TITLE:

Comparing the performance of flat and hierarchical habitat/land-cover classification models in a NATURA 2000 site.

AUTHORS:

Yoni Gavish[a,*] gavishyoni@gmail.com
Jerome O'Connell[b] jerome.oconnell@ucd.ie
Charles J. Marsh[a] charliem2003@gmail.com
Cristina Tarantino[c] cristina.tarantino@iia.cnr.it
Palma Blonda[c] palma.blonda@iia.cnr.it
Valeria Tomaselli[d] valeria.tomaselli@ibbr.cnr.it
William E. Kunin[a] W.E.Kunin@leeds.ac.uk

AFFILIATIONS:

[a] School of biology, University of Leeds, Leeds, LS2 9JT, United Kingdom

[b] School of Biosystems and Food Engineering, University College Dublin, D04 N2E5, Ireland

[c] Institute of Atmospheric Pollution Research (IIA), National Research Council (CNR), c/o Interateneo Physics Department, Via Amendola 173, 70126 Bari, Italy

[d] Institute of Biosciences and BioResources (IBBR), National Research Council (CNR-IBBR), via G.Amendola 165/A 70126, Bari, Italy

CORRESPONDENCE:

[*] Corresponding author: Yoni Gavish, School of biology, University of Leeds, Leeds, LS2 9JT, United Kingdom. Mobile: 00-44-75-9999-1988. Email: gavishyoni@gmail.com

Abstract

The increasing need for high quality Habitat/Land-Cover (H/LC) maps has triggered considerable research into novel machine-learning based classification models. In many cases, H/LC classes follow pre-defined hierarchical classification schemes (e.g., CORINE), in which fine H/LC categories are thematically nested within more general categories. However, none of the existing machine-learning algorithms account for this pre-defined hierarchical structure. Here we introduce a novel Random Forest (RF) based application of hierarchical classification, which fits a separate local classification model in every branching point of the thematic tree, and then integrates all the different local models to a single global prediction. We applied the hierarchal RF approach in a NATURA 2000 site in Italy, using two land-cover (CORINE, FAO-LCCS) and one habitat classification scheme (EUNIS) that differ from one another in the shape of the class hierarchy. For all 3 classification schemes, both the hierarchical model and a flat model alternative provided accurate predictions, with kappa values mostly above 0.9 (despite using only 2.2-3.2% of the study area as training cells). The flat approach slightly outperformed the hierarchical models when the hierarchy was relatively simple, while the hierarchical model worked better under more complex thematic hierarchies. Most misclassifications came from habitat pairs that are thematically distant yet spectrally similar. In 2 out of 3 classification schemes, the additional constraints of the hierarchical model resulted with fewer such serious misclassifications relative to the flat model. The hierarchical model also provided valuable information on variable importance which can shed light into “black-box” based machine learning algorithms like RF. We suggest various ways by which hierarchical classification models can increase the accuracy and interpretability of H/LC classification maps.

Key Words

Classification, machine-learning, hierarchical models, random forest, NATURA 2000, Habitat/Land-Cover

Abbreviations (footnote on first page)

H/LC	Habitat/Land-Cover
RF	Random Forest
FRF	Flat Random Forest
HRF	Hierarchical Random Forest
OoB	Out of Bag
H.Step	Hierarchical Stepwise Majority Rule
H.Mult	Hierarchical Multiplicative Majority Rule
FAO-LCCS	Food and Agriculture Organisation's Land Cover Classification System
EUNIS	EUropean Nature Information System habitat classification system
Hie.F	Hierarchical F measure

1. Introduction

Human-mediated changes in the distribution of habitats and land-cover types are one of the main drivers of the global biodiversity crisis. Consequently, providing reliable Habitat/Land-Cover (H/LC) maps for various conservation related issues is of high priority. For example, H/LC maps are used as input layers for species distribution models (Carlson et al., 2014; Coops et al., 2016; Thuiller et al., 2004) or as obligatory background layers for conservation of umbrella species with well-defined habitat requirements (Li and Pimm, 2016; Murphy and Noon, 1992). Furthermore, H/LC maps are fundamental for mapping ecosystem services (e.g., Koschke et al., 2012) and for natural capital assessments (Brown et al., 2016). Finally, in many cases, the habitats themselves are targeted for conservation and management. For example, as part of the EU Habitat Directive (EU, 2007), all member states of the European Union are required to periodically produce H/LC maps and use the maps for change detection and conservation status assessment. Hence further developing our ability to produce H/LC maps at fine thematic and spatial resolution over wide extents is essential for effective conservation, planning, monitoring, reporting and management of natural resources. As a consequence, there has been a recent surge of methodological and conceptual developments in the field of H/LC classification (Blaschke, 2010; Corbane et al., 2015; Lu and Weng, 2007; Lucas et al., 2015; Lucas et al., 2011; Myint et al., 2011; Tso and Mather, 2009; Xie et al., 2008).

In recent years the usage of machine-learning algorithms has become increasingly popular (Belgiu and Drăguţ, 2016) as these machine-learning algorithms are efficient at identifying complex classification rule sets, thus potentially providing accurate classification outputs with relatively little

86 investment of time and effort. In many cases, the H/LC classified by machine-learning algorithms rely
87 on a pre-defined national or international classification schemes (e.g., CORINE), to allow a common
88 language of communication between scientists, management agents and policy makers. Most
89 classification schemes adopt a hierarchical, tree-like structure due to several advantages of such
90 structures. Firstly, classes within a hierarchical classification scheme can be grouped into more
91 abstract classes based on semantic similarity criteria, i.e., a hierarchical H/LC class set comprises
92 several semantic granularities. Secondly, a hierarchical H/LC class set can be applied to a variety of
93 spatial scales (each spatial scale requiring the selection of a scale-specific semantic granularity). The
94 former characteristic is particularly useful to meet the minimum required accuracy standard when a
95 specific subclass accuracy is below this standard (Congalton, 1991) and/or when it is difficult to
96 differentiate between subclasses at a given spatial scale. For example, the European Nature
97 Information System habitat classification scheme (EUNIS) has a tree-like structure with up to eight
98 hierarchical levels, containing a total of 5282 habitat classes (at all levels). CORINE LC has three
99 hierarchical levels, with a total of 44 LC classes. Similarly, classification schemes invented ad-hoc for
100 more local studies may also have a hierarchical structure (e.g., Haest et al., 2017). However, most
101 machine learning algorithms follow a flat classification approach (sensu Silla and Freitas, 2011) in
102 which all H/LCs are classified simultaneously in a 'one-against-all' approach. In other words, machine
103 learning algorithms ignore information on the thematic hierarchy that forms the conceptual basis of
104 most classification schemes. Interestingly, many knowledge-based classifiers follow a top-down
105 approach, in which experts first provide rules (e.g., spectral) that separate general H/LC classes from
106 one another, and then move down the thematic tree while providing more specific rules for more
107 specific H/LC categories (e.g., Lucas et al., 2011; Lucas et al., 2007).

108 There are several reasons why incorporating such hierarchical information into the analytical
109 pathway may be beneficial. First, the rule-sets produced by most machine learning algorithms are a
110 'black-box' to the users because of their size and complexity. It is therefore, very difficult to
111 understand or visualise what variables are important in distinguishing between specific sets of
112 habitats. A hierarchical approach may shed some light into the 'black-box' by providing information
113 on variable importance in various locations along the class hierarchy. Second, habitats that are
114 thematically close to one another are not necessarily ecologically/spectrally similar. For example, a
115 forest and grassland may both be listed under the thematic group of 'non-crop' habitats while a
116 wheat field will occur under the thematic group of 'crops' habitats. However, ecologically and
117 spectrally, the grassland may resemble the wheat field more than the forest. A flat classification
118 approach ignores the thematic proximity altogether, while a hierarchical approach will first invest
119 considerable effort in distinguishing 'crop' from 'non-crop', thus potentially preventing confusion

between spectrally similar yet thematically distant habitats. Third, if the number of habitats is large, the flat approach may not be able to deal with the complexity of the thematic data, while a hierarchical approach could break the problem into manageable portions by partitioning the feature-space of each group into lower dimensions.

Finally, it has been shown that incorporating the hierarchical structure into the modelling framework can increase model accuracy (Thoonen et al., 2013). More specifically, Silla and Freitas (2011) found that various hierarchical approaches tended to increase model accuracy in a wide range of classification problems, especially when misclassifications are weighted by their distance along the classification hierarchy (Kiritchenko et al., 2005). Such hierarchical measures of accuracy acknowledge that not all misclassifications are as critical as the others, e.g., misclassifying one broadleaved-woodland habitat as an alternative closely-related woodland type is arguably a less critical mistake than misclassifying it as a grassland or saltmarsh. In addition, flat classification models only provide performance measures for the entire model or at the H/LC level (i.e., user and producer accuracies). Hierarchical classification models provide the same information with additional accuracy for each local model. That is, the hierarchical approach also provides accuracy for sets of H/LCs that share a common ancestor along the class hierarchy. This information may be crucial for decision makers that may be less interested in the overall accuracy of a map and more by its ability to provide reliable information on sets of H/LCs they care most about (e.g., how well does this model classify non-crop habitats?).

