WILL IT BLEND? VISUALIZATION AND ACCURACY EVALUATION OF HIGH-RESOLUTION FUZZY VEGETATION MAPS

A. Zlinszky [a]*, A.Kania [b],

[a] Balaton Limnological Institute, Centre for Ecological Research, Hungarian Academy of Sciences, Tihany, Hungary, zlinszky.andras@okologia.mta.hu
[b] Dept. of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria

Commission II, WG II/4

KEY WORDS: Fuzzy mapping, accuracy assessment, random forests, accuracy visualization, colour blending, vegetation mapping, quality control

ABSTRACT:

Instead of assigning every map pixel to a single class, fuzzy classification includes information on the class assigned to each pixel but also the certainty of this class and the alternative possible classes based on fuzzy set theory. The advantages of fuzzy classification for vegetation mapping are well recognized, but the accuracy and uncertainty of fuzzy maps cannot be directly quantified with indices developed for hard-boundary categorizations. The rich information in such a map is impossible to convey with a single map product or accuracy figure. Here we introduce a suite of evaluation indices and visualization products for fuzzy maps generated with ensemble classifiers. We also propose a way of evaluating classwise prediction certainty with "dominance profiles" visualizing the number of pixels in bins according to the probability of the dominant class, also showing the probability of all the other classes. Together, these data products allow a quantitative understanding of the rich information in a fuzzy raster map both for individual classes and in terms of variability in space, and also establish the connection between spatially explicit class certainty and traditional accuracy metrics. These map products are directly comparable to widely used hard boundary evaluation procedures, support active learning-based iterative classification and can be applied for operational use.

1. INTRODUCTION

1.1 The problem of fuzzy classification

The prevailing approach in remote sensing is that each output map pixel belongs to one and only one class, but the shortcomings of such classification have been identified early on (Foody, 1992; Mairota et al., 2015). Crisp classification includes a strong reduction of the information in the sensor data by binarizing gradients and omitting information on alternatives to the selected class. This often compromises the applicability of remote sensing derived vegetation maps (Townsend, 2000), but alternative approaches remain rare. Fuzzy mapping (also known as soft classification) assigns a probability of membership for each class to each pixel. It includes information on the sub-dominant classes, can handle smooth transitions and uncertain identification, and is therefore particularly well suited for vegetation mapping. Ensemble classifiers such as random forests or neural networks are becoming increasingly popular, but although these inherently output fuzzy data, they are mostly still used for creating hard-boundary maps.

Some reasons for this may be that many users ask for clear and unambiguous results even if this is not justified by the objects they are mapping. Quantitatively conveying the information in a fuzzy map is considered difficult towards non-specialists. Also, compatibility with standard data formats and especially vectorization remains problematic since several alternative approaches exist. But most important of all, fuzzy maps are regularly criticized because their accuracy is not straightforward to quantify. Many possible metrics of fuzzy classification accuracy exist, but most are difficult to compare with crisp classification maps and no standards have been accepted. Therefore, most fuzzy maps are converted to hard maps for the purpose of accuracy evaluation, but this approach loses the patterns contained in the class membership probabilities.

1.2 State of the art

The confusion matrix (Congalton, 1991) is the most widespread method for accuracy evaluation of hard classifications, since it gives a clear overview of the classwise accuracies and the individual frequencies of misinterpretations between classes. Modifications of the confusion matrix for compatibility with fuzzy classification have been proposed (Binaghi et al., 1999) but studies of fuzzy classification still continue to be published with non-fuzzy evalution (Du et al., 2012). Various indices for representing classification certainty from the class membership vector in each pixel have been proposed (Maselli et al., 1994; Prasad and Arora, 2014), which now allow spatially explicit representation of certainty or graphic visualization independent from position in space. However, these indices are not straightforward to interpret or link to the confusion matirx, and may require more detailed ground truthing (Townsend, 2000).

Hard boundary maps are usually visualized by assigning a colour to each class and rendering each pixel to the appropriate colour. Accuracy or certainty of classification is evaluated based on similarity to independent validation ground truths, and the assumption is made that the accuracy figures calculated from these samples are a representation of the overall accuracy of the classification for the whole map. Hard boundary classifiers do not allow direct calculation of classification quality outside areas where ground truths are present. Meanwhile, the need for spatially explicit evaluation across the whole surface of a map and not only within samples of ground truth has been identified

* Corresponding author

(Foody, 2002). Various alternatives for visualization of fuzzy vegetation maps have been used, the most popular being the calculation of class membership rasters for each category separately. These maps are informative if the user is interested in a single class, but are difficult to use if several classes are to be interpreted simultaneously, as this requires creating and visualizing a multi-band raster with a layer for each class. Blending the colours of the individual classes according to their probability has been suggested (Foody, 1996) but not applied in practice.

1.3 Objectives

Our objective was to propose an integrated solution for simple and straightforward accuracy evaluation metrics and visualization methods that unfold the information content of the fuzzy image. We also aimed to develop alternatives for visualization of fuzzy classification maps together with their accuracy that can be understood without high-level knowledge of image processing. Backwards-compatibility with indices and visualization used for hard-boundary maps and applicability in common GIS environments were also required.

