Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery

Fulgencio Cánovas-García[a,d,*], Francisco Alonso-Sarría[b], Francisco Gomariz-Castillo[b,c], Fernando Oñate-Valdivieso[a]

[a] *Departamento de Geología y Minas e Ingeniera Civil, Universidad Técnica Particular de Loja. San Cayetano Alto s/n, Loja, Ecuador*
[b] *Instituto Universitario del Agua y del Medio Ambiente, Universidad de Murcia, Edificio D Campus de Espinardo s/n 30100 Murcia, Spain*
[c] *Instituto Euromediterráneo del Agua, Universidad de Murcia, Edificio D s/n 30100 Murcia, Spain*
[d] *Departamento de Ingeniería Civil, Universidad de Cuenca, Av. 12 de Abril y Av. Loja s/n, Cuenca, Ecuador*

Abstract

Random forest is a classification technique widely used in remote sensing. One of its advantages is that it produces an estimation of classification accuracy based on the so called out-of-bag cross-validation method. It is usually assumed that such estimation is not biased and may be used instead of validation based on an external data-set or a cross-validation external to the algorithm.

In this paper we show that this is not necessarily the case when classifying remote sensing imagery using training areas with several pixels or objects. According to our results, out-of-bag cross-validation clearly overestimates accuracy, both overall and per class. The reason is that, in a training patch, pixels or objects are not independent (from a statistical point of view) of each other; however, they are split by bootstrapping into in-bag and out-of-bag as if they were really independent. We believe that putting whole patch, rather than pixels/objects, in one or the other set would produce a less biased out-of-bag cross-validation. To deal with the problem, we propose a modification of the random forest algorithm to split training patches instead of the pixels (or objects) that compose them. This modified algorithm does not overestimate accuracy and has no lower predictive capability than the original. When its results are validated with an external data-set, the accuracy is not different from that obtained with the original algorithm.

We analysed three remote sensing images with different classification approaches (pixel and object based); in the three cases reported, the modification we propose produces a less biased accuracy estimation.

*Corresponding author
Email address: fulgencio.canovas@um.es (Fulgencio Cánovas-García)

Keywords: Classification, random forest, object-based image analysis, bagging, statistical independence.

1. Introduction

Classification has been one of the most relevant practices in remote sensing; as a consequence, a great deal of effort has been devoted to developing and applying new techniques to classify remote sensing imagery, mainly based on artificial intelligence and machine learning [1]. Recently, ensemble learning techniques have received much attention. Such methods generate a large number of classifiers, which are later grouped, using a more or less complex procedure, to obtain a global classification. Decision trees are among the most suitable machine learning techniques used in ensembles; boosting, bagging and random forest (RF) are well known ensemble learning techniques used with decision trees [2].

RF has been used in medicine (e.g. Ghose *et al* (2012) [3]), ecology (e.g. Cutler *et al.* (2007) [4]), hydrology to classify groundwater samples (e.g. Baurdon *et al.* (2013) [5]), chemistry (e.g. Svetnik *et al.* (2004) [6]); in soil science (e.g. Schmidt *et al.* (2008) [7]), or to analyse land abandonment (e.g. Alonso-Sarría *et al.* (2016) [8]). The use of RF in image classification has undergone significant growth. Many research papers highlight its good performance compared with more traditional alternatives [4, 9]. It also outperforms more recent algorithms such as artificial neural networks or weighted k-nearest neighbors [10, 11], and has proved to be as powerful as support vector machines [12, 13, 14, 15]. Other advantages are that it is a non-parametric method, so no theoretical distribution is assumed in the training data; it is among the most accurate machine learning methods [16]; it provides a measure of the importance of variables; it is available as a package (`randomForest`) in the open-source program R [17]; it produces an internal measurement of the accuracy (out-of-bag cross-validation, OOB-CV); and it is less sensitive than other algorithms to the Hughes effect [11]. The main disadvantage of RF (at least in classification) is that the effect of the variables is not as easy to interpret as in other methods (e.g. decision trees or discriminant analysis). When used as a regression tool, partial dependence plots might be used to interpret the effect of the different variables, but the interpretation is not as straightforward in classification. However, when classifying images, the ability to predict is more important than the ability to explain.

1.1. The random forest algorithm

A clear and comprehensive description of classification trees and derived ensemble learning techniques can be found in Gao (2009) [1], Waske *et al.* (2012) [18], James *et al.* (2013) [19] or Kuhn and Jhonson (2013) [20]. Here we briefly describe the characteristics of the method to explain why we think OOB-CV may be biased in certain remote sensing applications.

Decision trees [1] are a non-parametric technique that can select, from among a wide set of features, those that best discriminate the dependent variable,

whether quantitative (regression) or qualitative (classification). One of the most popular decision trees algorithms is CART (Classification and Regression Trees) [21].

The calibration of a classification tree begins with a single node including all training cases. This node is then split into two nodes using the predictor feature and threshold value that minimise a heterogeneity measurement in the resulting nodes. This process continues until all terminal nodes are homogeneous. In a second step, the tree is pruned using an independent set of training data to obtain a balance between accuracy and parsimony [1] and to avoid overfitting. The Gini index [21] is used as heterogeneity measurement in CART and RF. The importance of a given feature in a tree is measured as the sum of the decrements in the Gini index attributed to that feature along the tree.

The main problem with decision trees is their high variance; they are very sensitive to slight differences in the training data that might drive the node-splitting process through a different path, leading to a completely different tree. Ensemble learning algorithms (boosting, bagging and RF) attempt to solve this issue.

In bagging, all trees are trained independently and simultaneously. Each tree is trained with a subset of cases obtained by bootstrapping, whereas the others (around 33% on average) form the so called out-of-bag. Each case appears in the out-of-bag of several trees, and these trees are used to predict its class by a vote system. Finally, the comparison of predicted and observed classes is used to obtain an estimation of the overall and per class accuracy, the so called out-of-bag cross-validation (OOB-CV).

RF [22] is one of the most used classification algorithms based on decision trees. This algorithm uses bagging, but includes another randomisation component: random feature selection. The split variable in each node of the decision trees is chosen from a random subset of the available features [18]. This seemingly counter-intuitive modification has proved to be a strategy that gives very good results [17]. It reduces correlation among trees, giving more sense to the whole ensemble learning concept [19].

RF provides measurements of the importance of variables. One of the most used is the mean decrease in the Gini index (MDGI), which is obtained for each feature by averaging its importance in all the trees [21].

The number of features randomly chosen to split each node (*Mtry*) is one of the parameters that the user must decide or optimise; however, the method is not very sensitive to this parameter, whose default value is obtained by truncating the square root of the number of available features [23]. Another configurable parameter is the number of trees generated (*Ntree*), 500 by default. Higher values do not significantly increase the accuracy of the classification [17, 16]. Ismail *et al.* (2010) [24] and Cánovas-García and Alonso-Sarría (2015b) [11] obtained good results using these default parameters.

1.2. The spatial dependence problem with out-of-bag cross-validation (OOB-CV) and leave-one-out cross-validation (LOO-CV)

All predictive models assume that calibration and validation cases are independent . When classifying remote sensing imagery, cases are obtained as training and validation areas. These areas are *patches* of pixels that do not present spatial discontinuities and are homogeneous enough for the photointerpreter to label them as the same class. The objective is to find patches that can be assimilated to the different classes in which we want to divide the image.

Spatial autocorrelation among reflectivity values has been largely studied and has been even used to create contextual features that improve classification accuracy [25]. However, because of this spatial autocorrelation, reflectivity values inside a patch are not independent of each other. So, we can consider that pixels in different training patches, and their reflectivity values, are statistically independent of each other, but pixels in the same training patch are not. This issue should be taken into account when doing cross validation, in order to avoid splitting pixels from the same patch into calibration and validation data-sets.

When analysing non-spatial data, it is usually considered that random forest OOB-CV provides an unbiased estimation of the overall classification accuracy, making an external cross-validation unnecessary [26, 22, 6].

However, we hypothesize that RF OOB-CV overestimates accuracy significantly, at least when classifying remote sensing imagery. In our view, the reason for this overestimation is that bagging assumes independence among the cases (pixels) in each calibration patch and, therefore, will split them between the bootstrapped and the out-of-bag subsamples. So the necessary independence between calibration and validation data is compromised and the OOB-CV accuracy estimation will overestimate the real accuracy of the model.

