Metrological evaluation and testing of robots in international competitions

Deliverable title	D5.1: ACRE Evaluation Plan
Deliverable lead	Politecnico di Milano
Related task(s)	T5.1 ACRE Competition definition
Author(s)	Riccardo Bertoglio, Giulio Fontana, Matteo Matteucci (POLIMI) Michel Berducat, Daniel Boffety, Rémi Rescoussie (INRAE) Davide Facchinetti, Stefano Santoro (UNIMI)
Dissemination Level	public
Related work package	WP5 : Agri-food
Submission date	July 7th, 2020
Grant Agreement #	871252
Start date of project	1st January, 2020
Duration	36 months
Abstract	Description of the features and timeline of the ACRE benchmarking competition (Agri-food Competition for Robot Evaluation)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Versioning and Contribution History

Version	Date	Modified by	Modification reasons
V1.0	July 7th 2020	all WP5 partners	Submitted for review to Coordinator
V1.1	July 7th 2020	Giulio Fontana	following review by Remi Regnier
V1.2	July 8th 2020	Giulio Fontana	following review by Anne Kalouguine

List of Abbreviations and Acronyms

Abbreviation/Acronym	Meaning
ACRE	Agri-food Competition for Robot Evaluation
FBM	Functionality BenchMark
TBM	Task BenchMark
EGER	Estimated Global Error Rate
LAI	Leaf Area Index

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Executive summary

ACRE (*Agri-food Competition for Robot Evaluation*) is one of the robot competitions designed and organised by the METRICS project. ACRE brings the idea of *benchmarking competition* to the applications of robotics in agriculture. Among these, ACRE devotes particular attention to autonomous weeding: in fact, by providing an alternative to the massive use of chemical products, weeding robots have the potential to bring environmental, societal, and economic benefit to Europe. A benchmarking competition evaluates robots according to scientifically sound protocols, employing quantitative and objective performance metrics. This document describes these elements of the ACRE competition. To maximise their impact on stakeholders, ACRE's field events take place in real-world agricultural environments located both in France and in Italy, and involve live crops and weeds chosen for their agricultural relevance. These elements are described by this document as well.

Compliance with METRICS methodology

Work Package 2 of METRICS provides the methodological background to the whole project. This document has been prepared in accordance with the guidelines set by WP2, and especially with METRICS' Common Evaluation Template as oulined in Deliverable D2.1. The following table is provided to help the reader in identifying the elements of the Template in the ACRE evaluation plan.

Topic	Taken into account	Detail
Organization of the evaluation		
The first occurrence of the competition is a dry-run	Yes	The dry run will take place in October 2020, as described in Section 2.1.1.1
The evaluation plan is formalized	Yes	The evaluation plan is presented in Section 2
Evaluation tasks		
Each evaluation task is relevant for industry	Yes	Relevance to industry has been ensured leveraging the extensive experience and contacts that two of the partners of WP5 (i.e., INRAE and UNIMI) have with industry; further contacts with stakeholders are foreseen to fine-tune the benchmark suite of ACRE
The dependent and independent variable of each evaluation are identified	Yes	They are described in Section 4
The evaluation is modular (FBM+TBM)	Yes	A high-level description of TBMs and FBMs is available in Section 3, while Section 4 provides further details
The constraints are adapted to the objective of the evaluation	Yes	Section 4 provides information about how this has been done
Testing environments		
Repeatability and reproducibility of the observations are maximized	Yes	Benchmarking procedures (concerning both setup and execution) are aimed at promoting repeatability and reproducibility
The accessibility of the test beds is maximized	Yes	Details are available in Section 2
A qualification procedure is defined and implemented	Yes	Benchmarking procedures are designed to support this, as explained in Section 4
Scoring		
Measurements and estimations are clearly identified	Yes	Details are provided in Section 4
Subjectivity is addressed in an appropriate way	Yes	Details are provided in Section 4
Metrics are properly designed	Yes	Details are provided in Section 4

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

1 Introduction

The goal of this Deliverable is to provide complete and precise information about the ACRE competition. The document is structured as follows. Section 2 presents the structure of the competition, its timeline (including the changes imposed by the 2020 Covid-19 pandemic) and the facilities where the field campaigns will take place; Section 3 presents the ACRE benchmarks; finally, Section 4 is devoted to the implementation of the ACRE field campaigns.

While Section 3 provides a high-level description of the benchmarks, highlighting their main features and the reasons for their relevance, Section 4 contains a preliminary version of the information needed to actually set up and execute the benchmarks, including: how the test environment is prepared; what is the protocol for benchmark execution; what data are used for performance evaluation; what evaluation metrics are applied to such data. Since the benchmarks for the ACRE cascade campaigns are directly derived from those of the ACRE field campaigns, the contents of Section 4 make reference to the second but are relevant for both.

2 Outline of the ACRE competition

The purpose of this Section is to provide an outline of the way the ACRE competition is structured, and to describe the features of the ACRE benchmarks. Details about the actual organisation and execution of the ACRE campaigns (such as how the test environments are prepared, or how participating teams are managed) can be found in other parts of this document.

2.1 Campaigns and timeline

The key events in ACRE's timeline are its **evaluation campaigns**. An evaluation campaign is a *benchmarking competition*: i.e., a competition where the participating teams are evaluated and ranked according to the results of the application of scientific benchmarks to their performance.

ACRE (as other METRICS competitions) comprises two kinds of evaluation campaigns:

1. **field evaluation campaigns**, which take place in real-world environments representative of the agri-food domain;
2. **cascade evaluation campaigns**, which are data-based competitions that teams participate remotely to.

The datasets that the cascade campaigns are based upon are collected during the field campaigns. Thus, when the same benchmark is used by both, it is possible to directly compare the performance of teams participating to field campaigns and cascade campaigns.

The campaigns that ACRE will organise, and their (provisional for the time being) location in time are described below. ACRE will also benefit from interaction with French national project ROSE. In ROSE, state-funded industrial players have worked on the development of robot systems for weeding, with operational capabilities that partially superimpose with those required by ACRE's benchmarks. The link to ROSE is provided by METRICS partners INRAE and LNE, who are also among the organisers of ROSE.