2. METHODS

2.1 Test site and example data

The methodology was developed on airborne LIDAR-based classification of protected grassland areas in the framework of a vegetation monitoring project. The example we present here is a classification of various grassland categories from a study site in Püspökladány, Hungary. Details about the sensor and field data collection, categories and processing are published in Zlinszky et al., (2015a). The example dataset we show here contains 6 classes and covers 88 088 000 pixels of 0.5×0.5 m resolution. Classes were selected to include all major land cover types together with the three most important grassland habitats that are the focus of the classification: alkali short grasslands, alkali open swards and tall grass alkali meadows (Deák et al., 2014).

2.2 Random forest classification of multi-band images

In this example, random forest machine learning was used for classification. Random Forest (Breiman, 2001) is a tree-based ensemble classifier: individual bootstrap subsamples of the training data are taken, and a number of decision trees are trained on these (100 in our case). All trees make independent predictions for the class membership of each pixel. Therefore, for each pixel of the predicted map, we obtain a vector with the respective predictions from each tree. The proportion of trees within the ensemble predicting a certain class for the pixel is interpreted as the probability of the pixel belonging to that class. The final class assigned to the pixel is decided by majority voting of the individual trees (ordering the vector), i.e. the class with the highest value in the membership probability vector. The information provided by the random forest procedure is therefore immediately suited for further processing in a fuzzy sense, but is also compatible through majority voting with classical crisp vegetation mapping where each pixel belongs to only one class.

The metrics we developed can be applied to any ensemble classifier as long as a number of classifiers is created using the same training data and the individual models (also known as base or weak classifiers) are diverse enough to learn patterns in data that might have been missed by some other base classifier in the ensemble. If the individual classifiers' errors are sufficiently uncorrelated, they compensate their individual errors, thus improving predictive power of the whole ensemble. Even if the individual classifiers output only a hard prediction, when the prediction from many such classifiers is merged, it can be evaluated as fuzzy in nature.

2.3 Visualization of fuzzy vegetation maps

Instead of a single product, we created a set of output tables, maps, and graphs that can be evaluated individually or together (Tab. 1, Figs 1, 2, and 3). Colouring the raster was extended to the fuzzy case by rendering each pixel mixing the respective colours of the classes that locally had non-zero probabilities, weighing the colours according to the probability of the corresponding class. The most simple colour blending method is based on the Red, Green and Blue (RGB) values of the image. Alternatively, we used hue-preserving rendering (Chuang et al., 2009) which avoids introducing new, synthetic hues, not existing in the original color scheme. Instead, when interpolating between distinct hues, the saturation of the first color is continually minimized until the color reaches gray tones, then the transition progresses towards the new hue of the other color of the mixture, increasing saturation until the destination color is reached. However, this blending mode only supports interpolation between two colors.

For three selected classes, their respective probability was also assigned directly to the R, G and B channels of an image. In this case, the map was dark wherever neither of the three classes had a high probability, and showed blended colours wherever two classes had similar probability. As such, this map already provides a spatially explicit representation of classification certainty. However, an additional map product was also created from the probability vector of each pixel, by defining the absolute difference between the number of ensemble votes received by the dominant class and the votes of the second most probable class as a metric of local classification certainty. This indicator we named "probability surplus" directly represents the certainty of categorization, and is calculated from the ensemble independently for each pixel. As the ensemble classifier produces a vector of probabilities for each pixel of the output, the expected reliability of the classification was visualized based on the "probability surplus". This output map allows spatially explicit interpretation of an accuracy indicator, including beyond validation samples.

2.4 Accuracy assessment of fuzzy vegetation maps

In order to evaluate fuzzy classification accuracy but preserve compatibility with the confusion matrix, the approach proposed by Lewis and Brown, (2001) was slightly reformulated. Lewis and Brown use sub-pixel area of each class in each pixel while in our case we use the ratio of classifiers in the ensemble to weigh the figure in each cell of the confusion matrix by its probability. In our case, this was achieved using the ensemble classifier, creating separate confusion matrices from the validation dataset for each base classifier. These individual confusion matrices were then added, and the resulting cell values normalized by the number of base classifiers, creating an "ensemble confusion matrix". The output includes fractional values of pixels wherever a fraction of the trees in the ensemble made different predictions, but the rows and columns still add up to real pixel counts and all confusion matrix based indices (producer's and user's accuracy, overall accuracy, Cohen's Kappa (Congalton, 1991), quantity and allocation disagreement (Pontius and Millones, 2011)) can be calculated and are meaningful.

However, summing the classwise accuracy into numeric figures is a strong over-simplification even where ensemble voting is taken into account. Adding further detail to the evaluation by exploring the probability relation between the dominant and the sub-dominant classes can help refine the classification process and better understand the classes in their context. We propose "dominance profiles", a quantitative graphical representation for the probability of the dominant class with respect to all other classes (Fig. 3). In the first step, all pixels are queried within the validation samples identified in the field as belonging to a class. On the X axis, the pixels are ranked according to the probability of this selected class, and grouped into bins that represent an equal number of pixels. On the Y axis, the probability of each class is plotted in stacked bars representing each bin. The order of classes in the stacked columns reflects their overall frequency in the matrix of class probabilities: the dominant class for which the graph was created is at the bottom (allowing class probabilities to be directly read from the Y-axis for this class), with the class occupying the second most area in the full graph above and so on until the least represented class. The bars representing the dominant class are ordered in a decreasing curve, starting from the pixels where the class in focus is predicted with the highest probability. The respective area of each class in this graph directly corresponds to their total number of base classifier votes within the validation sample, which is also the producer's accuracy figure in the ensemble confusion matrix. The first bin where a pixel occurs that is not dominated by the class in focus is marked by a line on the X-axis, the "dominance limit", corresponding to the hard producer's accuracy.

