Plant detection and counting from high-resolution RGB images acquired from UAVs: comparison between deep-learning and handcrafted methods with application to maize, sugar beet, and sunflower crops

Etienne David[1,*2], Gaëtan Daubige[2], François Joudelat[3], Philippe Burger[4],Alexis Comar[5], Benoit de Solan[2], Frédéric Baret[1]

[1]INRAe, UMR EMMAH, Avignon, France

[2]Arvalis – Institut du végétal, Avignon, France

[3]ITB – Institut Technique de la Betterave, Paris, France

[4]INRAe, UE Grandes Cultures Auzeville (GCA), Toulouse, France

[5]Hiphen, 22b Rue Charrue, Avignon, France

*** Correspondence:**
Corresponding Author
etienne.david@inrae.fr

Keywords: Plant counting, Deep-learning, Robustness, unmanned aerial vehicle, phenotyping. (Min.5-Max. 8)

1 Abstract

Plants density is a key information on crop growth. Usually done manually, this task can beneficiate from advances in image analysis technics. Automated detection of individual plants in images is a key step to estimate this density. To develop and evaluate dedicated processing technics, high resolution RGB images were acquired from UAVs during several years and experiments over maize, sugar beet and sunflower crops at early stages. A total of 16247 plants have been labelled interactively. We compared the performances of handcrafted method (HC) to those of deep-learning (DL). HC method consists in segmenting the image into green and background pixels, identifying rows, then objects corresponding to plants thanks to knowledge of the sowing pattern as prior information. DL method is based on the Faster RCNN model trained over 2/3 of the images selected to represent a good balance between plant development stage and sessions. One model is trained for each crop.

Results show that DL generally outperforms HC, particularly for maize and sunflower crops. The quality of images appears mandatory for HC methods where image blur and complex background induce difficulties for the segmentation step. Performances of DL methods are also limited by image quality as well as the presence of weeds. An hybrid method (HY) was proposed to eliminate weeds between the rows using the rules used for the HC method. HY improves slightly DL performances in the case of high weed infestation. A significant level of variability of plant detection performances is observed between the several experiments. This was explained by the variability of image acquisition conditions including illumination, plant development stage, background complexity and weed infestation. We tested an active learning approach where few images corresponding to the conditions of the testing dataset were complementing the training dataset for DL. Results show a drastic increase of performances for all crops, with relative RMSE below 5% for the estimation of the plant density.

2 Introduction

Plant density at emergence is a main yield component particularly for plants with reduced tillering or branching capacities such as maize, sugar beet and sunflower. The plant density at emergence is controlled by the seeding density and the emergence rate. Further, the seeding pattern defined by the distance between row and between plants influences the competition between plants and possibly with weeds. In addition to the estimation of plant density, the position of each plant can be documented to describe the local competitive environment (Godwin and Miller, 2003).Plant count is mainly used for experimental trial where it assess the quality of the microplot and thus the trial. It is also used by farmer to take decision of re-sowing the field in case of too much heterogenity. For both cases this measurement allows to save money by not continuing to grow plots/fields of low quality.

Plant density is currently mostly evaluated manually: operators count plants over a limited sampling area since this process is tedious, time-consuming and therefore expensive. Consequently, this method can lead to significant uncertainties due to the limited representativeness of the sampled area and human errors. Further, the position of plants is generally not documented because it would be even more tedious to measure each plant location.

The recent technological advances of crops phenotyping solutions including Unmanned Aerial Vehicles (UAV), sensors, computers and image processing algorithms, offer potentials to develop alternative methods to the manual counting. Several authors already reported accurate estimates of plant density from images acquired a few days after emergence, using high-resolution RGB image (Guo, Rage and Ninomiya, 2013), or vegetation indices (Liu *et al.*, 2008; Sankaran *et al.*, 2015). Plants are characterized either with handcrafted (HC) methods where a computer vision expert designs techniques to extract the plant features, or with plant features automatically provided by Deep-learning (DL) techniques. Then, these features are exploited through two main approaches: regression and localization. Regression approaches provide only the plant count number, using HC features (Liu *et al.*, 2008; Guo, Rage and Ninomiya, 2013; Sankaran *et al.*, 2015) or DL features (Lu *et al.*, 2017; Ribera *et al.*, 2017; Xiong *et al.*, 2019, 2020), while localization approaches provide additionally the position of each plant. Plant localization documents the sowing heterogeneity including missing plants, allowing to explore the competition between plants as outlined earlier. Jin et al. and Liu et al. (Jin *et al.*, 2017; Liu *et al.*, 2017) used HC methods to localize wheat plants: they classified vegetation pixels and then identified objects that were later assigned to the number of plants they contain. Koh et al. (Koh *et al.*, 2019) used an alternative algorithm to detect safflower from UAV images. Madec et al. (Madec *et al.*, 2019) demonstrated that the Faster RCNN DL model (Ren *et al.*, 2015) provides accurate localization of wheat ears with higher robustness than previous methods, including direct regression method and found a better heritability than with manual counting. DL applications to plant phenotyping are supervised learning methods, requiring large and diverse labelled datasets to converge to a generic solution. The recent progress in DL applied to detection/localization tasks beneficiated from the availability of large image collections such as ImageNet (Deng *et al.*, 2009) and COCO Dataset (Lin *et al.*, 2014) that are used to pre-train the DL model.

However, Geiros et al. (Geirhos *et al.*, 2020) raised the overfitting risk and the resulting lack of robustness associated with most DL algorithms: they can reach excellent performances for datasets similar to those used for their calibration, while often failing when applied to cases different from the training dataset. In comparison, HC methods are based on expert knowledge which select the main features to identify the target objects. This prevents the risk of overfitting but can hardly account for all specific cases. Performances of HC and DL methods are generally calibrated and evaluated over a single dataset, preventing to draw general conclusions on their accuracy, scalability and robustness.

The objective of this study is to evaluate the performances of state of the art HC and DL approaches to estimate plants position and density for maize, sugar beet and sunflower crops observed after emergence using high-resolution RGB images taken from UAVs. This will be achieved over an extended dataset including diverse image acquisition conditions, crop state, and stages. The dataset contains a total of 16247 labelled plants. An extra care on diversity was made as 3 species and 3 stages where sampled as followed. Further, we propose an hybrid (HY) method that takes advantage of complementing the DL expert knowledge.

3 Materials and methods

3.1 Dataset

3.1.1 Experiments

The dataset used was acquired over maize, sugar beet and sunflower from 2016 to 2019 in several experimental sites in France (Table 1). Chosen sites cover a large diversity of agronomic modalities. Therefore the number of plots vary from one site to another to get the same number of plants per site. In each site, a field trial included a set of microplots covering a large range of genotypic was conducted. For maize, measurements were acquired by HIPHEN and Arvalis (The French technical institute for Maize) over 9 sites with contrasted microplots sizes, row spacings (0.3-1.1m), plant density (5.1-11.2 plt.m^{-2}), and genotypes. A total number of 51 microplots was available. The sugar beet experiments took place in ITBtrials (French Technical Institute for sugar beet) in the North-East of France in 2017 and 2018, with similar farming practices and different genotypes. Microplot size, row spacing and plants density were varying within a small range (Table 1). Some sites were sampled two times to get a variation in the crop stage. A total number of 60 microplots was available. Eight sunflower trials were conducted by INRAE near Toulouse at the Phenotoul-Agrophen platform in 2017, 2018 and 2019, with an additional trial conducted by TI (Terres Inovia, the French technical institute for sunflower) in 2019 near Dijon. Note that five sessions were acquired in 2018 over the same trial to have a large variability of growth stages during the measurements. Microplot size, row spacing plant density and genotypes showed a large variability within the 78 microplots available.