We are aware of only a few published manuscripts that focused on hierarchical, machine-learning based classification methods in the remote-sensing literature. Melgani and Bruzzone (2004) found that several support-vector-machine based hierarchical models outperformed flat models when classifying 9 land-use classes in northwest Indiana. Thoonen et al. (2013) found that a tree-structure Markov random field (TS-MRF) method, which captures the hierarchical thematic structure as well as contextual information, outperformed flat classification methods for heathland areas in Belgium. O'Connell et al. (2015) accounted for spatial hierarchy (nested objects) and thematic hierarchy (2 levels). They reported slightly better classification outcomes (compared to a flat approach) when the probabilities from a Random Forest (RF; Breiman, 2001) model trained at the top level of the thematic hierarchy where included as predictors of RF models trained at the lower level of the thematic hierarchy. Pena et al. (2014) compared flat and hierarchical approaches (based on 4 different algorithms) for mapping cropland areas and found that the flat approach was slightly outperformed by a support-vector-machine based hierarchical model, which fitted a local classifier per parent node. They also found the hierarchical approach increased the minimum sensitivity at the crop level. Finally, Haest et al. (2017) applied an hierarchical classification along four thematic levels

when classifying heathland vegetation types for conservation status assessment. They followed a top-down approach such that the class selected for a given pixel in level 2 of the hierarchy could only be one of the children classes of the class selected in level 1 (with similar rules for levels 3 and 4). Haest et al. (2017) observed higher accuracies for the hierarchical approach compared to a flat approach.

In this paper we introduce a novel application of hierarchical classification based on the RF algorithm, which accounts for the pre-defined hierarchical structure of classification schemes. The application is available for use in a new R package, entitled '*HieRanFor*'. We tested the hierarchical approach in a NATURA 2000 study site from Italy, using three different classification schemes. Our main aim is to compare the performance of the hierarchical and flat approaches and to explore if the variables identified as important at various locations along the class hierarchy provide meaningful ecological knowledge of the system.

2. Flat and Hierarchical Random Forest

RF is a widely used and well-known classifier in remote sensing (Belgiu and Drăguţ, 2016; Bradter et al., 2011). When applied in the flat 'one against all' approach (Flat Random Forest, hereafter, FRF) the model uses a set of training cells and relevant explanatory variables to 'learn' the rules that distinguish one H/LC class from another (Figure 1A,B). The learning procedure is based on fitting a 'forest' of classification trees (Figure 1B). Each tree differs from others in the division of the training data to an 'in bag' set (used for tree growing) and an 'Out of Bag' set (OoB, not used for tree growing) and in the usage of variables during the tree growing procedure (Figure 1A). Each tree assigns each OoB case to a single H/LC class. The results are then usually translated into a soft classification output, by estimating for each case, the proportion of OoB votes that assigned the case to any of the H/LC classes (Figure 1B). The reliance on OoB votes and the uniqueness of the constituting trees result in the method being robust to over-fitting. In addition, as prediction are only provided from OoB votes, performance can be assessed without setting aside a considerable portion of the dataset as an external validation set.

The Hierarchical Random Forest (HRF) takes similar inputs as FRT along with additional information on the thematic class hierarchy (Figure 1C). Then, RF is used as the local classifier at every internal node of the class hierarchy that has at least two child nodes. The training set for each RF consists of the training cases of all descendants of the parent node. For example, in Figure 1C, training cases representing both H4 and H5 are used in classifier C1 to represent H1. The same cases are also used in local classifier C2 to represent H4 or H5, respectively. However, H4 and H5 training cases are not used in classifier C3 which aims to separate H6, H7 and H8 from one another.

After training all the local RFs, it is possible to predict the proportion of votes each case got for each class in each local classifier. Similar to FRF, it is important to ensure the usage of OoB trees when predicting the proportion of votes, otherwise we may overestimate the performance of the model and may be prone to overfitting. For example, a case from the H4 class of figure 1 is included in the training data of classifiers C1 and C2. For these two classifiers, we would only use the votes of trees in which the case was not randomly assigned to the in-bag set (tree training set, Fig. 1A). In classifier C3, which only included cases from H6, H7 and H8 in the training data, all trees can be used for predicting the H4 case probabilities. If a case is never predicted by a classification tree in which it is within the 'in-bag' subset, HRF should retain the robustness of FRF to over-fitting. Similarly, as accuracy is always based on OoB votes, there is no need for an external validation set. Nonetheless, similar to FRF, it is possible to run additional cases through the HRF model for prediction and producing maps or for additional assessment of performance using an external validation set. If the hierarchical model is used to predict for new data, all trees from all local classifiers can be used.

Both FRF and HRF produce for each case the proportion of OoB votes for each H/LC. To assess the performance (e.g., kappa) of the model and to produce maps, a single habitat needs to be selected according to the proportion of OoB votes (i.e., translating the 'soft' probabilities to a 'crisp' classification). For FRF, the flat majority rule is usually applied by selecting for a focal case the H/LC that received the highest proportion of OoB votes. For HRF, there are two different majority rule options – stepwise majority (H.Step) and multiplicative majority (H.Mult, Figure 1C). In stepwise majority rule, the flat majority rule is applied to each local classifier, and starting from the tree root, the selected H/LC is followed until a terminal H/LC (with no descendants lower down the hierarchy) is reached. In the case presented in Figure 1C (blue), H3 received the highest proportion of OoB votes at classifier C1 (0.5 versus 0.1 and 0.4), while H6 received the highest proportion of OoB votes in classifier C3 (0.6 versus 0.3 and 0.1). In the multiplicative majority rule, the proportion of OoB votes are multiplied along every path from the tree root until it reaches the terminal H/LC, and the flat majority rule is applied on the multiplicative proportions. In figure 1C, the multiplicative votes identify H4 (red) as the most probable H/LC. Interestingly, the multiplicative proportion of OoB votes are comparable to the proportion of OoB votes generated by FRF in a sense that both sum to 1 over all terminal nodes.

3. Methods

3.1 Study Site

We have focused our analysis on Le-Cesine, Italy (IT1 – SCI IT9150032; SPA IT9150014) – a Natura 2000 site and one of the oldest protected areas in Puglia. Le-Cesine covers an area of about 2148 ha.

The site is characterized by a high diversity in land-cover, habitat and vegetation types. The coastal wetland is characterized by a system of lagoons, ponds and marshes where several types of helophytic vegetation (reeds, sedges and rushes communities) are widespread. Further inland, the woody vegetation is composed by a mosaic of *Pinus halepensis* stands and different types of Mediterranean maquis and garrigues, while the agricultural areas are mainly composed of olive groves. The site is affected by marine erosion of the sandbank, resulting in reduction and fragmentation of the typical dune habitat types, as well as salinization of the lagoons and the related environments.

3.2 Ground-Truth Data and Classification Schemes

The entire extent of Le-Cesine was ground-truthed in earlier projects to two LC classifications schemes and one H classification scheme. The two LC classification schemes included CORINE (Bossard et al., 2000) and the UN Food and Agriculture Organisation's Land Cover Classification System (FAO-LCCS, Di Gregorio and Jansen, 2005), while for the H classification scheme we relied on EUNIS (Davies and Moss, 2002). The three classification schemes differ from one another in the number of internal and terminal nodes, in the number of hierarchical levels, and in the number of local classifiers required to run the HRF analysis.

A pre-existing validated LC map (scale 1:5000) in CORINE Land Cover was available from a previous Interreg (Nat Info) project. The selection of an appropriate LC classification system for habitat mapping applications is a crucial issue. The FAO-LCCS (Di Gregorio and Jansen, 2005) has been considered as an appropriate and user-friendly framework for long-term monitoring of the conservation status of habitats. LCCS allows the finest discrimination of natural and semi-natural types with respect to other widely used LC taxonomies (Tomaselli et al., 2013). Thus, CORINE classes were first converted to FAO-LCCS classes according to the LCCS2 software (Di Gregorio and Jansen, 2005) and the semantic heterogeneity issues were addressed through the expert knowledge of people with long experience on the monitoring of the study site. Then, FAO-LCCS classes were translated to EUNIS habitat classes by integrating the environmental attributes, i.e. lithology, soil group, soil-surface aspect and water quality which are the auxiliary data that can be used for habitat discrimination.

This information was provided from a previous INTERREG (Nat Info) project. As well known (Adamo et al., 2016; Tomaselli et al., 2013) translating LC to habitats classes may include one-to-many relations, so prior-knowledge from botanists was needed for selecting the environmental data useful to discriminate the different habitat classes that may correspond to a specific LC class. These environmental attributes help to resolve, in most cases, the challenge of one-to-many relationships

253 between LC and habitat classes.Field surveys were then carried out in 2011-2013 to validate both
254 FAO-LCCS and habitat classes, according to the EUNIS report (Davies and Moss, 2002) obtained from
255 each CORINE LC class and to select reference samples for training/testing the RF outputs. As a result,
256 the number of input CORINE classes is lower than the number of FAO-LCCS classes and EUNIS classes
257 (see Figure 2 in Tomaselli et al., 2013). A random sampling design within a 250 m cell regular grid, in
258 turn nested within a 1 km cell standard regular grid (INSPIRE), was selected; within this grid, 50
259 circular 50 m radius vegetation plots, randomly distributed throughout the site and covering all the
260 habitat types (according to EUNIS), were recorded and mapped. For each of these points,
261 information was collected on vegetation composition and structure, crop cover and habitat type
262 (Tomaselli et al., 2016). Such information, geocoded by GPS, was integrated into a GIS geo-database
263 using ArcGIS 9.2.

264 The CORINE classification scheme (Figure 2) for our study system contains 31 land cover classes
265 organized in three hierarchical levels, with 14 of the classes being terminal (i.e., not further
266 described as any other more detailed classes). A total of 8 local RF classifiers are required to fit the
267 HRF model according to the CORINE classification scheme. The FAO-LCCS classification scheme
268 (Figure 3) contains 49 land cover classes (18 terminal) organized along 6 hierarchical levels and
269 requiring 8 local RF classifiers. The EUNIS classification scheme (Figure 4) contains 60 habitats (23 of
270 which are terminal) organized along 5 hierarchical levels and requiring 11 local RF classifiers. Thus,
271 we can explore here the effect of the shape (number of hierarchical levels and distribution of nodes
272 along the tree structure) of the hierarchical classification scheme on the performance of the
273 different classification methods, in the same system using the exact same pixels and explanatory
274 variables.