Alternatively, dominance profiles were calculated not only for the pixels within each class of the validation dataset, but also for the full study area. In this case, the query is made for all pixels of the study area dominated by the class in focus, therefore the "dominance limit" is the edge of the graph.

For well-defined categories where the dominant class has a high margin of probability over the rest of the classes, the graph will have a large area occupied by the dominant class, with only the top right corner representing the rare cases where other classes also received some probability. For less certain classes, even the pixels where the probability was the highest would include considerable probability of other classes, which increases along the X axis to the point where the domination becomes marginal. In some cases, the area of the graph outside the dominant class is evenly distributed between several classes, in other cases, there is clearly a single sub-dominant class.

The level of similarity between the dominance profiles of each class within the validation data and the profile for the same class over the whole map allow estimating the representativeness of the validation samples. If the validation samples would be ideally distributed, the dominance profiles within the dominance limit would exactly match the profiles for the whole dataset.

2.5 Implementation

The backbone of the implementation is the Scikit-learn python library (Pedregosa et al., 2011) that implements a number of machine learning algorithms, including ensemble classifiers. GDAL was used to provide interoperability with the GIS raster and vector data formats used for mapping, and Scikit-image library for image processing tasks (Van Der Walt et al., 2014). Visualization rendering, colour blending and confusion matrices were computed directly in Python code, and dominance profiles were plotted with help of MatPlotLib library. These modules are part of a full data processing chain using Python as a glue language. The full software solution (under the working name "Vegetation Classification Studio") has been used for classification tasks in various habitats from airborne LIDAR data (Zlinszky et al., 2015a, 2014).

3. RESULTS

3.1 Alternatives for visualization

In case of the example dataset with 6 classes, colours were assigned to intuitively reflect the type of vegetation they correspond to. For a limited number of classes with carefully selected colours, the classical RGB colour mixing model produced results that are easily interpreted by the human operator. Results show that hue-preserving colour blending avoids generating new colours that are not among the pre-defined classes, and that the saturation of the individual colours can be used for inferring the level of classification certainty.

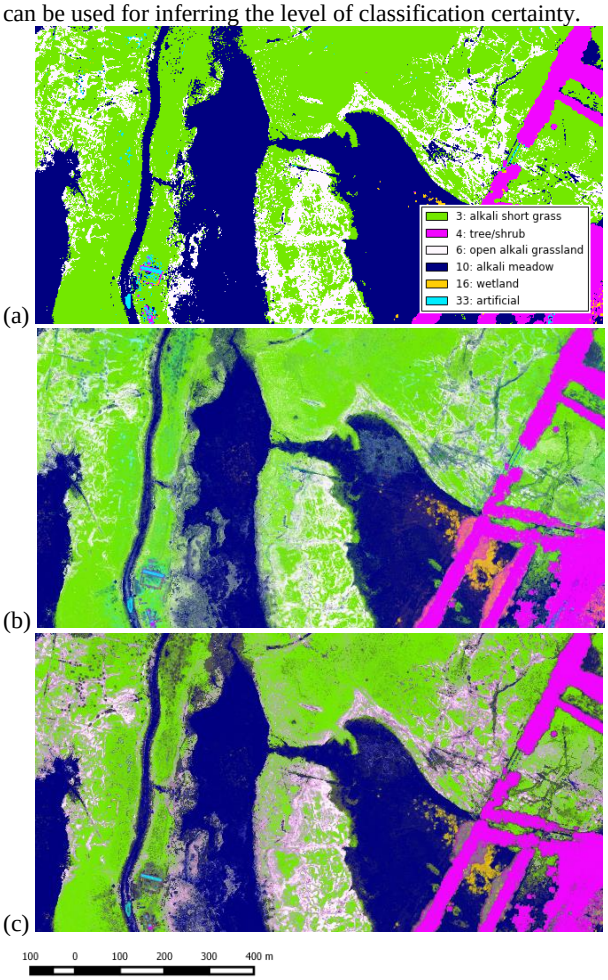

Figure 1. Classified vegetation map: hard-boundary (a), fuzzy blended in RGB (b), fuzzy blended using hue-preserving algorithm (c). The same colour scheme is used throughout the paper if not otherwise noted

Hard boundary classifications were created from the classification output, with different methods. The most simple method is majority voting but a hard-boundary map can also be generated by assigning all pixels to a certain class if the probability is above a given threshold regardless of whether it is dominant. This is especially relevant for mapping invasive species and other hazards or if focusing on a very rare class

where even predictions with low levels of certainty are of interest.

RGB renderings proved to be especially helpful if a subset of classes was of special interest, and if these classes were prone to create mixtures. If this was the case, various combinations of the classes could be recognized from this visualization.

For spatially explicit accuracy evaluation, the probability surplus map was used as this shows at the level of individual pixels the certainty of the class assigned to that pixel. Such a map allows identification of regions where the classification is less certain than other places. Reasons for this may include presence of a land cover type that was not included in the classification scheme, presence of some sort of noise in the sensor data, or local conditions producing an exceptions to the general rules used for classification.

