Badly Posed Classification of Remotely Sensed Images—An Experimental Comparison of Existing Data Labeling Systems

Andrea Baraldi, Lorenzo Bruzzone, *Senior Member, IEEE*, Palma Blonda, *Member, IEEE*, and Lorenzo Carlin

***Abstract*—Although underestimated in practice, the small/unrepresentative sample problem is likely to affect a large segment of real-world remotely sensed (RS) image mapping applications where ground truth knowledge is typically expensive, tedious, or difficult to gather. Starting from this realistic assumption, subjective (weak) but ample evidence of the relative effectiveness of existing unsupervised and supervised data labeling systems is collected in two RS image classification problems. To provide a fair assessment of competing techniques, first the two selected image datasets feature different degrees of image fragmentation and range from poorly to ill-posed. Second, different initialization strategies are tested to pass on to the mapping system at hand the maximally informative representation of prior (ground truth) knowledge. For estimating and comparing the competing systems in terms of learning ability, generalization capability, and computational efficiency when little prior knowledge is available, the recently published data-driven map quality assessment (DAMA) strategy, which is capable of capturing genuine, but small, image details in multiple reference cluster maps, is adopted in combination with a traditional resubstitution method. Collected quantitative results yield conclusions about the potential utility of the alternative techniques that appear to be realistic and useful in practice, in line with theoretical expectations and the qualitative assessment of mapping results by expert photointerpreters.**

***Index Terms*—Badly posed classification, competing classifier evaluation, clustering, curse of dimensionality, generalization capability, image labeling, inductive learning, map accuracy assessment, remotely sensed (RS) imagery, semilabeled samples, semisupervised learning, supervised learning, unsupervised learning.**

I. Introduction

In recent years, there has been a great development of new methods for (unsupervised) data clustering and (supervised) data classification in the image processing, pattern recognition, data mining, and machine learning literature [1]–[12]. Unfortunately, owing to their functional, operational, and computational limitations, most data classification techniques, both supervised and unsupervised, have had a minor impact on their potential field of applications [13]–[15]. For example, in remote sensing (RS) image understanding, we have the following.

1) An enhanced ability to detect genuine but small image structures, especially man-made objects, would increase the impact of data labeling methods in cartography, urban planning, and analysis of agricultural sites.
2) Improved data-driven learning capabilities (e.g., multiscale parameter estimate) would make image labeling algorithms easier to use, more robust with respect to noise and changes in input parameters, and more effective when little ground truth (prior) knowledge is available.
3) Computationally more efficient (e.g., noniterative) image analysis algorithms and architectures should be made available when training and processing time may still be considered a burden, e.g., in classification tasks at continental or global scale [16].

In recent years, the second potential improvement listed above has gained increasing importance as RS image understanding has had to come to grips with tremendous spatial and spectral complexity, such as with second- and third-generation satellite imagery (where spatial resolution may remain below one meter, while hyperspectral data may include hundreds of image bands). Thus, it has become increasing difficult, expensive (e.g., in hyperspectral images), and/or tedious (e.g., in high spatial resolution images) to collect independent reference samples having statistical properties appropriate for first generation classifiers [e.g., maximum *a posteriori* (MAP), mixture models, etc.] [2], [24]. In this image classification scenario, the well-known *small training sample size problem* (also called *Hughes phenomenon* or *curse of dimensionality*) is likely to occur, which causes inductive learning systems to be affected by poor generalization capability [2], [20]–[23]. Although well-studied in existing literature, the Hughes phenomenon is often underestimated in practice. This underestimation becomes even more severe in the field of RS image understanding, where the spatial autocorrelation reduces the informativeness of neighboring pixels by violating the assumption of sample independence [21], which may give rise to the so-called *unrepresentative sample problem*. A possible taxonomy of the bad-conditioning degree of predictive learning problems is proposed as follows (extending concepts proposed in [2]).

- *Ill-posed predictive learning problems*: where data dimensionality exceeds the total number of (independent) representative samples and, as a consequence, is much greater than the number of per-class representative samples.

Manuscript received April 15, 2004; revised June 28, 2005. This work was supported by the European Union under Contract EVG1-CT-2001-00055 LEWIS.

A. Baraldi is with the European Commission Joint Research Centre, I-21020 Ispra (Va), Italy (e-mail: andrea.baraldi@jrc.it).

L. Bruzzone and L. Carlin are with the Department of Information and Communication Technologies, University of Trento, I-38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

P. Blonda is with the Istituto di Studi su Sistemi Intelligenti per l'Automazione, Consiglio Nazionale delle Ricerche (ISSIA-CNR), 70126 Bari, Italy (e-mail: blonda@ba.issia.cnr.it).

Digital Object Identifier 10.1109/TGRS.2005.859362

- *Poorly posed predictive learning problems*: where data dimensionality is greater than or comparable to the number of (independent) per-class representative samples, but smaller than the total number of representative samples.

To mitigate the Hughes phenomenon, several data preprocessing and classification strategies can be adopted and, eventually, combined: 1) input space dimensionality can be reduced (feature extraction/selection); 2) robust statistics estimate techniques (e.g., covariance matrix regularization) can be applied in combination with first generation classifiers which in general are context-insensitive, i.e., not specifically developed to deal with two-dimensional (2-D) images[1] [20]–[22], [25] (also refer to further Section IV-A); and 3) a new (second) generation of context-sensitive (single- or multiscale) inductive learning classifiers suitable for dealing with a lack of reference samples in image data.

This work focuses on the third aforementioned issue. Conceived as a significant extension of three related papers [17]–[19], this paper compares a set of (five) advanced *semisupervised* data labeling systems (described in Appendix I) against a set of (nine) standard classifiers in the badly posed classification of RS images. Standard classifiers are selected from commercial image processing software toolboxes, like the nearest prototype (NP) and maximum-likelihood (ML) classifier, or among researchers and practitioners based on their degree of popularity, like the probabilistic neural network (PNN), multilayer perceptron (MLP), and support vector machine (SVM), to cover a wide range of well-known inductive learning principles and optimization algorithms.

It is worthwhile to note that this paper provides no original contribution in image classification system design. Rather, it compares twice as many classifiers (namely, 14 systems implemented in 20 versions) as its most closely related paper [19].

In line with Zamperoni's recommendations [13], the proposed classification assessment and comparison strategy seems a reasonable approximation of the operational characteristics of a large segment of real-world applications in RS image understanding where ground truth knowledge, if any, is expensive, tedious, or difficult to gather. This realistic framework allows to assess whether an existing classifier appears to be worthy of dissemination in commercial data/image processing all-purpose software toolboxes, in that it is presumably useful to a broad audience dealing with image/pattern recognition problems in general, with special emphasis on badly posed RS image labeling applications.

The rest of this paper is organized as follows: a taxonomy of statistical pattern recognition systems capable of learning from finite data is reviewed and a problem of terminology and notation is introduced in Section II. Section III provides a background in RS reference sample selection strategies capable of mitigating the small sample size problem. In Section IV, a set of existing data mapping systems is selected from existing literature for comparison purposes. The experimental session design is the subject of Section V. Experimental results are discussed in Section VI. Conclusions are reported in Section VII. To make this paper self-contained and provided with significant survey value, a synthesis of the selected nonstandard data labeling methods is proposed in Appendix I.

[1]Context-sensitive data mapping algorithms, either single- or multiscale, are specifically developed for 2-(spatial) dimensional image mapping tasks, whereas sample-based data mapping algorithms, employing no contextual information, are applicable to any 1-(spatial) dimensional sequence of multivariate input vectors.

II. Predictive Learning System Terminology and Notation

To make the conceptual framework of this work clear and explicit to RS experts and practitioners, this section provides some definitions adapted from pattern recognition and machine learning literature [26].

Classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [19]–[22]. In *supervised classification*, an observed set of unlabeled samples, $z = \{z_1, \ldots, z_N\}$, where vector $z_j \in \Re^D$, $j \in \{1, N\}$, such that D is the input space dimensionality and N is the total number of unlabeled samples, is mapped to a finite set of discrete class labels, $\{1, L\}$, where L is the total number of labels, such that $x = \{x_1, \ldots, x_N\}$ is an arbitrary labeling of the unlabeled dataset z. This discrete mapping is modeled in terms of a mathematical function $x_j = g(z_j, F)$, $j \in \{1, N\}$, where F is a vector of adjustable (free) parameters. The values of these parameters are determined (optimized) by an inductive learning algorithm (also termed *inducer*), whose aim is to minimize an *empirical risk functional* (related to an inductive principle) on a finite *reference dataset* of supervised samples (input-output examples selected by an external agent, supervisor, or oracle) $y_{i,k}$, $i \in \{1, L\}$, $k \in \{1, m_i\}$, where m_i is the number of labeled samples belonging to class i, assumed to be correct, such that $m_1 + \ldots + m_L = M$, where M is the reference sample set cardinality. In general, inequality $M \ll N$ holds [20]–[22], [27]. When the inducer reaches convergence or terminates, an *induced classifier* is generated [26], [28]. When a data mapping system provides an unlabeled sample z_j, $j \in \{1, N\}$, with a hard (crisp) implicit label $i \in \{1, L\}$, then the unlabeled sample becomes a *semilabeled sample*, identified as $z_{i,j}$. If n_i represents the cardinality of the set of semilabeled samples provided with implicit label i, then $n_1 + \ldots + n_L = N$. Thus, semilabeled samples are as many as the unlabeled samples and available at no extra classification cost. Since inequality $N \gg M$ typically holds, semilabeled samples are employed by the so-called *semisupervised learning algorithms* to mitigate the small/unrepresentative sample problem [2]–[4]. Implicit class labels, provided by the classifier, may be incorrect. The probability that an implicit label is incorrect is determined by the off-training set (generalization) error rate of the classifier [28]. On the contrary, explicit class labels of reference samples, provided by an external supervisor, are assumed to be (hardly, crisply) correct.

In *unsupervised classification*, called *clustering* or *exploratory data analysis* [21], no labeled data are available. The goal of clustering is to separate a finite unlabeled dataset into a finite and discrete set of "natural," hidden data structures [2], [20]. According to Backer and Jain [29], "in cluster analysis a group of objects is split up into a number of more or less homogeneous

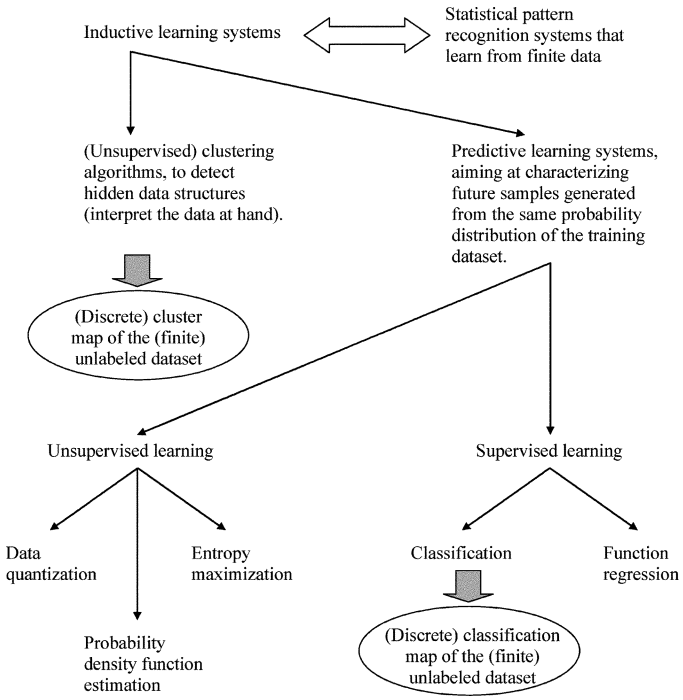


Fig. 1. Proposed taxonomy of statistical pattern recognition systems. Clustering and classification systems map an unlabeled input sample to a discrete label. These maps are called cluster maps and classification maps, respectively.

subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups." Since the goal of clustering is to group the data at hand rather than provide an accurate characterization of unobserved (future) samples generated from the same probability distribution, the task of clustering may fall outside the framework of unsupervised predictive learning problems, such as vector quantization [20], probability density function estimation [20]–[22], and entropy maximization [30], each of these classes of applications featuring a specific cost function. However, in practice, vector quantizers are also used for cluster analysis [21, p. 177]. To summarize, two main concepts are involved in estimating the accuracy of both unsupervised and supervised data labeling systems.

1) First, the subjective nature of the nonpredictive clustering problem precludes an absolute judgement as to the relative effectiveness of all clustering techniques [2], [3].
2) Second, it is well-known that "if the goal is to obtain good generalization performance in predictive learning, there are no context-independent or usage-independent reasons to favor one learning or classification method over another" [22, p. 454].

In line with Fig. 1, the rest of this paper deals exclusively with statistical pattern recognition systems capable of generating (2-D, discrete) maps starting from RS images, i.e., induced classifiers and clustering algorithms whose mapping results are called *classification maps* and *cluster maps*, respectively [31]. It is to be kept in mind that, hereafter, the generic term "map" is adopted to identify both cluster and classification maps.

III. Empirical Rules to Avoid the Badly Posed Classification of RS Images

In recent years, enhanced spectral, temporal, and spatial resolutions of RS sensors have increased the number of detectable land cover classes. These developments have dramatically increased the size of ground truth regions of interest (ROIs) required to be representative of the true class-conditional distributions. Hence, heuristic rules are traditionally adopted in both training and testing phases to avoid the reference data resampling scheme affected by bad-posedness.

Training Phase:

- Let us assume that $\sum_{i=1}^{L} m_i = M_{\text{train1}}$, where m_i is the number of independent training samples belonging to each class $i = 1, \ldots, L$. To avoid the curse of dimensionality, general rules of thumb to constrain m_i from below are the following.
  - —$m_i, i = 1, \ldots, L$, should be approximately proportional to the prior probability of that class if a maximum *a posteriori* (MAP) classification rule is adopted.
  - —m_i, $i = 1, \ldots, L$, should be capable of representing all possible variations in spectral response in each land cover type of interest.
  - —$m_i \in \{5{*}D, 100{*}D\}$ [23], [32], [33]. For example, this rule of thumb ensures an adequate estimation of nonsingular/invertible class-specific covariance matrices (typically required by some first generation classifiers, e.g., the maximum likelihood classifier). In fact, a class-specific covariance matrix in a D-dimensional feature space consists of $D(D+1)/2$ parameters to estimate and there must be at least $D+1$ observations of each class to ensure estimation of nonsingular-invertible class-specific covariance matrices [23].
  - —$m_i \geq 0 \div 0$, so that, according to a special case of the central limit theorem, the distribution of many sample statistics becomes approximately normal, which is a basic assumption employed by several traditional classifiers [24], [31].
- To avoid poor generalization capability of an induced classifier related to model complexity, the minimum number of per-class representative samples should be proportional to the number of the learning system's free parameters to be optimized during training. For example, an approximate worst-case limit on generalization is that correct classification of a fraction 1-ε of new examples requires a number of training patterns at least equal to $M_{\text{train2}} \approx F/\varepsilon$, where F is the total number of the system free parameters. If $\varepsilon = 0.1$, we need around ten times as many training patterns as there are free parameters in the learning system [20].
- As a corollary, if a holdout resampling method is adopted for the assessment of the generalization capability of competing classifiers where, typically, 2/3 of the available labeled dataset should be used for training and the remaining 1/3 for testing [28], then the reference sample set size becomes

$M_{\text{holdout}}(M_{\text{train}}) = (3/2) * M_{\text{train}}$, where $M_{\text{train}} = \max\{M_{\text{train1}}, M_{\text{train2}}\}$.

