Quality assurance and assessment framework for land cover maps validation in the Copernicus Hot Spot Monitoring activity

Zoltan Szántoi¹b, Gabriel Jaffrain¹, Heinz Gallauer¹, Conrad Bielski¹, Karl Ruf¹, Andrea Lupi², Petra Miletich³, Anne-Cécile Giroux³, Irina Carlan³, Willibald Croi³, Helene Augu⁴, Christian Kowalewski⁴ and Andreas Brink⁵

¹European Commission, Joint Research Centre, Ispra, Italy; ²Department of Geography and Environmental Studies, Stellenbosch University, Stellenbosch, South Africa; ³Institut Géographique National (IGN) FI, Paris, France; ⁴Joanneum Research, Forschungsgesellschaft mbH, Graz, Austria; ⁵EOXPLOR, Weil am Rhein, Germany; ⁶space4environment sàrl, Niederanven, Luxembourg; ⁷Officine National des Forêts International, Nogent sur Marne, France; ⁸Gisbox, Bucharest Romania; ⁹Luxspace Sàrl, R.C.S., Luxembourg B, Luxembourg; ¹⁰Sword Group, Paris, France

ABSTRACT
The Copernicus High-Resolution Hot Spot Monitoring activity (C-HSM) delivers a global dataset of Key Landscapes for Conservation (KLC), which are characterized by pronounced anthropogenic pressures that require high mapping accuracy. Detailed land cover and land cover change map products are freely available through the activity and include extensive map production accuracy assessments. Without a complete understanding of the map products’ spatial, temporal and logical consistencies, quality or quantified confidence levels, usability is reduced and can affect stakeholder decision-making and the implementation of sustainable solutions. For the quantitative accuracy assessment, a stratified random sampling approach was implemented where special emphasis was placed on (i) allocation of sampling units for rare land cover change categories; (ii) effective and accurate labelling of large numbers of sampling units; (iii) accuracy and area estimation in one consistent approach; and (iv) derivation of confidence intervals for all accuracy measures and area estimates. To handle correlations, large uncertainties, and complex probability density functions, bootstrapping was applied instead of analytical equations, which are based on normality assumptions. This paper presents the Quality Assurance and Quality Control framework applied to validate all the C-HSM thematic map products. The Kundelungu-Upemba KLC product results are presented as our use case.

Introduction
The free, full and open data policy of many current Earth Observation (EO) satellite systems makes it easy to produce satellite image-based thematic land cover (LC) and land cover change (LCC) map products (Turner et al., 2015). This has also encouraged the development of novel cloud processing services to produce LC products such as those based on data cubes (Strobl et al., 2017). Consequently, it has become relatively straightforward to produce thematic LC maps even at medium to high spatial resolution (i.e.10–30 m), but depending on the number of LC/ LCC classes discriminated, mapping quality can vary. Moreover, the information that is often missing from such LC products is a detailed description of the validation methodology applied and the assessment of a product’s accuracy (Morales-Barquero et al., 2019), both of which can be diverse and based on a variety of procedures (Qian et al., 2020; Stehman & Foody, 2019). Unverified LC map products or products validated without a fully transparent quality assurance (QA) scheme can not only affect usability but also influence trust in the products themselves. Thus, a rigorous and science-based QA and quantitative accuracy estimation procedure is required to guarantee the general quality of such products. To maximise usability and trust, not only is the overall LC and LCC map validation and assessment needed, but also the accuracies per thematic class as well as other quality assurance steps, such as the spatial, temporal and logical consistency checks. Many studies have focused on establishing a proper workflow to quantify thematic accuracies and area estimations of LC (Congalton, 1991; Olofsson et al., 2014; Pal, 2002), but most have concentrated solely on the direct determination of the overall and thematic accuracies, while confidence intervals as well as the underlying map product’s spatial, temporal and logical consistency (Burnicki, 2011; Gong & Mu, 2000; Stehman & Foody, 2019) are often overlooked or the applied procedures are not reported. However, both are vital pieces of information required to understand the overall effects of the map’s final accuracy, especially if they are used by decision makers and stakeholders.

CONTACT
Zoltan Szántoi  zoltan.szantoiremote-sensing-biodiversity.org  European Commission, Joint Research Centre, 21027, Ispra, Italy

© 2021 Informa UK Limited, trading as Taylor & Francis
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
The use of an accepted and portable classification scheme such as the Food and Agriculture Organization’s Land Cover Classification System (LCCS) (Di Gregorio, 2005), which was developed to provide a consistent, hierarchical framework for LC mapping, is necessary and, moreover, also complies with the Intergovernmental Panel on Climate Change (IPCC) recommendations (IPCC, 2006) as it contains no less than the six proposed LC/land use classes (forest, cropland, grassland, wetland, settlement and others). The LCCS, through its hierarchical structure, is suitable for the definition of a variety of LC classes and applicable to mapping LC on different map scales and to different extents. By applying such a common approach to LC map production, it also facilitates the assessment of accuracy evaluations. Accuracy evaluations, however, will also be affected by other factors such as the underlying data and the used classification approaches (Wulder et al., 2018) as well as the training/validation data and map interpreter’s experience and training (Pengra, 2020) or even the software used, i.e. practicability, ease of use.

The Copernicus Programme is the world’s largest civilian EO programme with operational services in six thematic areas – land, marine, atmosphere, climate change, emergency management and security. These services and the bulk of their products are available free under an open license (www.copernicus.eu/). The Copernicus Land Monitoring Service Thematic High-Resolution Hot Spot Monitoring activity (C-HSM, https://land.copernicus.eu/global/hsm) aims to produce detailed LC and LCC information for Key Landscapes for Conservation (KLC) which are primarily identified in African, Caribbean and Pacific regions. All C-HSM map products are based on the LCCS classification scheme and, therefore, the adapted validation protocol should also be the same across all map products regardless of location, landscape or cover type (Szantoi, Brink et al., 2020). A robust QA and Quality Control (QC) procedure is required to evaluate all these map products, because the QA process provides confidence that the final products satisfy the desired minimum and adequately defined map quality and the QC procedure ensures the accuracy and precision of the process throughout map production, including data collection.

The C-HSM maps are expected to be used in decision-making processes and are therefore required to be evaluated based on established methodologies that provide rigorous spatial-temporal-logical consistency (Wehmann & Liu, 2015) as well as quantitative accuracy measures. As noted by Stehman and Foody (2019), rigorous assessment of such map products is necessary to ensure trust in scientific results. Furthermore, the assessment procedure must be well documented and provide adequate explanations to help reduce the potential of the misuse of a map product due to lack of understanding or misinterpretation of the results (Stehman & Foody, 2019).

The objectives of this paper are (1) to present the theoretical basis of the entire QA/QC validation protocol used for evaluating the C-HSM products and (2) demonstrate the QA/QC validation protocol applied on the thematic LC and LCC maps of one of the C-HSM products: Kundelungu-Upemba KLC. We expect our experiences to encourage others to adapt the developed QA/QC protocol and that such an activity helps improve user confidence in produced map products.

**Study area**

**The Kundelungu-Upemba key landscape for conservation**

During the C-HSM activity, we mapped over 1.4 million square kilometers of land (Szantoi et al., 2021; Szantoi, Brink et al., 2020) using the same mapping methodology as well as the same validation protocol presented in this paper. To explain step-by-step in detail the QA/QC validation protocol here, we selected a specific KLC, the Kundelungu-Upemba. The Kundelungu-Upemba KLC (C-HSM code: CAF11) is located in the southeastern part of the Democratic Republic of the Congo and consists of two national parks and three hunting reserves and covers an area of 47,318 km². The KLC is part of the central Zambezian miombo woodland ecoregion, and is characterised by the presence of many different habitats, including wet and dry miombo, thicket, shrubland, gorges, savannah woodland, high altitude grasslands on the plateaus and flooded grasslands in the swamps. The CAF11 LC and LCC map products cover the entire area for the reference periods (2000 and 2016). Figure 1 presents the map of the various KLCs mapped so far by the C-HSM programme across Africa (Szantoi et al., 2021; Szantoi, Brink et al., 2020). The detailed land cover classes are presented in Table 1.

