Stoch Environ Res Risk Assess (2017) 31:2297–2312
DOI 10.1007/s00477-016-1310-y

ORIGINAL PAPER

An information-based criterion to measure pixel-level thematic uncertainty in land cover classifications

Patrick Bogaert[1] · François Waldner[1] · Pierre Defourny[1]

Published online: 31 August 2016

Abstract Traditional accuracy assessment of satellite-derived maps relies on a confusion matrix and its associated indices built by comparing ground truth observations and classification outputs at specific locations. These indices may be applied at the map-level or at the class level. However, the spatial variation of the accuracy is not captured by those statistics. Pixel-level thematic uncertainty measures derived from class membership probability vectors can provide such spatially explicit information. In this paper, a new information-based criterion—the equivalent reference probability—is introduced to provide a synoptic thematic uncertainty measure that has the advantage of taking the maximum probability value into account while committing for the full set of probabilities. The fundamental theoretical properties of this indicator was first highlighted and its use was afterwards demonstrated on a real case study in Belgium. Results showed that the proposed approach positively correlates with the quality of the classification and is more sensitive than the classical maximum probability criterion. As this information-based criterion can be used for providing spatially explicit maps of thematic uncertainty quality, it provides substantial additional information regarding classification quality compared to conventional quality measures. Accordingly, it proved to be useful both for end-users and map producers as a way to better understand the nature of the errors and to subsequently improve the map quality.

✉ François Waldner
patrick.bogaert@uclouvain.be

[1] Earth and Life Institute—Environment, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Keywords Information theory · Classification · Thematic uncertainty · Confidence · Remote sensing

1 Introduction

In the framework of classification, the most frequent way of assessing the performance of a classifier is to compare the labels of the classification with independent ground truth observations (Stehman 1997). Accuracy measures have been designed to report accuracy both at the map level and at the class level [see (Story and Congalton 1986) for examples] and are typically assumed to apply uniformly over the region of interest. Yet several studies have also demonstrated that errors vary spatially (Liu et al. 2004; Foody 2005; Comber et al. 2012; Renier et al. 2015; Liu et al. 2015; Waldner et al. 2015b; Feng et al. 2015).

As global accuracy statistics cannot model this spatial variation adequately, statistics describing the map quality at a more local level are thus necessary. Foody (2005) applied local accuracy assessment by constraining geographically the data used for accuracy assessment and showed that local accuracy assessment provides a more complete understanding of the quality of land cover maps derived from remote sensing. Nonetheless, to obtain sub-regional accuracy estimates using conventional design-based accuracy assessment, validation data must ensure a sufficient sample size within the region of interest for precise estimates. Unfortunately, sufficient sub-regional data are rarely available to support this (Strahler et al. 2006). Cripps et al. (2013) presented a Bayesian method for quantifying the uncertainty that results from potential misclassification in remotely sensed land cover maps. Discrete remote sensing classification neglects intrinsically the fuzzy character of the land surface and, as a consequence, leads to the inclusion of uncertainty in class assignments (Van der Wel et al. 1998). Lunetta et al. (1991) give an overview of the sources of errors and uncertainties in remote sensing classification. Accordingly, several studies are addressing quality issues by propagating uncertainties in spatial datasets (Pontius 2000; Atkinson and Foody 2002; Crosetto and Tarantola 2001; Liu et al. 2004), while others found that addressing classification uncertainty improved subsequent model calibration (Cockx et al. 2014).

In remote sensing, measures like the posterior probability of membership to the allocated class are often used as an indicator of uncertainty on a per-case basis (Foody et al. 1992). Probably the simplest approach for visualizing the uncertainties underlying a remote sensing classification is by the way of a gray-scale map depicting the maximum probability (MP) $\max(\mathbf{p})$ of a probabilistic output vector $\mathbf{p} = (p_1, \ldots, p_k)$ (Van der Wel et al. 1998) where k is the number of classes. The direct use of the MP from other probabilistic classifiers is a common practice; see for instance Mitchell et al. (2013), Dronova et al. (2011) and Polikar (2006). For non-probabilistic classifiers, soft outputs might also be used as a proxy to class membership probability. In the random forest framework it is defined as the number of trees in the ensemble voting for the final class (Loosvelt et al. 2012a). In support vector machine classifications it is based on the distances of the samples to the optimal separating hyperplane in the feature space (Giacco et al. 2010), while for the multi-layer perceptron it is based on the activation levels (Brown et al. 2009). If these measures are not posterior probabilities *per se*, they can be regarded as such.

Another criterion was proposed by Mitchell et al. (2008) to approximate the likelihood. It relies on Euclidean distance of each pixel from its closest class, and then account for differing class variability by standardizing by the variance of the respective class, i.e.,

$$d = \sqrt{\sum \left(\frac{r_i - \bar{m}_i}{s_i} \right)^2} \tag{1}$$

where r_i is the pixel value in the ith band, $\bar{m}_i$ is the class average in the ith band, and s_i is the standard deviation of the class in the ith band. The ratio d_1/d_2 between the distance to the closest/assigned class centroid d_1 and the distance to the second closest class centroid d_2 along with the magnitude of these distances (for each pixel) provide additional information about the reliability of class label assignment.

Gonçalves et al. (2009) investigated how the incorporation of uncertainty associated with the classification of surface elements into the classification of landscape units affects the accuracy. The uncertainty criterion they selected is given by (Eastman, 2006)

$$U = 1 - \frac{\max(\mathbf{p}) - \sum(p_i)/k}{1 - 1/k} \tag{2}$$

with values for U lying in [0,1] and only depending on the maximum probability and the total number of classes. The numerator of the second term expresses the difference between the MP assigned to a class and the probability that would be associated with the classes if a maximum dispersion for all classes occurred, that is, if a probability of 1/k was assigned to all k classes. The denominator corresponds to the extreme opposite case, where the MP is 1 (and thus a total commitment to a single class occurs). The ratio of these two quantities expresses the degree of commitment to a specific class relative to the largest possible commitment.

As the previously mentioned approaches neglect the whole probability distribution, another popular approach relies on the entropy (i.e., Shannon's measure of information), with

$$H(\mathbf{p}) = -\sum_{i=1}^{k} p_i \ln p_i \tag{3}$$

Brown et al. (2009) assessed the thematic error on a per-pixel basis based on the entropy of the outputs of a classifier in order to estimate thematic uncertainty. Loosvelt et al. (2012a) used entropy to compare the performance of different features for crop classification. McIver et al. (2001) demonstrated that classification errors tend to have low classification confidence while correctly classified pixels tend to have higher confidence. Class membership vectors and Shannon entropy were also combined with parallel coordinate plots to highlight the distribution of probability values of different land cover types for each pixel, and also reflect the status of pixels with different degrees of uncertainty (Ge et al. 2009). Loosvelt et al. (2012b) used the empirical shape of the distribution of two uncertainty indicators to assess the prediction strength of a classification model. The two indicators were the uncertainty defined as $U = 1 - \max(\mathbf{p})$ and the entropy $H(\mathbf{p})$, both being computed at the pixel-level. They concluded that, although entropy is a more representative evaluation of uncertainty than $1 - \max(\mathbf{p})$ as it includes the entire probability vector in its calculation, the uncertainty measure based on $\max(\mathbf{p})$ can be considered as an equivalent alternative to entropy since the uncertainty assessment performed on both measures was similar. As Shannon's entropy assumes values in the interval $[0, \ln k]$, Maselli et al. (1994) proposed a measure of the relative probability entropy (RPH) with

$$\mathrm{RPH} = \frac{-\sum_{i=1}^{k} p_i \ln p_i}{\ln k} \quad (4)$$

Similarly, Dehghan and Ghassemian (2006) defined the Normalized Uncertainty Criterion (NUC) based on the entropy and compared it to three criteria in order to evaluate classification performance: (i) the mean relative error (MRE), (ii) the root mean squared error (RMSE) of the average squared difference between two desired and actual membership vectors (Zhang and Sun 2002), and (iii) the linear correlation coefficient (LCC), with

$$\mathrm{NUC} = 1 - \frac{\ln k - H(\mathbf{p})}{\ln k} \quad (5)$$

They concluded that the MRE, RMSE and LCC criteria have been defined based on actual and desired outputs of classifier. Therefore, these criteria are dependent on the error of the results and sensitive to error variations. Waldner et al. (2015a) showed that correctly classified pixels tend to display a lower uncertainty NUC than misclassified pixels.

Studies such as those by Giacco et al. (2010) and Löw et al. (2013), (2015b) relied on the α-quadratic entropy $H_\alpha(\mathbf{p})$. This measure is based on the concept of the multiplicative class introduced by Pal and Bezdek (1994), with

$$H_\alpha(\mathbf{p}) = \frac{1}{k \times (2^{-2\alpha})} \times \sum p_i^\alpha (1-p_i)^\alpha \quad (6)$$

where α is an exponent which determines the behavior of the uncertainty measure. Indeed, if α is close to zero, the measure is not very sensitive to small changes in the components p_i, while for α close to one, the uncertainty is higher for p_i close to 0.5. The advantage of this measure is that it summarizes all the information contained in p and commits the probabilities of the other classes in the uncertainty evaluation. It has been suggested that it has a higher sensitivity compared to Shannon's entropy (Löw et al. 2013). Yet, its definition depends on α with values that are often set arbitrarily. Löw et al. (2013) proposed a normalized version of the α-quadratic entropy, the relative α-quadratic entropy that simply consists in dividing $H_\alpha(\mathbf{p})$ by the maximum possible $H_\alpha(\mathbf{p})$, that is when the probabilities are evenly distributed in all categories with $p_i = 1/k$ for all i.

Van der Wel et al. (1998) used an uncertainty measure that builds on the notion of weighted uncertainty as proposed by Glasziou and Hilden (1989), namely the quadratic score (QS), with

$$\mathrm{QS} = \sum p_i \times (1-p_i) \quad (7)$$

that exhibits the same behavior with respect to its minimum and maximum values as does the entropy measure. The entropy and the quadratic score differ, however, with respect to their slopes.

