Probabilistic unsupervised classification for large-scale analysis of spectral imaging data

Emmanuel Paradis

December 13, 2021

Probabilistic unsupervised classification for large-scale analysis of spectral imaging data

Emmanuel Paradis

ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

Email: Emmanuel.Paradis@ird.fr

Phone: +33 4 67 14 36 26

ORCID: 0000-0003-3092-2199

ABSTRACT

Land cover classification of remote sensing data is a fundamental tool to study changes in the environment such as deforestation or wildfires. A current challenge is to quantify land cover changes with real-time, large-scale data from modern hyper- or multispectral sensors. A range of methods are available for this task, several of them being based on the k-means classification method which is efficient when classes of land cover are well separated. Here a new algorithm, called probabilistic k-means, is presented to solve some of the limitations of the standard k-means. It is shown that the new algorithm performs better than the standard k-means when the data are noisy. If the number of land cover classes is unknown, an entropy-based criterion can be used to select the best number of classes. The proposed new algorithm is implemented in a combination of R and C computer codes which is particularly efficient with large data sets: a whole image with more than 3 million pixels and covering more than 10,000 km^2 can be analysed in a few minutes. Four applications with hyperspectral and multispectral data are presented. For the data sets with ground truth data, the overall accuracy of the probabilistic k-means was substantially improved compared to the standard k-means. One of these data sets includes more than 120 million pixels, demonstrating the scalability of the proposed approach. These developments open new perspectives for the large scale analysis of remote sensing data. All computer code are available in an open-source package called sentinel.

Keywords: Unsupervised classification; k-means; land cover; multivariate normal density; spectral imaging data

1 Introduction

Monitoring environmental changes has become critical for many issues related to sustainable development. Deforestation, wildfires, and other land use changes have profound impacts on human activities, our environment, and biodiversity (Pettorelli et al., 2005; Newbold et al., 2016; Betts et al., 2017). For instance, there is some evidence suggesting that pathogen outbreaks are linked to changes in land cover and particularly deforestation (Jones et al., 2013; Morand et al., 2019).

There are two main approaches to track on-going environmental changes: either by monitoring and measuring land uses directly in the field, or by remote sensing with satellites, aircrafts, or other airborne devices (e.g., unmanned aerial vehicles). Although the second approach has some limitations compared to the first one, it has some definite advantages that cannot be matched by field data. In particular, satellites can cover the whole surface of the Earth with a frequency of a few days or weeks (Li and Roy, 2017; Wulder et al., 2018). Furthermore, the most recent satellites are equipped with high-resolution sensors which are able to record a wide range of information such as reflectance at different wavelengths, altitude, or temperature (Fu et al., 2020). During the last decade, there has been a remarkable increase in the resolution of these sensors. To illustrate this progress, the University of Twente maintains a database listing 334 satellites (some being out of service) and 396 sensors with a number of bands ranging between 1 and 16,921 (`https://webapps.itc.utwente.nl/sensor/`; accessed 2021-08-31). Among these sensors, 43 (11%) are indicated to have a resolution of one meter or less (until 1.25 cm), and 90 (23%) others are listed with a resolution between 1 m and 10 m.

Spectral imaging sensors record electromagnetic waves and provide data in two broad categories: hyperspectral imaging (HSI) where reflectance is recorded for several hundreds of narrow bands (typically a few nanometres wide), and multispectral imaging (MSI) where reflectance is recorded for a few bands (usually less than 20) each with a width of few tens or hundreds nanometres. Both HSI and MSI usually

record wavelengths beyond visible light (e.g., ultraviolet, infrared). During the last decade, a range of open-access remote sensing data have been made available (e.g., `https://developers.google.com/earth-engine/`; see also Guo et al., 2020). As an example of these developments, the *Sentinel* program is made of seven satellites currently in orbit around the Earth (`https://sentinel.esa.int/`). Two of them, *Sentinel*-2A and *Sentinel*-2B, are equipped with an MSI sensor which records reflectance in thirteen bands from ultraviolet (UV) to infrared (IR) including three bands in visible light (Gascon et al., 2017). The *Sentinel* program stands apart from other similar programs because the data are available publicly in near-real time through the Copernicus datahub (`https://scihub.copernicus.eu/`). Each satellite covers the same location every two weeks, so the same location is potentially covered every week giving the opportunity to monitor environmental and land use changes at relatively high temporal resolution (Li and Roy, 2017).

One of the applications of spectral imaging data is to infer land cover and land use. Two types of approaches are used for this objective. In supervised classification methods, there is a reference sample with known land cover which is used to "train" the classification procedure in a first step, and the sample with unknown land cover is then classified in a second step. In unsupervised classification, there is no reference sample: classes or groups are defined following different criteria (see Wulder et al., 2018, for a recent review). As discussed below, both approaches have their respective advantages. For unsupervised classification, the k-means algorithm has been widely used in various contexts (see next section).

The objective of the present paper is to present a new method, called the probabilistic k-means, to analyse large-scale, spectral imaging data. The most important original feature of this method is to take into account variance heterogeneity among groups. Furthermore, a specific aim was to perform analysis of images with several millions of pixels in reasonable times. For instance, an image (or product) of *Sentinel*-2 (about 10,000 km^2) at a 10-m resolution has more than 120 million pixels.

The method is available in a computer package called sentinel (which also includes functions to query, manage, and download data from the Copernicus datahub). Before detailing the proposed methodological development, the next section presents a review of the recent literature on the applications of the k-means method to the analysis of remote sensing data. Four applications are then presented with two HSI data sets and two MSI data sets. The discussion gives further comparisons with previous contributions on unsupervised classification. The perspectives of current and future developments are also discussed.

2 Literature review

This review focuses on developments and applications of the k-means method in remote sensing data analysis published during the last ten years. Where possible, the sizes of the imaging data and the software used have been noted.