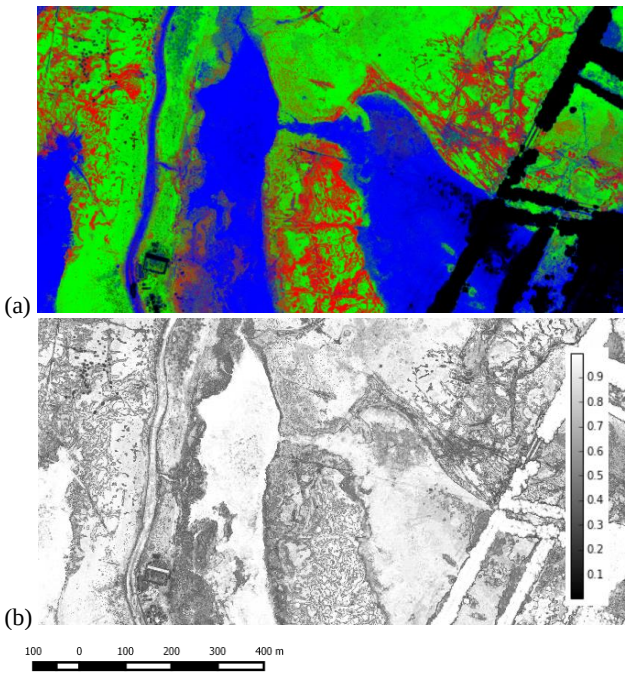

Figure 2. RGB three-class rendering (a) with 3 classes mapped to individual color channels, Red: class 6 (open alkali grassland), Green: class 3 (alkali short grass), and Blue: class 10 (alkali meadow); and probability surplus map (b) of the same area

3.2 Accuracy indices: the ensemble confusion matrix and the dominance profile graph

The "ensemble confusion matrix" we propose is sensitive enough to indicate the effect of even one decision tree that predicted differently than remaining hundred, but can still summarize the information from the whole ensemble. It will also show which classes are most frequently mistaken for each other even if these have only a low probability compared to the dominant class. The ensemble-based confusion matrix is directly compatible with the hard confusion matrix due to the normalization with the number of classifiers in the ensemble. In the theoretical case that all pixels would be classified with all classifiers in the ensemble predicting the same result (so total certainty and 100% probability surplus), the ensemble confusion matrix would be exactly the same as the hard confusion matrix. Where this is not the case, the numbers of pixels in each cell are modified according to the proportion of classifiers predicting that result. The row and column totals add up to the true number of pixels, but the accuracy figures are always smaller, the numbers in the main diagonal cells are always lower than for the hard-boundary confusion matrix.

The resulting differences in producer's and user's accuracy compared to the hard boundary matrix represent the level of certainty of the individual class assignments. The hard-boundary confusion matrix overestimates quality since it assumes every pixel has full certainty, the ensemble-based confusion matrix takes both correctness and certainty into account.

Confusion matrix

	(3)	(4)	(6)	(10)	(16)	(33)	Total	Prod.Acc.
alkali short grass (3)	12 981	6	2 788	2 362	0	20	18 157	71.5%
tree/shrub (4)	232	9 965	0	513	88	15	10 813	92.2%
open alkali grassland (6)	2 135	0	5 926	179	0	4	8 244	71.9%
alkali meadow (10)	376	8	241	38 938	158	0	39 721	98.0%
wetland (16)	775	2 274	1 006	2 457	6 178	0	12 690	48.7%
artificial (33)	223	1 378	12	668	18	18 655	20 954	89.0%
Total	16 722	13 631	9 973	45 117	6 442	18 694		
User Accuracy	77.6%	73.1%	59.4%	86.3%	95.9%	99.8%		

Cohen's Kappa: 0.79

Ensemble confusion matrix

	(3)	(4)	(6)	(10)	(16)	(33)	Total	Prod.Acc.
alkali short grass (3)	11 735.86	44.58	3 478.93	2 341.27	13.17	536.67	18 150.48	64.7%
tree/shrub (4)	203.82	9 346.55	24.08	642.80	286.62	307.40	10 811.27	86.5%
open alkali grassland (6)	2 574.08	0.83	5 024.31	474.59	0.27	164.52	8 238.60	61.0%
alkali meadow (10)	813.33	375.75	548.87	36 115.20	918.00	947.28	39 718.43	90.9%
wetland (16)	993.60	2 134.42	1 165.74	2 138.54	6088.92	161.09	12 682.31	48.0%
artificial (33)	1 068.76	1 799.40	313.83	580.02	99.85	17 089.82	20 951.68	81.6%
Total	17 389.45	13 701.53	10 555.76	42 292.42	7 406.83	19 206.78		
User Accuracy	67.5%	68.2%	47.6%	85.4%	82.2%	89.0%		

Cohen's Kappa: 0.707

Table 1. Confusion matrix and ensemble confusion matrix representing the hard and fuzzy accuracy figures of the same classifier

The dominance profile graphs we created allow exploring the prediction certainty reducing the information to individual classes, but dropping the spatially explicit dimension for better understanding. The dominance profile might be near horizontal (indicating that most pixels have the same distribution of probabilities), linear in shape (indicating that the various probability levels of the dominant class are evenly distributed) but always represents a monotone decreasing curve since the X axis is ordered by probability of the dominant class. The dominance profiles calculated within the validation pixels corresponding to each class are directly related to the figures of the ensemble confusion matrix: both the areas of the respective classes in the graph and the numbers in the corresponding row of the ensemble matrix represent the number of votes received by each class. The dominance profiles deliver more information than simply the count: the distribution of votes according to the level of class dominance can be read from the graph. This includes both the pixels in the validation sample that were assigned to the correct class but had a certain number of incorrect votes (and thus a level of certainty below 1) and the pixels where the correct class was sub-dominant. For the profiles representing the full dataset, the query was made according to the dominant class, therefore the dominance limit is also the limit of the graph.