Crop	Site Name	Year	Nb. sessions	Nb. microplots	Owner	Lat (°)	Long (°)	Microplot width (m)	Microplot length (m)	Row spacing (m)	Plant density (plt.m^{-2})
maize	Menainville	2016	1	6	Arvalis	47.92	1.46	2.2	7.0	1.10	5.1
	Nerac	2016	1	8	Arvalis	44.16	0.3	1.6	7.0	0.80	8.5
	Villedieu	2016	1	6	Arvalis	47.88	1.53	0.9	11.0	0.30	19.9
	Thenay	2017	1	6	Arvalis	47.38	1.28	4.4	6.0	0.63	7.3
	Blois	2019	1	7	Hiphen	47.74	1.21	1.7	7.0	0.83	9.5
	Castetis	2019	1	5	Arvalis	43.46	-0.71	2.8	4.0	0.70	11.2
	Ermine	2019	1	4	Hiphen	46.54	-1.06	3.2	5.5	0.80	8.6
	Selommes	2019	1	7	Hiphen	47.74	1.21	1.8	5.3	0.88	9.5
	Pleinefougeres	2020	1	2	Hiphen	48.56	-1.57	3.2	11.0	0.80	7.7
sugarbeet	Bucy	2017	2	7	ITB	49.67	3.91	1.4	6.2	0.45	11.1
	Charmont	2017	1	7	ITB	48.39	4.18	1.4	5.5	0.45	11.1
	Etienne	2017	1	6	ITB	49.28	4.39	1.2	7.6	0.40	15.6
	Memmie	2017	2	6	ITB	48.9	4.39	1.4	7.6	0.48	10.8
	Charmont	2018	2	8*	ITB	48.39	4.18	1.4	5.5	0.45	11.4
	Memmie	2018	1	6	ITB	48.96	4.39	1.4	7.6	0.45	11.4

sunflower	Rivière	2017	1	8	INRAE	43.5	1.53	3.0	4.1	0.50	7.1
	Auzeville	2018	2	3	INRAE	43.52	1.5	3.3	9.5	0.55	6.1
	Auzeville	2019	5	12	INRAE	43.52	1.5	2.9	9.0	0.96	3.7
	Epoisses	2019	1	4	TI	47.24	5.11	2.4	10.0	0.60	5.1

*For the first session, only 7 microplots are available

Table 1. Characteristics of the sites. For sugar beet, microplots from one session to another are the same. In the case of sunflower the microplots considered change between sessions.

3.1.2 Acquisition and labelling details

Field acquisitions were carried out by UAVs embedding three different RGB cameras: Sony Alpha 5100, Sony Alpha 6000, both with a resolution of 6024x4024 pixel, and the Zenmuse X7 (DJI) in the case of Epoisses site in 2019 with a resolution of 6016 x 4008 pixels. The cameras were fixed on a two axes gimbal to maintain the nadir view direction during the flight. The camera was set to speed priority of 1/1250 s to limit motion blur. The aperture and ISO were automatically adjusted by the camera. The camera was triggered by an intervalometer set at 1Hz frequency corresponding to the maximum value allowed to record the RGB images in JPG format on the memory card of the camera. Flights altitude above ground varied between 20 and 50m to get a ground sampling distance between 2 mm and 5 mm per pixel (Table 2). The flight trajectory was designed to ensure more than 70% overlap between images across and along tracks. Ground control points were placed in the field and their coordinates were measured with a real-time kinetic GPS device ensuring an absolute centimetric accuracy of their position

Session_name	plant number	plot number	Stage	ground resolution (mm)	typical BB size (cm)	typical BB size (pixel)	Weed infestation	Blur
Selommes_2019_1	510	7	1	3.5	6.5	26	2	233
Hermine_2019_1	542	4	1	3.5	7.8	31	1	79
Thenay_2017_1	617	6	1	2.5	8.5	34	1	1149
Castetis_2019_1	575	5	2	3.3	10.0	40	1	121
Pleinefougeres_2019_1	504	2	2	3.5	11.5	46	0	39
Blois_2019_1	579	7	2	3.3	12.3	49	1	346
Menainville_2016_1	620	6	3	3.4	12.3	49	1	78
Villedieu_2016_1	629	6	3	2.7	13.3	53	0	261
Nerac_2016_1	594	8	3	4.0	15.0	60	0	37
Total maize	**5170**	**51**						
Memmie_2017_1	667	6	1	4.5	8.0	32	0	26
Charmont_2018_1	556	7	1	4.2	11.5	46	0	93
Memmie_2018_1	602	6	1	4.3	11.5	46	0	77
Bucy_2017_1	634	7	2	5.3	12.8	51	0	25
Memmie_2017_2	679	6	2	5.7	14.8	57	0	72
Etienne_2017_1	635	6	2	4.5	16.0	64	0	27
Charmont_2017_1	669	8	3	3.4	20.5	82	0	191
Charmont_2018_2	647	8	3	4.1	20.5	82	0	102
Bucy_2017_2	558	6	3	4.5	23.0	92	0	31
Total sugar beet	**5647**	**60**						
Auzeville_2019_1	579	12	1	5.0	8.5	34	1	28
Auzeville_2019_2	640	12	1	5.0	13.5	54	1	510
Epoisses_2019_1	596	4	1	2.5	14.3	57	1	10

Auzeville_2018_1	596	3	2	2.3	14.3	57	1	488
Auzeville_2019_3	657	12	2	5.0	19.3	77	0	350
Auzeville_2019_4	603	12	2	5.0	24.5	98	0	221
Rivière_2017_1	634	8	3	5.2	25.0	100	2	42
Auzeville_2018_2	560	3	3	2.6	27.5	110	1	1286
Auzeville_2019_5	565	12	3	5	27.5	110	2	176
Total sunflower	**5430**	**78**						

Table 2. Characteristics of each measurement sessions. The typical bounding box (BB) size in pixels is computed after up-sampling the images at 2.5 mm resolution. The plant stage at the time of the session is quantified as: 1: early, 2: intermediate, 3: late. The weed infestation is scored from 0 (no weed) to 2 (many weeds observed). The image blur is quantified by the average variance of the Laplacian: high blur results in low value of the variance of the Laplacian.

Agisoft Photoscan Professional software was used to align the images. The high overlap between the images and structure from motion algorithm permit to compute the position and orientation of the cameras. The pipeline described in Jin et al. (Jin *et al.*, 2017) was then run to extract from each image the portion corresponding to the contained microplots. Using the original images avoids the possible distortions and artefacts observed in the orthomosaic. Several extracts may represent the same microplot viewed from different positions of the UAV. For each microplot, we selected the sharpest extract that contained the whole microplot. For each session, we randomly selected a few microplots for labelling (Table 1). Approximately 600 plants per session were labelled to ensure a consistent size of labelled data across sessions, leading to a various number of labelled microplots: from 2 for Pleinefougeres_2019_1 to 12 for Auzeville_2019_1. The dataset contains a total of 16247 labelled plants. We up-sampled images to get the same ground sampling distance (GSD) of 2.5 mm corresponding to the best resolution (Table 2). Then all images were labelled using the coco-annotator tool (Brooks, 2019) The labelling consisted in drawing a bounding box (BB) around each plant. Six different operators contributed to the labelling. The labelling from one operator was always reviewed at least once by a different operator.

The typical size of the BB for one session is computed as the square root of the mean area of all the BBs. This typical size is used to classify the sessions according to the growth stage of the plants (Table 2): 1 ,2 and 3 correspond respectively to early, intermediate, and late stages. The level of weed infestation was also visually evaluated from 0 (no weeds) to 2 where many weeds were observed. The level of blurriness for each session was evaluated by calculating the average variance of the Laplacian (Bansal, Raj and Choudhury, 2016).