Several papers attempted to develop methods aimed to improve the properties of the k-means method. Galluccio et al. (2012) developed a method which assumes there are modes (areas of highest densities of observations) in the distribution of reflectance. These modes are found in the multivariate density space using the link lengths of a minimum spanning tree. Basically, the goal of their method is to initialise the centres of the k-means algorithm. They applied it to image data from Paris (512×521 pixels, 7 bands) and from Mars (300×120 pixels, 256 bands). Another study found that the standard k-means algorithm usually performs poorly on HSI data (Zhang et al., 2013). These last authors define the pure neighbourhood index (PNI) to perform neighbourhood-constrained k-means which adds steps to the iterations of the standard k-means with a weight function defined with the PNI. They applied this method to a 200×200 pixels image with 80 bands. Haut et al. (2017) used the MapReduce computational framework to analyse two images from Indian Pines (145×145 pixels and 2678×614 pixels, both with 220 bands).

They programmed their analyses with Apache Spark for distributed computing and Python Scikit for the k-means. However, they did not assess the effect of different numbers of groups.

He et al. (2014) showed that support vector machine (SVM), a supervised classification method, performs very well even with a small training data set. On the other hand, fuzzy k-means (FKM) was found to have a reduced usefulness with large data sets. These authors proposed a fusion of the two methods where the entropy is used to find the appropriate number of groups (see below for details about the use of entropy). They applied their method on two SPOT6 images (1982×1630 pixels and 2113×2151 pixels) each with six reference classes. Their analyses were implemented in ENVI and IDL (ver. 4.8).

Zhang et al. (2016) used an object-based approach defining a hierarchy from the pixels up to the scene. Their analyses used a combination of principal component analysis (PCA) on HSI images, k-means with drop-out, and SVM. The code was implemented in LIBSVM. They applied their approach to the Indian Pines data (145×145 pixels, 220 bands) and the University of Pavia data (610×340 pixels, 103 bands). They concluded that the drop-out k-means improves efficiency of the standard k-means with a small computational burden. They also demonstrated that the spatial information contained in the neighbourhood of pixels is useful, although their results did not relate this improvement with the identification of physical objects on the ground. Similarly, in another study Kavzoglu and Tonbul (2018) used k-means to perform image segmentation in a framework of object-based image analysis. They applied their approach to an image with 5000×3700 pixels and 8 bands. They found that k-means generally performs well for image segmentation using different specific algorithms. They implemented their computations with ENVI and MATLAB.

Image matching and indexing are also applications of k-means. Cao et al. (2013) used k-means to perform image indexing based on the Kullback–Leibler discrepancy. They provided code in C++ and Matlab. Sedaghat and Ebadi (2015) performed

image matching using k-means in a second step to classify images into groups. They used MATLAB to implement their method.

Several papers used k-means to perform fine-scale spatial structure analyses. Kuo et al. (2019) analysed canopy structure by quantifying leaf angle distribution using a combination of k-means and an octree data structure: they analysed point cloud data (PCD), a kind of LiDAR (light detection and ranging) data which can reconstruct 3-D structures. The PCD were first split into octree subspaces so that each single octree unit contained no more than 1500 points. Each octree unit was then analysed with a standard k-means. Direct observations led these authors to infer that a leaf used between 500 and 1500 points, which helped them find the number of groups in the k-means analyses. Reza et al. (2019) used graph-cut and k-means to identify rice grains and estimate their sizes: they first applied k-means on the red-green-blue (RGB) image data after converting them to the Lab colour space, and then used a graph-cut algorithm to identify the rice grains. The best value of number of groups in the k-means analyses was found with the histogram method (Kanthana and Sujathab, 2013). The analysed images had 600×400 pixels. Wang et al. (2019) used k-means for image segmentation to identify roads from satellite images: the image data were converted from the RGB space into the the Lab colour space and then analysed with k-means fixing the number of groups to three (no information on image size was given).

Some authors used k-means to quantify temporal changes from several images. Kesikoglu et al. (2013) combined PCA with a fuzzy version of k-means called c-means to analyse temporal changes from image differencing, so there were effectively only two groups in their c-means analyses. Lv et al. (2019) used k-means with adaptive majority voting (AMV) to quantify change magnitude image (CMI). Their method starts from a "central" pixel, and builds a region around it. In a second step, a k-means analysis is done in the region with two groups (changed *vs.* unchanged pixels). In a third step, the region is extended with the AMV algorithm. They

analysed four images ranging in size from 412×300 to 950×1250 pixels.

Overall, k-means is a widely used method in image and remote sensing data analysis; it is often used in combination with other data analysis methods (e.g., PCA). A remarkable diversity of approaches have been developed during the past decade most of them with different objectives. The sizes of the data are generally moderate, and very little open-source software has been contributed by these studies.

3 Methods

3.1 Data

Remote sensing data are usually arranged in a rectangular raster with variables associated with each pixel of the raster. These variables may be univariate (i.e., a single value is associated to each pixel) or multivariate. In this paper, we consider a multivariate setting where these variables are the values of reflectance measured in different wavelengths (the bands). In the present study we do not consider the spatial arrangement of the pixels in their respective rasters, so that the pixels are assumed to be independent. Therefore, the data under consideration below are denoted as X with the values of reflectance arranged in a matrix with n rows and p columns, where n is the number of pixels of the raster (i.e., the product of the number of rows by the number of columns of the raster), and p is the number of bands of the image. Measures of reflectance are usually more or less noisy (Chavez, 1988; Zhang et al., 2018). The exact values measured by the sensor depend on land cover and also on several factors such as the satellite or aircraft position, the time of the day, the atmospheric conditions, and so on.

3.2 Probabilistic k-means

The k-means method is a widely used, unsupervised classification procedure (Hastie et al., 2009). It requires specification of the number of groups (or clusters), denoted

as K here, then the algorithm proceeds by assigning observations to a group depending on the distance to the group means (Lloyd, 1982). If the values of these means are unknown (which is the most common case), some initial values are chosen randomly, the observations are assigned as explained above, the group centres are recalculated, and the whole procedure is repeated until group assignments are stable. The method works with multivariate data using a multivariate distance such as the Euclidean distance.