**Methods**

**The Quality Assurance and Assessment Framework for land cover map validation**

Land cover/change map products were validated covering an area of approximately 1.4 million km² in the Caribbean, African and Pacific regions (Szantoi et al., 2021; Szantoi, Brink et al., 2020), where LC objects are represented by polygons and the minimum mapping unit for LC is 3 ha and for LCC is 0.5 ha. While the product quality is an end user concern, intermediate quality measurements and evaluations throughout the
production chain are just as important. Therefore, a comprehensive evaluation of different quality metrics at all stages of the production process is, in practice, an enabler of usability and can influence other map production processes including the need to provide new imagery or data. The Quality Assurance and Assessment Framework (QAAF) was applied to all the map products produced within the C-HSM programme.

**Utilising transferable location-independent land cover classification schemes**

The classification scheme for the C-HSM products is based on the Land Cover Classification System (LCCS, Di Gregorio, 2005). All C-HSM products, depending on the product’s classification level (dichotomous or modular), are mapped using these defined classes (Table 1).

The QC implemented here is essential because the LC changes do not generally represent more than 1% to 2% of a mapped KLC. Changes in LC are interpreted as categorical: one LC class is replaced by another LC class (Feranec et al., 2016) with a minimum mapping unit of 0.5 ha (C-HSM criteria). An important point to note is that a change in LC does not necessarily lead to a change in land use or land affectation. Land cover conversion is considered as a complete replacement of one type of LC by another. Other LCCs that affect the character of the LC without changing its overall classification are more subtle, and those that correspond to a modification. This is the case of agricultural intensification (small-sized fields of herbaceous crops into large-sized fields) or the progressive degradation of plant cover (closed forest into open forest) linked to gradual human pressure, climate change (desertification), soil erosion, over-grazing and the recurrence of bush fires, which is very common in subtropical regions.

The phenomenon of seasonality, which is marked in subtropical regions, further accentuates the impression of changes. For example, the physiognomy of a semi-deciduous evergreen mixed forest or a dry forest does not appear at all in the same way whether it is the rainy or the dry season, yet the nature of the object

Figure 1. Key Landscapes for Conservation mapped by the C-HSM across Africa.
remains the same, i.e. forest. The spectral, radiometric and phenological variations linked to these seasonal phenomena are therefore not taken into account as land cover changes. Changes of the same nature could be aggregated into categories of change. Based on the LCCS guideline, eight categories of change have been defined (Table 2).

Table 1. The Land Cover Classification System applied to the C-HSM map product specification.

| Dichotomous level | Modular level | Map Code | Class name |
|--------------------|---------------|----------|------------|
| A1 – Cultivated and managed terrestrial areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Tree Crop(s) Crop Cover: Plantation(s) | 31 | Tree Plantations Large |
| A1 – Cultivated and managed terrestrial areas | Continuous Small Sized Field(s) (<2 ha) Of Tree Crop(s) Crop Cover: Plantation(s) | 32 | Tree Plantations Small |
| A1 – Cultivated and managed terrestrial areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Tree Crop(s) Crop Cover: Orchard(s) | 33 | Tree Orchard Large |
| A1 – Cultivated and managed terrestrial areas | Continuous Small Sized Field(s) (<2 ha) Of Tree Crop(s) Crop Cover: Orchard(s) | 34 | Tree Orchard Small |
| A1 – Cultivated and managed terrestrial areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Shrub Crop(s) | 55 | Shrub Crops Large |
| A1 – Cultivated and managed terrestrial areas | Continuous Small Sized Field(s) (<2 ha) Of Shrub Crop(s) | 56 | Shrub Crops Small |
| A1 – Cultivated and managed terrestrial areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Herbaceous Crop(s) | 59 | Herbaceous Crops Large |
| A1 – Cultivated and managed terrestrial areas | Continuous Small Sized Field(s) (<2 ha) Of Herbaceous Crop(s) | 60 | Herbaceous Crops Small |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Closed (>70-60) Trees | 77 | Forest Closed |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Closed To Open (100–40)% Trees | 78 | Forest Open |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Closed To Open (100–40)% Shrub(s) | 112 | Shrub(s) Closed |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Closed To Open (100–40)% Herbaceous Vegetation | 148 | Grasslands Closed |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Open (40 – (20–10)% Shrub(s) | 116 | Shrub(s) Open |
| A12 – Natural or semi-natural terrestrial vegetation | Continuous Open (40 – (20–10)% Herbaceous Vegetation | 152 | Grasslands Open |
| A2 – Cultivated aquatic or regularly flooded areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Woody Crops – Aquatic Or Regularly Flooded | 155 | Agriculture flooded – Woody Large |
| A2 – Cultivated aquatic or regularly flooded areas | Continuous Small Sized Field(s) (<2 ha) Of Woody Crops – Aquatic Or Regularly Flooded | 156 | Agriculture flooded – Woody Small |
| A2 – Cultivated aquatic or regularly flooded areas | Continuous Large To Medium Sized Field(s) (>2 ha) Of Graminoid Crops – Aquatic Or Regularly Flooded | 159 | Agriculture flooded – Graminoid Large |
| A2 – Cultivated aquatic or regularly flooded areas | Continuous Small Sized Field(s) (<2 ha) Of Graminoid Crops – Aquatic Or Regularly Flooded | 160 | Agriculture flooded – Graminoid Small |
| A24 – Natural or semi-natural aquatic vegetation | Closed (>70-60)% Trees – Aquatic vegetation | 165 | Forest Closed Aquatic |
| A24 – Natural or semi-natural aquatic vegetation | Open General (70-60)-(20-10)% Trees – Aquatic vegetation | 166 | Forest Open Aquatic |
| A24 – Natural or semi-natural aquatic vegetation | Closed To Open (100–40)% Shrub(s) – Aquatic vegetation | 171 | Shrubs Closed Aquatic |
| A24 – Natural or semi-natural aquatic vegetation | Very Open (40 – (20–10) % Shrub(s) – Aquatic vegetation | 175 | Shrub(s) Open Aquatic |
| A24 – Natural or semi-natural aquatic vegetation | Closed To Open (100–40)% Herbaceous Vegetation – Aquatic vegetation | 178 | Grasslands Closed Aquatic |
| A24 – Natural or semi-natural aquatic vegetation | Very Open (40 – (20–10) % Herbaceous Vegetation – Aquatic vegetation | 182 | Grasslands Open Aquatic |
| B15 – Artificial surfaces and associates area(s) | Built Up Area(s) | 184 | Urban – Built Up |
| B15 – Artificial surfaces and associates area(s) | Non Built Up Area(s) | 185 | Urban – Not Built Up |
| B16 – Bare Area(s) | Bare Area(s) | 11 | Bare |
| B27 – Artificial water bodies, snow and ice | Artificial Waterbodies (Flowing) | 166 | Flowing Artificial Waterbodies |
| B27 – Artificial water bodies, snow and ice | Artificial Waterbodies (Standing) | 187 | Standing Artificial Waterbodies |
| B28 – Natural water bodies, snow and ice | Natural Waterbodies (Flowing) | 190 | River |
| B28 – Natural water bodies, snow and ice | Natural Waterbodies (Standing) | 191 | Lake |
| B28 – Natural water bodies, snow and ice | Snow | 192 | Snow |
| B28 – Natural water bodies, snow and ice | Ice | 193 | Ice |

**Image and data integrity used for map production**

Data integrity refers to the compliance of the output datasets with technical specifications (Tables 3, 4 and 5). These technical specifications can relate to file formats, attribution, projections, metadata and the spatial accuracy of the data produced. The QA process for data integrity is structured into three
sequential parts: logical consistency, temporal consistency and spatial consistency, and takes into account data quality components listed within existing geographic data quality guidelines. Each of these components is further detailed in the following sections. A quantitative summary of the results for data integrity is not implemented due to the varying importance of the different elements for the overall product quality and subsequent steps in the validation process. For example, whereas an incorrect file name will not impact upon any following assessment components, issues relating to projection prohibit further processing. Depending on the assessment results for the technical specification criteria (Tables 3, 4, and 5) the map product is categorised as:

- Compliant: All data integrity elements are compliant.
- Partially compliant: Few elements (2–5) have failed the check, but these do not affect the subsequent quality assessment components.
- Non-compliant: The detected quality issues are critical or more than five elements have failed the check. Criticality of a data integrity element is thereby defined as whether or not it’s non-compliance would require halting the validation process (e.g., corrupted file, major geometric shifts or broken projection).