All these considerations are also valid in Object Based Image Analysis (OBIA). The OBIA approach involves two steps: segmentation, which consists of dividing the image into spatially cohesive objects [27], and the posterior classification of such objects using a larger set of features that include spectral, textural, contextual and geometrical attributes. Objects within a training patch are more similar among themselves than to objects located in other patches, even if these patches belong to the same class, since intra-patch object homogeneity is greater than the inter-patch homogeneity.

The three different validation approaches that will be used are:

- VAL: Validation with a different and independent data-set.
- LOPO-CV: Leave-one-patch-out cross-validation: cross validation carried out leaving out not just one pixel or object, but all the pixels/objects in a training patch.
- OOB-CV: Out-of-bag cross-validation, the RF internal error estimation.

When analysing the results of the original RF algorithm, we will add an O in front of the validation method, and when using our modification we will add an M. Thus, M-LOPO-CV will mean leave-one-patch-out cross-validation of a

classification carried out with the modified algorithm, and O-VAL will mean validation with an independent data-set of a classification carried out with the original algorithm.

1.3. Objectives

The overall objective of this research is two-fold. Firstly, to demonstrate that lack of independence among elements (pixels or objects) in training patches may compromise the statistical independence between training and test elements when doing O-OOB-CV accuracy estimation. Secondly, to propose a modification of the original RF algorithm, more specifically the `randomForest` function in the `randomForest` R package [17]. This modified algorithm produces a modified RF out-of-bag cross-validation (M-OOB-CV) which is unbiased when analysing spatial data. These overall objectives involve several partial objectives:

1. To demonstrate that O-OOB-CV underestimates the prediction error measured by leave-one-patch-out cross-validation using the original RF algorithm (O-LOPO-CV).
2. To implement a modification of the original algorithm to guarantee the statistical independence of elements assigned internally to the in-bag and those assigned to the out-of-bag. The cross-validation performed by this modified algorithm is the above mentioned M-OOB-CV.
3. To demonstrate that M-OOB-CV error estimation is not as biased as O-OOB-CV, using a validation with a different dataset (VAL) as reference.
4. To demonstrate that M-VAL is equivalent to O-VAL. This would imply that the proposed modification does not involve a loss in the predictive capability of the modified algorithm.
5. To generate a modified version of the `randomForest` function [17] in an R package freely accessible to anyone interested.

2. Study areas and data sets

To verify our hypothesis, three study areas were analysed using different types of images and approaches; the objective was to test the generality of our hypothesis. The first image is an object-based case whereas the other two are pixel-based cases. One of the characteristics of the object based approach is that it produces a large amount of features, so a selection process is needed, and this process can also be affected by the lack of statistical independence.

2.1. Irrigation Unit 28 in south-eastern Spain (IU28)

The first study area (Figure 1 a), located in the Region of Murcia (south-east Spain, Figure 1 d), corresponds to Irrigation Unit 28, as defined in the *Plan Hidrológico de la demarcación del Segura 2015/2021* (River Segura Basin Hydrological Plan 2015/2021). In this area, a high resolution image was classified. It consists of a 2 m resolution multispectral (Blue, Green, Red and Near

Infrared) image and a 0.45 m panchromatic image acquired on 9,10 and 11 July 2008 with an Intergraph Z/I-Imaging Digital Mapping Camera.

The image was segmented using multiresolution segmentation [28], one of the most widely segmentation algorithms used in OBIA. The details can be consulted in Cánovas-García and Alonso-Sarría (2015) [29].

The objective of the classification was to produce a map of agricultural land cover types; the classes included in the classification scheme were: Almond trees (Alm); cereals (Cer); irrigated grassland (Igr); rural wasteland (Rws); irrigated fruit trees (Ifr); rainfed arable lands (Rar); olive trees (Oli); greenhouses (Gre); seedlings (See).

2.2. Vinalopó river basin (Vinalopó)

This study area (Figure 1 b) covers about 3000 km^2. It is a very anthropised coastal basin located in south-east Spain (south of Alicante province). Despite its small size, the variety of land-uses is large. Height ranges from 0 to 1600 m.a.s.l., giving a variety of natural environments. A Landsat 5 Thematic Mapper image (path 199, row 33) from 24 July 2009 was used. Visible and reflected infrarred bands were used to classify the image. Preprocessing of the image included atmospheric [30] and illumination [31] corrections. Additionally, terrain information from a 1:25,000 DEM from the Spanish *Instituto Geográfico Nacional* (National Geographical Institute) was used as ancillary data. The objective was to obtain a land-cover map using a pixel-based classification. The classification scheme includes: Forest (For); scrub (Scr); sparse tree crops (NDArb); dense tree crops (DArb); rainfed grass crops (NIGr); irrigated grass crops (IGr); impervious surfaces (Imp); water bodies (Wat); bare soil (BaSo); vineyards (Vin).

2.3. Zapotillo municipality (Zapotillo)

The Zapotillo municipality (Figure 1 c) is located in the south-west of Loja province (Ecuador) (Figure 1 e). The municipality covers an area of more than 1200 km^2. It is located in a transition zone between the inter-Andean region and the coastal region so that its climate is influenced by the Pacific Ocean, the warm Equatorial Countercurrent, and the movements of the intertropical convergence zone. Landsat 8 Operational Land Imager sensor data (path 011, row 063) were used to study the area. The image was taken on 12 June 2013, using eight out of the nine available bands (Table 3). The radiometric resolution was 16 bits. No pre-processing was carried out and digital counts, rather than reflectivities, were used. The scene was clipped according to the limits of the study area. The classification of this image was also based on pixel analysis. The objective was to produce a map of agricultural classes: Forest (For); scrub (Scr); rice (Ric); corn (Cor); fallow (Fall); associated crops (Asso); pastureland (Pas).

3. Methodology

3.1. Random Forest algorithm modification

We have created a new package called `SDRF` (Spatial Dependence Random Forest) including a modification of the original `randomForest` package [17]. In this last package, the R function `randomForest` calls a C function named `classRF` (located in the `rf.c` file in the `src` directory), which performs most involved calculations. Currently, this package works only in Linux systems (see supplementary material).

We have modified the `classRF` function to receive 2 additional arguments: a pointer to integer values that contain the numeric identifier of the training patch in which each case (pixel or object) is located, and an integer with the number of training patches. If this last argument is not equal to zero, a training patches bootstrapping is carried out instead of pixels/objects bootstrapping. In this way, all pixels/objects inside a training patch will be put in the same place: the in-bag or the out-of-bag. This modified function is named `classRF2` in the new package.

We have also created the SDRF function as a modification of the the `randomForest` function that receives a new argument called `areas`, with which the user can pass the identifiers of the training patches to the function. The function will internally calculate the number of training patches and will pass both arguments to the C function `classRF2`.

3.2. Training and validation datasets

The three datasets were obtained in different projects, so sampling procedures were also different. In IU28 and Vinalopó, sampling procedures derive from the objectives of such projects. Only Zapotillo data were collected specifically for this paper. Table 1 shows the main characteristics of the training and validation areas.

In IU28, training areas were collected using a not random stratified sampling, trying to properly represent all classes, and including 30 patches per class, except for seedlings (15 patches) as this class has very low frequency. Validation areas were collected using a random stratified sampling including 50 patches per class (15 in seedlings).

In Vinalopó, both training areas and validation areas were collected using a random stratified sampling. The sizes of the strata were proportional to the percentage of each class in the study area that was estimated using the 2006 CORINE Land Cover land use maps.

Finally, in Zapotillo a random stratified sampling were carried out both for training and validation areas with the same number of patches in each class (30) except rice (6) and fallow (20) that are quite infrequent in the study area.

In IU28 training and validation areas were identified and labelled by a combination of fieldwork and a descriptive statistical analysis to support photointerpretation. In addition, very high resolution imagery and thematic maps were used as ancillary data to help in the identification of the different land-covers.

In Vinalopó and Zapotillo, training and validation areas were identified using aerial photographs and land use/land cover maps.

3.3. Features obtained from the images

Table 2 shows the object features calculated using eCognition software in IU28. Features are grouped into six main categories and a short description is added when needed. The number of bands from which the features were calculated is indicated between parentheses. Technical details of every feature are described in DEFINIENS (2009) [32]. In summary, there are 356 features: 40 spectral features, 5 pixel-based features, 24 geometric features, 204 texture features and 83 context features.

Table 3 shows the pixel features for Vinalopó and Zapotillo calculated using GRASS GIS 7. Features are grouped into five main categories and a short description is added when needed. For Vinalopó, 14 spectral features, seven related with DTM and 34 texture features, were calculated. Finally, for Zapotillo 16 spectral features and 32 texture features were calculated.