Despite alternative approaches to characterize pixel-level thematic uncertainty with more elaborated criteria, the most popular way of assessing the performance of a classifier remains the rate of correctly classified items (or variations around this theme). Although the shortcomings of this simple approach have been clearly emphasized by many authors, it also remains true that most of the alternate way of assessing the accuracy that are proposed are based on *ad hoc* methods or indicators that lack strong epistemic grounds. As a direct consequence, this leads to a multiplication of these indicators, leaving the user without clear final guidelines.

It is true that a classifier which selects the category i^* that maximizes p_{i^*} over all possible other choices, i.e. $p_{i^*} = \max(\mathbf{p})$, is consistent with the maximum likelihood principle and nothing is intrinsically wrong either about considering p_{i^*} itself as uncertainty criterion. However, when it comes to comparing soft classification outputs, this leads to major difficulties. The limitation of the most probable category as an indicator of quality assessment is better illustrated with a very simple example, as given in Table 1 for $k = 4$ categories. When it comes to selecting at best the category c_i, the same choice c_1 would be made for all cases. However, when it comes to compare soft classification outputs, difficulties directly arise. Though cases from (*a*) to (*d*) share the same category c_1 as the most probable one, they widely differ with respect to probabilities p_2, p_3 and p_4. While (*a*) is concentrating the remaining probability $1 - p_1 = 0.3$ over a single category c_2 and (*b*) is distributing them evenly over these three categories, the corresponding p_1 is the same and does not allow to make a clear preference between these two cases. The same remark applies when comparing (*c*) with (*d*). A comparison between (*a*) and (*c*) would lead to the conclusion that (*c*) is more favourable, i.e. p_1 is higher while the remaining probability $1 - p_1$ is distributed over the same single category c_2. However, there is a major issue when it comes to comparing (*a*) with (*d*) and (*b*) with (*c*), as all probabilities are now different. Clearly, the difficulty of comparing these

Table 1 Illustrative examples when $k = 4$ for the values of p_1 and the way probabilities are distributed over the remaining categories

	p_1	p_2	p_3	p_4	max $(\mathbf{p})$
(*a*)	0.7	0.3	0	0	0.7
(*b*)	0.7	0.1	0.1	0.1	0.7
(*c*)	0.8	0.2	0	0	0.8
(*d*)	0.8	0.1	0.1	0	0.8

Each line corresponds to a distinct probability vector various cases is precisely coming from the necessity of accounting for the whole probability vector $\mathbf{p}$ based on a sound theoretical approach, so that meaningful comparisons can be made and clear conclusions can be reached afterwards. This is of course impossible when relying only on $\max(\mathbf{p})$.

Instead of discussing at length the benefits and limitations of all possible alternate approaches that have been advocated so far, we present in the present paper a way of assessing pixel-level thematic uncertainty by starting from scratch using information theory. To that aim, we will begin from the most elementary concept of information theory, i.e., the definition of information itself. It will be shown how an expected difference of information can account for the full set of probabilities, while remaining at the same time perfectly consistent with p_i^* when used as a simple assessment indicator or as a criterion for selecting the best category. The similarities and discrepancies with entropy-based criteria will also be emphasized. Following a rigorous statistical reasoning, one indicator is proposed: the equivalent relative probability derived from the information difference. Their use and usefulness is demonstrated with synthetic examples as well as with a real land cover classification case study.

2 The notion of difference of information

Let us consider a set of k non overlapping categories $\{c_1, \ldots, c_k\}$ with associated probabilities $\mathbf{p} = (p_1, \ldots, p_k)$ such that $\sum_i p_i = 1$. Let us consider that an arbitrary category c_i is observed. The information $I(p_i)$ which is gained by observing the occurrence of c_i is then given by

$$I(p_i) = -\ln p_i \geq 0 \tag{8}$$

i.e. the gain of information is equal to 0 when c_i is the sure event (i.e. when $p_i = 1$) and goes to infinity when $p_i = 0$. The information can be understood as measuring the surprise of seeing the outcome c_i, as the occurrence of a highly improbable event is very surprising, while the occurrence of a sure event does not cause any surprise (see Fig. 1).

Let us consider the information for a reference category c_{i^*} (where $i^* \in \{1, \ldots, k\}$) as the information to which the information for the other categories must be compared. Let us now define $D(i||i^*)$ as this difference of information, with

$$D(i||i^*) = I(p_i) - I(p_{i^*}) = \ln\left(\frac{p_{i^*}}{p_i}\right) \qquad \forall i \neq i^* \tag{9}$$

Clearly, if i^* is chosen as the most probable category so that $p_{i^*} = \max(\mathbf{p})$, then it comes directly that

$$D(i||i^*) \geq 0 \qquad \forall i \neq i^* \tag{10}$$

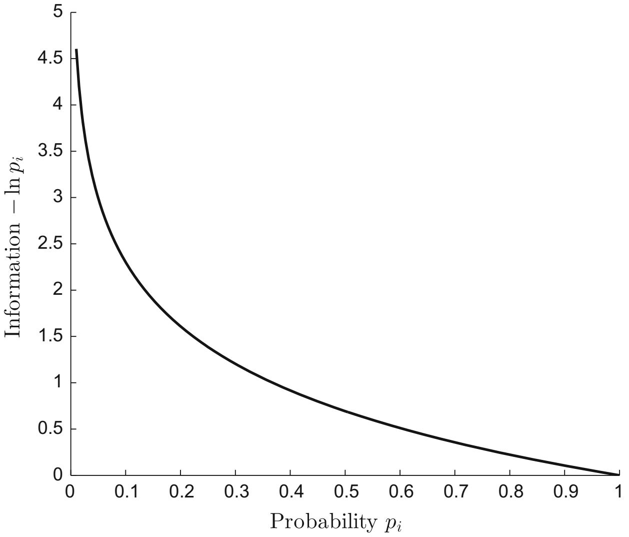

Fig. 1 Information $I(p_i) = -\ln p_i$ as a function of p_i

though choosing the most probable category as the reference category is not a mandatory choice for the forthcoming developments (e.g., the reference category can be the "true" category for the classification and its probability is not necessarily the highest one if the classification performs poorly).

As there are $k-1$ differences $D(i||i^*)$ to be accounted for when comparing each category c_i with the reference category c_{i^*}, one can summarize this set of differences by using their corresponding expected value. The expected difference of information when observing a category different from c_{i^*} is then given by

$$E[D(i||i^*)] = \sum_{i \backslash i^*} \frac{p_i}{1 - p_{i^*}} D(i||i^*) \tag{11}$$

where summation is done over all categories except c_{i^*} and where the values $p_i/(1 - p_{i^*})$ are the probabilities of observing the corresponding c_i's given the fact that $i \neq i^*$. Using Eq. (9) and because $\sum_{i \backslash i^*} p_i/(1 - p_{i^*}) = 1$, it comes directly that

$$\begin{aligned} E[D(i||i^*)] &= \sum_{i \backslash i^*} \frac{p_i}{1 - p_{i^*}} \ln\left(\frac{p_{i^*}}{p_i}\right) \\ &= \ln p_{i^*} - \frac{1}{1 - p_{i^*}} \sum_{i \backslash i^*} p_i \ln p_i \end{aligned} \tag{12}$$

2.1 The expected difference of information and its relationship with entropy

Before moving on with the interpretation of $E[D(i||i^*)]$ itself, it is worth noting that $E[D(i||i^*)]$ should not be confused with the expected gain of information *per se*, that corresponds to the traditional definition of entropy $H(\mathbf{p})$, with

$$H(\mathbf{p}) = E[I(\mathbf{p})] = \sum_{i=1}^{k} p_i I(p_i) = -\sum_{i=1}^{k} p_i \ln p_i \geq 0 \qquad (13)$$

Indeed, Eq. (13) is not associated with any specification for a reference category c_{i^*}, to the opposite of Eq. (12) where this choice is made explicit. Comparing Eqs. (12) and (13), it is however clear that these two quantities are directly related, as elementary algebraic manipulations from Eqs. (12) and (13) lead to the result

$$H(\mathbf{p}) = -\ln p_{i^*} + (1 - p_{i^*})E[D(i||i^*)] \qquad (14)$$

Clearly, $H(\mathbf{p})$ is a single value associated with a given probability vector $\mathbf{p}$ considered as a whole, while the values for $E[D(i||i^*)]$ are directly dependent both on $\mathbf{p}$ and on the choice that one makes for the reference category c_{i^*}, with its associated probability p_{i^*}.

It is also worth emphasizing that even if $E[D(i||i^*)]$ is measuring an expected difference of information, there is no direct connection with a classical measure of expected difference of information as given by the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951) nor with cross-entropy measures (Stehlík and Sivasundaram 2012). Indeed, while KL divergence and cross-entropies aim at comparing two distinct probability vectors, say p and q, in our case the comparison is always made with respect to a given reference category p_{i^*} belonging to a single probability vector p.

2.2 Fundamental properties

Resuming again from the interpretation of the probabilities as information, one can see that $E[D(i||i^*)]$ is measuring the average (difference of) surprise of observing any category c_i instead of the reference category c_{i^*}. If the issue is to select the reference category at best among a set of categories as represented by a probability vector $\mathbf{p}$, it is thus consistent to select c_{i^*} such that $E[D(i||i^*)]$ is maximized. When the problem at hand is to compare classifications as represented by two probability vectors $\mathbf{p}_j$ and $\mathbf{p}_{j'}$ with corresponding reference categories $c_{i_j^*}$ and $c_{i_{j'}^*}$, it is thus consistent to directly compare their corresponding expected difference of information $E[D(i_j||i_j^*)]$ and $E[D(i_{j'}||i_{j'}^*)]$ and to favor the classification which exhibits a higher expected difference of information.