Verifying logical consistency. The logical consistency addresses the degree of adherence to logical rules such as data format, structure, topology, attribution and logical relationship, whose characteristics are important for a base map product, but even more so for LCC procedure/products or for products covering different geographical regions. The logical consistency assessment is necessary to ensure machine readability.

For the C-HSM, it comprises 14 different data elements presented in Table 3 that have been organised into categories based on the format specifications, the mapping criteria, the metadata and the topology, following the guidelines established by the Infrastructure for Spatial Information in the European Community (INSPIRE, 2007). Data elements are checked in the order of the table. The compliance of each element to the specification criteria of logical consistency is recorded in a binary format. The full check is mainly implemented in the Python programming language.
using the Geospatial Data Abstraction Library (GDAL) and the OpenGIS Simple Features Reference Implementation (OGR) as well as the PyQGIS module. More detailed checking content is described in Table 3. This encompasses aspects of the syntactic correctness of file formats and field definitions. Mapping criteria relate to attribute entries and the defined Minimum Mapping Unit (MMU) checked in terms of polygon area. Metadata specification criteria are based on the technical guidelines for INSPIRE conformant metadata (INSPIRE 2007) and checked using the dedicated test suites available. Polygon topology rules are assessed utilising the Geometry Engine – Open Source (GEOS) software library based on the JTS Topology Suite and implemented via the dedicated PyQGIS tools.

Verifying temporal consistency. Land cover and LCC map products require that the input EO imagery for the desired time periods is available and suitable, i.e. LC information is not obscured. The temporal consistency assessment of the input imagery is evaluated on the basis of production image footprints and their accompanying image acquisition dates. The C-HSM check consists of two data integrity elements shown in Table 4. The first element refers to the reference period of the input imagery measured in years. The check consists of a calculation of the absolute and relative amount (%) of all imagery outside of this defined period. If the imagery outside the reference period exceeds an arbitrarily defined threshold of 35% of total imagery the element is deemed non-compliant. In general, establishing cut-off thresholds for temporal data quality elements is somewhat challenging as data availability over certain ecosystems may be restricted. The second data integrity element for the temporal product consistency focuses on the season from which imagery has been acquired. The automated check produces a basic overview of the count of images per month, which can be evaluated by an operator. Thereby the table highlights the coverage of certain seasons expected in the AOI. If the imagery only stems from a single season the checking script issues a warning and the result is deemed non-compliant.

Verifying spatial accuracy and consistency. Spatial (positional) accuracy and consistency ensure that there is little or no spatial difference between two measured point locations, e.g., differences in absolute satellite image geolocation, differences between products and/or imagery (Table 5). The defined requirement criteria are expressed in the Root Mean Square Error
Assessing qualitative thematic accuracy

The qualitative systematic accuracy (QSA) provides a benchmark that map users can use to understand the applicability of the information presented in the map product. To provide such information we employed the qualitative assessment grid approach (Büttner et al., 2004), where a 10 km by 10 km grid is generated and overlaid as shown in Figure 2. As described in Büttner et al. (2004), the grid size was chosen to provide sufficient area from which landscape-level information can be determined. This allows the visual interpretation to take into account the landscape-level context when assessing fine spatial resolution detail. This grid is used as the basis of the systematic qualitative control and such a grid also provides a good sense of the heterogeneity across the region being mapped because the qualitative validation makes it possible to examine all dominant classes with a density proportional to their occurrence as presented over the entire map area. Essentially, the procedure forces the assessment to look across the entire map product extent, providing a good understanding of the overall qualitative map issues as well as potentially identifying regions where systematic issues were produced. In order to achieve this, the grid cells are visually analysed to identify issues related to qualitative map characteristics such as mapping the wrong LC class, LC class doubt, the omission of a particular LC class (see Table 1 for map legend details), the delineation of LC classes and adherence to the MMU requirement (Table 3). For this QSA procedure, the same EO data/imagery and map sources that were utilised by the thematic map producers must be used. For the C-HSM, the QSA process had at least 50% of the grid cells evaluated with the following conditions respected:

- the qualitative assessment covers the entire area;
- the verification is balanced between areas where major LCCs are present as well as areas with no or limited changes;
- all LC classes are represented.

When a qualitative issue is observed concerning the mapped LC and LCC classes, the location of the identified error and the type of error are marked (spatial location) and described. These are all stored in a geospatial database for future reference and potential corrections.

Depending on the number of issues identified, one of three quality flags was assigned to each verified cell:

- accepted (green) – no qualitative thematic issues were identified;

Figure 2. The Qualitative Systematic Assessment sampling grid overlaying the Kundelungu-Upemba KLC land cover map for the year 2016. The left panel a presents the thematic map, while the right panel b presents the EO image mosaic.
• conditionally accepted (yellow) – one or two thematic issues were identified and logged;
• unaccepted (red) – three to five thematic issues were identified and logged.

**Quantitative assessment of the land cover and land cover change thematic accuracy**

A target overall accuracy for medium to high-resolution imagery-based products is often defined as >85% (Foody, 2002). Hence, the requirements for the overall C-HSM map accuracy were set to this value. However, thematic class accuracies (Table 1) rarely reach this value for all types of land cover (Gao et al., 2020), therefore a value of 75% was chosen as minimum individual class accuracy. Thematic per class accuracy for change mapping (Table 2) was then defined slightly below this value, at >72%, due to the additional complexity in assessing some of the land cover changes (Szantoi et al., 2021; Szantoi, Brink et al., 2020). In exceptional cases, the thematic accuracies might be lower than these thresholds due to the difficulty in distinguishing a particular class in a certain KLC. The C-HSM map accuracy requirements were established based on user’s needs, as accurately classified LC map products are needed for many applications, such as ecosystem modeling (Graﬁus et al., 2016) and ecosystem valuation (Foody, 2015), besides the general need for accurate representation of the ground cover for policy making. The Quantitative Assessment – Validation’s (QAV) general workflow is made up of three main parts: the sampling design, the response design and the quantitative analysis. Each of these parts is described in detail in the following sections.

**Sampling design.** The QAV sampling design is based on probability sampling where the two conditions defining probability sampling are (1) that the inclusion probability must be known for each unit selected in the sample and (2) the inclusion probability must be greater than zero for all units in the population (Stehman, 2001). Based on Olofsson et al. (2020), Olofsson et al. (2014), and Stehman (2009), our approach to achieve this uses the spatially explicit LC and LCC maps for stratification in order to increase the precision of estimated accuracies. Furthermore, this is used to gain control over the sample size allocation for infrequent LC or LCC categories in particular. The stratification includes the stable LC classes listed in Table 1 and the aggregated LCC classes listed in Table 2 as strata. The allocation of the number of sampling points for the different strata depends on the monitoring objectives. For the QAAF presented, the Neyman optimal allocation procedure (see e.g., Cochran, 1977) was employed because of its advantages for estimating the area of change and overall accuracy, as well as the allocation based on anticipated user accuracies as described in Stehman and Foody (2019).

The first step in the sampling design is to determine the sample size \( n_c \) for each stratum for an accepted standard error in the error of commission using equation 1:

\[
 n_c = \frac{d^2}{\left( p_c (1 - p_c) \right)} \tag{1}
\]

where \( d \) is the desired half-width of the confidence interval, and \( p_c \) is the anticipated user’s accuracy of category \( c \) (Stehman & Foody, 2019). However, focusing on the error of commission only would weaken the statistical analysis of infrequent LC classes. For example, LCC maps will often be made up of significant areas where no change has occurred and therefore would only be sampled by a low number of sampling points which impedes the detection of omission errors. To solve this potential shortcoming, a second step is introduced in order to select additional sampling points according to Neyman optimal allocation, which minimises the variance of the estimator of overall accuracy and area estimates for a given total sample size \( n \) (Ayala-Izurieta et al., 2017; Stehman, 2012) using equation 2:

\[
 n_k = \frac{n N_k S_k}{\sum_{k=1}^{L} N_k S_k} \tag{2}
\]

where \( n_k \) is the number of sample points in stratum \( k \), \( n \) is the total sample size, \( N_k \) is the population size of stratum \( k \), \( L \) is the number of strata and \( S_k \) represents the stratum’s standard deviation. The result of applying equation 1 and 2 is an uneven distribution of the number of sampling points for the different strata, with a defined minimum number of sampling points per stratum. The ramifications of this are further discussed below.