Testing phase:

- It is well known that any classification accuracy (precision) probability estimate $\hat{p}_c$ is a random variable (sample statistic) with a confidence interval (error tolerance) associated with it, i.e., it is a function of the specific training and testing sets being used [32]. The maximum-likelihood classification accuracy estimate $\hat{p}_c = c/M_{\text{test}}$, where c is the number of correctly classified samples out of M_{test} testing samples, is an unbiased and consistent estimator. The probability density function of c has a binomial distribution. When $M_{\text{test}} \geq 0$ (large testing sample set), a binomial sampling can be well approximated with a standardized normal distribution featuring mean $= M_{\text{test}} \cdot p_c$ and standard deviation $= sqrt(M_{\text{test}} \cdot p_c \cdot (1 - p_c))$. Thus, the testing sample set size M_{test} needed to estimate a specified classification accuracy probability p_c with a given error tolerance of $\pm\delta$ at a desired confidence level (e.g., if confidence level $= 95\%$ then the critical value is 1.96) becomes [18]

$$M_{\text{test}} = \frac{(1.96)^2 \cdot p_c \cdot (1 - p_c)}{\delta^2}.$$

  For example, if $p_c = 0.75$ with $\delta = 0.06$, then $M_{\text{test}} = 200$. If $p_c = 0.75$ with a confidence interval (error tolerance) $\delta = \pm 0.10$, then $M_{\text{test}} = 75$. If $p_c = 0.50$ with $\delta = 0.06$, then $M_{\text{test}} = 266$. If $p_c = 0.50$ with $\delta = 0.10$, then $M_{\text{test}} = 100$. It can be shown that [34]
  - —For a fixed precision level p_c, if δ increases then the number of required samples M_{test} decreases.
  - —Although it seems counterintuitive, if the confidence interval for all levels of precision p_c is fixed, then the number of reference samples required to achieve $p_c = 0.5$ (i.e., when the sample population is evenly split between the two classes) is much higher than when the level of precision p_c tends to 1 (i.e., when the sample population moves toward a dominant and rare two-class composition).
- As a corollary, if a holdout resampling method is adopted where 2/3 of the available labeled dataset is used for training [28], then the reference sample set size becomes $M_{\text{holdout}}(M_{\text{test}}) = 3 * M_{\text{test}}$, with M_{test} computed according to the foregoing equation.

Conclusion: Based on the aforementioned corollaries, if a holdout resampling method is adopted for the assessment of the generalization capability of competing classifiers, the overall size of the recommended reference dataset becomes $M_{\text{holdout}} = \max\{M_{\text{holdout}}(M_{\text{train}}), M_{\text{holdout}}(M_{\text{test}})\}$.

IV. Set of Classifiers to be Compared

Existing data labeling systems can be partitioned into different categories on the basis of their different functional/learning properties. For example, data labeling systems can be [20]–[22], [27], [32], [35]: 1) context-sensitive (i.e., specialized in dealing with images, based on either single-scale or multiscale image analysis mechanisms) and sample- (e.g., pixel-) based; 2) based on supervised learning (classifiers), unsupervised learning (clustering methods), and semisupervised learning (see Section II); 3) parametric and nonparametric (also called memory-based [27], whose computational complexity increases with the cardinality of the representative dataset); and 4) adaptive and nonadaptive (also called plug-in, where class-conditional parameters are estimated off-line, prior to the classification stage, by an external analyst [32]).

According to this terminology (refer to Fig. 1), adaptive labeling approaches are either supervised [e.g., multilayer perceptron (MLP), radial basis function (RBF) [36]] or unsupervised (e.g., hard C-means (HCM) [20], [21]), while plug-in approaches (like the Bayesian plug-in classifiers, either Bayes-normal-quadratic or Bayes-normal-linear [32], [37]) must be supervised and parametric.

In this section, a set of standard classification algorithms, well known to practitioners or adopted by commercial data processing software toolboxes to cover a wide range of alternative inductive learning principles and optimization procedures [38], and a set of nonstandard classification algorithms specifically developed to mitigate the small sample size problem are selected for comparison purposes.

A. Advanced Data Labeling Techniques

One possible solution to badly posed data classification consists of exploiting *semilabeled samples* (i.e., unlabeled samples after classification; refer to Section II) in adapted versions of the well-known iterative expectation–maximization (EM) maximum-likelihood (ML) estimator, like the sample-based (i.e., context-insensitive; refer to footnote 1) Semisupervised EM (SEM) classifier [2], which is theoretically well-founded,[2] and its heuristic context-sensitive single-scale version (i.e., suitable for dealing with images), hereafter referred to as contextual SEM (CSEM) [39]. More specifically, SEM's and CSEM's parameter estimation strategies combine the small set of labeled data, whose *explicit class labels* feature full weights, with a large set of semilabeled samples provided with reduced weights. Unfortunately, it is well known that when the normal model distribution estimated by the iterative (suboptimal) semisupervised mapping algorithms with EM does not match the true underlying distribution, the large amount of unlabeled data may have an adverse effect on classifier performance on labeled samples (i.e., while pursuing cost function reduction, these algorithms do not guarantee a better error rate for labeled samples in the next iteration) [41]. In practice, even though they rely on heavy class-specific normal density assumptions that may be untrue in many real-world image mapping problems (e.g., in highly textured images) and may require supervision to separate multimodal classes into unimodal subclasses [33], semisupervised classifiers based on the iterative EM Gaussian mixture model solution can be very powerful and easy to use [2], [19], [39].

In badly posed classification problems featuring a very high input space dimensionality D (ranging from 10 up to a few hundreds [25], e.g., in the case of RS hyperspectral data),

[2]According to [2], a (suboptimal) iterative predictive learning classifier is defined as theoretically well-founded if it is guaranteed to reach convergence at a (local) minimum of a known cost function.

TABLE I
TAXONOMY OF DATA MAPPING SYSTEMS ADOPTED FOR COMPARISON. LEGENDA: Y: YES, N: NO. SINGLE-SCALE°: MRF-BASED 8-ADJACENCY NEIGHBORHOOD. P: PARAMETRIC. M: MEMORY-BASED. NP: NONPARAMETRIC. ICM: ICM-BASED. EM: EM-BASED. ^: SYSTEM IMPLEMENTATIONS NOT CONSIDERED FOR COMPARISON IN [19]

Algorithm	Learning mode		Parametric vs non-parametric memory-based	Plug-in	Iterative	Soft/Crisp competitive learning	Context-sensitive
	Sup.	Unsup.					
NP	*	-	P	Y	N	-	N
ML	*	-	P	Y	N	-	N
PNN^	*	-	M-NP	Y	N	-	N
MLP^	*	-	NP	N	Y	-	N
SVM-OAA^	*	-	NP	N	Y	-	N
SVM-OAO^	*	-	NP	N	Y	-	N
ICM-MAP-MRF	*	-	P	N	Y-ICM	Crisp	Single-scale°
EM^	-	*	P	N	Y-EM	Soft	N
CEM^	-	*	P	N	Y-EM	Soft	Single-scale°
SEM	SEM2A, SEM2B^	SEM1	P	N	Y-EM	Semilabeled	N
CSEM	CSEM2A, CSEM2B^	CSEM1	P	N	Y-EM	Semilabeled	Single-scale°
MSEM	MSEM2A, MSEM2B^	MSEM1	P	N	Y-EM	Semilabeled	Multi-scale
MPAC	-	*	P	N	Y-ICM	Crisp	Multi -scale
MPACB^	-	*	P	N	Y-ICM	Crisp	Multi-scale

semilabeled samples alone may not be sufficient to reduce the variance of the covariance matrix estimation process where the number of free parameters increases dramatically (approximately to $0.5 * D^2$). In such cases, recent works recommend class-specific leave-one-out regularized covariance (LOOC) estimators initialized by training samples only. Next, these LOOC estimators are continuously updated using both semilabeled and training samples until a convergence is reached when a quadratic ML classification output changes very little [25]. Unfortunately, when the number of competing classifiers increases, the computational cost of leave-one-out estimation methods may soon become unaffordable (also refer to Section VI-B). Besides, this robust covariance matrix estimate is combined with a sample-based quadratic ML classifier which is not designed to specifically deal with images (i.e., it is neither multi- nor single-scale).

Unlike SEM and CSEM, which model class-specific distributions as Gaussian probability density functions (pdfs), the modified Pappas adaptive clustering (MPAC) algorithm, proposed as a multiscale adaptation of an original single-scale contextual clustering by Pappas [6], exploits contextual information in: 1) a multiscale class-conditional intensity average estimation (less sensitive than variance to the small sample size problem), and 2) an MRF-based regularization term to smooth the solution while preserving genuine but small image regions [7]. To avoid MPAC's tendency to generate artifacts while reducing the risk of filtering out genuine but small image regions, a modified version of MPAC, called MPAC with Backtracking (MPACB), was presented in [17].

Potentially superior to sample-based SEM and context-sensitive single-scale CSEM in detecting genuine but small image details, an original multiscale heuristic combination of the SEM classifier with MPAC, identified as multiscale SEM (MSEM), was recently proposed in image processing literature [19]. The capability of mitigating the small sample size problem while employing multiscale image analysis mechanisms makes MSEM potentially capable of detecting genuine, but small, structures in piecewise constant or slowly varying multispectral images when little prior knowledge is available. Thus, the potential applicability domain of MSEM is expected to range from, say, mapping the RS satellite imagery featuring low $(\approx 1$ km$)$ to medium $(\approx 30$ m$)$ spatial resolution that has been collected in massive amounts in recent years.

It is noteworthy that to date any multiscale image analysis adaptation (either empirical or well-founded) of existing sample-based semisupervised classification schemes potentially superior to SEM (like the recently published cost-effective semisupervised classifier CES^{2}C conceived as a semisupervised adaptation of the kernel Fisher's discriminant [41]) appears as an open problem of difficult solution. This is the case, for example, of the context-insensitive CES^{2}C classifier whose three system parameters (namely, the single-scale spread σ of Gaussian kernels in a vector space, a regularization term γ and a weighting coefficient C in a two-term cost function) are to be user-defined or estimated by cross-validation over the supervised training dataset.

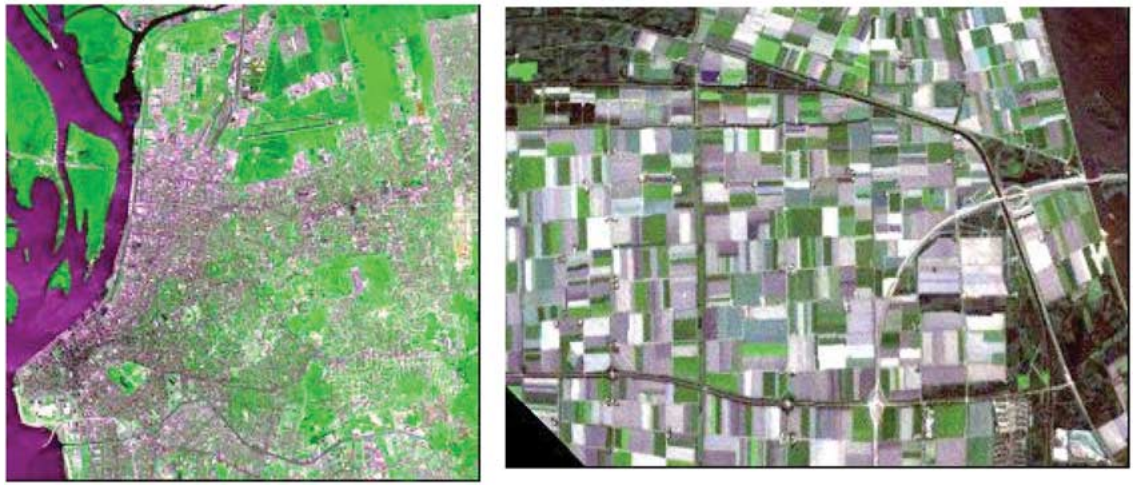
To summarize, MPAC, MPACB, SEM, CSEM, and MSEM are the advanced data labeling systems selected for comparison purposes. It is to be noted that all these systems, with the exception of SEM, are context-sensitive (either single- or multiscale).

B. Standard Data Labeling Techniques

Alternative standard data labeling techniques, well known to practitioners and/or implemented in commercial data processing software toolboxes, are selected for comparison purposes. These systems are: the PNN classifier [42], the MLP,[3] and SVM (in the one-against-all (OAA) and the one-against-one version (OAO),[4] the iterative conditional mode (ICM)-based MAP-

[3]Downloaded from http://fuzzy.cs.uni-magdeburg.de/~borgelt/mlp.html.

[4]Downloaded from http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM-OAA_LIGHT/SVM-OAA_light.html.



(a) (b)

Fig. 2. (a) Test case 1. False-color composition (B: VisBlue, G: NearIR, R: VisRed) of the SPOT image of Porto Alegre, Brazil, 512 × 512 pixels in size, three-band, 20-m spatial resolution, acquired on November 7, 1987. (b) Test case 2. True-color composition (B: VisBlue, G: VisGreen, R: VisRed) of the seven-band Landsat TM image provided by the GRSS Data Fusion Committee, 750 × 1024 pixels in size, 30-m spatial resolution.

Markov random field (MRF) classifier [43], the NP classifier (also called mimimum-distance-to-mean classifier [33]) and Gaussian ML classifier [20], the EM algorithm for density function estimation [20], [44], and the contextual EM (CEM) algorithm for image segmentation [45].

To summarize, 14 alternative data labeling approaches, implemented in 20 versions (refer to further Section VI-C), are selected to cover a wide range of inductive learning principles and optimization algorithms, as shown in Table I.

V. Badly Posed Image Classification Session Design: Test Images and Evaluation Measures

In line with [17]–[19], a realistic experimental framework is set up to adequately estimate and compare competing classification and clustering algorithms in badly posed image classification tasks. Thus, a test set of real and standard RS images provided with little representative ground truth regions of interest, a battery of measures of success and an ensemble of existing data mapping algorithms are selected for comparison purposes [35]. Starting from standards long established in natural and engineering sciences holding that only validated claims are published in journals, rather mild algorithm benchmarking rules proposed in computer science literature are the following: 1) at least two real and standard/appropriate datasets must be adopted to demonstrate the potential utility of an algorithm; 2) the proposed algorithm must be compared against at least one existing technique; and 3) at least one fifth of the total paper length should be devoted to evaluation [51].

A. Test Set of RS Images

According to [35], a test set of RS images suitable for comparing the performance of algorithms employed in image understanding tasks should be: 1) as small as possible; 2) consistent with the aim of testing; 3) as realistic as possible; and 4) such that each member of the set reflects a given type of image encountered in practice.

In line with [18] and [19], the test set of RS images consists of two satellite images, characterized by different sizes and spectral space dimensionality, fragmentation (i.e., visual complexity, related to the presence of genuine but small image details), and levels of prior knowledge, ranging from ill to poorly posed. The raw image adopted in test case 1 is shown in Fig. 2(a). This is a three-band SPOT image of the city area of Porto Alegre (Brazil), 512 × 512 pixels in size, featuring a spatial resolution of 20 m [17]. The image employed in test case 2 is shown in Fig. 2(b). It is a six-band Landsat Thematic Mapper (TM) image, 750 × 1024 pixels in size, with a spatial resolution of 30 m, depicting a country scene in Flevoland (The Netherlands). This second test image is extracted from the standard grss_dfc_0004 dataset provided by the GRSS Data Fusion Committee.[5] In visual terms, the presence of nonstationary image structures, such as step edges and lines, combined with many genuine but small image details, makes the town scene more fragmented than the country scene.