NOTE: COVID-19

The unfortunate situation in Europe caused by the ongoing pandemic and the consequent emergency regulations had an impact on the activities and timeline of the ACRE competition. In particular, the first event (i.e., the dry-run field campaign) and the ROSE event it is connected to (see below), both expected to take place in June 2020, have been postponed. At the moment of writing, both are planned for October

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

2020, though further changes may occur according to the evolution of the pandemic, of state regulations and of other factors (e.g., actual availability of project partner facilities, state of transport networks). These changes also have an impact on subsequent activities of ACRE, such as the dry-run cascade campaign (based on datasets collected during the dry-run).

The following of this section describes (in chronological order) ACRE's evaluation campaigns.

2.1.1.1 Dry-run field campaign [planned June 2020, now October 2020, Montoldre (FR)]

The goal of the dry-run campaign is to validate the evaluation plan and produce datasets for the upcoming cascade campaigns. It will be co-located in space and time with one of the field events of ROSE, taking place at INRAE's facility in Montoldre (France). Such co-location will help bootstrap ACRE by possibly involving already established ROSE teams and ensure the availability of datasets by leveraging ROSE teams' existing robots to collect them.

The ACRE field dry-run is co-located with the "second complete assessment" event of the pre-existing ROSE Challenge. At a meeting of ROSE Challenge participants that took place on June 17th 2020 it was decided that the ROSE second complete assessment will take place in October 2020 (between weeks 42 and 43, i.e., between 12th October and 23rd October). Consequently, and taking into account the ongoing Covid-19 health crisis, the ACRE dry-run field campaign has been moved accordingly.

The dry-run ACRE campaign has multiple objectives, i.e.: testing the environment, setup and execution of the benchmarks, identifying possible issues; collecting data to be used for subsequent activities of ACRE (such as cascade campaigns); establishing contacts with stakeholders to raise interest towards ACRE, promote the competition to possible participants, and collect feedback on the ACRE benchmarks and methodology.

2.1.1.2 Dry-run cascade campaign [October 2020 or January 2021, web-based]

NOTE: COVID-19

One of the key activities of the dry-run field campaign should have been to collect the datasets on which the dry-run cascade campaign will be based. Disruptions (such as the move to October 2020) due to the Covid-19 pandemic forced the ACRE organisation to consider a backup solution. In particular, permission has been asked from ROSE to use datasets generated by ROSE participants (and annotated by LNE) for the ACRE cascade campaign(s). The permission has been granted, at least partially, at the already cited ROSE meeting of June 17th; a precise definition of the extent of such permission is currently underway.

This data-based evaluation campaign will involve remotely participating teams. It will be based on datasets collected during the 2019 ROSE Challenge event, provided that ROSE's participants agree to grant access to them. In this case, ACRE dry-run cascade campaign will take place in October 2020.

In case it is not possible to use the 2019 ROSE datasets, new datasets will be provided by the October 2020 ROSE Challenge event (from one or more ROSE teams) and the October 2020 ACRE dry-run. In this case, the ACRE dry-run cascade campaign will take place in January 2021.

2.1.1.3 1st field campaign [June 2021, Montoldre (FR)]

This evaluation campaign takes the form of an infield competition among participating teams recruited by METRICS partners (particularly those involved in ACRE). It will provide a performance evaluation and a ranking for the participants, as well as datasets for the upcoming 2nd cascade

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

campaign. This campaign will be contiguous in time, and co-located in space, with the last field events of ROSE, taking place at INRAE's facility in Montoldre (France).

2.1.1.4 1st cascade campaign [September 2021, web-based]

As the dry-run one, the 1st cascade campaign will be data-based and involve remotely participating teams. It will be based on datasets collected during the 1st field campaign.

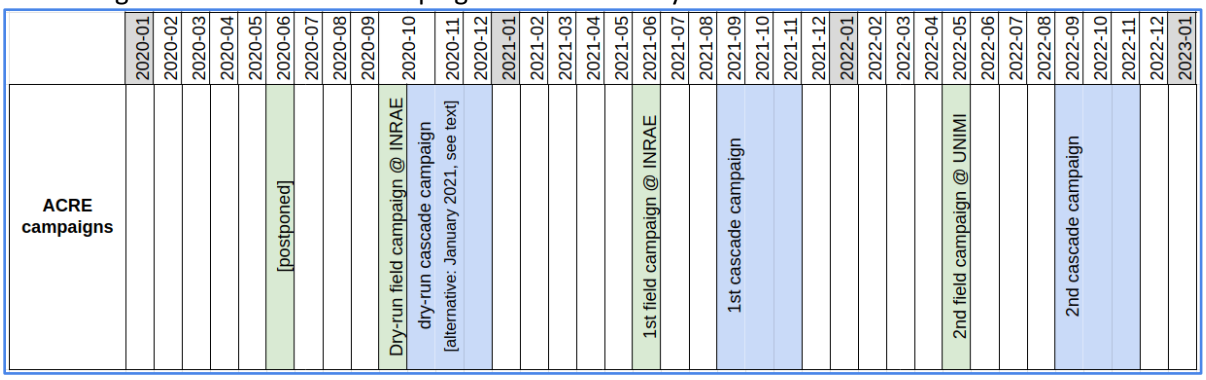
2.1.1.5 2nd field campaign [May 2022, Cornaredo (IT)]

Though this evaluation campaign will be similar to the 1st in features, structure and outcomes, it may benefit from improvements introduced thanks to that experience. The campaign will take place at UNIMI's facility in Cornaredo (Italy); if suitable sponsors will be identified, a change of location may be negotiated with them.

2.1.1.6 2nd cascade campaign [September 2022, web-based]

The 2nd cascade campaign will be data-based and involve remotely participating teams. It will be based on datasets collected during the 2nd field campaign.

The timing of ACRE evaluation campaigns is illustrated by the chart below.



NOTE: COVID-19.

Please note that the ACRE timeline may be subjected to further changes according to the evolution of the ongoing pandemic.

