European
Commission

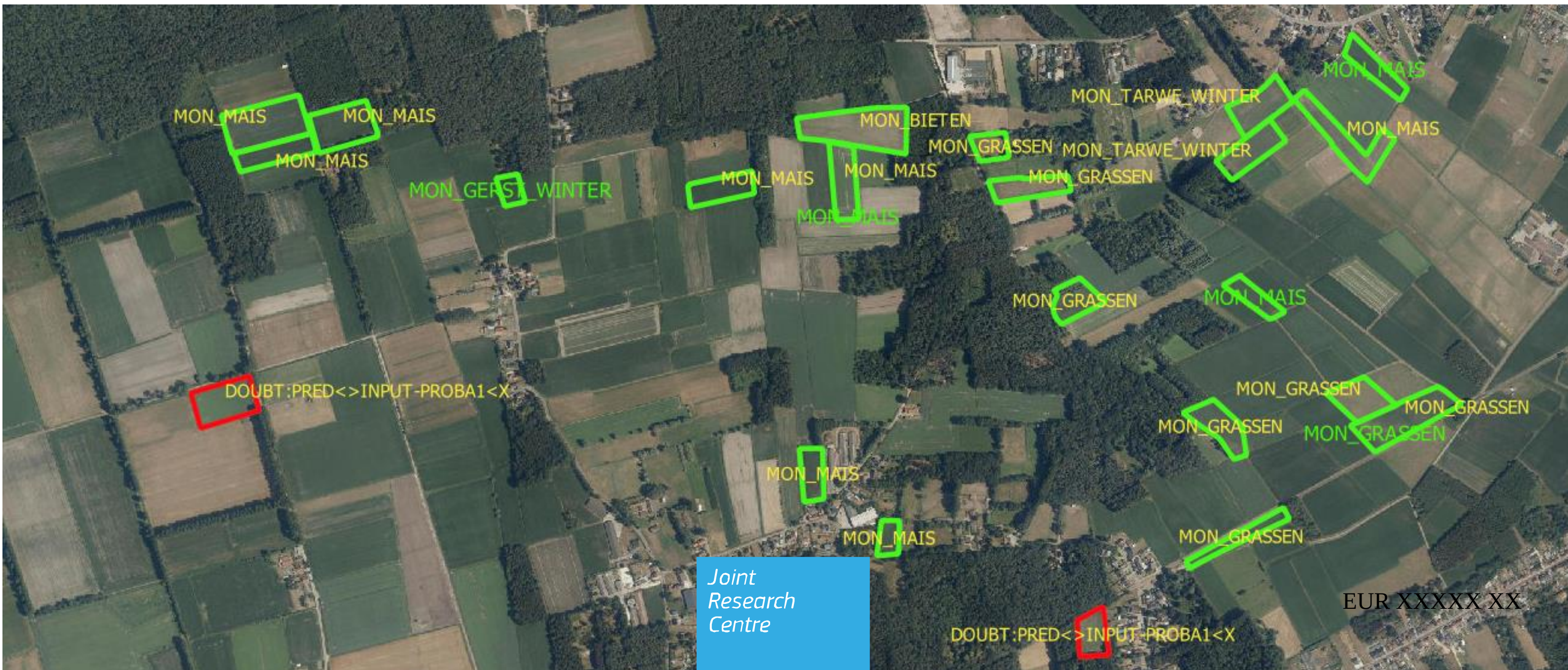
JRC TECHNICAL REPORT

Checks by monitoring Quality Assurance

Technical guidance v.1.2

June 2021

Nataša Luketić, Daniele Borio,
Wim Devos, Pavel Milenov



Joint
Research
Centre

EUR XXXXX XX

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information
Name: Wim Devos
Address: Via E. Fermi 2749, I-21027 Ispra/Italy
Email: wim.devos.@ec.europa.eu
Tel.: +39 0332-789655

EU Science Hub
https://ec.europa.eu/jrc

JRCXXXXXX

EUR XXXXX XX

PDF	ISBN XXX-XX-XX-XXXXX-X	ISSN XXXX-XXXX	doi:XX.XXXX/XXXXXX
Print	ISBN XXX-XX-XX-XXXXX-X	ISSN XXXX-XXXX	doi:XX.XXXX/XXXXXX

Ispra : European Commission, 2021

Contents

Foreword

The Joint Research Centre (JRC) is the European Commission's science and knowledge service, that supports EU policies with independent scientific evidence throughout the whole policy cycle. The work on preparing the technical guidance is normally carried out through collaborative meetings and discussion with DG Agri and Member State administrations.

This draft technical guidance serves to help CbM adopters set up and evaluate a quality assessment process, to gain and share experiences, and to point out weaknesses and possible improvements.

Highlighted elements indicate sections that are still provisional, missing or incomplete.

The feedback of these activities will be analysed and considered for subsequent versions of the Technical Guidance (TG).

The procedures described in this TG are expected to evolve taking into account annual results and feedback from the various stakeholders.

This is the fifth draft version of the same document produced in June 2020, May 2020, August 2020 and December 2020. All remarks have been collected and processed in this version (in blue).

The main changes compared to the previous version are as follows:

- Chapter 5.6 has been added - Inspection for the systems applying machine learning classification,
- Chapter 6.3 has been added – testing of machine learning classification,
- A new table 3 was inserted: former table numbers 3 and higher are recounted to 4 and higher,
- Chapter 6.2.2 Step 2 eligibility testing is modified to accommodate the reporting per scheme where the CbM implementation covers multiple schemes,
- Chapter 7 has been introduced on the calculation of financial impact.

Acknowledgements

The authors wish to acknowledge the paying agencies of Belgium-Flanders, Denmark, Spain, Italy and Malta for providing data and input. Also a special thanks to our former colleague Dominique Fasbender for providing knowledge on statistical input after he left our Unit.

Authors

Nataša Luketić, Daniele Borio, Wim Devos, Pavel Milenov

Abstract

Checks by Monitoring (CbM), in the current compliance-based CAP, use monitoring results to check the eligibility of parcels in area related payment schemes as defined in the Article 40a of Commission Implementing Regulation (EU) No 809/2014. The process itself consists of three different elements, the automatic detection/extraction of the information based on Sentinel data, eligibility checks and payment impact. In order to assure the effectiveness of the CbM process, a Quality Assurance (QA) procedure has been developed. This is the focus of this Technical Guidance (TG).

Methods and tools for verification of the automatic outputs obtained through the analysis of Sentinel or equivalent data are at first provided. This is step 1 of the CbM QA framework and the method is based on testing a random sample of paired lots. After determining the sampling plan, the quality of the Sentinel-based process is acceptable if the number of type I and type II errors (nonconformities) is less than the acceptance numbers (AC_1 and AC_0). Individual and complete test is done per type of information extraction.

Step 2 of the QA method is based on testing the random samples selected in step 1 for errors defined according to eligibility criteria. In particular, end-stage errors are defined as those that lead to an undue payment with no further intervention in the CbM process. Abatable errors are defined as those that prevent the payment of legitimate subsidies and can trigger reaction from the affected farmer. In this way, the Paying Agency (PA) will have to start a secondary procedure requiring human intervention. Both types of errors lead to financial losses and their number have to be kept within predefined quality limits. This TG specifies tests ensuring that the two types of errors are within the specified quality limits.

In step 3, the financial impact of end-stage errors found in the previous step is evaluated per scheme. This step takes into account only monitorable elements which was processed in the frame of the Sentinel based CbM. In such way a targeted evaluation of the CbM system can be reported as a single residual error expressed in EUR.

After inspection, CbM QA assessment report and remedial action plan should be produced and reported to the Commission.

1 Introduction

Checks by Monitoring (CbM) currently substitute, on an optional basis, the On-The-Spot Checks (OTSC) for direct payments totalling over €40B for 2019. These checks target the correctness of the payments to the farmers and cooperate with two additional components of the Integrated Control Administration Systems (ICASs): the Land Parcel Identification System (LPIS) and the Geo-Spatial Aid Application (GSAA). The additional systems developed are primarily to prevent incorrect aid applications and hence reduce the number of non-compliances detected during the OTSC.

The CbM option was introduced to take advantage of new technologies, in particular Sentinel 1 and Sentinel 2, towards reducing costs (higher cost-efficiency), reducing burden (by automation and reducing field visits), increasing fairness (by avoiding sampling) and increasing awareness and compliance (by a prevention mechanism).

The pre-conditions for monitoring are: effective system of GSAA, of administrative crosschecks and of retroactive recovery of undue payments combined with a good quality LPIS.

The CbM Quality Assurance (QA) is a framework intended to enable the Member State (MS) to report to the Commission about the state of one of the components inside the control and management system. The common methodology ensures an objective and comparable reporting from all MSs. For this purpose, templates and procedures are provided further in the document. There will be a set of documents, apart from this one, providing the full picture of the quality assurance framework.

The CbM process itself consists of three steps: the automatic detection/extraction of the information based on Sentinel or equivalent data, eligibility checks and payment/financial impact.

The QA procedure covers all three mentioned steps of CbM. In addition to these steps, the QA process deals with boundary conditions as validity tests. A first boundary, denoted as P1, deals with area management and aims at verifying the correctness of the location and area of the Feature Of Interest (FOI) as declared by the farmer. As the QA process directly depends on the good LPIS/GSAA quality, some elements of CbM testing are designed to assess the consistency/cardinality between the process-derived decisions and the reality in the field: this is the role of boundary P1.

