# Metrological evaluation and testing of robots in international competitions

| | |
|---|---|
| Deliverable title | D5.1: ACRE Evaluation Plan (v02 – April 2021 revision) |
| Deliverable lead | Politecnico di Milano |
| Related task(s) | T5.1 ACRE Competition definition |
| Author(s) | Riccardo Bertoglio, Giulio Fontana, Matteo Matteucci (POLIMI)<br>Michel Berducat, Daniel Boffety, Rémi Rescoussie (INRAE)<br>Davide Facchinetti, Stefano Santoro (UNIMI) |
| Dissemination Level | public |
| Related work package | WP5 : Agri-food |
| Submission date | July 7$^{th}$, 2020 (concerning version 01) |
| 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 |
| V2.0 | April 1st 2021 | Giulio fontana | Significant revision of benchmarks, following experience with dry-runs and discussion among WP5 partners |

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

### 2.1.1 1st Dry-run field campaign [October 2020, Montoldre (FR)]

*Note: a complete report on the activities and the results of this campaign is available in the form of Deliverable D5.2 of project METRICS.*

The goal of the first dry-run campaign has been to validate the evaluation plan. The campaign was co-located in space and time with the "second complete assessment" event of the pre-existing ROSE Challenge, taking place at INRAE's facility in Montoldre (France).

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

The activities of the 1st dry-run field campaign have been heavily impacted by the Covid-19 pandemic of 2020 and 2021. Nonetheless, it has been possible to validate the setup of the ACRE benchmarks, and to enlist some of the ROSE teams to validate their execution. In particular, the Field Navigation Functionality Benchmark has been successfully executed.
During and after the 1st dry-run field campaign contacts have been successfully established with stakeholders to raise interest towards ACRE, promote the competition to possible participants, and collect feedback on the ACRE benchmarks and methodology.

### 2.1.2 Dry-run cascade campaign [October 2020, web-based]

> *Note: a complete report on the activities and the results of this campaign is available in the form of Deliverable D5.3 of project METRICS.*

The competition took place from October 17th 2020, to January 22nd 2021. Being an online event, it suffered from the Covid-19 pandemic in a limited way. In practice, the only change with respect to the original plan has been to use datasets generated during previous editions of the ROSE Challenge instead of new datasets collected during the 1st ACRE field campaign. Apart from this, the cascade campaign has been a success, with a large and geographically diverse participation of teams and a good quality of the final outcomes. The competition was joined by 57 teams accounting for 457 individuals.

### 2.1.3 2nd dry-run 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 campaign. This campaign will be co-located in time and space with the last field events of ROSE, taking place at INRAE's facility in Montoldre (France). It is expected (and indeed this is an actively sought result) that some of ROSE's teams will choose to participate to ACRE's dry-run field campaign as well.
While originally planned as a full-fledged field campaign, this event has been repurposed as a second dry-run due to the lingering disruptions due to Covid-19, while maintaining at 2 the number of full-fledged field campaigns as originally planned. In this way, the field campaigns will be able to take place in periods when (hopefully) the effects of the pandemic are exhausted. The inclusion of a second dry-run field campaign will let ACRE organisers gain further insight on the benchmarks, their setup and execution, and obtain additional feedback from teams and stakeholders in general.
In order to cater with the necessities caused by Covid-19, a request for an extension of 12 months of the duration of the METRICS project has been submitted to the European Commission. The extension will also allow METRICS to accommodate the planned additional field campaign (see Section 2.1.1.7).

### 2.1.4 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 2nd dry-run field campaign.

### 2.1.5 1st field campaign [June 2022, Montoldre (FR)]

Though this evaluation campaign will be similar to the dry-runs in features, structure and outcomes, it may benefit from improvements introduced thanks to that experience.

### 2.1.6 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.

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

### 2.1.7 2nd field campaign [May 2023, Cornaredo (IT)]

This field campaign will be the final activity of ACRE within METRICS. Its details will be defined according to the experience and feedback collected during previous activities, and particularly the 1st field campaign. The campaign will take place at UNIMI's facility in Cornaredo (Italy).