2.2 Test facilities and locations

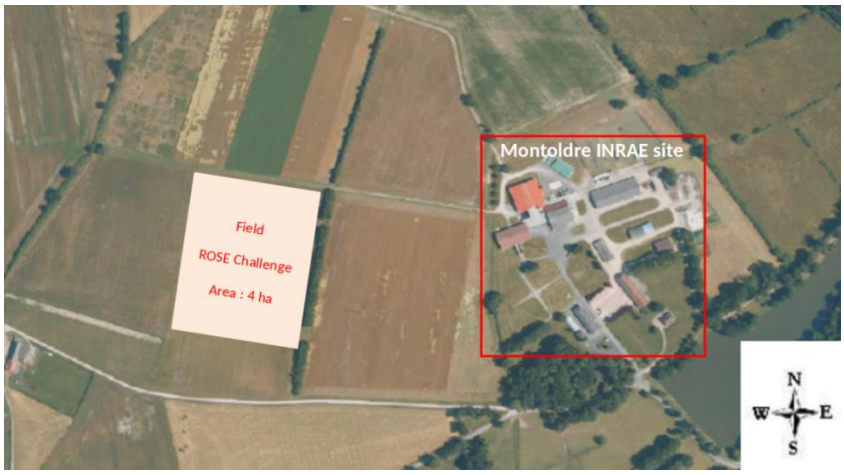
The test environments for ACRE are experimental plots presenting combinations of crops and weeds. Different weeds will be selected, which will represent some different growth patterns of weeds.
The crops considered for the competition will be selected among those that can be successfully cultivated in both France and Italy, to provide consistency between the field evaluation campaigns. Initial choices include maize and bean. Weeds will be selected to provide different patterns and shapes. ACRE will maintain the same crops of ROSE, at the same time expanding on the number and types of weeds. Differently from ROSE, in ACRE multiple weeds can be present in the same row at the same time.

2.2.1 Montoldre facility

The ACRE experimental field is located on the INRAE "AgroTechnoPole" site of Montoldre in Allier department in the centre of France. On this research and experimental site of Montoldre, a field of 4 hectares is already available for the participants of ROSE Challenge and additional areas can be easily available near in accordance with the experimental needs of ACRE. Some devices and developments have been installed for the management and the maintenance of the experimental field and its

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

surroundings and of access areas (Data capture, electric power supply, crops and environment remote monitoring by wireless sensors, field meteorological station) and are carried out by INRAE people. Several technics buildings and halls can welcome the participants with their robots and this INRAE experimental site has an experience in the organization of field meetings.



Aerial view of INRAE experimental field.

The soil of this experimental field is a sandy soil without stone suitable for weeding test and evaluation experiments with the news small prototypes solutions.

Crops:

- Maize (Zea Mays) (dry-run + 1st field campaign)
- Bean (Phaseolus vulgaris) (dry-run + 1st field campaign)

Crop plants: bean (left), and maize (right)

Weeds:

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- Ryegrass (Lolium perenne) (dry-run + 1st field campaign)
- Mustard (Sinapis arvensis) (dry-run + 1st field campaign)
- Lamb's quarter (Chenopodium album) (dry-run + 1st field campaign)
- Matricaria Chamomile (Matricaria Chamomilla) (dry-run + 1st field campaign)
- Hairy crabgrass (Digitaria sanguinalis) (1st field campaign)
- Green foxtail (Setaria viridis) (1st field campaign)

From left to right: Natural Weeds (Lamb's quarter - Chenopodium album, Matricaria - Matricaria chamomilla)

From left to right: Model Weeds (Wild mustard - Sinapis arvensis, Ryegrass - Lolium perenne, Hairy crabgrass – Digitaria sanguinalis, Green foxtail - Setaria viridis)

2.2.2 Cornaredo facility

The farm is in the municipality of Cornaredo (Milan) and occupies an area of about 23 Ha (see image below). The exact address is: *Via Cascina Baciocca - Cornaredo (MI) – ITALY*

Cornaredo is close to the area where Expo 2015 took place; there are many accommodation options in hotels located nearby. The farm is completely fenced, and in it there are numerous buildings equipped with surveillance cameras and alarm systems which will also be made available for the temporary storage of robots.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

The typical soil present in the farm is medium mixture with a high percentage of stones.

The test plots, suitably divided, will be sown with the two crops mentioned above according to conventional techniques. In addition to corn and beans, a horticultural crop (Lactuca sativa or Cucurbita pepo) will also be used, cultivated in rows 50 cm apart.

Crops:

- Maize (Zea Mays) (2nd field campaign, as in Montoldre)
- Bean (Phaseolus vulgaris) (2nd field campaign, as in Montoldre)
- Lettuce (Lactuca sativa) or Zucchini (Cucurbita pepo) (2nd field campaign – new)

Crop plants from left to right: bean, maize, and lettuce (or zucchini)

Weeds will be transplanted into plots in abundant numbers and subsequently selected in order to homogenize the test areas. The weeds proposed are typical of the Po Valley and include:

- Mustard (Sinapis arvensis) (2nd field campaign, as in Montoldre)
- Lamb's quarter (Chenopodium Album) (2nd field campaign, as in Montoldre)
- Ryegrass (Lolium perenne) (2nd field campaign, as in Montoldre)
- Johnson grass (Sorghum halepense) (2nd field campaign – new)
- Slender foxtail (Alopecurus myosuroides) (2nd field campaign – new)
- Sterile Oat (Avena sterilis) (2nd field campaign – new)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

From left to right: Johnson grass – Sorghum halepense, Lamb's quarter - Chenopodium album, Ryegrass - Lolium perenne, Slender foxtail – Alopecurus myosuroides, Wild mustard - Sinapis arvensis, Sterile Oat - Avena sterilis.

