Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Remote Sensing of Environment 103 (2006) 179 – 189

Remote Sensing of Environment

www.elsevier.com/locate/rse

The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM

Giles M. Foody [a,*], Ajay Mathur [b]

[a] *School of Geography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[b] *Punjab Remote Sensing Centre, PAU Campus, Ludhiana-141 004, India*

Received 16 December 2005; received in revised form 29 March 2006; accepted 2 April 2006

Abstract

The accuracy of a supervised image classification is a function of the training data used in its generation. It is, therefore, critical that the training stage of a supervised classification is designed to provide the necessary information. Guidance on the design of the training stage of a classification typically calls for the use of a large sample of randomly selected pure pixels in order to characterise the classes. Such guidance is generally made without regard to the specific nature of the application in-hand, including the classifier to be used. The design of the training stage should really be based on the classifier to be used since individual training cases can vary in value as can any one training set to a range of classifiers. It is argued here that the training stage can be designed on the basis of the way the classifier operates and with emphasis on the desire to separate the classes rather than describe them. An approach to the training of a support vector machine (SVM) classifier that is the opposite of that generally promoted for training set design is suggested. This approach uses a small sample of mixed spectral responses drawn from purposefully selected locations (geographical boundaries) in training. The approach is based on mixed pixels which are normally masked-out of analyses as undesirable and problematic. A sample of such data should, however, be easier and cheaper to acquire than that suggested by conventional approaches. This new approach to training set design was evaluated against conventional approaches with a set of classifications of agricultural crops from satellite sensor data. The main result was that classifications derived from the use of the mixed spectral responses and the conventional approach did not differ significantly, with the overall accuracy of classifications generally ~92%.

Keywords: Training set; Mixed pixel; Support vector machine; Classification

1. Introduction

Supervised classification is one of the most commonly undertaken analyses in remote sensing. The training stage of a supervised image classification provides the class descriptors upon which all allocations are based. Clearly, the quality of the training data set used is of fundamental importance to a classification and a major determinant of classification accuracy. Many studies have shown that the accuracy of a classification varies as a function of a range of training set properties (Fardanesh & Ersoy, 1998; Foody, 1999; Foody & Arora, 1997; Foody et al., 1995; Staufer & Fischer, 1997; Tsai & Philpot, 2002; Zhuang et al., 1994). The nature of an ideal training set is, however, unclear, possibly because of uncertainty about the aim of the training stage.

Fundamentally, the training stage seeks to provide descriptive statistics for each class in the image that may be used to direct the accurate determination of class membership by the selected classifier. The typical approach to training adopted involves the acquisition of a sample of pixels of known class membership from the image to characterise the classes. These selected pixels represent the training set upon which the remainder of the classification analysis is based. Typically, the descriptive statistics derived from the training pixels are used to characterise the classes and ultimately convey the information needed to partition feature space

[*] Corresponding author.
E-mail address: g.m.foody@soton.ac.uk (G.M. Foody).

0034-4257/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2006.04.001

so that class membership may be determined for all image pixels. Obtaining an accurate description of each class is seen as fundamental to the derivation of an accurate classification (Kuo & Landgrebe, 2002; Mather, 2004).

The way in which the training data are used varies between classifiers. For example, the maximum likelihood classifier uses parameters such as the mean and covariance matrix that summarise the spectral response of each class while a multi-layer perceptron neural network uses each training case directly. Classifiers can, therefore, differ greatly in terms of the training information they require for an accurate classification. A training set that could be used to derive a highly accurate classification from one classifier may yield a considerably lower accuracy if used with another classifier (Foody, 1999). Consequently, the nature of the classifier selected for a particular application should inform the design of the training data collection programme. The literature, however, generally promotes a relatively uniform or classifier-independent approach to training set design.

