Evaluation of urban sprawl from space using open source technologies

Gabriele Nolè, Beniamino Murgante, Giuseppe Calamita, Antonio Lanorte, Rosa Lasaponara

A R T I C L E   I N F O

Article history:
Accepted 6 May 2014
Available online xxxx

Keywords:
Urban sprawl
Planning
Sustainability
Satellite data
SVM
Open source software

A B S T R A C T

Up to nowadays, satellite data have become increasingly available, thus offering a low cost or even free of charge unique tool, with a great potential for quantitative assessment of urban expansion and urban sprawl, as well as for monitoring of land use changes and soil consumption. This growing observational capacity has also highlighted the need for research efforts aimed at exploring the potential offered by data processing methods and algorithms, in order to exploit as much as possible this invaluable space-based data source. The work herein presented concerns an application study on the process of urban sprawl conducted with the use of satellite ASTER data. The selected test site is highly significant, being it a coastal zone (with the presence of sand and rocks) characterized by a fragmented ecosystem and small towns, with an increasing rate of urbanization and soil consumption. In order to produce synthetic maps of urban areas, ASTER images were classified using two automatic classifiers, Maximum Likelihood (MLC) and Support Vector Machines (SVMs) applied by changing setting parameters, with the aim to compare their respective performances in terms of robustness, speed and accuracy. All process steps have been developed integrating Geographical Information System and Remote Sensing, and adopting free and open source software. Results pointed out that the SVM classifier with RBF kernel was generally the best choice (with accuracy higher than 90%) among all the configurations compared, and the use of multiple bands globally improves classification. One of the critical elements found in this case study is given by the presence of sand and mixed with rocks. The use of different configurations for the SVMs, i.e. different kernels and values of the setting parameters, allowed us to calibrate the classifier also to cope with a specific need, as in our case, to achieve a reliable discrimination of sand from urban area.

1. Introduction

In last centuries urbanization processes were strongly connected to industrial development, with related population growth; today, despite a profound revolution in labour market, this trend is still in place and can be more associated to the greater opportunities that cities offer compared to rural areas (Glaeser, 2011). Urban population will double within next twenty years and this phenomenon will lead the construction of new buildings and infrastructures. In many cases, urban growth is not due to a real need, but also to private interests, forms of rent, types of financing for local authorities and wrong lifestyles. The risk is that the application of urbanization policies without rules in the countryside, produces waterproofing of territory, wears out soils and distorts landscapes.

The confusing outward proliferation of buildings of a city to rural land is a phenomenon identified as “urban sprawl”. Outside the town centre, territory meshes become increasingly wider and less regular, entailing several inefficiencies, such as increase in travel time to move from home to work (and vice versa) and car-dependence, due to the growth of municipal extra mobility and, thus, increased road congestion and its related issues (Camagni et al., 2002). At the same time, costs of urbanization grow, making increasingly difficult to sustain investment for public transport sector, construction and maintenance of road infrastructure and public services (public lighting, garbage collection, etc.). Other criticisms linked to urban sprawl are aesthetic pollution, noise and air pollution, environmental impact, and soil consumption. Urban sprawl generates a highly fragmented agricultural and natural landscape (Murgante et al., 2011). This phenomenon produces natural islands, defined by the boundaries of new urbanized areas, which are too small to accommodate the life of certain animal species; consequently soil consumption has also a negative impact on biodiversity (Romano and Zullo, 2013). Soil is also linked to agricultural and zoo-technical production; consequently, human nutrition depends on it. From a sustainability point of view, soil has a strong relationship with cycles of water and carbon. Urbanization process produces soil sealing. The soil loses its biological value, becoming unable to absorb and filter rainwater.
This phenomenon is particularly intensive in Italy (Clementi et al., 1996; Indovina, 1990; Romano and Zullo, 2012), where the house has always been considered as a form of investment and homebuilders and concrete industry have always had a strong influence on governments.

In traditional approaches, studies on settlement system development are almost always dealt with, through the use of spatial data analysis methods (Cerreta and Poli, 2013; Modica et al., 2012; Murgante and Danese, 2011; Perchinunno et al., 2012).

