Group-Conditional Conformal Prediction via Quantile Regression Calibration for Crop and Weed Classification

Paul Melki
IMS, CNRS, University of Bordeaux
EXXACT Robotics
paul.melki@u-bordeaux.fr

Lionel Bombrun
IMS, CNRS, University of Bordeaux
Bordeaux Sciences Agro
lionel.bombrun@ims-bordeaux.fr

Boubacar Diallo, Jérôme Dias
EXXACT Robotics
boubacar.diallo@exxact-robotics.com
jerome.dias@exxact-robotics.com

Jean-Pierre Da Costa
IMS, CNRS, University of Bordeaux
Bordeaux Sciences Agro
jean-pierre.dacosta@ims-bordeaux.fr

Abstract

As deep learning predictive models become an integral part of a large spectrum of precision agricultural systems, a barrier to the adoption of such automated solutions is the lack of user trust in these highly complex, opaque and uncertain models. Indeed, deep neural networks are not equipped with any explicit guarantees that can be used to certify the system's performance, especially in highly varying uncontrolled environments such as the ones typically faced in computer vision for agriculture.

Fortunately, certain methods developed in other communities can prove to be important for agricultural applications. This article presents the conformal prediction framework that provides valid statistical guarantees on the predictive performance of any black box prediction machine, with almost no assumptions, applied to the problem of deep visual classification of weeds and crops in real-world conditions. The framework is exposed with a focus on its practical aspects and special attention accorded to the Adaptive Prediction Sets (APS) approach that delivers marginal guarantees on the model's coverage. Marginal results are then shown to be insufficient to guarantee performance on all groups of individuals in the population as characterized by their environmental and pedo-climatic auxiliary data gathered during image acquisition.

To tackle this shortcoming, group-conditional conformal approaches are presented: the "classical" method that consists of iteratively applying the APS procedure on all groups, and a proposed elegant reformulation and implementation of the procedure using quantile regression on group membership indicators. Empirical results showing the validity of the proposed approach are presented and compared to the marginal APS then discussed.

1. Introduction

Artificial intelligence has become an integral component of precision agriculture systems. It provides the "analytical" machinery that has allowed precision agriculture to adapt to the ever-increasing flow of data characterized by a high diversity of modalities (such as RGB images, LiDAR, text and GNSS) from multiple sources influenced by a large spectrum of natural and technical conditions. From this growing pool of raw data, machine learning algorithms, and particularly deep neural networks, have proven themselves to be the approach *par excellence* to extract useful information. This information will either be directly turned into useful insight and decisions by human actors, or will flow through fully-automated robotic pipelines in autonomous agricultural systems [30].

Complex machine learning models have replaced classical "handcrafted" models that were characterized by their well-defined interpretable features and their direct inspiration from agricultural and bio-environmental factors. Both practitioners and scientists in precision agriculture were comfortable with these classical approaches [20]. Deep learning models, on the other hand, with their complex components and architectures are not only relatively opaque to the agricultural community, but also, to a certain extent, to their own designers and developers [31, 24]. While their performance prowess has been and still is being proven in the lab and in the field, some important issues such as interpretability [31, 21, 7], generalization to new observations

and domains [28, 29, 27], robustness to noise and out-of-distribution observations [15, 17, 4], and uncertainty quantification [1, 10, 11] are yet to be solved or even understood properly. Indeed, while neural networks may be highly accurate on benchmark and test datasets, no formal guarantees on their behaviour "in the wild" can be provided to the end-user.

These shortcomings stand in the way of wider scale adoption of deep neural networks in industrialized precision agriculture solutions. The typical user, who does not fully understand the models nor is provided with guarantees on them, has difficulty in trusting the systems [7, 20].

To tackle one angle of this multi-dimensional problem, we propose to focus in this article on the problem of uncertainty quantification and control. Can we quantify the uncertainty of neural networks predictions? Can we provide valid guarantees on the performance of neural networks under real-world conditions so as to cultivate trust in systems that include these predictive models?

Conformal prediction [26, 22] offers an interesting framework for producing predictions with valid statistical guarantees, and quantifying a black box model's uncertainty [23, 5]. In a context of classification, this framework allows a predictive model to produce prediction sets for a given observation X, instead of point predictions, with guarantees that the true value Y is included in the prediction set with high probability. Concretely, given a specified error tolerance level $\alpha \in [0, 1]$, conformal prediction produces prediction sets $\mathcal{C}_{1-\alpha} \in \mathcal{Y}$ that satisfy the marginal coverage property

$$\mathbb{P}\big(Y \in \mathcal{C}_{1-\alpha}(X)\big) \geq 1 - \alpha \tag{1}$$

For example, if the user sets α to 10%, then the conformal model will produce prediction sets that guarantee that the true value is predicted 90% of the times.