The ruling paradigm in the design of the training stage appears to be strongly based on conventional statistics. In this, the training stage is viewed as aiming to derive an accurate description of each class. The descriptive statistics derived may then be used to allocate each pixel of unknown membership to the class with which it has the greatest similarity. Thus, for example, with the maximum likelihood classifier, descriptive statistics that summarize the spectral response of each class are used to determine the most likely class of membership for all pixels in the image. Critically, therefore, key issues in the design of the training stage relate to basic sampling theory and the literature urges the use of rigorous statistical approaches to derive the class descriptive statistics. Major concerns with the use of such an approach are that the training sample acquired should provide a representative and unbiased description of the classes. Thus, some variant of random sampling is typically promoted to help ensure the sample acquired satisfies the assumed conditions. In addition, the sample size required to characterise the spectral response of a class may, assuming it follows a normal distribution, be calculated from

$$n = \frac{\sigma^2 z^2}{h^2 + \frac{\sigma^2 z^2}{N}} \tag{1}$$

where h is a specified half-width of the confidence interval, σ is the planning or estimated value for the population standard deviation, z is the value of the z score at a specified level of confidence and N the size of the population. This equation indicates the sample size required to estimate the mean value of a distribution with a specified degree of precision. For large populations, which given the number of pixels in a typical image data set will generally be the case in remote sensing applications, the equation tends toward that for an infinite population which is

$$n = \frac{\sigma^2 z^2}{h^2} \tag{2}$$

The theory underlying this relationship may also be used to estimate the sample size required for the estimation of a proportion and has been used commonly as the basis for determining sample size requirements for classification applications in remote sensing (e.g. Nelson et al., 1984; Todd et al., 1980). Alternatively, simple heuristics are often used to determine training sample size (Mather, 2004; Van Niel et al., 2005). For example, it is often stated that the number of training samples for each class should comprise at least 10–30 times the number of wavebands, or other discriminating variables, used in the analysis (Mather, 2004; Piper, 1992) with a 'the larger the better' attitude often held.

Deviations from the widely used approaches outlined above tend to promote a focus on pure cases of the classes. For example, it is, of course, desirable that only pixels that actually represent an area of the class being described are used in its description. Thus, the pixels selected for training purposes should be pure members of the relevant classes. To achieve this researchers often deliberately mask out or exclude boundary regions where the mixing of class spectral responses may occur (e.g. Airkan, 2004). Furthermore, some analysts apply post-acquisition refinement operations to the training data that remove outliers or down-weight the contribution of cases perceived to be atypical of the class being characterised (Aria, 1992; Buttner et al., 1989; Ediriwickrema & Khorram, 1997; Mather, 2004). Alternatively, in acquiring training data, some researchers use seed functions (e.g. Sun et al., 2003), which do not allow the inclusion of pixels with values greatly different to the seed, to acquire the training set. The use of these post-acquisition refinement operations and seed functions will generally act to shift focus towards the purest exemplars of the classes, what may often be considered to be end members.

As a crude summary, the target in training has conventionally been to acquire a large sample of pure pixels to describe the classes. Moreover, it is typically argued that the bigger the training set the better, as this will increase the precision of the estimates made. This latter issue is evident from Eqs. (1) and (2) where the half-width of the confidence interval, which indicates the precision of the estimate, is inversely related to the sample size. While obtaining an accurate and representative description is typically seen as desirable some approaches for training set definition or refinement adopted may, however, act to bias the description by placing emphasis on cases typical of the class centroid. More fundamentally, however, the basis of the conventional approach to training set design may not be focused on the provision of the critical information needed for an accurate classification.

In this article, it is suggested that the conventional paradigm to the design of training data collection programmes may not always be the most appropriate to adopt. Critically, it is argued that the conventional paradigm appears to be focused on describing the classes and in particular the purest cases of the classes. Not only is this approach biased to the purest cases (and to the class centroid) but it also is not meeting the fundamental aim of training. The aim of training is not to accurately describe the classes but to provide information on the classes that will aid the fitting of classification decision boundaries or hyperplanes to separate them. From this perspective, alternative approaches to training may be promoted and are classifier-dependent. Here,

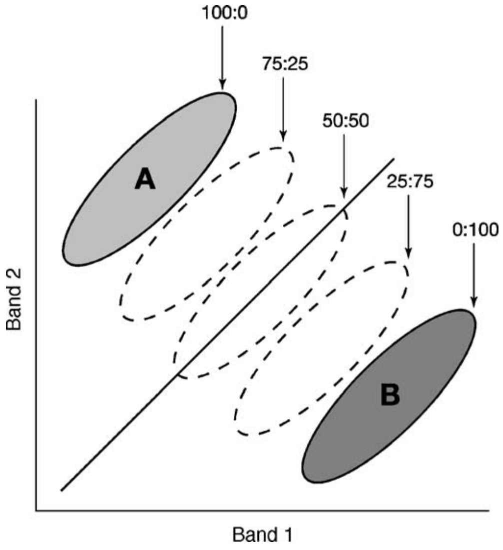

Fig. 1. A simplified view of the location of pure and mixed class spectral responses in a feature space. Note that between the distributions of the two pure classes, A and B, are distributions associated with simple spectral mixtures of the classes and the hyperplane that optimally separates the classes. Class composition is indicated as the percentage cover of class A:Class B.

we argue that an approach which is essentially the direct opposite to that conventionally used may be employed to yield accurate image classifications. This is argued with reference to classification by a support vector machine (SVM), which has been shown to be at least as accurate as other widely used classifiers in remote sensing (Foody & Mathur, 2004a; Huang et al., 2002).