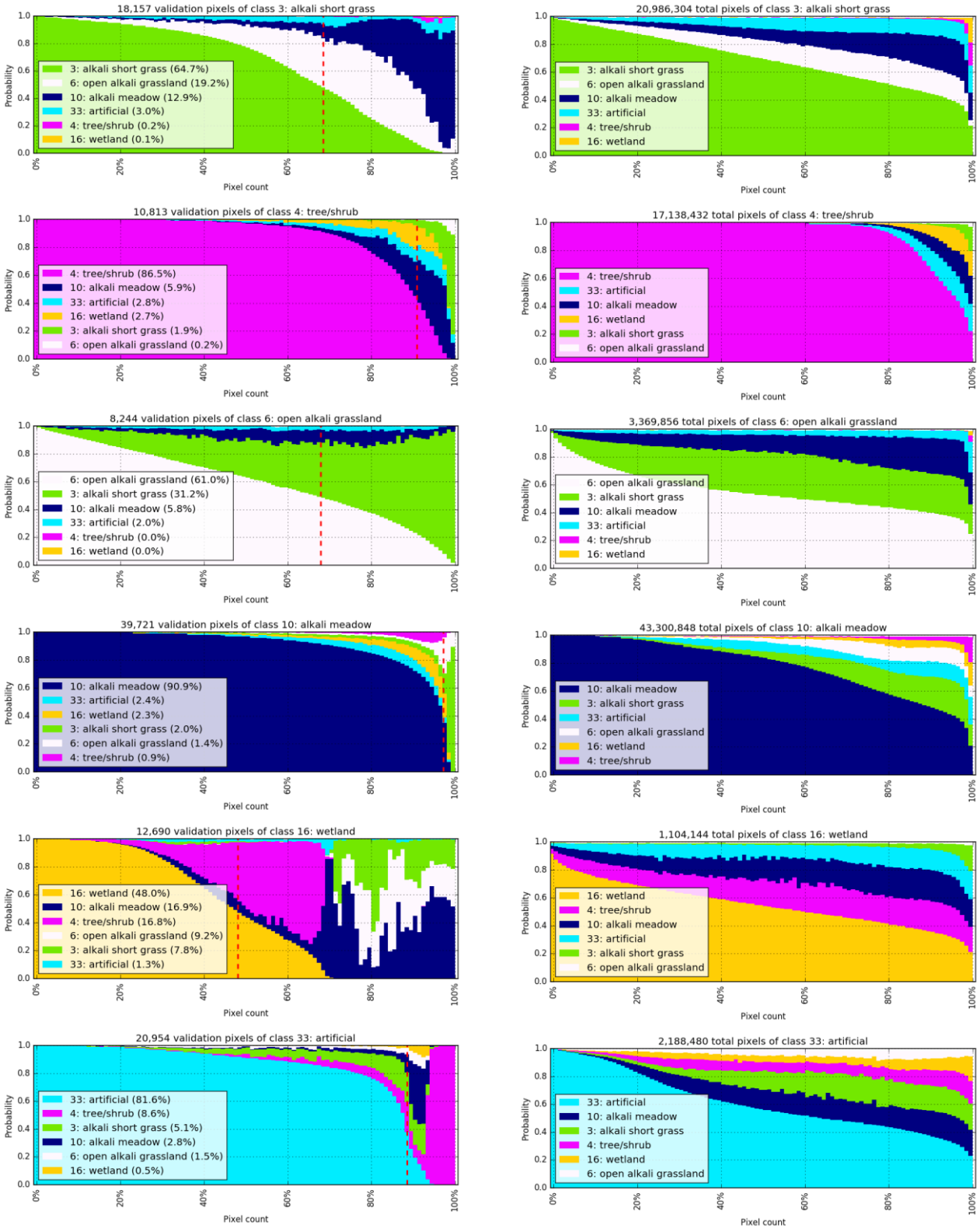

Figure 3. Dominance profile graphs for the validation pixels (left column)
and for the whole study area (right column) for respective classes. The “dominance limit” is drawn as the red dashed vertical line in left-column graphs.

For the profiles from the full dataset, the start and end points of the dominance profile represent the highest certainty with which the class was detected and the lowest probability that was still higher than any of the alternative classes.