3.4. Feature ranking and selection

A successful approach in machine learning is to consider feature selection as a heuristic procedure in which a subset of possible features is specified at each step of an iterative search [33]. Such a procedure involves 3 steps:

(1) Ranking all features in accordance with a criterion related to their relevance for classifying the dataset, in this case the mean decrease in the Gini index (MDGI) obtained for each variable aplying the original algorithm. Spearman correlation test was used to verify whether the modified algorithm might significantly modify the variable importance ranks.
(2) Iteratively modifying a classification model by removing features in reverse order to their MDGI-based rank.
(3) Selecting the best feature subset according to a classification accuracy measurement: the kappa index of the M-VAL curve.

Once all features were ranked, they were used to train both the original and the modified RF algorithms using the default *Ntree* and *Mtry* values. Kappa indices from O-OOB-CV, M-OOB-CV, O-VAL, M-VAL were calculated and the less important features were then eliminated from the dataset. The whole procedure was repeated recursively until only the most important feature was left. The evolution of the kappa indices obtained was then represented (Figures 2 to 4) to show how the accuracy of both the original and the modified RF algorithms evolve through a large number of classifications, the optimal number of features to minimise the classification error that are obtained with each validation method and the differences introduced by our modification in RF algorithm in the feature selection process.

This approach significantly reduces the number of features needed to train classification algorithms [34, 35, 36] and also serves to test the sensitivity of both algorithms to changes in the number of features.

In Vinalopó and Zapotillo a single feature was eliminated in each cycle; however, in IU28, to reduce the computational cost, five features were eliminated in each cycle due to the high dimensionality of this dataset.

3.5. Per-class accuracy analysis

Although confusion matrices are suitable tools for analysing in detail the results of a classification model, comparing six different matrices becomes quite cumbersome. Instead, per class accuracy statistics were compared to each other using pyramid graphs showing the omission and commission errors (Figures 5 to 7). These pyramids allow a per-class comparison of the results of two different classifications according to the two types of error that are usually studied in classification problems. To facilitate interpretation of the pyramids, the classes have been ordered according to O-LOPO-CV errors of omission.

4. Results

4.1. Feature ranking

Table 4 shows the 25 most important features according to MDGI using the original algorithm. Features related with height, when available, occupy the first ranks (IU28 and Vinalopó). Spectral features also appear in the first ranks in the three study areas. The values obtained with the Spearman correlation test were 0.99 in IU28 and Zapotillo and 0.96 in Vinalopó. These results show that our modification of the algorithm does not significantly change the feature importance rankings.

4.2. Feature-selection process

In Figures 2, 3 and 4, the lines represent the kappa indices obtained by OOB-CV and VAL in both the original and modified algorithm. It is clear that O-OOB-CV largely over-estimates the accuracy of the classification provided by O-VAL or M-VAL. In addition, M-OOB-CV is very similar to O-VAL or M-VAL, and only in the Zapotillo municipality is M-OOB-CV lower. The reason for this smaller accuracy estimation is probably the reduction in randomisation caused by the splitting by areas in M-OOB-CV, as there are less possible in-bag and out-of-bag combinations. Whatever the case, it implies a more conservative accuracy estimation. M-VAL kappa is, in general, slightly higher than O-VAL kappa. This demonstrates that the modified algorithm does not lose predictive capability.

Finally, these graphs allow us to select the smallest subset of variables that maximises the classification accuracy (rounded to two decimals) from a set of ordered features. From now on, we will continue analysing the per class results of the classification models generated with the first 95 features in IU28, the first 13 features in Vinalopó and the first 9 features in Zapotillo (blue vertical line in the graphs of Figures 2, 3 and 4). The selected features appear highlighted in red in table 4 (in IU28, there were 70 more features). The Kappa indices corresponding to these classifications are presented in Table 5.

4.3. Per-class accuracy analysis

Once the subset of features that maximises the classification accuracy was obtained, the corresponding model was analysed to obtain a per-class approach to the differences in accuracy estimation.

Figure 5 compares O-OOB-CV with O-LOPO-CV. In the three areas O-OOB-CV errors are much lower. In the IU28 study area (Figure 5 a) O-OOB-CV per-class error estimations are only similar to O-LOPO-CV when they are close to 0. The most obvious case of underestimation is presented in Vinalopó (Figure 5 b) where the O-OOB-CV error of commission for bare soil (BaSo) is close to 0, whereas O-LOPO-CV value is slightly above 0.8. Similar results were obtained with errors of omission, and also when analysing the class sparse tree crops (NDArb), where small differences between O-OOB-CV and O-LOPO-CV are only obtained in classes with O-LOPO-CV errors close to 0. In Zapotillo (Figure 5 c) both O-OOB-CV omission and commission errors are underestimated in all classes.

Figure 6 shows how M-OOB-CV produces results equivalent to M-LOPO-CV. In IU28, both omission and commission errors are virtually the same in both classifications. There are only minor differences in omission errors in cereals (Cer) and olive tress (Oli) and in commission errors in irrigated fruit trees (Ifr). RF is a stochastic model and does not generate two identical models from the same data, so there are always small differences in the results. In the Vinalopó and Zapotillo study areas (Figure 6 b and c), the results are similar; the omission and commission errors calculated by M-OOB-CV and M-LOPO-CV being similar.

Figure 7 compares the results of O-LOPO-CV and M-LOPO-CV. Differences are very small in IU28 for most of the classes (Figure 7 a), only two classes show slightly different values. The greenhouses class (Gre) for which the biggest omission errors are obtained with the modified algorithm, and the class almond trees (Alm), where the opposite is true. With respect to errors of commission, only one class (irrigated fruit trees) presents a noticeable difference, although it is still minimal. In the Vinalopó study area (Figure 7 b), we obtained similar results. Only the class bare soil (BaSo) presents significant differences in omission and commission errors. The modified algorithm produces slightly larger errors. In Zapotillo (Figure 7 c), only the class rice (Ric) showed different results. Commission errors were larger with the modified algorithm (0.25), the largest of the three study areas.

Finally, Figure 8 shows the comparison among M-VAL and O-VAL in the three study areas. Results are very similar for both accuracy estimations.

In summary, O-OOB-CV over-estimates classification accuracy, whereas M-OOB-CV does not. In addition, when the performance of both algorithm is tested using external cross-validation, the results are very similar. So, we conclude that the modification made in the RF algorithm does not affect its predictive capability.

5. Disscusion

In a recent review paper on RF applied to remote sensing, Belgiu and Dragut (2016) [16] pointed out that although some researchers have reported that the OOB error (equivalent to O-OOB-CV in this study) could be used as a reliable measurement of classification accuracy, very little work has been done on the topic and that the statement should be contrasted with more experiments using a variety of datasets in different application scenarios. Our research might be considered an answer to that call.

According to our interpretation of the literature on RF, O-OOB-CV and LOPO-CV should be similar [2]. However, our hypothesis is that, when classifying remote sensing imagery, the O-OOB-CV accuracy estimation might be biased when training patches are composed of several elements (pixels or objects) because of the statistical dependence between the elements in a single patch. This has been confirmed in the three study areas. When validating the classification models derived from the feature selection, in IU28 the O-OOB-CV kappa index is approximately 0.28 larger that for O-VAL, which is a very large difference. In Vinalopó, this deviation is 0.21, and in Zapotillo 0.14. These differences suggest that O-OOB-CV accuracy estimation is strongly overestimated.

This overestimation also appears when a per-class analysis is carried out. Figure 5 is quite convincing in this sense: all errors of omission and commission from OOB-CV data using the original algorithm are overestimated. Obviously this result is somewhat masked with classes whose LOPO-CV errors of omission and commission are close to zero. Hence, to study these issues in certain cases we have to use less than perfect classifications, otherwise it will be difficult to find bias in the accuracy or error estimations.

Other studies seem to reach different conclusions [37, 38]. A possible explanation for such disparate results may be that research of these authors was based on a data-set with very small validation errors, which might obscure accuracy differences.

We have also tested our modification to the RF algorithm, obtaining equivalent results with M-OOB-CV and M-LOPO-CV, both for omission and commission errors (Figure 6). There were only a few differences in accuracy values in classes with a low number of training patches. In such cases, the reduction in randomisation due to the LOPO-CV approach strongly affects the results, so a large number of small validation patches seems a better option than a small number of large validation patches.