Other approaches adopt different kinds of models, in order to predict urban sprawl phenomenon (Batty and Xie, 1994; Chaudhuri and Clarke, 2014; Clarke and Gaydos, 1998; Givertz et al., 2008; Kok et al., 2001; Makse et al., 1998; Martellozzo, 2012; Martellozzo and Clarke, 2011; Olofsson et al., 2013).

Up to nowadays, remotely sensed data with an acceptable spatial and spectral resolution, together with GIS software tools, have become increasingly available (Boyd and Foody, 2011; Mesev, 2007; Steiger and Hay, 2009) and provide a unique tool to retrieve a number of key variables, relevant to the quantification of urban sprawl and soil consumption from a local up to a global scale. However, this growing observational capacity is also increasing the need for research efforts aimed at exploring the potential offered by data processing methods and algorithms, in order to exploit as much as possible this invaluable space-based data source (Lasaponara and Lanotte, 2012; Lasaponara et al., 2014; Nolè et al., 2013).

The work here presented concerns an application study on the process of urban sprawl, carried out by the use of remote sensed information, from ASTER data which already proved to be suitable for these topics. In order to produce synthetic maps of urban areas of the territory, ASTER images were classified using two automatic classifiers, which are Maximum Likelihood Classifier (MLC) and Support Vector Machines (SVMs), the latter based on machine learning theory (Zhu and Blumberg, 2002). The aim was to compare performances in terms of robustness, speed and accuracy of the two classifiers, regarding urban pixels.

All process steps have been developed integrating Geographical Information System and Remote Sensing, and adopting free and open source software.

2. Materials and methods

2.1. Data and tools

Multi-temporal satellite data with medium spatial resolution is very suitable in urban sprawl phenomena evaluation. Data adopted in this case study is an image acquired by ASTER sensor on August 24, 2007 (source: http://glovis.usgs.gov/); additional details can be found in Abrams, 2000, on EOS-TERRA platform (http://asterweb.jpl.nasa.gov/).

ASTER captures data at 15 m spatial resolution in 14 bands, from the visible to the thermal infrared wavelengths.

ASTER data have been processed and handled with GRASS GIS (Neteler and Mitatosa, 2004), Quantum GIS and R (www.r-project.org). GRASS and Quantum GIS are two desktop GIS software, which are integrated by means of a specific plug in, avoiding to change environment in data manipulation. R, on the contrary, is a language and a software environment that can be interfaced with both GRASS and Quantum GIS. The database has been implemented in GRASS core and repository of the application, while Maximum Likelihood Classification and supervised classification algorithms are available in R.

2.2. Study area

The study area is located in Apulia region, more precisely in the area close to the southern part of Bari, the main town of the region. The area has a dimension of approximately 253 km² and a coastline of nearly 17 km. The analysis is mainly focused in Polignano a Mare and Monopoli municipalities and concerns also a portion of Conversano Municipality.

This region was characterized by a remarkable phenomenon of urban sprawl for many years (Romano and Zullo, 2014). According to recent data by the Italian National Institute of Statistics (ISTAT, 2011), the area is characterized by a population density included in a range between 140 and 319 inhabitants per km².

Apulia is one of the Italian regions with a high trend of urban sprawl (Romano and Zullo, 2014), that often has no correlation with population growth. Fig. 2 shows a substantially constant demographic trend. (See Fig. 1.)

2.3. Methodology

The semiautomatic methodology herein described has been implemented in order to carry out a quantitative analysis of the urban sprawl process using ASTER satellite imagery and open source software. The importance of our objective is connected with the fact that today satellite time series and ancillary data are currently available, even free of charge (from national and international spatial agencies) and offer a great potential for a quantitative assessment of urban expansion; nevertheless suitable data processing (free and open source easy to be applied and generalized) are needed to extract and map reliable information from space.

In this paper, in order to produce synthetic maps of urban areas for operative applications, ASTER images were classified using two automatic classifiers, MLC and SVMs, described in Sections 2.4 and 2.5, respectively, with the aim to compare their respective performances in terms of robustness, speed and accuracy. All process steps have been developed integrating Geographical Information System and Remote Sensing, and adopting free and open source software.