**Response design.** The procedures for assigning the class label to each sampling point are the basis for the QAV response design. The response design is based on the interpretation of the thematic class at a chosen point as the sampling unit taking into account product specifications (MMU, class definitions, etc.) and is made up of four steps.

In the first step, the review of the reference data labelling rules, for example, related to the minimum mapping units or seasonal water dynamics, is performed by the interpretation team. The team collectively works together on a subset of sampling points and discusses these cases to establish consistency in the interpretation of all LC and LCC classes that are part of the map being assessed. This is an important step before proceeding for obtaining the interpretations for the whole reference sample because all validation team members should understand the LC and LCC classification as it has been applied to the region at hand. This means that even though a map legend stays the same across map products, interpreters must re-establish consistency each time a new area has been mapped because the underlying landscape has changed.
In the second step, a “blind interpretation” is applied, without the knowledge of the product’s thematic classes for the selected sample points. One difficulty during this step is determining distinct boundaries between classes, in particular when the natural boundaries are not easily distinguished, for example, in a continuum of vegetation. Indeed, some elements in the landscape can have fuzzy limits in space or in time. Moreover, shifts can be observed when comparing medium spatial resolution satellite images (10–30 m) and finer spatial resolution imagery (<2.5 m), increasing the difficulty in determining a clear limit for small or narrow map elements, e.g., roads, rivers or settlements. Such limits are also difficult to determine when a specific class is affected by seasonal changes over time, e.g., aquatic landscapes and seasonal vegetation dynamics, and fall between the limits of the class definitions. In other cases, the feature boundaries in the image are very clear but the blind interpretation sample point fall on this boundary between both classes and could hence be classified in one or the other class. Another obstacle comes from the coarse spatial resolution imagery from past reference years where no other information is available and when the available imagery does not allow a clear interpretation. This blind approach may therefore underestimate the accuracy for complicated and difficult cases. In step three, a “plausibility interpretation” is therefore made afterwards the blind interpretation. The plausibility interpretation is performed for all sampling points where the blind interpretation results are in disagreement with the LC or LCC maps. Based on this information, the interpreter provides the plausibility information related to the following scenarios:

- both, the map and the blind interpretation are plausible;
- the blind interpretation is plausible, the map is not plausible;
- the map is plausible, the blind interpretation is not plausible;
- both, the map and the blind interpretation are not plausible.

During this plausibility step, the interpreter can thus also revise the blind interpretation in the event that an interpretation error is observed. The result of these interpretation steps are therefore two reference databases, one for the blind and one for the plausibility interpretation. The final QAV response design fourth step is a review of 10% of the sampling points to ensure consistent interpretation by the interpretation quality team.

Please note that the accuracy analysis described in the next section is performed twice, (i) for the blind and (ii) for the plausibility interpretation, as both provide relevant information for the use of the LC and LCC maps as described above.

To ensure that the validation data is of higher quality compared to the map data, very high-resolution satellite imagery and ancillary data are used together with the dense time-series of Sentinel-2 and Landsat satellite imagery over the monitoring period in one consistent visual interpretation step.

Analysis – map accuracy and class area estimation. A confusion matrix is generated based on the inclusion probability for each sampling point (Selkowitz & Stehman, 2011) and provides the basis for all subsequent analyses. The rows of the confusion matrix represent the map labels, the columns represent the reference data and the number of sampling points in the map and in the reference are described by \( n_i \) and \( n_j \). The confusion matrix is reported in terms of estimated area proportions, \( p_{ij} \), and not in terms of sample counts. For stratified random sampling in which the strata corresponds to the map classes, the direct estimator is presented in equation 3:

\[
\hat{p}_{ij} = W_i \frac{n_{ij}}{n_i}
\]

where \( W_i \) is the proportion of area mapped as class \( i \) in the confusion matrix, \( p_{ij} \) represents the proportion of area for the population that has map class \( i \) and reference class \( j \), where “population” is defined as the full region of interest (Olofsson et al., 2014). Two confusion matrices were calculated which are based on (i) the blind, and (ii) the plausibility reference data sets. All accuracy measures were derived separately from the two confusion matrices, and both, the blind and plausibility assessment results are provided.

The uncertainty of the accuracy measures and the area estimates are quantified by deriving confidence intervals that provide a range that includes the true (but unknown) value with a specified probability. To compute the confidence intervals, we applied bootstrapping as described e.g., in Schreuder et al. (2004) and Gallaun et al. (2015), where the selection of the bootstrap samples emulates the actual sample selection method. Bootstrapping is thereby performed by means of repeated sampling from the whole sample set with replacement. The percentile interval method is used as described in Efron and Tibshirani (1986), where the percentiles are estimated directly from the bootstrap distribution. For the 95% confidence interval, the interval between 2.5% (lower bound) and 97.5% (upper bound) of the bootstrap quantiles is determined. All elements of the error matrix \( p_0 \) and all derived accuracy measures and area estimates are calculated for each run. Confidence limits around the derived overall accuracy, error of omission and error of commission as well as around the area estimates are derived with this approach. A main advantage of the bootstrapping approach is that no assumption is required on normal distribution.
**Map accuracy assessment**

Once \( \hat{p}_{ij} \) is computed for all cells of the confusion matrix, the overall accuracy is estimated by summing the values on the main diagonal of the \( q \) classes of the confusion matrix using equation (4) (Olofsson et al., 2014):

\[
\hat{O} = \sum_{j=1}^{q} \hat{p}_{jj} \tag{4}
\]

Two common accuracy measures are then derived for each LC or LCC class.

**User's accuracy** of class \( i \) as the proportion of the area mapped as class \( i \) that has reference class \( i \), where \( \hat{p}_i \) are the row totals of the confusion matrix,

\[
\hat{U}_i = \frac{\hat{p}_{ii}}{\hat{p}_i} \tag{5}
\]

and **producer's accuracy** of class \( j \) as the proportion of the area of reference class \( j \) that is mapped as class \( j \), where \( \hat{p}_j \) are the column totals of the confusion matrix,

\[
\hat{P}_j = \frac{\hat{p}_{jj}}{\hat{p}_j} \tag{6}
\]

**Area estimation**

Due to the fact that errors of omission and errors of commission in general are not equal, the area of the respective LC and LCC would be biased if the mapping results are directly used to determine the area of the respective categories. In order to remove this bias, the confusion matrix is used for area estimation, where each cell of the confusion matrix gives the unbiased estimator of the proportion of area and the area proportions for each category \( j \) are directly estimated from the column totals. An unbiased estimator of the total area \( \hat{A}_j \) of category \( j \) is then, according to Olofsson et al. (2013):

\[
\hat{A}_j = \hat{A}_{\text{tot}} \times \hat{p}_j
\]

with \( \hat{A}_{\text{tot}} \) being the total area mapped.

This can be viewed as an "error adjusted" area estimator because it includes the area of map omission and leaves out the area of map commission.

**Results**

**Applying the quality assurance and assessment framework to large area mapping**

**Image and data integrity**

**Land cover map – validating logical, temporal and spatial consistencies.** For CAF11, the defined requirements thresholds (RMSE < 100 m) were achieved for Absolute Positional Accuracy (RMSE = 13.18 m CI [10.15 m, 18.98 m] and Relative Positional Accuracy (RMSE = 54.0 m CI [45.18 m, 67.13 m]. The local RMSE for the 2016 CAF11 LC map is illustrated in Figure 3, showing the absolute spatial accuracy. Twenty sampling points were selected and their local RMSE is presented as a bar graph (left side – Map a, Figure 3). The relative spatial accuracy of the LC classification polygons is also presented in Figure 3 (right side – Map b) for the 2016 LC map. The RMSE of the sampled nodes was computed by comparing it to the corresponding satellite imagery used for the map production. The entire evaluation results can be downloaded from: [https://land.copernicus.eu/global/hsm](https://land.copernicus.eu/global/hsm).

The imagery used to produce the LC map was within the correct reference period (2016) and also covered different periods in order to identify seasonality: 52 images acquired during the dry season (May 2016 to September 2016) and 58 images acquired during the wet season (October 2016 to April 2016), Figure 4.

