Euro-Mediterr J Environ Integr (2017) 2:24
DOI 10.1007/s41207-017-0036-7

ORIGINAL PAPER

Semiautomatic approach for land cover classification: a remote sensing study for arid climate in southeastern Tunisia

Moncef Bouaziz[1,2] · Stefanie Eisold[2] · Emna Guermazi[3]

Received: 24 November 2016 / Accepted: 4 September 2017 / Published online: 12 September 2017

Abstract The land use and land cover (LULC) classification has great potential to contribute to the monitoring of land degradation and climatic disasters. The purpose of this study was to assess the performance of parametric and non-parametric classification methods using remotely sensed Landsat satellite data of arid and semiarid areas, based on the computed producer's accuracy, user's accuracy, overall accuracy, and Cohen's kappa coefficient. Three LULC classes were identified, and supervised classifications were applied to Landsat 8 imagery. The results show that the support vector machines (SVM) classification method produced more accurate results, using two different kernel functions, compared with the maximum likelihood classification (MLC) and the minimum distance classification (MDC). The basis radial function affords the highest overall classification accuracy of 91.20% and a mean kappa coefficient of 0.87. This classification method is very well suited to accurately map LULC in arid and semiarid regions where the main vegetation type is oasis or steppes.

Keywords Land use · Land cover · Classification · Landsat imagery · Arid areas

✉ Emna Guermazi
guermazi.emna@ymail.com

[1] Department of Geology, Faculty of Sciences of Gafsa, University of Gafsa, Gafsa, Tunisia

[2] Faculty of Environmental Sciences, Institute of Geography, TU-Dresden, Helmholtzstr. 10, 1609 Dresden, Germany

[3] National School of Engineers, University of Sfax, Sfax, Tunisia

Introduction

Optical remote sensing data are an attractive source to generate land cover thematic maps, providing valuable information to determine extent of land cover classes as well as for performing temporal land cover change and risk analysis at different scales (Kavzoglu and Colkesen 2009). Land use and land cover (LULC) mapping is also relevant for the monitoring of desertification and land degradation. It is considered as key environmental parameter in arid areas such as the Mediterranean basin (Castillejo-González et al. 2009). Machine learning algorithms are used in remote sensing to generate and explore the LULC classification from multi-source remote sensing data. They are currently widely used in remote sensing classification (Wang et al. 2006).

Traditional approaches for automated land cover mapping using remotely sensed data have employed pattern recognition techniques including supervised and unsupervised approaches (Richards 1992). Both parametric and non-parametric classification techniques have been developed and applied for different remote sensing applications. A parametric classifier is based on the statistical probability distribution of each class (Kumar and Sahoo 2012), the most widely used parametric classifier is the maximum likelihood classification (MLC) (Guermazi et al. 2016). Non-parametric classifiers are used to estimate the probability density function when it is unknown (Kumar and Sahoo 2012) such as support vector machines (SVM) and artificial neural networks (ANN). Statistical classifiers depend on some predefined data model and the performance of these classifiers depends on how well the data match the predefined model (Pal and Mather 2004). Adam et al. (2014) confirmed the performance of machine-learning random forest (RF) and SVM classifiers to map

heterogeneous land in South Africa using RapidEye high resolution imagery. The same classifiers were applied to generate a global land cover using medium spatial resolution data of Landsat TM and ETM+ sensors (Gong et al. 2013). Senf et al. (2015) combined multi-seasonal images of MODerate-resolution Imaging Spectroradiometer (MODIS) and Landsat to produce a land cover map in Southern Portugal using the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM).

Paneque-Gálvez et al. (2013) demonstrated the high performance of SVM by classifying heterogeneous landscapes using Landsat TM images. A previous study by Pal and Mather (2005) showed that the SVM attains a higher classification accuracy than either the MLC or the ANN classifier applied for Landsat ETM+ and hyperspectral imagery, and that the SVM can be used with small training data sets and high-dimensional data. Previous work shows that SVM are not sensitive to training sample size. SVM have been improved to successfully work with limited quantity and quality of training samples (Mantero et al. 2005; Mountrakis et al. 2011).