275 *3.3 Data Preparation and Explanatory Variables*

276 Detailed information on the processing procedure of the images, on the derived indices and on each
277 of the other explanatory variables is found in supporting information S1. We based our analysis on
278 61,453 cells, at 10x10 m resolution covering the entire extent of the study-site. We used a 10x10 m
279 resolution, as preliminary analysis revealed it to produce similar results as finer resolutions (2x2 m
280 and 5x5 m) with considerably lower running time, while still providing detailed enough information
281 for management in this study system. For each cell we calculated 35 explanatory variables. The first
282 12 variables were radiometrically calibrated reflectance values of two Very High Resolution (VHR)
283 remotely sensed images: 1) a multispectral WorldView-2 image with 8 spectral bands at 2m spatial
284 resolution taken on October 9th 2010; 2) a multispectral Quickbird image with 4 spectral bands at 2.4
285 m spatial resolution taken on June 4th 2009. We applied radiometric correction in accordance with

O'Connell et al. (2013) where top-of-atmosphere reflectance followed by dark object subtraction correction was completed. The images were geometrically synced using 221 ground control using a direct linear transform model, giving a root-mean-squared-error of 0.6. Resampling was then done on both images to grids of 10 m using a cubic convolution interpolation, ensuring that both images were radiometrically and spatially aligned.

Five vegetation indices were also derived from each image; Difference Vegetation Index (DVI), Normalised Difference Vegetation Index (NDVI), Principal Components Analysis (PCA), Soil Adjusted Vegetation Index (SAVI) and Atmos Resistant Vegetation Index (ARVI). For the WorldView-2 image we added two additional derived indices- the WorldView-2 Soil Index (WVSI) and WorldView-2 improved Vegetation Index (WVVI) based on the additional red edge, coastal, yellow and near-Infra Red2 bands. For each of the four bands of the Quickbird image we further calculate the Tasselled Cap Index. Therefore a total of 16 spectrally derived variables were created from the two satellite datasets. Furthermore, we estimated 7 environmental variables including elevation, based on a 10 m resolution DEM model (Tarquini et al., 2007; Tarquini et al., 2012) and plant height layers obtained from a LiDAR campaign (canopy height model). The last five variables where categorical environmental variables covering the lithology (3 classes), soil group (6 levels), Soil-Surface aspect (Soil Bare, 6 levels), water quality (6 levels) and cadastral information (16 levels). Such information were collected from local authorities' archives and public repositories.

3.4 Creating Datasets for Classifications

Both the FRF and HRF require a training set containing cases (pixels or object) from each of the H/LC classes. To create a dataset, we first randomly selected from each H/LC 100 cases for the training set. The remaining cases were used as an external, independent validation set. If the total number of available cases for a given H/LC was lower than 100, we included all its cases in the training data. We repeated this procedure 15 times per classification scheme (45 datasets all together), with each dataset containing both a small training subset and a much larger independent validation subset.

3.5 Fitting the FRF and HRF Models for a Single Dataset

As part of the FP7 European project EU BON (Hoffmann et al., 2014) we created the '*HieRanFor*' R package which provides full functionality of applying the above framework (development version is available at https://r-forge.r-project.org/projects/hie-ran-forest/, while a working version is available as SI for this publication). The '*HieRanFor*' uses the RF algorithm as implemented in the '*randomForest*' package (Liaw and Wiener, 2002) in R (R Core Team, 2016). With the '*HieRanFor*'

318 package, users can fit an HRF, predict the proportion of OoB votes in each local classifier (for both
319 the training data and for external validation data) and translate the proportion of OoB votes to a
320 crisp class using either stepwise or multiplicative majority rule. The '*HieRanFor*' package further
321 includes functions to estimate the performance of the model using either flat or hierarchal indices.
322 Finally, the package allows extraction of variable importance values for each class in each local
323 classifier.

324 We first used the training data to fit a FRF model and an HRF model. For the FRF model we used
325 exclusively functions available in the '*randomForest*' package for modelling, extracting the
326 proportion of OoB votes for the training set, and predicting the proportion of votes for the validation
327 set. For the HRF model we first ran the model using the '*RunHRF'* function of '*HieRanFor*'. Then, we
328 used the function '*PerformanceHRF* to extract the proportion of OoB votes for each local classifier,
329 as well as the multiplicative proportion of OoB votes, and the '*PerformanceNewHRF*' function for
330 predicting the proportion of votes for the validation dataset. At this stage, we had for each case of
331 the training and validation set the proportion of OoB votes required to translate the soft
332 classification into a crisp one.

333 In FRF, we translated the soft classification into crisp by employing a simple majority rule, i.e.,
334 selecting for each case the H/LC that received the highest proportion of OoB votes. For HRF, we used
335 both the stepwise majority rule and the multiplicative majority rule (see above and in Figure 1C). The
336 crisp H/LC for each case were then used to assess the performance of the models separately for the
337 FRF, the stepwise HRF, and the multiplicative HRF. We estimated performance separately for the
338 training and validation set, using the overall accuracy (based on the diagonal of the error matrix),
339 Cohen's Kappa (Cohen, 1968), and the Hierarchical F measure (Kiritchenko et al., 2005), for a total of
340 18 performance values for each dataset. The unweighted kappa and accuracy were computed using
341 the *confusionMatrix* function of the '*caret*' R package (Kuhn et al., 2016). For the Hierarchical F
342 measure (Hie.F), we implemented within *'HieRanFor*' a function that calculates the Kiritchenko et al.
343 (2005) index directly from the confusion matrix. Additional information on Hie.F can be found in
344 supporting information 2.

345 *3.6 Variable Importance*

346 In FRF, variable importance is quantified as the mean decrease in accuracy of the entire model when
347 the values of a focal variable are permutated, and all other values remain unchanged. In HRF,
348 variable importance is estimated separately for each variable in each local classifier, thereby
349 allowing more information on why certain variables are more important than others. We explored

the variable importance separately for each classification scheme by taking the mean over all 15 runs.

3.7 Statistical Analysis

After fitting the FRF and HRF for the 45 datasets following the procedure outlined above, we explored the models performance using mixed-effect models. We used one of the model performance indices (kappa, accuracy or Hie.F measure) as the dependent variables. For fixed effects, we used the data type (2 levels- training or validation), the crisp rule (3 levels: simple majority for the FRF, stepwise for the HRF and multiplicative for the HRF) and their interaction. For random effects we used the data type, grouped according to the model run (i.e., the specific set of training and validation datasets created randomly -- a paired design). We used this model since it received the lowest AICc (Akaike Information Criteria corrected for small sample size values) from a set of 9 nested mixed effect models (supporting information S3). We repeated the analysis separately for each of the three H/LC classification schemes.

4. Results

In each run, we have used a total of 1333, 1661 and 1980 pixels in the training set for CORINE, FAO-LCCS and EUNIS, respectively -- only 2.2-3.2% of the study area's 61453 pixels (supporting information S4 for breakdown of the training set to classes and for examples of several confusion matrices). Despite the small training set, all methods in all classification schemes produced maps that are very similar to the observed map, both for the small training set and the large validation set (see examples in Figures 2-4). In all 3 classification schemes, we found the environmental variables to have considerably higher variable importance values relative to the spectral reflectance values and the vegetation indices derived from the remotely-sensed images (Figure 2-4, lower left panels). Among the environmental variables, water quality, soil group, soil bare and lithology showed the highest variable importance values, followed by the cadastral and elevation variables, and finally canopy height. This pattern was observed for both the FRF and HRF approaches in all 3 classification schemes. However, the HRF approach provided more detailed information on the importance of each variable in each local classifier, flagging different variables as important in different locations along the hierarchy, even if they had relatively low importance when averaged across all categories using FRF. For example, elevation was identified as the most important variable in local classifier C.8 for CORINE (Figure 2). Similarly, the cadastral variable received relatively low relative importance score in the flat model, but it was identified as the most important variable in one of the local classifiers of the FAO-LCCS (C.4 in Figure 3). This may be due to the fact that cadastral allows the discrimination of barren land from artificial structures for objects characterized by similar spectral

signature, when additional context-sensitive features are not considered (e.g., texture). For EUNIS, water quality was identified as having very low importance for some classifiers but was very important in several others, including being the only important variable for separating the two children nodes of habitat A2.52 (Figure 4, C.10). This is justified by the importance of water salinity for discriminating EUNIS habitats such as C3.421 (fresh water) and A2.51 (salt water) corresponding to the same FAO-LCCS class (A24/A2.A5.E7, i.e., inland water habitats (see Table 6 in Tomaselli et al., 2013).

The overall performance of both the FRF and HRF was very high. Kappa values were above 0.9 for CORINE and FAO-LCCS, with slightly lower values for EUNIS (Figure 5, middle column). Accuracy values were above 0.9 for CORINE and FAO-LCCS, i.e., more than 90% of the cells were correctly classified (Figure 5, left column). Accuracy for EUNIS was slightly lower, but still above 0.86 in all runs. Similar high values were also observed for the Hie.F measure (Figure 5, right column). The variance between the 15 runs for a given combination of performance index and classification scheme was very low, suggesting that the choice of training set had little effect on the results. Given the overall high performance values, it was hard to detect performance differences between the FRF and HRF approaches (Figure 5). In CORINE we observed a slight decrease in all three performance measures in the two HRF approaches relative to the FRF, for both the training and validation datasets. For EUNIS, FRF performed slightly better in the training set, yet the two HRF approaches slightly outperformed the FRF in the validation set. A similar pattern regarding the hierarchical stepwise approach was also observed for the FAO-LCCS classification scheme. When comparing the medians, while the FRF outperformed the two HRF in all 9 cases in the training set (3 classification schemes × 3 performance indices), the HRF outperformed the FRF in 6 out of 9 cases in the validation set. The hierarchical multiplicative approach outperformed the hierarchical stepwise approach in EUNIS and CORINE, while the hierarchical stepwise approach outperformed the hierarchical multiplicative approach in FAO-LCCS.