In order to assess capabilities and reliabilities of MLC and SVMs for the discrimination of urban from non urban areas, both of them were applied to the same spectral bands in two different ways: (i) firstly as binary classifications with only two classes (urban/not-urban) as output, and then extended to (ii) 8 classes (urban, agricultural soils, forests and green areas, coastal seawater, non-coastal seawater, bare soil, sand and rock, pure sand).

The process started with the identification of training areas (ROI – Region Of Interest) using Quantum GIS, with the creation of a spatial database in GRASS and subsequently the application of MLC and SVMs algorithms implemented in the rasclass package of R. A procedure of estimation accuracy has been applied to the results of supervised classification. Accuracy parameters are an important aid in evaluating the effectiveness of a classification algorithm. Accuracy means the quantitative assessment of the pixels correctly assigned to the corresponding classes on the ground. A criterion usually employed for the evaluation of accuracy is based on the comparison between the reference data to earth (test), identified independently from those used for classification (training), and classified data.

This method is based on the “confusion matrix” (Congalton, 1988; Congalton and Mead, 1983) where a number of classes identified by the classifier and from direct observation are listed on rows and columns, respectively. In practice, matrix elements are represented by the number of pixels that the procedure (classification or direct observation) assigns to a given class: on the main diagonal are pixels assigned simultaneously to the same class by both procedures and which, therefore, are correct; those outside of the diagonal are considered errors. The calculation of confusion matrix is done on a limited number of pixels and not on the whole image, otherwise this would paradoxically require the knowledge of ground truth of the whole examination area.

Then, it is possible to obtain from confusion matrix, the total accuracy (overall accuracy) which represents a parameter that indicates the number of pixels correctly classified (i.e. the sum of the cells in the
diagonal matrix of confusion) divided by total number of analysed
pixels:

\[ O.A. = \frac{1}{\sum_{i=1}^{r} X_{ii}} \]  

where \( X_{ii} \) is the number of pixels in the row \( i \) and in the column
(diagonal), a parameter that generally is expressed as a percentage
(with \( r \) = number of rows and columns of the square matrix).

Considering instead individual classes, it is possible to obtain other pa-
rameters such as errors of commission, through the confusion matrix.

Errors of commission occur when pixels of different classes are in-
cluded in the same category, errors of omission instead occur when a
pixel belonging to a class has not been included. Other parameters are
Producer's Accuracy (number of pixels correctly classified as class \( X \)) / (total number of pixels belonging to class \( X \)) and User's Accuracy
(number of pixels correctly classified as class \( X \)) / (number of pixels
classified as Class \( X \)).

Fig. 1. Study area (elliptical box).

Fig. 2. Demographic trends (ISTAT: Residents years 2002–2012).

Please cite this article as: Nolè, G., et al., Evaluation of urban sprawl from space using open source technologies, Ecological Informatics (2014),
http://dx.doi.org/10.1016/j.ecoinf.2014.05.005
Another useful indicator in estimating accuracy is the coefficient of concordance by Cohen, or Kappa coefficient (see, for example Congalton, 1988; Congalton et al. 1983), also called Kappa Statistics Index. K coefficient takes into account the part of correct classification due to chance and compares the error generated by the obtained classification to that of a classification done in a completely random way. The value of Kappa coefficient can range between 0 and 1, lower values mean no correspondence between certain and classified data, on the contrary, high values describe a good correspondence between certain and classified data. Usually values greater than or equal to 0.75 are considered from excellent to good results, while those less than or equal to 0.4 are considered poor. Kappa Index is often used to compare results of classification algorithms and different methodologies; for a reliable comparison it is appropriate, however, to compare classifications with similar classes.

K coefficient is expressed in the following way:

\[ K_{hit} = \frac{\sum_{i=1}^{N} X_{ii} - \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij} \cdot X_{ji}}{N - \sum_{i=1}^{N} X_{i\cdot} \cdot X_{\cdot i}} \]

where N is the pixel number of all classes of ground truth, \( X_{ii} \) represents pixels on the diagonal of the matrix, \( X_{i\cdot} \) and \( X_{\cdot i} \) represent total rows and columns for each class (with \( r \) = number of rows and columns of the square matrix).