Standard k-means algorithms work well when within-group variances are homogeneous so that group assignments using distances are likely to be valid. However, when variances are heterogeneous, this is likely to result in misclassification of observations. Figure 1 shows a small simulated example with two groups each with 200 observations drawn randomly from two normal distributions with means 0 and 6, and standard-deviations (SD) 2 and 0.1, both respectively for each group. Even though the two means are very different, the large SD of the first group is likely to result in mixing of observations from both groups, and thus a k-means-based classification may be in error for these observations. The standard k-means indeed resulted in 15 misclassified observations in this case.

A solution to this problem is to rely on a probabilistic approach when classifying observations in the different groups. In the above simple simulated case, it is straightforward to apply this approach: after running a standard k-means classification, the means and SDs of both groups are calculated, then the probability densities are calculated for all observations using parameters of both groups: each observation is reclassified to the group for which it has the highest density. This can be represented graphically with a classification limit where the inferred density curves intersect (Fig. 1B). Note, on the other hand, that the limit for the standard k-means is defined by the equidistant point between the two group means. The reclassification procedure can be repeated until the overall classification is stable. In this simple case, a single iteration is enough and results finally in four observations

misclassified.

This approach can be generalised to multivariate data using the densities of multivariate distributions. However, this requires estimation of a number of parameters which is likely to grow substantially with the number of variables. For instance, a multivariate normal distribution with p variables has $2p + p(p-1)/2$ parameters: p means, p SDs, and $p(p-1)/2$ covariances. Therefore, the number of parameters is proportional to p^2. A way to avoid having to estimate too many parameters when p increases is to first perform a PCA on the matrix X. PCA is usually used to perform dimension reduction in order to obtain a number of variables smaller than p that maximise the overall variation in X. Another property of PCA is that these principal components (PCs) are orthogonal: in other words, the coordinates of the observations (here the pixels of the image) on these PCs have zero covariances. We denote the matrix of these PC-based coordinates as Z. From a geometrical point of view, a PCA resulting in p PCs is a global rotation of the axes defined by the original p variables with the constraint that the covariances of the PCs are equal to zero. Therefore, this considerably simplifies the calculations of multivariate normal densities since it is now needed to estimate only $2p$ parameters (p means and p SDs) for each group of the classification.

Another crucial difference with PCA as commonly used in data analysis is that it is important here to not scale the original variables (i.e., divide them by their respective SD) before performing the PCA. If one of the variables has a large variance compared to the others, then it will contribute overwhelmingly to the PCA and will pull the overall variation in the data compared to the patterns from the covariances. This is the reason why variable scaling is usually recommended before running a PCA (e.g., Venables and Ripley, 2002). However, the present goal is to discriminate groups with the calculated PCs where the overall variance is actually the consequence of the existence of these groups. So, in order to not erase this overall variance, the variables should not be scaled.

A word should be said about the choice of the form of the density distribution. The present work assumes that the rows of Z (not X) follow a multivariate normal distribution. Furthermore, it is assumed that this distribution is non-homogeneous: its parameters (means and SDs) vary among the K groups and are assumed to be homogeneous within each group. In practice these assumptions may not be valid and other distributions may reflect more accurately the distribution of reflectance within each group. However, at the moment there is no theoretical or empirical justification for one distribution rather than another. Furthermore, the crucial point here is to assess variation among groups of different land cover and the normal distribution with its two parameters (mean and SD) may be flexible enough to accommodate such variation.

3.3 Selecting the number of groups

3.3.1 Likelihood-based information criteria

With unsupervised classification, there are two possible situations: the number of groups may be known *a priori*, or this number must be inferred from the data. In the second situation, a parametric, probabilistic approach makes possible to use standard statistical tools such as Akaike's information criterion (AIC, Akaike, 1973) which requires to compute the likelihood of the data. We must take care that group assignment is uncertain and has to be considered explicitly when calculating the likelihood function. Thus, we have to calculate the probability for the ith row of Z (z_i) using the parameters (means and SDs) estimated for group j multiplied by the probability that pixel i belongs to group j. These products are then summed over all K groups for each pixel. Finally, the log-likelihood is the sum of the log-transformed probabilities over all n pixels:

$$\mathcal{L} = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{K} \hat{f}_j \times \xi(z_i | \hat{\mu}_j, \hat{\sigma}_j) \right], \tag{1}$$

where ξ if the multivariate normal density function, $\hat{f}_j$ if the estimated proportion of pixels in group j, and $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated parameters for group j. There are thus $2pK + K - 1$ parameters estimated from the data: mean and SD for each column of Z and for each group, and $K - 1$ proportions (since $\sum_{j=1}^{K} f_j = 1$). We may now calculate the AIC:

$$\text{AIC} = -2\mathcal{L} + 2(2pK + K - 1). \tag{2}$$

The value of K resulting in the smallest value of AIC is to be preferred. Another criterion which can be used is the Bayesian information criterion (BIC) defined by (Schwarz, 1978):

$$\text{BIC} = -2\mathcal{L} + (2pK + K - 1) \times \ln n. \tag{3}$$

A simulation study is presented in the Supplementary Information which shows that both criteria are not robust to non-normality of the data. In particular, if the observations follow a uniform distribution and there is no heterogeneity (i.e., all observations are generated from the same distribution, thus $K = 1$), then both AIC and BIC will select a model with $K > 1$. Furthermore, in this situation the values of AIC and of BIC tend to decrease continuously when K is increased (see Supplementary Information for details).

3.3.2 Informational entropy

Another procedure for selecting the value of K is based on the principle of entropy (Burrough et al., 2000). This approach can be applied if there is a measure of uncertainty in the assignment of observations to the groups: in that case each observation is given a value of membership to each group with the constraint:

$$\sum_{j=1}^{K} m_{ij} = 1, \tag{4}$$

where m_{ij} is the membership value of observation i for group j. The entropy, H, for a given value of K, is then calculated with:

$$H = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} m_{ij} \times \ln m_{ij}. \tag{5}$$

The value of K resulting in the smallest value of H gives the best description of the data. Membership values have been defined in the context of FKM (Burrough et al., 2000; He et al., 2014), but they can be adapted in a straightforward way to the probabilistic k-means developed here using the probability densities. Furthermore, the computation of the multivariate normal densities on a log-scale (see next section) makes possible to calculate H even when densities can reach very low values (see Supplementary Information).