In the validation datasets, the pairs of H/LC that were responsible for the highest number of classification errors in each of the three schemes are summarized in table 1. In general, the main source of confusion over all schemes is amongst the olive groves and conifer plantations. The inclusion of more images covering the seasonal cycle and/or the introduction of context-sensitive features related to agricultural practices may improve the discrimination (e.g., olive trees are organized in parallel rows at regular distances). These habitats were either misclassified as one another, or as road or fields. In CORINE, confusion between *Coniferous Forest* (3.1.2) or *Olive Groves* (2.2.3) on one side and *Permanently Irrigated Lands* (2.1.2) or *Road and Rail Networks and Associated Land* (1.2.2) on the other side accounted for 71.5%, 76.9% and 75.0% of all classification

errors for the flat, H.Mult and H.Step approaches, respectively. In FAO-LCCS, confusion between *Plantations: needle-leaved evergreen tree crops- monoculture + rainfed* (A11_A1.B1.A8.A9.B3.W7.C1.D1) or *Orchards: broad-leaved evergreen tree crops- monoculture + rainfed* (A11_A1.B1.A7.A9.B4.W8.C1.D1) on one side and *Fields of irrigated no graminoid crops + one additional crop* (A11_A3.A5.B2.C2.D3) or *Paved roads* (B15_A1.A3.A7.A8) on the other side accounted for 55.8%, 55.5% and 47.2% of all classification errors for the flat, H.Mult and H.Step approaches, respectively. In EUNIS, confusion between *Native Conifer Plantations* (G3.F1) or *Olea europaea Groves* (G2.91) on one side and *Arable Land with Unmixed Crops Grown by Low-intensity Agricultural Methods* (I1.3) or *Road Networks* (J4.2) on the other side accounted for 40.3%, 40.0% and 38.4% of all classification errors for the flat, H.Mult and H.Step approaches, respectively. In addition, for EUNIS (and to a lesser extent, for FAO-LCCS) misclassification of different types of salt marshes as one another also contributed considerably to the overall error rate. These errors where avoided in CORINE due to the lower thematic resolution.

5. Discussion

5.1 Performance of the Flat and Hierarchical Models

In this manuscript we introduce the HRF: a machine learning classification algorithm that accounts for the pre-defined hierarchical structure of user-defined classification schemes. Our main objective was to explore whether this novel hierarchical approach provides better performance relative to the commonly used flat classification approach. We further explored whether the additional information provided by the HRF approach could shed light into the 'black-box' of RF. A single study site was used, for which we had complete independent coverage data based on two different LC and one H classification schemes. In general, we have found that for all three classification schemes, both the FRF and the HRF were able to predict the observed H/LC with high accuracy (Figure 5), despite using as low as 2.2% of the study site as training cells. This was observed not only in the training sets, but also in the external, much larger and independent validation sets.

Given the very high performance of both FRF and HRF, it is difficult to identify explicit differences between the two approaches. The difference in performance between the FRF, H.Mult and H.Step were very small. However there were some differences, with the flat approach best for CORINE, the H.Step best for FAO-LCCS and the H.Mult best for EUNIS (Figure 5). Although a single case-study is not enough for generalization, these differences may be attributed to differences in class hierarchy. The hierarchical approaches outperformed the flat one in the two classification schemes that have a complex hierarchical structure, with more nodes and more levels. In addition, the H.Step outperformed the H.Mult in FAO-LCCS, perhaps since FAO-LCCS classifies only two classes

in the local classifier closest to the tree root (C.1 in Figure 3) while EUNIS already has ten classes in C.1 (Figure 4). The small number of classes in the first classifier provides an advantage to the stepwise approach, with the higher weights it gives to the early classifier affecting the overall performance. Thus, based on this single case-study we suggest use of the flat approach when the hierarchy is relatively simple, the H.Step approach when the hierarchy is more complex but with simple early classifiers and the H.Mult when the hierarchy is complex right from the start.

In fact, as clearly noted by Haest et al. (2017), the main advantage of the H.Step approach may be its ability to separate thematically distinct but spectrally similar classes, assuming the classifier close to the tree-root performs well. However, when this classifier does not perform well, the errors will be carried down the class structure with no option to undo them. The H.Mult approach does not give higher weights to classifiers close to the tree root and thus errors made in certain classifiers have only a limited effect on the overall predicted probabilities. Instead, the probabilities are multiplied along the tree root along every path. Thus, probabilities of classes that have more children nodes further down the hierarchy may be diluted relative to other classes that terminate closer to the tree root. This can potentially bias the method toward shallower classes. If such a bias affected our results, we would expect the H.Mult to classify more cases to classes that are closer to the tree root than the flat and H.Step approaches. We have not found any evidence for such a bias when comparing the number of cases classified to each level in the two classification schemes that have terminal classes at different levels (FAO-LCCS and EUNIS).

5.2 Types of Misclassifications

Both FRF and HRF made most misclassification errors when trying to classify two main H/LC types – the conifer plantations and the olive groves (table 1). In these H/LC types, the actual cover of trees is considerably lower than 100%. The gaps between the trees are covered with bare soil or with green/dry annual vegetation (depending on the intensity of grazing, the agricultural practice and the time of the year). Thus, the main source of confusion was treating the gaps' bare soil area as roads or the gaps' annuals areas as agricultural fields. These misclassifications most likely represent cases that are thematically distant yet spectrally similar. Our general expectation was that HRF will make fewer such errors than FRF, since local classifiers closer to the tree root would already separate the H/LC at early stage. We further expected the stepwise approach to be better than the multiplicative approach in that respect, since the stepwise approach gives higher weights to local classifiers closer to the tree root while the multiplicative approach gives equal weights to all local classifiers. Indeed, we observed such a pattern for EUNIS, with 2669 vs. 2830 misclassifications of the above types for H.Step and FRF, respectively. Similarly, for the FAO-LCCS there were 2291 vs. 2851 such

misclassifications for the H.Step and FRF, respectively. For the FAO-LCCS, the difficultly in classifying gaps resulted for H.Step mainly in confusion between the orchards and plantations, that are classified at the same local classifier (C.4 in Figure 3). Interestingly, in CORINE, the flat approach outperformed the H.Step (2793 vs. 3178 misclassifications, respectively).

5.3 Variable importance

In all three classification schemes, the environmental variables were more important than the raw spectral variables or the variables derived from the remotely sensed images (lower left panels in figures 2-4). Perhaps, the poor performance of the spectral variables is due to a scale issue; spectral variables in high resolution images change over very short distances on the ground (even after resampling) but the environmental variables were based on relatively coarse resolution data and may have tied in better with the spatial scale of the classification. This can be a common issue in complex ecosystems or habitats. In addition, the remote sensing data were only from two dates (one of which was approaching winter leaf-off) and the spectral variables may increase in importance if more dates were available at key phenological stages in the season.

Nonetheless, the HRF approach also provided more detailed information on the importance of each variable in each local classifier. For example, in the EUNIS classification scheme (Figure 4) the local classifier C.10 reveals that distinguishing between the *Mediterranean Juncus maritimus and Juncus acutus saltmarshes* (A2.522) and *Mediterranean saltmarsh scrubs* (A2.526) requires information on water quality (salinity). However, distinguishing between *Marine saline beds of Phragmites australis* (A2.53C) and *Geolittoral wetlands and meadows: saline and brackish reed, rush and sedge stands* (A2.53D) in local classifier C.11 requires information on both water quality and soil groups, perhaps because A2.53C has a much wider ecological range with its distribution ranging from the dune area to the inland area. In FAO-LCCS (Figure 3) the first local classifier (C.1) reveals that none of the available explanatory variables are extremely important in distinguishing between primarily vegetated areas (habitat A) and primarily non-vegetated areas habitat (B). In the FAO-LCCS (Figure 3), the explanatory variables of cadastral information and the canopy height model were not identified as important by the flat model. However, these are the main explanatory variables that distinguish between the three descendent H/LC of *Cultivated and Managed Terrestrial Areas* (A11) in local classifier C.4. Similarly, in the CORINE classification scheme (Figure 2), the elevation variable is not identified as important by the flat model, yet the HRF analysis reveals it to have a relatively high effect on accuracy in at least 2 local classifiers (e.g., C.2 and C.8).

5.4 The Potential of Hierarchical Classification Models

The hierarchical approach we tested here is only one of several potential hierarchical approaches that have been used for classification problems in other research areas. In fact, in their review, Silla and Freitas (2011) list several types of hierarchical models including:

- *Local classifier per parent node* – a local classification model is fitted in every internal node that has more than one sibling. This is the approach followed here and in Pena et al. (2014).
- *Local classifier per node*- a binary model is fitted to each internal and terminal node, e.g., 9 local models in the example of Figure 1, each for one habitat (H1-H8).
- *Local classifier per level* – A local model is fitted for each level of the hierarchy, classifying all the habitats in the level. In the example of Figure 1, two classification models would be fitted, one for classifying H1, H2 and H3 and one classifying H4, H5, H6, H7 and H8.
- Global classifier – a single model is fitted which simultaneously captures the entire class hierarchy. As far as we know, this approach has never been tested using machine learning algorithm for remote-sensing. Interestingly, knowledge-based classifiers, which tend to follow a hierarchical classification structure, are the closest example we could think of.

It is probable that different approaches may be more suitable for particular applications. For example, the local classifier per node approach may be most suitable for situations in which not all cases in the training data are classified to the lowest level (e.g., some cases in Figure 1C are labelled as H1, without further labelling as either A4 or A5). Alternatively, local classifier per level may be most suitable for cases in which all terminal classes are at the same level (e.g., CORINE), especially if the probabilities of models closer to the tree root are entered as explanatory variables in models lower down the hierarchy (O'Connell et al., 2015). The local classifier per parent node approach may better suit cases in which the hierarchy is more complex (e.g., EUNIS).