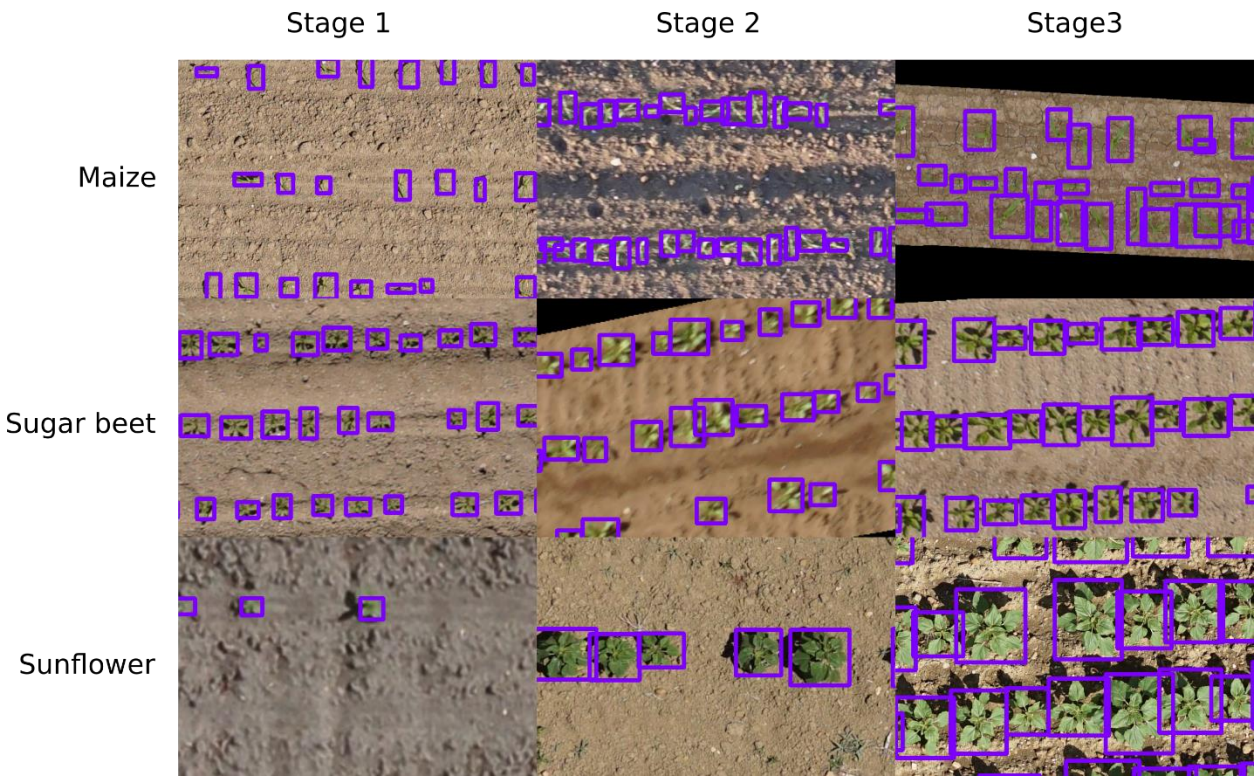

Figure 1: Samples of images for the three development stage. All images were resampled to 0.25mm.px^{-1}. The bounding boxes were drawn interactively around the plants.

3.2 Plant detection methods

3.2.1 Handcrafted method

The method developed here is based on several assumptions: (1) the plants are green and can be accurately separated from the background; (2) plants are sown in rows relatively evenly spaced and parallel; (3) the weeds are mainly located in between the rows and are not too dominant; (4) plants are relatively evenly spaced on the row and are not too variable in shape and size. The method first extracts each single row and then identify each individual plant on the row. The parameters of the below-described algorithm (**Erreur ! Source du renvoi introuvable.** and **Erreur ! Source du renvoi introuvable.**) are expressed in relative value to the row or plant spacing, to allow adaptation to a larger number of sowing patterns.

3.2.1.1 Row extraction

The original RGB images are first transformed into a black and white one (BW) using the excess green index (ExG =(2G-B-R)/G) where R, G, B correspond respectively to the red, green and blue colours of the original image (Meyer and Neto, 2008). Pixels are then assigned to the green (1) or background (0) classes using the ExG threshold value defined with the Otsu algorithm for each session (Otsu, 1975).

On this BW image an aperture morphological operation is applied to eliminate isolated small groups of pixels corresponding to small weeds or remaining artefacts of the binarization of the image. A closure morphological operation merges objects along the row and make rows more continuous. The Hough transform (Hough, 1962) is used to identify the main alignments corresponding to the rows and find their orientation. The image is then rotated to display the rows horizontally (Figure 2). The number of green pixels in each line is computed to obtain a profile of green pixels across the rows. The peaks

of the green pixel profiles are localized using the prior knowledge on row spacing (*Row_spacing_prior*) to prevent finding unexpected peaks between rows. The prior knowledge of the number of rows per microplot (*Row_number prior*) is also used when identifying the peaks. The prior values of row and plant spacing are not always known precisely. Therefore the row extraction pipeline (Figure 2) provides also updated and more accurate values of *Row_spacing_prior* for each session. Finally, each row is extracted using the fine tuned value of the row width.

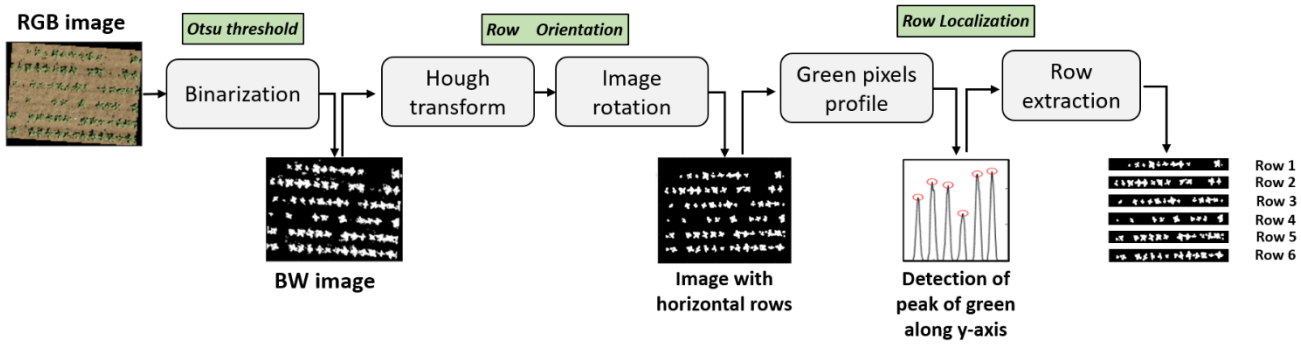

Figure 2. Flowchart of the rows extraction process from the original RGB image.

3.2.1.2 Plant identification with an object method

After the row extraction, the algorithm individualizes the objects (groups of connected pixels) in the image and classifies them as plants or weeds. Weeds that were not eliminated by the previous aperture operation are eliminated based on the distance to the row center: if the centroid of an object is located at a distance larger than a threshold value (*Minimum_distance_to_row)*, it is considered as a weed. The threshold value is expressed in relative value to the row spacing and set to 0.25 (Table S1). Objects with dimension along the row direction larger than the *Plant_spacing_prior* value (Table 1) are expected to include several plants. The number of plants contained in these big objects is derived from the number of peaks observed when summing the green pixels along the row direction: a peak may correspond to a plant position. Further, the number of plants found by the number of peaks is crosschecked with the expected number of plants computed by dividing the extension of the object by the *Plant_spacing_prior* value. Results are illustrated in Figure 3 for the two objects on the right of the bottom row.