Both test images are considered as piecewise constant or slowly varying intensity images featuring scarcely useful texture (correlation) information, i.e., ground truth ROIs localized and identified on test cases 1 and 2 correspond to spectrally, rather than texturally, uniform areas of interest. Moreover, in both test cases, each ground truth ROI identifies a distinct surface class of interest (which is a rather common practice in real-world RS applications[6]). Twenty-one ROIs/classes are identified on Fig. 2(a) [see Table II(a)], and 12 ROIs/classes are identified on Fig. 2(b) [see Table II(b)], respectively. It is noteworthy that according to Sections I and III, if class-specific mean and covariance matrix parameters are to be employed, then test problem 1, where the minimum number of independent identically distributed (i.i.d.) samples per class would be $10\times$ number of free parameters $\geq (10\times(D+0.5D^2+0.5D)) = (10\times(3+0.5\cdot 9+0.5\cdot 3) = 90$, is rather ill-posed, whereas test problem 2, where $10\times$number of free parameters$\geq (10\times(D+$

[5]http://www.dfc-grss.org

[6]See http://fuzzy.cs.uni-magdeburg.de/~borgelt/mlp.html

TABLE II
(a) TEST CASE 1. TWENTY-ONE ROIS SELECTED ON THE SPOT IMAGE DEPICTED IN FIG. 2(a). (b) TEST CASE 2. TWELVE ROIS SELECTED ON THE LANDSAT IMAGE DEPICTED IN FIG. 2(b)

ROI	Surface type	No. of pixels (%)
1	*dark artificial target 1*	11 (0.004)
2	*dark artificial target 2*	8 (0.003)
3	*bright artificial target*	9 (0.003)
4	*bridge*	21 (0.008)
5	*road 1*	23 (0.008)
6	*road 2*	35 (0.013)
7	*airport*	105 (0.040)
8	*sea port*	21 (0.008)
9	*building 1*	18 (0.006)
10	*building 2*	17 (0.006)
11	*building 3*	24 (0.008)
12	*vegetated area 1*	21 (0.008)
13	*vegetated area 2*	89 (0.033)
14	*vegetated area 3*	155 (0.059)
15	*vegetated area 4*	124 (0.047)
16	*vegetated area 5*	66 (0.025)
17	*grassland*	26 (0.009)
18	*baresoil*	47 (0.017)
19	*marine water 1*	141 (0.053)
20	*marine water 2*	304 (0.115)
21	*marine water 3*	63 (0.024)
TOTAL		1328

(a)

ROI	Surface type	No. of pixels (%)
1	*arable land 1*	689 (0.089)
2	*arable land 2*	571 (0.074)
3	*arable land 3*	787 (0.102)
4	*arable land 4*	238 (0.030)
5	*arable land 5*	1210 (0.157)
6	*vegetated agricultural area 1*	245 (0.031)
7	*vegetated agricultural area 2*	415 (0.053)
8	*vegetated agricultural area 3*	620 (0.080)
9	*vegetated agricultural area 4*	128 (0.016)
10	*scrub 1*	320 (0.041)
11	*scrub 2*	308 (0.040)
12	*marine water*	14900 (1.93)
TOTAL		20431

(b)

$0.5D^2+0.5D)) = (10\times(7+0.5\cdot 49+0.5\cdot 7)) = 350$, is rather poorly posed (eventually ill-posed if the image autocorrelation, superior to that in test case 1 which features finer spatial details, were considered in violating the hypothesis of i.i.d. samples).

The complexity of the classification problem is also increased by the partial overlap between spectral signatures. In test case 1, the minimum Jeffries–Matusita (JM) distance between ROI pairs [31], JM $\in [0, 2]$, is that between classes *vegetated area* 1 and *vegetated area 2*, equal to 0.50. In test case 2, the minimum JM distance is that between classes *scrub 1* and *scrub 2*, equal to 1.80.

B. Set of Measures of Success

It is well known that traditional supervised methods for estimating and comparing classifiers which employ a representative dataset are heuristic in nature. In the words of Duda *et al.*: "indeed, if there were a foolproof method for choosing which of two classifiers would generalize better on an arbitrary new problem, we could incorporate such a method into the learning ... Estimating the final generalization performance invariably requires making assumptions about the classifier or the problem at hand or both, and can fail if the assumptions are not valid ... Occasionally our assumptions are explicit (as in parametric models), but more often than not they are implicit and difficult to identify or relate to the final estimation (as in empirical methods)" [22, p. 482].

When the small/unrepresentative sample problem occurs, then we have the following.

- If the training set is small, then the induced classifier is not going to be robust (to changes in the training set) and will have a low generalization capability.
- When the test set is small, then the confidence in the estimated error rate is going to be low [32].

In general, when the small/unrepresentative sample problem occurs, traditional classification error estimation methods soon become unsuitable [20], [21], [27], [32]. In particular, we have the following.

- The resubstitution method increases its optimistic bias with the small sample size. For example, in [17], the resubstitution error was not in line with qualitative results by expert photointerpreters.
- The holdout method is inefficient in exploiting the available dataset for training, i.e., it is unfitted to deal with the small sample size problem.
- The leave-one-out method has a high computational cost even when the classification problem is badly posed. When the number of competing classifiers increases, the computational cost of the leave-one-out method may soon become unaffordable.
- In the n-fold cross validation, the computational load, which increases linearly with the number of competing classifiers by an n factor, may soon become unaffordable.
- The bootstrap method has the highest computational cost, which soon becomes prohibitive with the number of competing classifiers.

Last but not least, none of these reference dataset resampling methods allows estimation of the spatial distribution of classification errors (known as location accuracy [18]).

To avoid the aforementioned limitations of traditional resampling techniques, the recently published data-driven map accuracy assessment (DAMA) strategy can be employed to mitigate, with a mimimum of human intervention, the small and unrepresentative sample problems in estimating and comparing competing classifiers [18]. In general, DAMA provides a guideline in assessing the labeling and spatial fidelities of a map (under investigation) to a set of cluster maps capable of capturing genuine, but small, image details. By definition, multiple cluster maps are generated from separate clustering of nonoverlapping candidate representative subsets of the original (raw, input) image. In test cases 1 and 2, candidate representative areas are (subjectively) selected as, respectively, three image subsets of Fig. 2(a) (100×300 pixels each), and two image subsets of Fig. 2(b) (400×400 pixels each) (for implementation details; refer to [18]). In combination with the unsupervised DAMA strategy, additional measures of classification success can be conveniently computed in badly posed image classification problems, such as test cases 1 and 2. Since ground truth ROIs are available and fully employed for training the inducer, a confusion matrix, computed between the output map and

the available representative dataset, allows estimation of the so-called *resubstitution error* (upon the training dataset). If the resubstitution (learning) error is small, then bias is low, which means that the prior knowledge has been successfully passed on to the image mapping system. This is a desirable and necessary condition to keep the combination of bias with variance low [20].

A fourth feature that may be considered important in the assessment of competing classifiers is computation time, which affects the application domain of RS image mapping systems [17], and may determine whether or not an algorithm is capable of enriching a commercial image processing software toolbox, as required by Zamperoni [13].

C. Initialization Strategies

In order to guarantee a fair comparison between competing image mapping systems, prior knowledge, having the initial form of ground truth ROIs, must adapt its maximally informative representation to the learning properties of the system at hand. In our experiments involving parametric algorithms (either supervised or unsupervised), the number of template vectors (also called reference vectors, prototypes, or codewords) is assumed to be coincident with the number of surface types of interest (in a classification framework, these systems are known as one-prototype classifiers [43]). This implies that the distribution of class-specific representative samples is assumed to be consistent with the model of the class-specific spectral distributions adopted by the parametric labeling algorithm.

Let us model each supervised ROI (corresponding to a spectrally uniform surface area; see Section V-A) with a Gaussian distribution, parameterized by a (mean vector, covariance matrix) pair, identified as $(\mu_{i,0}, \Sigma_{i,0})$, $i = 1, \ldots, L$. Thus, ML is plugged in with estimates $(\mu_{i,0}, \Sigma_{i,0})$, $i = 1, \ldots, L$ (in this case, a semisupervised iterative method for combined covariance estimation and ML classification could be adopted to mitigate the problem of limited training samples [25]), whereas NP is plugged in with estimates $\mu_{i,0}$, $i = 1, \ldots, L$. The MPAC and MPACB clustering algorithms are initialized with mean template vectors $\mu_{i,0}$, $i = 1, \ldots, L$. With regard to parametric iterative learning systems that employ class-specific Gaussian distributions (which is the case of ICM-MAP-MRF, EM, CEM, SEM, CSEM, and MSEM), the empirical rules proposed in Section III recommend that a number of class-specific training samples equal, or possibly superior, to $10 \times$ number of free parameters $\geq [10 \times (D + 0.5\ D^2 + 0.5D)]$, be selected to ensure an adequate estimation of a per-class (mean vector, covariance matrix) pair (also refer to Section VI-A). To avoid poor generalization capability of the induced classifiers (related to model complexity), ICM-MAP-MRF, EM, and CEM, altogether with a specific implementation of the *partially semisupervised* classifiers SEM, CSEM, and MSEM (identified as version SEM2A, CSEM2A, and MSEM2A, respectively, where labeled samples with full weight are passed on to these semisupervised algorithms during their training phase), as well as their *purely semisupervised* versions (identified as SEM1, CSEM1, and MSEM1, respectively, where labeled samples with full weight are not passed on to these semisupervised algorithms during their training phase), employ the following initialization strategy (at iteration 0). First, supervised ROI-driven mean vector estimates $\mu_{i,0}$, $i = 1, \ldots, L$, are passed on to a nearest-prototype classification step, NP. Next, the crisp output map generated from NP provides image-wide category-specific estimates (μ_i, Σ_i), $i = 1, \ldots, L$, which are finally passed on, at iteration 1, to the iterative learning system at hand. In the case of poorly posed classification problems, due to the presence of many semilabeled samples (provided with partial weight) and of few labeled samples (provided with full weight, whose contribution to the system's free-parameter estimation may become negligible), partially semisupervised implementations SEM2A, CSEM2A, and MSEM2A are expected to behave somewhat similarly to their purely semisupervised counterparts.

A second supervised implementation of SEM, CSEM, and MSEM (identified as SEM2B, CSEM2B, and MSEM2B, respectively), is initialized with the supervised ROI-driven class-specific estimates $(\mu_{i,0}, \Sigma_{i,0})$, $i = 1, \ldots, L$. Supervised versions SEM2B, CSEM2B, and MSEM2B are expected to be more susceptible to poor initialization than their unsupervised counterparts (SEM1, CSEM1, and MSEM1, respectively) as well as their partially supervised alternatives (SEM2A, CSEM2A, and MSEM2A, respectively) since covariance matrices are more sensitive than intensity averages to the curse of dimensionality. Moreover, differences in performance between SEM2B, CSEM2B, and MSEM2B, and their purely semisupervised counterparts (SEM1, CSEM1, and MSEM1, respectively) are expected to be superior to those between partially supervised SEM2A, CSEM2A, and MSEM2A from purely semisupervised SEM1, CSEM1, and MSEM1, respectively. A complete list of the algorithms implemented for comparison purposes is proposed in Table I.

D. User-Defined Parameter Setting

Context-sensitive multiscale image mapping algorithms (namely, MPAC, MPACB, MSEM1, and MSEM2), adopt an application-independent battery of three local window sizes, equal to 3×3, 7×7, 11×11, to be employed in combination with the global (image-wide) scale, e.g., 512×512 in test case 1 (see Sections IV-A and Appendix I). Context-sensitive multiscale image clustering algorithms MPAC and MPACB employ a spatial continuity parameter $\beta = 0$ in (2), to inhibit their MRF-based contextual mechanism, such that its context sensitivity is exclusively due to multiscale intensity average estimation. The maximum number of iterations is set equal to 10 in the entire set of iterative algorithms (which are all, with the exclusion of NP, ML, and PNN). Context-sensitive single-scale MRF-based algorithms (namely, CEM, CSEM1, CSEM2, and ICM-MAP-MRF), employ [e.g., in (13)] two-point clique potential parameters $\beta = \beta_i = 1.0$, $i = 1, \ldots, L$. It is obvious that optimal smoothing parameters β_h, $h \in \{1, L\}$, are both class- and application-dependent. To avoid a time-consuming class-specific trial-and-error parameter selection strategy that would represent a degree of user's supervision superior to that required by the rest of the algorithms involved in our comparison, we set two-point clique potential parameters $\beta = \beta_i = 1.0$, $i = 1, \ldots, L$, independent of the class. This choice is in line with [43], where $\beta \in [1.0{-}1.6]$, independent of the dataset because larger values of β would lead to excessive smoothing of

(a) (b)

Fig. 3. (a) Test case 1. MPACB clustering of the three-band SPOT image, number of classes $L = 21$, shown in pseudocolors. (b) Test case 1. MSEM1 clustering of the three-band SPOT image, number of classes $L = 21$, shown in pseudocolors.[7]

regions. In PNN, spread parameter σ is data-driven based on a class-independent, trial-and-error selection procedure, which is fast and easy, due to the sensitivity of PNN to a small range of σ values. Unfortunately, MLP's model selection and parameter setting are application-dependent, based on a time-consuming trial-and-error strategy. A time-consuming trial-and-error approach is also adopted to select SVM's best pair of parameters (namely, a regularization coefficient and the spread of Gaussian kernels) according to map photointerpretation and resubstitution error quality criteria.

VI. Experimental Results

Of the systems compared in this experimental session, plug-in NP and ML, and nonparametric PNN, are expected to perform well in minimizing the resubstitution error (where bias must be low), whereas parametric iterative (adaptive) labeling algorithms, either unsupervised (EM, CEM, MPAC, and MPACB), supervised (ICM-MAP-MRF), purely semisupervised (namely, SEM1, CSEM1, and MSEM1), or partially semisupervised (namely, SEM2, CSEM2, and MSEM2, in both versions A and B), where all unlabeled samples contribute to the adaptation of category-specific template vectors, are expected to improve their generalization ability upon unobserved image areas (when the combination of bias with variance must be kept low) at the cost of a possible increase in their resubstitution error on ground truth ROIs (due to an increase in bias).

Based on model complexity, adaptive MPAC and MPACB should employ plug-in NP as a reference, whereas EM, CEM, and the supervised and unsupervised implementations of SEM, CSEM, and MSEM should employ plug-in ML as a reference.

All our experiments are conducted on a SUN Ultra 5 workstation with operating system SunOS 5.6, 64 MB of RAM, and a CPU UltraSPARC-IIi at 270 MHz. No optimization is employed at code compilation.

[7]Every class index is associated with a pseudocolor chosen to mimic the true color of that surface class (e.g., three shades of blue are adopted to depict labels belonging to classes *sea water 1* to *sea water 3*, etc.), to enhance human interpretability of mapping results.