Other boundary conditions, such as those on the discriminatory power of the Sentinel-based automatic extraction process are outside the scope of this Technical Guidance (TG). In particular, there are cases where the Sentinel-based automatic extraction process is unable to take a decision with the information provided by the end of the season (non-monitorable aspects) or cases where FOIs are intentionally non-monitored (non-monitored cases). These cases that are not targeted by the CbM flow will end up in the "pool of non-conclusive parcels" and will be subject to non-Sentinel follow-up including sampling. A dedicated test, called "P2", checks whether inclusion into that pool was appropriate. For practical reasons, the P2 test will not be required for the 2019 and 2020 campaigns.

In the first step of the QA process, a procedure is established to analyse the detection performance of Sentinel-based CbM. CbM is used to detect the presence of a specific phenomenon in the field and farmer's activity. The decision within the process is performed on a FOI and it concerns a specific marker/behaviour on a land parcel representing that FOI. For example, CbM can be used to detect ploughing. Two tests based on the ISO 2859/2 standard are developed in order to respond to the following questions:

- Is CbM detection able to inspect a full population (of size N) with a number of type I errors lower than Lq·N? Where Lq is the quality limit set to 10%.
- Is CbM detection able to inspect the full population (of size N) with a number of type II errors lower than Lq·N? In this case, the quality limit was set to 10%, as well.

Type I and type II errors are defined as false positive and false negative events, respectively. A false positive (type I) error occurs when a marker/behaviour is erroneously detected as present. Conversely, a false negative (type II) error occurs when a marker/behaviour is erroneously detected as not present.

Details on the tests developed and on the decision thresholds are provided in the following chapters.

While the tests developed in step 1 allow one to assess the quality of the Sentinel-based detection process, they do not consider eligibility conditions or the financial impact of these errors. The output of step 1 is thus related to eligibility conditions and tests to assess CbM performance in terms of eligibility errors are developed. This is the focus of CbM QA step 2. The document also combines the eligibility errors of step 2 with the payment rates and holding rules, allowing an assessment of the financial performance of the CbM operations. This propagation is the focus of CbM QA step 3.

In the framework of CbM QA step 2, this TG introduces the concepts of abatable/end-stage errors based on eligibility conditions. Tests, based on the sample, identified and inspected for step 1, are introduced and detailed along with processing examples.

Similarly, in the step 3, the end-stage errors per scheme will be a representative scope to calculate a residual risk in terms of the financial aspect. The area declared found to be erroneously detected will be converted into the payment amount and evaluated accordingly.

The outcome of QA framework should involve an honest self-assessment of the functioning of the system itself. This reflection involves investigating weaknesses both of the QA method and of the monitoring system itself. It will allow one to optimize operational processes or identify appropriate remedial actions.

2 Scope of the document

This document provides a general methodology to inspect the quality of the CbM process by Sentinel usage. It provides guidance on sampling principles, testing methods and acceptance criteria.

This methodology is applicable for the three steps of the CbM QA procedure to be implemented by the MS.

In particular,

- step 1 consists of validating the setup of the automatic algorithms for detecting specific phenomena by interpreting the Sentinel data (or equivalent).
- step 2 relates the output from the first step (Sentinel detection errors) and from other inspection data to eligibility cases. Inspection errors are linked to the corresponding agricultural parcels that quantify a distinct area for a particular scheme, allowing an assessment of the eligibility errors.
- step 3 combines the eligibility errors of the second part with the payment rates and holding rules, allowing an assessment of the financial performance of the CbM operations.

3 CbM population

CbM cover all aid applications within the given territory. However, a territory is rarely homogeneous regarding landscape, farming system, cultivated crops and management practices.

Quality inspection, e.g. as implemented in the LPIS QA, relies on an independent and external observation to verify the correctness of an item to be inspected. Conceptualization of the item and the external observation are therefore the first necessary steps in any quality inspection methodology.

3.1 Items for inspection

Although the outcome of any CbM process is to assess aid applications through the agricultural parcels therein, neither the application nor the parcels are the elementary component of that process. In fact, CbM is exclusively driven by elementary pieces of information that are extracted from a stack of Sentinel data.

The inspection addresses what was observed and how that observation is used on a given phenomenon represented by the FOI, (during a given application season). This represents the elementary **decision** of the CbM system. **A processing decision constitutes, therefore, the item for inspection** of the Sentinel component of the QA.

An item can thus be defined as an individual **processing decision** based on Sentinel observation of field conditions, that impacts an explicit conclusion.

In a given season, the processing of a FOI results either in one or more isolated decisions, depending on how many aspects have been processed and monitored for a single FOI. For example: was an arable crop observed? Y/N, was the ploughing date too early? Y/N, is the whole parcel covered? Y/N. In this theoretical example, there are three processing decisions and hence 3 items for inspection.

Note that all automated decisions in the process are ultimately expressed as a binary decision: i.e. some relevant phenomenon was either detected or not detected at a given moment.

The population is the set of all processing decision items subject to possible inspection. If a system contained only a single process thread with the 3 decisions above, the size of the population would be 3 times the number FOIs. In practice, different types of FOIs will be subject to different processing observations/decisions, depending on their land cover type and the expected scenario.

While the item of inspection is the combination of extraction and application (observation and decision), its observation component can be a composite of several distinct, individual observations, which alone do not constitute an item. E.g. hypothetical question "was there an arable crop? Y/N" could depend on observing both a ploughing event early in the season and a harvest event later on. Observing ploughing alone would be inconclusive (e.g. re-sown permanent grassland is also ploughed) and no decision can be made. To lead to a decision, both events would have to be detected in sequence and it is this sequence, not the individual event, with its subsequent decision that constitutes an item. Understanding the business logic of the CbM put in place (as described in the scenario) is therefore key.

As a result, the population of all items should be a collection of processing decisions whereby decisions should have a reference to the data set that led to this particular decision. Such data set will be re-used in a form of pre-defined tools for inspection process.

In the CbM QA discussion document (Devos, December 2020), the scope of the CbM QA is clearly defined. Non-monitored parcels and non-monitorable aspects do not contribute to the CbM population and do not lead to items subject to inspections. In addition to decision items, CbM QA has to monitor implied validity conditions. This is the case of the P1 test that is specifically considered in Section 6.4.

3.2 Inspection lot(s)

There are seven types of information extractions defined, which presumably encompass all possible spatio-temporal phenomena relevant in CbM context. Each of them applies particular techniques to observe given spatio-temporal aspects/behaviour of the real world phenomena and reflects the complexity of the associated extraction/observation process. The spatio-temporal phenomena and the associated types of information extraction are developed in depth in Annex III.

Such pre-defined typology of information extraction, related to a particular decision, allows a very efficient detection (and hence inspection) setup, limiting the necessary number and targeting the correct dates of images from the Sentinel stack. All observation/decision combinations in a CbM system can be assigned to such type.

Given that decisions based on particular type of information extraction are supposed to share some **common characteristics**, we assume that all related items exhibit a degree of homogeneity that would allow them to be considered as a single lot for quality inspection, regardless of an item’s position in the CbM process flow. Each lot then has to be individually sampled.

As a result, for the initial inspection, all items of a given information extraction type can be pooled together in a single lot. Thus, it follows that in a given system, there can initially be maximum 7 lots. To date, many CbM operate fewer types of information extraction to decide on eligibility, so will start with fewer lots.

If one of these initial lots would not meet the quality expectations, the obvious immediate consequence would be to discard the assumption of homogeneity and create individual sub-lots based on the particular position in the dataflow. This would allow one to identify which sub-procedures work and which don’t and allow to address the sub-procedures with issues.

Obviously, if the homogeneity cannot be assumed from the start, separate homogenous lots must be constructed. E.g. if it is already known that a particular sub-process with scarce items is rather ineffective, then those ineffective sub-processes should not be hidden in a remainder of effective sub-processes by merging all into one single lot. Such manipulation for intentional dilution is antithetical to the idea of QA. In addition, detecting which of the sub-processes are ineffective will assist in improving the overall performance of the CbM because it will allow focussing efforts to enhance the process exactly where needed and where most effectiveness gain can be achieved.

4 Sampling principle

4.1 General concept

A small sample size and an efficient inspection methodology guarantee low inspection costs. The remaining challenge is to make the inspection of that small sample effective.

Effectiveness of the inspection can be assured by avoiding bias. This is traditionally offered by random sampling. As indicated in Chapter 3, universal availability of Sentinel time stacks makes them an ideal external data source. Bias can completely be suppressed by a central sampling procedure.

The LPIS QA framework adheres to a similar approach by central sampling of reference parcels from their populations. The principle and the necessary data can therefore be directly applied in the CbM QA sampling, without any need for additional data deliveries by the MS.

While selection of the sampled parcels for the LPIS QA is driven by the random acquisition within the VHR satellite imagery zone for a particular year, sampling for CbM QA can be fully random over the entire area of the system.

For Basic Payment Scheme (BPS), the LpisPointZeroState (coordinates) will be randomly selected to produce point locations Ni(yi,xi) in a so-called pre-selection list. That list will then be sequentially processed until the necessary number of items have been inspected.

The sampling will occur for each lot, i.e. type of item or disambiguation thereof.

For Voluntary Coupled Support (VCS) or any other scheme under CbM, the CbM authority will have to provide a dedicated CbM population and deliver it to COM for sample generation. This can be the set of agricultural parcels requesting aid under the scheme for the inspection year.