**Land cover change map – validating logical, temporal and spatial consistencies.** The Positional Accuracy was verified by a visual comparison only to very high-

![Figure 3](image_url). Absolute (Map A – left side) and Relative (Map B – right side) spatial accuracy of the input imagery and classification polygons for the 2016 Kundelungu-Upemba KLC land cover map.
resolution data and the LC map. The Relative Positional Accuracy (RMSE = 73.12 m CI [61.18 m, 90.89 m] and the reference period of the input imagery (2000) and seasonality (35 images from the dry season, whereas 19 images from the wet season) are reported.

The following map (Figure 5) shows the relative spatial accuracy of the classified polygons for the CAF11 LCC map. The RMSE of the sampled nodes was calculated by comparing it to the data used for the production. The entire evaluation results can be downloaded from: https://land.copernicus.eu/global/hsm.

Qualitative assessment
Land cover map – qualitative systematic accuracy.
The QSA map presented in Figure 6 (right side) displays the results for each grid cell assessed according to the map legend. The QSA final grid statistics were the following based on a total of 540 grid cells assessed: accepted (green) = 42%, conditional (yellow) = 55%, and unaccepted (red) = 3%.

Inside each conditional and unaccepted grid cell, the map user will also see coloured dots (see map legend, Figure 6) that represent the identified qualitative issues within a particular cell. They are classified according to the five different categories (Table 6).

The QSA results map presents the spatial distribution of the identified issues, which the map user can link to specific areas and landscapes. However, not all issues are uniformly represented by either type or spatial distribution. The spatial distribution can be easily observed using the map and the type is summarised in Table 6. See also the bar graph on the left side of Figure 6.

Note that almost half (45.5%) of the issues identified for the CAF11 QSA were related to omitting an LC class, and 32.7% with LC class delineations as shown in Figure 7, where an area of the same land cover (77) has been wrongly delimited (yellow arrows), introducing class 78. The correct delineation should be at the green arrow, beyond where class 78 is prevalent.

These results can be applied for corrective purposes to either rectify the identified issues or improve the image classification and map production procedures.

| Issue          | Issue description                              | Count | Percentage (%) |
|----------------|-----------------------------------------------|-------|----------------|
| Omitted LC class | LC classes > MMU not delineated                | 209   | 45.5           |
| Delineation issue | Incorrect borders                        | 150   | 32.7           |
| Wrong LC class  | Incorrect class                               | 76    | 16.6           |
| Doubt in LC class | Thematic label in question              | 21    | 4.6            |
| Other remarks   | Other qualitative issues                      | 3     | 0.7            |

Land cover change map – qualitative systematic accuracy. The QSA approach was applied to the CAF11 LCC map product as well and Figure 8 presents the results. The LCC QSA grid cell results revealed that 73% were accepted, 24% were conditionally accepted and 3% were unaccepted – an in-depth discussion on the relevance of the QSA is provided in the Discussion (Practicality of the quality assurance and assessment framework section). Note that the total number of issues was lower for the LCC than for the LC QSA; 229 issues versus 459. Most of the issues are related to geometrical inconsistencies between the vector and reference images, but also commission and omission errors (Table 7).

Quantitative assessment
Validation.
Land cover and land cover change map – sampling design. To fulfil the map validation requirements for the stratified random sampling, the CAF11 LC and LCC maps were used as the basis for the stratification. Based on the LC class-specific user accuracies were estimated a priori to be 0.75 and 0.72 for each of the aggregated LCC classes, the number of sampling points per class are calculated using equation 1 in order to achieve an acceptable standard error in the error of commission of 0.05. Applying equation 1, the minimum number of sampling points was calculated to be 81 points per aggregated LCC class and 75 points per LC class. As described in the Sampling design section, this distribution of the sample points would weaken the statistical analysis of rare LC or LCC categories such as forest cover loss, as the dominating categories would be sampled with few sampling points only in relation to the proportion (see also in the Discussion section). To overcome this problem, additional sampling points were allocated using Neyman optimal allocation as defined in equation 2, which resulted in a total of n = 3,786 sampling points for CAF11 (Figure 9).

Land cover and land cover change map – response design. For each sampling point, the reference class was visually interpreted based on very high resolution (VHR) imagery over the monitoring period from a number of different sources, including Google Earth (https://earth.google.com), Microsoft BING

| Issue          | Issue description                              | Count | Percentage (%) |
|----------------|-----------------------------------------------|-------|----------------|
| Delineation issue | Incorrect borders                        | 64    | 27.9           |
| No change       | Commission error                            | 52    | 22.7           |
| Omitted LC class change | LCC classes > MMU not delineated | 44    | 19.2           |
| Wrong LC change | Incorrect class                               | 38    | 16.6           |
| Doubt LC class change | Thematic label in question              | 30    | 13.1           |
| Other remarks   | Other qualitative issues                      | 1     | 0.4            |

Table 7. Summary of the QSA issues identified for the CAF11 – Kundelungu-Upemba KLC thematic map product for the year 2000.
(https://www.bing.com/maps) and ESRI basemaps (https://www.esri.com/en-us/arcgis/products/data/data-portrait/basemaps), dense time-series from Copernicus Sentinel-2, Landsat and ASTER optical imagery (Figure 10) and contour lines that we derived from the ALOS global digital elevation model. Additionally, the seasonal matrix view for visualising phenology/seasonal developments over the years (Figure 4) for each sampling point was also checked. Through this procedure, a highly accurate labelling of the sampling points in one consistent interpretation step for LC and LCC was done.

**Map accuracy and class area estimation.**
As described in the Analysis - map accuracy and class area estimation section, the final steps in the quality assurance and assessment framework concerns the quantitative accuracy assessment and area estimations. The results presented here follow the descriptions from the Methods section, including the application of bootstrapping to derive the confidence intervals; all error measures and class area estimates were thereby computed 1,000 times by means of repeated sampling from the whole sample set with replacement. The percentile method was then applied to derive the 95% confidence intervals directly from the interval between 2.5% and 97.5% of the bootstrap quantiles. These values are applied to the final quantitative accuracy assessment and area estimations.

**Quantitative accuracy assessment.** The results of the quantitative accuracy assessment are presented in Tables 8 and 9 for the LC and LCC maps, respectively.

The assessment shows a high overall accuracy for the LC map of 95.3% for the plausibility approach. With the exception of the producer’s accuracy of the LC class “Shrubs Open Aquatic” (class code: 175), all other LC classes which are covered by more than 20 reference samples exceed the expected thematic accuracies of 75%. Furthermore, all aggregated LCC categories which are covered by more than 20 reference samples exceed the expected thematic accuracies of 72% with the exception of the producer accuracy of “Forest Cover Gain”.

**Area estimation.** The second part of the QAV is the area estimation, which is based on the results of the blind assessments. As presented earlier, errors of commission and omission are generally not the same. Therefore, LC maps tend to overestimate or underestimate the true area of the delineated classes, i.e. LC maps are biased. To adjust this bias, the error-adjusted area estimates are directly generated from the confusion matrix by applying equation 7. The resulting area estimates and confidence intervals are presented in Table 10 – for the year 2016, and in Table 11 for the areas changed from 2000 to 2016.

As described in the Methods - Area estimation section, there are significant differences in the area estimates between (i) mapped areas and (ii) area estimates derived by the sampling approach. This is particularly true for the evaluation of LCC. For example, “Artificial Surface Expansion” is mapped at 2,983 ha. In comparison, the sampling approach shows a much larger expansion of artificial surfaces of 3,825 ha. Please note that in addition to the area estimates, the sampling approach also provides confidence intervals which estimate the artificial surface expansion between 3,210 ha and 4,533 ha at the 95% confidence level. The wide confidence intervals are expected, as omissions of land change in maps can introduce large uncertainty in the accuracy as well as in area estimates (see e.g., Olofsson et al. (2020) for a detailed discussion of this topic).

![Figure 4. Seasonality matrix for the Kundelungu-Upemba KLC. For the specified time periods (y-axis: day of year, and x-axis: for the years 1999 to 2018) the scene with the lowest local cloud cover is selected.](image-url)
Discussion

Practicality of the quality assurance and assessment framework

Need for quality assessments

Thematic geographic databases/maps are used today in many sectors (forestry, agriculture, urban, environment, etc.), focusing on a variety of issues. Private consultancy firms, NGOs, universities and different institutions produce ever-increasing amounts of spatial and geospatial information but the degree of precision and reliability of this information is not always known (Qian et al., 2020). For example, the presented maps for the Kundelungu-Upemba KLC have been used to visualize and evaluate the threats and pressures on one of the biggest RAMSAR sites in the world (Bassin de la Lufira, https://risis.ramsar.org/ris/2318) and the progress to achieving the vision of the conservation of the last population of elephants in the Katanga province (DRC). Additionally, there is an ongoing hydropower project (Sombwe dam) in the area where the C-HSM 2016 map can be used as a base map to ensure that both aquatic and terrestrial biodiversity components are adequately addressed during and after the project.