TABLE III
Test Case 1. Resubstitution Overall Accuracy (Sum of Diagonal Elements of the Confusion Matrix) Between Labeling Results and Reference Data (ROIs). Number of Label Types (= number of ground truth ROIs) = 21. Rank1 Is Best When Smallest. *: Without Supervised (Training) Samples. **: With Supervised (Training) Samples

Classifier	Resubstitution overall accuracy %	Rank1
NP	87.8	5
ML	82.1	9
PNN	97.9	1
MLP	94.2	2
SVM-OAA	93.4	4
SVM-OAO	93.5	3
ICM-MAP-MRF	87.4	6
EM	83.6	8
CEM	87.3	7
SEM1*	66.3	18
SEM2A**	69.1	14
SEM2B**	75.8	11
CSEM1*	66.9	17
CSEM2A**	67.6	16
CSEM2B**	76.1	10
MSEM1*	68.4	15
MSEM2A**	70.9	13
MSEM2B**	71.8	12
MPAC	65.3	19
MPACB	63.0	20

A. SPOT Image Test Case

In PNN, spread parameter σ is set to 0.6 after a class-independent, trial-and-error selection procedure, which is fast and easy, also due to the sensitivity of PNN to a small range of σ values. By trial and error, MLP is selected with 15 hidden sigmoidal units, the learning rate is 0.01, the momentum equals 0.02 and the number of iterations is set to 10 000. In SVM-OAA, kernels are RBFs with the regularization parameter equal to 10 and Gaussian spread equals 10. In SVM-OAO, kernels are RBFs with the regularization parameter equal to 11 and Gaussian spread equals 16.

As two interesting examples of the mapping results obtained with the proposed parameter setting, Fig. 3(a) and (b) shows (in pseudocolors; refer to footnote 7) the maps generated with,

TABLE IV
TEST CASE 1. OVERLAPPING AREA (SUM OF DIAGONAL ELEMENTS OF THE CONFUSION MATRIX AFTER RESHUFFLING) BETWEEN x_i AND THE REFERENCE CLUSTER MAP x_i^*, $i = 1, \ldots, 3$. NUMBER OF LABEL TYPES (= number of ground truth ROIs) = 21. RANK2 IS BEST WHEN SMALLEST. ": WITHOUT SUPERVISED (TRAINING) SAMPLES. **: WITH SUPERVISED (TRAINING) SAMPLES

Classifier	Standard value of the overlapping area (%) on reference cluster map 1 (sample mean=40.9000, sample std=9.0708)	Standard value of the overlapping area (%) on reference cluster map 2 (sample mean=42.8857, sample std=6.6990)	Standard value of the overlapping area (%) on reference cluster map 3 (sample mean=35.6850, sample std=6.9130)	Average of standard mean values 1, 2 and 3	Rank2
NP	-1.3229 (28.9)	-0.6271 (37.6)	-1.1840 (27.5)	-1.0447	18
ML	-0.4630 (36.7)	-1.1417 (34.1)	-0.5909 (31.6)	-0.7319	14
PNN	-0.4740 (36.6)	-0.5830 (37.9)	-0.5620 (31.8)	-0.5397	13
MLP	-1.1686 (30.2)	-1.8475 (29.3)	-1.3431 (26.4)	-1.4531	20
SVM-OAA	- 1.5103 (27.2)	-0.9800 (35.2)	-1.4299 (25.8)	-1.3067	19
SVM-OAO	- 0.9481 (32.3)	-0.8623 (36.0)	-0.8513 (29.8)	-0.8872	15
ICM-MAP-MRF	-1.2347 (29.7)	-0.6271 (37.6)	-0.8947 (29.5)	-0.9188	16
EM	-0.9481 (32.3)	-0.3183 (39.7)	-0.0268 (35.5)	-0.4311	12
CEM	-1.1245 (30.7)	-0.8329 (36.2)	-0.9526 (29.1)	-0.9700	17
SEM1"	1.1906 (51.7)	0.4315 (44.8)	0.6820 (40.4)	0.7681	6
SEM2A**	0.9481 (49.5)	0.4462 (44.9)	0.6820 (40.4)	0.6921	7
SEM2B**	0.7056 (47.3)	0.0346 (42.1)	-0.5909 (31.6)	0.0497	10
CSEM1"	0.6835 (47.1)	0.4315 (44.8)	0.8267 (41.4)	0.6473	8
CSEM2A**	0.6504 (46.8)	0.3433 (44.2)	0.8556 (41.6)	0.6165	9
CSEM2B**	0.6725 (47.0)	-0.3624 (39.4)	-0.5475 (31.9)	-0.0792	11
MSEM1"	1.2347 (52.1)	0.8138 (47.4)	1.3909 (45.3)	1.1465	4
MSEM2A**	1.2678 (52.4)	0.7991 (47.3)	1. 3909 (45.3)	1.1526	3
MSEM2B**	0.9040 (49.1)	1.0050 (48.7)	0.4795 (39.0)	0.7962	5
MPAC	0.5953 (46.3)	1.8724 (54.6)	1.2896 (44.6)	1.2525	1
MPACB	0.3418 (44.0)	2.0048 (55.5)	1.3764 (45.2)	1.2410	2

respectively, clustering algorithms MPACB and MSEM1 (the other output maps are omitted to save presentation space). According to perceptual quality criteria adopted by expert photointerpreters, MPACB and MSEM1 appear to perform better than several other competing systems (in terms of genuine but small image detail detection), although their maps look different [e.g., in Fig. 3(a) and (b), note the different spatial distributions of water types].

In the framework of a resubstitution error estimation method, Table III reports the overall accuracy (sum of diagonal elements of the confusion matrix) between labeling results and ground truth ROIs. Table III shows that, in line with theoretical expectations, the resubstitution accuracy of some of the parametric, iterative labeling algorithms (namely, EM, CEM, SEM, CSEM, MSEM, MPAC, and MPACB), is largely inferior, or not superior, to that of traditional plug-in classifiers, whether nonparametric (PNN) or parametric (NP and ML), and to inductive learning classifiers (MLP and SVM). As expected, CEM performs better than EM in regularizing classification results upon training areas. Unexpectedly, MPACB performs worse than MPAC (in terms of salt-and-pepper classification noise effect on training areas). Partially semisupervised classifiers SEM2, CSEM2, and MSEM2 perform better than their purely semisupervised counterparts, in line with theoretical expectations. Although a low resubstitution error is a desirable property, optimistically biased estimates (refer to Section VI-B), provided by Table III, appear to be counterintuitive for expert photointerpreters employing perceptual quality criteria [e.g., see Fig. 3(a) and (b), ranked low in Table III].

To compare the generalization capabilities of inductive learning methods based on the DAMA strategy, Table IV shows the maximum sum (after reshuffling) of diagonal elements of the overlapping area matrix computed between the output map x and the multiple cluster maps x_i^* (generated by the enhanced Linde–Buzo–Gray (ELBG) vector quantizer [5]) $i = 1, 2, 3$, $L = 21$, from the raw image z. In line with qualitative photointerpretation of mapping results, Table IV reveals that labeling fidelities to multiple cluster maps of the MPAC, MPACB, and MSEM output maps appear to be superior to those of the other labeling approaches, including NP and ML (as theoretically expected). In line with theoretical considerations, MPAC detects fine image details better than MPACB; MSEM performs better than SEM, while SEM performs better than CSEM in preserving genuine but small image details, which is theoretically plausible but in contrast with conclusions found in [39]. Effectiveness of partially semisupervised versions SEM2B, CSEM2B, and MSEM2B appears to be slightly inferior to that of their partially semisupervised counterparts (SEM2A, CSEM2A, and MSEM2A, respectively) which in turn behave somewhat similarly to their purely semisupervised implementations (SEM1, CSEM1, and MSEM1, respectively), in line with theoretical expectations. Overall, in line with theoretical expectations, the poor correlation between Rank1 (from resubstitution) and Rank2 (from generalization) reveals the presence of the Hughes phenomenon.

To investigate the spatial fidelity of segmentation results to reference data according to the DAMA strategy, Table V reports the mean of the edge map difference computed between an edge map extracted from the output map and the one extracted from every multiple cluster map. Table V shows that multiscale labeling algorithms (namely, MPAC, MPACB, and MSEM), context-insensitive adaptive SEM and nonadaptive ML are superior to the other algorithms in preserving genuine but small image details, irrespective of their labeling. In particular, MPAC and MPACB outperform the other competing systems, whereas SEM performs better than MSEM, which is, in turn, better than CSEM. These spatial fidelity results appear to be

TABLE V
TEST CASE 1. MEAN AND STANDARD DEVIATION OF THE EDGE MAP DIFFERENCE COMPUTED BETWEEN THE TWO EDGE MAPS MADE FROM x_i^* AND x_i, $i = 1, \ldots, 3$. RANK3 IS BEST WHEN SMALLEST. *: WITHOUT SUPERVISED (TRAINING) SAMPLES. **: WITH SUPERVISED (TRAINING) SAMPLES

Classifier	Standard value of the mean (in [0,4]) of edge map difference 1 (sample mean = 1.0050, sample std = 0.1965)	Standard value of the mean (in [0,4]) of edge map difference 2 (sample mean = 0.9390, sample std = 0.2038)	Standard value of the mean (in [0,4]) of edge map difference 3 (sample mean = 1.1060, sample std = 0.2569)	Average of standard mean values 1, 2 and 3	Rank3
NP	0.7887 (1.16)	-0.1423 (0.91)	0.7553 (1.31)	0.4672	15
ML	-0.6360 (0.88)	-0.5348 (0.83)	-0.7631 (0.92)	-0.6446	6
PNN	0.0763 (1.02)	0.1030 (0.96)	-0.1012 (1.09)	0.0260	14
MLP	1.5010 (1.30)	1.2806 (1.20)	1.3004 (1.45)	1.3607	18
SVM-OAA	0.9922 (1.20)	1.0844 (1.16)	0.7553 (1.31)	0.9440	17
SVM-OAO	0.7887 (1.16)	1.2806 (1.20)	0.7164 (1.30)	0.9286	16
ICM-MAP-MRF	1.9589 (1.39)	1.9185 (1.33)	2.0790 (1.65)	1.9855	19
EM	0.2290 (1.05)	-0.2404 (0.89)	-0.0623 (1.10)	-0.0246	13
CEM	2.0098 (1.40)	2.0657 (1.36)	2.1958 (1.68)	2.0905	20
SEM1*	-0.8904 (0.83)	-0.9274 (0.75)	-0.9188 (0.88)	-0.9122	3
SEM2A**	-0.8904 (0.83)	-0.8783 (0.76)	-0.9188 (0.88)	-0.8958	4
SEM2B**	-0.7378 (0.86)	-0.7311 (0.79)	-0.7631 (0.92)	-0.7440	5
CSEM1*	-0.4325 (0.92)	-0.2404 (0.89)	-0.1791 (1.07)	-0.2840	10
CSEM2A**	-0.3816 (0.93)	-0.1423 (0.91)	-0.1402 (1.08)	-0.2214	11
CSEM2B**	-0.1781 (0.97)	0.0540 (0.95)	-0.0623 (1.10)	-0.0621	12
MSEM1*	-0.7378 (0.86)	-0.4858 (0.84)	-0.5684 (0.97)	-0.5973	7
MSEM2A**	-0.6869 (0.87)	-0.4367 (0.85)	-0.4906 (0.99)	-0.5380	8.5
MSEM2B**	-0.6869 (0.87)	-0.4367 (0.85)	-0.4906 (0.99)	-0.5380	8.5
MPAC	-1.1448 (0.78)	-1.2708 (0.68)	-1.3081 (0.78)	-1.2413	1
MPACB	-0.9413 (0.82)	-1.3199 (0.67)	-1.0356 (0.85)	-1.0989	2

in fairly strong agreement with the labeling accuracy results shown in Table IV, as confirmed by the Spearman correlation coefficient computed between Rank2 and Rank3, equal to 0.789 [24].

Computation time of the competing algorithms is proposed in Table VI, which shows that in this experiment the quality of labeling and segmentation results appears to be inversely proportional to computation time, with the notable exception of MLP and SVM. In particular, SEM, MPAC, and MPACB appear to be able to guarantee an interesting compromise between labeling and spatial fidelity of output results to reference data, with computation time.

Overall, these conclusions appear to be consistent with those obtained by expert photointerpreters and in line with the theoretical expectations about the algorithms' potential utility.

TABLE VI
COMPUTATION TIMES OF THE INDUCTIVE LEARNING ALGORITHMS IN THE SPOT IMAGE TEST CASE. RANK4 IS BEST WHEN SMALLEST. *: 21 LABEL TYPES, 1328 TRAINING PIXELS (0.5%). **: TEN MAX ITERATIONS

Classifier*	Computation time	Rank4
NP	4"	1
ML	27"	2
PNN	11' 50"	7
MLP	2h 35' 30"	20
SVM-OAA	2h 20' 10"	19
SVM-OAO	2h 10' 50"	18
ICM-MAP-MRF **	7' 30"	3
EM **	15' 00"	11
CEM **	16' 40"	12
SEM1, SEM2A, SEM2B **	10' 50"	4.5
CSEM1, CSEM2A, CSEM2B **	12' 10"	8.5
MSEM1, MSEM2A, MSEM2B **	58' 10"	15.5
MPAC **	28' 30"	13
MPACB **	30' 30"	14

B. Landsat Image Test Case

This test image is less fragmented than test case 1. As a consequence, in this experiment functional benefits deriving from the use of the context-sensitive single-scale ICM-MAP-MRF, CEM, and CSEM algorithms (provided with an MRF-based mechanism to enforce spatial continuity in pixel labeling) are expected to be greater than in test case 1. Moreover, the current space dimensionality, superior to that in test case 1, is expected to increase the undesirable effects due to the curse of dimensionality which may affect partially semisupervised implementations SEM2B, CSEM2B, and MSEM2B, as well as the plug-in ML classifier.

User-defined parameters are the same as those selected in test case 1, but spread parameter σ in PNN, which is set equal to 1.0 after a (category-independent, easy and fast) trial-and-error selection procedure. By trial and error, MLP is selected with 17 hidden sigmoidal units (the learning rate, the momentum, and the number of iterations are the same as in test case 1). In SVM-OAA, kernels are RBFs with the regularization parameter equal to 10 and Gaussian spread equal to 1. In SVM-OAO, kernels are RBFs with the regularization parameter equal to 10 and Gaussian spread equal to 3.

As in test case 1, interesting examples of the mapping results obtained with this parameter setting are shown in Fig. 4(a) and (b), where two maps generated by MPACB and MSEM1 respectively are depicted (in pseudocolors). In test case 2, due to its large fragmentation and the absence of easy-to-recognize built-up areas, it is rather difficult for expert photointerpreters to determine whether or not, for example, MPACB [see Fig. 4(a)] performs better than MSEM1 [see Fig. 4(b)].

In the framework of a resubstitution error estimation method, Table VII shows the overall accuracy (sum of diagonal elements

(a) (b)

Fig. 4. (a) Test case 2. MPACB classification of the seven-band Landsat TM image, number of classes $L = 12$, shown in pseudocolors. (b) Test case 2. MSEM1 classification of the seven-band Landsat TM image, number of classes $L = 12$, shown in pseudocolors. For both, refer to footnote 7.