We thus postulate that $E[D(i||i^*)]$ is a sound and natural way of assessing the quality associated with a probability vector $\mathbf{p}$ and the choice of a given c_{i^*} as reference category. In order to show this, the most important properties of $E[D(i||i^*)]$ will first be given. The corresponding proofs of the theorems are grouped in the appendices for the sake of conciseness. For the non-specialist reader, the proofs can thus be skipped without compromising the global understanding of the text. For each result, a special attention is also devoted to its interpretation and to the way it relates to specific and important cases. Furthermore, the use of $E[D(i||i^*)]$ will be illustrated using simple but carefully selected synthetic examples

Theorem 1 *Given any probability vector* $\mathbf{p} = (p_1, \ldots, p_k)$ *with* $\sum_i p_i = 1$ *and two possible reference categories* i^* *and* i^{**}, *with* $i^* \neq i^{**}$. *If* $p_{i^*} > p_{i^{**}}$, *then* $E[D(i||i^*)] > E[D(i||i^{**})]$.

This result states that, for any probability vector $\mathbf{p} = (p_1, \ldots, p_k)$, the values for the expected difference of information are sorted in the same order than the probabilities. A direct consequence is that the category with MP is also the reference category that maximizes the expected difference of information. For a given vector $\mathbf{p}$, selecting the most probable category as the reference category according to the maximum likelihood principle is thus equivalent to selecting c_{i^*} that maximizes $E[D(i||i^*)]$. In order to illustrate this property, Table 2 is presenting an arbitrary probability vector $\mathbf{p}$ when $k = 4$, along with the four possible choices for the reference category c_{i^*} and their associated $E[D(i||i^*)]$ values. Clearly, maximizing either p_{i^*} or $E[D(i||i^*)]$ will lead to the selection of the same category c_3, with a same ordering of values for the other possible choices.

Theorem 2 *If* $p_i \leq p_{i^*}$ $\forall i \neq i^*$, *then* $E[D(i||i^*)] \geq 0$ *with equality if and only if* $p_i = p_{i^*} = \frac{1}{k'}$ *for all* $k' \leq k$ *categories with associated non null probabilities.*

In other words, as long as the reference category is the most probable one, the expected difference of information is non-negative. In this case, the lowest possible value is equal to 0 and will only occur if there is a tie among all categories with non null probabilities, i.e. when there is an ambiguity when it comes to selecting at best a reference category among the set of k' candidate categories, considering that all $k - k' \geq 0$ categories with null probabilities are out of the competition. In order to illustrate these results, one can remark first from Table 2 that $E[D(i||i^*)]$ values can be negative (for our example this occurs for the two least probable categories), while the maximum value

Table 2 Synthetic example for a probability vector $\mathbf{p} = (0.1, 0.2, 0.4, 0.3)$, with c_3 as the most probable category, that also maximizes the value for $E[D(i||i^*)]$

i^*	p_{i^*}	$E[D(i\|\|i^*)]$
1	0.1	−1.1364
2	0.2	−0.4120
3	0.4	0.6059
4	0.3	0.1084

One can see that the values for $E[D(i||i^*)]$ are sorted in the same order than the p_{i^*}'s for $E[D(i||i^*)]$ is positive and occurs when c_3 is chosen as the reference category, i.e. when satisfying the condition $p_3 > p_i \; \forall i \neq 3$. Additionally, let us consider Table 3 where three different probability vectors sharing the same MP $p_1 = 0.5$ are considered, where it can be seen that $E[D(i||i^*)] \geq 0$ according to the theorem. The case $E[D(i||i^*)] = 0$ occurs when there is a tie with two categories sharing the same MP. It is worth noting too that this is not restricted to the case where the tie is only about two categories. Indeed, considering the probability vector $\mathbf{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$ would lead again to $E[D(i||i^*)] = 0$ by choosing any reference category among $\{c_1, c_2, c_3\}$.

Theorem 3 *Given a reference category i^* with probability p_{i^*}. The minimum possible value for $E[D(i||i^*)]$ is thus given by*

$$\mathcal{L}(p_{i^*}) \doteq \min_{\mathbf{p} \backslash p_{i^*}} E[D(i||i^*)] = \ln p_{i^*} - \ln(1 - p_{i^*}) \tag{15}$$

and it occurs if and only if there is a single non null $p_i = 1 - p_{i^*}$ with $i \neq i^*$.

For a given p_{i^*}, the lower bound is thus reached when the complementary probability $1 - p_{i^*}$ is concentrated over a single category (so all other categories have null probabilities). Considering $E[D(i||i^*)]$ as an accuracy assessment, this thus corresponds to the least favourable case. One can also remark that this lower bound corresponds to the expected difference of information when $k = 2$. Indeed, from Eq. (12) with $k = 2$, it comes that

$$\begin{aligned} E[D(i||i^*)] &= \ln p_{i^*} - \frac{1}{1 - p_{i^*}} (1 - p_{i^*}) \ln(1 - p_{i^*}) \\ &= \ln p_{i^*} - \ln(1 - p_{i^*}) \end{aligned} \tag{16}$$

This last result is also illustrated in Table 3(a) by considering c_1 as the reference category, where the lower bound is then precisely equal to $\ln \frac{1}{2} - \ln(1 - \frac{1}{2}) = 0$. However, Eq. (15) applies in a more general way even if the chosen reference category is not the most probable one (though the situation where the most probable category corresponds to the reference category is of particular interest, of course). Looking again at Eq. (15), it is worth noting that $\mathcal{L}(p_{i^*})$ is monotonically increasing with p_{i^*}, as seen from Fig. 2. Moreover, the value for $\mathcal{L}(p_{i^*})$ does not depend on the number k of categories. When combined with the results for the upper bound as given below, these remarks will prove to be useful for practical purposes.

Table 3 Synthetic example for three probability vectors $\mathbf{p}$ sharing the same MP value occurring for category c_1

	p_1	p_2	p_3	p_4	$E[D(i\|\|i^*)]$
(a)	0.5	0.5	0	0	0
(b)	0.5	0.3	0.1	0.1	0.9503
(c)	0.5	0.2	0.2	0.1	1.0549

When selecting c_1 as the reference category (i.e. as the most probable one), the lowest possible value $E[D(i||i^*)] = 0$ is reached when there is a tie between c_1 and c_2

Theorem 4 *Given a set of k categories and a reference category i^* with probability p_{i^*}. The upper bound for $E[D(i||i^*)]$ is then given by*

$$\mathcal{U}(p_{i^*}, k) \doteq \max_{\mathbf{p} \backslash p_{i^*}} E[D(i||i^*)] = \ln p_{i^*} - \ln\left(\frac{1 - p_{i^*}}{k - 1}\right) \tag{17}$$

and it occurs if and only if $p_i = \dfrac{1 - p_{i^*}}{k - 1} \; \forall i \neq i^*$.

For a given p_{i^*}, the highest possible value for the expected difference of information occurs when all other categories are equiprobable. This is an indirect consequence of the fact that equiprobable non-reference categories are maximizing the entropy over these categories. In order to illustrate the relationship between the values for $E[D(i||i^*)]$ and the way probabilities are distributed over the remaining categories, let us consider Table 4 with three different probability vectors $\mathbf{p} = (p_1, p_2, p_3, p_4)$ sharing the same maximum probability value $p_1 = 0.7$ for category c_1.

Clearly, when choosing c_1 as the reference category, the lowest possible value $E[D(i||i^*)]$ is reached for case (a) where the complementary probability $1 - p_1 = 0.3$ is concentrated over a single category, as previously stated by Eq. (15). It is worth noting too that this minimum value is higher than 0, as all non-null probabilities are not equal, to the opposite of Table 3(a). On the other side, the maximum possible value is reached for case (c) where $1 - p_1 = 0.3$ is distributed evenly over the three remaining categories.

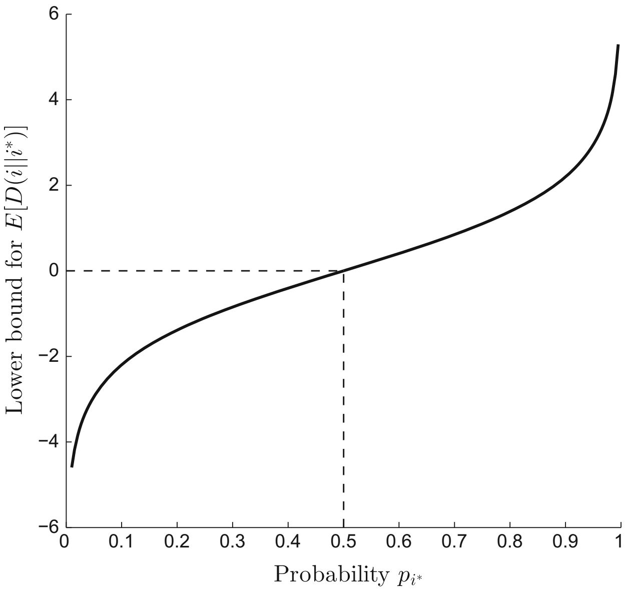

Fig. 2 Lower bound for $E[D(i||i^*)]$ as a function of the probability p_{i^*} of the reference category c_{i^*}

From Eq. (17), one can also see that the upper bound does depend both on p_{i^*} and k, where $\mathcal{U}(p_{i^*}, k)$ is monotonically increasing both with p_{i^*} and k. Combining the formulas for the lower and upper bounds as given by Eqs. (15) and (17) on the same graph leads directly to Fig. 3. For the special case $k = 2$, one can see from Eqs. (15) and (17) that the lower and upper bounds are identical. This is a direct consequence of the fact that p_{i^*} completely defines the distribution, as the only possible probability value for the single other category is $1 - p_{i^*}$, of course. One can also remark from Eqs. (15) and (17) that the difference between these bounds does not depend on p_{i^*}. Indeed,

$$\mathcal{U}(p_{i^*}, k) - \mathcal{L}(p_{i^*}) = \ln(k - 1) \tag{18}$$

and this can also be seen from Fig. 3 where all curves are parallel to each other.

The way the upper bound is changing with k can be illustrated with a simple example given in Table 5. Let us consider various probability vectors $\mathbf{p}$ sharing the same maximum probability $p_1 = \frac{1}{2}$ but where the complementary probability $1 - p_1 = \frac{1}{2}$ is evenly distributed over an increasing number $k - 1$ of remaining categories. Using the same reference category c_1, the upper bound is accordingly increasing with k.