Although useful and intuitive in its basic form, the original conformal approach only guarantees the results marginally; that is, on average over all observations. It does not provide any guarantees on specific subsets of observations: a property that would be quite useful in agricultural applications. Indeed, it is more important to provide performance guarantees for a given species or on the user's specific parcel and conditions rather than on average everywhere.

For this reason, "group-conditional conformal prediction" has been developed, with the aim of providing equalized coverage guarantees for all groups of individuals [25, 18]. Formally, let every individual be defined by the triplet $(X, Y, G) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G}$ where G is the group, then group conditional conformal prediction aims at producing prediction sets $\mathcal{C}_{1-\alpha,g}$ with the following group-conditional coverage guarantee:

$$\mathbb{P}\big(Y \in \mathcal{C}_{1-\alpha,g}(X) | G = g\big) \geq 1 - \alpha \qquad \forall g \in \mathcal{G} \tag{2}$$

The current article explores the application of group-conditional conformal prediction in an agricultural context; specifically, on the problem of crop and weed image classification using neural networks with the existence of auxiliary metadata describing various environmental and climatic characteristics of the image's content and context. The article's contributions can be summarized as being:

- introduction and presentation of conformal prediction methods to the agricultural community concerned by uncertainty in machine learning-based decision-making;
- application of the marginal adaptive prediction sets (APS) [19, 3] method to our classification use case, providing marginal coverage guarantees that will be shown empirically;
- simple description of the "classical" group-conditional APS approach via iterative group-specific calibration and prediction [19, 2];
- proposal of a simple and elegant alternative for a more efficient group-conditional calibration via quantile regression.

The article is structured as follows: Section 2 sets up the mathematical framework for the rest of the article then presents conformal prediction in its general form, with a focus on Adaptive Prediction Sets, a method that guarantees marginal coverage. Section 3 presents the experimental setup and the results of marginal APS on the problem of image classification into weed and multiple crops on a large dataset. The results are presented in the light of environmental auxiliary variables thus showing the insufficiency of marginal coverage for the agricultural applications of interest. Section 4 explores the group-conditional extension to conformal prediction and presents the group-conditional APS approach. A reformulation of the group quantile estimation procedure as quantile regression on group membership is then developed. The results of the classical iterative and quantile regression approaches are shown on a number of auxiliary variables chosen to form groups. Section 5 concludes with a discussion of the results and a future vision of conformal prediction, particularly for agricultural applications.

2. Conformal Prediction

2.1. Notation & Setup

Before we dive into the details of the conformal approach, we define the mathematical setup that will be used in the rest of the article. We are in a supervised learning context, whereby for each input (image) $X \in \mathcal{X}$ is associated a ground-truth class $Y \in \{1, ..., K\}$. As in a typical learning

framework, we observe a sample of N observations which we split into training and validation sets. Let $\mathcal{I}_1$ be the set of training observations. In the split-conformal framework [16], we further divide the validation set into two datasets; namely, the calibration set $\mathcal{I}_2$ and the proper test set $\mathcal{I}_3$. Note that $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$ are mutually exclusive.

We train on $\mathcal{I}_1$ a neural network classifier $\mathcal{B}$ that produces for each input $X \in \mathcal{X}$ a predicted output $\hat{y} \in \{1, ..., K\}$. We also have access to the softmax output for each class at the last layer of the neural network; we call them $p^{[1]}, ..., p^{[K]}$.

2.2. General Presentation

First formulated by Vovk *et al.* [26], conformal prediction is an uncertainty quantification and control technique based on frequentist statistics. Broadly speaking, it can be understood as a method that allows the prediction of "confidence intervals", instead of point predictions, at a specified level of significance $1 - \alpha$. These prediction sets are guaranteed to contain the true value at least $1 - \alpha$ of the times. This is the *marginal coverage guarantee* presented in Equation 1. The only condition required for the validity of these methods is the *exchangeability* of observations, which is a slightly weaker condition than the i.i.d. assumption commonly considered in statistical frameworks [22].

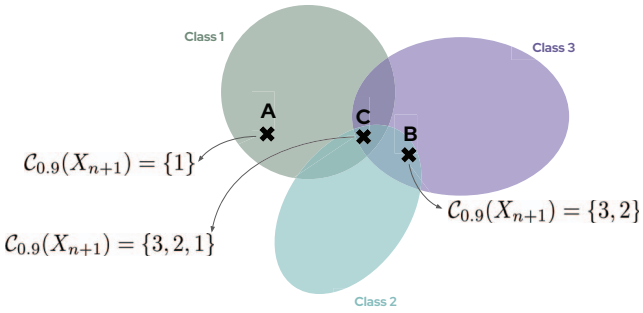

Figure 1. Representation of conformal prediction sets for three points, A, B and C, with different levels of uncertainty.

For a new input $X_{n+1} \in \mathcal{I}_3$, a conformal algorithm compares this input, using a measure of conformity (that will be defined later), to the calibration set $\mathcal{I}_2$ of observations that the conformal model has previously seen. Based on the conformity of X_{n+1}, the conformal model will be more or less confident in its prediction, as such predicting a conformal set that is more or less large in such a way as to guarantee the existence of the true value inside.