We can further envision usage of hierarchical models even in the absence of a pre-defined thematic scheme as a way to focus models on sets of classes that are not easily distinguished from one another. In the local classifier per parent approach this can be done by first fitting a flat model to produce a confusion matrix. Then hierarchical cluster analysis may be used to create the thematic hierarchy, such that classes that are commonly confused (e.g. olive and pine plantations) will be closer to each other in the thematic tree. A hierarchical model based on such a tree could then fit specific local classifiers to the most hard-to-distinguish sets.

Even if the marginal increase in performance of hierarchical vs. flat models would be low, the more detailed variable importance information of hierarchical models may provide information on the processes that govern the distribution of various habitats. Similarly, the accuracy of specific local classifiers may help identify sets of habitats that none of our predictors can distinguish between.

548 Finally, hierarchical models can be used as an automated tool to lower the costs of knowledge-based
549 classification models, by suggesting variables that are most suitable at distinguishing groups of
550 habitats from one another.

551 **6. Conclusion**

552 This is one of the first applications of a machine learning classification algorithm that accounts for a
553 pre-defined hierarchical thematic structure (but see: Haest et al., 2017; O'Connell et al., 2015; Pena
554 et al., 2014; Thoonen et al., 2013). We did not observe considerable differences in performance
555 between this new hierarchical approach and a typical flat approach. This is in accordance with Pena
556 et al. (2014) that followed a similar local classifier per parent node approach. However, our results
557 do suggest that the best strategy may depend on the complexity of the hierarchical structure.
558 Furthermore, we found that the hierarchical approach provides valuable information on variable
559 importance that shed some light into the usually 'black-box' of machine-learning algorithms. We
560 hope that this work will trigger additional works on the potential of using hierarchical models for
561 H/LC classification.

562

563 **Acknowledgment**

564 We wish to thank the associate editor prof. Jay Gao. We further thank Birgen Haest and an
565 anonymous reviewer for their valuable and constructive comments that improved the quality and
566 clarity of this manuscript.

567 **Funding**

568 This manuscript is a collaboration between two European projects. YG, JO, CJM and WEK were
569 financed by the EU BON project (www.eubon.eu) that is a 7th Framework Programme funded by the
570 European Union under Contract No. 308454. This work was also supported by the Horizon2020
571 ECOPOTENTIAL project: improving future ecosystem benefits through earth observations, grant
572 agreement 641762 (www.ecopotential-project.eu).

573

574 **References**

575 Adamo, M., Tarantino, C., Tomaselli, V., Veronico, G., Nagendra, H., Blonda, P., 2016. Habitat
576 mapping of coastal wetlands using expert knowledge and Earth observation data. J. Appl. Ecol.
577 53, 1521-1532.

578 Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future
579 directions. ISPRS-J. Photogramm. Remote Sens. 114, 24-31.
580 Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS-J. Photogramm. Remote
581 Sens. 65, 2-16.
582 Bossard, M., Feranec, J., Otahel, J., 2000. CORINE land cover technical guide-addendum 2000
583 Technical Report, No 40, European Environmental Agency.
584 Bradter, U., Thom, T.J., Altringham, J.D., Kunin, W.E., Benton, T.G., 2011. Prediction of National
585 Vegetation Classification communities in the British uplands using environmental data at multiple
586 spatial scales, aerial images and the classifier random forest. J. Appl. Ecol. 48, 1057-1065.
587 Breiman, L., 2001. Random forests. Mach. Learn. 45, 5-32.
588 Brown, C., King, S., Ling, M., Bowles-Newark, N., Ingwall-King, L., Wilson, L., Pietilä, K., Regan, E.C.,
589 Vause, J., 2016. Natural Capital Assessments at the National and Sub-national Level. UNEP-
590 WCMC, Cambridge, UK.
591 Carlson, B.Z., Georges, D., Rabatel, A., Randin, C.F., Renaud, J., Delestrade, A., Zimmermann, N.E.,
592 Choler, P., Thuiller, W., 2014. Accounting for tree line shift, glacier retreat and primary succession
593 in mountain plant distribution models. Diversity and Distributions 20, 1379-1391.
594 Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement
595 or partial credit. Psychological bulletin 70, 213-220.
596 Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data.
597 Remote Sensing of Environment 37, 35-46.
598 Coops, N.C., Waring, R.H., Plowright, A., Lee, J., Dilts, T.E., 2016. Using Remotely-Sensed Land Cover
599 and Distribution Modeling to Estimate Tree Species Migration in the Pacific Northwest Region of
600 North America. Remote Sens. 8.
601 Corbane, C., Lang, S., Pipkins, K., Alleaume, S., Deshayes, M., García Millán, V.E., Strasser, T., Vanden
602 Borre, J., Toon, S., Michael, F., 2015. Remote sensing for mapping natural habitats and their
603 conservation status – New opportunities and challenges. Int. J. Appl. Earth Obs. Geoinf. 37, 7-16.
604 Davies, C.E., Moss, D., 2002. EUNIS habitat classification. Final report to the European topic centre of
605 nature protection and biodiversity, European Environment Agency, Swindon.
606 Di Gregorio, A., Jansen, L.J.M., 2005. Land Cover Classification System (LCCS): classification concepts
607 and user manual. Food and Agriculture Organization of the United Nations, Rome.
608 EU, 2007. Habitats Directive. In: Commission, E. (Ed.), Article 10.
609 Haest, B., Vanden Borre, J., Spanhove, T., Thoonen, G., Delalieux, S., Kooistra, L., Mücher, C.,
610 Paelinckx, D., Scheunders, P., Kempeneers, P., 2017. Habitat Mapping and Quality Assessment of
611 NATURA 2000 Heathland Using Airborne Imaging Spectroscopy. Remote Sens. 9, 266.
612 Hoffmann, A., Penner, J., Vohland, K., Cramer, W., Doubleday, R., Henle, K., Kõljalg, U., Kühn, I.,
613 Kunin, W.E., Negro, J., Penev, L., Rodríguez, C., Saarenmaa, H., Schmeller, D., Stoev, P.,
614 Sutherland, W., Tuama, É.Ó., Wetzel, F., Häuser, C., 2014. The need for an integrated biodiversity
615 policy support process – Building the European contribution to a global Biodiversity Observation
616 Network (EU BON). . Nature Conservation 6, 49-65.
617 Kiritchenko, S., Matwin, S., Famili, F., 2005. Functional annotation of genes using hierarchical text
618 categorization, In: Proc. of the ACL Workshop on Linking Biological Literature, Ontologies and
619 Databases: Mining Biological Semantics.
620 Koschke, L., Fürst, C., Frank, S., Makeschin, F., 2012. A multi-criteria approach for an integrated land-
621 cover-based assessment of ecosystem services provision to support landscape planning. Ecol.
622 Indic. 21, 54-66.
623 Kuhn, M., Contributions from: Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper,
624 T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y.,
625 Candan, C., Hunt, T., 2016. caret: Classification and Regression Training. R package. version 6.0-
626 73. https://CRAN.R-project.org/package=caret.
627 Li, B.V., Pimm, S.L., 2016. China's endemic vertebrates sheltering under the protective umbrella of
628 the giant panda. Conserv. Biol. 30, 329-339.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18-22.
Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823-870.
Lucas, R., Blonda, P., Bunting, P., Jones, G., Inglada, J., Arias, M., Kosmidou, V., Petrou, Z.I., Manakos, I., Adamo, M., Charnock, R., Tarantino, C., Mücher, C.A., Jongman, R.H.G., Kramer, H., Arvor, D., Honrado, J.P., Mairota, P., 2015. The Earth Observation Data for Habitat Monitoring (EODHaM) system. Int. J. Appl. Earth Obs. Geoinf. 37, 17-28.
Lucas, R., Medcalf, K., Brown, A., Bunting, P., Breyer, J., Clewley, D., Keyworth, S., Blackmore, P., 2011. Updating the Phase 1 habitat map of Wales, UK, using satellite sensor data. ISPRS-J. Photogramm. Remote Sens. 66, 81-102.
Lucas, R., Rowlands, A., Brown, A., Keyworth, S., Bunting, P., 2007. Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping. ISPRS-J. Photogramm. Remote Sens. 62, 165-185.
Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sensing 42, 1778-1790.
Murphy, D.D., Noon, B.R., 1992. Integrating scientific methods with habitat conservation planning - reserve design for northern spotted owls. Ecol. Appl. 2, 3-17.
Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q.H., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sensing of Environment 115, 1145-1161.
O'Connell, J., Bradter, U., Benton, T.G., 2015. Wide-area mapping of small-scale features in agricultural landscapes using airborne remote sensing. ISPRS-J. Photogramm. Remote Sens. 109, 165-177.
O'Connell, J., Connolly, J., Vermote, E.F., Holden, N.M., 2013. Radiometric normalization for change detection in peatlands: a modified temporal invariant cluster approach. Int. J. Remote Sens. 34, 2905-2924.
Pena, J.M., Gutierrez, P.A., Hervas-Martinez, C., Six, J., Plant, R.E., Lopez-Granados, F., 2014. Object-Based Image Classification of Summer Crops with Machine Learning Methods. Remote Sens. 6, 5019-5041.
R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.
Silla, C.N., Freitas, A.A., 2011. A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. 22, 31-72.
Tarquini, S., Isola, I., Favalli, M., Mazzarini, F., Bisson, M., Pareschi, M.T., Boschi, E., 2007. TINITALY/01: a new Triangular Irregular Network of Italy.
Tarquini, S., Vinci, S., Favalli, M., Doumaz, F., Fornaciai, A., Nannipieri, L., 2012. Release of a 10-m-resolution DEM for the Italian territory: Comparison with global-coverage DEMs and anaglyph-mode exploration via the web. Computers & Geosciences 38, 168-170.
Thoonen, G., Spanhove, T., Vanden Borre, J., Scheunders, P., 2013. Classification of heathland vegetation in a hierarchical contextual framework. Int. J. Remote Sens. 34, 96-111.
Thuiller, W., Araujo, M.B., Lavorel, S., 2004. Do we need land-cover data to model species distributions in Europe? Journal of Biogeography 31, 353-361.
Tomaselli, V., Adamo, M., Veronico, G., Sciandrello, S., Tarantino, C., Dimopoulos, P., Medagli, P., Nagendra, H., Blonda, P., 2016. Definition and application of expert knowledge on vegetation pattern, phenology, and seasonality for habitat mapping, as exemplified in a Mediterranean coastal site. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology, 1-13.
Tomaselli, V., Dimopoulos, P., Marangi, C., Kallimanis, A.S., Adamo, M., Tarantino, C., Panitsa, M., Terzi, M., Veronico, G., Lovergine, F., Nagendra, H., Lucas, R., Mairota, P., Mucher, C.A., Blonda, P., 2013. Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a Mediterranean assessment. Landscape Ecology 28, 905-930.