For the most certain and also most accurate class, "trees/shrubs (4)", the validation pixel profile shows that 86.5% of the pixels were correctly classified. The graph of the total pixels suggests that a remarkable 60% of the class assignment is completely certain, and for the remaining pixels, all other categories had equal (but still rather low) certainty. The profile of all pixels dominated by this class is closely resembles the dominance profile of validation pixels, cut off at the dominance limit. This is the example of a strong and well-defined class, with the validation data providing a representative sample of the total. On the contrary, for the class "open alkali grassland (6)", only 72% of the validation pixels were correctly classified, and nearly all the rest were dominated by "alkali short grass (3)". The dominance profile for all pixels in this class shows a similar pattern: nearly 80% of the pixels had probabilities lower than 0.6. The second most probable class, "alkali short grass (3)" has probabilities around 0.2 in most pixels, much higher than any of the other sub-dominant classes, which suggests that the separation of these two classes is problematic not only within the validation sample but also over the whole area. Thus, the class has weaker accuracies, but the validation data seem representative. For the dominance profile for "alkali short grass (3)" most probable sub-dominant class is as expected, "open alkali grassland (6)", but is much higher and more linear in shape, indicating an even distribution of probabilities. However, while the validation pixels suggest that the class "artificial (33)" is hardly ever mixed with class (3), in the total pixels this sub-dominant class has nearly the same level of probability as the other two. The interpretation is that the validation data is somewhat underrepresenting the possibilities encountered in the full dataset. The full-area dominance profile of class "artificial (33)" is clearly bimodal, with varying steepness. This might indicate that a type of frequent artificial objects is recognized with high certainty, while another type is less well recognized. The total-pixel profile suggests that no sub-dominant class prevails and this is mirrored in the part of the validation dominance profile inside the dominance limit; however it is also shown that most of the misclassifications belonged to one class (4) with total certainty. In the confusion matrix, the class "wetland (16)" has the lowest producer's accuracies. This is further explained by the validation dominance profile: the discrimination from (4) tree/shrub is often uncertain, with a smooth transition between the dominance levels of these two classes inside the validation data. Additionally, for 30% of the validation pixels, this class had no probability at all, and overestimation of three other classes shared base classifier votes for these pixels. This is matched by full-area dominance profile of this class, which does not start at 1.0. The high certainty pixels in the validation sample apparently represent very rare cases. The bins of the total profile represent 1% each of the data, and even the most certain among these had probabilities only around 0.9. Based on the spatially explicit map of probability surplus, sub-samples within the image may be created where class dominance profiles can be investigated in order to better understand the reason for weak class prediction, or the dominance profile of the whole image irrespective of class can be plotted as an indicator of the overall probability surplus distribution and the certainty of the classification.

4. DISCUSSION

4.1 Fuzzy Visualization schemes

The multi-class fuzzy colour renderings show a wide range of patterns that are not visible or difficult to interpret in the hard-boundary maps. Especially in cases where the probability surplus of the dominant class is low and the second best-class has similar probability, fuzzy visualization allows recognition of the sub-dominant class. Especially the smooth transitions characteristic for grasslands were successfully visualized with this approach. Comparison of the map with field experience by expert ecologists has shown in many cases that the features represented by the sub-dominant classes resemble real patterns. Vegetation features that are not defined in a set of classes but have a characteristic shape, such has linear vehicle tracks can be instantly recognized even if they do not affect the dominant class. Instead of large homogeneous fields of colour that are typical for hard classification maps, patterns in vegetation can be identified even in areas that have the same dominant category. Finally, cases where large areas are occupied by a relatively even mixture of two classes can be identified, potentially leading to better definition of the class scheme. Since these maps were created using regular image formats used in GIS, their visualization or interoperability with other datasets was not problematic. All output maps are compatible with standard GIS image formats and involve either one greyscale or three RGB channels. Therefore even viewing in an office software environment for non-specialists is supported.

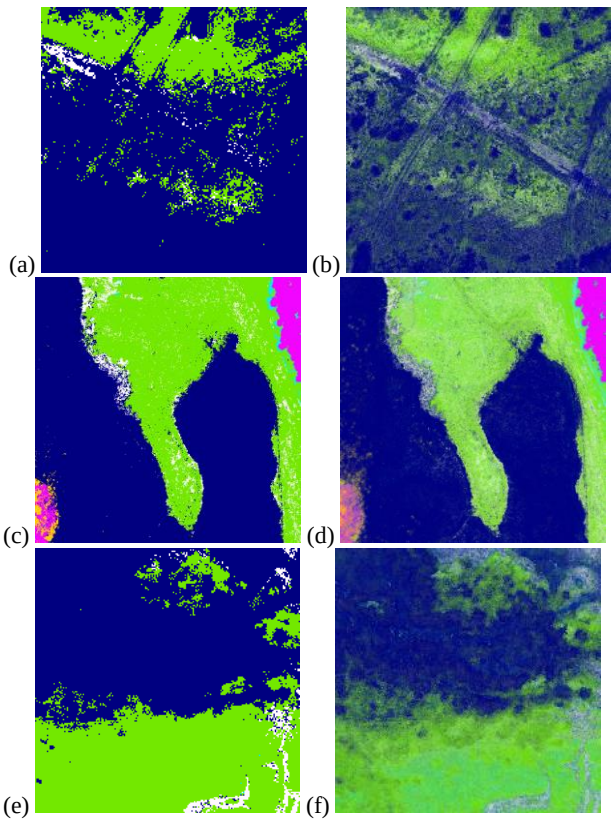

Figure 4: Hard boundary (left column) and fuzzy (right column) visualizations of the same area, showing clear linear features in (a, b);a large area covered by equal mix of two classes, suggesting the need to introduce a new class (c, d); and gradual transitions between several classes (e, f), all hardly detectable in the corresponding hard boundary map

However, simple colour blending visualization becomes less and less clear to interpret as the number of categories increases. The colour scheme must therefore be appropriately selected. Hue-preserving rendering is especially useful if many categories are present and their respective colours are not so distinctive. In this colour blending method, mixed pixels are always more grey than pure pixels, and no new colours are generated. If some classes are more interesting for the application of the map than others (such as the three main grassland classes in our case), the RGB rendering is useful for identifying their respective certainty and eventual mixing.