Consider the representation space shown in Figure 1 where we wish to predict conformal sets at the 90% level of confidence. If a new input X_{n+1} falls in position A, it is clearly in the domain of Class 1. The conformal model can predict with high confidence only one class while guaranteeing a high coverage at 90%. If the new input appears in the more ambiguous region at position B, then the conformal model, will produce a bigger prediction set with two classes in order to maintain the coverage guarantee at the desired 90% level. Finally, if the new input is a difficult example and falls in the region with high uncertainty at point C, then the model will predict all the classes in such a way as to guarantee predicting the true value.

2.3. Adaptive Prediction Sets (APS)

First proposed by [19] then improved and adapted to neural network classifiers in [3], the APS method not only provably achieves the marginal coverage guarantee but is also designed so that the size of the prediction sets *adapts* to the "difficulty" (think, *uncertainty*) of each example. As such, it provides both a global measure of model uncertainty and also an individual-level measure of uncertainty where bigger predicted sets indicate higher model uncertainty. The approach follows the typical split-conformal procedure of calibration and prediction:

1. CALIBRATION STEP
   After training the neural network on $x_j, j \in \mathcal{I}_1$, we now pass every individual $x_i, i \in \mathcal{I}_2$ into the network and compute its "conformity score" defined as:

$$E_i = \sum_{t=1}^{T} p_i^{(t)} \tag{3}$$

   where the softmax scores are ordered in decreasing order, t being the rank of the t^{th} class with highest softmax output, and T the rank of the ground-truth class. Accordingly, the conformity score is the cumulative softmax score of individual i until reaching its true class $y_i \in \{1, ..., K\}$. E_i is thus the cumulative pseudo-probability mass assigned to the true class by the neural network (see Figure 2(a)). In general, the bigger the probability mass, the more difficulty the neural network is having in finding the true class. For the specific case where the true class is predicted with a softmax score close to 1, see [3] for a regularized version that allows such a class not to be rejected.

   After obtaining the scores on the calibration set, we estimate $\hat{Q}_{1-\alpha}$, the $1-\alpha$ quantile of the empirical distribution of these scores, as can be seen in Figure 2(b). This quantile is the maximum score among the $1 - \alpha$ lowest scores assigned to the true class by the neural network. It will be used to construct the prediction sets in the next step of the procedure.

2. PREDICTION STEP
   For the previously unseen inputs from the prediction set $\mathcal{I}_3$, we can now construct conformal sets by passing each individual in the neural network and comparing the score $p^{(k)}$ of each class k to the estimated quantile

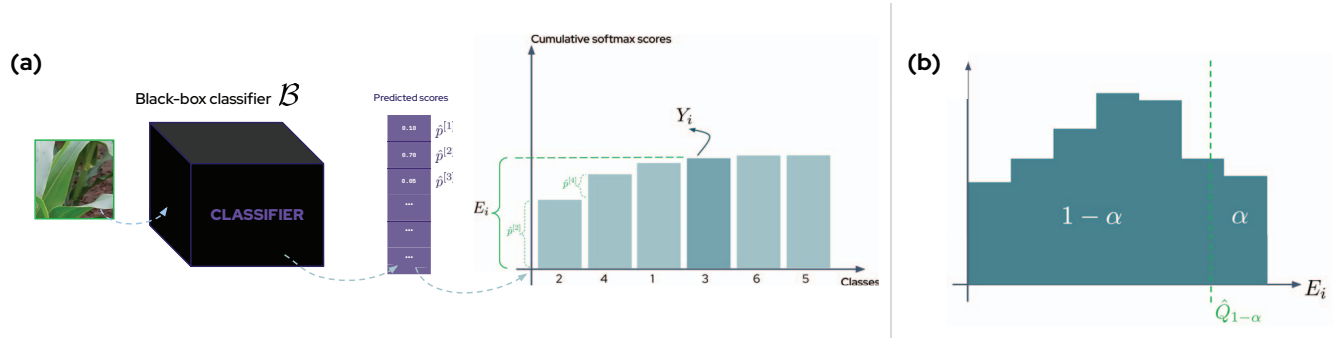

Figure 2. APS calibration: (a) Computation of the E score for a given input; (b) Estimation of the decision quantile on the empirical distribution of scores.

$\hat{Q}_{1-\alpha}$. The classes that will form the prediction set are those classes whose cumulative softmax scores do not exceed $\hat{Q}_{1-\alpha}$ (Figure 3)[1]. These are the classes that are considered "probable" enough to be predicted. The notion of "conformity score" discussed in Section 2.2 appears here: indeed, scores that are higher than $\hat{Q}_{1-\alpha}$ are considered *non-conformal* (think, "too extreme", or "too improbable"), and as such are not considered to be valid predictions.