2.4. Maximum Likelihood Classifiers and Support Vector Machines

2.4.1. Maximum Likelihood Classifiers (MLC)

MLC is one of the most commonly used supervised classification algorithms (see, for example, Richards and Jia, 2006; Richards and Xiuping, 1999).

MLC, as other conventional classifications, assumes that data sets (multispectral pixels in remote sensing applications) after examination are characterized by an explicit underlying probability model. On the basis of this model, the algorithm provides a probability for each pixel of being in each class.

These classification methods are also known as parametric (i.e. based on distribution parameters) and the classification is performed on the basis of parametric signatures defined by statistical parameters (e.g., mean and covariance matrix) and attributes, such as, the number of spectral bands, mean, minimum and maximum value in each band, as well as the number of pixels and covariance matrix for each training cluster. Due to the a-priori assumption of underlying probability model, performance and reliability of these classifications depend on how well data (multispectral pixels in our cases) match the given pre-defined model. MLC is based on the assumption that each spectral class can be described by a multivariate normal distribution. Therefore, the probability distribution for each spectral class is based on a multivariate normal model with dimensions equal to the number of spectral bands. Assuming that the probability distributions are normally (Gaussian) distributed with some unknown mean and variance, the method is fed by mean and variance estimated using known values of some samples (selected areas/pixel of interest) of the overall population (multi-spectral images). Therefore, MLC requires a preliminary knowledge necessary to generate representative parameters for each class of interest and to carry out the training stage. The whole procedure, as for all the supervised classification algorithms, requires two basic steps: (i) clustering or training which consists in providing known areas for each class, generally identified through in situ analysis, and (ii) classification which is carried out by comparing the spectral signature to each pixel (under investigation) with the spectral signature of the training cluster. Finally, the categorization is carried out according to the Bayesian rule, for which pixels are assigned to the class which exhibits the higher probability, so that the final results are obtained minimizing the errors in terms of probability. The metric (confusion matrix, overall accuracy, K index, etc.) enables us to assess the reliability of the obtained classes.

The effectiveness of MLC, as for other parametric classifiers, depends on reasonably accurate estimation of the mean vector \( \mu \) and the covariance matrix for each spectral class data (Richards and Xiuping, 1999).

2.4.2. Support Vector Machines (SVMs)

In the last ten years, different approaches as for example, SVMs have been proposed. They are generally non parametric classifiers which show a great ability to optimize classification issues, minimizing the empirical classification errors while maximizing the classification separations. Details are in Section 2.4.2.

SVM models and algorithms are based on supervised learning approaches, also called machine learning, that are considered to be very powerful and widely used within both industry and academia (Dixon and Candace, 2008). Machine learning is a branch of Artificial Intelligence (AI) defined (by Arthur Samuel one of the pioneers of AI) as the field of study that gives to computers the ability to learn without being explicitly programmed or in other words systems that can learn from data. During the last ten years, SVMs have been showing a great

![Fig. 3. Comparison of accuracy parameters.](http://dx.doi.org/10.1016/j.ecoinf.2014.05.005)
potential for data classification also in satellite data processing (Mountrakis et al., 2011) and for these reasons has been adopted in this study. It is expected that the use of SVMs in remote sensing should enable us to successfully handle small training data sets and outperform other more traditional classifiers (Mountrakis et al., 2011). The SVMs are non-parametric classifiers, namely, no prior assumptions are made on the probability distribution function of the data. This characteristic of the SVM classifiers made them very appealing also in remote sensing applications since data acquired from remotely sensed imagery usually have unknown distributions (Mountrakis et al., 2011).

Even though other non-parametric classifiers like Neural Network are likely to work fine as well, several studies demonstrate worse performances in terms of speed of training and accuracy classification (Mountrakis et al., 2011). SVM approaches, as for other supervised learning algorithms, are fed with a training set, consisting of m training samples made up of n vectors of input, named features X_i of m dimension and labels of “correct answers” that we desire the algorithm predicts on that example. In the case of satellite pictures, the training sets are pixels (chosen to be training samples) described as vectors, whereas the labels are the given classes assigned for those pixels.