TABLE VII
TEST CASE 2. RESUBSTITUTION OVERALL ACCURACY (SUM OF DIAGONAL ELEMENTS OF THE CONFUSION MATRIX) BETWEEN LABELING RESULTS AND REFERENCE DATA (ROIS). NUMBER OF LABEL TYPES (= number of ground truth ROIs) = 12. *: WITHOUT SUPERVISED (TRAINING) SAMPLES. **: WITH SUPERVISED (TRAINING) SAMPLES. RANK5 IS BEST WHEN SMALLEST

Classifier	Resubstitution overall accuracy %	Rank5
NP	99.3	5.5
ML	99.8	2.5
PNN	100.	1
MLP	98.7	9
SVM-OAA	99.6	4
SVM-OAO	99.8	2.5
ICM-MAP-MRF	99.3	5.5
EM	91.1	20
CEM	99.3	5.5
SEM1*	96.7	13.5
SEM2A**	96.8	12
SEM2B**	94.8	19
CSEM1*	96.6	15
CSEM2A**	96.7	13.5
CSEM2B**	94.9	18
MSEM1*	97.3	10
MSEM2A**	96.4	16
MSEM2B**	96.9	11
MPAC	96.1	17
MPACB	99.3	5.5

of the confusion matrix) between labeling results and ground truth ROIs. In this experiment, the performance of nontraditional algorithms (namely, MPAC, MPACB, SEM, CSEM, and MSEM) is more competitive with those of traditional labeling approaches (namely, NP, ML, PNN, MLP, SVM, and ICM-MAP-MRF) than in test case 1 (refer to Table IV). In line with theory, MPACB performs better than MPAC in smoothing out classification results on training areas. The same consideration holds for CEM with respect to EM.

To compare the generalization capabilities of competing classifiers, Table VIII shows the maximum sum (after reshuffling) of diagonal elements of the overlapping area matrix computed between the reference cluster map x_i^* (generated by the ELBG vector quantizer [5]) with the corresponding submap $x_i \subseteq x$, $i = 1, 2$, with $L = 12$ (see Section VI-B). In Table VIII, where ML shows the worst performance (as expected), the labeling fidelities to multiple cluster maps of output results provided by clustering algorithms MPACB, SEM1, CSEM1, and MSEM1, as well as their partially semisupervised implementations SEM2A, CSEM2A, and MSEM2A, are superior to those of the other labeling approaches, which is consistent in part with test case 1 (refer to Table IV). Due to the curse of dimensionality with respect to model complexity (in test case 2, the ratio between the number of per-class samples with the number of class-specific free parameters is relatively high for several classes, but spatial autocorrelation is superior to that in test case 1; refer to Section VI-A), partially semisupervised versions SEM2B, CSEM2B, and MSEM2B, as well as ML, perform rather poorly, which was not always true in test case 1 (refer to Table IV). It is noteworthy that in line with theoretical expectations MPAC (which is prone to detect artifacts) performs more poorly in the less fragmented test case 2 than in test case 1. Actually, in test case 2, MPAC performs even worse than plug-in NP. Overall, in line with theoretical expectations and in line with test case 1 (refer to Section VI-A), the scanty correlation between Rank5 (from resubstitution) and Rank6 (from generalization) reveals the presence of the Hughes phenomenon.

To investigate spatial fidelity of segmentation results to reference data (irrespective of their labeling), Table IX reports the mean of the edge map difference computed between an edge map extracted from the system's output map and the one extracted from every reference cluster map. In contrast with results shown in Table VIII, Table IX reveals that although SEM1, CSEM1, and MSEM1 perform better than ML (in line with theoretical expectations), they are ranked average in preserving genuine but small image details irrespective of their labeling. These clustering algorithms are outperformed by MPACB and MPAC, which also perform better than NP (in line with theoretical expectations). Partially semisupervised implementations SEM2B, CSEM2B, and MSEM2B perform poorly even with respect to ML. Single-scale MRF-based contextual algorithms ICM-MAP-MRF, CEM, and CSEM perform better than in test case 1 (refer to Table V), in line with theoretical expectations, which proves the strong application-dependency of

TABLE VIII
TEST CASE 2. OVERLAPPING AREA (SUM OF DIAGONAL ELEMENTS OF THE CONFUSION MATRIX AFTER RESHUFFLING) BETWEEN x_i AND THE REFERENCE CLUSTER MAP x_i^*, $i = 1, 2$. NUMBER OF LABEL TYPES (= number of ground truth ROIs) $= 12$. ": WITHOUT SUPERVISED (TRAINING) SAMPLES. **: WITH SUPERVISED (TRAINING) SAMPLES. RANK6 IS BEST WHEN SMALLEST

Classifier	Standard value of the overlapping area (%) on reference cluster map 1 (sample mean=47.5685, sample std=5.7095)	Standard value of the overlapping area (%) on reference cluster map 2 (sample mean=56.7745, sample std=6.8924)	Mean standard value	Rank6
NP	-0.2747 (46.00)	0.8771 (62.82)	0.3012	10
ML	-1.9859 (36.23)	-1.0148 (49.78)	-1.5004	19
PNN	0.3400 (49.51)	0.3969 (59.51)	0.3685	9
MLP	-0.5234 (44.58)	-0.8494 (50.92)	-0.6864	15
SVM-OAA	-0.1241 (46.86)	0.0124 (56.86)	-0.0558	13
SVM-OAO	0.3471 (49.55)	0.2315 (58.37)	0.2893	11
ICM-MAP-MRF	-0.6565 (43.82)	1.0889 (64.28)	0.2162	12
EM	-1.6776 (37.99)	-1.5284 (46.24)	-1.6030	20
CEM	0.4486 (50.13)	1.1180 (64.48)	0.7833	6
SEM1"	1.1107 (53.91)	0.6044 (60.94)	0.8575	5
SEM2A**	1.0932 (53.81)	0.6261 (61.09)	0.8597	4
SEM2B**	-1.0016 (41.85)	-1.8346 (44.13)	-1.4181	17
CSEM1"	0.7236 (51.70)	0.4970 (60.20)	0.6103	7
CSEM2A**	0.7359 (51.77)	0.4796 (60.08)	0.6077	8
CSEM2B**	-1.0156 (41.77)	-1.8853 (43.78)	-1.4505	18
MSEM1"	1.2298 (54.59)	0.7146 (61.70)	0.9722	1
MSEM2A**	1.6852 (57.19)	0.2547 (58.53)	0.9699	2
MSEM2B**	-0.5024 (44.70)	-1.1599 (48.78)	-0.8312	16
MPAC	-0.6373 (43.93)	0.1952 (58.12)	-0.2210	14
MPACB	0.6851 (51.48)	1.1760 (64.88)	0.9305	3

MRF-based image mapping approaches on the optimization of class-specific clique potentials. The Spearman correlation value between Rank6 and Rank7 is 0.437, revealing poor agreement [24] (which justifies the separate, independent computation of indexes of labeling and segmentation fidelity of a map to reference data pursued by DAMA).

Computation time of the labeling algorithms is reported in Table X. These results are in line with those shown in Table VI with the exception of PNN. In this experiment, the large computational load of PNN, which depends on the cardinality of the training dataset, makes the exploitation of this algorithm (quite) impracticable.

Overall, conclusions about test case 2 seem to be fairly consistent with those of test case 1 and with theoretical expectations about the algorithms' potential utilities.

C. Discussion of Experimental Results

In the (subjective) assessment of quantitative experimental results proposed in this section, the evaluation criterion proposed in [13], where Zamperoni considers any new image processing algorithm worth disseminating among a broad audience if it may enrich a commercial image processing software toolbox, is taken into consideration.

Let us collect results of test cases 1 and 2 in Table XI, where we have the following.

- Column *Total* (learning + generalization + computational load) quality index (*Tot.*, best when smallest) is computed as $Rank1 + \ldots + Rank8$. *Score1* is the rank of column *Tot*.
- Column *Accuracy* (learning + generalization) index (best when smallest) is computed as $Rank1 + Rank2 + Rank3 + Rank5 + Rank6 + Rank7$, i.e., *Accuracy* ignores the computational costs of the compared algorithms. *Score2* is the rank of column Accuracy.
- *Generalization Capability* (generalization) index (best when smallest) $= Rank2 + Rank3 + Rank6 + Rank7$. *Score3* is the rank of column *Gen.Cap*.

The arbitrary and problem-specific nature of the map quality measures *Score1*, *Score2*, and *Score3* does not allow the reaching of any final conclusion about the accuracy and efficiency of the algorithms involved in the comparison (i.e., other empirical evaluation criteria, such as considering the best classifier the one whose largest rank number is the smallest, may provide different subjective conclusions). Nonetheless, the analysis of Table XI yields some relative (subjective) conclusions about the potential usability of the tested classifiers in dealing with the badly posed classification of piecewise constant or slowly varying color images (i.e., where texture information is negligible). These relative conclusions are interesting as they are based on weak (arbitrary, subjective) but numerous measures of image mapping quality that reasonably approximate the real-world characteristics of new generation image mapping applications.

1) Subjective but numerous measures of image mapping quality collected in *Score2* and *Score3* reveal that when computational costs are ignored (which may be reasonable in a technological scenario where processing speed increases dramatically each year):
   - In line with theoretical expectations (see Sections IV and Appendix I), nontraditional data mapping approaches (namely, SEM, CSEM, MSEM, MPA,C and MPACB) appear to be capable of guaranteeing image labeling performance superior (on average) to those of first generation classifiers. Among nontraditional classification approaches, in line with theoretical expectations (see Appendix I), clustering algorithms MPACB (ranked first in *Score2* and *Score3*) and MSEM1 (ranked second in *Score2* and third in *Score3*) appear (on average) to be superior to or competitive with the other competing

TABLE IX
TEST CASE 2. MEAN AND STANDARD DEVIATION OF THE EDGE MAP DIFFERENCE COMPUTED BETWEEN THE TWO EDGE MAPS MADE FROM x_i^* AND x_i, $i = 1, 2$. *: WITHOUT SUPERVISED (TRAINING) SAMPLES. **: WITH SUPERVISED (TRAINING) SAMPLES. RANK7 IS BEST WHEN SMALLEST

Classifier	Standard value of the mean (in [0,4]) of edge map difference 1 (sample mean = 0.6118, sample std = 0.0376)	Standard value of the mean (in [0,4]) of edge map difference 2 (sample mean = 0.5688, sample std = 0.0478)	Mean standard value	Rank7
NP	-0.2872 (0.601)	-0.9357 (0.524)	-0.6115	6
ML	-0.4202 (0.596)	0.5280 (0.594)	0.0539	12
PNN	-0.2872 (0.601)	0.3816 (0.587)	0.0472	11
MLP	0.4574 (0.629)	0.7998 (0.607)	0.6286	15
SVM-OAA	-0.6861 (0.586)	-0.7475 (0.533)	-0.7168	5
SVM-OAO	-0.6063 (0.589)	0.0261 (0.570)	-0.2901	7
ICM-MAP-MRF	-0.7127 (0.585)	-1.1866 (0.512)	-0.9497	4
EM	2.2924 (0.698)	1.9917 (0.664)	2.1420	20
CEM	-0.7393 (0.584)	-1.1657 (0.513)	-0.9525	3
SEM1*	0.0585 (0.614)	-0.0157 (0.568)	0.0214	10
SEM2A**	0.0053 (0.612)	-0.1202 (0.563)	-0.0575	8
SEM2B**	0.9361 (0.647)	1.2598 (0.629)	1.0980	18
CSEM1*	0.1915 (0.619)	0.0261 (0.570)	0.1088	14
CSEM2A**	0.1915 (0.619)	0.0052 (0.569)	0.0984	13
CSEM2B**	0.6436 (0.636)	1.0298 (0.618)	0.8367	17
MSEM1*	0.2181 (0.620)	-0.2457 (0.557)	-0.0138	9
MSEM2A**	1.0159 (0.650)	0.3816 (0.587)	0.6987	16
MSEM2B**	1.4148 (0.665)	1.0507 (0.619)	1.2328	19
MPAC	-1.6701 (0.549)	-0.7894 (0.531)	-1.2297	2
MPACB	-2.0158 (0.536)	-2.2740 (0.460)	-2.1449	1

TABLE X
COMPUTATION TIMES OF THE INDUCTIVE LEARNING ALGORITHMS IN THE LANDSAT IMAGE TEST CASE. RANK8 IS BEST WHEN SMALLEST. *: 12 LABEL TYPES, 20431 TRAINING PIXELS (2.6%). **: TEN MAX ITERATIONS

Classifier*	Computation time	Rank8
NP	23”	1
ML	1’ 31”	2
PNN	9h 22’ 30”	17
MLP	15h 35’ 20”	20
SVM-OAA	10h 45’ 30”	19
SVM-OAO	10h 12’ 10”	18
ICM-MAP-MRF **	15’ 50”	3
EM **	1h 20’ 10”	10
CEM **	1h 23’ 50”	11
SEM1, SEM2A, SEM2B **	32’ 30”	4.5
CSEM1, CSEM2A, CSEM2B**	36’ 30”	7.5
MSEM1, MSEM2A, MSEM2B**	4h 4’ 30”	14.5
MPAC**	1h 40’ 10”	12
MPACB **	1h 43’ 40”	13

mapping approaches, especially when considering map quality indexes *Rank2* and *Rank6*, which appear to be highly correlated to the qualitative map assessment criteria adopted by human photointerpreters. In particular, MSEM appears to be largely superior to CSEM, but only slightly superior to SEM (ranked third in *Score2* and fourth in *Score3*). This experimental superiority of SEM with respect to CSEM is somehow in contrast with results reported in [39]. While MPACB seems to be superior to MSEM (also in terms of an inferior computation overhead), MSEM features an application domain, ranging from unsupervised to supervised image mapping, which is wider than MPACBs. It is noteworthy that theoretical limitations of MPAC (tendency to generate artifacts, see Appendix I), known from existing literature, are confirmed by experimental numerical results (MPAC ranks second in *Score3* where generalization capability is considered irrespective of learning ability, but it ranks fifth in *Score2* where the combination of learning and generalization capabilities is examined).