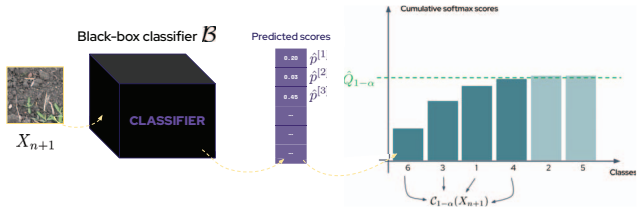

Figure 3. APS prediction.

where there is no plant. The distribution of the images over the different classes can be seen in Figure 4.

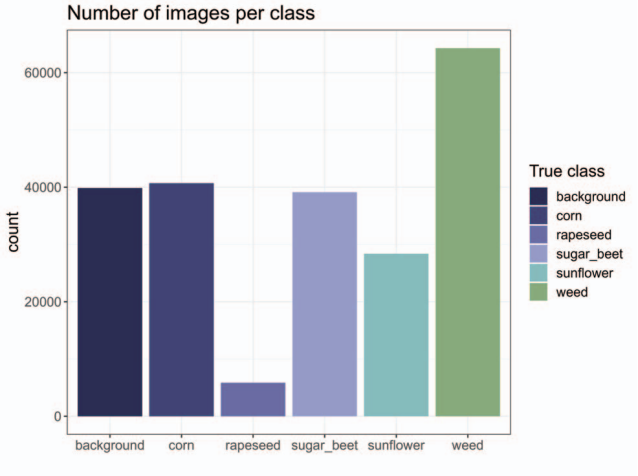

Figure 4. Distribution of images over the 6 classes.

3. Experimental Setup

3.1. Data

To demonstrate the conformal approaches on an agricultural use case, we work on a specialized proprietary dataset gathered in multiple locations around the world, under real-world uncontrolled conditions, for the problem of visual identification of crop and weed via image classification. The dataset consists of 218 thousands RGB images of size 224×224 annotated internally. Associated to each image is one of six classes specifying the crop type of the largest "object" in the image: *corn*, *rapeseed*, *sugar beet*, *sunflower* or *weed*. A final class *background* is assigned to the images

[1]This is a slight simplification of the procedure, refer to [3, 19] on the importance of randomizing the inclusion of the classes around the decision quantile.

3.2. Auxiliary Data

To each image is associated a number of auxiliary variables ("metadata") that describe different factors related to the image. Some of these factors can be considered "intrinsic" to the visual scene – that is, visible – such as some pedoclimatic characteristics like the color, texture and humidity of the soil. Other variables describe the broader environmental characteristics that may have direct or indirect influence on the image such as the conditions of the sky and the wind or the geographical location of the acquisition. These metadata are entered at the moment of image capture by the data acquirers based on their qualitative evaluation of the conditions following well-defined criteria. The visually-verifiable metadata are also reviewed during the annotation process. Other metadata such as geo-location, time and sensor conditions are automatically captured and saved.

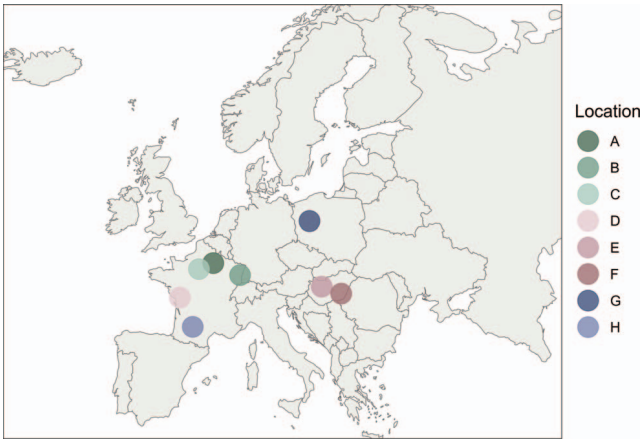

Figure 5. Locations of data acquisition in Europe.

For the purpose of the current article, with the aim of keeping the presentation as concise and clear as possible, we focus on only two auxiliary variables that are particularly interesting for practical use cases:

- *Location:* it is the location of the acquisitions as defined using GPS coordinates. From a broad perspective, our data can be divided into eight different locations across Europe denoted A to H. High-level positioning of these locations can be seen in Figure 5. Given that each location is characterized by largely different environmental and pedo-climatic conditions, this auxiliary variable can be considered a proxy for multiple other characteristics and holds high practical interest: it is important to guarantee acceptable levels of detection in all locations where the system is to be deployed.

- *Sky:* this variable represents the "perceived" condition of the sky at the moment of data acquisition. It can take one of two values, of each an example is shown in Figure 6: `overcast` (a) and `sunny` (b). The condition of the sky has an interesting impact on the visual characteristics of the image, such as luminosity, color temperature and shadows. Since the system is to be deployed in uncontrolled environments, detection results should be guaranteed regardless of the sky and ambient light.