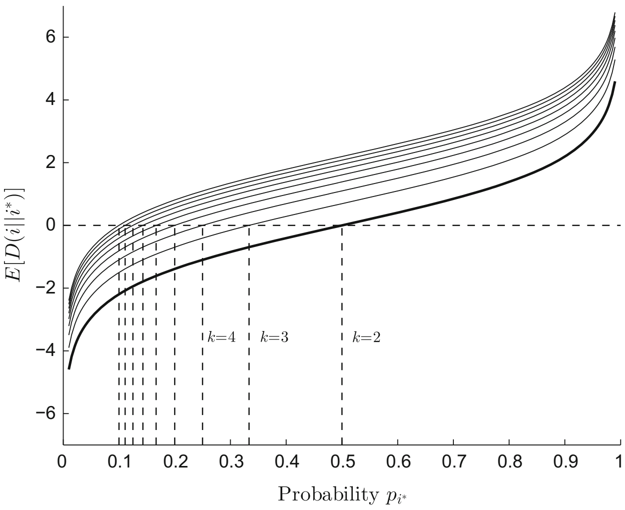

Fig. 3 Lower bound (*thick line*) and upper bounds (*thin lines*) for $E[D(i||i^*)]$ as a function of the probability p_{i^*} for the reference category and the number k of categories (when $k = 2$, the upper and lower bounds are identical and equal to $E[D(i||i^*)]$). *Vertical lines* specify the value of $p_{i^*} = \frac{1}{k}$ for which the upper bound is equal to 0

Table 5 Synthetic example for three probability vectors $\mathbf{p}$ sharing a same MP value occurring for category c_1 and remaining probabilities that are evenly distributed over an increasing number of categories k

k	p_1	p_2	p_3	p_4	$E[D(i\|\|i^*)]$
2	$\frac{1}{2}$	$\frac{1}{2}$	–	–	0
3	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	–	0.6931
4	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1.0986

2.3 Categories with null probabilities

Though this might not appear as an obvious result from the previous developments, it is worth remarking that the only categories that are playing an effective role in Eq. (12) are those that are associated with a non-null probability of occurrence, as all categories with null probabilities p_i's will be filtered out. Indeed, from the result $\lim_{p_i \to 0} p_i \ln p_i = 0$, the only categories that are accounted for in $E[D(i||i^*)]$ are those for which $p_i \neq 0$. However, though $E[D(i||i^*)]$ does not depend on these null probabilities, this is not the case for the upper bound $\mathcal{U}(p_{i^*}, k)$ as given by Eq. (17), where it is the total number of categories that need to be accounted for. In order to illustrate this subtlety, let us consider Table 6 where two probability vectors have been given, with $k = 3$ categories for the first one and $k = 4$ for the second one. Both vectors are identical with respect to the three first categories and share the same value $E[D(i||i^*)] = 1.0986$, along with the same lower bound $\mathcal{L}(p_{i^*}) = 0.455$. However, they do not share the same upper bound. Clearly, the case $k = 4$ is far from the upper bound that would be reached if the complementary probability $1 - p_1 = 0.4$ would be evenly spread over the remaining categories p_2, p_3 and p_4. Considering $E[D(i||i^*)]$ as a measure of quality, the case where $k = 3$ is thus much more favourable (it reaches the upper bound) than the case where $k = 4$.

Table 4 Synthetic example for three probability vectors $\mathbf{p}$ sharing the MP value occurring for category c_1

	p_1	p_2	p_3	p_4	$E[D(i\|\|i^*)]$
(*a*)	0.7	0.3	0	0	0.8473
(*b*)	0.7	0.2	0.1	0	1.4838
(*c*)	0.7	0.1	0.1	0.1	1.9459

When selecting c_1 as the reference category (i.e. as the most probable one), the highest possible value for $E[D(i||i^*)]$ is reached when all other categories are equiprobable

This also emphasizes that, as soon as one wants to compare classifiers over a distinct number of categories, the value for $E[D(i||i^*)]$ cannot be interpreted as is without referencing it to the way $E[D(i||i^*)]$ is located with respect to the corresponding upper bound (the lower bound remaining the same as it does not depend on k). It will be shown a little bit further that a way of accounting for this

Table 6 Synthetic example for two probability vectors **p** sharing the same MP value occurring for category c_1

k	p_1	p_2	p_3	p_4	$\mathcal{L}(p_{i^*})$	$E[D(i\vert\vert i^*)]$	$\mathcal{U}(p_{i^*})$
3	0.6	0.2	0.2	–	0.4055	1.0986	1.0986
4	0.6	0.2	0.2	0	0.4055	1.0986	1.5041

When selecting c_1 as the reference category, the case where $k = 3$ is maximizing $E[D(i||i^*)]$ while the case where $k = 4$ does not reach the corresponding upper bound

relative location of $E[D(i||i^*)]$ is through the use of an "upper bound equivalent probability".

2.4 Negative values for $E[D(i||i^*)]$

As seen from the curves in Fig. 3, the lower and upper bounds for $E[D(i||i^*)]$ are monotonically increasing functions of p_{i^*} and they can also be negative. From Eq. (17), negative values for $E[D(i||i^*)]$ will necessarily occur as soon as the corresponding upper bound $\mathcal{U}(p_{i^*},k)$ is lower or equal to 0. From Eq. (17), it comes that

$$\mathcal{U}(p_{i^*},k) = 0 \iff p_{i^*} = \frac{1-p_{i^*}}{k-1} \iff p_{i^*} = \frac{1}{k} \tag{19}$$

as also seen from Fig. 3. If the reference category c_{i^*} exhibit a probability $p_{i^*} < \frac{1}{k}$, then there is obviously at least another probability for a category c_i such that $p_i > \frac{1}{k}$. Stated in other words, negative values for $E[D(i||i^*)]$ will automatically arise when the chosen reference category c_{i^*} is not the most probable one (though the opposite is not true : a negative value for $E[D(i||i^*)]$ does not necessarily imply that the chosen reference category is not the most probable one).

2.5 Equivalent reference probability

As is, the value for $E[D(i||i^*)]$ can be used for comparing different classifiers as long as they share the same number k of categories. Its values are necessarily lying in the interval $[\mathcal{L}(p_{i^*},k),\mathcal{U}(p_{i^*},k)]$, so that for any probability vectors **p** and a chosen reference category c_{i^*} one can see how the corresponding value $E[D(i||i^*)]$ is close or far from these lower and upper bounds. However, for people used to deal with probabilities, the interpretation of the $E[D(i||i^*)]$ values are made more difficult due to the fact that both $\mathcal{L}(p_{i^*})$ and $\mathcal{U}(p_{i^*},k)$ are unbounded. Indeed, both from Eqs. (15) and (17) along with Fig. 3, it is clear that

$$\begin{aligned} \lim_{p_{i^*}\to 0} \mathcal{L}(p_{i^*}) &= \lim_{p_{i^*}\to 0} \mathcal{U}(p_{i^*},k) = -\infty \\ \lim_{p_{i^*}\to 1} \mathcal{L}(p_{i^*}) &= \lim_{p_{i^*}\to 1} \mathcal{U}(p_{i^*},k) = +\infty \end{aligned} \tag{20}$$

so that $E[D(i||i^*)]$ is taking its value over the real line from $-\infty$ to $+\infty$. In order to circumvent this problem and to ease the interpretation of $E[D(i||i^*)]$, its value can be converted in an "upper bound equivalent probability". Indeed, for an arbitrary probability vector **p** with a chosen reference category c_{i^*} with associated probability p_{i^*}, let us look in Eq. (17) for a corresponding value of probability p^* so that $E[D(i||i^*)]$ would match the upper bound, i.e.

$$E[D(i||i^*)] = \ln p^* - \ln\left(\frac{1-p^*}{k-1}\right) \tag{21}$$

Solving now for p^*, i.e., the equivalent reference probability, with respect to $E[D(i||i^*)]$ and k leads to the result

$$p^* = \frac{\exp(E[D(i||i^*)])}{\exp(E[D(i||i^*)]) + k - 1} \tag{22}$$

where p^* has the meaning of an "upper bound equivalent probability" that one can associate with any value for $E[D(i||i^*)]$. As p^* is a probability, its values are now restricted to the [0, 1] interval. From the monotonic property of $\mathcal{U}(p^*,k)]$ as a function of p^*, it is also clear that $p^* \le p_{i^*}$, i.e. this equivalent probability is always lower or equal to the reference probability p_{i^*}, with equality if and only if $E[D(i||i^*)]$ is precisely corresponding to the upper bound.

In order to illustrate this, let us consider Table 7 where c_1 is chosen as the reference category for the probability vector in (a). Solving for p^* using Eq. (22) leads to the result $p^* = 0.5$. Accordingly, the probability vector in (b) where $p_1 = p^* = 0.5$ can be viewed as an equivalent case, in the sense that it has the same $E[D(i||i^*)]$ value but this value now corresponds to the upper bound when c_1 is chosen as the reference category (note however that any permutation of the probabilities in (b) would lead to the same result as long as the same probability value is used for the reference category, of course, so that p^* is not intended to be associated with any specific category). Focusing now on the graphic representation of this equivalence between $E[D(i||i^*)]$ and p^* as given in Fig. 4, it can be seen that looking for the value of p^* is done by moving horizontally leftwards from the point $(p_{i^*}, E[D(i||i^*)])$ up to the curve corresponding to the upper bound $\mathcal{U}(p^*,k)]$, making also clear that the result $p^* \le p_{i^*}$ necessarily holds true. Clearly too, the closer $E[D(i||i^*)])$ is from the upper bound $\mathcal{U}(p_{i^*},k)]$, the closer p^* will be from p_{i^*}.