3.4 Computational details and implementation

The overall workflow is summarised on Figure 2. The whole procedure was implemented in code written in the R and in C computer languages. The PCA was performed by singular value decomposition (SVD) which is faster and numerically more stable than the usual eigendecomposition (Venables and Ripley, 2002). With HSI data, it was observed that a relatively substantial number of PCs had nearly zero variance so that keeping all PCs made the computations much slower for no benefit: the number of PCs selected was set to keep at least 99% of the overall variance. For MSI data, all PCs are kept. The coordinates on the p PCs are first analysed with a standard k-means using Hartigan and Wong's (1979) algorithm which is particularly efficient and fast. The means and SDs are calculated for the p PCs and each group. The multivariate normal densities are calculated on a logarithmic scale which avoids numerical underflows and considerably simplifies the calculations (the overall densities are calculated with sums instead of products if full densities were used). Furthermore, the mathematical expression is factorised to avoid repeating

redundant computations (e.g., the terms $-\ln(\sqrt{2\pi}\sigma_j)$ were computed once for all observations). These factorisations result in running times around 2.5 times faster than using the internal log-density function. Finally, the densities are evaluated separately for each pixel and only its classification is stored, avoiding to store all densities which would require an array of npK real values (amounting to 4.1 GB of memory with $n = 3.3 \times 10^6$, $p = 13$, and $K = 12$). Furthermore, this makes the overall memory requirement independent of the value of K. The running times are predicted to be proportional to n, p, and K (i.e., $\mathcal{O}(npK)$). It was evaluated that a single iteration of the algorithm takes $\approx \frac{K}{5}$ sec on a standard laptop with $n = 3{,}348{,}900$ and $p = 13$. On the other hand, the number of iterations required to reach convergence depends on the data: analyses of data sets with strong structure converge quickly (typically less than 10 iterations with $K = 2$), whereas if there is no structure convergence takes longer to reach.

The probabilistic reclassification is iterated until convergence. Furthermore, two stopping criteria have been defined: the maximum number of iterations can be fixed by the user (e.g., 200); or the procedure can be stopped when less than a fixed proportion of pixels are reclassified (e.g., if this proportion is zero, then iterations are stopped when no pixel is reclassified). This probabilistic k-means has been coded in a C routine called from R.

All code is available in a package named sentinel distributed on GitHub (`https://github.com/emmanuelparadis/sentinel`). Some code is also provided to use the standard k-means method in order to ease comparisons with the present method. This package includes code to query the SciHub repository where the *Sentinel* data are stored.

3.5 Applications

Five data sets were analysed (Table 1). They are described in details below.

3.5.1 Hyperspectral Data: Pavia University and Okavango

Two hyperspectral data sets were considered. The Pavia University and Okavango data are two HSI data sets that have been preprocessed (`http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`; accessed 2021-07-07). Both data sets are associated with reference data defined as "ground truth" with 9 and 14 classes of land cover, respectively (Tables S1–S2). Two analyses were performed with both data sets. First, the ground truth data only were analysed with the standard and the probabilistic k-means with K set equal to the known number of classes of land cover for each data set. The classification performance of each method was quantified with the overall accuracy as defined by Olofsson et al. (2014). Because both k-means algorithms are unsupervised, the reference and inferred land cover values were matched with the Hungarian algorithm, a method which aims to maximise the values on the diagonal of a matrix, as implemented in the package RcppHungarian (Silverman, 2019); the diagonal values of the matrix output were used to calculate the accuracy. Second, the complete data set was analysed with the probabilistic k-means using increasing values of K: the value of H was computed for each value of K and the final maps were drawn for both standard and probabilistic k-means using the value of K giving the smallest value of H.

3.5.2 Southern France

An image data taken by the satellite SPOT6 was analysed. The image was taken on 2019-06-27 above the South of France and had no cloud cover. The area of the image was estimated to be 3207 km^2. A preliminary analysis of the CORINE land cover database over this area found that it is covered by 34 distinct land classes (as defined by the CORINE database). Out of these 34 classes, 17 were represented by less than 0.5% of the area, whereas 14 classes were represented by at least 1% of the area (Tables S3). The data were analysed with the probabilistic k-means with increasing values of K: the value of H was computed for each value of K and the

final maps were drawn for both standard and probabilistic k-means using the value of K giving the smallest value of H.

3.5.3 Eastern Thailand

One area was selected in Thailand extending from N 14°27′56″ to N 13°27′49″, and from E 100°51′19″ to E 101°51′39″. A single *Sentinel*-2 image taken on 2021-02-05 was selected with 0% cloud cover. The whole product ($109.8 \times 109.8\,\text{km} = 12{,}056.04\,\text{km}^2$; Table 1) was analysed with the same procedure than for the Southern France data. With *Sentinel*-2 data, four bands are available at a resolution of 10 m, six at 20 m, and three at 60 m. Two data sets were built from this image: using the highest resolution bands (10 m, 4 bands) and using all bands aggregating the highest resolution bands at 60 m (13 bands). Similarly to the Southern France data, there was no ground truth data for this data set. An analysis of land cover data from the European Spatial Agency Climate Change Initiative (ESA/CCI; `http://maps.elie.ucl.ac.be/CCI/viewer/index.php`; accessed 2019-11-27) for the period 2016–2018 identified twelve main land cover classes (Tables S4).

4 Results

4.1 Pavia University

The overall accuracies were 0.55, and 0.62 for the standard and probabilistic k-means, respectively. The maps drawn with the ground truth data only show that some areas are not correctly identified with both methods (Fig. 3). However, some areas look more homogeneous with the probabilistic than with the standard k-means which is consistent with with the better overall accuracy of the former.

The smallest value of H was observed for $K = 2$, and then the value of H increased when the value of K increased as well (Fig. 4). However, two local minima of H were observed for $K = 6$ and $K = 9$. Figure 5 shows the overall maps inferred

from the k-means methods assuming $K = 9$.