To assure feasibility of the inspection, the so-called pre-selection list of coordinates will be at least 3 times bigger than the sample needed for inspection. COM will analyse the results of the QA and verify if additional criteria would be needed to deal with sub-parcelling issues.

4.2 Paired observations in a “paired lot”

Any observation/decision in the CbM process inevitably deals with a binary outcome (Y/N) based on detection or non-detection of a physical phenomenon. However, both detection and non-detection outcomes are in turn subject to a binary reality: the phenomenon occurred or did not occur. This makes sure that each item belongs to one and only one of four outcomes:

- it occurred and it was detected (the true positives)
- it occurred and it remained undetected (the false negatives) a.k.a. beta or type II errors
- it did not occur but it was nevertheless detected (the false positives) a.k.a. alpha or type I errors
- it did not occur and nothing was detected (the true negatives)

Table 1. Confusion matrix in a paired lot

Paired lot	**CbM values**	

		CbM detected	**CbM not detected**	sub-population sizes
Actual values	**True presence**	true positive N_{11} GREEN	false negative N_{01} or β False GREEN	N_1
	True absence	false positive N_{10} or α False RED	true negative N_{00} RED	N_0
sum		$N_{.1}$	$N_{.0}$	N

Note that the outcome and the abundance of the phenomenon are of no relevance. Observing an activity as ploughing to confirm arable activities is good and expected to occur on all arable lands, however observing that very ploughing under a ban for ESPG is bad and expected to be very rare. This evidences that the inspection is truly about paired observations and no easy shortcut is possible.

So the N elements of the lot (see Table 1) are essentially partitioned in two mutually exclusive populations of FOIs: those where the phenomenon occurred (true positives and false negatives) counting N_1 FOIs, and those where it did not occur (false positives and true negatives) counting N_0 FOIs. The sum of N_1 and N_0 is obviously N, but their exact values are unknown; they cannot be directly calculated from the CbM results but an unbiased estimate will be produced by the inspection itself.

As a result, each lot of items for inspection in step 1 can be considered as being composed of two complementary sub-populations, the FOIs where the phenomenon occurred and the FOIs where did not.

If a processing line provides the possibility to output an inconclusive option (phenomenon's observation is not considered reliable), and will thus be followed up by another data sources of check than Sentinel (field visits, farmer's inputs, etc.), such decisions are not in scope for this Sentinel based QA procedure.

4.3 A single sample size for the paired lot

Given the large lot size, N, only a small sample of size n is inspected during the CbM QA process. Given the random nature of the sampling procedure, it is not possible to control the number of items in the sample where the phenomenon occurred/not occurred. While this sampling process could lead to a situation where one of the two subpopulations (N_o and N_1) is underrepresented, this is only an apparent paradox. Justifications for using a single sample with paired observations are provided in Annex V and in the CbM QA discussion document (Devos, December 2020). Moreover, it is shown that $n \leq 365$ is sufficient for the CbM QA process.

The sample size, n, is provided in Table 2 as a function on N, the population size. The table is derived in Annex V.

Table 2: Tabulated values for the sample size, n, as a function of the population size, N.

N	[1-124]	[125-199]	[200-399]	[400-2100]	> 2100
n	N	125	200	315	365

The same sample of size n will also be used for step 2.

4.4 Lot subdivision depending on compliance

In previous chapters (chapter 4.2 and 4.3) the lot for inspection was divided on the basis of the actual phenomenon occurrence. While this population partition is suitable and effective for the QA of the Sentinel-based automatic decision process (step 1), it does not take into account eligibility conditions of the scenarios analysed.

Consider the case of ploughing, the impact of a detection error is different if the detection of the ploughing marker is performed on an Environmentally Sensitive Permanent Grassland (ESPG, where ploughing is banned) or on an Arable Land (AL, where ploughing is a boundary condition) population. In an ESPG, ploughing should not occur. The farmer should not be paid (or penalized) if ploughing occurs and its marker is detected. The converse is true for AL: the farmer should not be paid if ploughing is not detected. In light of these considerations, it is possible to define two types of scenario:

- **manifestation scenario:** that requires the manifestation of the behaviour/ marker(s) detected in order to confirm payment eligibility (compliance rules).
- **absence scenario:** that expects the absence of manifestation of the behaviour/marker(s) in order to confirm payment eligibility (non-compliance rules).

Depending on the manifestation/absence of the scenario, detection errors can be further divided as

- **abatable**: if the farmer is expected to come forward and demand correction, i.e. the applicant has an interest or incentive to have the error reversed because the current state is disadvantageous for him/her. A false negative is abatable if it occurs on a manifestation scenario where the behaviour should be present to confirm payment eligibility. A false positive is abatable if it occurs on an absence scenario where a behaviour should not be present.
- **end-stage**: if the farmer has no incentive to contest. End-stage errors lead to undue payments and to direct financial losses, because the applicant has no interest or incentive to have the error reversed. A false negative is end-stage if it occurs on an absence scenario, which would lead to an undue payment. A false positive is end-stage if it occurs on a manifestation scenario where a behaviour should be present to confirm eligibility.

  The type of compliance error (abatable vs. end-stage) is determined by the unique combination of land use/ land cover and type of detection error (false positive/negative).

  Using these definitions, the false positives and false negatives in the full lot of size N can be further divided as

$$N_{01} = N_{a01} + N_{e01} \tag{1}$$

$$N_{10} = N_{a10} + N_{e10}. \tag{2}$$

Where subscripts “a” and “e” are used to denote abatable and end-stage compliance errors, respectively.

Finally, the total number of abatable/end-stage errors is given by

$$N_a = N_{a01} + N_{a10} \tag{3}$$

$$N_e = N_{e01} + N_{e10}. \tag{4}$$

The goal of step 2 is to verify that N_a and N_e are below predefined limiting qualities (LQs) using information extracted from the samples inspected during step 1.

5 Inspection

CbM QA inspection is carried out on boundary condition elements of the CbM, on the automated outcomes (items) of the Sentinel (or equivalent) data on whether they are confirmed or not confirmed.

The purpose of the QA inspection is to evaluate the correctness of the outputs by using a reference data set as a pointwise inspection of confirmation/rejection.

The inspection is done per parcel/FOI, and yields a binary "pass/fail" or "yes/no" verdict applied to every single inspection.

The inspection is applicable for the items of any lot determined as G1, G2, T1, …or C1. If boundary validation (cardinality test) is not integrated into the CbM process, then the P1-lot will be determined for a mandatory test where the sample size and the FOI in subject will be equal to the identified T-type lot with the highest importance (in that order: T4, T3, T2, T1, C1).

The assessment is done by means of visual photo-interpretation in combination with pre-defined tools for spatio-temporal assessment. Assessment is in principle an independent "blind check" interpreting the presence of a particular phenomenon (as defined per lot type) without an a-priori information from the detection outcome. The applicable methods of information extraction and visual interpretation are given in Annex III.

Inspection of step 1 focuses on the presence/absence of the targeted phenomenon, following further analysis in step 2 of the QA where the focus is on whether the presence or absence meets eligibility criteria. This implies that the same reference data set/sample is used for both step 1 and step 2, the sample preparation needs to include also all the information required for performing the analysis on step 2 (manifestation/absence type).

In the following chapters, the inspection procedure for both step 1 and 2 is described.

Inspection protocol for P1 lot is given in chapter 6.4.

5.1 Sentinel imagery selection

The results of the assessment depend on the ability of the imagery to be relevant and suitable to provide conclusive observations on the item of concern.

The relevance of the selected Sentinel imagery is driven by the specific scenario. More than one image will be used in some cases to obtain the information for phenomena verification. The suitability of the selected imagery depends on the following basic qualities:

- type of signal (optical or radar),
- information content: minimum resolvable object on the ground, and
- timing: date of the reference imagery in relation to the date of the item to be assessed.

For each type of information extraction, a minimum set of reference data will be defined to make photo-interpretation feasible (see Annex I).

5.2 Preparation of the decision items

All decisions participating in the CbM processes, should be identified and grouped into the separate populations depending on the type of information they generate. To identify them, follow these instructions:

- In the CbM process flow: Identify each decision (derived from given processing option) that
  - depends on Sentinel information, and
  - leads to a partial or complete conclusion of the defined scenario(s).
- For each of these decisions in the flow
  - determine which of the 7 pre-defined types of information extraction it relies on,
  - identify all items for inspection for a given campaign year: i.e. the involved FOIs that were subject to the partial or complete conclusion of the defined scenario(s).

- Compile inspection packages by grouping all decision items per type of information extraction; count the total (N) in each lot,
- From the second year onwards: if the resulting lot (or one of the two sub-lot therein, see Chapter 6.2) has failed the quality inspection, then disambiguate that lot or sub-lot into either individual decision switches (based on the location in the process flow) or, where if appropriate in terms of any other homogeneity consideration, into smaller groups of decisions.

5.3 Declaring the lot(s) for sampling

Sampling is based on the Postgres SQL RANDOM() function of the population in subject. Depending on the declared scheme type and the population size by the Member States, the sampling will be executed by the JRC in three procedures:

1. For the BPS, the sample pre-selection will be based on the annual LPIS population received in the current year (LpisPointZeroState.gml file);
2. Alternatively, the sample pre-selection could be based on the specific CbM population for a given lot delivered by the MS to the JRC (extracted LpisPointZeroState.gml file). Generation of specific CbM population is mandatory for VCS;
3. For the zones where any of the scheme monitored is smaller than the national/regional LPIS, for testing purposes and pilot projects, sample pre-selection will be based on the specific CbM population for a given lot delivered by the MS to the JRC (extracted LpisPointZeroState.gml file).