Let the following be the training set to be classified:

\[
\{ (X_i, Y_i) | X_i \in \mathbb{R}^m, Y_i \in \{0, 1\} \}_{i=1}^n
\]

that is a selection of m pixels, considering n bands, and two classes (0 or 1). The supervised algorithm searches for a function that takes values from \( \mathbb{R}^m \) and goes to Y, here \([0,1]\) being the two possible values. The basic approach to choose the function is a structural risk minimization. This implies to seek the best fitting function for the training data taking into account a term to control bias/variance trade-off.

In the case of SVMs, the minimization issue leads basically to the definition of a set of separating planes, in a \((n-1)\)-dimensional space, with \( n \) being the number of features. Among the possible hyperplanes that can classify data, the choice is done by means of an optimization procedure that leads to maximize the minimum distance (called vector margin) from any of the training samples to the decision boundary (support vector). This is why SVM is also often called “large margin classifier”.

The robustness of SVM is actually due to its slightly more sophisticated characteristics than this large margin view might suggest. In particular, notwithstanding the large margin reliability of the classifier, the algorithm can still be sensitive to outliers, but SVMs offer the chance to control the risk of overfitting, also indicated as bias-variance trade off (Mountrakis et al., 2011), through the tuning of one parameter, called C. If the value of C is set very large, then the learned hypothesis may fit the training set very well, but it will fail to generalize to new examples (predict classes on new pixels); this can be translated also as a hypothesis with higher variance and lower bias. Whereas, it can do a better job ignoring few (eventually present) outliers maintaining C to moderate high values. On the opposite, very low values can add a risk of underfitting.

In its simplest form, SVMs are linear binary classifiers capable to assign the features of an input test sample to one of the two possible classes/labels (Mountrakis et al., 2011). Even though SVMs were originally developed (Cortes and Vapnik, 1995) for binary classification, nowadays many SVM packages already have built-in multi-class classification functionalities, through the one-against-one (or one-against others) technique by fitting all binary sub-classifiers and finding the correct class by a voting mechanism (Meyer, 2012). Whereas concerning the linearity of classifiers, SVMs can be adapted in order to develop complex nonlinear classifiers mainly by means of technique called kernel tricks. In order to solve for the parameters in a computationally efficient way, only kernel functions that satisfy Mercer’s Theorem are valid kernel in SVMs (Scholkopf and Smola, 2001).
The most widely used one is the Gaussian kernel also called Radial Basis Function (RBF). RBF kernel requires only one parameter (generally denoted as Gamma) to be set, which makes it more robust in its implementation in contrast to other kernels (Petropoulos et al., 2010). Generally speaking the Gamma parameter defines the influence that a single training example reaches, i.e. low (high) values of Gamma means far (close). In other words, as Gamma is set to larger values (sigma squared small), Gaussian kernel would tend to fall off abruptly resulting in a function that varies rapidly, and so this will give hypothesis with lower bias and higher variance similar to larger values of C, which trades off misclassification of training examples against simplicity. In the case of C, lower C makes the decision surface smoother, whereas higher C aims at classifying all training examples correctly but must be carefully set to avoid the risk of overfitting.

As a rule of thumb, the use of SVMs with “linear kernel” (that means no kernel) is going to be useful whenever, in presence of a large number of features, \( n \) is large, and the number of training examples \( m \) is small. In fact, trying to fit a very complicated function in a very high dimensional feature space, with a small training set might risk overfitting. Whenever, the number of features is relatively small, \( n \approx 1–1000 \), and the training samples are not very big, \( m \approx 10–10000 \), Gaussian RBF will outperform other classifiers that will result to be slower. Only when the size of samples is about 50,000–100,000, Gaussian RBF will start to become slow.

Other than Gaussian RBF kernel, there are some other less popular kernels, sometimes found in the literature, like polynomial kernel. This kind of kernel almost always or usually performs worse than RBF and high order polynomial becomes easily computational demanding/expensive.

As a whole, a Support Vector Machine is a maximal margin hyperplane in feature space built by using a kernel function in spectral space. The selection of SVM kernel classifier (along with kernel parameters) is considered as one of the most important steps in the implementation of SVM classifier. The lower the number of parameters to be defined, the higher the robustness of the SVM implementation will be. Moreover, the selection of the proper kernel functions or proper parameters of a kernel function is extremely time consuming.