3 ACRE benchmarks

METRICS incorporates the "*Benchmarking through Competitions*" methodological framework devised by European project RoCKIn and further developed by European projects RockEU2 and SciRoc. This is the same framework underpinning the *European Robotics League* robot competitions. In a nutshell, it is based on the definition of two types of benchmarks:

- **Functionality Benchmarks** (**FBMs**), focused on specific capabilities of a robot and designed to make the benchmark as independent as possible from other features of the robot not directly involved in the functionality under examination.
- **Task Benchmarks** (**TBMs**), evaluating the execution of complex tasks involving multiple functionalities, where the final result depends both on these individually and on system-level properties of the robot such as integration between functionalities.

3.1 Benchmarks for the field evaluation campaigns

Field campaigns of the ACRE competition will involve at least two FBMs and one TBM. Discussion about what benchmarks will be chosen for ACRE is ongoing and will also incorporate feedback from stakeholders, e.g., sponsors and possible participants to the competition. This choice will be subjected to revision after every campaign, and possibly changed during the life of the METRICS project. Participating teams will have the possibility to execute a subset of the available benchmarks.
Candidate Functionality and Task Benchmarks for ACRE are listed below. For each one, a brief description is provided, subdivided into parts as follows:

Goal: the objective of the robot in executing the benchmark
Rationale: the reason why the benchmark is relevant to stakeholders
Execution: a synthesis of the benchmark protocol
Evaluation: a brief summary of the process to assess robot performance
Caveats: aspects of the benchmark that may make it difficult to execute (e.g., cost), if any
Notes: additional observations, if any

The reader is invited to read Section 4 of this document (preliminary Rulebook for the ACRE field campaigns) for additional information about the execution of the benchmarks. The reader is invited to note that -if needed- additional TBMs can be defined by adding the requirement of fully autonomous navigation to several of the Functionality Benchmarks.

3.1.1 Plant discrimination FBM

Goal: decide which plants of a row are weeds and which are crops (intra-row detection).
Rationale: being able to differentiate crops from weeds is essential to the task of autonomous weeding; more generally, the ability to distinguish one type of plant from another is important in many agricultural applications.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Execution: the robot is required to make a pass over a prepared row containing both useful crops and weed plants, using its sensors (e.g., vision) to perceive the plants. The output of the robot is a classification of the crops and weeds present in the field. To decouple the plant discrimination functionality from others, during image acquisition the robot is not required to move autonomously.
Evaluation: performance metrics compare plant classification produced by the robot with ground truth provided by qualified human.
Caveats: requires labour-intensive human classification.
Notes: n/a

3.1.2 Field navigation FBM

Goal: move through a cultivation without damaging the crop
Rationale: being able to navigate through a field, rows, or other cultivated area without causing damage to the crop is a key functionality for an agricultural robot.
Execution: predefined destination locations are identified by the organisers within a cultivated area. The robot under test is assigned one of these locations and required to reach it within a timeout.
Evaluation: Performance metrics take into consideration the accuracy of the final location and the amount of damage to the crops caused by the robot.
Caveats: areas damaged by a robot cannot be reused for other benchmarks, so special (additional) areas should be used, which increases the necessity for prepared cultivated areas.
Notes: n/a

3.1.3 Leaf area estimation FBM

Goal: estimate the leaf area of the plants along a cultivated row.
Rationale: while applying treatments to a crop, knowing how much leaf surface must be treated would allow modulation of the treatment, lowering cost and pollution; also, some treatment requires that (or are most effective when) leaves are at a specific growth stage.
Execution: the test environment for this FBM is a linear row in which there is a cultivation of approximately 30cm-50cm high plants. Leaf area is variable along the row. The robot under test is required to move along the row and use its own perception to estimate leaf area along the length of the row. The resulting estimate will be a 1-dimensional function of location along the row.
Evaluation: performance metrics are based on a comparison between the ground truth leaf area function estimated by human experts with a measurement tool.
Caveats: generating precise leaf area ground truth requires much labour; simplified methodologies can probably be adopted (e.g., subdividing leaves in classes and counting the items in each class) but they must be developed and validated.
Notes: The robot devices and ground truth acquisition systems might be different or might be the same, e.g., cameras. We need to certify the accuracy of the instrument used for the evaluation. Artificial plant preliminary scanned could be used in case we want a highly accurate measurement, but this might hinder the sensing device of the robot. Ex-post destruction possible, but tests should be done in the same day for all the teams. A decision will be made according to the registered participant teams' sensors.

3.1.4 Weed destruction FBM

Goal: destroy unwanted plants (weeds) in intra-row without damaging wanted ones (crops).
Rationale: being able to destroy specific plants in intra row while not damaging other in the vicinity is necessary to intelligent weeding robots.

Execution: evaluation takes place in a prepared plot containing crops and weeds in the rows and consists of a comparison of the state of the intra row area in the plot before and after the weeding. In order to make this evaluation as independent as possible from other functionalities, visual markers

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

will be used to identify crop and weed plants in the prepared plot; additionally, the robot is not required to drive autonomously along the row.
Evaluation: scoring relies on crop and weed count before and after the weeding action. To assess the effectiveness of the weed destruction and the impact on the surrounding crops, the observation of the plot is not performed just immediately after the benchmark. Instead, robot performance is assessed according to the results of one or more delayed observations of weeds (which may show regrowth) and crops (which may suffer from not immediately obvious damage).
Caveats: each prepared row can be used only once. Assessing weed and crop quantities in the row, both before and after weeding, is labour-intensive.
Notes: no constraints are imposed to the method used to destroy weeds, provided that they do not pose any risk of causing harm to people, both during and after the execution of the benchmark. In particular, use of chemicals with any level of toxicity is forbidden. Evaluation over a prolonged period is necessary to assess the capability of the robot of killing the weed (as opposed to the mere destruction of its upper part) and of not causing harm to the crop.