Table 7 Synthetic example for two probability vectors **p** sharing the same $E[D(i||i^*)]$ value but where the last vector corresponds to the upper bound when $k = 4$

	p_1	p_2	p_3	p_4	$E[D(i\vert\vert i^*)]$	$\mathcal{U}(p_{i^*})$
(a)	0.6	0.2	0.2	0	1.0986	1.5041
(b)	0.5	0.16	0.16	0.16	1.0986	1.0986

3 Synthetic examples

In order to illustrate now the use of $E[D(i||i^*)]$ as an uncertainty, let us use an augmented version of Table 1, as given by Table 8. Without loss of generality, let us consider $k = 4$ and a reference category c_1 chosen here as the most probable one. Clearly, cases (a) and (b) are respectively the lower and upper bounds for $E[D(i||i^*)]$ when $p_{i^*} = 0.7$. As a consequence, any intermediate case sharing the same p_{i^*} value will have a value $E[D(i||i^*)] \in [0.85, 1.95]$ as, e.g., for case (c). Comparing now case (a) with case (d) and case (c) with case (e) for which the probabilities are distributed with the same logic over c_2, c_3 and c_4, it can be seen that increasing p_{i^*} will lead to an increase for $E[D(i||i^*)]$, as expected. However, higher p_{i^*}'s do not necessarily correspond automatically to higher $E[D(i||i^*)]$'s. Indeed, comparing directly cases (b) and (d) which are respectively the most favorable case when $p_{i^*} = 0.7$ and the least favorable one when $p_{i^*} = 0.8$, $E[D(i||i^*)]$ is still favouring case (b) over case (d), as the even distribution of the probabilities over categories c_2, c_3 and c_4 in (b) does compensate the higher probability for c_1 in (d). Using $E[D(i||i^*)]$ as a sorting criterion from the most favourable to the least favourable case, the ordering is now (e), (b), (c), (d), (a). Clearly, $E[D(i||i^*)]$ allows us to directly compare here the various case using a single criterion that simultaneously accounts for the effect of the reference category probability and the way other probabilities are distributed over the remaining categories.

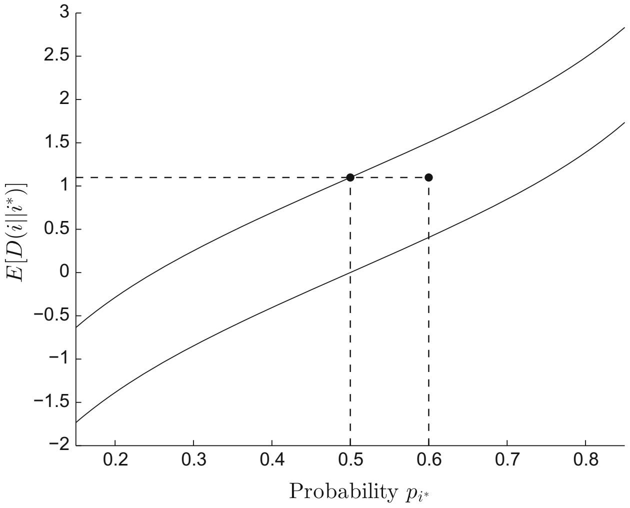

Fig. 4 Upper bound equivalent probability $p^* = 0.5$ for $E[D(i||i^*)] = 1.0986$ when $p_{i^*} = 0.6$ and $k = 4$. The *lower curve* is the lower bound $\mathcal{L}(p_{i^*})$ while the *upper curve* is the upper bound $\mathcal{U}(p_{i^*}, k)$ when $k = 4$

Table 8 Illustrative examples when $k = 4$ for the values of p_{i^*} and $E[D(i||i^*)]$, where (a) and (b) are the lower and upper bounds when $p_{i^*} = 0.7$, while (d) is the lower bound when $p_{i^*} = 0.8$ (the value for the upper bound is equal to 2.48)

	p_1	p_2	p_3	p_4	p_{i^*}	$E[D(i\|\|i^*)]$
(a)	0.7	0.3	0	0	0.7	0.85
(b)	0.7	0.1	0.1	0.1	0.7	1.95
(c)	0.7	0.15	0.15	0	0.7	1.54
(d)	0.8	0.2	0	0	0.8	1.39
(e)	0.8	0.1	0.1	0	0.8	2.08

4 Evaluation using remote sensing data

Satellite images were downloaded over a 30 × 30 km^2 area in Belgium centered on 50.60°N, 4.68°E from which land/crop cover maps were derived by a random forest classifier. It should be emphasized here that the purpose was not to achieve the highest level of accuracy but rather to demonstrate (1) how the equivalent reference probability (ERP) as defined by Eq. 22 can complement traditional accuracy assessments and (2) how ERP criterion compares with the MP criterion.

4.1 Study area and data

The study site is located in central agricultural loamy region of Belgium. The typical field size ranges from 3 to 15 ha and the dominant crop types are winter wheat, winter barley, potatoes, sugar beet, and corn. Winter crops are generally sown in October and harvested in August at the latest whereas summer crops are sown in April and harvested from September onward. Other dominant land covers include pastures, forests, artificial lands and water bodies. The landscape topography is flatlands and hills. The climatic zone is temperate with annual rainfall of about 780 mm that are relatively well distributed over the year, therefore irrigation is not frequent.

Two cloud-free SPOT-4 images and one cloud-free Landsat-8 image were at hand: the SPOT-4 imagery was acquired during the spring season (2014-04-02 and 2014-05-27) while the Landsat imagery was acquired at the end of the summer season (2014-09-30) (Fig. 5). Therefore, the Landsat-8 image is critical to discriminate between summer crops such as corn, potato and sugar beet. Both the SPOT-4 and the Landsat-8 data were calibrated, orthorectified and corrected for the atmosphere (Hagolle et al. 2008, 2015). The Landsat-8 image was resampled to SPOT-4's resolution and only the first seven spectral bands were kept.

The targeted legend includes eleven classes: six crop types [winter barley (WB), winter wheat (WW), sugar beet (SB), potato (Po), corn (C) and other crops (OC)], pasture

(Pa), forest (F), artificial areas (A) and water bodies (W) [see Radoux et al. (2016) for a separability analysis of the main land cover classes in the area]. One thousand calibration samples were randomly extracted from a data set combining the land parcel identification system and the land cover map of Wallonia. Similarly, 2000 samples independent from the training data were randomly selected to constitute the validation dataset.

4.2 Evaluation methodology

Based on the training data set, a random forest classifier was trained and applied on the three collected images (Fig. 6a). Random forest is an ensemble learning method for classification that operates by constructing a multitude of decision trees and outputting the class that is the mode of the decision of all the trees. Random forest have been widely used to derive land cover maps from remotely sensed data (Gislason et al. 2006; Rodriguez-Galiano et al. 2012; Waldner et al. 2015c; Löw et al. 2015a). Rodriguez-Galiano et al. (2012) demonstrated that random forest does not overfit and offers several advantages such as (1) the low number of user-defined hyper-parameters, (2) the estimation of the importance of variables (bands) for the general classification of the land-cover categories and for the classification of each category by means of the Gini Index, and (3) its robustness to noise and training data set size reduction. The reference samples were then used to derive accuracy measures corresponding to the classification and the thematic uncertainty was assessed by means of the ERP (Fig. 6c, f, j) and the MP (Fig. 6d, g, j). Pixel-level equivalent reference probability were also computed to assess the thematic uncertainty of the classifications using Eq. 22.

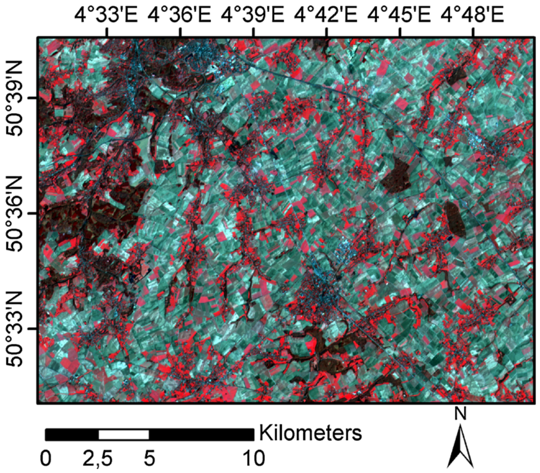

Fig. 5 False color infrared image of the SPOT-4 acquisition of the 2014-04-02. Forests appear in *dark red*, winter crops in *light red*, summer crops in *green* and built up areas in *dark blue*

4.2.1 *Qualitative analysis and spatial patterns*

Pixel-level thematic uncertainty measures are useful to underline patterns of uncertainty in the map. As seen from Fig. 6, classes seem to be associated with similar uncertainty level which allows users to recognize class-specific spatial patterns. Linear class transitions (mixed pixels) such as field boundaries and roads are especially well identifiable. The ERP images are more contrasted than MP images (darker areas) as ERP can be seen as a penalized version of MP as a function of the membership probability vector distribution. This is especially visible comparing forest uncertainty in Fig. 6c, d. To better highlight patterns in the spatial distribution of the uncertainty, the average equivalent reference probability was computed for each class and for different distances to the class boundary (Fig. 7). Two main conclusions can be drawn. First, edge pixels are classified with a higher uncertainty (low ERP) which is explained easily as a result of mixing the spectral signature at class transition. This effect tends to vanish after 40 m (two pixels), except for the other crop class which gather marginal crops that may have diverse spectral signatures. Second, the average thematic uncertainty depends of the class considered. The high uncertainty of the water body class may be explained by the small size of the water features in the landscape.

4.2.2 *Quantitative analysis and relationship to class-level accuracy measures*

To quantitatively evaluate the proposed indicator, thematic uncertainty measures and classification errors were compared. The results from this comparison were then used to establish if thematic uncertainty is positively correlated with classification accuracy and can therefore indicate classification quality. Results demonstrate that the proposed approach successfully predicts the quality of the classification and is more sensitive than MP.