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


## 2.2 Test facilities and locations

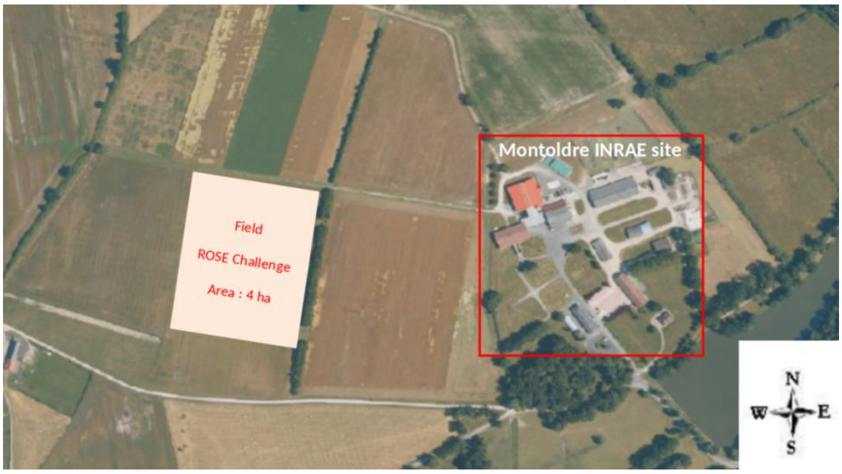
The test environments for ACRE are experimental plots presenting combinations of crops and weeds. Different weeds will be selected, which will represent some different growth patterns of weeds.

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 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.

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



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) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)
- Bean (Phaseolus vulgaris) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)


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

Weeds:

- Ryegrass (Lolium perenne) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)
- Mustard (Sinapis arvensis) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)
- Lamb's quarter (Chenopodium album) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)
- Matricaria Chamomile (Matricaria Chamomilla) ($1^{st}$ and $2^{nd}$ dry-runs + $1^{st}$ field campaign)

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

- Hairy crabgrass (Digitaria sanguinalis) (possibly used in 2nd dry-run)
- Green foxtail (Setaria viridis) (possibly used in 2nd dry-run)

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

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

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 consider the amount of damage caused by the robot to the crops and the time to complete the task.
**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: Evaluating the field plots over a prolonged period is necessary to assess the robot's capability to kill the weed (instead of the mere destruction of its upper part) and not causing harm to the crop. No constraints are imposed to the method used to destroy weeds, provided that they do not pose any risk to people, both during and after the execution of the benchmark. In particular, the use of chemicals with any level of toxicity is strictly forbidden. In case the weeding unit of a robot is based on a sprayer or other system for the distribution of chemicals, the chemical is replaced by a non-harmful water-based solution provided by the organizers. Additional information is provided in Section 4.

### 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 plans 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 species of weeds and must proceed autonomously to weeding the row. There are no markers on 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

plants to facilitate their detection and identification. Thus, the task involves the detection system and the weeding effector, and all the intervention decisions. 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. Plant positions will be a set of points on a Cartesian coordinate system. Each plant will be uniquely determined by its UTM coordinates.
**Evaluation**: the map produced by the robot is compared (by suitable software) to a ground truth map created by human experts. The software will compute a mapping error as the discrepancy between the two.
**Caveats**: time consuming ground truth reconstruction and crop plants removal.
**Notes**: we do not put any limitation on the robot to be terrestrial in this case, so unmanned aerial vehicles can participate. We will remove part of the crop plants thus to have a variable plant density.

## 3.2 Benchmarks for the cascade evaluation campaigns

ACRE cascade campaigns will involve <u>at least one benchmark</u>, 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.
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.

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

- **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 campaigns 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.

#### *4.1.1.1 For the 1st and 2nd dry-run field campaigns and the 1st field campaign at Montoldre, 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 is 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.

#### *4.1.1.2 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 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.

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 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 time of the tests will be like 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.