3.1.5 Biomass estimation FBM

Goal: estimate above-ground crop biomass
Rationale: the above-ground biomass of a crop is a good indicator of the nutritional status and nitrogen utilization of a plant. Moreover, this indicator of the crop development can permit to evaluate the level of the competition between weeds and crops in intra row and the effectiveness of previous weeding actions.
Execution: the robot is required to make a pass over a prepared field composed of one or more rows, using its sensors to perceive the plants. The robot must provide an estimate of the fresh weight of the above-ground parts of the plants (without distinguishing between types of plant). To decouple the biomass estimation functionality from others, the robot is not required to move autonomously.
Evaluation: the estimate provided by the robot is compared with the ground truth obtained by destroying the cultivation (after all participating robots have executed the benchmark) and weighing the plants.
Caveats: since the crop is destroyed for the evaluation, each run (possibly involving multiple robots, since benchmark execution itself is non-destructive) of this benchmark requires a new row.
Notes: n/a

3.1.6 Intra-row weeding TBM

Goal: perform fully autonomous intra-row weeding of a row (i.e., eliminate the weeds located among the crop plants of a row without damaging the crop).
Rationale: weeding is crucial to cultivations. However, the best method -i.e., manual weeding- is very labour-intensive, while currently available weeding machines lack accuracy and performance and are very expensive. While today the main weeding practice is the use of chemical products, UE and UE countries plan to reduce the use of chemicals[1] (which can cause environmental pollution and sanitary issues): therefore, the availability of accurate autonomous weeding machines would bring great advantages to agriculture.
Execution: the robot is placed at the beginning of a cultivated row containing a crop and one or more weeds and must proceed to autonomously weeding the row. There are no markers on the plants to facilitate their detection and identification. Thus, the task involves the detection system and the

[1] An example of this type of action is the Ecophyto II plan in France http://www.dextrainternational.com/french-government-launches-ecophyto-ii-consultation-to-reduce-reliance-on-pesticides/.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

weeding effector, but also all the decisions taken during the intervention. For this task, robot navigation is required to be fully autonomous.
Evaluation: the criterion for the evaluation is the number of weeds destroyed and crops plants uprooted during the intervention of the weeding system being assessed. Both crop plants and weeds are precisely counted before the execution of the task. To assess the effectiveness of the weed destruction and the impact on the surrounding crops, the observation of the plot is not performed just immediately after the benchmark. Instead, robot performance is assessed according to the results of one or more delayed observations of weeds (which may show regrowth) and crops (which may suffer from not immediately obvious damage).
Caveats: each prepared row can be used only once. Assessing weed and crop quantities in the row, both before and after weeding, is labour-intensive.
Notes: no constraints are imposed to the method used to destroy weeds. Evaluation over a prolonged period is necessary to assess the capability of the robot of actually killing the weed (as opposed to the mere destruction of its upper part) and of not causing harm to the crop.

3.1.7 Crop mapping TBM

Goal: produce a map of an entire cultivation by exploring it autonomously.
Rationale: variations in plants spacing can affect the canopy density and this leads to an uneven distribution of moisture and light that in turn results to lower yields. Thus, plant localization can be exploited to measure the in-plant space. Also, plant localization allows to count the crop plants. It is interesting to establish the relation between the final crop yield and the number of crop plants or a plant density mapping. The experiments could be repeated after various harvests to study several agronomic contexts. Thus, Crop mapping gives to researches a useful information to be studied.
Execution: the robot is required to explore a multi-row cultivated plot autonomously and to provide a map of crop plants. The robot will have to recognize single plants and provide their positions. Plants positions will be a set of points on a Cartesian coordinate system. Thus, each plant will be uniquely determined by its (x,y) coordinates relative to the participant's reference frame.
Evaluation: the map produced by the robot is compared (by suitable software) to a ground truth map created by human experts. The software will automatically find the best rigid alignment between the participant's map and the ground truth map and will then compute a mapping error as the discrepancy between the two.
Caveats: time consuming ground truth reconstruction and need for alignment between the ground truth and the map provided by the participants.
Notes: we do not put any limitation on the robot to be terrestrial in this case, so unmanned aerial vehicles can participate. On a first run we allow teams to have their robot teleoperated and not autonomous although the long term aim is to have them autonomous; thus two different ranking will be done, i.e., for autonomous robots and for teleoperated nes. To ease the task we could limit it to crop and not both crop and weeds.

3.2 Benchmarks for the cascade evaluation campaigns

ACRE cascade campaigns will involve at least one benchmark, preferably selected among those also implemented in field campaigns. The main features of such benchmarks have been already described in Section 3.1; details will be provided by Section 4.
The key limitation of cascade evaluation campaigns is that they are by necessity based on pre-recorded datasets. For this reason, the only robot activities that can be benchmarked by a cascade campaign are those where decisions taken by the robot during the execution of the benchmark cannot influence the data collected by the robot during the remaining part of the activity. In other terms, *closed-loop* benchmarks are not suitable for the cascade campaigns.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Considering this constraint, the following benchmarks are being considered for ACRE cascade campaigns, all of which belong to the set of candidate benchmarks for the field campaigns.

- **Plant discrimination FBM**: identical to the version described in Section 3.1. Input data is a dataset collected during an ACRE field campaign, provided to participants by the organisers.
- **Leaf area estimation FBM**: identical to the version described in Section 3.1. Input data is a dataset collected during an ACRE field campaign, provided to participants by the organisers.
- **Biomass estimation FBM**: identical to the version described in Section 3.1.
- **Crop mapping TBM**: to be used for the cascade campaign, this TBM must be transformed into an *open-loop* benchmark by considering the robot's trajectory predefined.

For all the benchmarks above, input data is composed of one or more datasets collected during one of the ACRE field campaign, provided to participants by the organisers.
For the dry-run cascade campaign, ACRE will focus on the Plant discrimination FBM, especially in case the campaign will be based on data collected during the ROSE 2019 event (as explained in Section 2.1).

4 Execution of ACRE benchmarks

The following is a description of the way that the ACRE benchmarks are planned to be organised and executed in practice. As such, it represents the first draft of the ACRE field campaign rulebook. This draft will be presented to relevant stakeholders prior and during to the dry-run field campaign and discussed as widely as possible, to identify possible issues and make the benchmarks maximally relevant. Since ACRE cascade campaigns are based on the same benchmarks of the field campaigns, most of the contents of this sections are also applicable to cascade campaign benchmarks.