Finally, to check whether the proposed modification reduces the predictive capability of the algorithm, we compared O-LOPO-CV with M-LOPO-CV, on the one hand, and O-VAL with M-VAL, on the other. The differences were negligible, being only slightly higher in classes with fewer training patches.

Another common practice when classifying images with RF is to use O-OOB-CV to identify the feature subset and parameter values that maximise the classification accuracy (e.g. Puissant *et al.* 2014 [39]). According to our data, at least when identifying the optimal subset of variables, this strategy would not have been successful using the original algorithm in two of the three study

areas. The number of variables selected would have been much lower than the number that maximises accuracy classification.

6. Conclusions

A modification of the random forest algorithm is proposed to perform a ~~an~~ patch-based split rather than a pixel-based split when calculating out-of-bag cross-validation.

The modification is performed in the `randomForest` function of the `randomForest` R package [17] (we are not aware if the independence issue is tackled in other random forest implementations). The result is a function called SDRF (Spatial Dependence Random Forest) inside an homonym package that can be downloaded from `https://github.com/pacoalonso/SDRF`. It should be emphasised that we have introduced only a slight modification in a very large and powerful package.

This modification does not affect feature ranking based on MDGI importance. Spearman coefficients among the different rankings were equal to or larger than 0.96.

Neither does the modification produce a loss in prediction capability. Both algorithms were used to classify the same three data-sets; when the results were validated with an external validation set, the results were equivalent.

When the results of the out-of-bag cross-validation in the original algorithm (O-OOB-CV) are compared with a validation with an external data-set or with the results of a leave-one-patch-out cross-validation (LOPO-CV) external to the algorithm, it is clear that O-OOB-CV overestimates accuracy and underestimates both omission and commission errors.

On the other hand, when using the modified algorithm (M-OOB-CV) in the same way, there is neither accuracy overestimation nor error underestimation.

The only drawback of this modification is that if a class is represented by a very small number of training patches the results are strongly affected because of the randomisation reduction inherent in the M-OOB-CV approach.

The feature selection process, the accuracy analysis and the omission and commission errors analysis allow us to reach the aforementioned conclusions.

We think that the results have both a theoretical and a practical interest. We have shown how OOB-CV, as it is currently performed by the random forest algorithm, does not necessarily produce reliable accuracy or error estimations in a remote sensing imagery classification. However, our modification seem to do so.

The implications of this statistical dependence of the elements that form a patch goes beyond the empirical results exposed in this research and are worth to be investigated as wrong conclusions can be reached otherwise.

Acknowledges

This research has been funded by Prometeo Project, Secretariat of Higher Education, Science, Technology and Innovation, Gobierno de Ecuador. We also

thank the four anonymous reviewers whose suggestions have substantially improved this manuscript.

References

[1] H. Gao, Digital Analysis of Remotely Sensed Imagery, McGraw-Hill, 2009.

[2] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, 2009.

[3] S. Ghose, J. Mitra, A. Oliver, R. Martí, X. Lladó, J. Freixenet, J. Vilanova, D. Sidibe, F. Meriadeau, A Random Forest Based Classification Approach to Prostate Segmentation in MRI, 2012, pp. 20–27.

[4] D. Cutler, T. Edwards, K. Beard, A. Cutler, K. Hess, J. Gibson, J. Lawler, Random Forest for Classification in Ecology, Ecology 88 (11) (2007) 2783–2792.

[5] P. Baudron, F. Alonso-Sarría, J. L. García-Aróstegui, F. Cánovas-García, D. Martínez-Vicente, J. Moreno-Brotóns, Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification, Journal of Hydrology 499 (2013) 303–315.

[6] V. Svetnik, A. Liaw, C. Tong, T. Wang, Application of Breiman's Random Forest to modeling structure-activity relationships of pharmaceutical molecules, in: F. Roli, J. Kittler, T. Windeatt (Eds.), MCS, Springer-Verlag, 2004, pp. 334–343.

[7] K. Schmidt, T. Behrens, T. Scholten, Instance selection and classification tree analysis for large spatial datasets in digital soil mapping, Geoderma 146 (1-2) (2008) 138–146.

[8] F. Alonso-Sarría, C. Martínez-Hernández, A. Romero-Díaz, F. Cánovas-García, F. Gomariz-Castillo, Main environmental features leading to recent land abandonment in Murcia Region (Southeast Spain), Land Degradation & Development 27 (3) (2016) 654–670.

[9] A. O. Ok, O. Akar, O. Gungor, Evaluation of random forest method for agricultural crop classification, European Journal of Remote Sensing 45 (2012) 421–432.

[10] A. Maxwell, T. Warner, M. Strager, J. Conley, A. Sharp, Assessing machine-learning algorithms and image- and lidar-derived variables for GEOBIA classification of mining and mine reclamation, International Journal of Remote Sensing 36 (2015) 954–978.

[11] F. Cánovas-García, F. Alonso-Sarría, Optimal Combination of Classification Algorithms and Feature Ranking Methods for Object-Based Classification of Submeter Resolution Z/I-Imaging DMC Imagery, Remote Sensing 7 (2015) 4651–4677.

[12] M. Pal, Random forest classifier for remote sensing classification, International Journal of Remote Sensing 26 (1) (2005) 217–222.

[13] A. Ghosh, P. Joshi, A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery, International Journal of Applied Earth Observation and Geoinformation 26 (0) (2014) 298 – 311.

[14] S. Sesnie, B. Finegan, P. Gessler, S. Thessler, Z. Bendana, A. Smith, The multispectral separability of Costa Rican rainforest types with support vector machines and Random Forest decision trees, International Journal of Remote Sensing 31 (2010) 2885–2909.

[15] E. Adam, O. Mutanga, J. Odindi, E. Abdel-Rahman, Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers, International Journal of Remote Sensing 35 (2014) 3440–3458.

[16] M. Belgiu, L. Dragut, Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing 114 (2016) 24–31.

[17] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002) 18–22.

[18] B. Waske, A. Benediktsson, J. Sveinsson, Random Forest Classification of Remote Sensing Data, in: C. Chen (Ed.), Signal and Image Processing for Remote Sensing, CRC Press, 2012.

[19] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2013.

[20] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013.

[21] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and Regression Trees, Chapman y Hall/CRC, 1984.

[22] L. Breiman, Random Forest, Machine Learning 45 (2001) 5–32.

[23] P. Gislason, L. Benediktsson, J. Sveinsson, Random Forests for land cover classification, Pattern Recognition Letters 27 (2006) 294–300.

[24] R. Ismail, O. Mutanga, L. Kumar, Modeling the Potential Distribution of Pine Forests Susceptible to Sirex Noctilio Infestations in Mpumalanga, South Africa, Transactions in GIS 14 (5) (2010) 709–726.

[25] B. Ghimire, J. Rogan, J. Miller, Contextual land-cover classification: Incorporating spatial dependence in land-cover classification models using random forests and the getis statistic, Remote Sensing Letters 1 (1) (2010) 45–54.

[26] B. Efron, R. Tibshirani, Improvements on Cross-Validation: The .632+ Bootstrap Method, Journal of the American Statistical Association 92 (438) (1997) pp. 548–560.

[27] S. Ryherd, C. Woodcock, Combining Spectral and Texture Data in the Segmentation of Remotely Sensed Images, Photogrammetric Engineering & Remote Sensing 62 (2) (1996) 181–194.

[28] M. Baatz, A. Schape, Multi-resolution Segmentation: an optimization approach for high quality multi-scale image segmentation, in: J. Strobl, T. Blaschke, G. Griesebner (Eds.), Angewandte Geographische Informationsverarbeitung XIII, Wichmann Verlag, 2000, pp. 12–23.

[29] F. Cánovas-García, F. Alonso-Sarría, A local approach to optimise the scale parameter in Multiresolution Segmentation for multispectral imagery, Geocarto International 30 (8) (2015) 937–961.

[30] P. Chávez, An improved dark-object substraction technique for atmospheric scattering correction of multispectral data, Remote Sensing of Environment 24 (1988) 459–479.

[31] P. Teillet, B. Guindon, D. Goodenough, On the slope-aspect correction of multispectral scanner data, Rev. Canadian Journal of Remote Sensing 58 (1982) 84–106.

[32] DEFINIENS, eCognition Developer 8. Reference Book, DEFINIENS, 2009.

[33] A. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245–271.

[34] Q. Yu, P. Gong, N. Clinton, G. Biging, M. Kelly, D. Schirokauer, Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery, Photogrammetric Engineering & Remote Sensing 72 (7) (2006) 799–811.