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.

### 4.1.2 Benchmarking protocol

For this task the different steps are:

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

1) Collecting pictures taken 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.

Teams are required to provide the ACRE organisation with the complete dataset covering the row where the benchmark takes place, i.e. to provide images covering the whole row. However, to allow a quick evaluation of results, only a small subset (not known in advance) of these images will be used for performance evaluation. Alongside each image, the team must provide the classification of the plants identified by the robot in the image.

To support the selection of the images used for performance evaluation, markers (e.g., QR codes, or ARUCO or APRIL Tags), will be put within the row (e.g., every 15-50cm), elevated relative to the soil, at a height like that of the plants so that they are visible from above. The markers will be glued to 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.
Just before the execution of the benchmark, the organisers will randomly choose a small subset of the markers (called "selected markers" in the following). Among the dataset of images provided by a robot, those considered for performance evaluation (called "candidate images" in the following) will be those that include in full at least one of the selected markers.
The evaluation of the performance of each team in the Plant Discrimination FBM will be done using one image for each selected marker. Such image will be picked from the candidate ones by choosing the one where the marker is closer to the geometric centre of the image.

It may happen that, due to a particularly narrow field of view of the camera used to collect the images provided to the organisers, the overall number of plants visible in images picked for performance evaluation is considered too low by the organisers to be representative of the performance of a robot. When this happens, the organisers will add other images from the set of candidate images to supplement those initially selected for performance evaluation. If even by using all the candidate images the number of featured plants will still be considered too low, other images will be picked randomly from the dataset produced by the robot.

### 4.1.3 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.

### 4.1.4 Evaluation metric

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

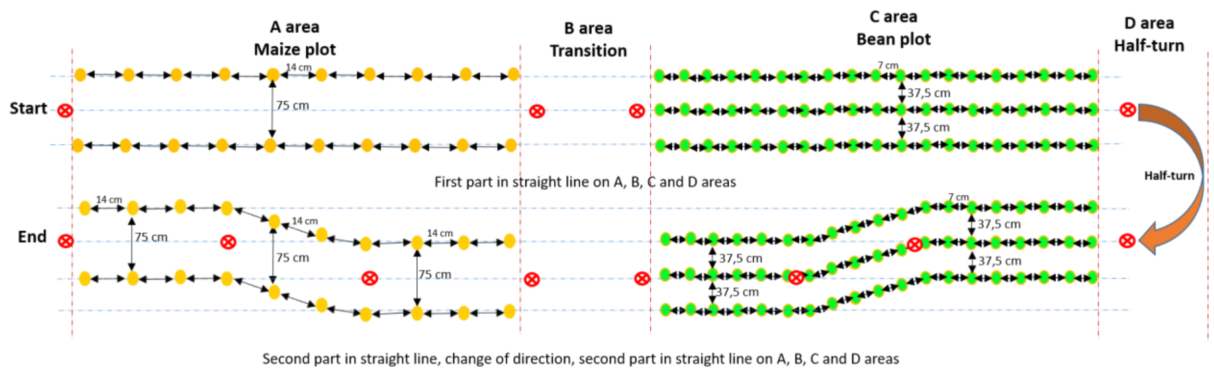
$$EGER = \frac{\#crop\_as\_weed + \#weed\_as\_crop + \#missed\_crop + \#missed\_weed}{\#crop + \#weed}$$

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

## 4.2 Execution of the Field navigation FBM

### 4.2.1 Test environment

The test fields used for the competition will comprise several rows. Each row can be straight, or it can include a "shift", i.e., an offset that shifts the median axis of the row (see image below).
The specific number, type and relative position of the rows, as well as the type of plants in the rows, will change from one campaign to the other. For instance, for the 2nd dry-run field campaign of June 2021 the arrangement will be the one depicted below.




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.