Tso, B., Mather, P., 2009. Classification methods for remotely sensed data. CRC Press, Taylor & Francis Group.
Xie, Y., Sha, Z., Yu, M., 2008. Remote sensing imagery in vegetation mapping: a review. J. Plant Ecol. 1, 9-23.

SUPPORTING INFORMATION

- Supporting Information S1 – The image processing procedure, the derived indices and the environmental explanatory variables.
- Supporting Information S2 – Information on the Hierarchical F measure
- Supporting Information S3 –Results from the nine mixed effect models.
- Supporting information S4 – excel file with example confusion matrices and frequency of classes in the training and validation sets.
- Supporting information S5 – zip file with the '*HieRanFor*' package, verified to work on R version 3.3.1 (July 2017).

695 TABLES

696 **Table 1:** The top 5 pairs of habitat/land-cover that were responsible for the highest number of classification
697 errors in the validation sets in each of the three classification schemes. Results for the flat, hierarchical
698 multiplicative (H.Mult) and hierarchical stepwise (H.Step) are given as the mean over all 15 runs. Number of
699 errors is the total number of mix-up (i.e., number of times A was classified as B plus the number of times B was
700 classified as A).

Habitat/Land-cover 1	Habitat/Land-cover 2	Number of errors (rank)		
		Flat	H.Mult	H.Step
CORINE	Total errors:	3907	3956	4239
Coniferous forest	Road and rail networks and associated land	1718 (1)	1890 (1)	1964 (1)
Olive groves	Permanently irrigated land	438 (2)	488 (2)	508 (2)
Olive groves	Road and rail networks and associated land	412 (3)	400 (3)	389 (3)
Coniferous forest	Discontinuous urban fabric	312 (4)	181 (6)	299 (5)
Coniferous forest	Olive groves	245 (5)	253 (5)	250 (6)
Coniferous forest	Permanently irrigated land	225 (6)	265 (4)	318 (4)
FAO-LCCS	Total errors:	5107	5310	4857
Plantations: needleleaved evergreen tree crops- monoculture + rainfed	Paved roads	1415 (1)	1448 (1)	877 (2)
Plantations: needleleaved evergreen tree crops- monoculture + rainfed	Orchards: broadleaved evergreen tree crops- monoculture + rainfed	1144 (2)	1213 (2)	1469 (1)
Orchards: broadleaved evergreen tree crops- monoculture + rainfed	Paved roads	773 (3)	752 (3)	537 (4)
Orchards: broadleaved evergreen tree crops- monoculture + rainfed	Fields of irrigated no graminoid crops + one additional crop	500 (4)	544 (4)	602 (3)
Orchards: broadleaved evergreen tree crops- monoculture + rainfed	Scattered industrial or other areas	312 (5)	294 (6)	271 (7)
Temporarily flooded land with perennial closed tall grasslands	Temporarily flooded land with *Aphyllous* closed dwarf shrubs	299 (6)	369 (5)	369 (5)
EUNIS	Total errors:	7025	6885	6955
Fen *Cladium mariscus* beds	Marine saline beds of *Phragmites australis*	3177 (1)	3115 (1)	3129 (1)
Native conifer plantations	Road networks	1711 (2)	1635 (2)	1635 (2)
Olea europaea groves	Road networks	467 (3)	425 (4)	442 (3)
Olea europaea groves	Arable land with unmixed crops grown by low-intensity agricultural methods	421 (4)	431 (3)	389 (5)
Native conifer plantations	Scattered residential buildings	316 (5)	208 (7)	278 (6)
Native conifer plantations	*Olea europaea* groves	235 (6)	362 (5)	407 (4)

701

702

703

FIGURE LEGENDS

Figure 1: Main concept of hierarchical random forest.

Flow chart for creating: (A) a single classification tree, (B) a single randomForest model, and (C) a single hierarchical randomForest model. See text for details. Prop' – proportion; OoB-- Out-of-Bag.

Figure 2: The results for CORINE.

The top panel is an example of the observed map and the maps predicted by a flat RF model, by Hierarchical multiplicative RF model and by hierarchical stepwise RF. Land-covers legend is given in the class hierarchy (lower right panel) alongside the location of seven local classifiers (e.g., C.1). The lower left panel shows the variable importance values in each local classifier and in the flat model. The variables are divided to environmental variables (blue), variables derived from the remote sensing images (black) and the raw reflectance values of the two remotely-sensed images (red).

Figure 3: The results for FAO-LCCS.

The top panel is an example of the observed map and the maps predicted by a flat RF model, by Hierarchical multiplicative RF model and by hierarchical stepwise RF. Land-covers legend is given in the class hierarchy (lower right panel) alongside the location of eight local classifiers (e.g., C.1). The lower left panel shows the variable importance values in each local classifier and in the flat model. The variables are divided to environmental variables (blue), variables derived from the remote sensing images (black) and the raw reflectance values of the two remotely-sensed images (red).

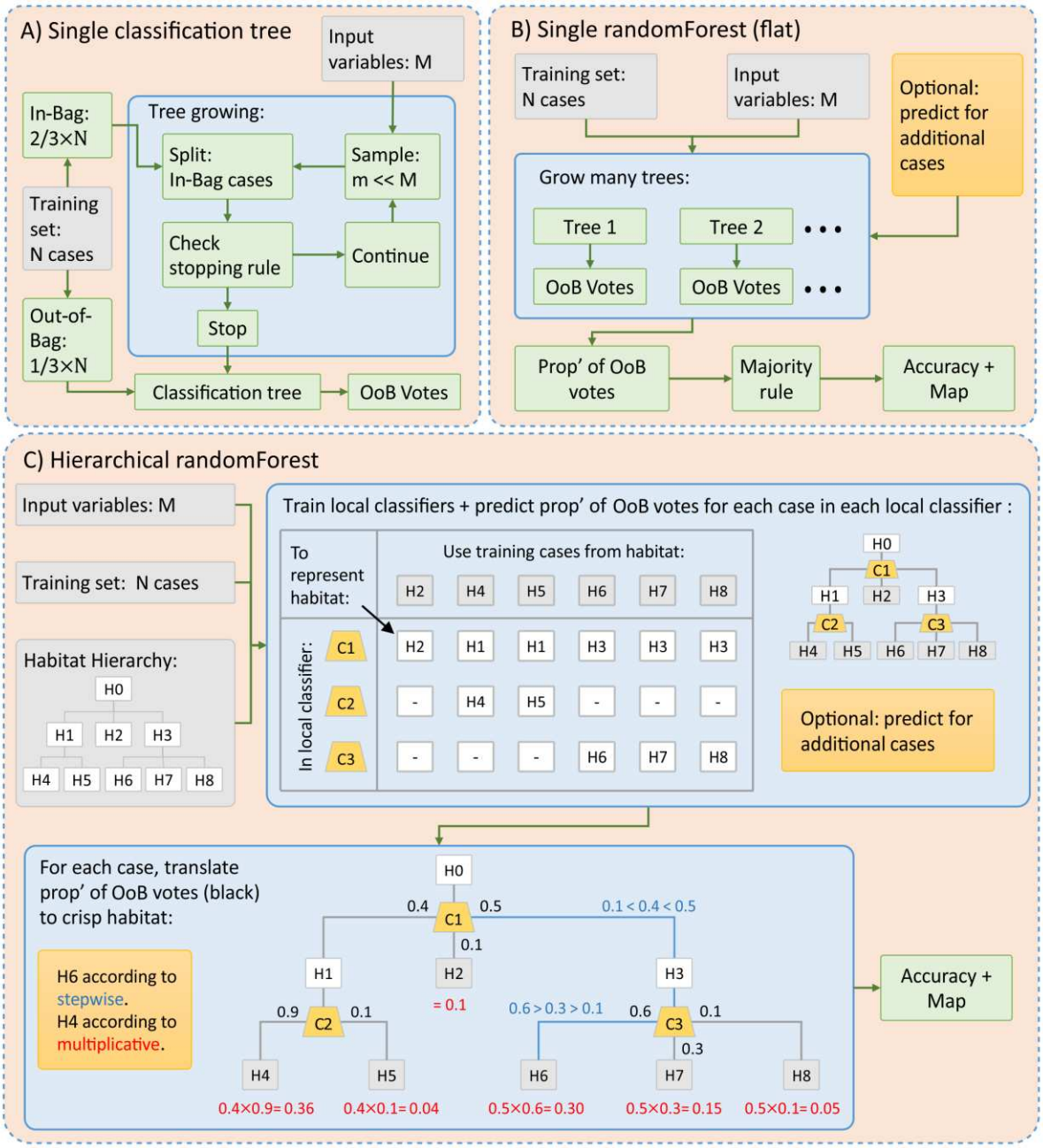
Figure 4: The results for EUNIS.

The top panel is an example of the observed map and the maps predicted by a flat RF model, by Hierarchical multiplicative RF model and by hierarchical stepwise RF. Habitats legend is given in the class hierarchy (lower right panel) alongside the location of eleven local classifiers (e.g., C.1). The lower left panel shows the variable importance values in each local classifier and in the flat model. The variables are divided to environmental variables (blue), variables derived from the remote sensing images (black) and the raw reflectance values of the two remotely-sensed images (red).