We have introduced probability surplus as a way of spatially explicit classification certainty mapping. This is a longtime demand towards remote sensing and has only been made possible using ensemble classifiers. Contrary to the confusion index proposed by (Burrough et al., 1997), this is not normalized by the maximum probability, and investigates only the first two dominant classes contrary to the probability entropy index (Maselli et al., 1994). We believe this allows more straightforward interpretation of the results, especially for non-specialists. Even then, the assumption that class certainty represents accuracy or quality has to be made. The validity of this assumption is confirmed by the link between confusion matrix-based producer's accuracies and the respective class dominance profiles within the validation samples.

Calculating categorization certainty for each pixel also allows processing classification tasks through "active learning" (Tuia et al., 2011). Compared to the mainstream workflow of collecting a static set of ground truths and applying it in a single step for training and evaluation, active learning is an iterative approach. After every round of classification, the input ground truth data is extended by adding reference data in the areas where the classification has the least certainty, and ideally for classes that are identified with the least accuracy. This allows an optimized use of the fieldwork effort, and can be iterated until the desired accuracy is reached. Of course, the criteria of evenly distributed reference data are still to be taken into account, and can be checked by comparing validation sample based and full-area classwise dominance profiles.

4.2 Accuracy evaluation of fuzzy maps

Various approaches to evaluating or visualizing fuzzy classification have been proposed in the literature. The novelty of our approach is that instead of a single indicator or visualization, we propose a set of map products and graphics. Each of these reduces the multi-dimensional information in the output matrix of the ensemble classifier in a different way, from 2+3 dimensions (RGB rendering, blending in colourspace) through 2+1 (probability surplus map) and two dimensions (dominance profile graphs, ensemble confusion matrix) to one dimension (ensemble-based overall accuracy or Cohen's Kappa). Each index grasps a different aspect of the fuzzy map and can be used to answer a different question. Together, these data products allow an in-depth understanding of any ensemble classification and a much more thorough use of the immense information contained in high-resolution airborne sensor datasets such as hyperspectral imaging or full-waveform LIDAR (Zlinszky et al., 2015b).

One important and hitherto unevaluated question is how much the accuracy of the classification in the sense of agreement with ground truths is closely related to the probability surplus. Comparing the hard boundary confusion matrix, the ensemble confusion matrix and the class dominance profiles calculated within the validation samples helps answer this question. The producer's accuracy figures in the ensemble confusion matrix directly represent the number of votes from base classifiers that each class received within its own validation pixels. The area occupied by each class in the dominance profile also equals the proportion of votes within the respective validation pixels. The user's accuracy is also represented in the graphs: the area occupied by a class in its own dominance profile (the number of "correct" votes) divided by the total area of the same class in the profiles of all classes (total number of votes) gives the ensemble-based user's accuracy. Therefore, classes where dominance is stronger also receive higher user's accuracies in the ensemble confusion matrix. Based on this, it can be assumed that the probability surplus map is a valid representation of classification correctness (and not "only" certainty), as it shows how strong the majority of the dominant class is in each pixel. The link between probability surplus and hard-boundary confusion indices is weakened by the binarization of class membership using majority voting: classical producer's and user's accuracies are therefore higher than indicated by the local probability surplus. Still, areas of the map that have high probability surplus figures are expected to be locally more accurately classified than areas with lower class dominance.

4.3 Discussion: how to use these indices

The proposed indices work with any ensemble-based classifier, such as random forests, neural networks, probabilistic decision trees (Du et al., 2012). In a production environment, it is possible to train an ensemble classifier and evaluate its accuracy using only the ground truth data as a subset of the original. This ensures high-speed processing compared to the regular practice of directly working with the whole dataset. For each model learning run, the regular (hard-boundary) confusion matrix can be generated as a first indicator of accuracy. Where the figures in the confusion matrix are favourable, the ensemble confusion matrix can also be generated for more detailed analysis of classification quality, and further changes to the classifier or the class definitions can be made if necessary. In the next step, a colour-blended fuzzy visualization is created to check an overall impression of the classification and how pure the individual pixels are. The probability surplus visualization can quantify the level of certainty for the whole study area and support recommendations of locations where additional field references should be collected (active learning). If certain classes are more important, an RGB rendering can inform on their occurrence even in the sub-dominant probability levels. The next step is to create dominance profile graphs for each class as these will allow a detailed understanding of the similarity between classes, and alternative sub-dominant categories. The level of probability typical for each class can be inferred, and if unsatisfactory, additional ground truths may again be added or classes merged. Finally, if all these indices are satisfactory, hard-boundary maps may be created based on majority vote or other output products (such as the probability of a critical class, eg Zlinszky et al. 2015b) may be delivered.