Exchange of data are carried out via a secured ftp account provided by the JRC.

After declaring the lot size, applicable scheme, year of assessment and MS region by the MS, the JRC will randomly generate the sample pre-selection (in principle a list three times bigger than the minimum sample size) and return it as a list of points with the ordinal numbers. The sample pre-selection will be a point collection spatial gml file.

Each lot declared by the MS will have corresponding sample pre-selection returned by the JRC, except the P1-lot that will be the same as one of the declared T-type.

5.4 Inspection flow

After receiving the sample pre-selection from the JRC, follow the next instruction steps:

1. For each lot (per type of information extraction) retrieve the sample pre-selection list produced by the JRC,
   a. If sample pre-selection was generated from the LpisPointZeroState.gml, identify the final parcel/FOI to inspect by spatial intersection of the point from the sample with the prepared items (polygons of the FOI’s) of that lot. In case there is no intersection found for a particular point, execute the selection of the first “free” nearest centre point from the FOIs of a given lot,
   b. if sample pre-selection was generated from the targeted CbM population of a particular lot, identify the parcel/FOI to inspect by spatial intersection of the point from the sample pre-selection with the prepared items (polygons of the FOI’s) of that lot.
2. Start inspecting a first item identified of that list and stop inspecting when reaching the minimum required sample size (maximum of 365 items):
   a. Identify the FOI representation that covers the coordinate provided in the list,
   b. Perform a blind test that feeds into a paired observation (detected vs not detected). Additional details are provided in Annex II,
   c. Determine the type of scenario for the FOI (manifestation vs absence scenario). This action is required to record eligibility conditions.
3. If the inspection is not feasible, skip the item and go to the next from the pre-selection list,

4. Analyse the results from the QA with the CbM detection outcome (step 1):

   a. Determine the paired lot size of n_1 and n_0 in the bundle,

   b. Determine the acceptance numbers, AC_1 and AC_0, for positive and negative observations,

   c. Compare the number of false outcomes in each lot to that acceptance number and count for the alpha and the beta per lot.

Definitions and details about n_0, n_1, AC_0 and AC_1 are provided in Chapter 6.2.1.

5. Perform the eligibility check (step 2) considering all lots:

   a. Form the system sample as the union of the samples used for the individual lots. Determine the size of this sample, n^s

   b. Determine the number of (not waivered) abatable and end-stage errors in the system sample. These numbers are denoted n_a^s and n_e^s, respectively

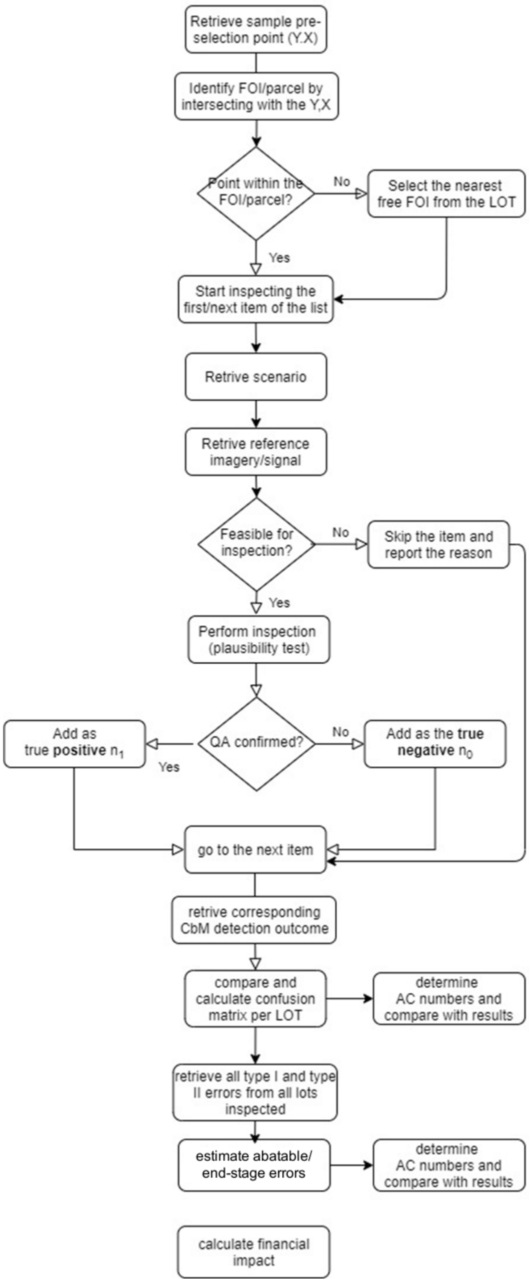
   c. Determine the acceptance numbers, AC_a and AC_e, for abatable and end-stage observations,

   d. Compare the total number of abatable and end-stage errors with the acceptance numbers.

Definitions and details about n_a^i, n_e^i and AC_e and AC_e are provided in Chapter 6.2.2.

The processing flow for the selection and inspection of a lot is shown in Figure 1. The box “retrieve scenario” implies the collection of manifestation/absence information for the determination of abatable/end-stage errors.

Figure 1: Inspection diagram applicable for lots: G1, G2, T1, T2, T3, T4, and C1


Retrieve sample pre-selection point (Y.X)
Identify FOI/parcel by intersecting with the Y,X
Point within the FOI/parcel?
No
Select the nearest free FOI from the LOT
Yes
Start inspecting the first/next item of the list
Retrive scenario
Retrive reference imagery/signal
Feasible for inspection?
No
Skip the item and report the reason
Yes
Perform inspection (plausibility test)
QA confirmed?
No
Add as the **true negative** n_0
Yes
Add as true **positive** n_1
go to the next item
retrive corresponding CbM detection outcome
compare and calculate confusion matrix per LOT
determine AC numbers and compare with results
retrieve all type I and type II errors from all lots inspected
estimate abatable/ end-stage errors
determine AC numbers and compare with results
calculate financial impact

5.5 Feasibility for inspection and skipping

Analyse visually if the agricultural land corresponding to the item (AP/FOI) can be inspected based on the available imagery set:

- check for the presence of any technical conditions on image data (i.e. artefacts, data gaps, geometric shifts), which prevent inspection of the phenomenon on the FOI representation,
- check for presence of any external (to the EO sensor) conditions (i.e systematic presence of clouds in the image time series, flooding, snow, etc.) which prevents inspection of the phenomenon on the FOI representation.

If the operator finds any of these two issues, first verify if alternative Sentinel imagery (or equivalent dataset) could address this. If no alternative is found, then the inspection is not feasible and the item is skipped from further inspection. The reporting package will contain specific tagging of such parcels.

Further skipping of items is allowed in case more items refer to the same FOI representation. In such case only the item with the lowest ordinal number should be inspected, while other items with higher ordinal numbers should be skipped.

5.6 Inspection for the systems applying classification by machine learning

While the marker-based inspection routine implies interpreting the actual phenomena observed from the suitable imagery and checking against the comparable class occurrence as declared in the GSAA, in case of machine learning classification such inspection needs extended observation against the predicted class where the class with accepted probability is taken as the most probable class present in the reality.. The extended procedure is necessary to find a possible mismatch between the predicted by the algorithm and the class found by the inspection which can go beyond a paired binary context and may result in a potential not-accounted eligibility error in the step 2 of the eligibility check (see Id 3,5,6 in table 3. below). Such multiple-combination situation is generated mostly by classification algorithms of machine learning processes where classification may predict more than two classes. In that case there is a need to introduce a plausibility check related not only to the declared class, but also related to the predicted class. In particular, both declared and predicted class can be erroneous and different from the one found during the QA process (found class).

Note that crop classification processing becomes even more challenging when dealing with crops that have similar temporal profiles. The clarification provided below addresses in a general manner classification approaches as adopted and proposed by the MS. Assuming that only a limited number of crop classes is considered and that each class grouping several crop types, has a clear signature, reliable interpretation of the data for the QA process is feasible. In this condition, the found class is assumed to represent the actual truth. The eligible character of the found class will be judged in step 2 - after identification of the error found in the Step 1.

The next table represents all possible variations between the three outcomes: declared class (GSAA), predicted class (CbM) and found class (QA). The table also provides the result from a single marker detection/prediction (item status): 1 – positive detection of the declared class and 0 – negative detection of the declared class, and the final traffic light result for the payment: 1 or green – declared crop is compliant for the payment and 0 or red - declared crop is not compliant for the payment. All inspection procedure variants together with the QA step 1 findings are represented.

Table 3: Inspection variants for machine learning classification

Id	Declared class (GSAA)	Predicted class (CbM)	Found class (QA)	Item status (CbM)	Traffic light (CbM)	Step 1 QA
1	C1	C1	C1	1	1	1 1
2	C1	C1	C2	1	1	1 0
3	C1	C2	C1	0	0	0 1
4	C1	C2	C1	0	1	0 1
5	C1	C2	C2	0	0	0 0
6	C1	C2	C2	0	1	0 0
7	C1	C2	C3	0	0	0 1
8	C1	C2	C3	0	1	0 1

The different cases summarized in Table 3 are better discussed in the following.