Figure 5: Summary of models' performance.

Performance of the FRF (Flat), HRF with multiplicative majority rule (H.Mult) and HRF with stepwise majority rule (H.Step) in each classification scheme, for the training and validation sets. Box plots represent the 25, 50 and 75% percentiles of the 15 runs, while the whiskers give the 1.5 IQR. Outliers are given as points while the mean is given as a triangle.

738 Figure 1:



739

740

741

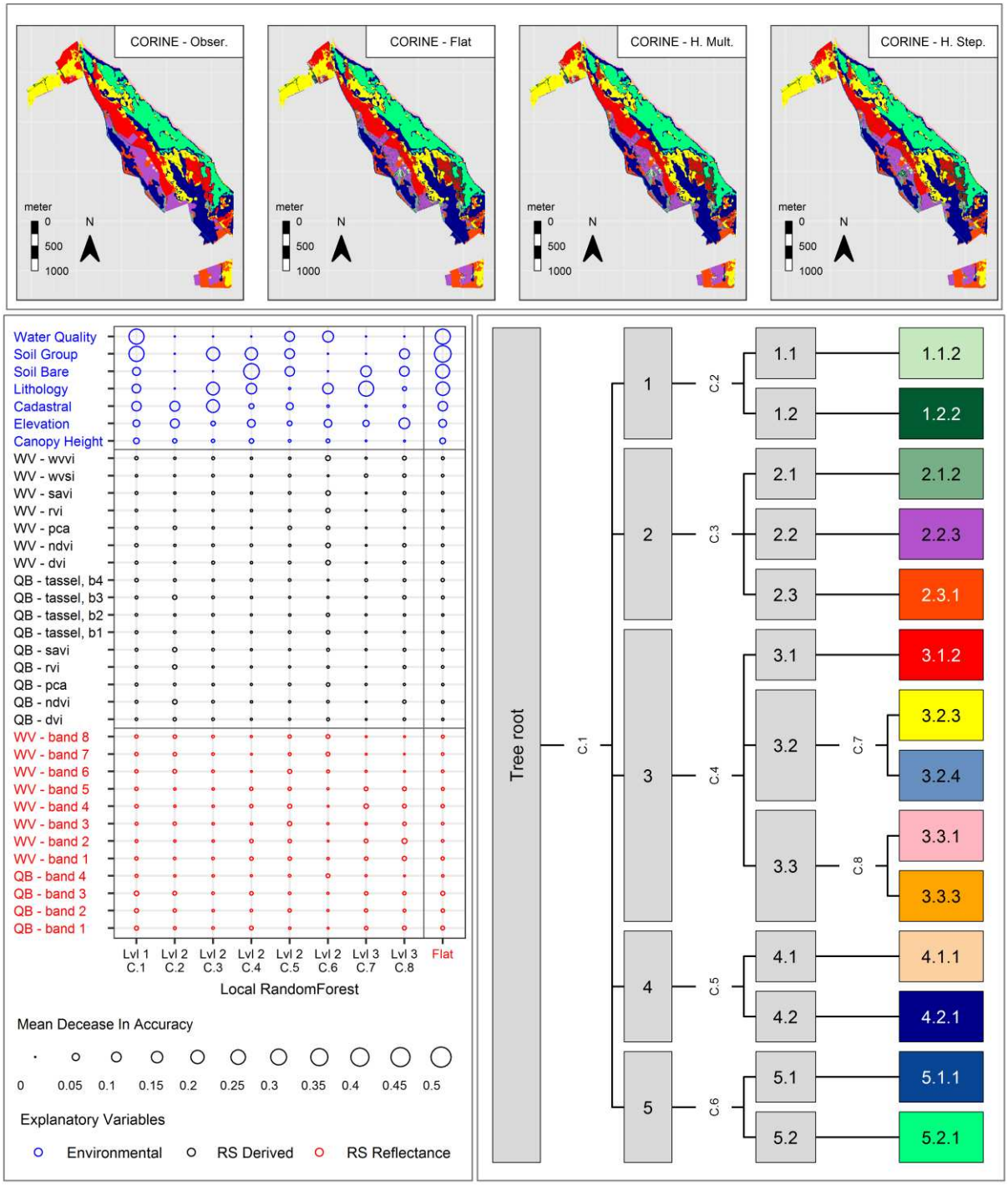
742

743

744

745

746 Figure 2:



747

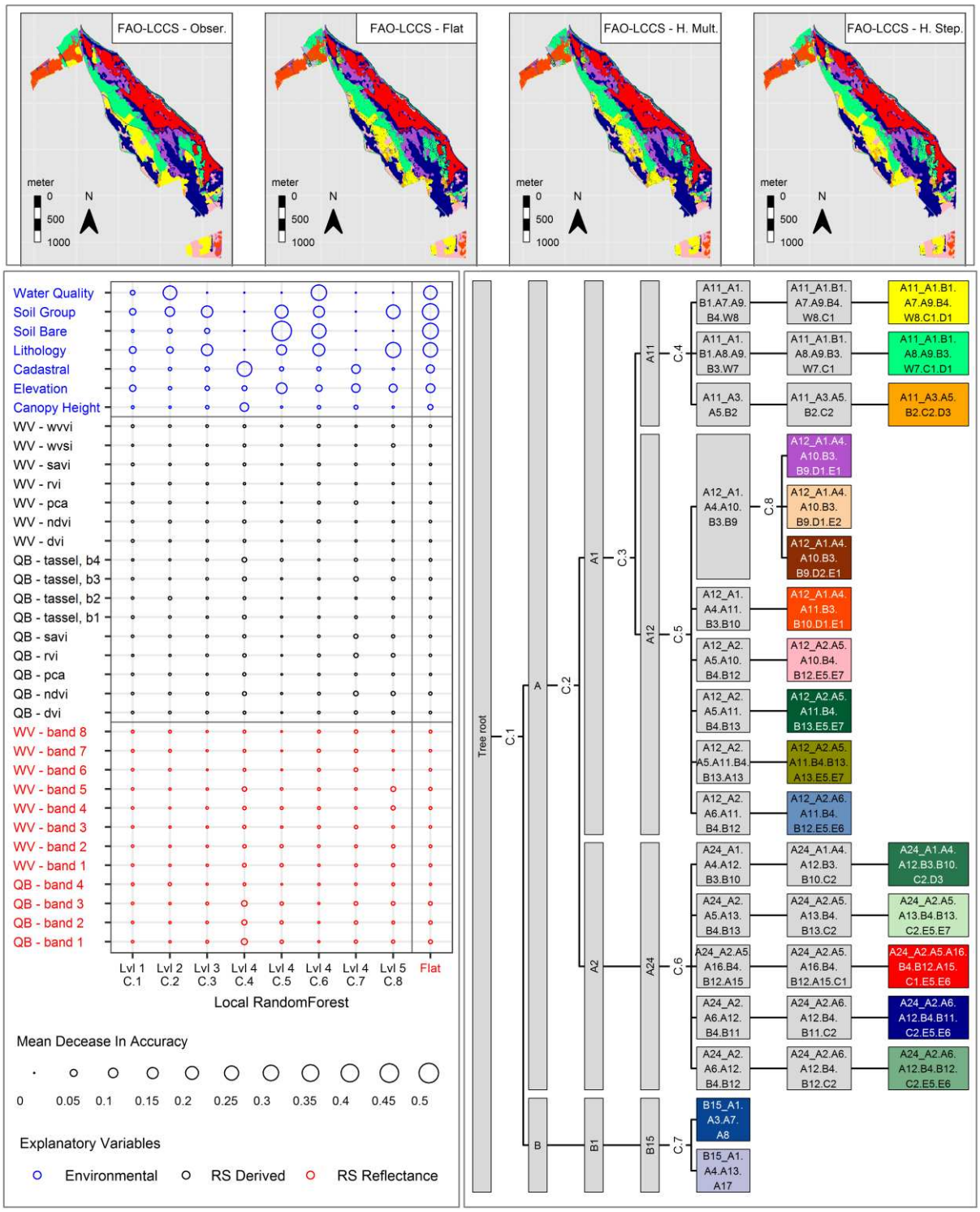
748

749

750

751

752 Figure 3:



753

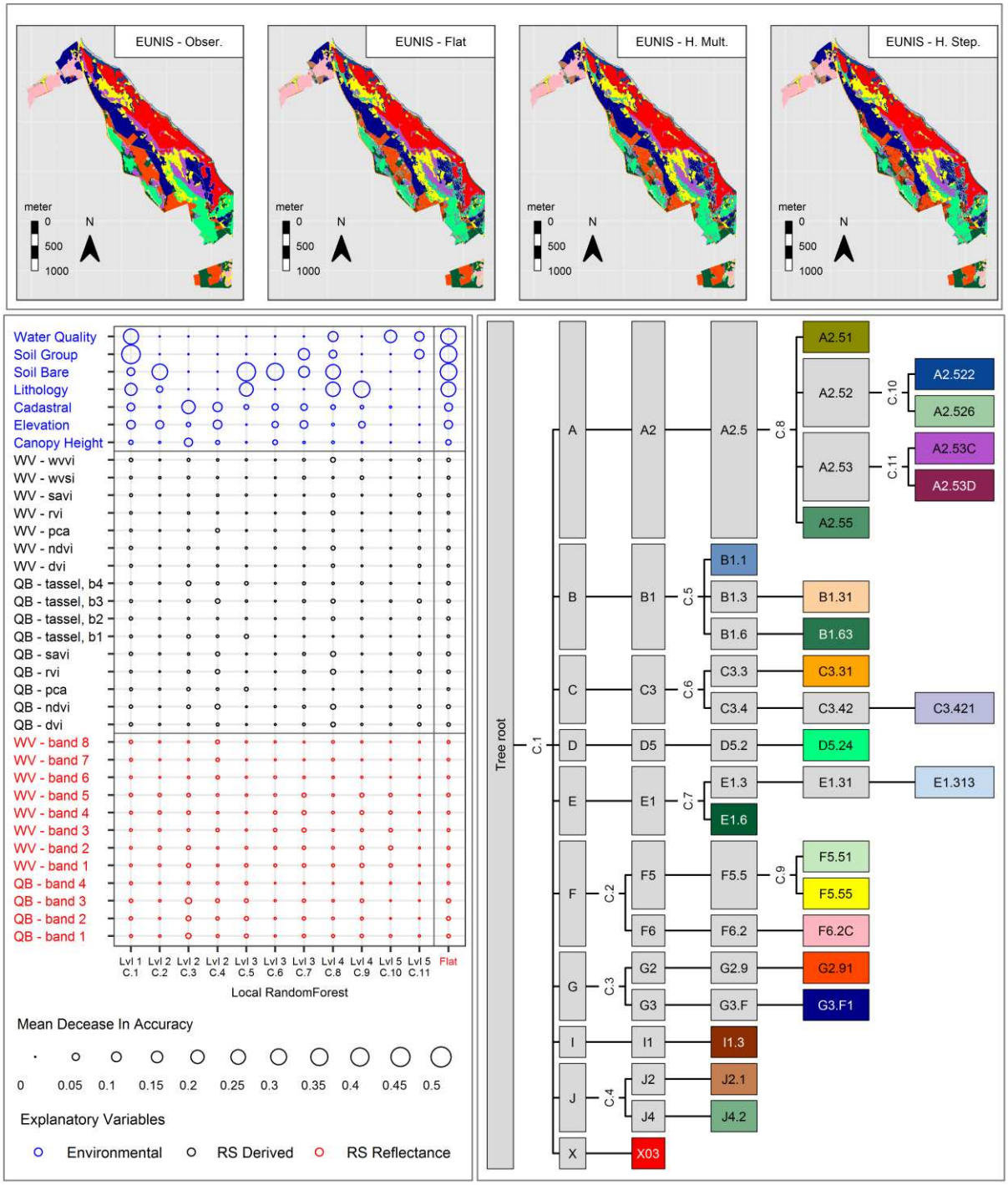
754

755

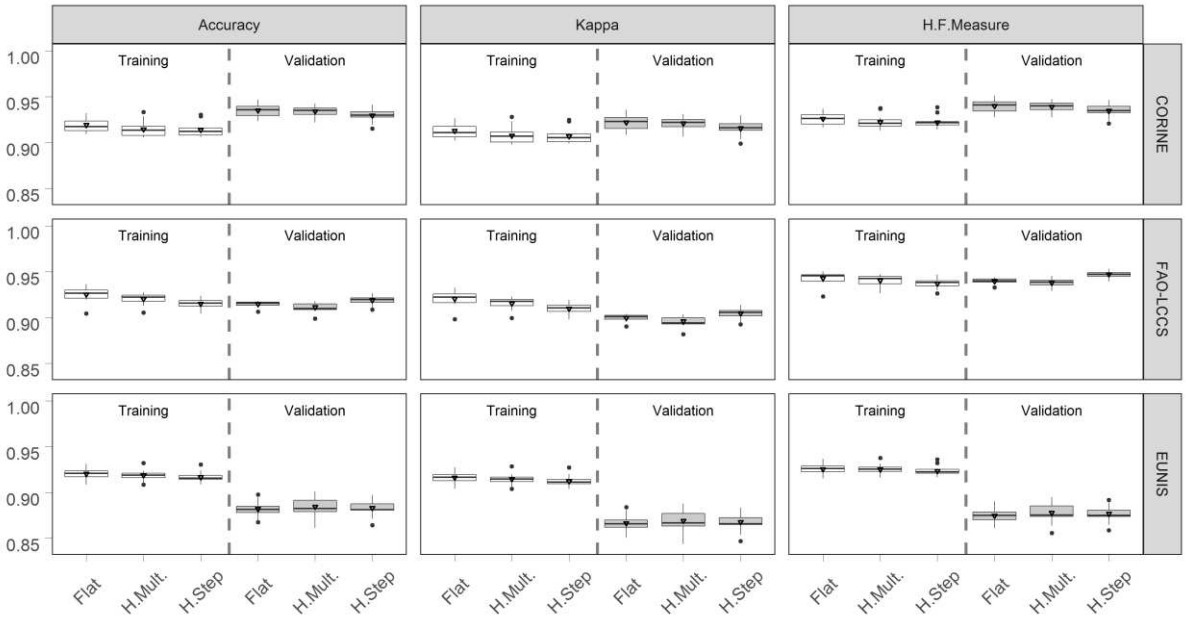
756

757

Figure 4:


EUNIS - Obser.
EUNIS - Flat
EUNIS - H. Mult.
EUNIS - H. Step.
meter
0
500
1000
N
Water Quality
Soil Group
Soil Bare
Lithology
Cadastral
Elevation
Canopy Height
WV - wvvi
WV - wvsi
WV - savi
WV - rvi
WV - pca
WV - ndvi
WV - dvi
QB - tassel, b4
QB - tassel, b3
QB - tassel, b2