With a SVM the most useful training cases are those that lie close to where the hyperplane is to be fitted (Foody & Mathur, 2004b; Vapnik, 1995). The hyperplane, of course, lies between the classes and so in the region in which mixtures of the class spectral responses occur (Fig. 1). On this basis, mixed spectral responses and mixed pixels may represent the useful and informative training cases for a classification. Moreover, as these are the informative training cases and others, located closer to the class centroids, may have no impact at all on the fitting of the hyperplane only a very small training sample may be required for accurate classification (Foody & Mathur, 2004b; Pal & Mather, 2005). Thus, for classification by techniques such as SVM we propose that a small training sample, based around mixed pixels, can be used to derive an accurate classification. This proposal will be evaluated by a series of SVM classifications of a remotely sensed data set.

2. SVM classification

SVM based approaches have considerable potential for the supervised classification of remotely sensed data. Comparative studies have shown that classification by a SVM can be more accurate than popular contemporary techniques such as neural networks and decision trees as well as conventional probabilistic classifiers such as the maximum likelihood classification (Foody & Mathur, 2004a; Huang et al., 2002; Melgani & Bruzzone, 2004). Although SVMs were designed for binary classification various methods exist to extend the binary approach to multi-class classification (Hsu & Lin, 2002; Huang et al., 2002; Melgani & Bruzzone, 2004). In all situations, however, the fundamental nature of the SVM is the same.

Classification by a SVM is based on the fitting of an optimal separating hyperplane between classes by focusing on the training samples that lie at the edge of the class distributions and between the class centroids, the support vectors (Fig. 2). All of the other training samples are effectively discarded as they do not contribute to the estimation of hyperplane location (Belousov et al., 2002; Brown et al., 2000; Wang et al., 2005). Thus, only the training samples that lie close to the location of the optimal separating hyperplane are used in its establishment. With SVM based classification, therefore, not only is an optimal hyperplane fitted, in the sense that it is expected to have a large degree of generalizability, but also a high accuracy may be obtained with the use of a small training set (Foody & Mathur, 2004b; Wang et al., 2005). Given the costs of training data acquisition is often noted as a concern (e.g. Tadjudin and Landgrebe, 2000; Chi and Bruzzone, 2005), the potential to limit training set size may be an advantageous feature, especially if it is achieved without any significant loss of discriminatory power in the final classification.

In order to appreciate the potential value of mixed pixels as a resource in training it may be helpful to outline the basic features of classification by a SVM. Extensive discussion of

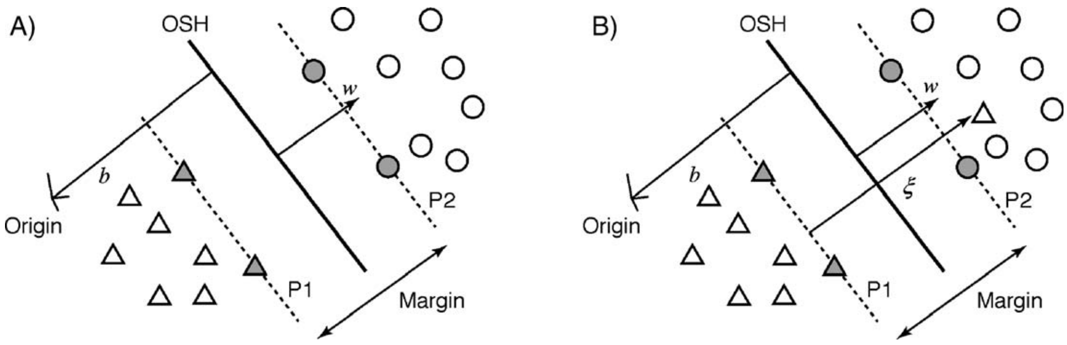

Fig. 2. Basic issues in classification SVM for (A) linearly separable and (B) partially inseparable situations. Note that the support vectors have been highlighted in grey and lie on two hyperplanes that are parallel to the optimal separating hyperplane (OSH).

