflict in the dataset in relation to its spatial data's correctness in representing the real world. Inconsistencies occur when two or more pieces of data or information disagree or are incompatible with each other.	Domain consistency	Explains data value consistency (actual data) or the rule(s) governing the data value (statement of data value)
	Format consistency	Explains the consistency in the format when data is stored in a database
	Topological consistency	Explains the degree of adherence of a feature's geometry characteristics to specified rules (for instance, a road should use a connected line) during the mapping process. This includes transformations, which often omit the geometry characteristics of an object
Positional accuracy Positional accuracy can be defined as how well the data represent positional accuracy (the level of accuracy of any adjustments must also be known)	Absolute (external accuracy)	Explains how close the coordinates (on a map) are to the location of the object in the real world, and in the event of any error, that error should be within the acceptable value
	Relative (internal accuracy)	Explains how close the relative position of a feature in a dataset is to its position in the real world
	Gridded data (pixel) position accuracy	Explains how close the centre of a pixel is when ground truthing
Temporal accuracy The temporal information that accurately describes a geographic phenomenon.		There are three essential time elements for geospatial information: event time, i.e., when a change occurred; observation time, i.e., when the geographic phenomenon was observed; and transaction time, i.e., when the data/information was stored in the database. Quality in temporal accuracy can be defined as: a. Accuracy of temporal references b. Consistency in relation to the order of events c. Validation in relation to time
Thematic accuracy Thematic accuracy is correctness in classifying features and the accuracy of data attributes.	Classification correctness	Accuracy based on a comparison of the class assigned to a feature (real-world representation) or its data attributes to the real world
	Attribute accuracy	The accuracy of facts about features, including agreed history, information on measurement, and results of interpretation/analysis. Attributes are usually represented in tables where rows correspond to features and columns present accurate information on an attribute
	Qualitative data accuracy	Accuracy is gauged based on the correctness of qualitative data, such as name or class
	Quantitative data accuracy	Accuracy is gauged based on the correctness of quantitative data, such as level of elevation