The classification accuracy assessment is an essential application of the thematic mapping. Previously, accuracy assessment was based on a visual interpretation of the derived map (Foody 2002). Currently, researchers use the confusion or error matrix by comparing class in the predicted thematic map and ground truth data. Several classification accuracies may be derived from the information content of the confusion matrix. The most popular are user and producer accuracy and overall accuracy (Yuan et al. 2005; Myint et al. 2011; Jia et al. 2014). Nevertheless, Smits et al. (1999) define Cohen's kappa coefficient as a standard measure of classification accuracy.

This research has focused on the application and assessment of different classification algorithms of LULC in Southern Tunisia. It was conducted to test the potential of machine learning algorithms to classify LULC in arid regions.

Data and methodology

Study area and data sets

The area investigated was located in southeastern Tunisia between Jeffara plain and the Gulf of Gabes. The study area was chosen because of the important agriculture interests in this region and also the environmental problems related to soil such as salinization and polluted soils due to the chemical industries in this region. The geographic locations correspond to 33.92°N and 33.78°N and 10.01°E and 10.10°E (Fig. 1). The study area is characterized by an arid to semiarid climate. The rainfall is irregular and ranges between 150 and 240 mm per year with 6 months dry season (April–September), when the rain does not exceed 4 mm per month.

The Landsat 8 imagery used in this study was acquired from https://earthexplorer.usgs.gov on 24 May 2013 with 1.75% cloud cover. The Landsat 8 satellite includes two sensors: the Operational Land Imager (OLI), which provides nine spectral bands, and the Thermal Infrared Sensor (TIRS), which consists of two thermal bands. The product Level 1T was used, which refers to orthorectified data. OLI data is converted into radiance and subsequently into atmospherically corrected surface reflectance using the fast line-of-sight atmospheric analysis of spectral hypercubes (FLAASH), a MODTRAN4-based algorithm (Felde et al. 2003).

Classification algorithms

A methodological framework in the context of remote sensing and geographic information system (GIS) techniques was considered to classify Landsat 8 imagery into three different LULC classes identified in this study area, namely bare soil, vegetation, and urban area.

Training areas were selected on the basis of knowledge and information available from the region that were acquired during fieldwork and then combined with available information from Landsat.

For better discrimination in selecting training samples from image classes, we used the results of unsupervised algorithms to help define the training features. The *k*-means clustering method was selected to obtain probable clusters. In addition, visual interpretations from RGB composition and reference maps were considered for a better selection of the training sets.

The access to the sites in the field was in some places limited and therefore we could not obtain specific spatial structure of the sampling network. The 68 sites were visited in the field and recognized in the Landsat 8 image.

The spectral plots and ground observations were used to define the regions of interest (ROIs) (105 pixels for each class).

A brief description of maximum likelihood classification, SVM, and minimum distance classification are given below.

Support vector machines

In universal learning machines, SVM are supervised learning models, used for pattern recognition and originally designed to solve binary classification problems (Wijaya et al. 2008).

The SVM technique uses hyperplanes to separate data points into several classes. In doing so, support vectors ensure that the margin width will be maximized (Fig. 2). A