classification by a SVM is available in the literature (e.g. Huang et al., 2002; Melgani & Bruzzone, 2004; Vapnik, 1995) but here we draw on an earlier discussion (Foody & Mathur, 2004a) to highlight the key issues. The simplest starting point for outlining the nature of SVM based classification is to consider the situation in which there are two classes that are linearly separable in q dimensional space. For each of the r training cases there is a vector, $\mathbf{x}_i$, that represents the spectral response of the case and its location in feature space together with a definition of class membership, y_i. Using the training data represented by $\{\mathbf{x}_i, y_i\}$, $i=1,\ldots,r$, $y_i \in \{1, -1\}$, the aim is to develop a classifier that generalizes accurately. For this, a decision boundary or hyperplane that separates the classes in feature space is required. A large number of candidate hyperplanes could be fitted to separate the classes but there is only one optimal separating hyperplane, which is expected to generalize well in comparison to other possible hyperplanes. This optimal separating hyperplane should lie between the two classes in feature space and be positioned such that all of the samples of a class are located to one side of it and it is located such that the distance to the closest training data samples in both of the classes is as large as possible. In that way the margin between the two classes is maximised.

A hyperplane in feature space is defined by the equation $\mathbf{w}\cdot\mathbf{x}+b=0$, where $\mathbf{x}$ is a point lying on the hyperplane, $\mathbf{w}$ is normal to the hyperplane and b is the bias (Fig. 2). A separating hyperplane can be defined for the two classes as: $\mathbf{w}\cdot\mathbf{x}_i+b\geq 1$ (for the class $y_i=+1$) and $\mathbf{w}\cdot\mathbf{x}_i+b\leq -1$ (for the class $y_i=-1$). These two equations may be combined to give

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b)-1\geq 0 \tag{3}$$

The training samples on these two hyperplanes are termed the support vectors and are central to the establishment of the optimal separating hyperplane on which SVM classification is based.

The support vectors of the two classes lie on two hyperplanes, P1 and P2, which themselves are parallel to the optimal hyperplane and are defined by $\mathbf{w}\cdot\mathbf{x}_i+b=\pm 1$ (Fig. 2). The margin between these planes is $\frac{2}{|\mathbf{w}|}$ and the analysis aims to maximise this margin through the constrained optimization problem,

$$\min\left\{\frac{1}{2}|\mathbf{w}|^2\right\} \tag{4}$$

under the inequality constraints of Eq. (3).

If the classes are not linearly separable, slack variables, $\{\xi_i\}_{i=1}^{r}$, that indicate the distance the sample is from the hyperplane P1 or P2 passing through the support vectors of the class to which the sample belongs (Fig. 2), and so the amount of violation of the constraints allowed, are introduced (Cortes & Vapnik, 1995). With this addition, Eq. (3) may be rewritten as,

$$y(\mathbf{w}\cdot\mathbf{x}_i + b)>1-\xi_i \tag{5}$$

If outliers exist in the data set, Eq. (5) can always be satisfied by making ξ_i very large and so, a penalty term, $C\sum_{i=1}^{r}\xi_i$ is added to penalize solutions for which ξ_i are very large. The constant C controls the magnitude of the penalty associated with training samples that lie on the wrong side of the hyperplane. The value of C must be set by the analyst and requires careful selection to avoid problems such as over-fitting that may limit the classifier's generalization capacity. Although the parameter C must, therefore, be selected with care SVM based classification has been shown to display a large degree of robustness to variation in parameter values (Belousov et al., 2002). With the addition of the penalty term to the analysis, the optimization problem becomes,

$$\min\left[\frac{|\mathbf{w}|^2}{2} + C\sum_{i=1}^{r}\xi_i\right] \tag{6}$$

under the constraints of Eq. (5). Note that the first part of Eq. (6) seeks to maximize the margin between the classes while the second part aims to penalize the samples located on the incorrect side of the hyperplane with C controlling the relative balance of these two competing objectives. If the classes overlap considerably in feature space, then $C\sum_{i=1}^{r}\xi_i$ can be very large and the optimal hyperplane may be expected to have limited generalization ability.