Other areas of application are for automatic optimization of machine learning algorithm settings, hyperparameter optimization, various settings related to pre-processing of remote sensing and reference data and in any processing regimes that rely on building hundreds or thousands of classification models searching for optimal settings. Such optimization approaches might use genetic algorithms, non-linear optimization algorithms or automatic algorithm configuration approaches to search over multi-dimensional space of possible parameters influencing the classification process, in an attempt to determine parameter values yielding classification models with the best accuracy and reliability. In

most such contexts classification is treated as a "black box" process, therefore an automatic optimization algorithm relies on some target function that is able to evaluate accuracy or "goodness" of a model. A number of optimization approaches also use gradients of target function to guide the algorithms towards search areas with more promising quality.

In this context, the crisp accuracy metrics show considerably worse "sensitivity" (or "resolution") of the accuracy, compared to the fuzzy ones. This effect is based on an inherent loss of information in traditional metrics that rely on reducing all the rich probability vector information generated for every pixel by an ensemble classifier to just one single class value.

So the fuzzy accuracy metrics – while being always lower that their hard-boundary counterparts – are much more sensitive to the quality of classification and are able to guide the optimization algorithm to obtain much faster convergence and find models with better properties.

5. CONCLUSIONS

Spatially explicit accuracy evaluation for classified maps was so far mostly done inside ground truth areas, and fuzzy classification was rarely used due to the perceived difficulty of accuracy assessment. Here we suggest a set of accuracy indicators that work on ensemble-based fuzzy maps. These data products allow understanding various aspects and levels of map quality, from single accuracy figures for the whole map on an ensemble basis, through various colour blending-based spatially explicit visualizations of classification and its accuracy, to classwise dominance profiles that are an intuitive but also quantitative way of evaluating the accuracy of class prediction. These visualization products can be linked in an efficient workflow for stepwise improvement of the classifier. We expect that the proposed techniques will facilitate a wider uptake of fuzzy classification in operational remote sensing.

REFERENCES

Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern Recognit. Lett. 20, 935–948.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/a:1010933404324

Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma, Fuzzy Sets in Soil Science 77, 115–135. doi:10.1016/S0016-7061(97)00018-9

Chuang, J., Weiskopf, D., Moller, T., 2009. Hue-preserving color blending. Vis. Comput. Graph. IEEE Trans. On 15, 1275–1282.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46.

Deák, B., Valkó, O., Alexander, C., Mücke, W., Kania, A., Tamás, J., Heilmeier, H., 2014. Fine-scale vertical position as an indicator of vegetation in alkali grasslands – Case study based on remotely sensed data. Flora - Morphol. Distrib. Funct. Ecol. Plants 209, 693–697. doi:10.1016/j.flora.2014.09.005

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., Liu, S., 2012. Multiple Classifier System for Remote Sensing Image Classification: A Review. Sensors 12, 4764–4792. doi:10.3390/s120404764

Foody, G., 1996. Fuzzy modelling of vegetation from remotely sensed imagery. Ecol. Model. 85, 3–12.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.

Foody, G.M., 1992. A fuzzy sets approach to the representation of vegetation continua from remotely sensed data : an example from lowland health. Photogramm. Eng. Remote Sens. 58, 221–225.

Lewis, H.G., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. Int. J. Remote Sens. 22, 3223–3235.

Mairota, P., Cafarelli, B., Didham, R.K., Lovergine, F.P., Lucas, R.M., Nagendra, H., Rocchini, D., Tarantino, C., 2015. Challenges and opportunities in harnessing satellite remote-sensing for biodiversity monitoring. Ecol. Inform. 30, 207–214.

Maselli, F., Conese, C., Petkov, L., 1994. Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. ISPRS J. Photogramm. Remote Sens. 49, 13–20. doi:10.1016/0924-2716(94)90062-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 32, 4407–4429. doi:dx.doi.org/10.1080/01431161.2011.552923

Prasad, G., Arora, M.K., 2014. Assessing uncertainty in fuzzy land cover classification by confusion index. Int. J. Geomat. Geosci. 5, 332–344.

Townsend, P.A., 2000. A quantitative fuzzy approach to assess mapped vegetation classifications for ecological applications. Remote Sens. Environ. 72, 253–267.

Tuia, D., Pasolli, E., Emery, W.J., 2011. Using active learning to adapt remote sensing image classifiers. Remote Sens. Environ. 115, 2232–2242. doi:10.1016/j.rse.2011.04.022

Van Der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in Python. PeerJ 2, e453.

Zlinszky, A., Deák, B., Kania, A., Schroiff, A., Pfeifer, N., 2015a. Mapping Natura 2000 Habitat Conservation Status in a Pannonic Salt Steppe with Airborne Laser Scanning. Remote Sens. 7, 2991–3019. doi:10.3390/rs70302991

Zlinszky, A., Heilmeier, H., Balzter, H., Czúcz, B., Pfeifer, N., 2015b. Remote sensing and GIS for habitat quality monitoring: New approaches and future research. Remote Sens. 7, 7987–7994. doi:doi:10.3390/rs70607987

Zlinszky, A., Schroiff, A., Kania, A., Deák, B., Mücke, W., Vári, Á., Székely, B., Pfeifer, N., 2014. Categorizing Grassland Vegetation with Full-Waveform Airborne Laser Scanning: A Feasibility Study for Detecting Natura 2000 Habitat Types. Remote Sens. 6, 8056–8087. doi:doi:10.3390/rs6098056

ACKNOWLEDGEMENTS

Data collection and software development for this study was supported by the Changehabitats2 EU FP7 Marie Curie Industry-Academia Partnership Project (Contract No. 251234) and by the OTKA PD 115833 grant.