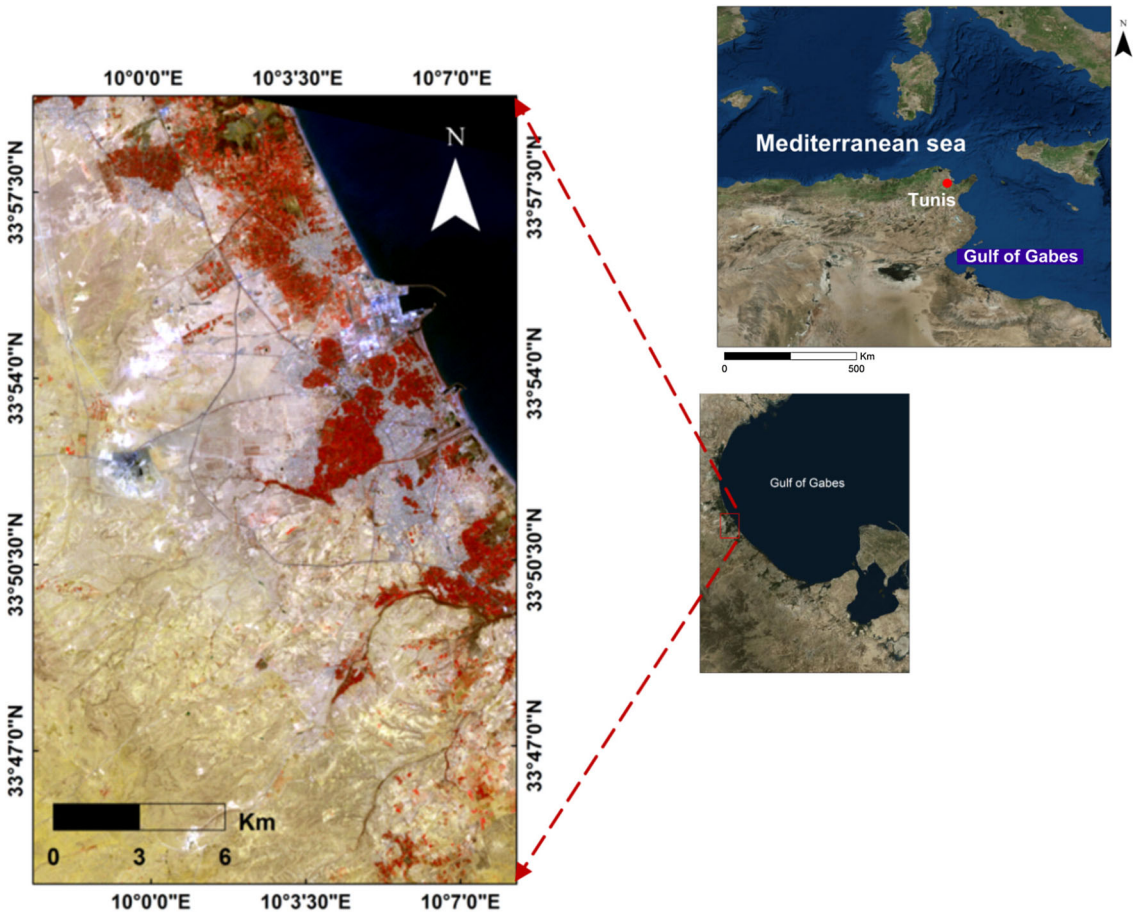

Fig. 1 Location of the study area: Landsat 8 false colors (left); Google earth imagery of the Gulf of Gabes (right)

single supplementary data point can notably effect the location of the hyperplane. In its original conception the method is presented with a set of labeled data and the SVM training algorithm objective is to find a hyperplane that separates the data set into a discrete predefined number of classes in a fashion consistent with the training examples (Vapnik 1982). Figure 2 illustrates a simple scenario of a two-class separable classification problem in a two-dimensional input space.

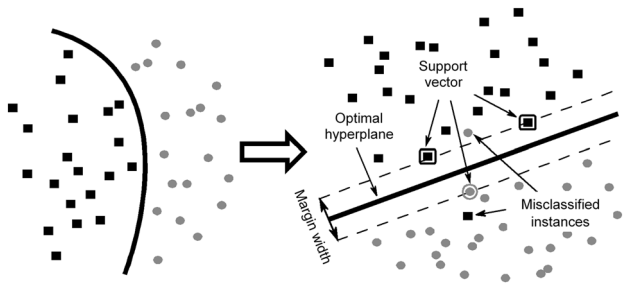

Fig. 2 Linear support vector machine example presenting a simple scenario of a two-class separable case; optimal hyperplane separates the data set into two classes by a defined margin width. Adapted from Burges (1998)

SVM aim to determine the optimal separating hyperplane (OSH) among all the possible hyperplanes (Srivastava et al. 2012) and this is done through an optimization problem using Lagrange multipliers and quadratic programming methods (Pal and Mather 2004).

The SVM classification requires a proper selection of kernel function to establish accurate hyperplanes that minimize misclassification error (Wijaya et al. 2008). The kernel function allows the training data to project the training data in a larger space where it may be increasingly possible to discover a superior separating margin for the OSH.

An important aspect of the SVM technique is the type of kernel that is used. Two SVM kernels were selected for the classification cases in this study: the radial basis function and a polynomial function, which are given in Eqs. (1) and (2), respectively:

$$K(x_i, x_j) = \mathrm{e}^{-\gamma(x_i, x_j)^2}, \quad \gamma > 0, \tag{1}$$

$$K(x_i, x_j) = (\gamma x_i^{\gamma} x_j + r)^d, \quad \gamma > 0, \tag{2}$$

where x_i and x_j, represent feature vectors in some input space, γ is the width of the kernel function, d is the

polynomial degree term, and r is the bias term in the kernel function (Srivastava et al. 2012).