3.3. Base Model

As mentioned previously, conformal prediction requires a base predictor that produces point predictions, which will be "transformed" via the conformal procedure into a conformal predictor producing sets of prediction points. For the purpose of this study and without loss of generality, the base classifier used is a classic ResNet18 network [9] pre-trained on ImageNet [8] and fine-tuned on our training data. It is important to note that the proposed conformal approaches are independent of the chosen base classifier. It can be any neural network architecture or other model such as random forests or support vector machines [19].

Figure 6. Examples of images taken in different *sky* conditions in the same *location*: (a) `overcast`, (b) `sunny`.

3.4. Experimental Results: Marginal APS

Method	Coverage	Set Size
Base (Top-1 Accuracy)	0.680	1.000
Marginal APS Classifier	0.896	2.566

Table 1. Comparison of Base & Conformal ResNet18 classifier.

We finetune the ResNet18 network on the training set $\mathcal{I}_1$ (50% of the database), then calibrate and predict respectively on $\mathcal{I}_2$ and $\mathcal{I}_3$ (45% and 55% of the remaining individuals) respectively following the APS procedure with an error tolerance level fixed at $\alpha = 0.1$.

Table 1 shows a comparison between the coverage obtained for the base classifier with its conformal version. The coverage of the base point predictor (which corresponds to its Top-1 overall accuracy) is 68%, with a unique set size of 1, since we only predict the top class. The APS procedure maintains the coverage exactly at the required $1 - \alpha = 0.9$ level, with an average prediction set size of 2.6. That is, by calibrating the predictive system on a dataset that resembles the population on which we want to predict and permitting the network to predict, on average, between 2 and 3 classes, we guarantee finding the true class 90% of the time.

Although the coverage is perfectly maintained marginally, the picture changes when we look at the conditional coverage per group. What if we like to guarantee the $1-\alpha$ coverage for each possible agro-environmental condition in our data?

As Table 2 shows, the coverage is not maintained at the desired level but is highly varying among the groups (Note that the group `H,overcast` is not included in the table because this combination does not exist in the data). Although the group-conditional coverage criterion defined in Equation 2 does not seem, empirically, to be violated for a number of groups, we cannot say that the condition is guaranteed since there are no explicit constraints on the estima-

Group	*Location*	*Sky*	Coverage	Set Size
1	A	overcast	0.935	2.823
2	A	sunny	0.872	2.614
3	B	overcast	0.926	1.560
4	B	sunny	0.914	1.953
5	C	overcast	0.966	2.625
6	C	sunny	0.877	2.412
7	D	overcast	0.891	2.476
8	D	sunny	0.901	2.437
9	E	overcast	0.944	2.105
10	E	sunny	0.908	2.450
11	F	overcast	0.959	2.454
12	F	sunny	0.937	2.741
13	G	overcast	0.990	2.728
14	G	sunny	0.943	2.348
15	H	sunny	0.943	2.477
Marginal APS Classifier			**0.896**	**2.566**

Table 2. Results of the Marginal APS classifier, per group.

tion of the quantile or the construction of the prediction sets in such a way as to provide such a guarantee. Indeed, for such groups as Group 2 and Group 6, the coverage is far from being maintained; while for other groups we see that the coverage is overly conservative leading to bigger prediction sets (in size) that may be required.

4. Group-conditional Conformal Prediction

The marginal coverage guarantee may not be useful in a number of use cases since it does not imply validity on all individuals; that is, conditional on their idiosyncratic characteristics. While the coverage is maintained on average, it is not guaranteed on certain groups of individuals; usually those that are not represented enough in the data [18]. In a number of use cases, such as the deployment of an autonomous weed detection system in new environments or the detection of diseases in plants, it is required to provide guarantees on all groups of individuals so that the system may be deemed reliable. Group-conditional conformal prediction has been developed for this purpose, providing the conditional coverage guarantee defined in Equation 2.

Now that we have defined the notion of auxiliary variables in Section 3.2, we can refine the definition of a "group." Assume that for each individual we observe an image X to which we associate a ground-truth label Y, and a number of auxiliary variables $\{M_L \in \mathcal{M}_L, M_S \in \mathcal{M}_S, ...\}$. An individual's group is thus defined as being its observed combination of auxiliary data: $G \in \mathcal{G}$, where $\mathcal{G} = \mathcal{M}_L \times \mathcal{M}_S \times ...$. For the sake of simplicity and without loss of generality, we assume that we only observe the two auxiliary variables *location* and *sky*. For example, one group can be defined as $G_1 = \{M_L = \texttt{A}, M_S = \texttt{overcast}\}$. We can thus provide the coverage guarantee:

$$\mathbb{P}\big(Y \in \mathcal{C}_{1-\alpha,G_1}(X) | M_L = \texttt{A}, M_S = \texttt{overcast}\big) \geq 1 - \alpha \tag{4}$$

4.1. Iterative Group-conditional APS

The "classical" approach to produce prediction sets that satisfy the group coverage guarantee consists of iteratively conducting the APS Calibration procedure described in Section 2.3 and Figure 2 on each group $g \in \mathcal{G}$ separately [2]. A conformal decision quantile $\hat{Q}^{(g)}_{1-\alpha}$ is estimated separately for each group g on the individuals in $\mathcal{I}_2$ that satisfy the conditions of group g.