- Among traditional algorithms (namely, NP, ML, PNN, MLP, SVM, ICM-MAP-MRF, EM, CEM):
  - — Only the nonparametric PNN is ranked high in *Score2*, due to its favorable resubstitution error (see *Rank1* and *Rank5*), whereas its more relevant generalization capability appears to be either poor or average (e.g., refer to columns *Rank2*, *Rank3*, and *Rank6*, *Rank7*), as reflected by *Score3*.
  - — In line with theoretical expectations, single-scale MRF-based image mapping approaches (e.g., ICM-MAP-MRF, CEM, and CSEM) require accurate application-dependent fine tuning of class-specific clique potentials to become effective.
  - — Although it requires time-consuming model selection and parameter fine tuning procedures, context-insensitive MLP is slow to reach convergence and performs quite poorly in both image mapping experiments.
  - — Although SVM classifiers are intrinsically more robust than other algorithms with respect to the Hughes phenomenon [53], supervised context-insensitive SVM is slow to reach convergence and performs rather poorly in both image mapping experiments. As expected, SVM-OAO performs better than SVM-OAA at a lower computational cost.
  - — Context-insensitive EM (conceived as a pdf estimator; see Fig. 1) performs rather poorly in image mapping tasks. While this result is somehow in contrast with common practice in RS image mapping

TABLE XI
SUMMARY OF EXPERIMENTAL RESULTS. A) TOTAL: (learning + generalization + computational load) QUALITY INDEX (best when smallest) = $Rank1 + \ldots + Rank8$. *Score1* IS THE RANK OF COLUMN Tot. B) ACCURACY: (learning + generalization) INDEX (best when smallest) = $Rank1 + Rank2 + Rank3 + Rank5 + Rank6 + Rank7$. Score2 IS THE RANK OF COLUMN *ACCURACY*. C) GENERALIZATION CAPABILITY: (GENERALIZATION) INDEX (best when smallest) = $Rank2 + Rank3 + Rank6 + Rank7$. Score3 IS THE RANK OF COLUMN GenCap

Classifier	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Tot.	Score1	Accuracy	Score2	Gen.Cap.	Score3
NP	5	18	15	1	5.5	10	6	1	61.5	3	59.5	10	49	12.5
ML	9	14	6	2	2.5	19	12	2	66.5	5	62.5	12.5	51	15.5
PNN	1	13	14	7	1	9	11	17	73	7	49	3.5	47	10
MLP	2	20	18	20	9	15	15	20	119	20	79	17	68	20
SVM-OAA	4	19	17	19	4	13	5	19	100	16	62	11	54	17
SVM-OAO	3	15	16	18	2.5	11	7	18	90.5	15	54.5	6	49	12.5
ICM-MAP-MRF	6	16	19	3	5.5	12	4	3	68.5	6	62.5	12.5	51	15.5
EM	8	12	13	11	20	20	20	10	114	19	93	20	65	19
CEM	7	17	20	12	5.5	6	3	11	81.5	10	58.5	8.5	46	9
SEM1	18	6	3	4.5	13.5	5	10	4.5	64.5	4	55.5	7	24	5
SEM2A	14	7	4	4.5	12	4	8	4.5	58	1	49	3.5	23	4
SEM2B	11	10	5	4.5	19	17	18	4.5	89	14	80	18	50	14
CSEM1	17	8	10	8.5	15	7	14	7.5	87	12	71	15	39	7
CSEM2A	16	9	11	8.5	13.5	8	13	7.5	86.5	11	70.5	14	41	8
CSEM2B	10	11	12	8.5	18	18	17	7.5	102	18	86	19	58	18
MSEM1	15	4	7	15.5	10	1	9	14.5	76	8	46	2	21	3
MSEM2A	13	3	8.5	15.5	16	2	16	14.5	88.5	13	58.5	8.5	29.5	6
MSEM2B	12	5	8.5	15.5	11	16	19	14.5	101.5	17	71.5	16	48.5	11
MPAC	19	1	1	13	17	14	2	12	79	9	54	5	18	2
MPACB	20	2	2	14	5.5	3	1	13	60.5	2	33.5	1	8	1

applications, it is in line with theoretical considerations based on results reported in [2]–[4], where a predated version of SEM, called the unlabeled expectation–maximization (UEM) classifier, is proposed. By employing each unlabeled sample, weighted by its posterior probability, in the estimation of mean and covariance statistics of all classes, UEM may cause estimated statistics to deviate from the true ones (if we assume that each unlabeled sample has a unique explicit class label), especially when a large number of unlabeled samples (with respect to the number of supervised samples) are used [2]. This limitation is potentially identical to that which may affect EM when it is employed in classification tasks.

2) Subjective but numerous measures of image mapping quality, collected in *Score1*, reveal the following.
   - Among nontraditional labeling strategies, the contextual clustering MPACB algorithm (ranked second) and the noncontextual SEM classifier (ranked first and fourth as SEM2A and SEM1, respectively) provide an interesting compromise between labeling and spatial fidelity of results to reference data, with ease of use and low computational costs. However, while SEM features a rigorous statistical foundation (unlike MPACB, CSEM, and MSEM), it can be employed in either supervised or unsupervised learning modes, and it does not apply exclusively to (2-D) images; on the other hand, MPACB is heuristic, unsupervised, and specifically developed to deal with images. Computation time of MPACB is about three times superior to that of its most competitive (in terms of mapping accuracy) alternative, SEM.
   - Traditional plug-in classifiers, namely, ML and NP, provide an acceptable tradeoff between labeling and spatial fidelity of results to reference data, ease of use and computational costs. This consideration justifies their diffusion in commercial image processing software toolboxes [38].

Overall, these conclusions appear to be consistent both with theoretical considerations and subjective (perceptual) evaluations of output maps by expert photointerpreters.

VII. CONCLUSION

As a significant extension of a related paper [19], 14 data labeling approaches (partitioned between advanced data labeling systems, namely, SEM, CSEM, MSEM, MPAC, and MPACB, and standard approaches, namely, NP, ML, PNN, MLP, SVM-OAA, SVM_OAO, ICM-MAP-MRF, EM, CEM), implemented in 20 versions and selected from existing literature and/or commercial image processing software toolboxes to cover a wide range of predictive learning principles and parameter optimization algorithms, are compared in the badly posed classification of two RS images featuring little useful texture information. In this context, a heuristic unsupervised procedure for the quality assessment of image mapping techniques (originally proposed in [18]) is adopted to provide subjective, but numerous quantitative measures of the labeling and spatial fidelity of a test map to multiple reference cluster maps (the latter fidelity estimate being ignored in practice in RS literature [24]). This empirical protocol combines an unsupervised DAMA strategy, capable of capturing genuine but small image details in multiple reference cluster maps, with traditional supervised resampling techniques (e.g., the resubstitution method). Experimental results reveal that overall, MPACB appears to be superior to or competitive with the other competing mapping systems (refer to *Score2* and *Score3* in Table XI) in terms of learning ability (refer to *Rank1* and *Rank5* in Table XI) and generalization capability (refer to the labeling fidelity of the map to reference data estimated as *Rank2* and *Rank6* in Table XI, which appear to be highly correlated to empirical map quality criteria adopted by expert

photointerpreters, as well as spatial fidelity indexes estimated as *Rank3* and *Rank7* in Table XI), at the cost of a computational overhead 50% lower than that of MSEM (which ranks second in *Score2* and third in *Score3*), but three times higher than what seems to be its most competitive semisupervised alternative, SEM (namely, SEM2A, ranked first in *Score1*, third in *Score2* and fourth in *Score3*).

In the light of Zamperoni's recommendations, additional realistic, useful and relative conclusions about the set of competing mapping systems are yielded by the collected set of experimental results. In particular, sample-based (i.e., context-insensitive) SEM, which is provided with a rigorous statistical foundation and capable of dealing with generic one-dimensional (1-D) sequences of multivariate data samples, appears to be worthy of dissemination in commercial data processing all-purpose software toolboxes, in that it is presumably useful to a broad audience dealing with pattern recognition problems, which may or may not involve images, whether unsupervised or supervised, well or badly posed.

Among traditional noncontextual classifiers, NP, ML and PNN appear to be able to justify their diffusion in commercial data processing software toolboxes, owing to their theoretical simplicity, acceptable performance and competitive computational load when they deal with real-world RS image classification problems. On the contrary, exploitation of the EM probability density function estimator is discouraged in RS image mapping tasks. The same consideration holds for context-insensitive neural network models, such as MLP and SVM, which require time-consuming model selection and parameter fine tuning procedures, that are slow in reaching convergence and perform poorly in both image mapping experiments. This observation is by no means in contrast with (rather it is complementary in nature to) the strong evidence that context-insensitive SVM classifiers, while dealing with 1-D (i.e., nonpictorial, noncontextual) sequences of multivariate data samples or when contextual information is neglected, are intrinsically more robust than other algorithms with respect to the Hughes phenomenon [53], [54].

To date, these conclusions are important in practice because context-insensitive MLP and EM are, indeed, widely adopted standard classifiers in the field of RS image understanding.

Finally, single-scale MRF-based image mapping systems appear to depend on accurate, application-dependent, user-defined parameters' fine tuning to become effective.

With no exception, these conclusions are supported by: 1) a theoretical analysis of potential advantages and drawbacks of the tested classifiers, and 2) (subjective) map assessment criteria adopted by expert photointerpreters. On an *a posteriori* basis, this overall consistency justifies the rationale behind the DAMA strategy which, despite its intrinsically subjective nature, appears to be capable of providing useful and reliable evidence about the relative assessment of competing mapping systems when an image classification problem is badly posed.

APPENDIX I

For the sake of completeness, this section reviews the nonstandard clustering and classification algorithms, namely, MPAC, MPACB, SEM, CSEM, and MSEM, selected for comparison purposes. Description of systems MPAC, SEM, and MSEM is taken from [19]. MPACB is shortly revised from [17]. The implemented CSEM is a variation of that proposed in [39].

A. Multiscale Contextual Clustering

The MPAC and MPACB algorithms are summarized below.

MPAC: An induced image classifier generated from a maximum *a posteriori* (MAP) inductive principle aims at maximizing a posterior joint probability, $p(x|z)$. If simulated annealing is adopted to learn the system parameters from a finite labeled dataset, then the global maximum of $p(x|z)$ can be approached slowly. In practice, to reduce computation time while guaranteeing suboptimal convergence, an ICM algorithm is adopted instead [1], [6], [7], [46]. Assuming that observed pixel gray values are conditionally independent and identically distributed (i.i.d), given their (unknown) labels, posterior joint probability, $p(x|z)$, can be expressed as [1], [6], [7], [46], [47]

$$p(x|z) \propto p(z_j|x_j) p\left(x_j|\hat{x}_{N_j}\right), \qquad j \in \{1, N\}, x_j \in \{1, L\}. \quad (1)$$

where $\hat{x}_{N_j}$ is the scene reconstruction in neighborhood N_j centered on pixel j. Equation (1) shows that suboptimal convergence to a local maximum of $p(x|z)$ is guaranteed if, for each pixel $j \in \{1, N\}$, ICM estimates label x_j that maximizes the right side of (1), where only the class-conditional probability $p(z_j|x_j)$ and labels of the pixel neighbors $\hat{x}_{N_j}$ are required. In practice, ICM enforces batch label updating at the end of each raster scan to alternate between the pixel labeling and category-specific model parameter estimation required to compute class-conditional probabilities [7], [47].

In [6], after speculating that an MRF model of the labeling process is not very useful unless it is combined with a good model for class-conditional densities, Pappas presents an ICM-based context-sensitive single-scale algorithm for quantization error minimization, hereafter referred to as the Pappas adaptive clustering (PAC) algorithm. PAC adopts a context-sensitive single-scale class-conditional intensity average estimate based on a slowly varying or piecewise constant image intensity model. To overcome PAC's well-known limitation, which is that of removing genuine, but small, image regions [6], [7], MPAC pursues a multiscale adaptation of the single-scale category-specific intensity average estimation strategy proposed by PAC (see Fig. 5), where texture (correlation) information is assumed to be negligible. In other words, MPAC (like PAC) is exclusively applicable to piecewise constant or slowly varying color images, eventually affected by an additive white Gaussian noise field independent of the scene [7]. Let us consider pixel $j \in \{1, N\}$ and identify with symbol $\hat{\mu}_{Wj,s}(x_j)$ the slowly varying intensity function estimated as the average of the gray levels of pixels that belong to region type $x_j \in \{1, L\}$ and fall inside an adaptive (local) window $W_{j,s} \subset W$, centered on pixel j at spatial scale $s \in \{1, S\}$, where the nonadaptive window W may overlap with the whole image z. The width of window $W_{j,s}$, $\forall j \in \{1, N\}$, $s = 1, \ldots, S$, is identified with symbol WW_s, such that window width WW_s increases with spatial scale $s \in \{1, S\}$, i.e., $WW_1 < WW_2 < \ldots < WW_S < WW_W$. Symbol

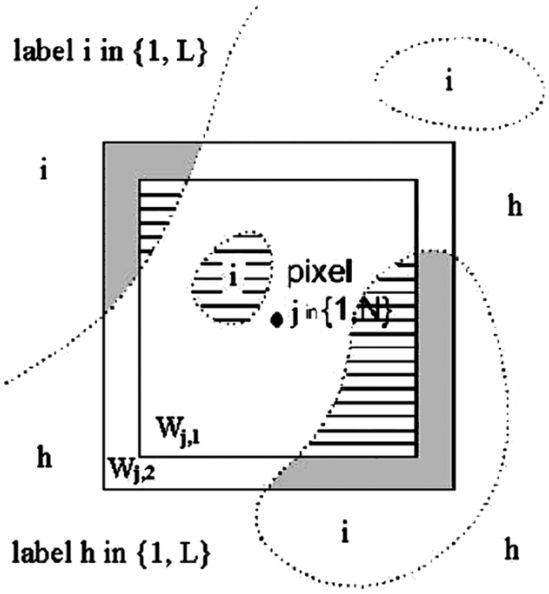


Fig. 5. Multiscale MPAC intensity average estimation strategy, whose soft-competitive adaptation is employed by the MSEM technique. In this example, category-conditional intensity averages are extracted from adaptive neighbors within windows $W_{j,s}$ at spatial scale $s \in \{1, 2\}$, centered on pixel $j \in \{1, N\}$, where window width $WW_2 > WW_1$ refers to the textured area with horizontal lines at spatial scale 1, to be added to the gray area at spatial scale 2 for label type $i \in \{1, L\}$, with $i \neq h \in \{1, L\}$.

β identifies a user-defined (free) parameter (MRF two-point clique potential) enforcing spatial continuity in pixel labeling, such that $\beta \propto \sigma^2$, where σ is the additive white Gaussian noise standard deviation [7]. Given these symbols, the MPAC cost function to be minimized becomes

$$\hat{x}_j = \arg \min_{x_j \in \{1,L\}} \{\Delta(x_j) + \beta \cdot \hat{v}_j(x_j)\}, \qquad j = 1, \ldots, N \tag{2}$$

where the second-order MRF-based cross-aura measure, $\hat{v}_j(x_j) \in \{0, 8\}$, computes the number of 8-adjacency neighbors of pixel j whose label is different from pixel status x_j (refer to Section III), while [see (3), shown at the bottom of the page] where any local estimate $\hat{\mu}_{Wj,s}(x_j)$ is (empirically) considered unreliable if the number of pixels of type x_j, within window $W_{j,s}$ is less than window width WW_s. In cascade to the crisp label assignment rules (2) and (3), the second stage of MPAC, which performs multiscale estimation of category-conditional intensity averages $\hat{\mu}_{Wj,s}(x_j)$, $j = 1, \ldots, N$, $x_j = 1, \ldots, L$, $s = 1, \ldots, S$, is shown in Fig. 5.

According to (2) and (3) and to the multiscale intensity average estimation stage shown in Fig. 5, MPAC may tolerate the same label type to feature different intensity averages in parts of the image separated in space by more than $WW_1/2$, i.e., half of the width of the investigation window that works at the finest resolution (i.e., at spatial scale $s = 1$). While this property guarantees that MPAC is less sensitive to changes in the user-defined number of input clusters than traditional noncontextual (i.e., sample-based) clustering algorithms, like the HCM vector quantizer [20] (refer to Section II), when MPAC reaches convergence, separate image areas featuring different spectral responses may be associated with the same label type. This may lead MPAC to detect artifacts, i.e., to generate an oversegmented output map [17].

Experimental results show that in comparison with alternative sample-based labeling algorithms, like the well-known HCM vector quantizer [20], [21], or well-known single-scale context-sensitive image labeling algorithms, like Rignot and Chellappa's MRF-based classifier [43], MPAC performs well in detecting genuine but small image structures [6].

MPAC With Backtracking (MPACB): To remove artifacts detected by MPAC, MPACB enforces consistency between local (segment-based) and global (image-wide) category-conditional intensity averages. To reach this objective, MPACB employs MPAC [refer to (2) and (3), plus Fig. 5] in cascade with a segment-based label backtracking module. In each iteration of MPACB, the segment-based label backtracking module works as described hereafter. After the MPAC labeling step generates an output map, this temporary map is partitioned into segments, where each segment (also called a region, or "blob" [38]) is: 1) made of connected pixels featuring the same label type and 2) provided with a unique (segment-based) identifier (digital number) [38]. Next, each segment is spectrally parameterized by its within-segment intensity average. Finally, a new output map is generated, where all pixels belonging to a segment are relabeled with the index of the category whose image-wide (i.e., global) Gaussian distribution features the shortest Mahalanobis distance from the segment's intensity average.