4.2 Okavango

The overall accuracies were 0.62, and 0.79 for the standard and probabilistic k-means, respectively. The ground truth data were more dispersed than for the previous data so that the improvement accuracy of the probabilistic k-means was not so clearly visible (Fig. 6).

A pattern similar to the previous data was observed with respect to the relation between H and K: the smallest value of H was observed for $K = 2$, and then the former increased when the latter increased (Fig. 4). Local minima were observed for $K = 6$, $K = 9$, and $K = 11$. Figure 7 shows the overall maps inferred from the k-means methods assuming $K = 11$.

4.3 Southern France

The value of entropy was very low for $K = 2$ ($H = 0.004$) meaning that classification between the two groups was practically almost perfect (Fig. 8). However, a very substantial portion of the study area was covered by water which could readily explain this result. Indeed, as for the previous data sets, the value of H increased when K increased. However, a local minimum was observed for $K = 15$. The maps inferred from both k-means methods assuming $K = 15$ show some interesting differences (Fig. 9). Particularly, the coastal lagoons which were found to be covered with different classes by the standard k-means were all grouped in the same class by the probabilistic k-means (Fig. 9).

4.4 Eastern Thailand

The values of H varied with respect to K in the same way than for the previous data sets: the smallest value was observed for $K = 2$ and local minima were observed for

$K = 12$ at the finest resolution (10 m) and for $K = 15$ at the coarsest resolution (60 m; Fig. 10). These two values of K were selected to infer the maps at their respective resolutions (Figs. 11–12). Overall, the maps inferred with probabilistic k-means show better delimitation of the fields compared to the results obtained with the standard k-means, particularly for the paddy fields on the west of the study area.

4.5 Computational efficiency

With 3,348,900 pixels and 13 variables, each iteration with $K = 12$ took around 2.4 sec. Therefore, 200 iterations (the default limit set in the code of sentinel) took 8 min. Furthermore, it was observed that in all cases, with either real or simulated data, there was convergence to a stable classification with no further reassignment. In all cases reassignment was around 0.001% of the pixels after 200 iterations.

5 Discussion

The present work has contributed a new k-means method which appears as an improvement compared to currently available implementations with respect to three points: better accuracy, possibility to identify the number of groups, and ability to handle and analyse very large data sets. Each point is discussed below.

With both HSI data sets for which ground truth data were available, the method proposed here showed better overall accuracy compared to the standard k-means. The improvement was particularly substantial for the Okavango data set. Although no ground data were available for the two MSI data sets analysed here, the maps suggest improved classification with the probabilistic method compared to the standard k-means. These results clearly suggest that the proposed probabilistic method has improved performance compared to the standard k-means for land cover classification of spectral imaging data. The fact that the assumption of homogeneous

variance is relaxed in this method is certainly an important factor to explain this improvement (see further below).

The present results emphasise the importance of selecting the number of groups, K. The above review suggests that this issue did not receive a lot of attention in the recent literature. Although the information criteria presented above seem good candidates to select the best value of K in a probabilistic framework, this was not conclusive (see Supplementary Information). Clearly, the lack of statistical robustness shown by this approach is problematic and needs to be investigated further. On the other hand, the entropy-base criterion, H, proposed by Burrough et al. (2000, and previous references therein) appears a good alternative. However, some care must be taken when using it. It was observed that the smallest value of H was always obtained with $K = 2$ groups. This could make sense considering that spectral imaging data often show a strong discrimination between two broad classes of land cover (e.g., land *vs.* water, urban *vs.* vegetation), so that it is expected that classification with $K = 2$ yields essentially very good results so that all values of membership, m_{ij}, are either zero or one. On the other hand, in all applications the values of H showed a local minimum for more realistic values of K. Therefore, it is suggested here that the entropy-based criterion is useful provided it is used within a range of realistic values of K (i.e., avoiding too small values).

The main feature of the approach adopted in this paper is to relax the assumption of homogeneous variance which underlies the standard k-means algorithm. The assumption of homogeneous variance is an important feature of the ISODATA method (Ball and Hall, 1965) which is derived from the standard k-means. Memarsadeghi et al. (2007) when implementing the ISODATA assumed that 'the clusters are well-separated, that is, the probability that a point belonging to one cluster is closer to the centre of another cluster than to its own cluster centre is negligible.' As illustrated above, if the variances are homogeneous this is likely to result in misclassifications. Interestingly, Memarsadeghi et al. (2007) made no parametric assumption on the

distribution of the data within groups (or clusters). Indeed, if the groups are well separated and their variances small enough, there is no need to make any such assumption and the standard k-means algorithms are very likely to perform very well. The DBSCAN method (Ester et al., 1996; Li et al., 2019) is another unsupervised clustering method related to k-means which takes into account the spatial configuration of the data as well as noise. However, the DBSCAN, although closely related to probabilistic k-means, is more complex and current implementations have been explored only with limited data sizes, typically with a few ten thousand observations (Hahsler et al., 2019).

There has been substantial research on applying the k-means method to the analysis of remote sensing data (e.g., Lv et al., 2010; Pascucci et al., 2018, and the above review of the recent literature). Besides these applications to remote sensing, an approach has recently been developed to take noisy data into account in the context of quantum computing (Kerenidis et al., 2019; Khan et al., 2019): these proposals can be compared to the method proposed here in the sense that they aim to deal with overlapping clusters; however, they treat this issue quite differently. Ma et al. (2016) proposed an elaborate method named spectral clustering which seems to outperform other classification methods. However, spectral clustering appears to be a computationally costly method and seems unfeasible even with a few ten thousands pixels (Pascucci et al., 2018). On the other hand, the method proposed in this paper is economical in terms of computations as it only requires evaluation of densities for each pixel and each group. For instance, Rodriguez and Laio (2014) developed a clustering method based on densities but also requiring calculation of distances among observations. More recently, (Liu et al., 2021) proposed a method, with a name similar to that presented here, which is based on a probabilistic modelling of the standard k-means which is solved by numerical optimisation (see also Li et al., 2020). However, similarly to the standard k-means, and by contrast to the present method, they assumed homogeneous variance among groups. These authors

implemented their method in MATLAB and presented several applications with artificial and real data sets of modest sizes (several thousands observations or less) with known number of groups.