[35] D. C. Duro, S. E. Franklin, M. G. Dubé, Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests, International Journal of Remote Sensing 33 (14) (2012) 4502–4526.

[36] F. Löw, U. Michel, S. Dech, C. Conrad, Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines, ISPRS Journal of Photogrammetry and Remote Sensing 85 (2013) 102 – 119.

[37] R. Lawrence, S. Wood, R. Sheley, Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest), Remote Sensing of the Environment 100 (2006) 356–362.

[38] V. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS Journal of Photogrammetry and Remote Sensing 67 (2012) 93–104.

[39] A. Puissant, S. Rougier, A. Stumpf, Object-oriented mapping of urban trees using random forest classifiers, International Journal of Applied Earth Observation and Geoinformation 26 (2014) 235–245.

[40] R. Haralick, K. Shanmugan, I. Dinstein, Textural Features for Image Classification, IEEE Tr. on System, Man and Cybernetics SMC-3 (6) (1973) 610–621.

Table 1: Summary statistics of the training and validation samples of the three study areas: almond trees (Alm), cereals (Cer), irrigated grassland (Igr), rural wasteland (Rws), irrigated fruit trees (Ifr), rainfed arable lands (Rar), olive trees (Oli), greenhouses (Gre), seedlings (See), forest (For), scrub (Scr), sparse tree crops (NDArb), dense tree crops (DArb), rainfed grass crops (NIGr), irrigated grass crops (IGr), impervious surfaces (Imp), water bodies (Wat), bare soil (BaSo), vineyards (Vin), rice (Ric), corn (Cor), fallow (Fall), associated crops (Asso), and pastureland (Pas).

Irrigation unit 28					Vinalopó river basin					Zapotillo municipality				
	Training		Validation			Training		Validation			Training		Validation	
Class	Patches	Objects	Patches	Objects	Class	Patches	Pixels	Patches	Pixels	Class	Patches	Pixels	Patches	Pixels
Alm	30	3853	50	7610	For	19	5267	10	1563	For	30	5669	30	4560
Cer	26	830	50	3714	Scr	22	4841	12	3410	Scr	30	2381	30	2682
Igr	33	1187	50	3544	NDArb	13	1241	7	828	Ric	6	134	6	162
Rws	29	1309	50	1365	DArb	14	2374	8	636	Cor	30	605	30	864
Ifr	30	3113	50	4347	NIGr	15	3715	8	1774	Fall	20	480	20	522
Rar	30	985	50	1818	IGr	10	4695	5	1653	Asso	30	2005	30	2267
Oli	30	2568	50	4593	Imp	16	6783	7	1798	Pas	30	1340	30	1633
See	20	1476	15	907	Wat	11	6262	6	3327					
Gre	30	311	50	1192	BaSo	4	118	2	129					
					Vin	17	3177	8	928					
Total	258	15,632	368	25,925		141	38,473	73	16,046		176	12,614	176	12,690
% Sup	0.73%		2.81%			0.032%		0.013%			0.018%		0.018%	

Table 2: Summary of the calculated object features [32] for the irrigation unit 28. Textural features are calculated for several directions. The total number of features appears in parentheses. DTM: digital terrain model, DSM: digital surface model.

Original bands		**Spectral features**	
B1	red	MEAN (10)	
B2	green	SD (10)	standard deviation
B3	blue	MAX (1)	maximum value
B4	near-infrared	MIN (1)	minimum value
C5	DTM	ASYM (10)	skewness
C6	DSM	INTENSITY (1)	IHS transformation
C7	DSM-DTM	HUE (1)	IHS transformation
C8	slope	SATURATION (1)	IHS transformation
C9	aspect	NDVI (1)	normalized difference vegetation index
C10	convexity	RATIO (4)	percentage of total brightness
Geometric features		**Texture features**	
PERIM (1)	including inner borders	GLCM.homo (26)	homogeneity
LENGTH (1)		GLCM.cont (26)	contrast
WIDTH (1)		GLCM.dis (26)	dissimilarity
L/W (1)	LENGTH/WIDTH	GLCM.ent (26)	entropy
ASYM02 (1)	asymmetry	GLCM.asm (26)	angular second moment
BORDER.i (1)	$PERIM/perimeter_{SR}$	GLCM.mean (26)	mean
COMPACT (1)	$LENGTH \cdot WIDTH/AREA$	GLCM.sd (26)	standard deviation
DENSITY (1)	similarity to a square	GLCM.corr (26)	correlation
ELLIPTIC.fit (1)	similarity to a ellipse	**Pixel-based features**	
MAIN.dir, (1)	main direction	MEAN.int.bor (1)	mean reflectivity of the inner border
RADIUS.largest (1)	radius of the largest enclosed ellipse	MEAN.ext.bor (1)	mean reflectivity of the outer border
RADIUS.smallest (1)	radius of the smallest enclosed ellipse	BOR.cont (1)	difference between MEAN.int.bor and the borders of the surrounding objects
RECT.fit (1)	similarity to a rectangle	SD.rec (1)	standard deviation of pixels not in the object but in the SR
ROUNDNESS (1)		NEIGH.cont (1)	difference between MEAN and the mean of pixels not in the object but in the surrounding rectangle
SHAPE.i (1)	$PERIM/(4 \cdot \sqrt{AREA})$	**Context features**	
AREA.excl (1)	area excluding inner polygons	NUM.c (1)	number of neighboring objects
AREA.incl (1)	area including inner polygons	MEAN.c (2)	neighboring objects' mean
LENGTH.arc (1)	average length of arcs	MEAN.d.c (10)	mean difference to neighboring objects, using objects' means
LONGEST.arc (1)	length of longest arc	MEAN.d.c.dr (10)	mean difference to darker neighboring objects
COMPACT.p (1)	AREA divided by the area of a circle with the same perimeter	MEAN.d.c.dr2 (10)	modified mean difference to darker neighboring objects when the darker object is being analyzed
NUMBER.arcs (1)		MEAN.d.c.br (10)	mean difference to brighter neighboring objects
NUMBER.int (1)	number of inner objects	MEAN.d.c.br2 (10)	modified mean difference to brighter neighboring objects when the brighter object is being analyzed
PERIMETER.p (1)	excluding inner borders	NUM.dr (10)	number of darker neighboring objects
SD.edges (1)	standard deviation of length of arcs	NUM.br (10)	number of brighter neighboring objects
		POR.bor.br (10)	relative border to brighter neighboring objects

Table 3: Summary of the calculated object features for the study area Vinalopó river basin and Zapotillo area. Textural features are calculated for several directions. The total number of features appears in parentheses. L5: Landsat 5, L8: Landsat 8, GLCM: Grey level coocurrence matrix.

Original bands		**Derived from the DTM***	
B1 (1)	blue (L5), coastal/aerosol (L8)	SLOPE (1)	slope
B2 (1)	green (L5), blue (L8)	ASP (1)	aspect
B3 (1)	red (L5), green (L8)	CURV.perp (1)	perpendicular curvature
B4 (1)	near infrared (L5), red (L8)	CURV.tang (1)	tangencial curvature
B5 (1)	sort wavelength infrared (L5), near infrarred (L8)	ASP.sin (1)	sin aspect
B6 (1)	short wavelength infrared (L8)	ASP.cos (1)	cosine aspect
B7 (1)	short wavelength infrared (L5 & L8)	**Index and transformations**	
B9 (1)	cirrus (L8)	NDVI (1)	normalized difference vegetation index
DTM* (1)	digital terrain model	INTENSITY (1)	IHS transformation
Texture layers based on the spectral semivariogram		HUE (1)	IHS transformation
VARIO.tc.1 (1)	empirical semivariogram calculated on the first layer of the Tasselled Cap transformation	SATURATION (1)	IHS transformation
VARIO.ndvi (1)	empirical semivariogram calculated on the NDVI layer	TC (4)	Tasselled Cap transformation
Haralick's texture features [40] calculated on the first layer obtained with the Taselled Cup transformation			
GLCM.homo (5)	homogeneity	GLCM.asm (5)	angular second moment
GLCM.cont (5)	contrast	GLCM.coor (5)	correlation
GLCM.ent (5)	entropy	GLCM.var (5)	variance

Table 4: Ranking of the 25 most relevant features according to mean decreasy Gini index. The selected features are highlighted in red (in the case of IU28 there was 70 features more). Features that were calculated with more than one of the original bands are followed by a colon and the band that was used. In textural features, the direction is indicated between parentheses. dir means direccionally-invariant (details in tables 2 and 3).