Linear spectral unmixing

Every image pixel is always a mixture of different components (Tompkins et al. 1986). The idea behind the linear spectral unmixing (LSU) method is to decompose the pixel spectra into a collection of constituent spectra, or spectral signatures, and their corresponding fractional abundances that quantify the proportion of end-members present in the pixel. The technique is useful for extracting information from data with low spatial resolution.

Minimum distance classification

Supervised minimum distance classification (MDC) was also applied in this study. It is a non-parametric classifier, which uses the mean vectors of each end-member and calculates the Euclidean distance from each unknown pixel to the mean vector for each class (Richards and Jia 1999). All pixels are classified to the nearest class unless a standard deviation or distance threshold is specified, in which case some pixels may be unclassified if they do not meet the selected criteria. The MD algorithm is fast and one of the more commonly used algorithms because of its mathematic simplicity, only requiring the mean vectors for each band from the training data. This method does not consider class variability; thus, large differences in the variance of the classes often lead to misclassification (Lu et al. 2004).

Maximum likelihood classification

Maximum likelihood classification (MLC) is the most widely adopted parametric classification algorithm (Jensen 2005). The MLC algorithm is based on probability distributions and decision rules, which assume the data values to be a set of multivariate normal distributions (Manandhar et al. 2009). The algorithm classification assigns a particular class to each pixel on the basis of the shortest modified distance of the pixel from the class mean. It also considers shape, size, and orientation of the training samples.

MDC, MLC, and classification by SVM was applied to the Landsat 8 imagery. In this study the standard deviation was varied to find out the optimal MDC for the selected Landsat 8 imagery. In one case LSU was applied to the data before MDC. Table 1 shows the different LULC classifications applied and the input parameters tested for each method.

Accuracy assessment

Classification accuracies were assessed using confusion matrices. Sample data of salt-affected soils were initially selected from the satellite image and confirmed with the information collected during fieldwork, which was then used to run the classification.

A confusion matrix was used to calculate the producer's accuracy, user's accuracy, and overall accuracy. Also the kappa coefficient was assessed.

The producer's accuracy is calculated by dividing the amount of pixels in a particular class classified correctly by the total of ground truth pixels of this class. The user's accuracy indicates the percentage of probability that the class in which a pixel is classified to on an image actually represents that class on the ground. It is the ratio between the number of correctly classified pixels of a class and the number of pixels the classifier labeled into this class. The overall accuracy is the total percentage of pixels correctly classified.

The measurements of user's and producer's accuracies are related to commission and omission errors (Gupta and Srivastava 2010).

Cohen's kappa coefficient (K), which considers all of the elements of the error matrix, is computed as given by Eq. (3) (Bishop et al. 1977):

$$K = \frac{N\sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r}(x_{i+})(x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+})(x_{+i})}, \quad (3)$$

where N is the number of observations, X_{ii} is the number of observations in row i and column i (the major diagonal in the confusion matrix), x_{+i} and x_{i+} are the marginal totals of row r and column i, respectively, and r is the number of rows in the matrix.

Results and discussion

To compare the SVM, MLC, and MD classifications, producer's accuracy, user's accuracy, overall accuracy, and kappa coefficient were computed (Fig. 3). The results show better performances of SVM (with two different kernel functions; radial basis and polynomial functions) to classify the soil than the other classifiers. The use of the radial basis as the kernel function in the SVM classifier affords the most accurate results with 91% overall accuracy. The MD classifier shows the lowest accuracy with 78% overall accuracy and kappa coefficient of 0.68. The literature does not provide an obvious explanation of this behavior and mainly links the observations to the components of the algorithms (Bouaziz et al. 2011).