QB - tassel, b1
QB - savi
QB - rvi
QB - pca
QB - ndvi
QB - dvi
WV - band 8
WV - band 7
WV - band 6
WV - band 5
WV - band 4
WV - band 3
WV - band 2
WV - band 1
QB - band 4
QB - band 3
QB - band 2
QB - band 1
Lvl 1 C.1
Lvl 2 C.2
Lvl 2 C.3
Lvl 2 C.4
Lvl 3 C.5
Lvl 3 C.6
Lvl 3 C.7
Lvl 4 C.8
Lvl 4 C.9
Lvl 5 C.10
Lvl 5 C.11
Flat
Local RandomForest
Mean Decease In Accuracy
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Explanatory Variables
Environmental
RS Derived
RS Reflectance
Tree root
C.1
A
A2
A2.5
C.8
A2.51
A2.52
C.10
A2.522
A2.526
A2.53
C.11
A2.53C
A2.53D
A2.55
B
B1
C.5
B1.1
B1.3
B1.31
B1.6
B1.63
C
C3
C.6
C3.3
C3.31
C3.4
C3.42
C3.421
D
D5
D5.2
D5.24
E
E1
C.7
E1.3
E1.31
E1.313
E1.6
F
C.2
F5
F5.5
C.9
F5.51
F5.55
F6
F6.2
F6.2C
G
C.3
G2
G2.9
G2.91
G3
G3.F
G3.F1
I
I1
I1.3
J
C.4
J2
J2.1
J4
J4.2
X
X03

764 Figure 5:



765

766

767

768

769

770

771

772

773

774

775

776

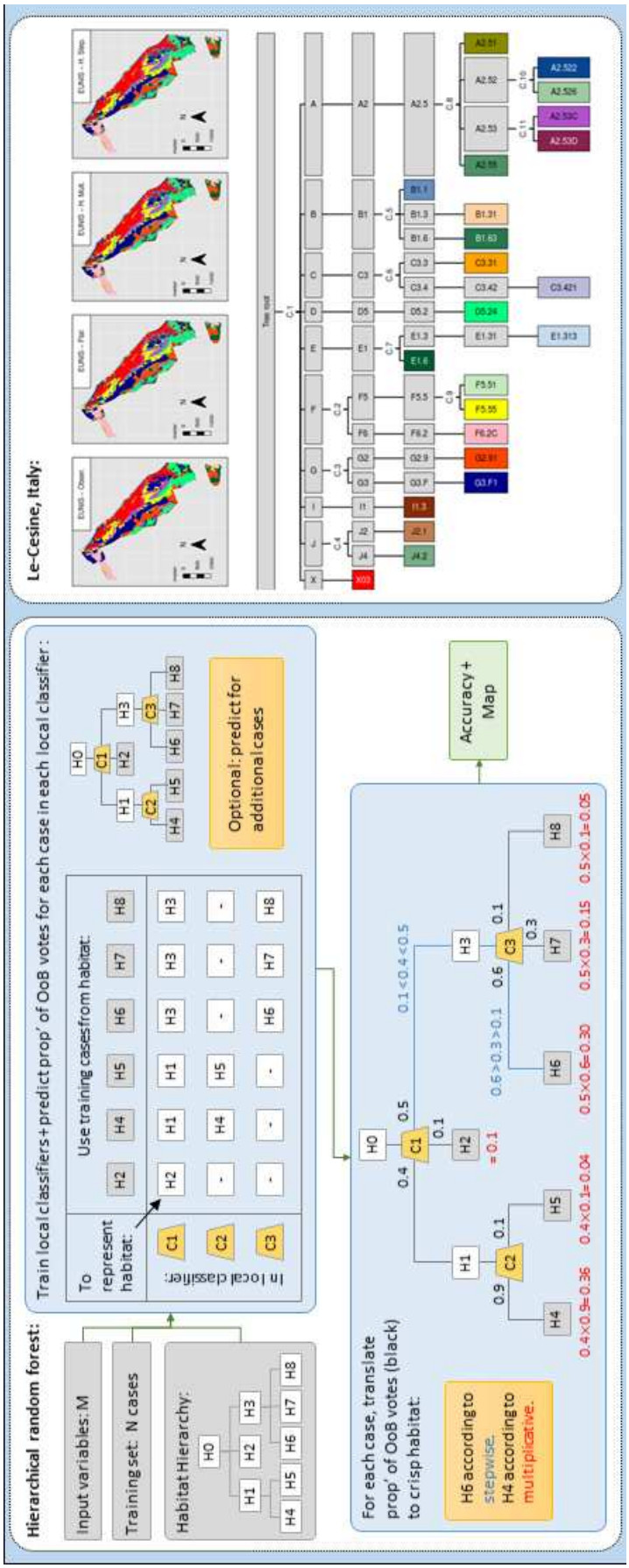
777

778

779

780

781 Graphical abstract:



782