Annex 2. Geospatial dataset product specification

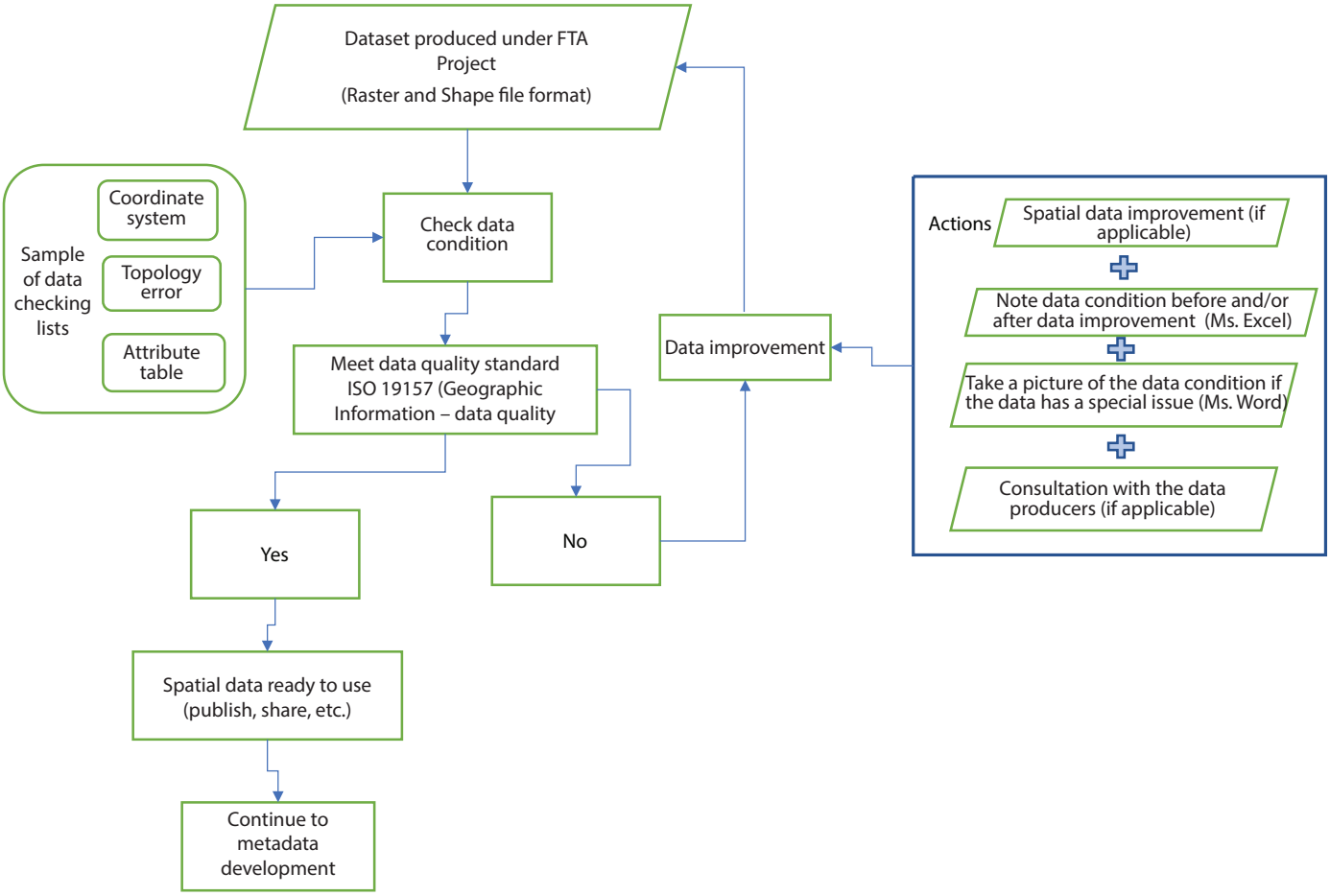
Data quality sub elements	Product output specification	Scope
Logical consistency (Format consistency)	Datum Horizontal: WGS 1984 Projection: Geographic Database structure: Shapefile format for vector data and tiff format for raster Assurance: Verification and review from researcher/project leader	The dataset
Positional accuracy (Absolute)	Requirement: The positional accuracy should be accurate and similar to the base map (i.e., Google base map, ESRI base map). Assurance: Check plotting and verification. If necessary, a check plot should be taken more than once. Researcher/project leader will conduct a cross check for authorization	The dataset
Completeness (Commission)	Requirement: All data topology should be complete Assurance: Exhaustive comparisons of base maps will be conducted. Researcher/project leader will arrange cross checks	The dataset
Logical consistency (Domain consistency)	Requirement Unit: Units of measurement should use the International System of Units (SI) system Assurance: All attribute names and definitions should be verified	The dataset
Logical consistency (Topological)	Requirement: Point: All points should have an ID value Lines: All lines should have an ID value Polygon: Polygons should have only one level point. Assurance: Verification from researcher/project leader	The dataset
Thematic accuracy (Classification accuracy)	Line/Polygon: Boundary class will be given a code list. For example: If a line represents an international border, provincial or district boundary, it will be assigned a code list based on international standard. Assurance: Verification from researcher/project leader	The dataset

Annex 3. Assessment methods

There are two quality evaluation methods: direct and indirect. The direct method involves either sampling or a full inspection of the data in the dataset. The indirect method involves evaluating quality of data and/or information from sources other than the data in the dataset.

Assessment Method	Procedures
Direct evaluation	There are two types of direct evaluation: the full inspection and sampling methods. Features in the dataset should be tested according to the standard defined by data quality elements, data quality sub-elements, and data quality measures. A full inspection involves checking all items in a population to determine their quality. The sampling method involves checking some items in a population to determine overall quality. (See Annex 4) The evaluation method can be an automated or non-automated process. Automated evaluations can use ArcGIS topology checks, while non-automated assessments can be conducted manually. Elements that can be evaluated include: Logical consistency: topological consistency (polygon closure); Completeness: (checking street names against other maps); Temporal: temporal consistency.
Indirect evaluation	Indirect evaluation procedures include: Defining the elements and sub-elements of geospatial data quality; Estimating the quality of data from other sources, including, but not limited to metadata, knowledge of the dataset's purpose, dataset lineage documentation, and report quality.

Annex 4. Data quality assessment workflow (sampling method)



RESEARCH PROGRAM ON Forests, Trees and Agroforestry
CGIAR

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/