Table 1 Algorithm and input parameters for the different LULC classifications

	Algorithm	Parameters	Input parameters	
SVM1	SVM	Kernel function	Radial basis function	
SVM2	SVM	Kernel function	Polynomial function	
MDC1	LSU/MDC	–	–	
MDC2	MDC	Max Std for each class of ROIs	Bare soil	30
			Urban	2.3
			Vegetation	10
MDC3	MDC	Max Std for each class of ROIs	Bare soil	40
			Urban	4.5
			Vegetation	15
MLC	MLC	Probability thresholds	Soil	0.01
			Urban	0.99
			Vegetation	0.01

Max Std maximum standard deviation

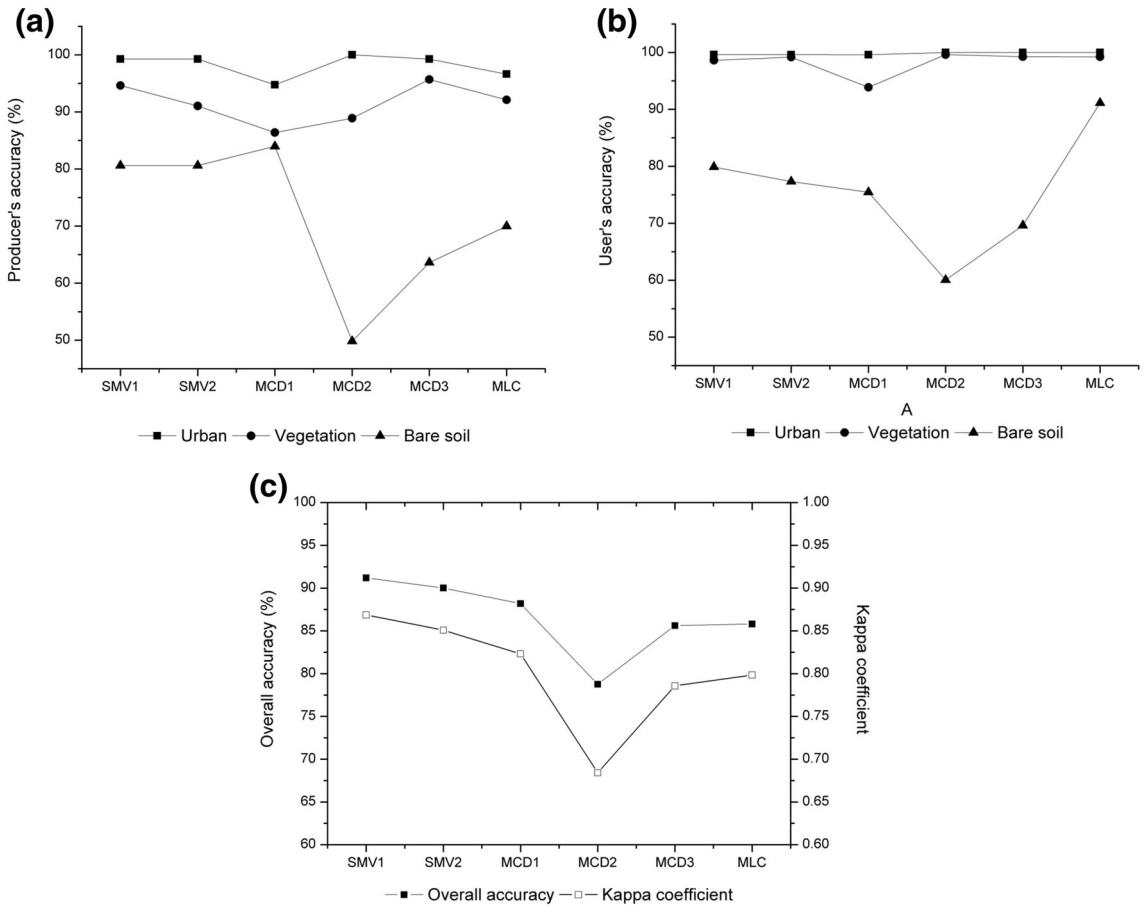

Fig. 3 Producer's accuracy, user's accuracy, overall accuracy, and kappa statistics of LULC classifications

Overall accuracy is highest for SVM1 with 91.20%, closely followed by SVM2 with 90.01%. The lowest value is for MDC2 with 78.75%. MDC6 and MLC have values in the middle range with circa 86%. MDC1 is represented with 88.19% and therefore more like the results of the SVM classifications.