Finally, some objects may be located too close together to be considered as separate plants: these objects correspond to several parts of the same plant. Figure 3 illustrates it with the second plant starting from the left on the top row, where a leaf and the main plant are separated. If the distance between the centroids of the closest object is smaller than the maximum acceptable distance, *Big_plants_tolerance x Plant_spacing_prior*, the two objects are merged as a single plant. Table S1 in the supplementary materials presents the value used for each parameter. The centroid (center of mass of the object), and the bounding box (smallest rectangle that contains all object's pixels) of the objects are finally computed.

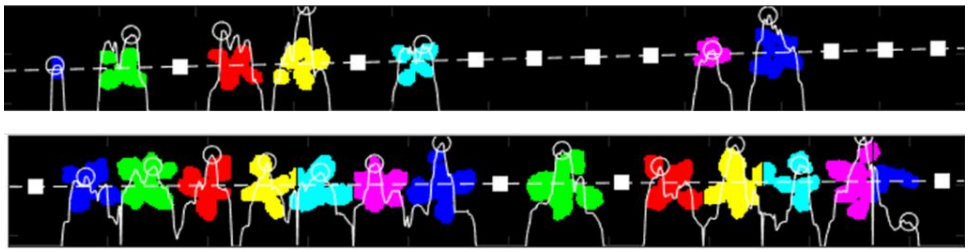

Figure 3. Typical output of the HC algorithm illustrated for two sugar beet rows. The dashed white line indicates the row. The white curve represents the profile of number of green pixels perpendicular to the row, with peaks identified by a circle. The object method is illustrated by the colours assigned to each identified plant. Note that big objects have been split into individual plants (bottom row, the four last plants) and isolated plant parts have been reconnected to form a single plant (top row, fourth plant starting from the left). The white squares correspond to the position of missing plants.

3.2.2 Deep-learning method

3.2.2.1 Model architecture

We selected an object detection method to predict the bounding box around each plant. This information can then be used to derive more traits to characterize every individual plant. Object detection is a fast-growing area within DL techniques since the emergence of networks such as R-CNN(Girshick *et al.*, 2013) or SSD (Liu *et al.*, 2016). Most DL object detection models fall into one-stage or two-stage models. In the one-stage model, the object is localized and categorized in a single step. In the two-stage model, a first stage detects possible objects and a second stage categorizes them. We used the Fast-RCNN two-stage model (Ren *et al.*, 2015) because it performs well in the context of plant phenotyping. Madec et al. (Madec *et al.*, 2019) used it successfully for counting wheat heads. It allows also to analyze the nature of the possible errors by visualizing them.

In our implementation, a backbone network computes a feature map from the original image. The ResNet-34 model was used as the backbone network because it offers a good compromise between accuracy and speed of training. A Feature Pyramidal Network (FPN) is then used to provide object proposition at different scales. All other architectural details come from the default configuration provided by the library mmdetection (Chen *et al.*, 2019) based on PyTorch (Paszke *et al.*, 2019). We also choose to train one model by crop: preliminary tests show lower performances when mixing the three crops datasets.

3.2.2.2 Pre-processing and data augmentation

We set the input image size of the network to 512 x 512 pixels to match memory constraints during training. However, images from the microplots are larger. A preprocessing step first splits them randomly into patches of 512 x 512 pixels. For each session in the training dataset, we randomly selected 100 patches which results in a total of 900 patches to train the model for each crop over the nine available sessions. Randomly sampled patches provide more diverse than regurly sampled with a fixed overlap. Data augmentation is applied to artificially extend the diversity of situations by applying random geometric distortions, blur, noise, changes in hue value and contrast. Once trained, the model is applied to all the patches. Predictions from the overlapping patches are finally merged together by using the Non-Max-Suppression algorithm (Ghosal *et al.*, 2019) with an Intersection over Union (IoU) threshold of 0.7.

3.2.3 Hybrid method

DL methods are supposed to be more efficient to detect individual plants than HC methods because they implement more detailed analysis of their distinguitive features. However, HC methods exploit experts prior knowledge on the sowing pattern to eliminate plants located at a non-expected positions such as between rows. We propose therefore a hybrid method that combines the benefits of both HC and DL methods. DL method is first applied to detect plants. Then, the HC method presented earlier is

used to identify the row position and eliminate all remaining weeds corresponding to plants with centroids located at a larger distance to the row than a threshold value *distance_to_row* (Table S1).

3.3 Evaluation strategy for plant detection

3.3.1 Strategies for training and evaluation

Detection models were developed and evaluated independently for each crop. DL method requires an extensive training dataset that should represent the expected diversity of situations. Due to the limited amount of labelled images, we defined two strategies: “Out-Domain” and “In-Domain”. “Out-Domain” is the standard strategy where the performances of the DL method are evaluated over sessions not used during the training process. For each crop and each stage, we used two sessions for training and the remaining one for testing. This allows to balance the stages between the training and testing datasets. We used a three-fold cross-validation strategy that exploits all sessions while providing relatively independent test cases: three different models were trained for each crop using six sessions, representing about 3800 plants, and tested on the remaining three sessions representing around 1900 plants. The “In-Domain” strategy is based on adding few images randomly selected in the testing datatsets to the training dataset. It aims at reducing possible lack of representativeness in the training dataset. The same three-fold cross-validation process was used for each crop, except that 1/3 of the 600 plants used previously as testing datasets were added to the training dataset. The remaining 2/3 images (400 plants) are used to evaluate the performances of the models for each. The same test data set (1200 plants corresponding to the 400 test plants for each of the three test sessions) is finally used to compare the Out-domain and In-domain approaches.

3.3.2 Evaluation metrics

Detection

Two metrics were used to evaluate wether a plant was correctly detected: the Intersection over Union matching strategy (IoU_MS) and the “Centroid matching strategy” (C_MS). The IoU_MS is based on the Intersection over Union between the detected and labelled BB with a standard threshold of 0.5. A detected plant is considered true positive (TP) if its IoU is larger than 0.5. Otherwise, it is a false positive (FP). If a labeled BB has no overlap with any detected BB, it is classified as false negative (FN). The C_MS is based on the distance between the centroids of the plants. If the distance between centroids of the detected plant and the closest labelled one is smaller than *Plant_distance_prior* / 2 it is considered as TP. Otherwise it is a FP. If a labelled plant has no detected plant within a distance smaller than *Plant_distance_prior* / 2, it is a false negative (FN).

The plant detection performance was quantified per session with the terms of the confusion matrix normalized by the number of labelled plants (TP+FN) for easier comparison between crops and stages: they correspond to rates of TP (TP%), FP (FP) and FN (FNR). We additionally used the accuracy as TP/(TP+FN+FP). For the DL and HY methods, these metrics were computed for a score of 0.5 associated to the BB.**Plant density**

Plant density (PD) was calculated by dividing the number of plants in the microplot by its area. The area is computed as the number of rows multiplied by the row spacing. The relative root mean square error (rRMSE) is used to compare the estimated and the reference PD values and assess the accuracy of the method. Additionally, to trust the estimation breeders and farmers requires a robust predictions.

Therefore, we introduced a robustness metrics based on the frequency of microplots achieving better rRMSE than standard values: rRMSE<5% were considered as robust, between 5%<rRMSE<10% as satisfactory, between 10%<rRMSE< 20% as poor, and rRMSE>20% as very poor.