Richards et al. (2010) present a method that shares some similarities with the present one as it is based on the multivariate normal distribution. However, these authors proposed to maximise the log-likelihood function by expectation–maximisation (Dempster et al., 1977). Besides, they weight the contributions of the pixels to the likelihood function with respect to their spatial contiguity, which was not considered in the present work (but see the perspectives below).

In terms of running times, the probabilistic k-means has an attractive feature. As reported above, the analysis of a complete *Sentinel*-2 data set covering more than 12,000 km^2 at a resolution of 60 m with 13 bands takes around 8 min. Additionally, three analyses with several million pixels (including one with more than 10^8 pixels) are reported showing how the method presented here is scalable to very large data sets.

In addition to the developments on k-means, the package sentinel presented in this paper adds to the software tools available for the analysis of *Sentinel*-2 data. Ranghetti et al. (2020) presented another package written in R, sen2r, to handle and manage *Sentinel*-2 data. By comparison, sentinel makes it possible to search, download, and manage products and data from all *Sentinel* satellites. Besides, the R environment makes it possible to read the different file formats used in *Sentinel* products thanks to the packages rgdal (Bivand et al., 2018) and ncdf4 (Pierce, 2019). These integrated tools have the potential to contribute to a software environment for time-series analysis of remote sensing data (Gray and Song, 2013; Cai et al., 2014; Gómez et al., 2016).

The present paper aims at developing and implementing a fast unsupervised classification method to analyse multispectral data. The ultimate goal of this work is to be able to analyse large scale remote sensing data to infer changes in forest cover

over large areas (see e.g., Hermosilla et al., 2018; Paradis, 2020b,a). The approach presented in this paper offers several perspectives of future development in several directions. Although k-means is basically an unsupervised method, it could be extended into a supervised method by defining known groups and evaluating the *a priori* distribution of reflectance. This poses some difficulties since it is difficult to find reference sites with relevant information to use as 'training' data. Another direction which is currently pursued by the author is to include spatial contiguity in the model. Richards et al. (2010) used this information to calculate weights in their likelihood function. Currently, an approach using edge detection with Prewitt's and Sobel's operators (Wang et al., 2006) is under study.

Acknowledgements

The author is grateful to two reviewers for their constructive comments. The calculations used for data analyses benefited from the ISEM computing cluster platform. This is publication ISEM 2021-332.

Declaration of data availability

The data that support the findings of this study are openly available from `http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes` (Pavia University and Okavango), from `https://sso.theia-land.fr/` (Southern France), and from European Space Agency's Copernicus SciHub at `https://scihub.copernicus.eu/` (Eastern Thailand).

Declaration of interest statement

The author declares no conflict of interest.

Code availability

https://github.com/emmanuelparadis/sentinel.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akadémia Kiado, Budapest, pp. 267–281.

Ball, G.H., Hall, D.J., 1965. ISODATA, a novel method of data analysis and pattern classification. Technical Report, Stanford Research Institute, Menlo Park, California. URL: https://apps.dtic.mil/dtic/tr/fulltext/u2/699616.pdf.

Betts, M.G., Wolf, C., Ripple, W.J., Phalan, B., Millers, K.A., Duarte, A., Butchart, S.H.M., Levi, T., 2017. Global forest loss disproportionately erodes biodiversity in intact landscapes. Nature 547, 441–444. doi:10.1038/nature23285.

Bivand, R., Keitt, T., Rowlingson, B., 2018. rgdal: bindings for the 'Geospatial' Data Abstraction Library. URL: https://CRAN.R-project.org/package=rgdal. R package version 1.3-6.

Burrough, P.A., van Gaans, P.F.M., MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets and Systems 113, 37–52. doi:https://doi.org/10.1016/S0165-0114(99)00011-1.

572 Cai, S.S., Liu, D.S., Sulla-Menashe, D., Friedl, M.A., 2014. Enhancing MODIS land cover product with a spatial-temporal modeling algorithm. Remote Sensing of 574 Environment 147, 243–255. doi:10.1016/j.rse.2014.03.012.

Cao, J., Wu, Z., Wu, J.J., Liu, W., 2013. Towards information-theoretic K-means 576 clustering for image indexing. Signal Processing 93, 2026–2037. doi:10.1016/j.sigpro.2012.07.030.

578 Chavez, Jr, P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environ-580 ment 24, 459–479. doi:10.1016/0034-4257(88)90019-3.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incom-582 plete data via the *EM* algorithm (with discussion). Journal of the Royal Statistical Society. Series B. Methodological 39, 1–38.

584 Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 586 Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon. pp. 226–231.

588 Fu, J., Ma, J., Chen, P., Chen, F., 2020. Remote sensing satellites for Digital Earth, in: Guo, H., Goodchild, M.F., Annoni, A. (Eds.), Manual of Digital Earth. 590 Springer, Berlin, pp. 55–123.

Galluccio, L., Michel, O., Comon, P., Hero, A.O., 2012. Graph based k-means 592 clustering. Signal Processing 92, 1970–1984. doi:10.1016/j.sigpro.2011.12.009.

594 Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A., Languille, F., Alhammoud, 596 B., Viallefont, F., Pflug, B., Bieniarz, J., Clerc, S., Pessiot, L., Trémas, T., Cadau,

E., De Bonis, R., Isola, C., Martimort, P., Fernandez, V., 2017. Copernicus Sentinel-2A calibration and products validation status. Remote Sensing 9, 584. doi:10.3390/rs9060584.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: a review. ISPRS Journal of Photogrammetry and Remote Sensing 116, 55–72. doi:10.1016/j.isprsjprs.2016.03.008.

Gray, J., Song, C.H., 2013. Consistent classification of image time series with automatic adaptive signature generalization. Remote Sensing of Environment 134, 333–341. doi:10.1016/j.rse.2013.03.022.