### 4.2.2 Benchmarking protocol

For this benchmark, robots can use both local and absolute (e.g., GNSS) localisation systems. The robot is asked to navigate along a trajectory involving multiple rows, from a starting location to a destination location. The trajectory is specified via waypoints (shown as red crossed circles in the image above), provided both in terms of GPS coordinates and of physical positions in the field (thus leaving teams free to use any localisation system). For each row where a shift takes place, the shift is described via a couple of waypoints, located in the inflection points (flexes) of the central line of the row when considered as a 2D curve.

The robot is free to choose the best trajectory that passes, in order, through the specified waypoints. Such trajectory will usually include the necessity to invert the direction of motion, i.e., to perform U-turns, as illustrated in the image above.

### 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 damaged. A distance penalty is subtracted from the total distance traveled based on the number of damaged plants: number of damaged plant times

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

the average distance between plants. Teams that will show the same distance score will be evaluated by the time they used to perform the benchmark.

## 4.3 Execution of the Leaf area estimation FBM

### 4.3.1 Test environment

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 weed destruction efficiency will 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. The need for a second count after a period of time is due to the fact that, for some weed destruction methods, the actual effect on the plant cannot be immediately assessed with accuracy.
The need to provide a different cultivated plot for each execution of this benchmark makes it especially challenging to the organisers.

### 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, as shown in the image below. The robot is

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

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.

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

One of the goals of ACRE is to promote a reduced and focused use of chemicals in agriculture. Therefore, no robots or implements based on bulk distribution of chemicals are admitted to the competition. However, participation from robots or implements that use focused application of chemicals is possible, though subject to an individual decision by the organizers.
For such chemical-based solutions, the key requisite is that their participation to ACRE will not disperse in the environment of the competition any quantity of toxic chemicals, since no safety measures are foreseen to manage such substances. The robot or implement must be prepared in order to not show a residue of chemicals from previous use. For example, the weeding system will need to undergo a thorough internal and external washing *before* arriving at the place of the competition. It will not be allowed to distribute in the competition fields anything except non-harmful water-based solutions provided by the organizers.
The solutions will be chosen to interact with the plants and/or the colored markers in a way that makes it possible for the organizers to assess if each plant has received a quantity of solution compatible with its focused treatment with a weeding chemical solution. Therefore, for chemical-based weeding robots performance evaluation will be based on the above-described inspection by the organizers, not on actual damage inflicted to the plants. For instance, the solution may include substances that stain the leaves in a visible way.

### 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)

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

## 4.5 Execution of the Biomass estimation FBM

### 4.5.1 Test environment

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.

### 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

This task requires the functionalities of plant discrimination, field navigation and weed destruction. Its execution takes place in fields prepared in the same ways used for the Field Navigation FBM and/or the Plant Discrimination FBM, but does not involve the use of the colored markers of the Weed Destruction FBM. Also, the fields used for this TBM will only include straight (i.e., without “shifts”) rows.

Intra-row weeding efficiency is evaluated with the same method used for the Weed Destruction FBM, i.e., by counting the weeds destroyed immediately after the weeding action and after a longer time (several days) to check and confirm the results obtained.

The need to provide a different cultivated plot for each execution of this benchmark makes it especially challenging to the organisers.

### 4.6.2 Benchmarking protocol

The robot is required to move autonomously along a cultivated row where both crop and weeds are present, destroying as efficiently as possible the weeds while minimising the damage inflited to the crops. Robots are required to complete this without any human intervention.

### 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

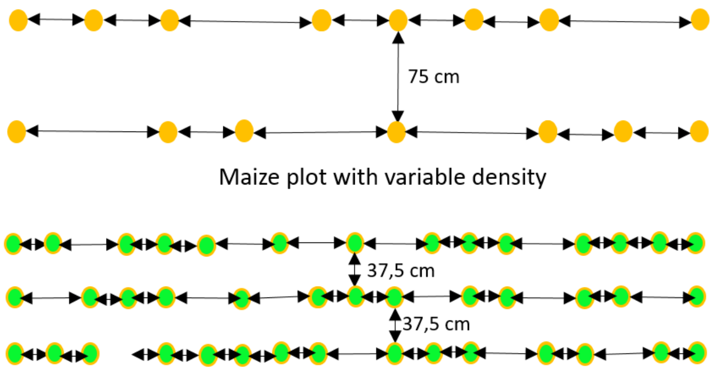
This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 871252