4 Results and discussion

4.1 Centroid matching strategy (C_MS) is preferred to the IoU one (IoU_MS)

the BB of plants detected by the HC method have different dimensions as compared to the labelled BB (Figure 4, left): The distribution of the size of BB for HC is gaussian, while that of labelled, DL and HY are very similar and skewed with significantly smaller BBs as well as larger ones. This resulted into lower values of accuracy computed with IoU_MS (Figure 4, right) because of a significant amount of mismatch between the predicted and reference BBs at IoU=0.5. Rather than adapting the IoU threshold level, we preferred using the distance between centroids to evaluate the match between predicted and interactively labeled plants. The accuracy computed with C_MS (Figure 4, right) is significantly larger than that computed with IoU_MS, particularly for the low accuracy values as well as for the HC method for the reasons exposed above. Therefore, in the following, we will use the centroid distance to compute the terms of the confusion matrix and the accuracy. Detailed metrics can be found in Table S2.

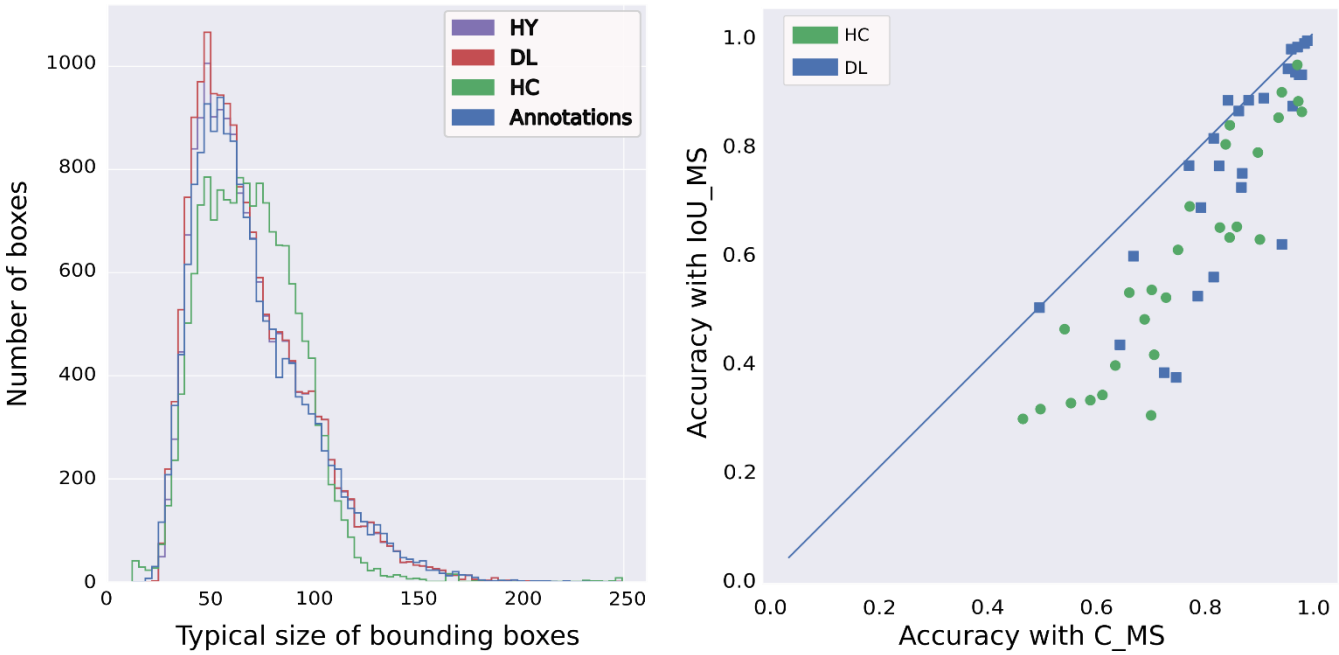

Figure 4: Left: distribution of the typical size of BB annotated interactively and those defined around the plants identified by the HC method. Right: comparison of Accuracy computed either with IoU_MS, and with C_MS for HC (green discs), and DL methods (blue squares).

4.2 DL and HY methods detect better plants than the HC one

Performances are very different depending on the crops (Table 3 and Figure 5). Detection of maize plants appears difficult for the three methods and particularly for HC with a low rate of true positive and a high rate of false negative (Table 3). This is mostly explained by the small size of the plants that overlap: groups of overlapping plants are interpreted as a sigle plant (Figure 6, HC of the green segmentation used in the first step to identity objects (Figure 6a). However, a high rate of false positive

is also observed for the first stage with HC method, due to the poor quality of the green segmentation where background artifacts were interpreted as plants (Figure 6b). A large variability between the three instances of the three-fold cross validation is observed for this early stage (Figure 5) due to the variability in image quality. Marginal differences are observed between DL and HY methods. They both show relatively balanced FPR and FNR. False positives are mostly explained by possible confusions between plants and their shadows or soil artifacts (Figure 6c) while false negatives are explained by the small size of the plants that are difficult to detect (Figure 6d). The resulting in accuracy values between 0.77 to 0.80, with little variation between stages (Table 3). However, a larger variability across the three instances of the three-fold cross validation is observed for the late stage (Figure 5).

Crop	Stages	N	TPR			FPR			FNR			Accuracy		
			HC	DL	HY	HC	DL	HY	HC	DL	HY	HC	DL	HY
Maize	1	1669	0.61	0.88	0.86	0.27	0.12	0.07	0.39	0.12	0.14	0.56	0.79	0.80
	2	1658	0.70	0.92	0.92	0.03	0.18	0.16	0.30	0.08	0.08	0.68	0.78	0.79
	3	1930	0.70	0.88	0.86	0.05	0.15	0.14	0.30	0.12	0.14	0.68	0.77	0.77
Sugar beet	1	1825	0.95	0.98	0.98	0.04	0.01	0.01	0.05	0.02	0.02	0.93	0.97	0.97
	2	1948	0.95	0.99	0.99	0.01	0.03	0.03	0.05	0.01	0.01	0.93	0.97	0.97
	3	1874	0.94	0.99	0.99	0.06	0.04	0.04	0.06	0.01	0.01	0.88	0.95	0.95
Sunflower	1	1603	0.80	0.87	0.86	0.17	0.06	0.04	0.20	0.13	0.14	0.75	0.80	0.81
	2	1856	0.82	0.94	0.94	0.15	0.08	0.07	0.18	0.06	0.06	0.72	0.87	0.88
	3	1759	0.86	0.97	0.97	0.42	0.43	0.21	0.14	0.03	0.03	0.61	0.74	0.81

Table 3: Terms of the confusion matrix for the three methods the three crops, and the three stages. True Positive Rate (TPR), False Positive Rate (FPR), and False Negative Rate (FNR) are displayed. N is the true number of plants (N=TP+FN). Green color corresponds to good metrics values (high for TPR, low for FPR and FNR), and red for poor metrics values (low for TPR, high for FPR and FNR).

Detection of sugarbeet plants appears to be much easier, with performances similar between the three methods. A small rate of false positive is observed for the three methods, particularly for the latest stage, which explains the decrease in accuracy (Table 3). This is due to difficulties when plants are overlapping (Figure 6e). Slightly more false negative are observed for HC corresponding to non-detected plants in the case of small plants and image of poor quality. This is also observed with DL for the very early stages (Figure 6f). The variability across the three instances of the three-fold cross validation is also small (Figure 5). Marginal differences are observed between DL and HY methods.