Experimental results show that MPACB removes artifacts but also some of the genuine image details detected by MPAC [17].

B. Semisupervised Sample-Based or Context-Sensitive Classifiers

The SEM, CSEM, and MSEM algorithms are summarized below.

SEM and CSEM Classifiers: To mitigate the small training sample size problem, SEM relies on an original iterative algorithm for ML estimation of Gaussian mixture parameters, where (few) labeled samples are given full weight, and (many) semilabeled samples (refer to Section II) are given partial weight. Thus, semilabeled samples are: 1) as many as the unlabeled samples, and 2) available at no extra classification cost. The SEM algorithm is as follows [2].

$0)_{\mathrm{SEM}}$. Initialize Gaussian mixture parameters ϕ_i^c, $i \in \{1, L\}$. Set $c = 0$.

$1)_{\mathrm{SEM}}$. E-Step: compute class-conditional probabilities, $f(z_j|\phi_i^c)$, $i \in \{1, L\}$, $j \in \{1, N\}$, and weighting factors $w_{i,j}^c$,

$$\begin{cases} \Delta(x_j) = \min\left\{[z_j - \hat{\mu}_{Wj,1}(x_j)]^2, \ldots, [z_j - \hat{\mu}_{j,S}(x_j)]^2, [z_j - \hat{\mu}_W(x_j)]^2\right\}, \text{ if local statistic } \hat{\mu}_{Wj,s}(x_j) \text{ exists} \\ \qquad\qquad \text{and is considered reliable, } \forall\, s \in \{1, S\} \\ \Delta(x_j) = [z_j - \hat{\mu}_W(x_j)]^2, \text{ if no local statistic } \hat{\mu}_{Wj,s}(x_j) \text{ does exist and is considered reliable, } \forall\, s \in \{1, S\} \end{cases} \tag{3}$$

equivalent to relative memberships $r_{i,j}^c$ (which employ neither global nor local priors), $i \in \{1, L\}$, $j \in \{1, N\}$

$$r_{i,j}^c = w_{i,j}^c = \frac{f\left(z_j|\phi_i^c\right)}{\sum_{k=1}^{L} f\left(z_j|\phi_k^c\right)}, \qquad r_{i,j}^c \in [0,1]$$

$$\sum_{i=1}^{L} r_{i,j}^c = 1, \qquad i \in \{1, L\},\ j \in \{1, N\}. \tag{4}$$

$2)_{\mathrm{SEM}}$. Crisp labeling based on the ML assignment rule

$x_j \in$ class label i, i.e., (unlabeled)$z_j \Rightarrow$ (semilabeled)$z_{i,j}$ to be employed in Equations (7) and (8) if $i = \underset{1 \le h \le L}{\arg\max} \left\{r_{h,j}^c\right\}$, $\qquad j \in \{1, N\}$. (5)

$3)_{\mathrm{SEM}}$. M-Step: maximize the mixed log-likelihood

$$\phi_i^+ = \left(\mu_i^+, \Sigma_i^+\right) \in \underset{\phi_i \in \Omega}{\arg\max} \left\{ \sum_{k=1}^{m_i} ln\left(f\left(y_{i,k}|\phi_i^c\right)\right) + \sum_{j=1}^{n_i} w_{i,j}^c ln\left(f\left(z_{i,j}|\phi_i^c\right)\right) \right\}. \tag{6}$$

Thus, the Gaussian mixture parameter update equations become as in (7) and (8), shown at the bottom of the page.

$4)_{\mathrm{SEM}}$ Check for convergence. If convergence is reached, stop. Otherwise: $c = c + 1$, and goto Step $1)_{\mathrm{SEM}}$.

Suitable for image mapping applications, a heuristic context-sensitive single-scale adaptation of SEM, identified as contextual SEM (CSEM), was proposed by the same authors in a recent paper [39]. Like SEM, CSEM is an ICM-based algorithm (i.e., it alternates between parameter estimation and pixel labeling based on (1); see Section IV) where [39]:

1) A first-stage, context-insensitive, ML classifier provides crisp membership values (labels) to a second-stage, context-sensitive, MAP classifier, in which an 8-adjacency MRF is adopted to compute "local" (i.e., per-pixel) priors. The aim of the ML classification stage is to recover more image details, as it is less likely to bias the minority class (i.e., the class featuring a small number of pixels) than the MAP classifier [39].
2) Because the accuracy of statistics estimation is strongly related to the accuracy of classified samples, Gaussian mixture parameters are computed according to the SEM update (7) and (8), where:
   - Weights are no longer computed as context-insensitive relative memberships [refer to SEMs (4)]. Rather, these weighting factors are computed as posterior probabilities by the MAP classifier. The reason for this is that contextual information is expected to enhance the performance of semilabels in terms of their influence on class-conditional statistic estimation [39].
   - Semilabeled samples are those detected by the MAP classifier, rather than the ML classifier. The reason for this is that semilabeled samples generated from the MAP classifier should contain more correctly classified samples, as contextual information is expected to reduce salt-and-pepper classification noise [39].

As a potential improvement over the original CSEM algorithm proposed in [39], our implementation of CSEM replaces the ML crisp membership values (labels) with ML (soft) relative memberships [computed via (4)] as inputs to the CSEM's second-stage (MAP classifier). In other words, our version of the CSEM's second-stage (MAP classifier) computes per-pixel prior probabilities based on an MRF exploiting ML-soft, rather than ML-crisp, membership values. To summarize, our CSEM implementation is as follows (see Fig. 6).

$0)_{\mathrm{CSEM}}$. Initialize Gaussian mixture parameters ϕ_i^c, and per-pixel priors $\alpha_{i,j}^c$, $i \in \{1, L\}$, $j \in \{1, N\}$. Set $c = 0$.

$1)_{\mathrm{CSEM}}$. E-Step: compute class-conditional probabilities, $f(z_j|\phi_i^c)$, $i \in \{1, L\}$, $j \in \{1, N\}$, weighting factors $w_{i,j}^c$, equivalent to posterior probabilities $\tau_{i,j}^c$ (employing per-pixel priors), $i \in \{1, L\}$, $j \in \{1, N\}$, plus relative memberships $r_{i,j}^c$ (employing neither global nor local priors), $i \in \{1, L\}$, $j \in \{1, N\}$, as follows:

$$\tau_{i,j}^c = w_{i,j}^c = \frac{f\left(y_j|\phi_i^c\right)\alpha_{i,j}^c}{\sum_{k=1}^{L} f\left(y_j|\phi_k^c\right)\alpha_{k,j}^c}, \quad \tau_{i,j}^c \in [0,1]$$

$$\sum_{i=1}^{L} \tau_{i,j}^c = 1, \qquad i \in \{1, L\}, j \in \{1, N\} \tag{9}$$

$$\mu_i^+ = \frac{\sum_{k=1}^{m_i} y_{i,k} + \sum_{j=1}^{n_i} w_{i,j}^c z_{i,j}}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} = \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{k=1}^{m_i} y_{i,k}}{m_i} + \frac{\sum_{j=1}^{n_i} w_{i,j}^c}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{j=1}^{n_i} w_{i,j}^c z_{i,j}}{\sum_{j=1}^{n_i} w_{i,j}^c} \qquad i \in \{1, L\} \tag{7}$$

$$\sum_i^+ = \frac{\sum_{k=1}^{m_i} \left(y_{i,k} - \mu_i^+\right)\left(y_{i,k} - \mu_i^+\right)^T + \sum_{j=1}^{n_i} w_{i,j}^c \left(z_{i,j} - \mu_i^+\right)\left(z_{i,j} - \mu_i^+\right)^T}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c}$$

$$= \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{k=1}^{m_i} \left(y_{i,k} - \mu_i^+\right)\left(y_{i,k} - \mu_i^+\right)^T}{m_i} + \frac{\sum_{j=1}^{n_i} w_{i,j}^c}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{j=1}^{n_i} w_{i,j}^c \left(z_{i,j} - \mu_i^+\right)\left(z_{i,j} - \mu_i^+\right)^T}{\sum_{j=1}^{n_i} w_{i,j}^c}$$

$$i \in \{1, L\} \tag{8}$$

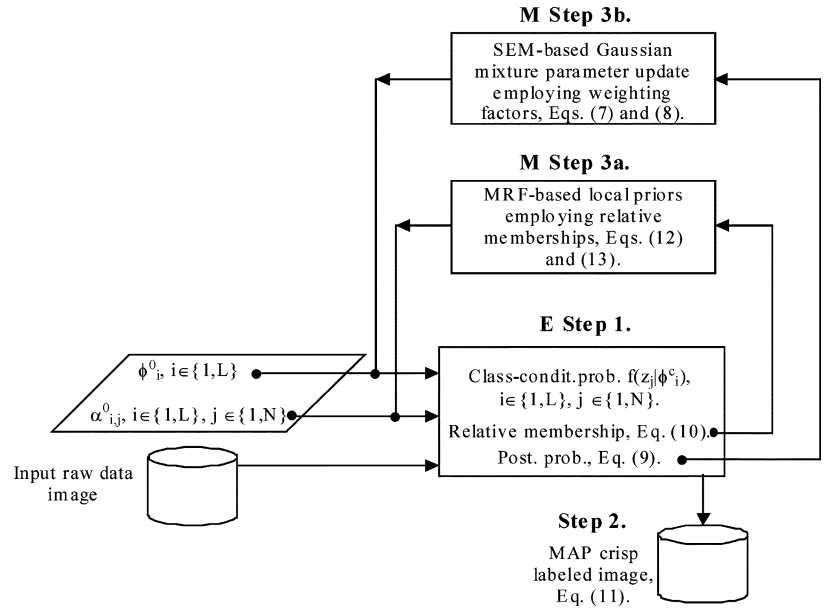


Fig. 6. Block diagram of the implemented "enhanced" CSEM algorithm.

$$r^c_{i,j} = \frac{f\left(z_j|\phi^c_i\right)}{\sum_{k=1}^{L} f\left(z_j|\phi^c_k\right)} = (4), \qquad r^c_{i,j} \in [0,1]$$

$$\sum_{i=1}^{L} r^c_{i,j} = 1, \qquad i \in \{1,L\}, j \in \{1,N\}. \tag{10}$$

2)$_{\mathrm{CSEM}}$. Crisp labeling based on the MAP assignment rule

$$x_j \in \text{class label } i, \text{i.e., (unlabeled)} z_j \Rightarrow \text{(semilabeled)} z_{i,j}$$
$$\text{to be employed in Equations (7) and (8)}$$
$$\text{if } i = \arg\max_{1 \le h \le L} \left\{\tau^c_{h,j}\right\},\ j \in \{1,N\}. \tag{11}$$

3a)$_{\mathrm{CSEM}}$. M-Step: update Gaussian mixture parameters (mean vectors and covariance matrices) according to (7) and (8), where weighting factors of semilabeled samples are computed as MAP posterior probabilities by (9).

3b)$_{\mathrm{CSEM}}$ M-Step: MRF-based updating of local (per-pixel) prior probabilities $\alpha^+_{i,j}$, $i \in \{1,L\}$, $j \in \{1,N\}$, computed according to [45], where posterior probabilities are replaced with ML relative memberships, $r^c_{i,j}$, computed by (10), such that

$$\alpha^+_{i,j} = \frac{\exp\left(\sum_{p \in \mathrm{Neigh}_{j,8}} \varepsilon_{i,(j,p)} \cdot r^c_{i,p} + \beta_i \cdot r^c_{i,j}\right)}{\sum_{h=1}^{L} \exp\left(\sum_{p \in \mathrm{Neigh}_{j,8}} \varepsilon_{i,(j,p)} \cdot r^c_{h,p} + \beta_i \cdot r^c_{h,j}\right)}, \qquad i \in \{1,L\}, j \in \{1,N\} \tag{12}$$

where the two-point clique potential $\varepsilon_{i,(j,p)}$ is defined as [45]

$$\varepsilon_{i,(j,p)} = \begin{cases} +\beta_i, & \text{if pixel pair } (j,p) \text{ is aligned either horizontally or vertically,} \\ +\frac{\beta_i}{\sqrt{2}}, & \text{if pixel pair } (j,p) \text{ is aligned neither horizontally nor vertically.} \end{cases} \tag{13}$$

In (13), category-specific two-point clique potentials β_i, $i \in \{1,L\}$, are user-defined, to enforce spatial continuity in pixel labeling.

4)$_{\mathrm{CSEM}}$ Check for convergence. If convergence is reached, stop. Otherwise: $c = c + 1$, and goto Step 1)$_{\mathrm{CSEM}}$.

MSEM for Image Clustering and Classification: MSEM aims at improving MPAC, which is susceptible to detecting image artifacts, by means of a learning strategy quite different from MPACBs. To combine the MPAC capability of detecting genuine but small image details with the SEM ability to mitigate the Hughes phenomenon, MSEM is conceived as a heuristic combination of the SEM class-conditional parameter update equations with a soft-competitive version of the multiscale objective function adopted by MPAC [refer to (2) and (3)]. Let us identify with $\mathrm{Neigh}_{j,s}$ the adaptive window chosen at spatial scale $s \in \{1,S\}$ and centered on the jth pixel, $j \in \{1,N\}$ (refer to Fig. 5). In the place of symbol $W_{j,s}$ adopted by the crisp-competitive MPAC algorithm (see Fig. 5), symbol $\mathrm{Neigh}_{j,s}$ is adopted herein, to identify a neighborhood centered on pixel j, at spatial scale s, featuring soft (relative), rather than crisp (hard, binary), membership values. The proposed MSEM algorithm consists of the following blocks.

0)$_{\mathrm{MSEM}}$ Initialize Gaussian mixture parameters ϕ^c_i, $i \in \{1,L\}$. Set $c = 0$.

1)$_{\mathrm{MSEM}}$ E-Step: compute class-conditional probabilities, $f(z_j|\phi^c_i)$, $i \in \{1,L\}$, $j \in \{1,N\}$, and weighting factors $w^c_{i,j}$, equivalent to relative memberships $r^c_{i,j}$ (that employ neither global nor local priors), $i \in \{1,L\}$, $j \in \{1,N\}$, computed via (4) of SEM.