- ID1 – CbM predicted C1 class with high probability. The C1 class is the same as that declared in the GSAA. Since there is a match between predicted and declared class the CbM item status (decision) was set to 1 (true or positive match) and the traffic light result was set to green as class C1 is eligible for payment and the declaration was compliant. The QA inspection confirmed the presence of C1 class from the imagery time series (positive observation - 11) in step 1. Hence, no error found.
- ID2 - CbM predicted C1 class with high probability. The C1 class is the same as that declared in the GSAA. Since there is a match between predicted and declared class the CbM item status (decision) was set to 1 (true or positive match) and the traffic light result was set to green as class C1 is eligible for

payment and the declaration was compliant. The QA inspection didn't confirm C1 class, but confirmed the presence of some other C2 class from the imagery time series. C2 class is different from the C1 class, hence there is an error in prediction and declaration. The error found is a false positive (10).

- ID3 - CbM predicted C2 class with high probability. The C2 class is not the same as the declared C1 class in the GSAA. Since there is a mismatch between predicted and declared class the CbM item status (decision) was set to 0 (false or negative match) and the traffic light result was set to red as the declared class C1 didn't meet the eligibility compliance with respect to the scenario of the payment scheme (C1 is a different land cover type than C2). The QA inspection didn't confirm predicted C2 class, but confirmed the presence of declared C1 class from the imagery time series. Although C1 class is the same as declared C1 class, there is a mismatch with the predicted C2 class. This mismatch is a false negative (01) error.

- ID4 – CbM predicted C2 class with high probability. The C2 class is not the same as the declared C1 class in the GSAA. Since there is a mismatch between predicted and declared class the CbM item status (decision) was set to 0 (false or negative match), however differently from case ID3, this time the traffic light result was set to green since also class C2 is compliant with the eligibility rules with respect to the scenario of the payment scheme. Since the QA inspection did not confirm predict class C2, a step 1 error (01, false negative) was committed. While this error has to be accounted for in the step 1 evaluation, it will not propagate to step 2 since also C2 meets eligibility requirements and thus no eligibility error is committed.

- ID5 - CbM predicted C2 class with high probability. The C2 class is not the same as declared C1 class in the GSAA. Since there is a mismatch between predicted and declared class the CbM item status (decision) was set to 0 (false or negative match) and the traffic light result was set to red as the predicted class C2 didn't meet the eligibility compliance with respect to the C1 class. Predicted C2 class is the same as found in the QA, hence the negative detection was confirmed as true negative (00). Hence, no error is found.

- ID6 - CbM predicted C2 class with high probability. The C2 class is not the same as declared C1 class in the GSAA. Since there is a mismatch between predicted and declared class the CbM item status (decision) was set to 0 (false or negative match). Differently from case ID5, the traffic light result was however set to green since the predicted class C2 is also meeting the eligibility criteria with respect to the scenario considered. Predicted C2 class is the same as found in the QA, hence the negative detection was confirmed as true negative (00). Hence, no step 1 error is found. Since the traffic light result is green, eligibility criteria will be confirmed and no abatable error is committed. In this case, the error is not propagated to step 2.

- ID7 - CbM predicted C2 class with high probability. The C2 class is not the same as declared C1 class in the GSAA. Since there is a mismatch between predicted and declared class the CbM item status (decision) was set to 0 (false or negative match) and the traffic light result was set to red as the predicted class C2 didn't meet the eligibility compliance with respect to C1 class. The QA inspection found a third class C3, that is different from both declared C1 and predicted C2 classes. Although class C3 found in the QA is equally not matching the declared C1 class, the classification error should be accounted since the true class is not matching the predicted class. A false negative error is found.

- ID8 - CbM predicted C2 class with high probability. C2 class is not the same as declared C1 class in the GSAA. Even there is a mismatch between predicted and declared class, according to the eligibility rules they both can be compliant with respect to the payment scheme (lane). Therefore, as CbM item status is set to 0 (false or negative match) but the traffic light result is set to green as the declared class C1 meet the eligibility compliance. The QA inspection found third class C3, that is different from both declared C1 and predicted C2 classes. A false negative error is found.

If a classification approach with more than 2 classes is considered, then there is the possibility that the found class is different from both the declared and the predicted ones. In this case, the CbM algorithm has committed a step 1 error since it did not identify the correct class. This observation is relevant and important because step 1 deals with the actual performance of the automatic detection (and eventually classification) process and should serve as investigation tool for the MS. This section of the TG clarifies that a step 1 error is committed every time the outcome of the automatic detection process is different from the one found during the QA irrespectively of the eligibility conditions that are verified in step 2. The eight cases considered above allow to deal with potential step

1 errors that would not be accounted for by considering a different approach as considering for example a correct decision when both the predicted and found classes differ from the declared one.

Note that Table 3 also considers the impact of eligibility conditions. In particular, cases ID3 and ID4, ID5 and ID6, and ID7 and ID8 are paired in the sense that they lead to the same step 1 outcome. However, different results in terms of eligibility are observed. In the analysis, a declared class, C1, is always meeting eligibility conditions.

5.7 Inspection set up

JRC has developed a collection of notebooks to support the inspection set up. They are currently accessible for the MSs, free of charge, on the following link: http://jrc-ntb.vm.cesnet.cz/GTCAP/cbm.

The user needs to have an account to access the repository. To get an account please send an email to: Konstantinos.ANASTASAKIS@ext.ec.europa.eu.

6 Testing

6.1 ISO validation test

The sampling methodology for the CbM QA uses the same statistical assumptions of the binomial distribution underlying the international standard ISO document 2859/2 for inspection by attributes. ISO 2859/2 provides a sampling plan applicable for use when individual or isolated lots are to be sampled.

The purpose of that acceptance sampling inspection is to assure that the producer submits lots of a quality that is not worse than a level demanded by the consumer. That quality level is LQ (limiting quality) and the sample plan is based on the mathematical theory of probability of a worse lot slipping through. A lot at LQ has a low probability of acceptance, any better quality has a much higher probability.

6.2 Determining the acceptance numbers

In the following, the acceptance numbers for the tests developed for step 1 and 2 are provided along with illustrative examples for the implementation of the tests.

6.2.1 Step 1

As the sampling plan has been determined (random selection of the parcels in paired lots), the quality is acceptable if the number of type I and type II errors (nonconformities) is less than the acceptance number (AC_1 and AC_0) specified in the plan.

Once the sample for inspection per information extraction type detection has been determined, and inspection revealed the numbers of the matrix for positives, false positives (α), negatives and false negatives (β), the acceptance values should be compared with the outcome in the confusion matrix.

Table 4. QA confusion matrix of the sample (n).

Paired lot	**CbM Detected**	**CbM not detected**	sum
QA Confirmed	true positive n_{11}	false negative n_{01} or β'	n_1
QA not confirmed	false positive n_{10} or α'	true negative n_{00}	n_0
Sum	$n_{.1}$	$n_{.0}$	n

When determining the acceptance number, AC, the number of positive inspection outcomes, n_1, and the number of negative inspection outcomes, n_0, confirmed/not confirmed by the QA should be taken separately. AC numbers for alpha' and beta' errors are pre-defined based on the binomial distribution (mathematical derivations are provided in Annex V) and are provided in Table 5 below.

The testing flow is the following:

1. populated confusion matrix table allows one to calculate n_1 and n_0 as the number of the observed and missing phenomena during the blind/ test (not the CbM detection outcome):
   - $n_1 = (n_{11} + n_{01})$
   - $n_0 = (n_{10} + n_{00})$

- verify if $n = n_1 + n_0$

2. use Table 5 to calculate AC_1 and AC_0 respectively; for both n_{10} and n_{01} the LQ is set to 10%[1] . n_{10} is compared with AC_0 and used to test the presence of an excessive number of type I errors (false positives). If n_{10} is lower than or equal to AC_0 the test is passed and the CbM process is deemed of sufficient quality (type I errors less than 10%). Otherwise, the CbM process fails the QA test. A similar process is performed on n_{01} that is compared with AC_1 to verify the presence of type II errors (false negatives).

Table 5. Acceptance sampling for α and β with LQ= 10%. The acceptance size, AC_0/AC_1 has to be compared with n_{10}/n_{01}, the number of false positives/negatives.

Sample size n_0/n_1	AC_0/AC_1	*Sample size* n_0/n_1	AC_0/AC_1	*Sample size* n_0/n_1	AC_0/AC_1	*Sample size* n_0/n_1	AC_0/AC_1
0-21	0*	116-127	7	210-221	15	301-311	23
22-37	0	128-139	8	222-232	16	312-323	24
38-51	1	140-151	9	233-244	17	324-334	25
52-64	2	152-163	10	245-255	18	335-345	26
65-77	3	164-174	11	256-266	19	346-356	27
78-90	4	175-186	12	267-278	20	357-365	28
91-103	5	187-198	13	279-289	21		
104-115	6	199-209	14	290-300	22		

In Table 5, AC_0 and AC_1 values indicated with “0*” mean that while no decision errors should be made by the CbM test, the sample size is too small to perform QA with a quality risk lower than 10% as required by the ISO 2859/2 standard. In this case, a higher consumer risk is accepted.

In the current version of the TG, the LQ for both false positives and false negatives is set to 10%. In previous versions, LQ was set to 5% for α errors and to 10% for β errors. These values were selected considering the industry practice suggested by the ISO 2859/2 standard. The symmetry and interchangeability between α and β errors, which depend on the formulation of the detection process (detecting the presence in opposition to the detection of the absence of a phenomenon) has been however recognized. For this reason, both LQs are now set to 10%. Having both LQs set to 10% also provides better protection against excessively strict tests in the case of small sample sizes (zero acceptant tests in the ISO 2859-2).