	Irrigation unit 28	**Vinalopó river basin**	**Zapotillo municipality**
1	MEAN:C6	MDE	B1
2	MEAN:C5	SLOPE	B2
3	dMIN	TC.2	B4
4	MEAN.d.c:B1	B5	B7
5	NEIGH.cont	B4	B3
6	NDVI	TC.1	SATURATION
7	RATIO:B4	NDVI	TC.2
8	RATIO:B1	B7	NDVI
9	MEAN:B1	TC.3	B5
10	MEAN.int.bor	B3	TC.3
11	MEAN.ext.bor	TC.4	VARIO.ndvi
12	MEAN.d.c:B2	INTENSITY	TC.1
13	MEAN:B2	B1	B6
14	RATIO:B2	B2	TC.4
15	INTENSITY	ASPECT	HUE
16	MEAN.d.c:B3	SATURATION	VARIO.tc
17	MEAN.d.c.br2:B1	ASP.sin	GLCM.cont(90)
18	HUE	ASP.cos	GLCM.val(dir)
19	RATIO:B3	HUE	GLCM.var(0)
20	MEAN.d.c.dr:B1	VARIO.ndvi	GLCM.idm(90)
21	MEAN:B4	CURV.tang	GLCM.cont(dir)
22	MEAN.d.c:B4	VARIO.tc	GLCM.idm(dir)
23	MEAN.d.c.br:B1	CURV.perp	GLCM.var(90)
24	MEAN.d.c.dr:B3	GLCM.cont(dir)	GLCM.cont(45)
25	MEAN.d.c.dr:B2	GLCM.var(90)	GLCM.var(45)

Table 5: Kappa indices obtained after feature selection and number of selected features. O-VAL: Validation with a different and independent data-set using original algorithm, M-VAL: Validation with a different and independent data-set using modified algorithm, O-OOB-CV: Out-of-bag cross-validation using original algorithm, M-OOB-CV: Out-of-bag cross-validation using modified algorithm.

	O-VAL	**M-VAL**	**O-OOB-CV**	**M-OOB-CV**	**Features**
Irrigation Unit 28	0.73	0.73	0.97	0.73	95
Vinalopó river basin	0.84	0.86	0.99	0.84	13
Zapotillo area	0.59	0.61	0.76	0.58	9

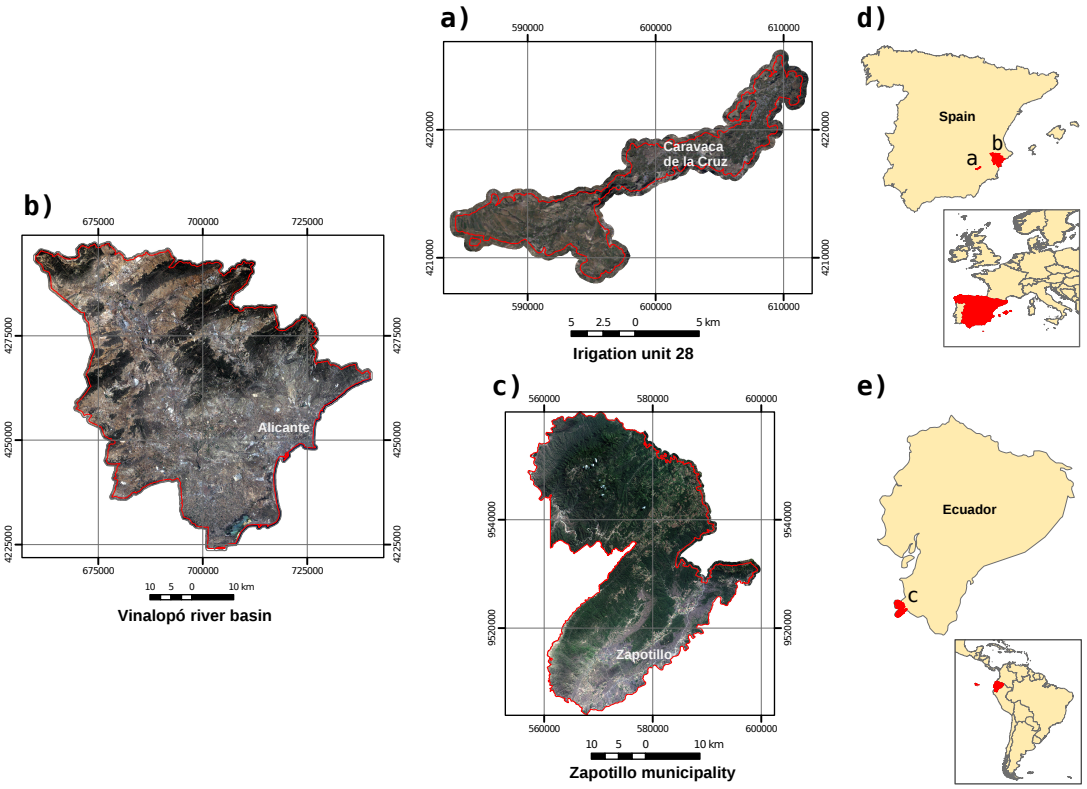

Figure 1: Location of the three study areas. a) Irrigation unit 28. b) Vinalopó river basin. c) Zapotillo municipality. d) Location of irrigation unit 28 and Vinalopó river basin in Spain and Europe. e) Location of Zapotillo municipality in Ecuador and South-America.

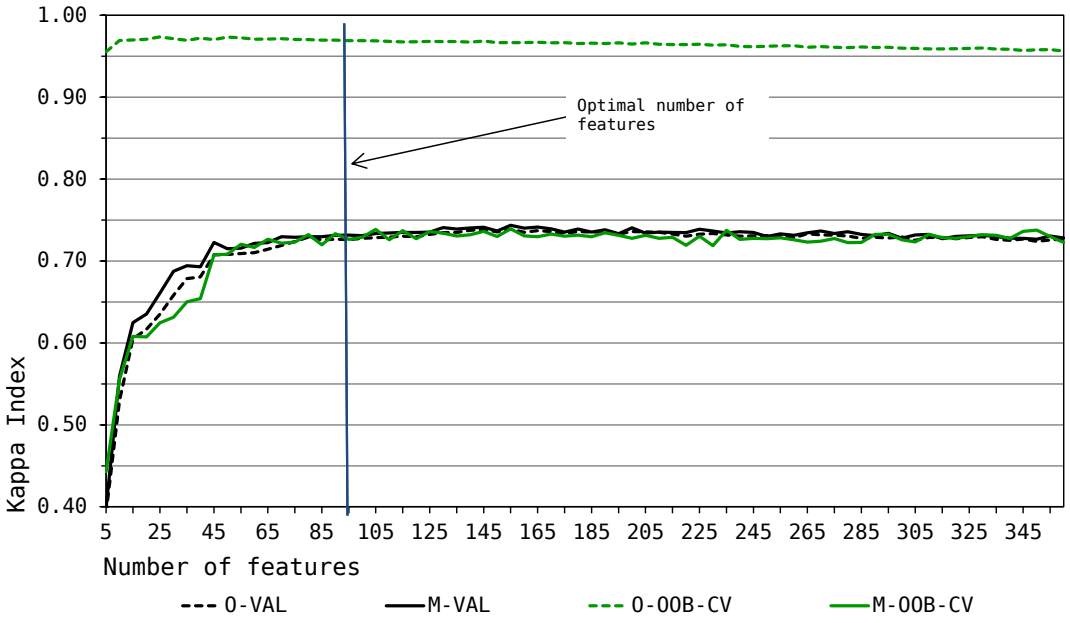

Figure 2: Irrigation unit 28. Kappa indices obtained with original and modified random forest algorithm both using OOB-CV and an external validation data-set. O-VAL: Validation with a different and independent data-set using original algorithm, M-VAL: Validation with a different and independent data-set using modified algorithm, O-OOB-CV: Out-of-bag cross-validation using original algorithm, M-OOB-CV: Out-of-bag cross-validation using modified algorithm.