The basic approach to SVM classification may be extended to allow for non-linear decision surfaces. For this situation, the input data are mapped into a high dimensional space through some non-linear mapping which has the effect of spreading the distribution of the data points in a way that facilitates the fitting of a linear hyperplane. The classification decision function is then

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{r}\alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{7}$$

where α_i, $i=1,\ldots,r$ are lagrange multipliers and $k(\mathbf{x},\mathbf{x}_i)$ is a kernel function (Wang et al., 2005). The magnitude of α_i is determined by the parameter C and indicates the contribution made by the training samples to the fitting of the optimal separating hyperplane with values and lying within the range of 0–C (Belousov et al., 2002). The kernel used must meet Mercer's condition (Vapnik, 1995) and one such kernel which is used in this research is the radial basis function,

$$k(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma|(\mathbf{x}-\mathbf{x}_i)|^2} \tag{8}$$

the width of which is controlled by the parameter γ.

The basic SVM approach for binary classification may be extended for multi-class classifications. Commonly, this has been achieved using either the one-against-one or one-against-all strategies in which the multi-class classification problem is reduced to a set of binary problems (Huang et al., 2002; Melgani & Bruzzone, 2004). Alternatively, however, a multi-class SVM may be used (Hsu & Lin, 2002). One multi-class SVM which is similar to the one-against-all approach involves the solution of a single optimisation problem and was used in the research reported below. With this approach, for an c class problem, c two class rules where the mth function $\mathbf{w}_m^T\varphi(\mathbf{x})+b$ separates the training data vectors of class m from that of others are constructed. Hence,

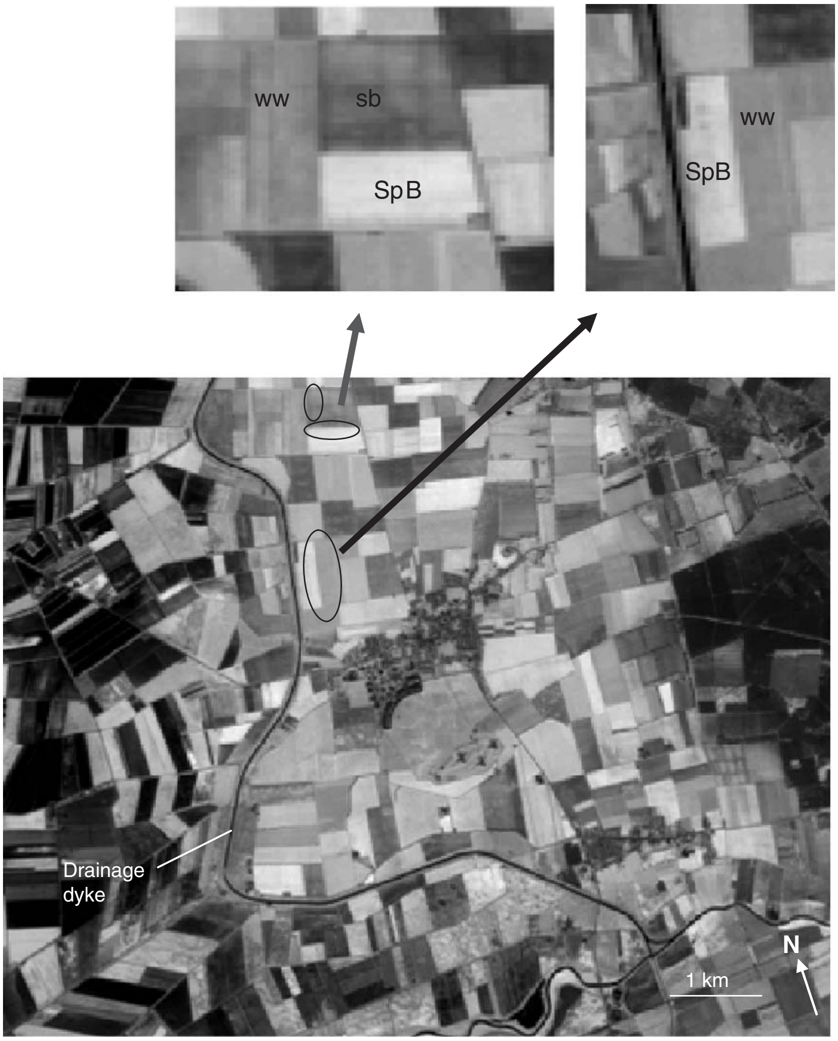

Fig. 3. SPOT HRV near-infrared image of the study area with the location of the three field boundaries used in the derivation of the mixed spectral response training set. Class labels are winter wheat (ww), spring barley (SpB) and sugar beet (sb).