As a first way of assessing the ability of ERP to relate with classification accuracy, frequency distributions were plotted for correctly and incorrectly classified samples, respectively, regardless of their class (Fig. 8). The shape of these distributions was then analyzed to evaluate if ERP is a reliable spatial measure to predict errors in the land cover map. The underlying assumption is that high ERP values are associated with correctly classified samples. Similarly, wrongly classified samples should in principle be characterized by low ERP, that is when the classifier algorithm had substantial doubt about the final class decision. If high ERP values indicate correct classification, and low ERP incorrect classification, then ERP successfully indicates the spatial distribution of misclassification (or correct classification) in the map. It can be seen that the two distributions

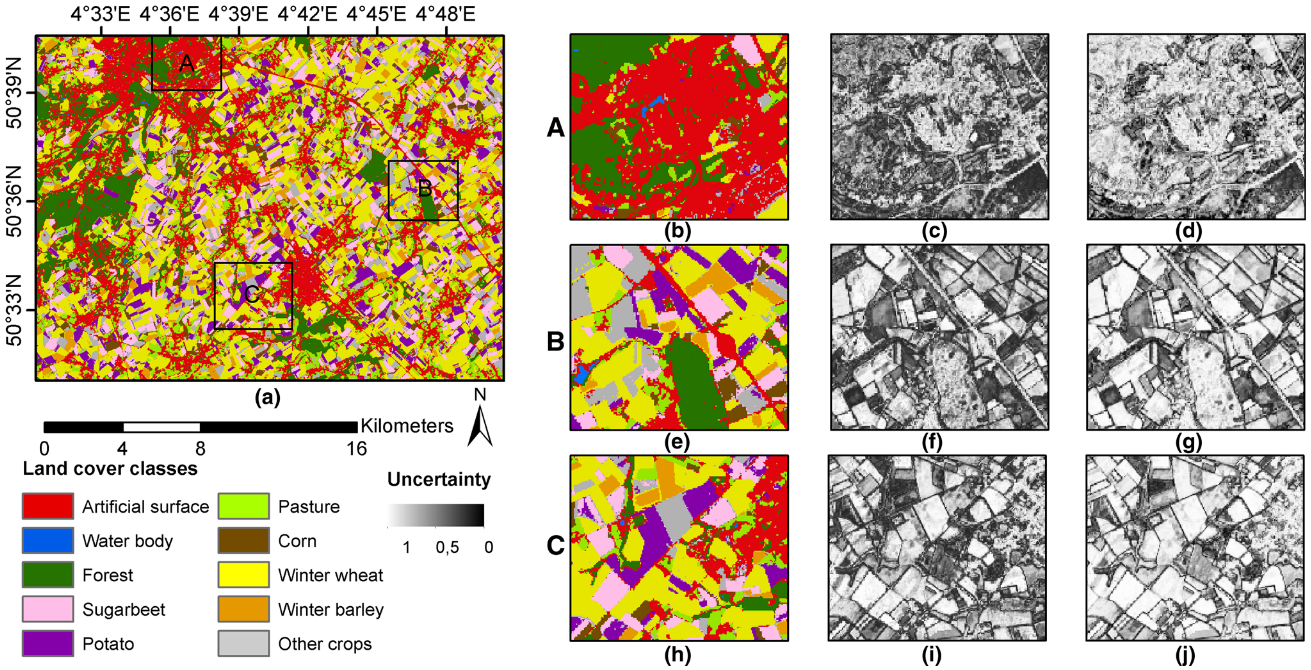

Fig. 6 Land cover classification of the study area (**a**) and zooms on three areas of interest (**b**, **e**, **h**), including their associated ERP (**c**, **f**, **i**) and MP (**d**, **g**, **j**) spatial distributions

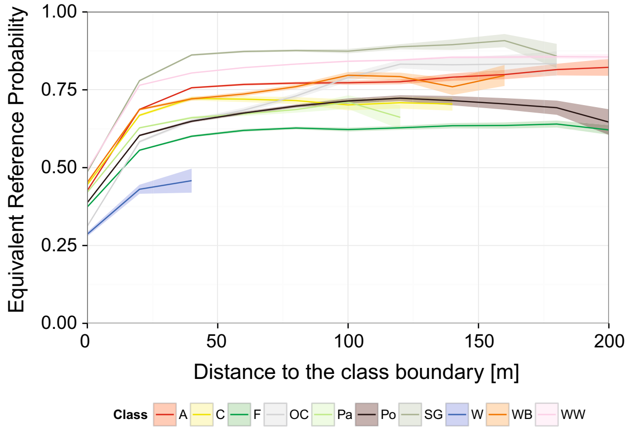

Fig. 7 Average equivalent reference probability by class as a function of the distance to the class boundary. The average ERP by class varies as a function of the distance and of the class itself

respect this hypothesis: correctly classified samples are associated with high ERP (high confidence) and conversely.

In a second test aiming at the utility of the class confidence estimated by ERP, we compared class-specific accuracies derived from the confusion matrix with mean predicted classification confidences (Fig. 9). Results show that for all classes the mean class confidence seems to slightly underestimate the proportion of well classified except for the Water body class. Nonetheless, the mean classification confidence for each class remains in general closely correlated with the accuracy (Pearson-R = 0.8). Therefore, these results suggest that the mean confidence provides a reliable indicator of the proportion of correctly classified pixels.

A final important consideration for the information-based criterion presented in this paper is its sensitivity to accuracy compared to the MP approach. A closer inspection of the differences between the uncertainty assessed with the MP and the equivalent reference probability further supports the validity of the newly introduced measure. In areas of high disagreement between the two indicators, i.e., when MP is substantially larger than ERP, ERP better captures variations in accuracy (Fig. 10). On the contrary, the maximum probability appears mostly insensitive to variations in accuracy once a certain accuracy threshold is reached ($\sim$0.7). This enhanced sensitivity results from the fact that ERP commits for the whole class membership vector. ERP is a thematic uncertainty criterion that is more sensitive than MP and its sensitivity allows a better representation of the class accuracy.

5 Discussion and conclusions

This paper presents a new criterion to derive thematic uncertainty measures from pixel-level class membership outputs as provided by classifiers. This indicator—the equivalent reference probability—is built on the concept of information as defined in information theory. Its derivation from the expected difference of information has been

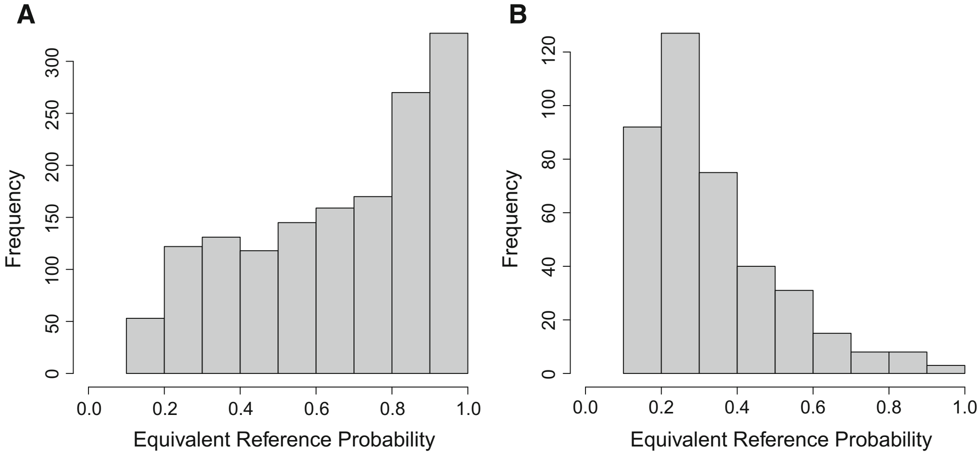

Fig. 8 Frequency distributions of ERP of, **a** correctly, **b** incorrectly classified samples. Note the difference of scale between the two abscissa axes of the histograms

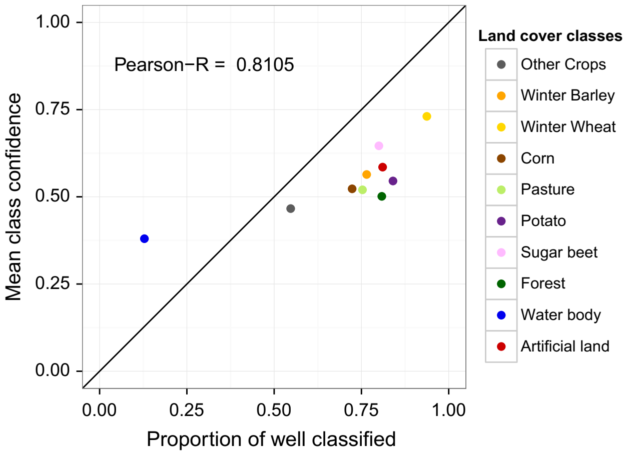

Fig. 9 Comparison of the mean class confidence expressed as mean ERP by class and proportion of correctly classified samples for each considered land cover class

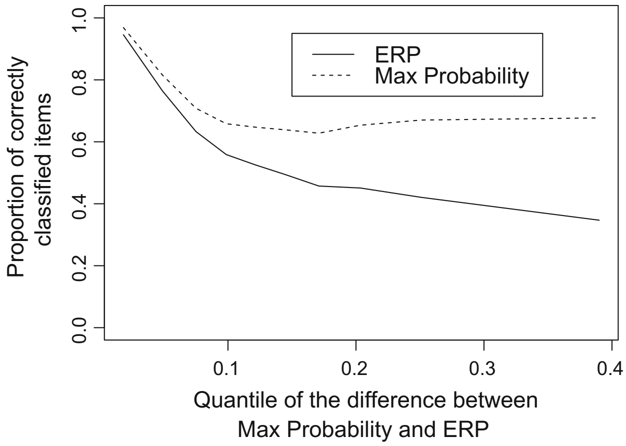

Fig. 10 Sensitivity of the equivalent reference probability and the MP to the proportion of correctly classified items. The thematic uncertainty expressed with ERP better captures variations in accuracy than MP

demonstrated. Theorems and simple synthetic examples illustrated how it can account for the full set of probabilities, while remaining at the same time perfectly consistent with the MP both when used as a simple assessment indicator or as a criterion for selecting the best category. Additionally, the ERP does not rely on any tuning procedure, and it can be derived from any classifier that provides soft outputs, either probabilistic or based on probability membership proxies—number of trees, distance to the separating plane, activation level, etc.