4.1 Execution of the Plant Discrimination FBM

4.1.1 Test environment

The test fields used for the competition will be prepared by INRAE for the dry-run and 1st field campaign and by UNIMI for the 2nd field campaign in such a way as to make the characteristics almost homogeneous with each other.

4.1.2 For the dry-run and 1st field campaign at Montoldre in France

The dimension of each experimental plot is 46,5m of length and from 2m of width (2 rows for maize and 3 for beans). The distance between sowing rows are 75cm for maize and 37,5cm for beans. Four to six weeds are sown on the row of crops simultaneously with the crop sowing. These specifications are in line with the ROSE challenge to promote an initial synergy between the ACRE and ROSE teams.

4.1.3 For the 2nd field campaign at Cornaredo in Italy

The dimension of each test field will be agreed well in advance and will still be from 3 to 15 m of width and from 50 to 100 m of length. The distance between the sowing rows will be agreed well in advance and will still be between 50 and 70 cm. The width will be agreed with prospector participants to allow the largest participation possible.
The field will be subjected to preventive chemical weeding techniques, which will be followed by manual weeding carried out 2-3 days before the competition, to make it almost free of any visible form of weeds. In this campaign, "normalized" weeds will be manually transplanted immediately after manual weeding. By operating in this way, it will be possible to be sure that the number of weeds for each one of the selected species will be exactly the same for each test field but if in theory this method appears as very well, some difficulties can maybe to be met to set up on the field in real

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

conditions (several manual operations of work, a lot of labour and some difficulties with a dry and hot weather for weeds transplanted). Otherwise will be also possible to put the same number of weeds on the sown row and between the sown row, always to ensure the same conditions for everyone. Also, the age of the transplanted weeds, and their dimensional class, will be the same in every test field.
In the case of maize, the condition of the crop at the time of the tests will be between the case shown on the right and the one on the left of the following photo. The medium seeding distance on the row will be agreed in advance and will be selected from 8 to 15 cm.

In the case of beans, the condition of the crop at the timeof the tests will be similar to the one on the following photo. The medium seeding distance on the row will be agreed in advance and will be selected from 3 to 5 cm.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

In the case of the horticultural crop, e.g., lettuce or zucchini, it is necessary to take an agreement in advance among the kind of the crop. Depending on the choice, we will also take another agreement later about the correct age of the crop necessary for carrying out the tests.

4.1.4 Benchmarking protocol

For this task the different steps are:

1) Collecting pictures/data with each team's robot sensor systems. The robot is not required to move autonomously. It can be set up on other vector than the weed control system to collect data
2) Labelling manually by human experts each team's picture/data with DIANNE software developed by METRICS partner LNE
3) Assessing team's classification system against the DIANNE classification reference
4) Collecting results according to EGER methods
5) Analysis the origins of mistakes or differences: which type of plants is correctly or badly classified? Provide score and synthesis of evaluation.

QR codes, or markers such as ARUCO or APRIL Tags, will be put every 15-50cm, elevated relative to the soil, at a height similar to that of the plants so that they are visible from above. Only the images that contains one or more marker will be evaluated. For each marker, just one image will be picked, at random, if more images that contains the same marker are provided. Marker will be glued at the top of stakes like the one in the image. The height of the stakes could be easily adjusted by choosing how deep they are hammered in ground. The position and orientation of the marker at the top of the stakes should be studied to provide the best visibility. This procedure is aimed at guaranteeing an equal number of images per team, and a fair comparison by an automated image selection procedure which has as limited subjectivity as possible. It will be validated during the dry run.

4.1.5 Output data

Database with images labelled by the robot vision system and the human labelled ground truth. Each image should contain one or more markers to ease the proper referencing of results to ground truth.

4.1.6 Evaluation metric

Compare classification performance intra-row with EGER (Estimated Global Error Rate) Metrics

$$EGER = \frac{\#crop\_as\_weed + \#weed\_as\_crop + \#missed\_crop + \#missed\_weed}{\#crop + \#weed}$$

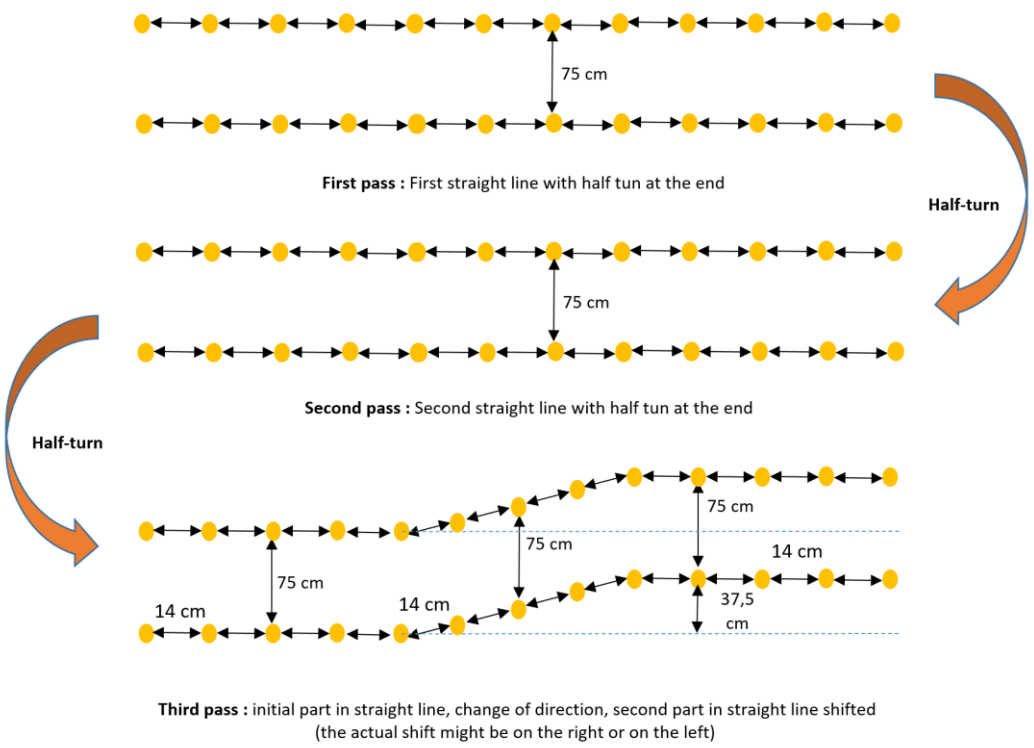
4.2 Execution of the Field navigation FBM

4.2.1 Test environment

The test fields used for the competition will be prepared by INRAE for the dry-run (straight line) and for the 1st field evaluation campaign in France (straight lines and shift) and by UNIMI for the 2nd field evaluation campaign in Italy (straight lines and shift).