This proliferation of spatial databases increases the urgency to link quality measures to these products in order to understand their accuracy for both data producers and for end users because otherwise it will be impossible to compare and develop an understanding of the underlying phenomena. Consequently, a map product must meet certain quality criteria and provide information for users to assess its accuracy together with the associated confidence interval. Moreover, map quality control must take into account the realities and goals of the mapping procedure including both technical and financial aspects. Reference data is crucial (Szantoi, Geller et al., 2020) and consists of already existing maps, fine spatial resolution EO images or field surveys, all of which can be costly. Field surveys are sometimes difficult or even impossible due to accessibility constraints or simply because of the mismatch between the in-situ field check and the data/image acquisition date. Another practical difficulty relates to the field controls of large extents – in the case of CAF11, tens of thousands of km² – that cover a diverse, sometimes complex and heterogeneous set of landscapes. The choice of the control

![Map legend](image)

**Figure 5.** Relative spatial accuracy of the classification polygons for the 2000–2016 Kundelungu-Upemba KLC land cover change map.
method had to take into account the time difference between the map production date (2018/2019) and the map reference dates (2000 and 2016) used to produce the LC and LCC maps. Finally, a significant constraint was to assess the quality of LCC estimated between 2000 and 2016, given that this phenomenon of change represents less than 2% of the territory.

Figure 6. Summary statistics of the Qualitative Systematic Accuracy results for the 2016 Kundelungu-Upemba KLC land cover map.

Figure 7. Example of a delineation issue between closed forest (77) and open forest (78), as identified on a Landsat 8 reference image. Yellow arrows -> wrong delineation, green arrows -> delineation missing. Modular LCCS classes: 77 – continuous closed trees, 78 – continuous open general trees, 178 – closed to open herbaceous vegetation.
The two types of controls (QSA and QAV) set up in this map validation chain were established to meet the specific requirements of the map owners. That said, it is impossible to imagine all the potential future uses of the thematic maps produced especially given the free and open license associated with so many community driven mapping projects. Therefore, it is imperative that map products are linked to transparent and open quality assessments because users need to know the quality of the information on which they are making decisions. The purpose of the QSA based on visual verification presented in the context of this framework is twofold: on one side it can be applied to be corrective (reproduction if it is necessary) and to help make improvements to the map production and interpretation procedures; on the other hand it could be used as a record of map quality so that users can understand where the map has potential quality issues. The QSA and results together paint a valuable picture that we hope will support all types of map users and help them understand not only the map’s thematic information but the quality of that information as well. We encourage other map owners and producers to follow a similar framework so that not only map information but also map quality can be compared.

Sampling design
The most delicate stage of quantitative quality control consists of sampling the LC classes to be controlled. For error statistics to be reliable, the sampling must be statistically representative of the entire area mapped, both in terms of the proportion of classes sought and in terms of classification error. We have implemented a stratified random sampling approach based on inclusion probabilities by using the maps of LC and LCC to define the strata. Specific attention is thereby to be paid to the sampling design in the event that the area of interest represents very small strata in combination with very large strata as analysed in detail by Olofsson et al. (2020). As for most operational LC and LCC monitoring applications, this is also the situation in the Copernicus High-Resolution Hot Spot Monitoring activity. The main problem thereby arises when omission errors are observed at sample locations in a very large stratum. Such omissions tend to carry very large area weights and may therefore result in large uncertainties and wide confidence intervals in the accuracy measures and area estimates for LC and LCC categories covering small areas only. To reduce this effect, a large number of additional sampling points were sampled based on Neyman optimal allocation. For the LCC, the sampling plan is specifically adapted because the sample points must fall on areas, which
represent less than 1% to 2% of the entire territory. For the stratification, the mapping results have been aggregated to those strata, which are required to derive the change categories listed above (Image and data integrity used for map production section). This sampling approach proved to be effective in reducing the uncertainty of accuracy measures and area estimates for all KLCs mapped by the C-HSM.

**Response design**
To allow an accurate interpretation of the complex LC and LCC transitions, in addition to very high-resolution imagery, full time series of high-resolution satellite imagery covering the monitoring periods are required. In order to take full advantage of very high-resolution imagery, as the full time series of freely available Sentinel-2 time series, the historic Landsat and SPOT archive imagery and of other ancillary data, we implemented a web-based interpretation tool, which allows the comprehensive data sets to be viewed and annotated in a web browser. Import and export functionality was implemented utilizing a PostgreSQL database with PostGIS extensions, which ensures scalability and security, and provides a multi-user manageable environment for a team of interpreters. The interpretation tool includes a trajectory viewer, which presents the full temporal trajectory of the bottom of the atmosphere reflectance imagery and all the sampling points to the interpreters (see Figure 10). Further, the tool provides a seasonal matrix viewer, which presents the images with the lowest cloud cover over 40-day intervals (Figure 4). The rows of the matrix provide
Figure 10. Trajectory view for visualising the complete time-series of bottom of atmosphere reflectances (mean value of 3 by 3 pixels), and corresponding images in a 2 km by 2 km neighborhood.

Table 8. User’s and producer’s overall land cover map accuracy for the blind and plausibility assessments with 95% confidence intervals for upper and lower bounds. Accuracy values presented in percent (OA – Overall accuracy, PA – Producer’s accuracy, UA – User’s accuracy, LB – lower bound, UB – upper bound). Classes with less than 20 reference samples are omitted.

| LC Class      | Blind Assessment (OA = 91.81 ± 0.99) | Plausibility Assessment (OA = 95.27 ± 0.77) |
|---------------|-------------------------------------|---------------------------------------------|
|               | PA       | LB       | UB       | UA       | LB       | UB       | PA       | LB       | UB       | UA       | LB       | UB       |
| 11            | 96.8     | 89.7     | 100      | 93.3     | 84.8     | 100      | 100      | 100      | 100      | 100      | 100      | 100      |
| 32            | 100      | 100      | 100      | 99.2     | 96.8     | 100      | 100      | 100      | 100      | 100      | 100      | 100      |
| 59            | 92.5     | 77.1     | 100      | 94.3     | 94.2     | 91.4     | 96.7     | 94.6     | 92.2     | 96.8     | 95.2     | 93.3     |
| 60            | 89.6     | 84.3     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 92.4     | 90.6     | 94.2     | 97.1     | 95.5     |
| 77            | 91.6     | 88.7     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 96.8     | 94.8     | 98.7     | 86.9     | 93.6     |
| 78            | 89.6     | 87.4     | 100      | 95.6     | 94.3     | 92.3     | 96.3     | 97.7     | 95.7     | 99.2     | 94.3     | 91.6     |
| 112           | 92.9     | 89.9     | 100      | 95.6     | 94.3     | 92.3     | 96.3     | 98.5     | 97.1     | 99.7     | 97.2     | 95.4     |
| 116           | 92.2     | 88.5     | 100      | 95.6     | 94.3     | 92.3     | 96.3     | 98.5     | 97.1     | 99.7     | 97.8     | 95.8     |
| 148           | 96.4     | 94.4     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 97.9     | 96.2     | 99.3     | 98.6     | 96.6     |
| 165           | 79       | 51.6     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 79.3     | 48.3     | 100      | 96.2     | 91.7     |
| 166           | 89.8     | 73.2     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 96.8     | 88.6     | 100      | 99.2     | 97.3     |
| 171           | 67.3     | 50.5     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 75       | 58.2     | 100      | 99.2     | 97.3     |
| 175           | 55.9     | 31.7     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 57.8     | 33.9     | 100      | 99.2     | 97.3     |
| 178           | 95.8     | 93.7     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 97.9     | 96.2     | 100      | 99.2     | 97.3     |
| 182           | 90.3     | 75       | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 95.1     | 85       | 100      | 94.9     | 83.3     |
| 184           | 77.2     | 54.4     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 87.9     | 68.7     | 100      | 98.1     | 94.9     |
| 185           | 100      | 100      | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 100      | 100      | 100      | 98.1     | 94.9     |
| 190           | 74       | 51.5     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 77.8     | 59.3     | 100      | 94.3     | 92.3     |
| 191           | 98.3     | 96.8     | 100      | 95.3     | 94.3     | 92.3     | 96.3     | 99.8     | 96.2     | 100      | 99.2     | 97.3     |