The fundamental theoretical properties of the expected difference of information leading to the definition of the ERP were first demonstrated. In particular, it has been shown that the expected difference of information (i) is bounded, (ii) is consistent with the initial order of the input probability vector, and that (iii) as long as the reference category is the most probable one, the expected difference of information is non-negative. To ease the interpretation and comparison of the information-based criterion, we introduce the notion of equivalent reference probability, that bounds the expected difference of information between zero and one. Using synthetic examples, it has been shown how this index allows us to directly compare various cases of probability membership outputs using single values that simultaneously accounts for the effect of the reference category probability and for the way the other probabilities are distributed over the remaining categories. The usefulness and complementary information brought by the criterion was successfully highlighted in both synthetic and real data sets. Based on a case study, it has been shown that they provide a way of obtaining per-pixel classification confidences that are strongly correlated with classification accuracy (Pearson-R = 0.8).

The ERP criterion has been shown to be more sensitive than maximum probability criterion. For a given MP, the ERP varies as a function of the distribution of the remaining class membership probability vector which permits a finer characterization of the uncertainty. This exacerbated sensitivity highlighted in the real case study makes the ERP the fittest indicator for classification comparisons and benchmarking activities.

Reliable pixel-level thematic uncertainty indicators are critical because they provide a means of producing classification confidence that convey considerably more information about classification quality than traditional accuracy assessment measures. As classifying large areas repeatedly over time with high spatial resolution images is becoming more and more frequent, the local/regional relevance of simple global confusion matrices and their derived measures are continuously reduced.

This type of approach is interesting for providing a deeper and spatially explicit understanding of the quality of land cover maps as derived from remote sensing. Additionally, the indicator is also useful to visualize the uncertainty, to ease the monitoring of ecological conditions (Dronova et al. 2011) and to further improve the classification accuracy (Foody 2008; Gonçalves et al. 2009), e.g., by combining different classifier outputs (Liu et al. 2004) and fusing classifier decisions (Löw et al. 2015a). Such criterion could also inform about sampling strategies for selecting reliable pixels in the framework of vegetation monitoring, area estimates or subsequent classifications. Further research will focus on the link between uncertainty, class proportion and purity as well as on the way to integrate these information-based criteria within the classifiers themselves for optimal class selection.

Acknowledgments This research was funded in the framework of the Seventh Programme for research, technological development, and demonstration under Grant Agreement No. 603719. The Landsat data were obtained through the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP 440 DAAC), USGS/Earth Resources Observation and Science (EROS) Center, 441 Sioux Falls, South Dakota (https://lpdaac.usgs.gov/get_data). The SPOT4 imagery was obtained under the SPOT4/Take5 programme. Imagery is copyrighted to CNES under the mention: "CNES 2013, all rights reserved. Commercial use of the product prohibited".

Appendix 1: Proof of the first theorem

Let us consider that $p_{i^*} > p_{i^{**}}$ and let us write both $E[D(i||i^*)]$ and $E[D(i||i^{**})]$ by restricting the sum notation over the subset of identical categories, so that

$$\begin{aligned} E[D(i||i^*)] &= \ln p_{i^*} - \frac{\sum_{i\setminus\{i^*,i^{**}\}} p_i \ln p_i}{1-p_{i^*}} - \frac{p_{i^{**}} \ln p_{i^{**}}}{1-p_{i^*}} \\ E[D(i||i^{**})] &= \ln p_{i^{**}} - \frac{\sum_{i\setminus\{i^*,i^{**}\}} p_i \ln p_i}{1-p_{i^{**}}} - \frac{p_{i^*} \ln p_{i^*}}{1-p_{i^{**}}} \end{aligned} \tag{23}$$

After rearranging terms, the difference is thus given by

$$E[D(i||i^*)] - E[D(i||i^{**})] = A + B \tag{24}$$

where

$$\begin{aligned} A &= \left(\frac{1}{1-p_{i^*}} - \frac{1}{1-p_{i^{**}}}\right)\left(-\sum_{i\setminus\{i^*,i^{**}\}} p_i \ln p_i\right) \\ &= \frac{p_{i^*} - p_{i^{**}}}{(1-p_{i^*})(1-p_{i^{**}})}\left(-\sum_{i\setminus\{i^*,i^{**}\}} p_i \ln p_i\right) > 0 \end{aligned} \tag{25}$$

because all factors are positive, and

$$B = (\ln p_{i^*})\left(1 + \frac{p_{i^*}}{1-p_{i^{**}}}\right) - (\ln p_{i^{**}})\left(1 + \frac{p_{i^{**}}}{1-p_{i^*}}\right) \tag{26}$$

and so we need to prove that $B > 0$. After reducing to the same denominator and simplifying,

$$\begin{aligned} B > 0 &\Longleftrightarrow \frac{\ln p_{i^*}}{\ln p_{i^{**}}} > \frac{1-(p_{i^{**}})^2}{1-(p_{i^*})^2} \\ &\Longleftrightarrow \ln p_{i^*}(1-(p_{i^*})^2) > \ln p_{i^{**}}(1-(p_{i^{**}})^2) \end{aligned} \tag{27}$$

subject to the conditions $p_{i^*} > p_{i^{**}}$. As the function $(\ln p)(1-p^2)$ is monotonically increasing over [0, 1], this is always true and so $B > 0$, as requested. □

Appendix 2: Proof of the second theorem

From Eq. (12), it is clear that the second term is the expectation of the various $\ln p_i$'s when $i \neq i^*$. For the sake of conciseness, let us define

$$E_{\mathbf{p}\setminus p_{i^*}}[\ln \mathbf{p}] \doteq \frac{1}{1-p_{i^*}} \sum_{i\setminus i^*} p_i \ln p_i \tag{28}$$

where all possibly null probabilities are filtered out from the computation of $E_{\mathbf{p}\setminus p_{i^*}}[\ln \mathbf{p}]$ because $\lim_{p_i \to 0} p_i \ln p_i = 0$. From the properties of an expectation, it comes too that

$$\min_{\mathbf{p}\setminus p_{i^*}}(\ln \mathbf{p}) \le E_{\mathbf{p}\setminus p_{i^*}}[\ln \mathbf{p}] \le \max_{\mathbf{p}\setminus p_{i^*}}(\ln \mathbf{p}) \tag{29}$$

If $p_i \le p_{i^*}\ \forall i \neq i^*$, it thus comes that

$$E_{\mathbf{p}\setminus p_{i^*}}[\ln \mathbf{p}] \le \max_{\mathbf{p}\setminus p_{i^*}}(\ln \mathbf{p}) \le \ln p_{i^*} \tag{30}$$

leading to $E[D(i||i^*)] \ge 0$, as stated. Clearly, this also shows that, from Eqs. (12) and (30),

$$E[D(i||i^*)] = 0 \iff E_{\mathbf{p}\backslash p_{i^*}}[\ln \mathbf{p}] = \ln p_{i^*}$$
$$\iff \begin{cases} E_{\mathbf{p}\backslash p_{i^*}}[\ln \mathbf{p}] = \max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p}) \\ \ln p_{i^*} = \max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p}) \end{cases} \qquad (31)$$

Let us now consider the following possibilities for p_{i^*} subject to the condition $p_i \leq p_{i^*}\ \forall i \neq i^*$:

i. if $p_{i^*} > \frac{1}{2}$, then $p_i < p_{i^*}\ \forall i \neq i^*$ and it thus comes that $\max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p}) < \ln p_{i^*}$, i.e. $E[D(i||i^*)] > 0$;

ii. if $p_{i^*} = \frac{1}{2}$, then $\ln p_{i^*} = \max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p})$ implies that one and only one p_i (with $i \neq i^*$) is equal to $\frac{1}{2}$ and thus, when $k > 2$, all other p_i's must be null ;

iii. if $p_{i^*} < \frac{1}{2}$, then $\ln p_{i^*} = \max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p}) \iff p_{i^*} = \max_{\mathbf{p}\backslash p_{i^*}}(\mathbf{p})$ is impossible for $k = 2$, as for $k = 2$ we have under this condition

$$\sum_{i=1}^{k} p_i = p_{i^*} + \max_{\mathbf{p}\backslash p_{i^*}}(\mathbf{p}) = 2p_{i^*} < 1 \qquad (32)$$

while $\sum_{i=1}^{k} p_i = 1$ by definition. For $k > 2$, there are at least two p_i's > 0 (with $i \neq i^*$) with the highest one equal to $\max_{\mathbf{p}\backslash p_{i^*}}(\mathbf{p})$. On the other side, $E_{\mathbf{p}\backslash p_{i^*}}[\ln \mathbf{p}]$ reaches its upper bound $\max_{\mathbf{p}\backslash p_{i^*}}(\ln \mathbf{p})$ when all non null $\ln p_i$'s, are equal to $\max_{\mathbf{p}\backslash p_{i^*}}(\mathbf{p})$. It thus comes that, for all non null probabilities, $p_i = p_{i^*} = \frac{1}{k'}$ where $k' \leq k$ is the number of categories with non null probabilities so that $\sum_{i=1}^{k} p_i = 1$, as required.