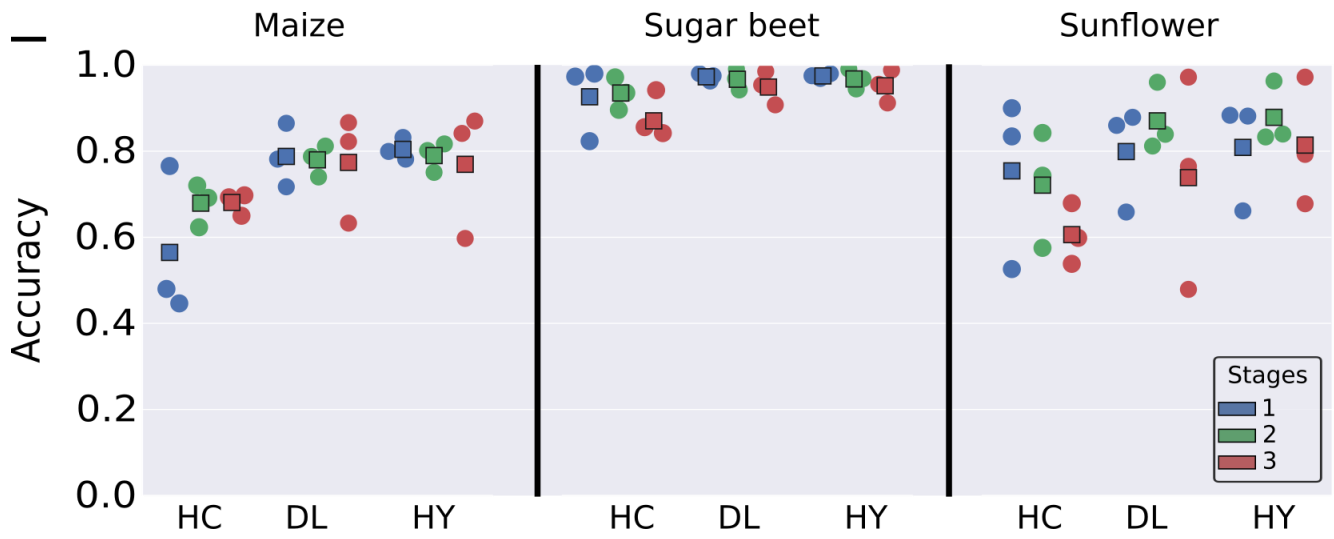

Figure 5: Accuracy for all methods and crops. For each crop and method, stages are represented by a specific color. Each point corresponds to a test session used in the three-fold validation process. The squares represent the average of the three points .

Detection of sunflower plants shows accuracy values intermediate between maize and sugar beet (Table 3 nd Figure 5). The HC shows lower TPR and higher FPR and FNR as compared to DL and HY. In the late stage, the HC shows very high rate FPR corresponding to problems of plant separation when they are overlapping. Further, the weeds close to the row line are not well eliminated and confounded with plants (Figure 6g). Similar problems are observed for the DL method, with weeds confounded with the crop. However, the HY methods allows to eliminate part of the weeds that are located in between rows (Figure 6h). and HY shows high and similar TPR (Table 3). However, larger FNR are observed for the first stage where background artifacts were detected as plants. Conversely, high FPR are observed for the late stage where DL shows difficulty to detect plants in a group of overlapping ones and confounds weeds with the crop. A large variability between the three instances of the three-fold cross validation is observed for sunflower (Figure 5). It is explained by a high degree of heterogenety in the microplots and between them, as well as between sessions.

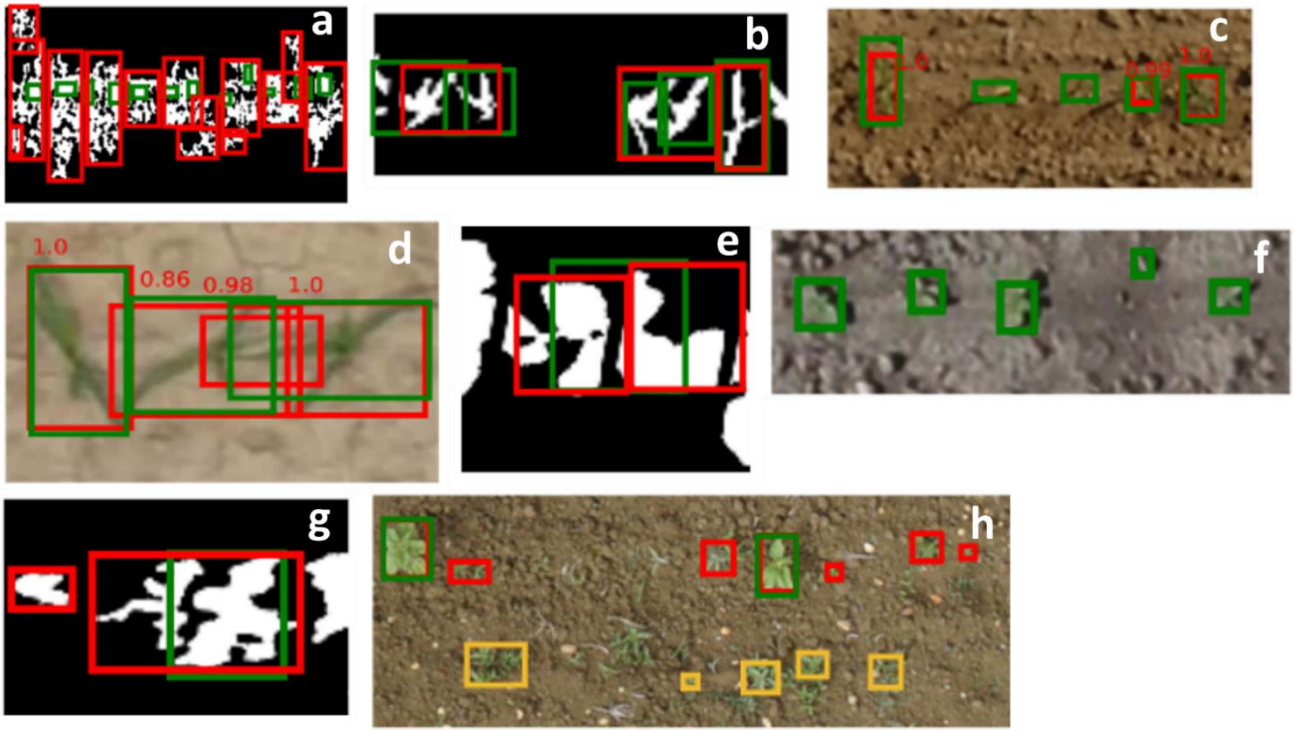

Figure 6: Possible detection errors for HC and DL methods. green BBs correspond to labelled plants. Red BBs correspond to the detected plants and yellow boxes correspond to weeds detected as crop. RGB images are displayed for DL method.BW images are displayed for the HC method. A, b, c, d corresponds to maize, e, f, to sugar beet and g, h to sunflower.

Image quality appears therefore mandatory for HC methods to get a good segmentation. HC methods appears also limited to eliminate weeds on the rows and to separate efficiently the overlapping plants. DL methods are also limited in separating crops from weeds. Additional images would have been necessary to get a better model training for weed detection or elimination. Data augmentation techniques could have been used by artificially pasting weed images in actual images. However, the HY methods allows to eliminate part of the weeds. DL methods show also some difficulties in detecting plants when they are small or when their shadows or other soil artifacts such as cracks are present. Nevertheless, DL methods outperform HC ones in most cases.

4.3 Plant density is better estimated with DL and HY methods

HC provides the poorest performances of plant density estimation for maize, with rRMSE generally higher than 0.2 (Figure 7), which is consistent with the detection performances (Figure 5). Image acquisition during the early stages tends to degrade the performances conversely to what was observed for the detection (Figure 6). This may be explained by the balance between false positives and negatives (Table 3). Marginal differences are observed between DL and HY methods for maize where weeds were not the main issue.