Example 1:

Tell-tale marker related to “grazing” scenario For a population of 1.500.000 parcels/FOIs an inspection sample n of 365 items has been determined (see Table 2) and inspected. The outcome of the inspection is:

Phenomenon	CbM detected	CbM not detected	∑
QA found	182	85	n_1=267
QA not found	3	95	n_0=98

Number of type I errors (n_{10}=3), acceptance number is set to 5 (AC_0=5), hence the test passes.

Number of type II errors (n_{01}=85), acceptance number is set to 20 (AC_1=20), hence the test fails.

Example 2:

[1] The LQ values have been progressively updated through the development of the TG guidance. Previous LQ values were LQα =5% and LQβ= 10% for 2019 year.

Tell-tale marker on ploughing related to "annual herbaceous crop" scenario. For a population of 900.000 parcels/FOIs an inspection sample n of 365 items has been determined and inspected. The outcome of the inspection is:

Phenomenon	CbM detected	CbM not detected	∑
QA found	281	1	n_1=282
QA not found	28	55	n_0=83

Number of type I errors (n_{10}=28), acceptance number is set to 4 (AC_0=4), hence the test fails.

Number of type II errors (β=1), acceptance number is set to 21 (AC_1=21), hence the test passes.

6.2.2 Step 2 – Eligibility check

In step 2, a CbM system level verdict is performed by parcel-level eligibility assessment. By "system" is meant here the set of all items in the scope of the CbM process as described in the CbM discussion document (Devos, 2020, Figure 3). P1 and P2 are outside the boundaries of the system. For this reason, the items from several lots (G1, G2, T1-T5) are jointly considered. The following procedure is defined.

Consider the samples for the different lots (G1, G2, T1-T5) and create the union sample:

$$\Sigma = \bigcup_i \Sigma_i \tag{5}$$

where Σ_i is the sample (set of FOIs) selected for the inspection of the i^th^ lot. Note that the union operator is used to avoid that a FOI inspected in more than one sample (i.e. for more than one lot) is double-counted in the system level analysis of step 2. Σ is the system level sample and it is a set of unique elements: each FOI is listed only once. Σ has size n^s: this size is greater or equal to the size of the largest sample, Σ_i.

For each element in Σ verify if it led to an error (false positive/negative) during step 1 and determine if this error is abatable or end-stage (this depends on the eligibility conditions). If an element $f \in \Sigma$ was originally present in a single sample (say Σ_k), it is an abatable/end-stage error at system and scheme level if it is a false positive/negative in Σ_k and meets manifestation/absence conditions collected during the preparation of the sample Σ_k. If an element $f \in \Sigma$ was originally present in more than one sample, it is counted as a system abatable/end-stage error if it is determined as abatable/end-stage error in at least one of the original samples where it was present. This procedure avoids double counting abatable/end-stage errors at the system level.

For systems, where sequential processing of the final parcel/FOI eligibility is applied, abatable and end-stage errors can be waivered. A waiver can be applied only in case when, for a single FOI, the QA confirms the correctness of the final eligibility decision generated by the CbM, within a given scenario. Such sequential processing should be documented to provide the applicable eligibility or scenario rules for each combination of information extraction and their connection in the system.

Example: Rice crop detection scenario is processed sequentially in a chain for a given FOI: 1) event detection, 2) second event detection, 3) crop classification. the first and the second event detections (T3) generated a negative detection of typical rice harvest and water presence on the FOI. Only after a third sub-processing, a rice crop was detected correctly by a crop classification algorithm (T4). The final traffic light for payment for the FOI has been assigned as GREEN. In the QA, step 1: the negative detections in the lots T3 yield a false negative, step 2: false negative errors were classified as abatable errors. But the final eligibility for that FOI was assigned as "green", hence the abatable errors found in T3 can be waivered.

The total number of abatable/end stage errors (at the system level) is then obtained as by counting the abatable/end-stage errors in Σ. n_a^s and n_e^s are used to denote the number of abatable/end-stage errors, respectively. Subscript 's' denotes system level quantities.

In order to test is the number of abatable errors is below the prescribed limiting quality (LQ = 5%), a test of the form

$$n_a^s \leq AC_a \qquad (6)$$

is performed.

Similarly, to verify that the number of end-stage errors is below a limiting quality LQ = 5%, a test of the form

$$n_e^s \leq AC_e \qquad (7)$$

should be performed. AC_a and AC_e are the acceptance numbers and are obtained similarly to AC_0 and AC_1 (see Annex V) with limiting qualities both equal to 5%. In this case, the acceptance numbers are obtained as a function of the size of the system sample, Σ. If all the lots have the same size (maximum 365) and no intersection between samples occurs, the total size, n^s, will be a multiple of this basic size.

Values for AC_a and AC_e are provided in Table 6.

Table 6: Acceptance numbers for Lq = 5%. The acceptance size, AC_a and AC_e, have to be compared with n_a^s and n_e^s, the number of abatable and end-stage errors.

Sample size, n^s	AC_a/ AC_e	Sample size, n^s	AC_a/ AC_e	Sample size, n^s	AC_a/ AC_e	Sample size, n^s	AC_a/ AC_e	Sample size, n^s	AC_a/ AC_e
1-44	0*	785-806	31	1484-1504	63	2169-2189	95	2846-2867	127
45-76	0	807-828	32	1505-1526	64	2190-2210	96	2868-2888	128
77-104	1	829-850	33	1527-1547	65	2211-2231	97	2889-2909	129
105-131	2	851-872	34	1548-1569	66	2232-2253	98	2910-2930	130
132-157	3	873-895	35	1570-1590	67	2254-2274	99	2931-2951	131
158-183	4	896-917	36	1591-1612	68	2275-2295	100	2952-2972	132
184-208	5	918-939	37	1613-1633	69	2296-2316	101	2973-2993	133
209-233	6	940-961	38	1634-1655	70	2317-2338	102	2994-3014	134
234-257	7	962-983	39	1656-1676	71	2339-2359	103	3015-3035	135
258-281	8	984-1005	40	1677-1698	72	2360-2380	104	3036-3056	136
282-305	9	1006-1026	41	1699-1719	73	2381-2401	105	3057-3077	137
306-329	10	1027-1048	42	1720-1741	74	2402-2422	106	3078-3098	138
330-352	11	1049-1070	43	1742-1762	75	2423-2444	107	3099-3120	139
353-376	12	1071-1092	44	1763-1783	76	2445-2465	108	3121-3141	140
377-399	13	1093-1114	45	1784-1805	77	2466-2486	109	3142-3162	141
400-422	14	1115-1136	46	1806-1826	78	2487-2507	110	3163-3183	142
423-445	15	1137-1158	47	1827-1848	79	2508-2528	111	3184-3204	143
446-468	16	1159-1179	48	1849-1869	80	2529-2550	112	3205-3225	144
469-491	17	1180-1201	49	1870-1890	81	2551-2571	113	3226-3246	145
492-514	18	1202-1223	50	1891-1912	82	2572-2592	114	3247-3267	146
515-537	19	1224-1245	51	1913-1933	83	2593-2613	115	3268-3285	147
538-560	20	1246-1266	52	1934-1954	84	2614-2634	116		
561-582	21	1267-1288	53	1955-1976	85	2635-2655	117		
583-605	22	1289-1310	54	1977-1997	86	2656-2676	118		
606-627	23	1311-1331	55	1998-2018	87	2677-2698	119		
628-650	24	1332-1353	56	2019-2040	88	2699-2719	120		
651-672	25	1354-1375	57	2041-2061	89	2720-2740	121		
673-695	26	1376-1396	58	2062-2082	90	2741-2761	122		
696-717	27	1397-1418	59	2083-2104	91	2762-2782	123		
718-739	28	1419-1439	60	2105-2125	92	2783-2803	124		
740-762	29	1440-1461	61	2126-2146	93	2804-2824	125		
763-784	30	1462-1483	62	2147-2168	94	2825-2845	126		

As for step 1, the value "0*" in Table 6 indicates that the sample size is too small to produce a statically significant decision. Indeed, the sample size is too small to guarantee a CR lower than 10%. In this case, if n_a^s=0 and n_e^s=0, it not possible to declare that the test is passed with a CR < 10%. However, if n_a^s>0 or n_e^s>0, the test fails.

Finally, if the sample size, n^s, is not present in the table above, the acceptance thresholds can be computed using the following formula:

$$AC_a/AC_e = BINOM.INV(n^s, 0.05, 0.1) - 1 \qquad (13)$$

where BINOM.INV is the inverse cumulative binomial distribution as implemented in Microsoft Excel and returns the smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value.

Thus, the testing flow is as follows:

1. Take the samples of step 1 for all lots applicable and construct the system/scheme sample Σ as the union of the individual samples
2. For each FOI in Σ determine if it is an abatable/end-stage error,
3. Sum up per scheme and determine n_a^s and n_e^s
4. Determine the total sample size (to be used for determining the acceptance numbers), n^s, as the size of Σ.
5. Use Table 6 to calculate AC_a and AC_e respectively; the LQ is set to 5%, for both cases . n_a^s is compared with AC_a and used to test the presence of an excessive number of abatable errors. If n_a^s is lower than or equal to AC_a the test is passed and the CbM process is deemed of sufficient quality. Otherwise, the CbM process fails the QA test for step 2. A similar process is performed on n_e^s that is compared with AC_e to verify the presence of end-stage errors.