The field navigation evaluation could be set up in the same conditions that the classic weeding conditions with the presence of real rows of crop plants (maize) or in artificial conditions where the rows of crop plants are simulated by means of specifics landmarks. The field comprises two straight paths and a path which includes a shift as reported in the picture below.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252



Third pass : initial part in straight line, change of direction, second part in straight line shifted (the actual shift might be on the right or on the left)

The conditions of distance between the real sowing rows or simulated sowing rows will be between 37,5 and 75 cm in the dry-run and the 1st field campaign while it will be revised for the 2nd campaign. In the figure above, an example of field navigation test possibility with two rows of crop (example of two maize rows) per pass. In these conditions, it is possible to have the same number of crop plants in each row and the same spacing between the rows to ensure the same conditions for everyone. Distances in the figure and number of rows are preliminary and will be discussed in advance with relevant stakeholders including the distance between the passes after the half turn .

In the case of maize crop, the stage of the maize crop will be with three or four leaves and an approximative size of 10 height centimetres. We expect robots to perform the benchmark without the actual implements, as only damages caused by tyres are of interest.

4.2.2 Benchmarking protocol

The robot is placed in front of the first row and it has to follow it until the end without damaging the crop. The row is straight. Then it turns and enters a second row which is 4-10 meters apart from the previous one and it follows it back to the end. It is again a straight line. Now it can decide to stop or perform second half-turn and enter a third row which is not straight anymore and follows it until the end.

4.2.3 Output data

Time/speed of the robot in performing the task, number of plants destroyed, total length of the robot trajectory as measured by the number of plants the robot has been able to surpass.

4.2.4 Evaluation metric

Teams are evaluated by the length of the trajectory they have been able to perform without damaging any plant. If the robot touches a plant we consider it as damaged. A distance equal to the number of damaged plant times the average distance between plants is subtracted from the total

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

trajectory before the team ranking. If all teams succeed in the three passes length they will be evaluated by the time they used to perform the benchmark. Two different ranks will be done considering only the straight parts and the whole trajectory.

4.3 Execution of the Leaf area estimation FBM

4.3.1 Test environment

The test fields used for the competition will be prepared by INRAE for dry-run and for the 1st field evaluation campaign in France and by UNIMI for the 2nd field evaluation campaign in Italy.
The Leaf area estimation could use either the same test bed of plant discrimination task or a new test bed without weeds. If the same test bed of Plant discrimination FBM is used, then, also weeds leaves should be accounted in the LAI estimation, thus, to decouple this FBM to the Plant discrimination one. However, a test bed free of weeds should be better to ease the LAI computation task both for participants and organizers (depending on the ground truth measurements system).

4.3.2 Benchmarking protocol

The participants are required to scan one or multiple crop rows and retrieve the LAI per unit ground area (depending on the crop rows distance, the ground unit could be 50x50 cm^2 - 1 m^2) at given positions identified via markers. Scanning systems could be equipped with every kind of sensor they want (cameras, LiDARs, spectral sensor/camera, etc.). The continuous LAI will be compared to the ground truth. The ground truth could be that obtained with a measurement tool or with an ad-hoc system composed by custom hardware/software architecture. However, organizers plan to use measurements methods of which the accuracy is known and possibly has been certified. There are different kind of tools for measuring the LAI: flatbed or handle-held in-field scanners or out-field scanning machines. An example of out-field scanning machine is the LI-3100C; Li-Cor, Lincoln, NE, USA (https://www.licor.com/env/products/leaf_area/LI-3100C/). This is system has an adjustable resolution of 0.1 mm^2 or 1 mm^2. Depending on the chosen instrument, plants could be defoliated to count and measure each single leaf.
Possible enhancements to the DIANNE software could be used to measure the LAI by image inspections, ether from team's images or from scans of the leaves.

4.3.3 Output data

The (geo)referenced LAI per unit ground area of one or more crop rows in a selected number of parcels, e.g., 10 parcels of 1 m^2 collected along a 10 m row.

4.3.4 Evaluation Metric

The evaluation metric will be a correlation measure (R^2) between the actual values of leaf area coming from the ground truth instrument, and the predicted leaf area values calculated by the participants.

4.4 Execution of the Weed destruction FBM

4.4.1 Test environment

The test fields used for the competition will be prepared by INRAE for dry-run and for the 1st field evaluation campaign in France and by UNIMI for the 2nd field evaluation campaign in Italy.
The weed destruction efficiency can be evaluated in the real conditions in the field inside each row of crop by counting the weeds destroyed immediately after the weeding action and after a longer time (several days) to check and confirm the results obtained. In a first time in the field, to avoid sowing crops and weeds, another possibility is to realise a simulation of the rows crop and weeds by means of specifics landmarks.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252


Examples of intra-row weeding assessment task with color-coded markers indicate weeds to be weeded (yellow) and crops to be preserved (blue).

4.4.2 Benchmarking protocol

Based on the ROSE challenge experience, for intra-row weeding, easily detectable markers will be placed at the base of crop and weeds to be destroyed. The robot is placed at the beginning of the row and it has to destroy all weeds according to the markers placed on the ground without damaging the crop. Only plants and weeds identified with colored markers will count to the success of the benchmark.

4.4.3 Output data

The data about the number of weeds destroyed or not destroyed and the number of crop plants damaged (collected manually). This measurement performed immediately, and it is repeated after a week.