In the following discussion, in order to refer to the parameter C the term Cost is going to be used.
3. Application

The result of classification from MLC has been compared with those obtained from different configurations of SVMs obtained by changing setting parameters, i.e. different kernels and different values for Cost and Gamma coefficients.

In order to perform accuracy evaluation automatically after classifications, a script in the R language has been implemented calling, among others, functions from the package rasclass.

In this paper, a lot of configurations have been tested and the comparisons using accuracy parameters of the most significant configurations have been reported. Moreover a separate evaluation has been made running the script and working on groups of more than one band at a time (the first 3 bands, the first 6 bands and the first 9-bands). The bands with a spatial resolution of 30 m have been re-sampled to 15 m.

3.1. First case: ASTER classification (2 classes using 3, 6 or 9 bands)

Fig. 3 shows the accuracy coefficients obtained considering the classification based on only two classes, urban and non urban areas. The synthesis of accuracy parameters shows that the values of the coefficient $K$ are significantly lower than the overall accuracy (Fig. 3) and no significant differences are evident changing the number of the spectral bands (3, 6 or 9) under processing.

Comparison of the accuracy parameters obtained from MLC and the different configuration of SVMs clearly shows that the best results are obtained from the SVMs compared to MLC (especially in the case of 9-bands). Among the SVM configurations the best results are achieved by using the RBF kernel. Whereas, the number of the processed spectral bands 3, 6 or 9 bands does not significantly impact the accuracy parameters. One of the main problems founded in this case concerns the...
coastal area which tends to be misclassified almost always it is categorized as urban (area included in circle in Fig. 4).

Fig. 5 offers a visual comparison of the results obtained from the classification made using 6 ASTER bands applying both the MLC and SVMs. In this condition, the best performance is provided by SVM with RBF kernel.

Fig. 6 shows a comparison of the resulting maps from MLC applied to 3 bands, 6 bands and 9 bands (respectively, shown from left to right) made re-sampling to 15 m all the satellite pictures. The case of 9-band classification is very interesting, Fig. 6 clearly shows that the MLC classification with 9-band is explicitly worse than the same MLC classification with 3 or 6 band (Fig. 6).

Fig. 7 shows results obtained from the diverse SVM configurations. In this case, using a linear kernel the class "urban" does not cover sand pixels but the classification still contains errors: it is visible that the roads are not identified. The best compromise is obtained using a configuration based on RBF kernel with the Cost parameter equal to 10 or 100 (Fig. 7).

3.2. Second case: ASTER classification (8 classes using 3, 6 or 9 bands)

The classifications with only two classes (urban/not-urban) as output, have been extended to 8 classes (urban, agricultural soils, forests and green areas, coastal seawater, seawater not coastal, bare soil, sand and rock, pure sand). Fig. 8 shows the accuracy parameters obtained using 3, 6 or 9 bands. Figs. 8 shows that using 8 classes the parameter accuracy is in general a little bit improved and the greater number of bands contributes to improve the classification as evident by the increasing values of the accuracy coefficients.
3.2.1. ASTER classification (8 classes using 6 bands)

As shown in Table 3, the agricultural areas have high accuracy (81.82% and coefficient k 71.44%).

Analyzing the results of various configurations, we can argue that none of them fully satisfies the expectations because of the difficulty of discerning the sand from urban areas. Analysing the results of the overall accuracy and kappa coefficients, we can affirm that the SVMs provide higher values for the configurations with RBF kernel. Fig. 10 highlights an area on the coast which is characterized by sand and sand mixed with rock. The MLC on this small area ranking quite well, but the rest of the territory, also non-coastal, is “dirtied” by class sand/rock. Even if the road is not well defined, the best compromise is obtained in the case of SVMs with RBF kernel configuration and Cost = 10; in circled area the sand (yellow) and the mixture of sand and rock (light pink) are identified correctly.