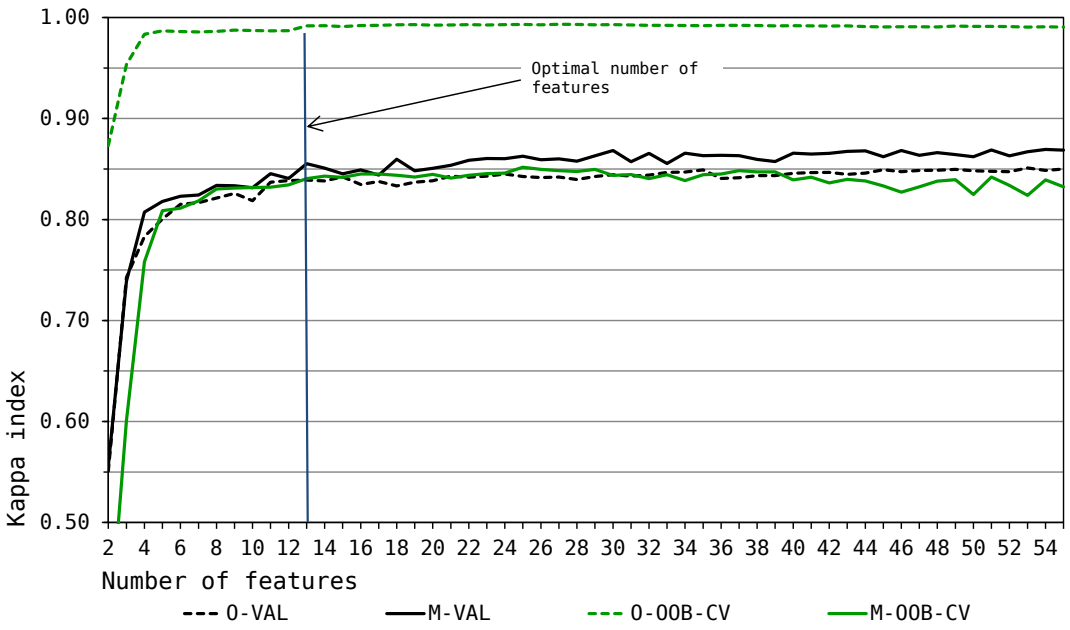

Figure 3: Vinalopó river basin. Kappa indices obtained with original and modified random forest algorithm both using OOB-CV and an external validation data-set. O-VAL: Validation with a different and independent data-set using original algorithm, M-VAL: Validation with a different and independent data-set using modified algorithm, O-OOB-CV: Out-of-bag cross-validation using original algorithm, M-OOB-CV: Out-of-bag cross-validation using modified algorithm.

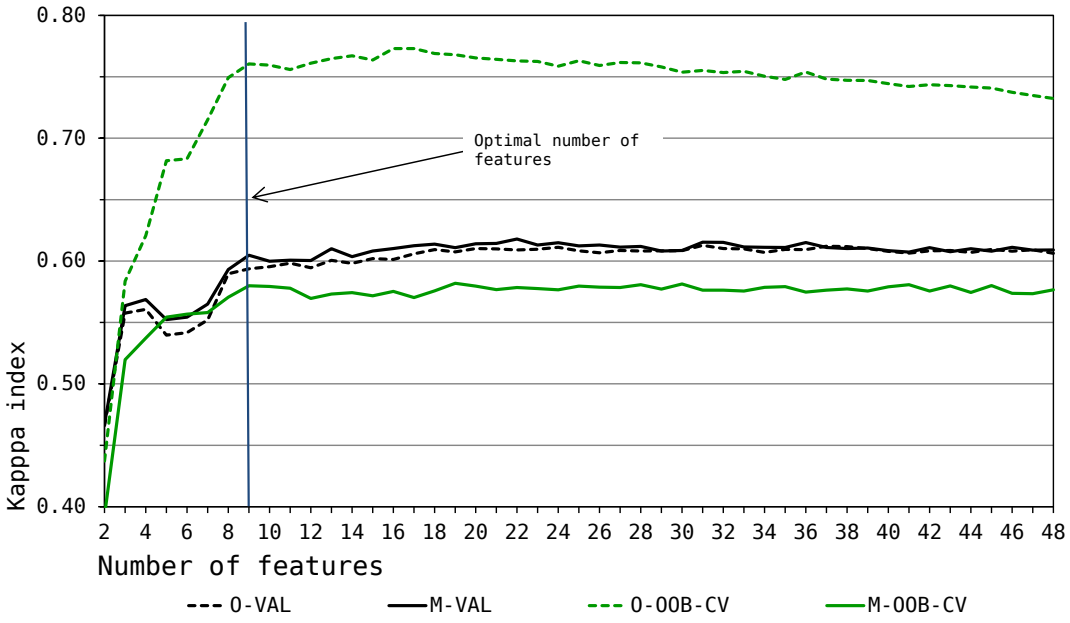

Figure 4: Zapotillo municipality. Kappa indices obtained with original and modified random forest algorithm both using OOB-CV and an external validation data-set. O-VAL: Validation with a different and independent data-set using original algorithm, M-VAL: Validation with a different and independent data-set using modified algorithm, O-OOB-CV: Out-of-bag cross-validation using original algorithm, M-OOB-CV: Out-of-bag cross-validation using modified algorithm.

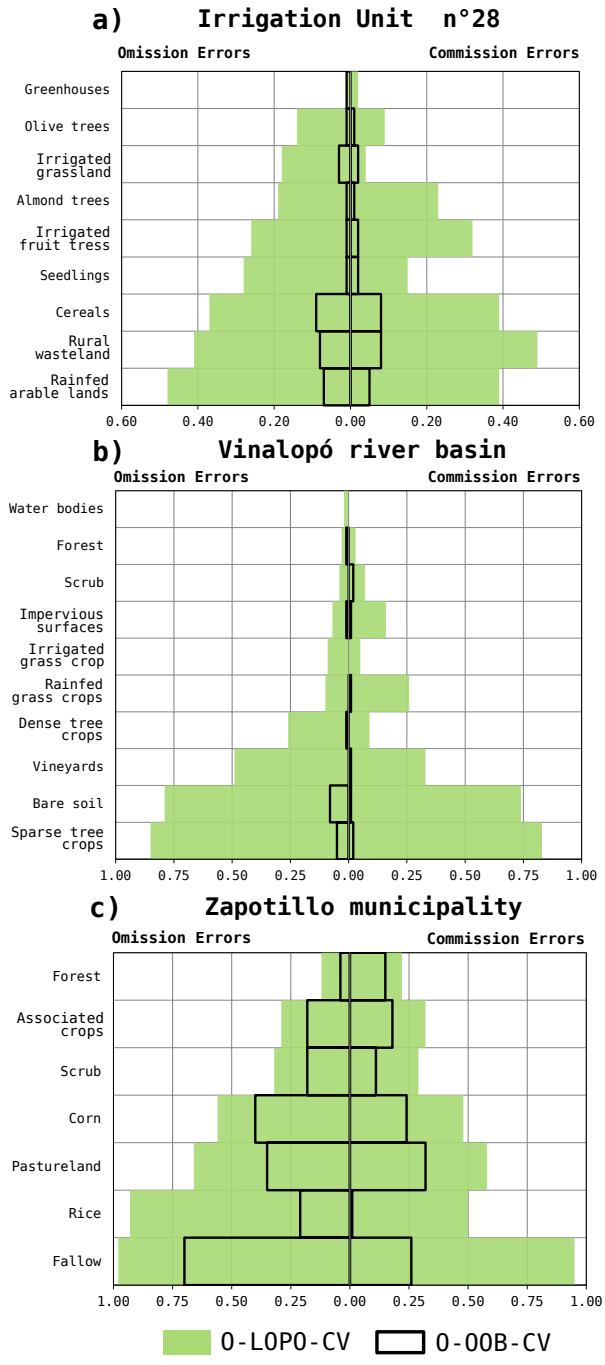

Figure 5: Error pyramids in the three study areas (a) Irrigation Unit 28, b) Vinalopó river basin and c) Zapotillo municipality). Omission and commission errors are compared for O-LOPO-CV and O-OOB-CV. O-LOPO-CV: Leave-one-patch-out cross-validation with original algorithm, O-OOB-CV: Out-of-bag cross-validation using original algorithm.