EGER Metric (Estimated Global Error Rate)

## 4.7 Execution of the Crop mapping TBM

### 4.7.1 Test environment

This task requires the functionalities of plant discrimination and field navigation. Its execution takes place in fields prepared in the same ways used for the Field Navigation FBM and/or the Plant Discrimination FBM, but will only include straight (i.e., without “shifts”) rows.
In the test rows, crops and weeds will be present. The robot or smart implement executing the benchmark is required to map only the crops, ignoring any weed.
Some crop plants will be removed just before the execution of the benchmarks, thus creating a variable density of plants along the row. This is done to ensure that robots produce their maps based on perception without relying on a priori knowledge about plant location. Presence in the output map of “phantom” crop plants (i.e., plants that would have been there if the crop was regularly planted but were indeed absent in the actual test field) will be penalized.
The following image illustrates an example of test setup for the Crop Mapping TBM.




Maize plot with variable density

Bean plot with variable density

### 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 UTM coordinates.

### 4.7.3 Output data

A map with plant positions uniquely determined by their (x,y) UTM coordinates. The reference frame origin and orientation is the one of the UTM absolute reference system.
In order to let teams that do not use GPS and/or absolute localisation execute the benchmark, at least two points of the cultivated plot will be marked with their UTM locations (collected with the same RTK GPS system used to provide ground truth plant locations: see Section 4.7.4). In this way, it will be possible for each team to translate the local coordinate system of their robots into the UTM reference system, and thus produce maps in UTM coordinates even if the robot is not fitted with a GPS receiver. The setup for such translation must be completed before the execution of the benchmark: no manual intervention on the maps created by the robot is allowed.

### 4.7.4 Evaluation Metric

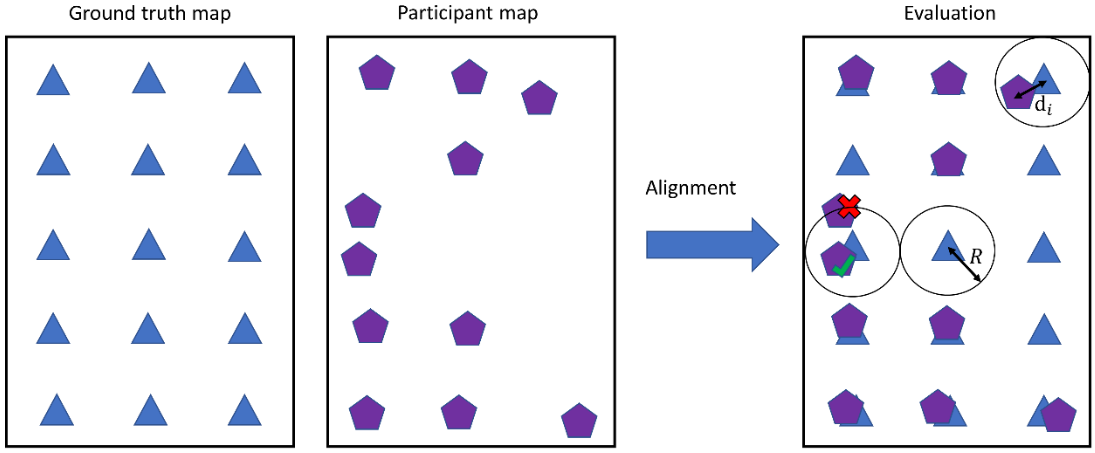
The map produced by the robot is compared (by suitable software) to a ground truth plant map created by human experts. The ground truth map will be constructed using RTK-GPS. The

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

software will compute a mapping error. The error metric is explained in the following. For each plant in the ground truth map we consider a circle of ray $R$. If the participant map shows a plant 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