This completes the proof, as the second case is consistent with the third one, i.e. $k' = 2$ and so $p_i = p_{i^*} = \frac{1}{2}$. □

Appendix 3: Proof of the third theorem

In order to prove this, let us remember that

$$\frac{1}{1 - p_{i^*}} \sum_{i\backslash i^*} p_i = 1 \qquad (33)$$

so that using this property,

$$\begin{aligned} \ln(1 - p_{i^*}) &= \ln(1 - p_{i^*}) \frac{1}{1 - p_{i^*}} \sum_{i\backslash i^*} p_i \\ &= \sum_{i\backslash i^*} \frac{p_i}{1 - p_{i^*}} \ln(1 - p_{i^*}) \end{aligned} \qquad (34)$$

From Eqs. (12) and (34), one can thus write

$$\begin{aligned} E[D(i||i^*)] &= E[D(i||i^*)] + \ln(1 - p_{i^*}) - \ln(1 - p_{i^*}) \\ &= \ln\left(\frac{p_{i^*}}{1 - p_{i^*}}\right) - \sum_{i\backslash i^*} \frac{p_i}{1 - p_{i^*}} \ln\left(\frac{p_i}{1 - p_{i^*}}\right) \\ &= \ln\left(\frac{p_{i^*}}{1 - p_{i^*}}\right) + H(\mathbf{p}\backslash p_{i^*}) \end{aligned} \qquad (35)$$

where p_{i^*} is given and where $H(\mathbf{p}\backslash p_{i^*}) \geq 0$ is the entropy of the subset of p_i's when $i \neq i^*$. From the properties of the entropy, the minimum of $H(\mathbf{p}\backslash p_{i^*}) = 0$ is reached if and only if there is a single $p_i/(1 - p_{i^*}) = 1$ (i.e. the other probabilities are null) and, accordingly under this condition,

$$E[D(i||i^*)] = \ln\left(\frac{p_{i^*}}{1 - p_{i^*}}\right) = \ln p_{i^*} - \ln(1 - p_{i^*}) \qquad (36)$$

is the minimum possible value, as stated. □

Appendix 4: Proof of the fourth theorem

Starting again from Eq. (35), the entropy $H(\mathbf{p}\backslash p_{i^*})$ reaches its maximum possible value if and only if all probabilities are equal over the $k - 1$ categories, i.e.

$$\frac{p_i}{1 - p_{i^*}} = \frac{1}{k - 1} \quad \forall i \neq i^* \qquad (37)$$

and so it comes that

$$H(\mathbf{p}\backslash p_{i^*}) = \ln(k - 1) \qquad (38)$$

so that the maximum possible value for $E[D(i||i^*)]$ is

$$\begin{aligned} E[D(i||i^*)] &= \ln\left(\frac{p_{i^*}}{1 - p_{i^*}}\right) + H(\mathbf{p}\backslash p_{i^*}) \\ &= \ln\left(\frac{p_{i^*}}{1 - p_{i^*}}\right) + \ln(k - 1) \\ &= \ln p_{i^*} - \ln\left(\frac{1 - p_{i^*}}{k - 1}\right) \end{aligned} \qquad (39)$$

as stated by the theorem. □

References

Atkinson P, Foody G (2002) Uncertainty in remote sensing and GIS. Wiley, Chichester, pp 1–18

Brown K, Foody G, Atkinson P (2009) Estimating per-pixel thematic uncertainty in remote sensing classifications. Int J Remote Sens 30(1):209–229

Cockx K, Van de Voorde T, Canters F (2014) Quantifying uncertainty in remote sensing-based urban land-use mapping. Int J Appl Earth Obs Geoinf 31:154–166

Comber A, Fisher P, Brunsdon C, Khmag A (2012) Spatial analysis of remote sensing image classification accuracy. Remote Sens Environ 127:237–246

Cripps E, OHagan A, Quaife T (2013) Quantifying uncertainty in remotely sensed land cover maps. Stoch Environ Res Risk Assess 27(5):1239–1251

Crosetto M, Tarantola S (2001) Uncertainty and sensitivity analysis: tools for GIS-based model implementation. Int J Geogr Inf Sci 15(5):415–437

Dehghan H, Ghassemian H (2006) Measurement of uncertainty by the entropy: application to the classification of MSS data. Int J Remote Sens 27(18):4005–4014

Dronova I, Gong P, Wang L (2011) Object-based analysis and change detection of major wetland cover types and their classification uncertainty during the low water period at Poyang Lake, China. Remote Sens Environ 115(12):3220–3236

Eastman JR (2006) Idrisi andes. Guide to GIS and image processing. Clark University, Worcester, pp 87–131

Feng Y, Liu Y, Batty M (2015) Modeling urban growth with GIS based cellular automata and least squares SVM rules: a case study in Qingpu–Songjiang area of Shanghai, China. Stoch Environ Res Risk Assess 30:1–14

Foody G (2005) Local characterization of thematic classification accuracy through spatially constrained confusion matrices. Int J Remote Sens 26(6):1217–1228

Foody GM (2008) RVM-based multi-class classification of remotely sensed data. Int J Remote Sens 29(6):1817–1823

Foody GM, Campbell N, Trodd N, Wood T (1992) Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification. Photogr Eng Remote Sens 58(9):1335–1341

Ge Y, Li S, Lakhan VC, Lucieer A (2009) Exploring uncertainty in remotely sensed data with parallel coordinate plots. Int J Appl Earth Obs Geoinf 11(6):413–422

Giacco F, Thiel C, Pugliese L, Scarpetta S, Marinaro M (2010) Uncertainty analysis for the classification of multispectral satellite images using SVMs and SOMs. IEEE Trans Geosci Remote Sens 48(10):3769–3779

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recognit Lett 27(4):294–300

Glasziou P, Hilden J (1989) Test selection measures. Med Decis Mak 9(2):133–141

Gonçalves LM, Fonte CC, Júlio EN, Caetano M (2009) A method to incorporate uncertainty in the classification of remote sensing images. Int J Remote Sens 30(20):5489–5503

Hagolle O, Dedieu G, Mougenot B, Debaecker V, Duchemin B, Meygret A (2008) Correction of aerosol effects on multi-temporal images acquired with constant viewing angles: application to formosat-2 images. Remote Sens Environ 112(4):1689–1701

Hagolle O, Huc M, Villa Pascual D, Dedieu G (2015) A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of formosat-2, landsat, venμs and sentinel-2 images. Remote Sens 7(3):2668–2691

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86

Liu R, Chen Y, Wu J, Gao L, Barrett D, Xu T, Li L, Huang C, Yu J (2015) Assessing spatial likelihood of flooding hazard using Naive Bayes and GIS: a case study in Bowen Basin, Australia. Stoch Environ Res Risk Assess 30:1–16

Liu W, Gopal S, Woodcock CE (2004) Uncertainty and confidence in land cover classification using a hybrid classifier approach. Photogr Eng Remote Sens 70(8):963–971

Loosvelt L, Peters J, Skriver H, De Baets B, Verhoest NE (2012a) Impact of reducing polarimetric sar input on the uncertainty of crop classifications based on the random forests algorithm. IEEE Trans Geosci Remote Sens 50(10):4185–4200

Loosvelt L, Peters J, Skriver H, Lievens H, Van Coillie FM, De Baets B, Verhoest NE (2012b) Random forests as a tool for estimating uncertainty at pixel-level in sar image classification. Int J Appl Earth Obs Geoinf 19:173–184

Lunetta RS, Congalton RG, Fenstermaker L, Jense J, McGwire K, Tinney L (1991) Remote sensing and geographic information system data integration: error sources and reseach issues. Photogr Eng Remote Sens 57(6):677–687

Löw F, Conrad C, Michel U (2015a) Decision fusion and non-parametric classifiers for land use mapping using multi-temporal rapideye data. ISPRS J Photogr Remote Sens 108:191–204

Löw F, Knöfel P, Conrad C (2015b) Analysis of uncertainty in multi-temporal object-based classification. ISPRS J Photogr Remote Sens 105:91–106

Löw F, Michel U, Dech S, Conrad C (2013) Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines. ISPRS J Photogr Remote Sens 85:102–119

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 Photogr Remote Sens 49(2):13–20

McIver DK, Friedl M et al (2001) Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods. IEEE Trans Geosci Remote Sens 39(9):1959–1968

Mitchell SW, Remmel TK, Csillag F, Wulder MA (2008) Distance to second cluster as a measure of classification confidence. Remote Sens Environ 112(5):2615–2626

Mitchell JJ, Shrestha R, Moore-Ellison CA, Glenn NF (2013) Single and multi-date landsat classifications of basalt to support soil survey efforts. Remote Sens 5(10):4857–4876

Pal NR, Bezdek JC (1994) Measuring fuzzy uncertainty. IEEE Trans Fuzzy Syst 2(2):107–118

Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45

Pontius RG (2000) Quantification error versus location error in comparison of categorical maps. Photogr Eng Remote Sens 66(8):1011–1016

Radoux J, Chomé G, Jacques DC, Waldner F, Bellemans N, Matton N, Lamarche C, dAndrimont R, Defourny P (2016) Sentinel-2s potential for sub-pixel landscape feature detection. Remote Sens 8(6):488

Renier C, Waldner F, Jacques DC, Babah Ebbe MA, Cressman K, Defourny P (2015) A dynamic vegetation senescence indicator for near-real-time desert locust habitat monitoring with MODIS. Remote Sens 7(6):7545–7570

Rodriguez-Galiano V, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez J (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogr Remote Sens 67:93–104

Stehlí KM, Sivasundaram S (2012) Decompositions of information divergences: recent development, open problems and applications. In: AIP conference proceedings, vol 1493. American Institute of Physics, p 972

Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ 62(1):77–89

Story M, Congalton RG (1986) Accuracy assessment-a user\'s perspective. Photogr Eng Remote Sens 52(3):397–399

Strahler AH, Boschetti L, Foody GM, Friedl MA, Hansen MC, Herold M, Mayaux P, Morisette JT, Stehman SV, Woodcock CE (2006) Global land cover validation: recommendations for evaluation and accuracy assessment of global land cover maps. European Communities, Luxembourg 51

Van der Wel FJ, Van der Gaag LC, Gorte BG (1998) Visual exploration of uncertainty in remote-sensing classification. Comput Geosci 24(4):335–343

Waldner F, Canto GS, Defourny P (2015a) Automated annual cropland mapping using knowledge-based temporal features. ISPRS J Photogr Remote Sens 110:1–13

Waldner F, Lambert MJ, Li W, Weiss M, Demarez V, Morin D, Marais-Sicre C, Hagolle O, Baret F, Defourny P (2015c) Land cover and crop type classification along the season based on biophysical variables retrieved from multi-sensor high-resolution time series. Remote Sens 7(8):10400–10424

Waldner F, Ebbe MAB, Cressman K, Defourny P (2015) Operational monitoring of the desert locust habitat with earth observation: an assessment. ISPRS Int J GeoInf 4(4):2379. doi:10.3390/ijgi4042379. http://www.mdpi.com/2220-9964/4/4/2379

Zhang J, Sun J (2002) The survey of accuracy analysis of remote sensing and GIS. Int Arch Photogr Remote Sens Spat Inf Sci 34(2):581–584