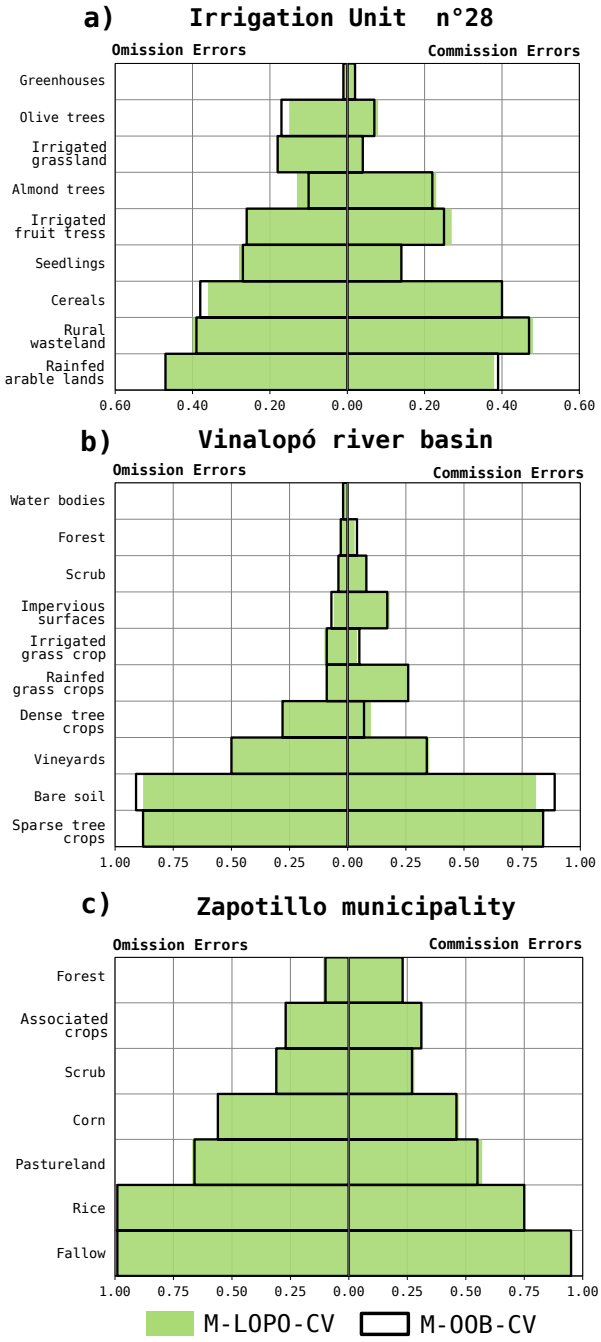

Figure 6: Error Pyramids in the three study areas (a) Irrigation Unit 28, b) Vinalopó river basin and c) Zapotillo municipality). Omission and commission errors are compared for M-LOPO-CV and M-OOB-CV. M-LOPO-CV: Leave-one-patch-out cross-validation with modified algorithm, M-OOB-CV: Out-of-bag cross-validation using modified algorithm.

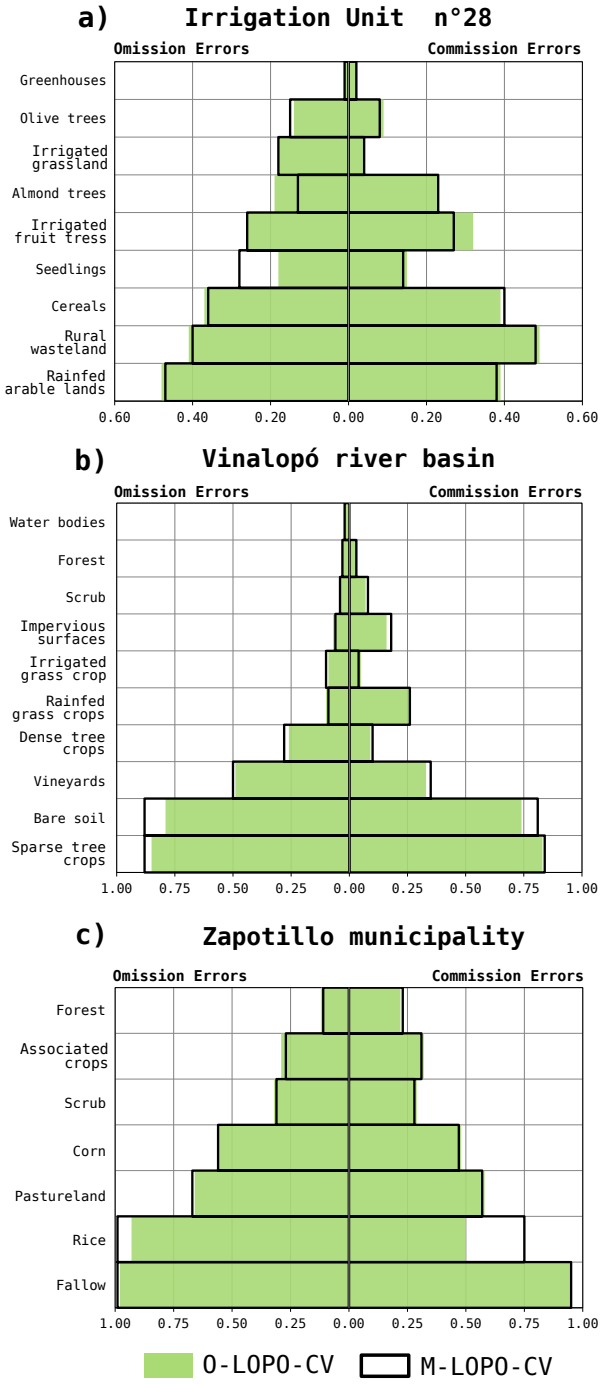

Figure 7: Error Pyramids in the three study areas (a) Irrigation Unit 28, b) Vinalopó river basin and c) Zapotillo municipality). Omission and commission errors are compared for M-LOPO-CV and O-LOPO-CV. O-LOPO-CV: Leave-one-patch-out cross-validation with original algorithm, M-LOPO-CV: Leave-one-patch-out cross-validation with modified algorithm.

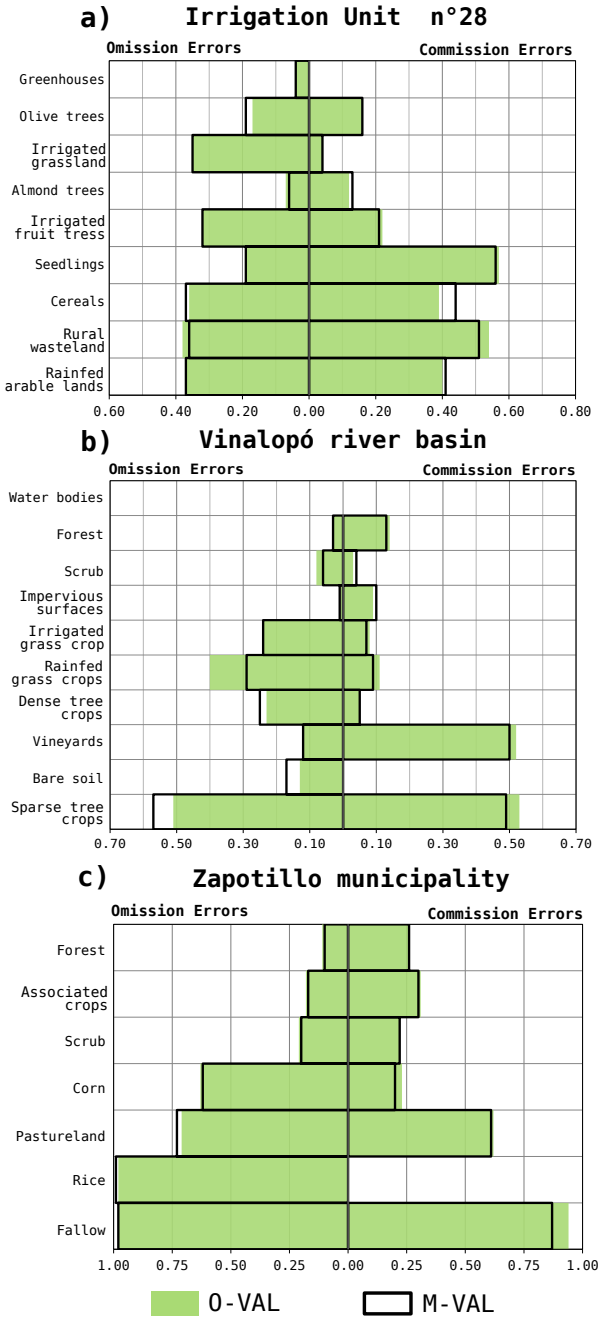

Figure 8: Error Pyramids in the three study areas (a) Irrigation Unit 28, b) Vinalopó river basin and c) Zapotillo municipality). Omission and commission errors are compared for M-VAL and O-VAL. O-VAL: Validation with a different and independent data-set using original algorithm, M-VAL: Validation with a different and independent data-set using modified algorithm.