2)$_{\mathrm{MSEM}}$ Per-pixel crisp labeling based on an objective function maximization where multiscale, class-specific intensity averages are weighted by their reliability factors. Compute the following:

$$\text{absolute membership } \eta^c_{i,j,s} = \frac{1}{\mathrm{EuclDis}^c_{i,j,s} + 1} \in [0,1], \qquad i \in \{1,L\}, j \in \{1,N\}, s \in \{1,S\} \tag{14}$$

where

$$\text{EuclDis}^{c}_{i,j,s} = \left\| z_j - \mu^{c}_{i,j,s} \right\|^2 \tag{15}$$

such that symbol $||.||$ identifies the Euclidean distance (L_2-norm), and

$$\text{intensity average estimate } \mu^{c}_{i,j,s} = \frac{\sum_{p \in \text{Neigh}_{j,s}} r^{c}_{i,p} x_p}{\text{SumR}^{c}_{i,j,s}} \tag{16}$$

where $\text{Neigh}_{j,s}$ is the neighborhood centered on pixel j at spatial scale s, $r^{C}_{i,p}$ is a relative membership value computed according to (4), while normalization factor $\text{SumR}^{C}_{i,j,s}$ is computed as

$$\text{SumR}^{c}_{i,j,s} = \sum_{p \in \text{Neigh}_{j,s}} r^{c}_{i,p}, \text{ such that}$$

$$\sum_{i=1}^{L} \text{SumR}^{c}_{i,j,s} = |\text{Neigh}_{j,s}| \text{ holds true}$$

$$\text{with } j \in \{1, N\}, s \in \{1, S\} \tag{17}$$

where $|\text{Neigh}_{j,s}|$ is the cardinality of neighborhood $\text{Neigh}_{j,s}$. Moreover, reliability factors of multiscale class-specific intensity averages are computed as

$$\text{reliability factor } \text{RF}^{c}_{i,j,s} = \frac{1}{|\text{Neigh}_{j,s}|} \text{SumR}^{c}_{i,j,s}$$

$$\text{such that } \text{RF}^{c}_{i,j,s} \in [0, 1]$$

$$\sum_{i=1}^{L} \text{RF}^{c}_{i,j,s} = 1, \qquad j \in \{1, N\}, s \in \{1, S\}. \tag{18}$$

Equations (14) and (18) are combined into the MSEM objective function as follows:

$$x_j \in \text{class label } i, \text{i.e., (unlabeled)} z_j \Rightarrow \text{(semilabeled)} z_{i,j}$$

see Equations (7) and (8)

$$\text{if } i = \arg\max_{1 \le h \le L} \left\{ \sum_{s=1}^{S} \text{RF}^{c}_{h,j,s} \cdot \eta^{c}_{h,j,s} \right\}, \; j \in \{1, N\}. \tag{19}$$

Thus, the MSEM objective function, (19), consists of a soft (weighted) combination of multiscale category-specific intensity averages, where weighting coefficients are the estimates' reliability factors. These reliability factors take their inspiration from those adopted in multitemporal/multisource optimization problems, where data sources are weighted depending on their different discrimination ability (e.g., refer to [48]). In the case of the MSEM objective function, the role of reliability factors is to measure the degree of compatibility of class-specific statistics estimated at local spatial scales, inherently prone to the small sample size problem, with class-specific statistics estimated at the global (image-wide) spatial scale. In other words, during pixel labeling, MSEM requires multiscale class-specific intensity averages to be consistent through scale. It is noteworthy that objective function (19) employs absolute rather than relative memberships [computed by (4)] to avoid the well-known "probabilistic (relative) membership problem." From fuzzy set theory, it is well known that an outlier tends to have small "possibilistic" (absolute) membership values with respect to all category prototypes (models), while its "probabilistic" (relative) membership values may be high [49]–[52].

3)$_{\text{MSEM}}$ M-Step: update Gaussian mixture parameters according to SEM's equations (7) and (8).

4)$_{\text{MSEM}}$ Check for convergence. If convergence is reached, stop. Otherwise: $c = c + 1$, and goto Step 1)$_{\text{MSEM}}$.

Potential advantages and limitations of MSEM with respect to its most similar alternatives, namely, SEM, CSEM, and MPAC, are expected to be the following: a first competitive advantage of MSEM over MPAC is that objective functions (14)–(19), consisting of a weighted combination of class-specific multiscale intensity average estimates, should avoid the detection of artifacts (which rather affects MPAC as it requires no consistency between interscale category-specific mean intensity estimates, see Appendix I); a second competitive advantage of MSEM over MPAC is that the former applies to both unsupervised and supervised image labeling tasks, i.e., MSEM can be employed with or without a reference labeled dataset; in the case of supervised learning tasks, MSEM is expected to mitigate the small sample size problem, in line with SEM and CSEM, by adopting Gaussian distribution parameter update (7) and (8); another interesting feature of MSEM is that by combining MPACs with SEM's learning strategies it pursues robust statistics estimation at local as well as global (image-wide) spatial scales. In particular: 1) MSEM employs multiscale intensity averages which are less sensitive than variance to the small sample size problem (see Fig. 5), in line with MPAC, and 2) MSEM exploits semilabeled samples to mitigate the small sample size problem in the estimation of Gaussian mixture parameters at the global (image-wide) scale, in line with SEM.

A first disadvantage of MSEM with respect to SEM, CSEM, and MPAC is its superior computational load. A second drawback is that, unlike SEM, it benefits from no rigorous statistical foundation. In fact, per-pixel crisp labeling equations [(14)–(19)] and the Gaussian mixture parameter update rules [(7) and (8)] are based on heuristics rather than being derived from an objective function minimization, e.g., see (6).

ACKNOWLEDGMENT

P. C. Smits, as a member of the GRSS-DFC, is acknowledged for providing us with the grss_dfc_0004 Landsat TM image. The authors wish to thank the Associate Editor and anonymous reviewers for their helpful comments.

REFERENCES

[1] S. Krishnamachari and R. Chellappa, "Multiresolution Gauss-Markov random filed models for texture segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 251–267, Feb. 1997.

[2] Q. Jackson and D. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.

[3] B. M. Shahshahani, "Classification of multispectral data by joint supervised-unsupervised learning," Ph.D. dissertation, School Elect. Eng., Purdue Univ., West Lafayette, IN, Jan. 1994. Tech. Rep. TR-EE-94-1. [Online]. Available: http://dynamo.ecn.purdue.edu/~landgreb/publications.html.

[4] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.

[5] G. Patanè and M. Russo, "The enhanced-LBG algorithm," *Neural Netw.*, vol. 14, no. 9, pp. 1219–1237, 2001.

[6] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Process.*, vol. 3, no. 2, pp. 162–177, Feb. 1992.

[7] A. Baraldi, P. Blonda, F. Parmiggiani, and G. Satalino, "Contextual clustering for image segmentation," *Opt. Eng.*, vol. 39, no. 4, pp. 1–17, Apr. 2000.

[8] Y. Jhung and P. H. Swain, "Bayesian contextual classification based on modified M-estimates and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 67–75, Jan. 1996.

[9] C. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Proces.*, vol. 3, no. 2, pp. 162–177, Feb. 1994.

[10] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[11] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, 1991.

[12] E. Binaghi, I. Gallo, and I. Pepe, "A cognitive pyramid for contextual classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2906–2922, Dec. 2003.

[13] P. Zamperoni, "Plus ça va, moins ça va," *Pattern Recognit. Lett.*, vol. 17, no. 7, pp. 671–677, 1996.

[14] R. C. Jain and T. O. Binford, "Ignorance, myopia and naivetè in computer vision systems," *Comput. Vision, Graphics, Image Process.: Image Understanding*, vol. 53, pp. 112–117, 1991.

[15] M. Kunt, "Comments on "dialogue," a series of articles generated by the paper entitled "Ignorance, myopia and naivetè in computer vision systems'," *Comput. Vision, Graphics, Image Process.: Image Understanding*, vol. 54, pp. 428–429, 1991.

[16] M. Sgrenzaroli, A. Baraldi, G. De Grandi, H. Eva, and F. Achard, "A novel operational approach to tropical vegetation mapping at regional scale using the GRFM radar mosaics," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 11, pp. 2654–2668, Nov. 2004.

[17] A. Baraldi, M. Sgrenzaroli, and P. Smits, "Contextual clustering with label backtracking in remotely sensed image applications," in *Geospatial Pattern Recognition*, E. Binaghi, P. Brivio, and S. Serpico, Eds. Kerala, India: Research Signpost/Transworld Research, 2002, pp. 117–145.

[18] A. Baraldi, L. Bruzzone, and P. Blonda, "Quality assessment of classification and cluster maps without ground truth knowledge," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 857–873, Apr. 2005.

[19] ——, "A multiscale expectation-maximization semisupervised classifier suitable for badly posed image classification," *IEEE Trans. Image Process.*, 2004, submitted for publication.

[20] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[21] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory, and Methods*. New York: Wiley, 1998.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[23] J. T. Morgan, A. Henneguelle, J. Ham, J. Ghosh, and M. M. Crawford, "Adaptive feature spaces for land cover classification with limited ground truth," *Int. J. Pattern Recognit. Artif. Intell.*, 2003, to be published.

[24] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL: Lewis, 1999.

[25] Q. Jackson and D. Landgrebe, "An adaptive method for combined covariance estimation and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1082–1087, May 2002.

[26] R. Kothari and V. Jain, "Learning with a minimum amount of labeling effort," *IEEE Trans. Neural Netw.*, 2006, to be published.

[27] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[28] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," presented at the *Int. Joint Conf. Artificial Intelligence*, 1995.

[29] E. Backer and A. K. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-3, no. 1, pp. 66–75, Jan. 1981.

[30] B. Fritzke. (1997) Some competitive learning methods. [Online]. Available: http://www.ki.inf.tu-dresden.de/~fritzke/JavaPaper.

[31] P. H. Swain and S. M. Davis, *Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 1978.

[32] A. K. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[33] T. Lillesand and R. Kiefer, *Remote Sensing and Image Interpretation*, 3rd ed. New York: Wiley, 1994.

[34] R. Lunetta and D. Elvidge, *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*. London, U.K.: Taylor & Francis, 1999.

[35] L. Delves, R. Wilkinson, C. Oliver, and R. White, "Comparing the performance of SAR image segmentation algorithms," *Int. J. Remote Sens.*, vol. 13, no. 11, pp. 2121–2149, 1992.

[36] G. M. Foody, "Thematic mapping from remotely sensed data with neural networks: MLP, RBF and PNN based approaches," *J. Geograph. Syst.*, vol. 3, pp. 217–232, 2001.

[37] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 4, pp. 2113–2118, Jul. 1999.

[38] *ENVI User's Guide*, Res. Syst. Inc., Boulder, CO, 2003.

[39] Q. Jackson and D. A. Landgrebe, "Adaptive Bayesian contextual classification based on a Markov random field," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.

[40] L. Bruzzone, F. Roli, and S. Serpico, "Structured neural networks for signal classification," *Signal Process.*, vol. 64, pp. 271–290, 1998.

[41] M. M. Durat and D. A. Landgrebe, "A cost-effective semisupervised classifier approach with kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[42] D. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, pp. 109–118, 1990.

[43] E. Rignot and R. Chellappa, "Segmentation of polarimetric synthetic aperture radar data," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 281–300, Mar. 1992.

[44] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Ser. B*, vol. 39, pp. 1–38, 1977.

[45] J. Dehemedhki, M. F. Daemi, and P. M. Mather, "An adaptive stochastic approach for soft segmentation of remotely sensed images," in *Series in Remote Sensing: 1 (Proc. of the Int. Workshop on Soft Computing in Remote Sensing Data Analysis, Milan, Italy, Dec. 1995)*, E. Binaghi, P. A. Brivio, and A. Rampini, Eds, Singapore: World Scientific, 1996, pp. 211–221.

[46] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, Jun. 1984.

[47] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[48] A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.

[49] R. N. Davè and R. Krishnapuram, "Robust clustering method: A unified view," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 2, pp. 270–293, May 1997.

[50] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks—Part I," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 645–661, May 2002.

[51] ——, "Constructive feedforward ART clustering networks—part II," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 662–677, May 2002.

[52] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[53] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[54] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

Andrea Baraldi was born in Modena, Italy, in 1963. He received the laurea degree in electronic engineering from the University of Bologna, Bologna, Italy, in 1989. His master's thesis focused on the development of unsupervised clustering algorithms for optical satellite imagery.

He is currently with the IPSC-SES unit of the Joint Research Centre, Ispra, Italy, involved with optical and radar image interpretation for terrestrial surveillance and change detection.From 1989 to 1990, he was a Research Associate at CIOC-CNR, an Institute of the National Research Council, Bologna, and served in the army at the Istituto Geografico Militare in Florence, working on satellite image classifiers and GIS. As a consultant at ESA-ESRIN, Frascati, Italy, he worked on object-oriented applications for GIS from 1991 to 1993. From December 1997 to June 1999, he joined the International Computer Science Institute, Berkeley, CA, with a postdoctoral fellowship in artificial intelligence. From 2000 to 2002, he was a Post-Doc Researcher with the European Commission Joint Research Center, Ispra, Italy, where he worked on the development and validation of algorithms for the automatic thematic information extraction from wide-area radar maps of forest ecosystems. Since his master thesis, he has continued his collaboration with ISAC-CNR, Bologna and ISSIA-CNR, Bari, Italy.His main interests center on image segmentation and classification, with special emphasis on texture analysis and neural network applications in computer vision.

Dr. Baraldi is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS.

Lorenzo Bruzzone (S'95–M'99–SM'03) received the laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently Head of the Remote Sensing Laboratory in the Department of Information and Communication Technologies at the University of Trento, Trento, Italy. From 1998 to 2000, he was a Postdoctoral Researcher at the University of Genoa. From 2000 to 2001, he was an Assistant Professor at the University of Trento, and from 2001 to February 2005 he was an Associate Professor of telecommunications at the same university. Since March 2005, he has been a Full Professor of telecommunications at the University of Trento, where he currently teaches remote sensing, pattern recognition, and electrical communications. His current research interests are in the area of remote sensing image processing and recognition (analysis of multitemporal data, feature selection, classification, regression, data fusion, and neural networks). He conducts and supervises research on these topics within the frameworks of several national and international projects. He is the author (or coauthor) of more than 140 scientific publications, including journals, book chapters, and conference proceedings. He is a referee for many international journals and has served on the Scientific Committees of several international conferences. He is a member of the Scientific Committee of the India–Italy Center for Advanced Research.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was a recipient of the *Recognition of IEEE Transactions on Geoscience and Remote Sensing Best Reviewers* in 1999 and was a Guest Editor of a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of multitemporal remote sensing images (November 2003). He was the General Chair and Co-chair of the First and Second, respectively, IEEE International Workshop on the Analysis of Multitemporal Remote-Sensing Images. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. From 2004 to 2005, he was an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. Since 2005, he has been an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a member of the International Association for Pattern Recognition (IAPR) and of the Italian Association for Remote Sensing (AIT).

Palma Blonda (M'93) received the Ph.D. degree in physics from the University of Bari, Bari, Italy, in 1980.

In 1984, she joined the Institute for Signal and Image Processing (now the Institute of Intelligent Systems for Automation), Italian National Research Council (ISSIA-CNR), Bari. Her research interests include digital image processing, fuzzy logic and neural networks, and soft computing applied to the integration and classification of multisource remote sensed data. She has recently been involved with the Landslide Early Warning Integrated System (LEWIS) Project, founded by the European Comunity in the framework of Fifth PQ. In this project, her research activity focuses on the application of multisource data integration and classification techniques for the extraction of EO-detectable superficial changes of some landslide-related factors to be used in early-warning mapping.

Lorenzo Carlin received the laurea (B.S.) and the laurea specialistica (M.S.) degrees in telecommunication engineering (summa cum laude) from the University of Trento, Trento, Italy, in 2001 and 2003, respectively. He is currently pursuing the Ph.D. degree in information and communication technologies at the University of Trento.

He is with the Pattern Recognition and Remote Sensing group, Department of Telecommunication and Information Technologies, University of Trento. His main research activity is in the area of pattern recognition applied to remote sensing images. In particular, his interests are related to classification of very high resolution remote sensing images. He conducts research on these topics within the frameworks of several national and international projects.

Mr. Carlin is a referee for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.