Table 9. User’s and producer’s accuracy of land cover change categories for the blind and plausibility assessments with 95% confidence intervals for upper and lower bounds. Accuracy values presented in percent (PA – Producer’s accuracy, UA – User’s accuracy, LB – lower bound, UB – upper bound). Classes with less than 20 reference samples are omitted.

| LC Class                                | Blind Assessment | Plausibility Assessment |
|-----------------------------------------|------------------|-------------------------|
| Artificial surface expansion            | PA       | 79.2       | 64.3       | 94.3       | 94.1       | 89.4       | 98.4       | 99.8       | 99.8       |
| Cropland expansion                      | PA       | 77.6       | 68.6       | 85.7       | 93.7       | 90.2       | 97         | 96.4       | 100        |
| Forest cover gain                       | PA       | 54.4       | 30.1       | 87.5       | 82.3       | 67.2       | 91.8       | 77.3       | 96         |
| Forest cover loss                       | PA       | 70.9       | 59.3       | 83.9       | 91.8       | 85.3       | 97.5       | 97.7       | 100        |
| Gain of natural vegetation              | PA       | 81.8       | 64.2       | 98         | 79.6       | 66.6       | 90.3       | 90.7       | 98.8       |
| Loss of natural vegetation              | PA       | 77.8       | 68.9       | 86.4       | 93.7       | 89.9       | 97         | 96.6       | 99.9       |

Information on the seasonal changes (vegetation phenology) and the columns provide information on the yearly developments over the monitoring period. These time-series visualizations allow efficient and accurate interpretation also for difficult LCC processes such as those related to forest degradation.

Blind and plausibility procedures. The interpretation of samples within the first control is performed by an interpreter who is not able to access information of the designated map class (hence “blind”). Subsequently, all samples which show a deviation between the blind interpretation and the map class are re-interpreted in
Table 10. Area estimation results and confidence intervals at the 95% level in ha for the 2016 land cover. Classes with less than 20 reference samples are omitted.

| LC Class | Mapped area | Estimated area | Bootstrapping |
|----------|-------------|----------------|---------------|
| 11       | 99          | 88             | 85            | 92            |
| 32       | 77          | 77             | 78            | 80            |
| 59       | 3483        | 3552           | 2393          | 5286          |
| 60       | 241,616     | 262,373        | 247,806       | 278,931       |
| 77       | 805,328     | 786,738        | 851,841       | 900,543       |
| 78       | 1,246,188   | 1,384,234      | 1,348,989     | 1,421,984     |
| 112      | 782,100     | 579,686        | 599,496       | 599,496       |
| 116      | 441,233     | 392,814        | 376,574       | 408,020       |
| 148      | 532,647     | 511,015        | 499,367       | 522,069       |
| 165      | 10,302      | 14,639         | 8653          | 22,591        |
| 166      | 1959        | 5186           | 2869          | 9749          |
| 171      | 21,505      | 27,531         | 20,637        | 35,485        |
| 175      | 7604        | 17,192         | 9511          | 27,567        |
| 178      | 445,747     | 457,345        | 433,462       | 481,219       |
| 182      | 165         | 178            | 175           | 181           |
| 184      | 15,095      | 17,959         | 13,581        | 25,423        |
| 185      | 451         | 453            | 326           | 716           |
| 190      | 9921        | 16,118         | 10,723        | 23,989        |
| 191      | 165,223     | 162,342        | 153,268       | 170,629       |

Table 11. Area estimation results and confidence intervals at the 95% level in ha for land cover changes from 2000 to 2016. Classes with less than 20 reference samples are omitted.

| LCC Class                  | Mapped area | Estimated area | Bootstrapping |
|----------------------------|-------------|----------------|---------------|
| Artificial surface expansion | 2983        | 3825           | 3210          | 4533          |
| Cropland expansion         | 95,431      | 141,544        | 126,952       | 157,108       |
| Forest cover gain          | 12,823      | 13,581         | 7476          | 20,663        |
| Forest cover loss          | 66,399      | 93,963         | 79,958        | 109,185       |
| Gain of natural vegetation | 25,821      | 22,771         | 17,955        | 28,878        |
| Loss of natural vegetation | 98,197      | 145,075        | 128,336       | 163,370       |

a second step (plausibility). For example, in the case of transitional zones between natural and managed areas, both class assignments can be considered as plausible. Furthermore, potential geometric shifts are taken into account, such as if a sample point is located close to the border of a lake, it can be considered as mapped plausible, even if the VHR validation data show that the plot is located one pixel (i.e. 30 m or 10 m) outside of the lake because of a possible geometric shift of Landsat or Sentinel imagery. Thus, the analysis of plausibility is an important step because, in general, it provides flexibility in the interpretation of certain LC classes after the blind assessment. The class assessment related to the cover density, the height of vegetation, is difficult to evaluate when objects are smaller than the spatial resolution of the images (i.e. 10 m or 30 m) and involves the usage of higher resolution data. As can be seen from the accuracy results, an evaluation based on validation datasets revised through the plausibility process led to higher accuracies compared to the “blindly” collected dataset-based ones. To avoid systematic interpretation errors, we recommend the involvement of several experts for the generation of interpretation keys, for performing the actual interpretation and for an independent quality control by comparing an independently interpreted sub-sample.

**Statistical analysis**

For deriving confidence intervals for accuracy measures as well as for area estimation, we recommend applying bootstrapping instead of applying analytical equations, as bootstrapping can handle correlations, large uncertainties and complex probability density functions, which are not based on normality assumption. This is of specific importance in the event that quantitative information on LC and LCC is used for international reporting as, for example, within the framework of the reducing emissions from deforestation and forest degradation (REDD https://redd.unfccc.int/) programme, where unbiased area estimates and confidence intervals are required to determine the most reliable minimum estimate of changes in carbon stocks over time.

**Conclusions**

The QAAF presented has been applied as part of the Copernicus thematic High-Resolution Hot Spot Monitoring activity in order to strengthen the reliability of thematic map products produced for developing regions (e.g., sub-Saharan Africa), and to build trust in the applications where such map products are used, such as policy-making. Both are vital characteristics for users. The approach presented ensures transferability, as the methodology is applicable to any location in the world and to local, regional or global extents. Furthermore, these quality control procedures are reliable, robust and suitable for other biogeographical regions, applicable to future updates of the current KLCs and can be evaluated and analysed together due to their unique quality of using the same classification system as well as quality evaluation methodology. It is expected that regular updates of the LC thematic layers will be produced based on the long-term acquisition of Copernicus Sentinel-2 images, together with other complementary geospatial data in order to continue the assessment of anthropogenic pressures on natural environments around the world.

Reliable and accurate LC and LCC maps are required to support sustainable decision-making and to assist in the monitoring of environmental policies. The QAAF demonstrates that it can support local, regional, national and international governments and authorities in applying transparent and open geospatial information. The QAAF was developed to manage a variety of thematic map products and therefore LC and LCC layers validated through the framework can be easily compared, as well as information shared because the user is able to discern the issues between the products and understand where errors can
influence decisions and recommendations. Mapping technology continues to evolve and it is imperative that the maps of today remain usable to help solve the problems of tomorrow.

It is impossible to envisage all the applications of our thematic maps at the time of production and the QAAF therefore provides the details necessary for users outside the domain of KLCs to not only understand the thematic layers produced, but also assess their applicability in other domains. Thus, this type of quality assurance guarantees an ensured quality of the cartographic maps essential to end users by following IPCC recommendations.

Thematic map production and validation are therefore two successive but inseparable steps for producing reliable geospatial products with accompanying statistics. Moreover, validation data sharing through data archives, especially those supported by public funds, should be a routine procedure, and it would further increase trust in map products as well as the usage of such invaluable datasets (Szanto, Geller et al., 2020). Future work on the QAAF and the required methods will focus on increased automation and distributed teams in order to make the map validation process more fluid and optimised. We encourage that other thematic LC mapping initiatives will adopt the QAAF in order to help in improving the sharing and usability of LC maps to tackle current issues at hand.