4.4.4 Evaluation Metric

Weeding effectiveness: ratio weed counting before and after weeding
Crop saving: ratio crop counting before and after weeding
EGER Metric (Estimated Global Error Rate)

4.5 Execution of the Biomass estimation FBM

4.5.1 Test environment

The test fields used for the competition will be prepared by INRAE for the dry-run and the first field evaluation campaign in France and by UNIMI for the second field evaluation campaign in Italy.
The Biomass estimation could use either the same test bed of plant discrimination task or a new test bed without weeds. However, a test bed free of weeds should be better to ease the Biomass computation task both for participants and organizers.

4.5.2 Benchmarking protocol

The participants are required to scan one or multiple crop rows and retrieve the Biomass estimation per unit ground area (depending on the crop rows distance, the ground unit could be 50x50 cm^2 - 1 m^2) at given positions identified via markers. Robot could be equipped with every kind of sensor they want (cameras, LiDARs, spectral sensor/camera, etc.). The reconstructed Biomass will be compared to the ground truth; yet to be decided with relevant stakeholders if the ground truth will be the dry or wet biomass. In both cases, the crop is manually harvested and weighted before and after a drying in steamer during 24 hours at 70°C (in case the dry weight is of interest).

4.5.3 Output data

The (geo)referenced Biomass weight per unit ground area of one or more crop rows in a selected number of parcels, e.g., 10 parcels of 1 m^2 collected along a 10 m row.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

4.5.4 Evaluation Metric

The evaluation metric will be a correlation measure (R^2) between the actual values of biomass weight coming from the ground truth weighting procedure, and the predicted biomass weight calculated by the participants.

4.6 Execution of the Intra-row weeding TBM

4.6.1 Test environment

The test fields used for the competition will be prepared by INRAE for dry-run and for the first field evaluation campaign in France and by UNIMI for the second field evaluation campaign in Italy.
This intra-row weeding task includes the precedent functionality weed destruction. The intra-row weeding efficiency can be evaluated with the same method, in the real conditions in the field inside each rows of crop by counting the weeds destroyed immediately after the weeding action and after a longer time (several days) to check and confirm the results obtained.

4.6.2 Benchmarking protocol

The same method that the Weed destruction FBM described will be used but without markers; robots are required to complete the task fully autonomously.

4.6.3 Output data

Counting of the weeds and crops before and after weeding carried out by robots immediately and after several days.

4.6.4 Evaluation Metric

Weeding effectiveness : ratio weed before and after weeding
Crop saving : ratio crop counting before and after weeding
EGER Metric (Estimated Global Error Rate)

4.7 Execution of the Crop mapping TBM

4.7.1 Test environment

The test fields used for the competition will be prepared by INRAE for dry-run and for the first field evaluation campaign in France and by UNIMI for the second field evaluation campaign in Italy.

4.7.2 Benchmarking protocol

The robot is required to explore a multi-row cultivated plot autonomously and to provide a map of crop plants. The robot will have to recognize single plants and provide their positions. Plants positions will be a set of points on a Cartesian coordinate system. Thus, each plant will be uniquely determined by its (x,y) coordinates relative to the participant's reference frame.

4.7.3 Output data

A map with plants positions uniquely determined by their (x,y) coordinates. The reference frame origin and orientation can be arbitrarily chosen by the participant.

4.7.4 Evaluation Metric

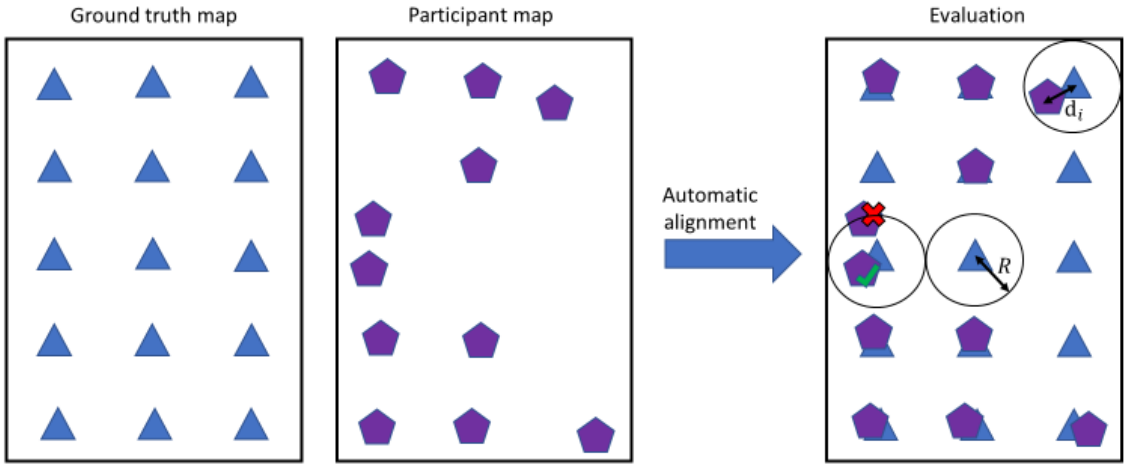
The map produced by the robot is compared (by suitable software) to a ground truth map created by human experts. The ground truth map could be constructed by using accurate positioning systems like RTK-GPS. The software will automatically find the best alignment of the participants' maps with the ground truth map and will then compute a mapping error. An example of error metric could be the following. First, the participant map is automatically aligned to the ground truth map. Then, for each plant in the ground truth map we consider a circle of ray R. If the participant map shows a plant

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

in that circle, we compute the error as the Euclidean distance (d_i) between the two. If there are no plants in the circle, this will count as a not recognized plant. If more than one plant in the circle is found, just the closer plant is considered. Then, the global error is computed as in the following:

$$error_R = \frac{\sum_{i=0}^{n} d_i + R * \max(\#not\_associated\_plants, \#not\_recognized\_plants)}{\#plants},$$

where *n* is the *#recognized_plants*.



#plants = 15
#not_recognized_plants = 5
#recognized_plants = 10
#not_associated =1

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252