Then, for a new individual whose auxiliary variables are observed, we simply produce a prediction set following the APS Prediction procedure using the group-specific $\hat{Q}^{(g)}_{1-\alpha}$ quantile. Although quite simple to implement and understand, such a method may prove to be time inefficient, especially for a large number of groups, since it requires an iterative traversing and quantile estimation on each group separately.

4.2. Calibration by Quantile Regression

We propose a simple and more elegant reformulation of the group-conditional conformal calibration procedure via quantile regression. Quantile regression [13, 6] is a method that allows the estimation of a desired $\tau \in [0, 1]$ quantile of a dependent variable Y based on a set of explanatory variables X [2]. It can be understood as the counterpart of linear regression – that estimates the mean of the output variable – for the estimation of the quantiles, a special case of which is the median for $\tau = 0.5$. For an output variable Y and explanatory variables X, a generic formulation of the quantile regression is given by:

$$Q_{Y|X}(\tau) = X\beta_\tau \tag{5}$$

where $Q_{Y|X}(\tau)$ is the τ quantile of the conditional distribution of Y given X, assuming a linear relationship between the conditional quantile and the explanatory variables. The estimated coefficient $\hat{\beta}_\tau$ is solution to the following optimization problem:

$$\hat{\beta}_\tau = \underset{\beta \in \mathbb{R}^d}{\arg\min} \big[(\tau - 1) \sum_{Y_i < X_i\beta} (Y_i - X_i\beta) + \tau \sum_{Y_i > X_i\beta} (Y_i - X_i\beta)\big] \tag{6}$$

where d is the dimension of the vector X. This minimization problem can be efficiently solved using linear programming approaches [14, 12].

[2]Note that X and Y here are not as defined previously but are generic variable names in keeping with common definitions of regression models.

4.2.1 Calibration

We can thus estimate the group-conditional $1-\alpha$ quantiles of the scores by regressing them on group-membership indicator variables. This constitutes the calibration of the conformal procedure.

To illustrate how the approach works, we consider the two previously described auxiliary variables, *location* and *sky*. To simplify the presentation, we consider that the variable *location* has only two levels: $M_L \in \{\texttt{A}, \texttt{B}\}$, and *sky*: $M_S \in \{\texttt{sunny}, \texttt{overcast}\}$. The regression model

$$\begin{aligned} Q_{E|\{M_L,M_S\}}(1-\alpha) = \beta_0 + \\ \beta_{\texttt{A}}\mathbb{1}_{\{M_L=\texttt{A}\}} + \\ \beta_{\texttt{sunny}}\mathbb{1}_{\{M_S=\texttt{sunny}\}} + \\ \beta_{\texttt{A, sunny}}\mathbb{1}_{\{M_L=\texttt{A}\}}\mathbb{1}_{\{M_S=\texttt{sunny}\}} \end{aligned} \tag{7}$$

where $\mathbb{1}_{\{\ \}}$ is the indicator function, thus allows us to estimate the $1-\alpha$ quantile of the scores for all the possible groups defined by these two auxiliary variables. Notice that all the groups are identified in this model: $\hat{\beta}_0$, the estimated intercept, is the estimated $1-\alpha$ quantile of the score for the baseline group, defined by the conditions that are not explicitly specified in the regression equation; in this case $\{M_L = \texttt{B}, M_S = \texttt{overcast}\}$. The other estimated coefficients $\hat{\beta}_{\texttt{A}}$, $\hat{\beta}_{\texttt{sunny}}$ and $\hat{\beta}_{\texttt{A, sunny}}$ are to be interpreted as the difference in quantiles from the baseline $\hat{\beta}_0$. Hence, the estimated quantile for the group $\{M_L = \texttt{B}, M_S = \texttt{sunny}\}$ is $\hat{\beta}_0 + \hat{\beta}_{\texttt{sunny}}$, just as the estimated quantile of the group $\{M_L = \texttt{A}, M_S = \texttt{sunny}\}$ is $\hat{\beta}_0 + \hat{\beta}_{\texttt{A}} + \hat{\beta}_{\texttt{sunny}} + \hat{\beta}_{\texttt{A, sunny}}$.

This methodology can be simply expanded for the case where more auxiliary variables are considered or where the auxiliary variables have more than two levels, or are continuous [13] – unlike the classical approach.