The kappa coefficient has its lowest value for MDC2 with 0.68. SVM1 and SVM2 achieved the highest results

with 0.87 and 0.85. The remaining values fluctuate around 0.80 with a maximal variance of 0.20.

All classification results of the producer's accuracy show lowest results for the classification of pixels in the bare soil class. MDC2 has the lowest value with 49.83%. SVM1 and SVM2 have the best results with 80.61% each. The vegetation class has higher values with a minimum of 86.38% for MDC1 and a maximum of 94.62% for SVM1. The urban class has best results with a minimum of 94.74% for MDC1 and a maximum of 100% for MDC2.

The highest values of user's accuracy were obtained for the urban class. MDC2, MDC3, and MLC show 100%, followed by 99.16% for SVM2, 98.75% for SVM1, and 93.89% for MDC1. The results for the vegetation class show only values over 99%. The bare soil class has the lowest values for user's accuracy with a maximum of 91.13% for MLC and a minimum of 60.05% for MDC2.

The SVM algorithm using the radial basis as kernel function (SVM1) seems to produce the most stable results. Overall accuracy and kappa coefficient are the highest from all classification results. Also SVM1 not only produced the highest values for all classes, it has a minimum of 79.88%. Other classifications afford better results in some classes, but the values fluctuating more. For example, MLC has very high values for all classes except for producer's accuracy for the urban class. Also overall accuracy and kappa coefficient are not as high as in the SVM results.

A chief advantage of the SVM technique is the reduced necessity, compared to others used techniques, for obtaining a complete training data set. Even with sparse sampling the SVM technique permits informative classification. This was a benefit to the present study as training samples were not easily obtained. This is in part the reason why the SVM technique produced the best LULC classification. As highlighted by Mantero et al. (2005), SVM are particularly appealing in the remote sensing field owing to their ability to successfully handle small training data sets, often producing higher classification accuracy than the traditional methods.

Conclusion

The conducted research shows the high potential of machine learning algorithms applied on Landsat data to classify and discern patterns of different land cover. Computed accuracy from the classification results gave different but encouraging accuracy results varying between 78.75 and 91.20%. Therefore, SVM was more efficient than maximum likelihood (ML) and minimum distance (MD) classifications in this study. The SVM classifier with the radial basis function showed the best performance in extracting patterns and features of LULC classes over an arid region. Thus, we have demonstrated that the SVM classifiers based on carefully selected input variables are suitable for classifying land.

Acknowledgements The authors gratefully acknowledge the financial support of German Academic Exchange. Service (DAAD).

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

Adam E, Mutanga O, Odindi J, Abdel-Rahman EM (2014) Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers. Int J Remote Sens 35:3440–3458. doi:10.1080/01431161.2014.903435

Bishop YMM, Fienberg SE, Holland PW et al (1977) Book review: discrete multivariate analysis: theory and practice. Appl Psychol Meas 1:297–306. doi:10.1177/014662167700100218

Bouaziz M, Wijaya A, Gloaguen R (2011) Remote gully erosion mapping using ASTER data and geomorphologic analysis in the main Ethiopian rift. Geo Spat Inf Sci 14:246–254. doi:10.1007/s11806-011-0565-1

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167. doi:10.1023/A:1009715923555

Castillejo-González IL, López-Granados F, García-Ferrer A et al (2009) Object- and pixel-based analysis for mapping crops and their agro-environmental associated measures using QuickBird imagery. Comput Electron Agric 68:207–215. doi:10.1016/j.compag.2009.06.004

Felde GW, Anderson GP, Cooley TW, Matthew MW, Adler-Golden SM, Berk A, Lee J (2003) Analysis of hyperion data with the FLAASH atmospheric correction algorithm. IEEE Trans Geosci Remote Sen 3:90–92

Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sens Environ 80(1):185–201

Gong P, Wang J, Yu L et al (2013) Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. Int J Remote Sens 34:2607–2654. doi:10.1080/01431161.2012.748992

Guermazi E, Bouaziz M, Zairi M (2016) Water irrigation management using remote sensing techniques: a case study in Central Tunisia. Environ Earth Sci 75:202. doi:10.1007/s12665-015-4804-x

Gupta M, Srivastava PK (2010) Integrating GIS and remote sensing for identification of groundwater potential zones in the hilly terrain of Pavagarh, Gujarat, India. Water Int 35:233–245