3.2.2. ASTER classification (8 classes using 9 bands)

Table 3 lists the accuracy coefficients which is generated automatically by the Rascall package and provides, for each classifier, information on the accuracy for each class. Analysing, for example the SVM configuration with Cost 10 and considering the class 1 (urban-asphalted areas), low accuracy values can be noticed. In particular, 54 pixels of the class 1 (sample 1) are predicted as class 2 and 21 pixels of the ROI belonging to the agricultural class are predicted as class 1; therefore for the sample 1 a value Producer’s Accuracy equal to 0.932 and User’s accuracy equal to 0.886 has been achieved. There is therefore a problem of misclassification between urban and agriculture classes. (See Tables 1 and 2.)

As shown in Table 3, the agricultural areas have high accuracy values. However, also the urban areas and the coastal zones are well classified using SVMs with RBF kernel and Cost parameter low enough (Fig. 11).

Differently to the classification urban/non-urban, the use of the first 9-bands does provide improvement in the classifications, compared to the use of the first 3 or 6 bands. In addition, good results can be achieved in sand discrimination with RBF kernel and lower Cost value parameter.

Fig. 12 shows how the different configurations of the SVMs give different results. In general, the different land uses are better classified with RBF kernel rather than linear kernel. The most appropriate compromise in the choice of parameters seems to be the SVM with RBF kernel parameter and Cost equal to 10 or 100.

4. Conclusions

This work shows an application of a SVM algorithms for satellite image classification with the final aim of the assessment of urban sprawl completely realized with open-source software. The configurations and processing in this application have been included within a script based on the Grass GIS software and R-statistics and this is very useful in the first phase of testing. Results confirm the effectiveness of SVM algorithms compared to a more “classic” one, as the MLC.

In an area so fragmented the choice of only 2 macro-classes returns satisfactory results. However, better results are obtained using multiple classes, 8 in the examined case. In general, the use of multiple bands, considering that in any case there is a re-sampling to 15 m, improves the classifications. An exception is the use of 9 bands in the case of 2 classes urban/non-urban which does not show better accuracy parameters compared to the use of less bands. One of the critical elements found in this case study is given by the presence of sand and sand mixed with rock.

Often it is this element that does not allow in obtaining a land use classification very “clean”. In this case a very appreciable is the flexibility of the SVMs that, in relation to the possibility of using different kernels and different values of the parameters Cost and Gamma, allows in calibrating the classifier also for a specific need, as in our case, to discern the classes related to the sand from the urban class.

The SVM classifier with RBF kernel is generally the best choice among the configurations being compared, though, with respect to specific land use classes, a given configuration can provide better results than the other. To be able to distinguish the classes sand and urban (and sometimes even the agricultural areas class) the increase in the Cost parameter leads to significant improvements, but it is always wise to keep this value as low as possible in order to control the risk of overfitting.

The result assessment is carried out by analysing the maps of output together with the accuracy values. In this way it is possible to choose a configuration that works for each class of land use or apply a criterion linked to one or more bands at a time. In the first case, by means of map-algebra operations, it is possible to make choices about the individual classes, for example by selecting the urban pixels from the SVM configurations with RBF kernel standard, or the sand pixels along the coast from the SVM configuration RBF kernel and Cost = 10.

Ultimately in the various tests carried out it can be seen that the use of SVMs allow you to have good results for the classification of urban areas, low accuracy values can be noticed. In particular, 54 pixels of the class 1 (sample 1) are predicted as class 2 and 21 pixels of the ROI belonging to the agricultural class are predicted as class 1; therefore for the sample 1 a value Producer’s Accuracy equal to 0.932 and User’s accuracy equal to 0.886 has been achieved. There is therefore a problem of misclassification between urban and agriculture classes. (See Tables 1 and 2.)

As shown in Table 3, the agricultural areas have high accuracy values. However, also the urban areas and the coastal zones are well classified using SVMs with RBF kernel and Cost parameter low enough (Fig. 11).

Differently to the classification urban/non-urban, the use of the first 9-bands does provide improvement in the classifications, compared to the use of the first 3 or 6 bands. In addition, good results can be achieved in sand discrimination with RBF kernel and lower Cost value parameter.

Fig. 12 shows how the different configurations of the SVMs give different results. In general, the different land uses are better classified with RBF kernel rather than linear kernel. The most appropriate compromise in the choice of parameters seems to be the SVM with RBF kernel parameter and Cost equal to 10 or 100.