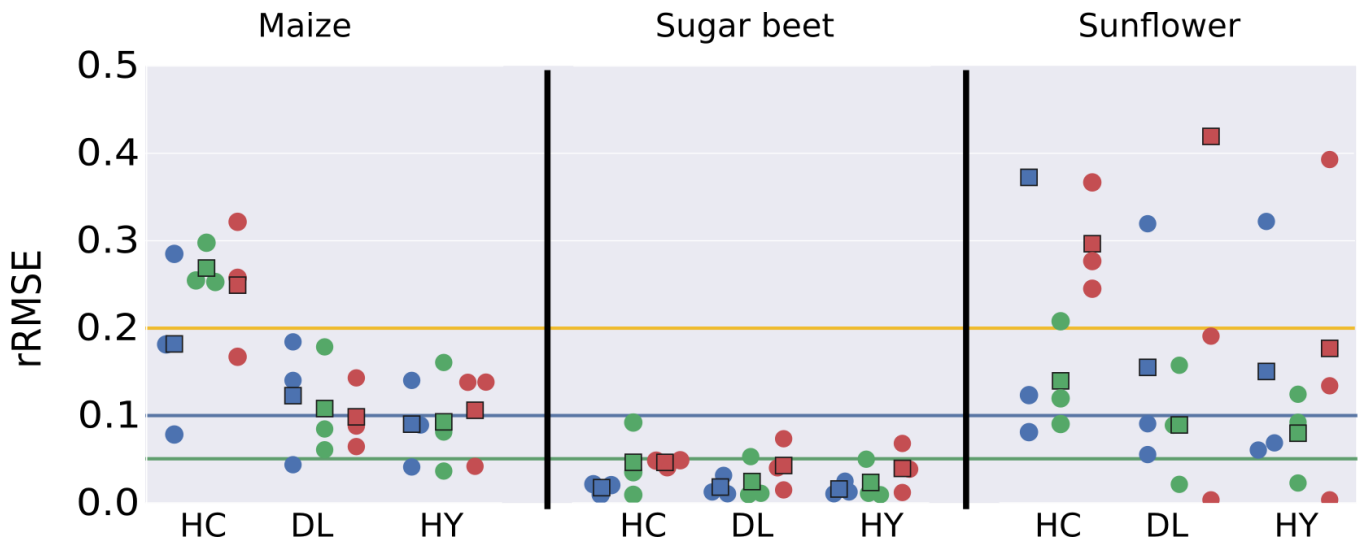

Figure 7: rRMSE for plant density estimation for all methods crops, and stages. Results obtained over the testing dataset. For each crop, method and stage, the three instances (corresponding to three testing sessions) of the three-fold cross validation process are displayed as colored disks, while the corresponding average is represented by a colored square. Colors correspond to stages.

All the methods reach the desired breeders performances (rRMSE<0.05) for sugarbeet, with even better performances for the two first stages when plants are easilly identified and weeds not too developped (Figure 7). The poorer detection performances noticed earlier for HC (Figure 5) do not impact the density estimation because the false positives are well compensated by the false negatives.

Sunflower shows more variability between sessions and stages, with rRMSE around 0.1 for the intermediate development stage showing better performances than the early one and moreover than the late one (Figure 7). The models for sunflower are very poor for the session 3_auzeville_2019_5 (Figure 7), mainly because of weeds infestation. DL performs better than HC while HY improves marginally the performances for the two early stages, but significantly for the late stage where significant amount of weeds were observed.

4.4 Improving DL on out-domain data is more promising than adding HC rules to reach breeders requirements

The performances of DL method are closely related to the number of images used in the training database and their representation of possible situations (Geirhos *et al.*, 2020). DL method works very well for sugarbeet where all the images were relatively similar across sessions for each development stage. However, the acquisition conditions were quite different from the ones experienced in the other sessions for the sunflower on Epoisses_2019_1, explaining why the DL models were failing to detect plants for this particular session. The “Out-domain” strategy used previously was compared here to the “In-domain” one where 1/3 of the images of the initial testing sessions were used to finetune the model. Results are presented on the remaining 2/3 images of the initial testing sessions.

Note first that the density estimation performances (Figure 8) evaluated on a limited test data set (1200 images) are very consistent with the ones presented previously over the full test dataset including 1800 images (Figure 7). Results show that the additional images used in the training process and having the same characteristics as those in the testing dataset decreased significantly the rRMSE for all crops

(Figure 8). This outperforms the marginal gain observed with the HY method on few sessions. Training with the In-domain strategy reduces the variability of performances across sessions. The 5% rRMSE value expected by breeders is reached for all crops except maize, where performances are anyway close to this target. Plant overlapping and the small leaf size makes the DL method for maize more challenging. However, there are still some outliers for Maize and Sunflower, corresponding to Pleinefougeres_2019_1 and Epoisses_2019_1 sessions. The images of these two sessions are highly blurred (Table 2) explaining most of their poor detection performances.

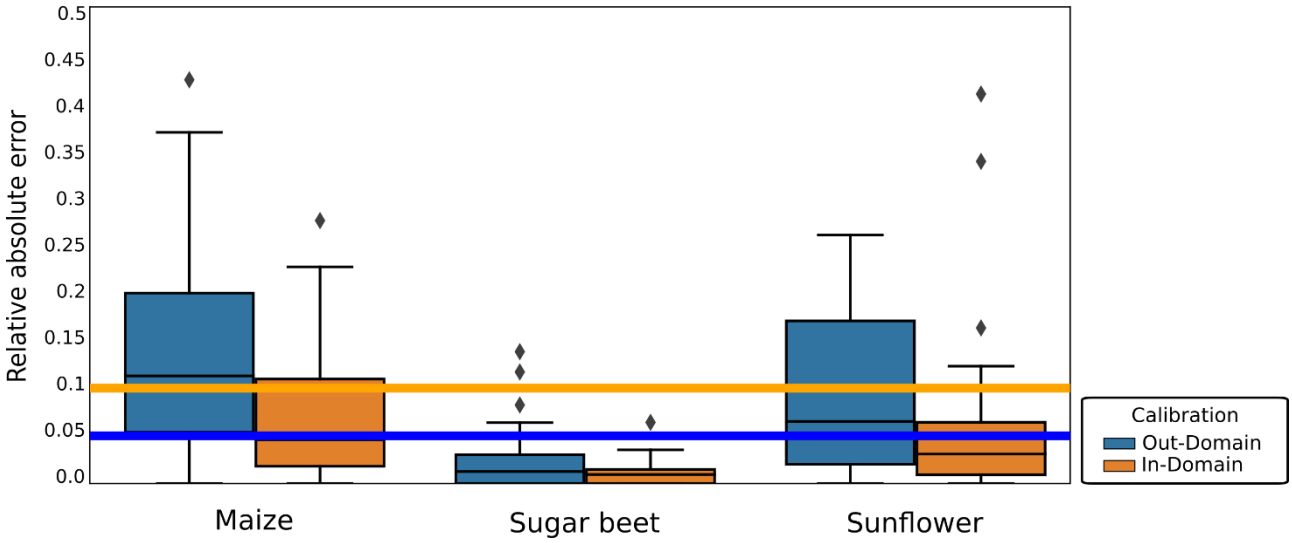

Figure 8: Distribution of relative absolute error for each microplots for the Out-Domain and In-Domain approaches for DL. Box-plot representation where the black horizontal bar represents the median, the box represents ±25%, the wiskers while the whiskers extend to the the lowest (highest) data point still within 1.5 interquantile range of the lower (upper) quartile. Diamonds are outliers. 1 outlier for Out-domain Maize and 3 outliers for Out-Domain Sunflower are above 0.5 and are not presented on the graph.