Example 1:

Consider the two following lots:

- Tell-tale event marker (T3) on ploughing related to "annual herbaceous crop" scenario. An inspection sample of 365 items has been determined and inspected. The outcome of the inspection is:

Phenomenon	Scenario	CbM detected	CbM not detected	Σ
QA found	Manifestation	180	30	243
	Absence	24	9	
QA not found	Manifestation	5	73	122
	Absence	44	0	

- Identification of "annual herbaceous crop" (T4). An inspection sample of 365 items has been determined and inspected. The outcome of the inspection is:

Phenomenon	Error Type	CbM detected	CbM not detected	Σ
QA found	Manifestation	295	6	304
	Absence		3	
QA not found	Manifestation	1	56	61
	Absence	4		

- The two lots do not intersect and different FOIs are used for their analysis (this is the expected condition for most cases).

The tables above are provided in a form such that the information already determined for step 1 can be easily identified. Indeed, the main difference between the previous step is that now errors, of both type I and II, are split depending on the type of scenarios (manifestation vs. absence). The colours in the tables are used to identify end-stage (light red) and abatable (light green) errors as defined by the type of scenario.

Step 1. Total sample size: since two lots are considered (T3 and T4), each of 365 samples, the total sample size is n^s = 2 x 365 = 730

Step 2. Number of (not waivered) abatable errors: the number of abatable errors is computed considering the values reported in all the lots considered:

n_a^s = (30 + 44) + (6 + 4) = 84

Parentheses are used to identify errors coming from the first (T3) and from the second (T4) lot.

Step 3. Number of (not waivered) end-stage errors: the number of end-stage errors is computed considering the values reported in all the lots considered:

n_e^s = (5 + 9) + (1 + 3) = 18

Step. 4. Test on the number of abatable errors and comparison with the acceptance threshold. The acceptance threshold is derived from Table 6 and determined by n = 730. In this case, AC_a= 28

n_a^s=84 > AC_a=9 --> **the test fails**

Step. 5. Test on the number of end-stage errors and comparison with the acceptance threshold. The acceptance threshold is derived from Table 6 and determined by n = 730. In this case, AC_e=28

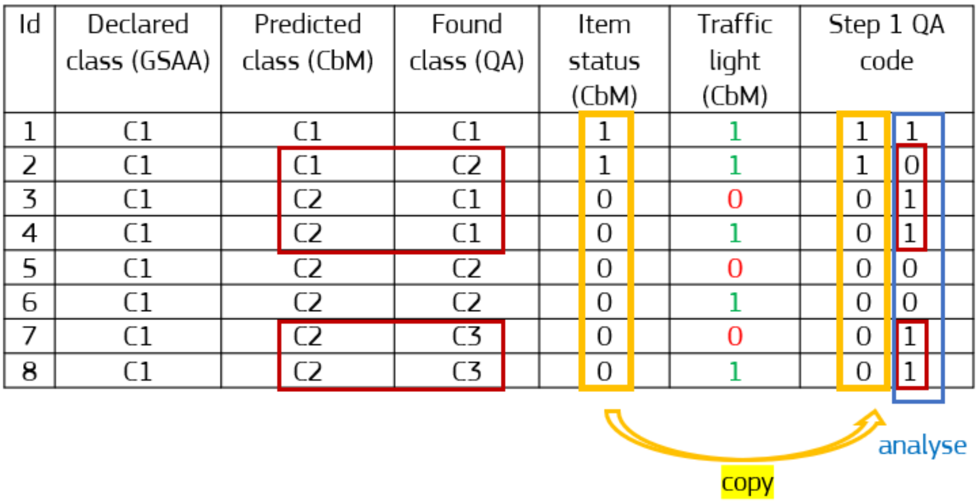
n_e^s=18 < AC_e=28 --> **the test passes**

6.3 Testing of ML classification

In section 5.6, step 1 analysis was extended to include classification approaches and evaluate them in a binary context. It was highlighted that a step 1 error is committed every time the class found by the inspection differs from the one predicted by the CbM algorithm.

For the purpose of determining confusion matrix results (as described above for the binary case) for classification approaches, and in order to simplify the relationship among declared, predicted and QA found classes, the practice summarized in Figure 2 can be applied.

Figure 2: Post-inspection error mapping



Id	Declared class (GSAA)	Predicted class (CbM)	Found class (QA)	Item status (CbM)	Traffic light (CbM)	Step 1 QA code	
1	C1	C1	C1	1	1	1	1
2	C1	C1	C2	1	1	1	0
3	C1	C2	C1	0	0	0	1
4	C1	C2	C1	0	1	0	1
5	C1	C2	C2	0	0	0	0
6	C1	C2	C2	0	1	0	0
7	C1	C2	C3	0	0	0	1
8	C1	C2	C3	0	1	0	1

analyse
copy

The step 1 QA code (last column in Figure 2) is obtained as follows:

1. Copy the item status code (0,1) as the first code digit. The item status is equal to 1 if the declared and predicted class coincide and 0 otherwise,

2. Determine the second code digit (0,1) by assessing the match/mismatch between predicted and found classes:

- If predicted class = QA found class then the detection status preserves the initial binary code (00, 11) with no error,
- If predicted class ≠ QA found class then the detection status changes the initial binary code (01, 10) and a classification error is found (red boxes in the table above).

The step 1 codes correspond to the four cases described above and included in the confusion matrix. In particular, the final QA outcomes are:

- ID1 – 11 or true positive,
- ID2 – 10 or false positive,
- ID3, ID4, ID7 and ID8 – 01 or false negative,
- ID5 and ID6 – 00 or true negative.

Thus, in the classification case n_{00}, n_{01}, n_{10} and n_{11} should be determined assigning items according to the resulting Step1 QA code according to the Figure 2. If an item has a code ij, then it should be included in the n_{xy} count. After determining n_{00}, n_{01}, n_{10} and n_{11} the procedure is the same as that described for the binary case in the previous section.

The table above also clarifies how to propagate errors from step 1 to step 2 in the general classification case.

In particular, when the found class is different from both the declared and the predicted one, two eligibility conditions may occur: found class is also eligible (as declared class) or not. Since found class is different from both declared and predicted (see ID2, ID7 and ID8), an error should be propagated in step 2 where the determination of the end-stage or abatable character will be judged against the found class eligibility. For case ID2, a green light was provided on the basis of the predicted class, C1. If the found class is also eligible, the step 1 error is waived and not propagated to step 2, otherwise an end-stage error is committed. In case ID7, a red light was obtained: an abatable error is committed if on the contrary the C3 found class is eligible. The step 1 error is not propagated if the C3 class also not eligible. Finally, for case ID8 the error is waived if the found class was also eligible. Otherwise an end-stage error is committed. Cases ID3 and ID4 also consider the cases where the predicted class is erroneous. In these cases, the found class is equal to the declared one which is assumed always eligible (a farmer will never declare a non-eligible class. So, for case ID3 an abatable error is committed whereas for case ID4, the error is waived since both class C1 and C2 are eligible. Waiving an error is used in situations when both eligible found class is matching the green traffic light and when the ineligible found class is matching the red traffic light. It makes no sense to account these “errors” if the traffic light would yield a correct decision for payment to the farmer.

EXAMPLES

The following examples present a hypothetical classification algorithm used to determine four land cover classes: permanent grassland (PG), arable land (AL), permanent crop (PC) and other (O). Class O considers all other cases such as non-agricultural areas etc. During the QA, the following outcomes are found:

1) EXAMPLE (ID7)

	Class	ItemStatus (CbM) 1st digit	Traffic light (CbM) 2nd digit	Step 1 QA code	Step 2
Declared (GSAA)	AL	0	1		

Predicted (CbM)	O			01 (false negative)	For BPS, a PG if eligible – error is abatable
Found (QA)	PG				

- since declared different from predicted, the item status will be set to 0 (first digit),
- since the found is different from the predicted, the QA negates the predicted outcome and a step 1 error occurs. Thus, the second digit should be different from (negation of) the first one and a 1 is found.

Step 1 outcome: 01 (conventionally) a false negative.

In terms of eligibility we assume that both AL and PG are eligible for payment. In this case, the step 2 error will be abatable.

2) EXAMPLE (ID2)

	Class	ItemStatus (CbM) 1st digit	Traffic light (CbM) 2nd digit	Step 1 QA code	Step 2
Declared (GSAA)	AL	1	0	10 (false positive)	For BPS, a O is ineligible – error is counted as end-stage
Predicted (CbM)	AL				
Found (QA)	O				

- since declared equal to predicted, the item status will be set to 1 (first digit),
- since the found is different from the predicted, the QA negates the predicted outcome and a step 1 error occurs. Thus, the second digit should be different from (negation of) the first one and a 0 is found.

Step 1 outcome: 10 a false positive.

In terms of eligibility, class O is found as not eligible. In this case, error is propagated to step 2.

6.4 P1 cardinality test

Given the boundary condition related to the continuous validity of each of the FOIs used in CbM process, the P1 type test will be mandatory for all systems that didn't integrate such test (type G1 – spatial cardinality) within the overall CbM process. This is a key boundary ensuring that the area component is dealt with upfront and that the observations made on each FOI are meaningful. This makes population of P1 type testing mandatory (see CbM QA discussion document).