there are c decision functions or hyperplanes but all are obtained by solving one problem,

$$\min_{w,b,\xi}\frac{1}{2}\sum_{m=1}^{c}\mathbf{w}_m^T\mathbf{w}_m+C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_{i,m}, \tag{9}$$

under the constraints,

$$\mathbf{w}_{y_i}^T\varphi(\mathbf{x}_i)+b_{y_i}\geq\mathbf{w}_m^T\varphi(\mathbf{x}_i)+b_m+2-\xi_{i,m},$$

$$\xi_{i,m}\geq 0, i=1,\ldots..,l, m\in\{1,\ldots.c\}\setminus y_i$$

where $i=1,\ldots,l$ are the training data vectors. The decision function is then,

$$\mathrm{argmax}_{m=1,\ldots c}(\mathbf{x}_\mathrm{i})+\mathrm{b}_\mathrm{m}) \tag{10}$$

By reducing the classification to a single optimization problem this approach may also require fewer support vectors than a multi-class classification based on the combined use of many binary SVMs (Hsu & Lin, 2002), an advantage when wishing to constrain training set size. Additionally, with the multi-class SVM approach to classification the values for the parameters C and γ need only to be defined once.

The accuracy with which a SVM may classify a data set is dependent on the magnitude of the parameters C and γ. With a large value of γ and/or C, there is a tendency for the SVM to over-fit to the training data, yielding a classifier that may generalize poorly. In this situation it may be possible to classify the training data set accurately but the accuracy with which an independent testing set is classified may be low. Since the interest in remote sensing is typically to derive a classifier from a limited training set which may be applied usefully to other samples, the generalization ability of the classifier is of fundamental importance and hence over-fitting is undesirable. Consequently, the magnitude of C and γ must be determined carefully for the task in-hand. To help reduce the subjectivity in setting the SVM parameter values, a cross-validation approach is commonly adopted for their determination (Belousov et al., 2002).

The critical feature to note about SVM classification, however, is that only the support vectors contribute to the fitting of the optimal separating hyperplane. It is, therefore, possible to derive an accurate classification from a small and unrepresentative sample of training cases. The strict demands imposed by random sampling are no longer necessary as samples can be deliberately selected by the analyst. To fully exploit this feature, a means to identify sites likely to furnish support vectors in advance of the classification is required (Foody & Mathur, 2004b). Such judgemental sampling can provide the training information needed for a classification and should be easier to acquire than a random sample. As a further refinement or as an alternative, the analyst could try and focus attention onto the general region around the expected location of a suitable hyperplane. In feature space, the support vectors lie on the edges of the class distributions and in between the class centroids. One would also expect to find mixed pixels, the spectral response of which being a composite of the pure class responses involved, lying in this region between the class centroids. Broadly one would expect that a mixed pixel comprising equal proportions of the classes would lie roughly mid-way between the class centroids (Fig. 1). Indeed this is the basis of using a SVM for spectral unmixing (Brown et al., 2000). The spectral response of mixed pixels of imbalanced class composition would generally be located between this mid-way position and the centroid of the class which made the dominant contribution to the pixel's spectral response, the precise position in feature space being a function of the proportional cover of the classes and getting closer to the class centroid the purer the pixel's composition. Critically, mixed spectral responses would generally be expected to lie between the class centroids and be closer than the class centroids to the location of a hyperplane that can accurately separate the classes. Moreover, the spectral response derived from an imbalanced mixture may be expected to lie close to locations of hyperplanes such as P1 and P2 used in SVM classification. One might expect, therefore, that mixed spectral responses are actually more useful than those from pure pixels in fitting a hyperplane. Furthermore, as only support vectors contribute to the fitting of the hyperplane it should be possible to use a small training set of cases comprised of mixed spectral responses to classify a data set as accurately as a large, conventionally defined, training set of pure cases. It should be possible, therefore, to use a small training set comprising of data on mixed spectral responses to derive an accurate classification. This paper seeks to test this hypothesised scenario and illustrate the potential for using mixed spectral responses in training a supervised image classifier.