Acknowledgments

This study was performed within the framework of the Copernicus Global Land Hot Spot Monitoring Service, which is managed by the European Commission’s DG Joint Research Centre (JRC). We would like to thank the JRC Team and all partners involved: IGN FI, Joanneum Research, Space4Environment, EXOPIRE, GiSbox, ONFI and LuxSpace. Also, we would like to thank the reviewers for their time spent reviewing our manuscript and their comments, helping us to improve the technical quality and presentation of our article.

Data availability

The corresponding LC and LCC maps and the validation datasets are freely available from the dedicated Copernicus website: https://land.copernicus.eu/global/hsm, as well as from an open data repository: https://doi.org/10.1594/PANGAEA.920847. Additionally, other KLCs and accompanying validation datasets (12 areas) are available here: https://doi.pangaea.de/10.1594/PANGAEA.914261 and here (10) https://doi.org/10.1594/PANGAEA.931968.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Ayala-Izurieta, J. E., Márquez, C. O., García, V. J., Recalde-Moreno, C. G., Rodríguez-Llerena, M. V., & Damián-Carrón, D. A. (2017). Land Cover Classification in an Ecuadorian Mountain Geosystem Using a Random Forest Classifier, Spectral Vegetation Indices, and Ancillary Geographic Data. Geosciences, 7 (2), 34. https://doi.org/10.3390/geosciences7020034

Burnicki, A. C. (2011). Modeling the Probability of Misclassification in a Map of Land Cover Change. Photogrammetric Engineering & Remote Sensing, 77(1), 39–49. https://doi.org/10.14358/PERS.77.1.39

Cochrane, W. G. (1977). Sampling Techniques (3rd ed.). Wiley.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37(1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Di Gregorio, A. (2005). Land cover classification system: Classification concepts and user manual: LCCS (Software version 2). Food and Agriculture Organization of the United Nations.

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) | INSPIRE. (2007). https://inspire.ec.europa.eu/documents/directive-20072ec-european-parliament-and-council-14-march-2007-establishing.

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science, 1(1), 54–75. https://doi.org/10.1214/ss/1177033815

Feranec, J., Soukup, T., Hazeu, G., & Jaffrain, G. (Eds.). (2016). European Landscape Dynamics. CRC Press. https://doi.org/10.1201/9781315372860

Foody, G. M. (2002). Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80 (1), 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Foody, G. M. (2015). Valuing map validation: The need for rigorous land cover map accuracy assessment in economic valuations of ecosystem services. Ecological Economics, 111, 23–28. https://doi.org/10.1016/j.ecolecon.2015.01.003

Gallau, H., Steinegger, M., Wack, R., Schardt, M., Kornberger, B., & Schmitt, U. (2015). Remote Sensing Based Two-Stage Sampling for Accuracy Assessment and Area Estimation of Land Cover Changes. Remote Sensing, 7(9), 11992–12008. https://doi.org/10.3390/rs70911992

Gao, Y., Liu, L., Zhang, X., Chen, X., Mi, J., & Xie, S. (2020). Consistency Analysis and Accuracy Assessment of Three Global 30-m Land-Cover Products over the European Union using the LUCAS Dataset. Remote Sensing, 12 (21), 3479. https://doi.org/10.3390/rs12213479

Gong, P., & Mu, L. (2000). Error Detection through Consistency Checking. Annals of GIS, 6(2), 188–193. https://doi.org/10.1080/1082400009480549

Gräfus, D. R., Corstanje, R., Warren, P. H., Evans, K. L., Hancock, S., & Harris, J. A. (2016). The impact of land use/land cover scale on modelling urban ecosystem services. Landscape Ecology, 31(7), 1509–1522. https://doi.org/10.1007/s10980-015-0337-7

IPCC. (2006). 2006 IPCC guidelines for national greenhouse gas inventories. Institute for Global Environmental Strategies. http://www.ipcc-nggip.iges.or.jp/public/2006gl/index.htm
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(19), 2305. https://doi.org/10.3390/rs11192305

Olofsson, P., Arévalo, P., Espejo, A. B., Green, C., Lindquist, E., McRoberts, R. E., & Sanz, M. J. (2020). Mitigating the effects of omission errors on area and area change estimates. Remote Sensing of Environment, 236, 111492. https://doi.org/10.1016/j.rse.2019.111492

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sensing of Environment, 129, 122–131. https://doi.org/10.1016/j.rse.2012.10.031

Pal, M. (2002). Factors influencing the accuracy of remote sensing classifications: A comparative study. PhD Thesis, University of Nottingham. http://eprints.nottingham.ac.uk/10314/

Pengra, B. W. (2020). Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. Remote Sensing of Environment, 10. https://doi.org/10/ggfq5k

Qian, T., Kinoshita, T., Fujii, M., & Bao, Y. (2020). Analyzing the Uncertainty of Degree Confluence Project for Validating Global Land-Cover Maps Using Reference Data-Based Classification Schemes. Remote Sensing, 12 (16), 2589. https://doi.org/10.3390/rs12162589

Schreuder, H., Ernst, R., & Ramirez-Maldonado, H. (2004). Statistical techniques for sampling and monitoring natural resources (Gen. Tech. Rep. RMRS-GTR-126; p. 111). U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Selkowitz, D. J., & Stehman, S. V. (2011). Thematic accuracy of the National Land Cover Database (NLCD) 2001 land cover for Alaska. Remote Sensing of Environment, 115(6), 1401–1407. https://doi.org/10.1016/j.rse.2011.01.020

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment. Statistical Rigor and Practical Utility in Thematic Map Accuracy Assessment, 67(6), 727–734. https://www.asprs.org/wp-content/uploads/pers/2001journal/june/2001_jun_727-734.pdf

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing, 30(20), 5243–5272. https://doi.org/10.1080/01431160903131000

Stehman, S. V. (2012). Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. Remote Sensing Letters, 3(2), 111–120. https://doi.org/10.1080/01431161.2010.541950

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

Strobl, P., Baumann, P., Lewis, A., Szantoi, Z., Killough, B., Purss, M. B. I., Craglia, M., Nativi, S., Held, A., & Dhu, T. (2017). The six faces of the data cube. Proc. of the 2017 Conference on Big Data from Space (BiDS’17), Toulouse, France: Publications Office of the European Union. 32–35. https://doi.org/10.2760/383579

Szantoi, Z., Brink, A., & Lupi, A. (2021). An update and beyond: key landscapes for conservation land cover and change monitoring, thematic and validation datasets for the African, Caribbean and Pacific regions. Earth System Science Data, 13(8), 3767–3789. https://doi.org/10.5194/essd-2021-98

Szantoi, Z., Brink, A., Lupi, A., Mammonne, C., & Jaffrain, G. (2020). Key landscapes for conservation land cover and change monitoring, thematic and validation datasets for sub-Saharan Africa. Earth System Science Data, 12(4), 3001–3019. https://doi.org/10.5194/essd-12-3001-2020

Szantoi, Z., Geller, G. N., Tsendbazar, N.-E., See, L., Griffiths, P., Fritz, S., Gong, P., Herold, M., Mora, B., & Obregon, A. (2020). Addressing the need for improved land cover map products for policy support. Environmental Science & Policy, 112, 28–35. https://doi.org/10.1016/j.envsci.2020.04.005

Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A. K., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L. P., Leimgruber, P., Taubenboeck, H., Wegmann, M., Wikelski, M., & Woodcock, C. (2015). Free and open-access satellite data are key to biodiversity conservation. Biological Conservation, 182, 173–176. https://doi.org/10.1016/j.biocon.2014.11.048

Wehmann, A., & Liu, D. (2015). A spatial–temporal contextual Markovian kernel method for multi-temporal land cover mapping. ISPRS Journal of Photogrammetry and Remote Sensing, 107, 77–89. https://doi.org/10.1016/j.isprsjprs.2015.04.009

Wulder, M. A., Coops, N. C., Roy, D. P., White, J. C., & Hermosilla, T. (2018). Land cover 2.0. International Journal of Remote Sensing, 39(12), 4254–4284. https://doi.org/10.1080/01431161.2018.1452075