4.2.2 Prediction

For a new observation for which we observe the auxiliary data, we can easily plug-in its values in the regression model and obtain its corresponding quantile estimation. It is the estimated quantile of the group to which the observation belongs. The obtained $\hat{Q}^{(g)}_{1-\alpha}$ will then be used following the APS prediction procedure previously described in Section 2.3 and Figure 3 to produce prediction sets for this new observation.

4.2.3 Experimental Results

The proposed approach is compared to the Marginal APS. The validation set is split into a calibration set $\mathcal{I}_2$ (45%) and prediction set $\mathcal{I}_3$ (55%) following a stratified proportion sampling scheme where each group is sampled according to its proportion in the validation set. The two methods are calibrated and tested on the same data. In order to validate the results, we implement a resampling scheme over 100 iterations leading to a different split of the validation set at each iteration.

Figure 7 shows the boxplots of the obtained coverage per group for the 100 resamplings for the two methods, with the groups sorted by decreasing order of number of individuals. While the Marginal APS shows, generally, a smaller variance per group, its group-specific coverages are highly biased. We observe a high variability in the group coverages, echoing the results previously presented in Section 3.4. On the other hand, our proposed group-conditional method stably maintains the group coverage at the required 0.9 level, on average, for all groups. Even though the variance of the observed coverage is naturally higher for less-represented groups, it is still acceptably maintained over the 100 iterations.

Table 3 shows the average empirical coverage and set size for each group over the 100 resamplings. The proposed approach by quantile regression leads, on average, to smaller prediction sets without compromising on coverage. The importance of such a result may not be obvious in use cases with few classes like the current one. However, on datasets with a large number of classes, valid prediction sets with smaller size are largely preferred for use cases of automated decision making based on the predicted sets, and applications requiring a study of the prediction sets by a human agent [3].

5. Conclusion

In this article, we introduced and presented the conformal prediction framework from a practical perspective with a special focus on its importance to the agricultural community. Indeed, as deep learning black box methods become the go-to approaches in a large spectrum of automated agricultural tasks, methods that provide valid guarantees on their performance – or, at least, quantify the uncertainty associated to their predictions – are important to certify their quality. Here, the work was demonstrated on the task of weed and crop classification in real-world conditions. Special attention has been accorded to the recently developed Adaptive Prediction Sets (APS) method which was shown to empirically maintain the marginal coverage guarantee as defined in Equation 1. However, the marginal guarantee is not enough to ensure the required coverage is maintained on all possible individuals or groups of individuals (in our case defined by auxiliary data acquired during image acquisition): it is thus not enough for multiple agricultural use cases.

This motivated our presentation of group-conditional conformal prediction; first, via the classical approach that consists of iteratively applying the APS procedure on each group separately; then using our proposed "elegant" approach via quantile regression of calibrated softmax scores

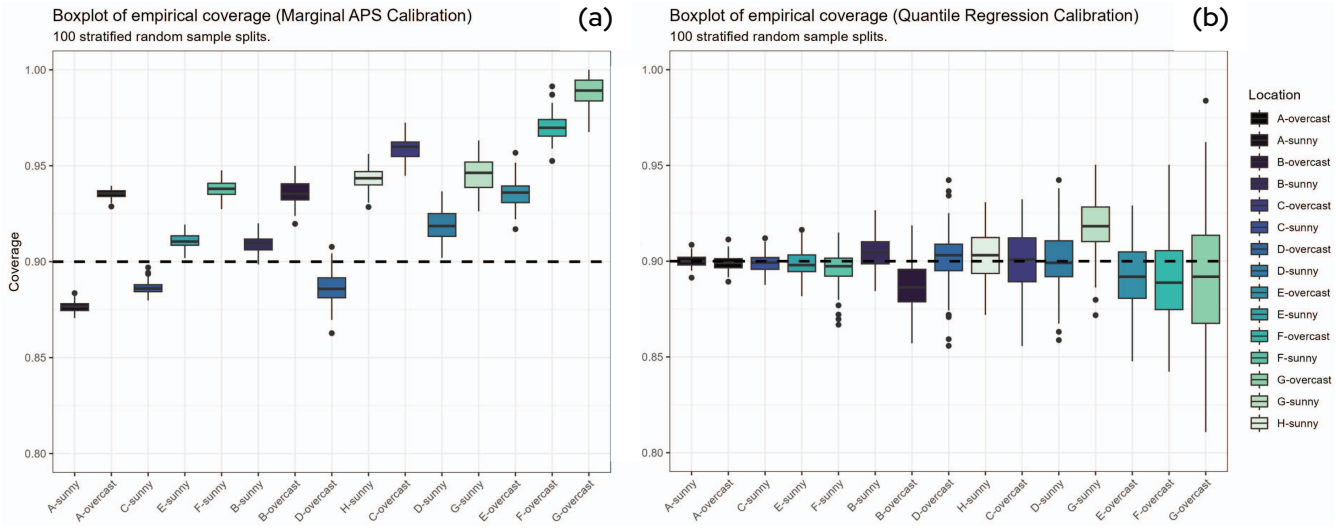

Figure 7. Boxplot of the empirical coverage per group over 100 different splits of the validation set: (a) Marginal APS. (b) Quantile Regression Calibration (ours). Groups are sorted in decreasing order of number of individuals.