Guo, H., Goodchild, M.F., Annoni, A. (Eds.), 2020. Manual of Digital Earth. Springer, Berlin. doi:10.1007/978-981-32-9915-3.

Hahsler, M., Piekenbrock, M., Doran, D., 2019. dbscan: fast density-based clustering with R. Journal of Statistical Software 91, 1–30. doi:10.18637/jss.v091.i01.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. Applied Statistics 28, 100–108. doi:10.2307/2346830.

Hastie, T.J., Tibshirani, R.J., Friedman, J., 2009. The elements of statistical learning. Data mining, inference, and prediction (second edition). Springer, New York.

Haut, J.M., Paoletti, M., Plaza, J., Plaza, A., 2017. Cloud implementation of the K-means algorithm for hyperspectral image analysis. Journal of Supercomputing 73, 514–529. doi:10.1007/s11227-016-1896-3.

He, T., Sun, Y.J., Xu, J.D., Wang, X.J., Hu, C.R., 2014. Enhanced land use/cover classification using support vector machines and fuzzy k-means clustering algorithms. Journal of Applied Remote Sensing 8, 083636. doi:10.1117/1.JRS.8.083636.

Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., 2018. Disturbance-informed annual land cover classification maps of Canada's forested ecosystems for a 29-Year Landsat time series. Canadian Journal of Remote Sensing 44, 67–87. doi:10.1080/07038992.2018.1437719.

Jones, B.A., Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M.Y., McKeever, D., Mutua, F., Young, J., McDermott, J., Pfeiffer, D.U., 2013. Zoonosis emergence linked to agricultural intensification and environmental change. Proceedings of the National Academy of Sciences USA 110, 8399–8404. doi:10.1073/pnas.1208059110.

Kanthana, M.R., Sujathab, S.N.N., 2013. Automatic grayscale classification using histogram clustering for active contour models. International Journal of Current Engineering and Technology 3, 369–373.

Kavzoglu, T., Tonbul, H., 2018. An experimental comparison of multi-resolution segmentation, SLIC and K-means clustering for object-based classification of VHR imagery. International Journal of Remote Sensing 39, 6020–6036. doi:10.1080/01431161.2018.1506592.

Kerenidis, I., Landman, J., Luongo, A., Prakash, A., 2019. q-means: a quantum algorithm for unsupervised machine learning, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/16026d60ff9b54410b3435b403afd226-Paper.pdf.

Kesikoglu, M.H., Atasever, U.H., Ozkan, C., 2013. Unsupervised change detection in satellite images using fuzzy c-means clustering and principal component analysis. International Archives of the Photogrammetry, Re-

mote Sensing and Spatial Information Sciences XL-7, 129–132. doi:10.5194/isprsarchives-XL-7-W2-129-2013.

Khan, S.A., Awan, A.J., Vall-Llosera, G., 2019. K-means clustering on noisy intermediate scale quantum computers. URL: http://arxiv.org/abs/1909.12183.

Kuo, K.T., Itakura, K., Hosoi, F., 2019. Leaf segmentation based on k-means algorithm to obtain leaf angle distribution using terrestrial LiDAR. Remote Sensing 11, 2536. doi:10.3390/rs11212536.

Li, J., Roy, D.P., 2017. A global analysis of Sentinel-2A, Sentinel-2B and Landsat-8 data revisit intervals and implications for terrestrial monitoring. Remote Sensing 9, 902. doi:10.3390/rs9090902.

Li, X., Zhang, P., Zhu, G., 2019. DBSCAN clustering algorithms for non-uniform density data and its application in urban rail passenger aggregation distribution. Energies 12, 3722. doi:10.3390/en12193722.

Li, Y., Liu, B., Liu, Z., Zhang, T., 2020. Probabilistic k-means clustering via nonlinear programming URL: https://arxiv.org/abs/2001.03286.

Liu, B., Li, Y., Zhang, T., Liu, Z., 2021. L_p-norm probabilistic K-means clustering via nonlinear programming. International Journal of Machine Learning and Cybernetics 12, 1597–1607. doi:10.1007/s13042-020-01257-6.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28, 129–137. doi:10.1109/TIT.1982.1056489.

Lv, Z., Hu, Y., Zhong, H., Wu, J., Li, B., Zhao, H., 2010. Parallel K-means clustering of remote sensing images based on MapReduce, in: Wang, F.L., Gong, Z., Luo, X., Lei, J. (Eds.), Web Information Systems and Mining. Proceedings of the International Conference, WISM 2010 Sanya, China, October 23–24, 2010. Lecture Notes in Computer Science 6318. Springer-Verlag, Berlin, pp. 162–170.

Lv, Z.Y., Liu, T.F., Shi, C., Benediktsson, J.A., Du, H.J., 2019. Novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images. IEEE Access 7, 34425–34437. doi:10.1109/ACCESS.2019.2892648.

Ma, A.L., Zhong, Y.F., Zhang, L.P., 2016. Spectral-spatial clustering with a local weight parameter determination method for remote sensing imagery. Remote Sensing 8, 124. doi:10.3390/rs8020124.

Memarsadeghi, N., Mount, D.M., Netanyahu, N.S., Le Moigne, J., 2007. A fast implementation of the isodata clustering algorithm. International Journal of Computational Geometry & Applications 17, 71–103. doi:10.1142/S0218195907002252.

Morand, S., Blasdell, K., Bordes, F., Buchy, P., Carcy, B., Chaisiri, K., Chaval, Y., Claude, J., Cosson, J.F., Desquesnes, M., Jittapalapong, S., Jiyipong, T., Karnchanabanthoen, A., Pornpan, P., Rolain, J.M., Tran, A., 2019. Changing landscapes of Southeast Asia and rodent-borne diseases: decreased diversity but increased transmission risks. Ecological Applications 29, e01886. doi:10.1002/eap.1886.