Our results demonstrate that active learning techniques (Ghosal *et al.*, 2019) could greatly improve DL model performances for these new sessions: a small sample of images coming from the new sessions to be processed has to be labelled to complement the training dataset. Alternative techniques could be also tested to bypass the tedious and expensive interactive labelling process. This includes obviously data sharing between several organizations as this was done for the head counting problem (David et al., 2020). Data augmentation (Kuznichov *et al.*, 2019) could improve greatly the generalization performances of DL methods. It would consist in manipulating the quality of the images, while creating synthetic images where a wide diversity of plants and weeds would be placed over different backgrounds with variation in the development stages and sowing pattern.

5 Conclusion

We investigated different techniques to detect plants and estimate the plant density. We first discussed the metrics used to detect plants and concluded that the distance between centroids of the detected and labelled plants provided more consistent results when the detected bounding boxes have not the same size as the labelled ones. HC needs prior knowledge on the sowing pattern, DL needs a large volume of images to be labelled, and HY exploits both sources of information. HC performs well in the case of sugar beet with high quality images, but DL outperforms the HC in most cases. However, DL fails to generalize well on new sessions different from the ones used for training. We identified the source

of errors coming either from the quality of the images, plant size or weed infestation. HY improves marginally DL performances. To reach high level of accuracy for plant detection and good estimates of plant density, the training database for DL should be very large and diverse. We demonstrated that active learning where few images from the application domain are labelled to complement the training dataset could greatly improve the performances.

6 Acknowledgments

The work received support from ANRT for the CIFRE grant of Etienne David, co-funded by Arvalis. The study was partly supported by several projects, including ANR PHENOME. Many thanks to the dataset contributors: Arvalis, Terres Inovia, Institut Technique de la Betterave (ITB), Hiphen Plant.

7 Bibliography

Bansal, R., Raj, G. and Choudhury, T. (2016) 'Blur image detection using Laplacian operator and Open-CV', in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, pp. 63–67. doi: 10.1109/SYSMART.2016.7894491.

Brooks, J. (2019) *COCO Annotator*. Available at: https://github.com/jsbroks/coco-annotator/.

Chen, K. *et al.* (2019) 'MMDetection: Open MMLab Detection Toolbox and Benchmark', *arXiv preprint arXiv:1906.07155*.

Deng, J. *et al.* (2009) 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Geirhos, R. *et al.* (2020) 'Shortcut Learning in Deep Neural Networks', *arXiv preprint arXiv:2004.07780*.

Ghosal, S. *et al.* (2019) 'A weakly supervised deep learning framework for sorghum head detection and counting', *Plant Phenomics*, 2019, p. 1525874.

Girshick, R. B. *et al.* (2013) 'Rich feature hierarchies for accurate object detection and semantic segmentation', *CoRR*, abs/1311.2524. Available at: http://arxiv.org/abs/1311.2524.

Godwin, R. J. and Miller, P. C. H. (2003) 'A Review of the Technologies for Mapping Within-field Variability', *Biosystems Engineering*, 84(4), pp. 393–407. doi: 10.1016/S1537-5110(02)00283-0.

Guo, W., Rage, U. K. and Ninomiya, S. (2013) 'Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model', *Computers and Electronics in Agriculture*, 96, pp. 58–66. doi: 10.1016/j.compag.2013.04.010.

Hough, P. V. (1962) *Method and means for recognizing complex patterns*. Google Patents.

Jin, X. *et al.* (2017) 'Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery', *Remote Sensing of Environment*, 198, pp. 105–114. doi: 10.1016/j.rse.2017.06.007.

Koh, J. C. O. *et al.* (2019) 'Estimation of crop plant density at early mixed growth stages using UAV imagery', *Plant Methods*, 15(1), p. 64. doi: 10.1186/s13007-019-0449-1.

Kuznichov, D. *et al.* (2019) 'Data Augmentation for Leaf Segmentation and Counting Tasks in Rosette Plants', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA: IEEE, pp. 2580–2589. doi: 10.1109/CVPRW.2019.00314.

Lin, T.-Y. *et al.* (2014) 'Microsoft coco: Common objects in context', in *European conference on computer vision*. Springer, pp. 740–755.

Liu, J. *et al.* (2008) 'Crop fraction estimation from casi hyperspectral data using linear spectral unmixing and vegetation indices', *Canadian Journal of Remote Sensing*, 34(sup1), pp. S124–S138. doi: 10.5589/m07-062.

Liu, S. *et al.* (2017) 'A method to estimate plant density and plant spacing heterogeneity: application to wheat crops', *Plant methods*, 13(1), p. 38.

Liu, W. *et al.* (2016) 'Ssd: Single shot multibox detector', in *European conference on computer vision*. Springer, pp. 21–37.

Lu, H. *et al.* (2017) 'TasselNet: counting maize tassels in the wild via local counts regression network', *Plant methods*, 13(1), p. 79.

Madec, S. *et al.* (2019) 'Ear density estimation from high resolution RGB imagery using deep learning technique', *Agricultural and Forest Meteorology*, 264, pp. 225–234. doi: 10.1016/j.agrformet.2018.10.013.

Meyer, G. E. and Neto, J. C. (2008) 'Verification of color vegetation indices for automated crop imaging applications', *Computers and Electronics in Agriculture*, 63(2), pp. 282–293. doi: 10.1016/j.compag.2008.03.009.

Otsu, N. (1975) 'A threshold selection method from gray-level histograms [J]', *Automatica*, 11(285–296), pp. 23–27.

Paszke, A. *et al.* (2019) 'PyTorch: An Imperative Style, High-Performance Deep Learning Library', in Wallach, H. et al. (eds) *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8024–8035. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ren, S. *et al.* (2015) 'Faster r-cnn: Towards real-time object detection with region proposal networks', in *Advances in neural information processing systems*, pp. 91–99.

Ribera, J. *et al.* (2017) 'Counting plants using deep learning', in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1344–1348. doi: 10.1109/GlobalSIP.2017.8309180.

Sankaran, S. *et al.* (2015) 'Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review', *European Journal of Agronomy*, 70, pp. 112–123. doi: 10.1016/j.eja.2015.07.004.

Xiong, H. *et al.* (2019) 'TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks', *Plant Methods*, 15(1), p. 150. doi: 10.1186/s13007-019-0537-2.

Xiong, H. *et al.* (2020) 'From Open Set to Closed Set: Supervised Spatial Divide-and-Conquer for Object Counting', *arXiv preprint arXiv:2001.01886*.

8 Data Availability Statement

The datasets analyzed for this study will be available on Zenodo after publication.

9 Supplementary material

Rules	**Parameter name**	**Operations**	**Definition**	**values**
Get BW mask	*Excess Green threshold*	Segmentation	The threshold used to transform the image into a vegetation mask	Determined by the otsu method
	Size of smaller objects	Aperture	size used to eliminate small objects just after the segmentation for row detection	3 cm
Find row	*Row number spacing*	Aperture	Expected number of rows	Determined in Table 1
	Row spacing prior	Row detection	Prior value of the row spacing as defined in Table 1	Determined in Table 1
	Peak prior	Row detection	The fraction of the maximum height of the peaks used to consider a peak as corresponding to a row	0.5
	Plant spacing prior	Split object	Prior value of the plant spacing as defined in Table 1	Determined in Table 1
Find plant	*Minimum distance to row*	Weed elimination	Minimum distance from the row centre (expressed relatively to *Row spacing prior*) to consider the objects as weeds	0.25
Remove false positives	*Big Plants Tolerance*	Leaves detection	All centroids under *Big Plants Tolerance* x *Plant spacing prior* are considered to belong to the same plant	0.9

Table S1. List of parameters used for row extraction and plant identification

Table S2. Complete results for the three methods on all sessions. Accuracy, precision and recall are presented with the IoU matching strategy.