			Marginal APS		**Quantile Regression**	
Group	*Location*	*Sky*	Coverage	Set Size	Coverage	Set Size
1	A	overcast	0.935	2.838	0.899	2.415
2	A	sunny	0.876	2.627	0.900	2.893
3	B	overcast	0.936	1.566	0.887	1.354
4	B	sunny	0.909	1.985	0.904	1.939
5	C	overcast	0.959	2.615	0.900	1.929
6	C	sunny	0.886	2.431	0.899	2.560
7	D	overcast	0.886	2.482	0.901	2.683
8	D	sunny	0.918	2.439	0.901	2.286
9	E	overcast	0.936	2.157	0.892	1.735
10	E	sunny	0.911	2.480	0.899	2.364
11	F	overcast	0.970	2.494	0.892	1.699
12	F	sunny	0.938	2.788	0.897	2.295
13	G	overcast	0.989	2.707	0.891	1.682
14	G	sunny	0.945	2.358	0.918	2.150
15	H	sunny	0.943	2.493	0.903	2.066
Marginal Results			**0.900**	**2.570**	**0.898**	**2.137**

Table 3. Comparison of average empirical coverage and prediction set size over 100 different splits.

on group membership indicators. The proposed approach allows for the joint estimation of the $1-\alpha$ decision quantiles of all groups. Quantile regression calibration has been shown empirically to maintain the $1-\alpha$ coverage level for all groups, even those that are not largely represented in the dataset. This approach also provided smaller prediction sets, on average, per group being thus more useful from a decisional perspective – simply because it is easier to take a decision when fewer classes are predicted.

This article is the first work, to the authors' knowledge, to introduce these notions and methods to the agri-tech community. It constitutes a first step in a research direction aiming at developing reliable and trustworthy machine learning systems on which the farmers can rely and have confidence in, even without fully understanding all their intricacies. Future work aims at extending the current methods to the more realistic scenario in which the auxiliary data are not, or only partially, observed on prediction images; at developing theoretical guarantees of the maintenance of group coverage by quantile regression; and finally at adapting and presenting the conformal methodology on other computer vision tasks such as object detection and image segmentation.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021-12. 2

[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. Publisher: Now Publishers. 2, 6

[3] Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *International Conference on Learning Representations (ICLR)*, 2021, 2021. 2, 3, 4, 7

[4] Martin Arjovsky. Out of distribution generalization in machine learning. *PhD Thesis, Courant Institute of Mathematical Sciences, New York University*, 2019. 2

[5] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Newnes. Google-Books-ID: pgfUAgAAQBAJ. 2

[6] A. Beyerlein. Quantile regression–opportunities and challenges from a user's perspective. *American Journal of Epidemiology*, 180(3):330–331, 2014. 6

[7] Sarah Condran, Michael Bewong, Md Zahidul Islam, Lancelot Maphosa, and Lihong Zheng. Machine learning in precision agriculture: A survey on trends, applications and evaluations over two decades. *IEEE Access*, 10:73786–73803, 2022. 1, 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[10] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021-03. 2

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[12] Roger Koenker. quantreg: Quantile regression. *http://CRAN. R-project. org/package= quantreg*, 2009. 6

[13] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. Publisher: [Wiley, Econometric Society]. 6, 7

[14] Roger W. Koenker and Vasco D'Orey. Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):383–393, 1987. Publisher: [Wiley, Royal Statistical Society]. 6

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018. 2

[16] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, Lecture Notes in Computer Science, pages 345–356. Springer, 2002. 3

[17] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 2

[18] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2), 2020-04-30. 2, 6

[19] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020. 2, 3, 4, 5

[20] Sheila M. Saia, Natalie Nelson, Anders S. Huseth, Khara Grieger, and Brian J. Reich. Transitioning machine learning from theory to practice in natural resources management. *Ecological Modelling*, 435:109257, 2020-11. 1, 2

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020-02. 1

[22] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. 2, 3

[23] Fan Shi, Cheng Soon Ong, and Christopher Leckie. Applications of class-conditional conformal predictor in multi-class classification. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 235–239, 2013. 2

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014-02-19. 1

[25] Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2):349–376, 2013-09. 2

[26] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. 2, 3

[27] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019. 2

[28] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon,*

France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 2

[29] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021-03. 2

[30] Qin Zhang and Manoj Karkee. Agricultural and field robotics: An introduction. In Manoj Karkee and Qin Zhang, editors, *Fundamentals of Agricultural and Field Robotics*, Agriculture Automation and Control, pages 1–10. Springer International Publishing, 2021. 1

[31] Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021-10. 1