Jensen JR (2005) Thematic map accuracy assessment. In: Introductory digital image processing: a remote sensing perspective, 3rd edn. Geographic Information Science Series. Prentice Hall, Upper Saddle River, pp 495–515

Jia K, Liang S, Zhang N et al (2014) Land cover classification of finer resolution remote sensing data integrating temporal features from time series coarser resolution data. ISPRS J Photogramm Remote Sens 93:49–55. doi:10.1016/j.isprsjprs.2014.04.004

Kavzoglu T, Colkesen I (2009) A kernel functions analysis for support vector machines for land cover classification. Int J Appl Earth Obs Geoinf 11:352–359. doi:10.1016/j.jag.2009.06.002

Kumar Y, Sahoo G (2012) Analysis of parametric & non parametric classifiers for classification technique using WEKA. Int J Inf Technol Comput Sci 4:43–49. doi:10.5815/ijitcs.2012.07.06

Lu D, Mausel P, Batistella M, Moran E (2004) Comparison of land-cover classification methods in the Brazilian Amazon Basin. Photogramm Eng Remote Sens 70:723–731. doi:10.14358/PERS.70.6.723

Manandhar R, Odeh IOA, Ancev T (2009) Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement. Remote Sens 1:330–344. doi:10.3390/rs1030330

Mantero P, Moser G, Serpico S (2005) Partially supervised classification of remote sensing images through SVM-based probability density estimation. IEEE Trans Geosci Remote Sens 43:559–570

Mountrakis G, Im J, Ogole C (2011) Support vector machines in remote sensing: a review. ISPRS J Photogramm Remote Sens 66:247–259. doi:10.1016/j.isprsjprs.2010.11.001

Myint SW, Gober P, Brazel A et al (2011) Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sens Environ 115:1145–1161. doi:10.1016/j.rse.2010.12.017

Pal M, Mather PM (2004) Assessment of the effectiveness of support vector machines for hyperspectral data. Future Gener Comput Syst 20:1215–1225. doi:10.1016/j.future.2003.11.011

Pal M, Mather PM (2005) Support vector machines for classification in remote sensing. Int J Remote Sens 26:1007–1011. doi:10.1080/01431160512331314083

Paneque-Gálvez J, Mas J-F, Moré G et al (2013) Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity. Int J Appl Earth Obs Geoinf 23:372–383. doi:10.1016/j.jag.2012.10.007

Richards JA (1992) Remote sensing digital image analysis. Springer, Cambridge

Richards JA, Jia X (1999) Remote sensing digital image analysis, 3rd edn. Springer, Berlin

Senf C, Leitão PJ, Pflugmacher D et al (2015) Mapping land cover in complex Mediterranean landscapes using Landsat: improved classification accuracies from integrating multi-seasonal and synthetic imagery. Remote Sens Environ 156:527–536. doi:10.1016/j.rse.2014.10.018

Smits PC, Dellepiane SG, Schowengerdt RA (1999) Quality assessment of image classification algorithms for land-cover mapping: a review and proposal for a cost-based approach. Int J Remote Sens 20:1461–1486

Srivastava PK, Han D, Rico-Ramirez MA et al (2012) Selection of classification techniques for land use/land cover change investigation. Adv Space Res 50:1250–1265. doi:10.1016/j.asr.2012.06.032

Tompkins S, Mustard JF, Forsyth DW (1986) Optimization of endmembers for spectral mixture analysis. Remote Sens Environ 59:472–489. doi:10.1016/S0034-4257(96)00122-8

Vapnik V (1982) Estimation of dependences based on empirical data. Springer, New York

Wang L, Zhu J, Zou H (2006) The doubly regularized support vector machine. Stat Sin 16:589–615

Wijaya A, Marpu PR, Gloaguen R (2008) Geostatistics texture classification of tropical rainforest in Indonesia. In: Stein S, Shi W, Bijker W (eds) Quality aspects in spatial data mining. CRC, Boca Raton, pp 199–210

Yuan F, Sawaya KE, Loeffelholz BC, Bauer ME (2005) Land cover classification and change analysis of the Twin Cities (Minnesota) metropolitan area by multitemporal Landsat remote sensing. Remote Sens Environ 98:317–328. doi:10.1016/j.rse.2005.08.006