4. Conclusions

This work shows an application of a SVM algorithms for satellite image classification with the final aim of the assessment of urban sprawl completely realized with open-source software. The configurations and processing in this application have been included within a script based on the Grass GIS software and R-statistics and this is very useful in the first phase of testing. Results confirm the effectiveness of SVM algorithms compared to a more “classic” one, as the MLC.

In an area so fragmented the choice of only 2 macro-classes returns satisfactory results. However, better results are obtained using multiple classes, 8 in the examined case. In general, the use of multiple bands, considering that in any case there is a re-sampling to 15 m, improves the classifications. An exception is the use of 9 bands in the case of 2 classes urban/non-urban which does not show better accuracy parameters compared to the use of less bands. One of the critical elements found in this case study is given by the presence of sand and sand mixed with rock.

Often it is this element that does not allow in obtaining a land use classification very “clean”. In this case a very appreciable is the flexibility of the SVMs that, in relation to the possibility of using different kernels and different values of the parameters Cost and Gamma, allows in calibrating the classifier also for a specific need, as in our case, to discern the classes related to the sand from the urban class.

The SVM classifier with RBF kernel is generally the best choice among the configurations being compared, though, with respect to specific land use classes, a given configuration can provide better results than the other. To be able to distinguish the classes sand and urban (and sometimes even the agricultural areas class) the increase in the Cost parameter leads to significant improvements, but it is always wise to keep this value as low as possible in order to control the risk of overfitting.

The result assessment is carried out by analysing the maps of output together with the accuracy values. In this way it is possible to choose a configuration that works for each class of land use or apply a criterion linked to one or more bands at a time. In the first case, by means of map-algebra operations, it is possible to make choices about the individual classes, for example by selecting the urban pixels from the SVM configurations with RBF kernel standard, or the sand pixels along the coast from the SVM configuration RBF kernel and Cost = 10.

Ultimately in the various tests carried out it can be seen that the use of SVMs allow you to have good results for the classification of urban areas, low accuracy values can be noticed. In particular, 54 pixels of the class 1 (sample 1) are predicted as class 2 and 21 pixels of the ROI belonging to the agricultural class are predicted as class 1; therefore for the sample 1 a value Producer’s Accuracy equal to 0.932 and User’s accuracy equal to 0.886 has been achieved. There is therefore a problem of misclassification between urban and agriculture classes. (See Tables 1 and 2.)

As shown in Table 3, the agricultural areas have high accuracy values. However, also the urban areas and the coastal zones are well classified using SVMs with RBF kernel and Cost parameter low enough (Fig. 11).

Differently to the classification urban/non-urban, the use of the first 9-bands does provide improvement in the classifications, compared to the use of the first 3 or 6 bands. In addition, good results can be achieved in sand discrimination with RBF kernel and lower Cost value parameter.

Fig. 12 shows how the different configurations of the SVMs give different results. In general, the different land uses are better classified with RBF kernel rather than linear kernel. The most appropriate compromise in the choice of parameters seems to be the SVM with RBF kernel parameter and Cost equal to 10 or 100.
areas with few training sites and, consequently, this decreases the error that can be made in the identification of these sites. Moreover, the use of the SVM algorithms requires processing time much faster than the MLC and, especially in the case of classifications of areas with complex land-use, it is better able to discriminate between classes of objects with similar spectral signature.

Fig. 11. Monopoli (BA) — SVMs RBF Cost = 10.

Fig. 12. Monopoli (BA) (clockwise from top) orthophotos, Linear SVM, SVM RBF, SVM RBF Cost = 100 Gamma = 1. Even in the case of using 9-bands the best values of K and OA are related to the SVMs with the RBF kernel.

Please cite this article as: Nolè, G., et al., Evaluation of urban sprawl from space using open source technologies, Ecological Informatics (2014), http://dx.doi.org/10.1016/j.ecoinf.2014.05.005
The ASTER data and the application of the SVM algorithms are adequate for a large-scale study and the choice of free/open source technologies does not put technological limits for replicability of classifiers in other contexts.

References


Romano, B., Zullo, F., 2014. The urban transformation of Italy’s Adriatic Coast Strip: fifty years of unsustainability. Land Use Policy 38, 26–36.