In this case, the testing procedure aims at detecting matches between the FOIs represented by the GSAA declaration and the Sentinel data. Such interpretation doesn't generate the type I nor type II errors per se, but rather a match or a no-match between GSAA and Sentinel-derived FOI representations. More information on the spatial cardinality can be found in Annex III.

The inspection protocol is as follows:

- Sample for P1 cardinality testing is equal to the sample obtained from the LOT created for other types of information extraction (no separate LOT will have to be created for P1). In case of two or more lots, select the first one present in the following order: T4, T3, T2, T1, C1.
- Check for spatial match (1-1 spatial cardinality) between both FOI representations,
- Count and report the number of errors (QA no matches) found and compare the AC only from Table 5. (complete confusion matrix cannot be compiled). LQ10% is applicable.

Example:

P1 test - cardinality. In this example, a sample of n=125 items has been determined and inspected. The outcome of the inspection is:

P1 cardinality	QA match	QA no match	∑
GSAA/CbM	123	2	n=125

Number of cardinality non-conformances equal to 2, acceptance number is set to 7 (Table 5: for sample size 125, AC=7), hence the test passes.

7 Financial Impact

This is the third step of the quality assessment for the conclusive and automatic part of the CbM. Note that elements that are not monitored, not monitorable or came as inconclusive out from the CbM detection are not in the current scope of this chapter 7. These elements will be taken into account and discussed in a separate chapter/document, still under development. The fact that all other elements will not be taken into account, could be interpreted as financial underestimation of the CbM.

This final step of the CbM QA analyses the parameters of the end-stage errors found in the previous step and translates them into the financial impact regardless of the outcome of the step 2 (conformance test). This involves adding declared area and applicable payment rates to the agricultural parcel, and determine the amounts involved. Regardless of the fact that some end-stage errors found individually will not have an impact on the holding level, for the CbM QA they qualify and are taken into account as they indicate a propagated system error.

To calculate the financial impact, apply the following procedure:

- retrieve end-stage errors per aid scheme n_e^s and identify applicable FOI IDs,
- determine the end-stage area amount A_e by summing up the area declared for the identified FOIs,
- determine the average payment amount P_{av} (in national currency per hectare) for a given aid scheme taking into account the assessment year,
- calculate the affected financial amount AFA for the error rate by multiplying average payment amount P_{av} with the end-stage area amount A_e: AFA = $A_e * P_{av}$,
- calculate the total financial amount TFA for the given aid scheme by multiplying average payment amount P_{av} with the total area amount per aid scheme from the QA sample, TA_{QA}: TFA = $P_{av} * TA_{QA}$,
- Test the financial amounts: affected financial amount against the total financial amount: AFA / TFA < 2%.

The final financial impact per aid scheme is expressed as AFA/TFA times the total amount in EUR reported on a given date for a given assessment year.

Example:

Five end-stage errors have been identified for the BPS for 2020 assessment year:

FOI id	Declared area (ha)	P_{av}	**Affected financial amount (AFA)**
X1	1,52	110 eur	
X5	0,76		
X8	2,32		
X99	0,92		
X154	1,05		
Sum A_e	6,57 ha	110 eur	**722,70 Eur**

The total BPS area declared in the QA sample is (TA_{QA}) = 6.050,56 ha, and the total financial amount (TFA) for the BPS is calculated to be 665.561,60 Eur.

Since AFA / TFA is less than 2% (0,11 %), the residual CbM error is not significant. This does not mean that in certain subpopulations there is a material systemic issue to be addressed.

8 Reporting

The number of inspected items (paired sample size n) has been determined to be maximum of 365. By indexing on LQ5 as required for type I errors and by indexing on LQ10 as required for type II errors, separate determinations yield a different sample sizes (n_0 and n_1) and the inspection should continue until the number of inspected items reaches the minimum required sample size without skipped ones.

The MS, per type of information extraction:

- use Table 2 and Table 5 to determine the sample size and acceptance number (AC_1 and AC_0) for each error type,
- when for any decision type, the estimated sample size is significantly small (in the tables marked with *), and in the same time the sample size makes 10% of the population (lot size), MS can choose not to report the results or they can do the in-depth analysis and report it in the assessment report,
- when the observed number on non-conforming items (abatable and end-stage errors) per scheme exceed the acceptance numbers derived in point above, assign 'non-conforming' for each error of that decision type in the CbM QA scoreboard. On this basis, remedial actions could be considered.

As at May 2021, no reporting deadlines are yet set. The CbM QA reporting (per lot) shall hold:

- a CbM quality assessment report,
- if appropriate, suggested remedial action,
- CbM QA reporting data package.

Delivery instructions

The CbM QA report and, where appropriate, the remedial action plan shall be emailed to agri-implementation-support@ec.europa.eu.

The CbM QA reporting package shall be uploaded on the CbM QA portal.

8.1 Scoreboard TBD

Scoreboard to jointly developed and agreed by the CbM adopter paying agencies.

8.2 Reporting data package

To enable verification of the inspection method applied and the content expressed in the textual document, the CbM QA reporting data package, shall be provided to the European Commission. It shall hold:

Phase	**File name**	**Description**	**Note**
LOT definition	LpisPointZeroState.gml	Point representation of reference parcels (point inside a parcel)	MS
Sample pre-selection	CbmQAsamplePreselection.xml	Sample pre-selection list of coordinates	JRC
	CbmScenarioProfile.xml	Catalogue of scenario definitions	MS

CbM QA reporting package	CbmEligibilty Rules.xml	Combination of items at lot level and final eligibility decision per FOI for sequential CbM processing	
	CbmItemLog.xml	List of all items/decisions in scope for the CbM	
	CbmQAsamplePreselectionStatus.xml	Final inspection status of all parcels in the preselection sample	
	CbmPolygonZeroState.gml	Polygons representing GSAA parcels or FOIs	
	CbmItemStatus.gml	List of all items/decisions in scope for the QA inspection (marker type)	
	CbmQAobservations.xml	Observation log	
	CbmReferenceImagery.xml	List of reference imagery taken for the observations	
	CbmQAassessmentReport.xml	Report with scoreboard, acceptance and testing result	

Each of the files is documented in Annex VI.

References

- DS-CDP-2018-18 technical discussion: 2nd discussion document on the introduction of monitoring to substitute OTSC: rules for processing applications in 2018-2019

(https://marswiki.jrc.ec.europa.eu/wikicap/images/b/b9/JRC112913.pdf)

- Annex IX - Technical guidance on LPIS population for LPIS QA inspection
- ISO 2859/2: Sampling procedures for inspection by attributes – Part 2
- Devos Wim, “Checks by Monitoring quality inspection: EU requirements and methodology”, JRC Technical Report, Aug 2020

List of abbreviations and definitions

Abatable error
Type of false compliance output where the farmer is expected to come forward and demand correction, i.e. the applicant has an interest or incentive to have the error reversed because the current state is disadvantageous for him/her.

Absence scenario
Scenario where the absence of a behaviour and/or its markers is required to meet eligibility conditions

AC
Acceptance number;

AL
Arable Land

CAPI
Computer assisted photo-interpretation;

CbM
Checks by monitoring; a process substituting the OTCS; a procedure of regular and systematic observation, tracking and assessment of all eligibility criteria, commitments and other obligations which can be monitored by Copernicus Sentinels satellite data or other data with at least equivalent value (as defined in Art 40a of Regulation EU No 809/2014);

Compliance
Accordance with the eligibility rules within an aid application or a system;

Confusion matrix
Is a tool to determine the performance of a classifier. It contains information about actual and predicted classifications;

CR
Consumer Risk

End-stage error
Type of compliance error where the farmer has no incentive to contest. End-stage errors lead to undue payments and to direct financial losses, because the applicant has no interest or incentive to have the error reversed.

FOI
The Feature of Interest determines spatial “footprint” of the observed land phenomenon; i.e. the space occupied by the (bio)physical object on the earth. Its spatial representation in the CbM system is derived/constructed from GSAA/LPIS. On individual Sentinels images it is captured, as a continuous patch of pixels associated with (bio)physical object.

Ground truth
Data and information on physical reality obtained by direct observation or measurement in the field; used to derive the rules and parameters for extraction of the relevant information from remote sensing data.

GSAA
Geospatial aid application;

i.i.d.
independent and identically distributed

Item
Individual processing decision based on Sentinel observation of field conditions, that impacts an explicit conclusion within the CbM process flow or within its sub-process;

a unit of the assessment/inspection;

LOT
A quantity produced together and sharing the same production costs and specifications;

LPIS
Land parcel identification system;

Manifestation scenario
Scenario where the manifestation/presence of a behaviour and/or its markers is required to meet eligibility conditions

Monitorable	It refers to either an eligibility condition, a parcel or other element that is a subject of automated process in the CbM and as such can produce observable verdicts on Sentinel based processing;
MS	Member State
N	Population size;
n	Sample size;
N_0	Sub population of negatives;
N_1	Sub population of positives;
OTSC	On the spot check; control process of the farmers' applications on 5% sample in a given year;
PA	Paying Agency
QA	Quality Assurance
TG	Technical Guidance
Type I error	α, false positive of the automated process in the system; on a sub process level a type I error occurs when a decision (or a switch) identifies falsely presence of phenomenon
Type II error	β, false negative of the automated process in the system; on a sub process level a type II error occurs when a decision (or a switch) identifies a falsely absence of phenomenon

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

XX-NA-XXXXX-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:XX.XXXX/XXXXXX

ISBN XXX-XX-XX-XXXXX-X