Newbold, T., Hudson, L.N., Arnell, A.P., Contu, S., De Palma, A., Ferrier, S., Hill, S.L.L., Hoskins, A.J., Lysenko, I., Phillips, H.R.P., Burton, V.J., Chng, C.W.T., Emerson, S., Gao, D., Pask-Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B.I., Whitmee, S., Zhang, H.B., Scharlemann, J.P.W., Purvis, A., 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. Science 353, 288–291. doi:10.1126/science.aaf2201.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42–57. doi:10.1016/j.rse.2014.02.015.

Paradis, E., 2020a. Forest gains and losses in Southeast Asia over 27 years: the slow convergence towards reforestation. Forest Policy and Economics 122, 102332. doi:10.1016/j.forpol.2020.102332.

Paradis, E., 2020b. Modelling transition in land cover highlights forest losses and gains in Southeast Asia. Biodiversity and Conservation 29, 2539–2551. doi:10.1007/s10531-020-01987-7.

Pascucci, S., Carfora, M.F., Palombo, A., Pignatti, S., Casa, R., Pepe, M., Castaldi, F., 2018. A Comparison between standard and functional clustering methodologies: application to agricultural fields for yield pattern assessment. Remote Sensing 10, 585. doi:10.3390/rs10040585.

Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J.M., Tucker, C.J., Stenseth, N.C., 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in Ecology & Evolution 20, 503–510. doi:10.1016/j.tree.2005.05.011.

Pierce, D., 2019. ncdf4: interface to Unidata netCDF (version 4 or earlier) format data files. URL: https://CRAN.R-project.org/package=ncdf4. R package version 1.17.

Ranghetti, L., Boschetti, M., Nutini, F., Busetto, L., 2020. "sen2r": an R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. Computers & Geosciences 139, 104473. doi:10.1016/j.cageo.2020.104473.

Reza, M.N., Na, I.S., Baek, S.W., Lee, K.H., 2019. Rice yield estimation based on K-means clustering with graph-cut segmentation using low-altitude UAV images. Biosystems Engineering 177, 109–121. doi:10.1016/j.biosystemseng.2018.09.014.

Richards, J.W., Hardin, J., Grosfils, E.B., 2010. Weighted model-based clustering for

remote sensing image analysis. Computational Geosciences 14, 125–136. doi:10.1007/s10596-009-9136-z.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344, 1492–1496. doi:10.1126/science.1242072.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464. doi:10.1214/aos/1176344136.

Sedaghat, A., Ebadi, H., 2015. Very high resolution image matching based on local features and k-means clustering. Photogrammetric Record 30, 166–186. doi:10.1111/phor.12101.

Silverman, J., 2019. RcppHungarian: solves minimum cost bipartite matching problems. URL: https://CRAN.R-project.org/package=RcppHungarian. R package version 0.1.

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S (fourth edition). Springer, New York.

Wang, S., Ge, F., Liu, T., 2006. Evaluating edge detection through boundary detection. EURASIP Journal on Applied Signal Processing 2006, 76278. doi:10.1155/ASP/2006/76278.

Wang, Y., Li, D., Wang, Y., 2019. Realization of remote sensing image segmentation based on K-means clustering. IOP Conference Series: Materials Science and Engineering 490, 072008. doi:10.1088/1757-899x/490/7/072008.

Wulder, M.A., Coops, N.C., Roy, D.P., White, J.C., Hermosilla, T., 2018. Land cover 2.0. International Journal of Remote Sensing 39, 4254–4284. doi:10.1080/01431161.2018.1452075.

Zhang, F., Du, B., Zhang, L.P., Zhang, L.F., 2016. Hierarchical feature learning

with dropout k-means for hyperspectral image classification. Neurocomputing 187, 75–82. doi:10.1016/j.neucom.2015.07.132.

Zhang, G.J., Cowled, C., Shi, Z.L., Huang, Z.Y., Bishop-Lilly, K.A., Fang, X.D., Wynne, J.W., Xiong, Z.Q., Baker, M.L., Zhao, W., Tachedjian, M., Zhu, Y.B., Zhou, P., Jiang, X.T., Ng, J., Yang, L., Wu, L.J., Xiao, J., Feng, Y., Chen, Y.X., Sun, X.Q., Zhang, Y., Marsh, G.A., Crameri, G., Broder, C.C., Frey, K.G., Wang, L.F., Wang, J., 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339, 456–460. doi:10.1126/science.1230835.

Zhang, H.K.K., Roy, D.P., Yan, L., Li, Z.B., Huang, H.Y., Vermote, E., Skakun, S., Roger, J.C., 2018. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. Remote Sensing of Environment 215, 482–494. doi:10.1016/j.rse.2018.04.031.

Table 1: Main features of the data analysed in this study.

Site	Number of pixels		Resolution (m)	Bands
Pavia University	207,400	(610×340)	1.3	103
Okavango	377,856	(1476×256)	30	145
Montpellier	89,161,101	(9799×9099)	6	4
Eastern Thailand	120,560,400	(10980×10980)	10	4
Eastern Thailand	3,348,900	(1830×1830)	60	13

Figure 1: (A) Two normal distributions with mean and standard-deviation 0 and 2 (red) and 6 and 0.1 (blue). (B) Two hundred observations, shown under the x-axis, were simulated from each distribution in A. After a standard k-means classification, 15 observations were misclassified. After a probabilistic k-means, four observations remained misclassified. The curves show the densities inferred from the observations.

Figure 2: Workflow of the probabilistic k-means for the analysis of spectral imaging data.

Figure 3: Maps of the Pavia University data set considering only the reference data.

Figure 4: Values of entropy (H) with different numbers of groups (K) for the Pavia University and Okavango data sets.

Figure 5: Maps of the Pavia University data set.

Figure 6: Maps of the Okavango data set considering only the reference data.

Figure 7: Maps of the Okavango data set.

Figure 8: Values of entropy (H) with different numbers of groups (K) for the Southern France data set.

Figure 9: Maps of the Southern France data set.

Figure 10: Values of entropy (H) with different numbers of groups (K) for the Eastern Thailand data set.

Figure 11: Maps of Eastern Thailand data set at the 10 m resolution. Scales are UTM-based (in km).

Figure 12: Maps of Eastern Thailand data set at the 60 m resolution. Scales are UTM-based (in km).