3. Data and methods

The potential to focus the training stage of a SVM classification on mixed spectral responses was evaluated from a series of classification analyses of imagery acquired of an agricultural test site. The study area selected was located around the village of Feltwell, UK (Fig. 3). This region is topographically flat and most of the land has been divided up into large agricultural fields. For the purposes of this research, the effect of marked variations in soil type in the region on class separability was reduced by focusing on only land to the east of the drainage dyke that runs through the region. In this study area the fields had been planted mainly to sugar beet, winter wheat or spring barley on a brown renzina soil. All analyses focused on the classification of these three classes.

A three waveband multispectral SPOT HRV image of the test site acquired on 16 June 1986 was used for the analyses. This image provided data in three spectral wavebands: red, green and near-infrared (Fig. 3). No pre-processing of the image was undertaken with actions such as atmospheric correction viewed as unnecessary for the classification of a single image (Song et al., 2001) and the agricultural fields were readily identifiable in the imagery removing the need for a geometric correction. The classification analyses were, therefore, undertaken using image DN values. At the time of image acquisition, the winter wheat and spring barley classes were at a mature stage of growth, with typically >95% ground cover. However, the fields that had been planted to sugar beet were essentially bare at the time of image acquisition. A map showing the crop type that had been planted in each field that had been produced by conventional methods was available as ground data. For simplicity, the class membership of each field will be described in terms of the crop which had been planted within its area.

A variety of training sets were formed. To provide a basic benchmark to aid evaluation of the results, one training set contained 90 randomly selected pixels of each crop type. The pixels selected were all drawn from large areas of homogeneous class coverage. This training set corresponded to that which would be acquired through adoption of the standard heuristic that 30 training samples per-class per-waveband should be acquired for a classification. This is referred to later as the training set formed from the use of the 30*p* heuristic, where *p* indicates the number of wavebands used.

A training set based on mixed spectral responses was acquired. Here, the aim was to use a small number of mixed spectral responses to provide the training information for the classification. For this, 3 field boundaries were selected (Fig. 3). These boundaries were selected relatively arbitrarily, they were simply a set of easy to identify and reasonably long boundaries that, critically, lay between every possible combination of the classes. That is, the data set contained a boundary that separated each possible pair of crops. Thus, in this simple three class scenario, there was the potential to obtain information on the way the spectral response of each class mixed with each other class.

The boundaries between fields were clearly visible in the imagery and could be more objectively located by inspection of the DN values if required. For each selected boundary, a transect of pixels that ran along the length of the boundary was identified. The pixels in this transect would all be mixed pixels. The spectral response of these mixed pixels may convey all of the information that is needed to locate the hyperplane between the classes. However, this might require detailed and difficult to acquire information on the exact nature of the mixing (e.g. the proportion of the area represented by the pixel covered by the component classes). Here, an approach which requires no such detailed information was adopted. Specifically, the approach simply required mixed spectral responses. Critically, however, these mixed responses needed to be biased towards a class. There also needed to be cases for each mixing scenario in which a class was the major component of the mixture and cases when it was the minor component of the mixture. In this way two sets of mixed responses for the class pair would be derived. These mixed spectral responses were derived since it would be expected that the hyperplane in feature space to separate the pair of classes would lie between the locations of the two mixed spectral responses derived for that pair. To remove the need to know the class composition of the mixed pixels, the DN value of the pixel on the transect was averaged with the DN value of its neighbour to one side of the boundary in one of the fields. This yielded a mixed spectral response that was dominated by the class associated with the field from which the neighbouring pixel was drawn. This process was then repeated but by averaging the mixed pixel response with the corresponding pixel on the opposite side of the boundary. In this way, for each mixed pixel straddling a boundary, two mixed spectral responses were derived, one dominated by each of the classes separated by that boundary. It would be expected that the hyperplane to separate the classes would lie between these two sets of mixed spectral responses in feature space. Moreover, these mixed responses would be closer to the hyperplane than the responses of the pure classes and this may help in define an appropriate separating hyperplane. For training the classification itself, each mixed spectral response was associated fully with its dominant class (i.e. labelled as a member of the dominant class) in order to establish the location of the separating hyperplanes. A simple generalization of the process is given in Fig. 4.

The mixed spectral response data for training were generated from 3 boundaries. Thus only ground data on the land cover of a maximum of 6 fields was required. The actual number of fields required to provide the required number of boundary types is a function of the crop mosaic on the ground. For the simple 3 class scenario considered here it would be possible to need information on just 3 fields if appropriately located relative to

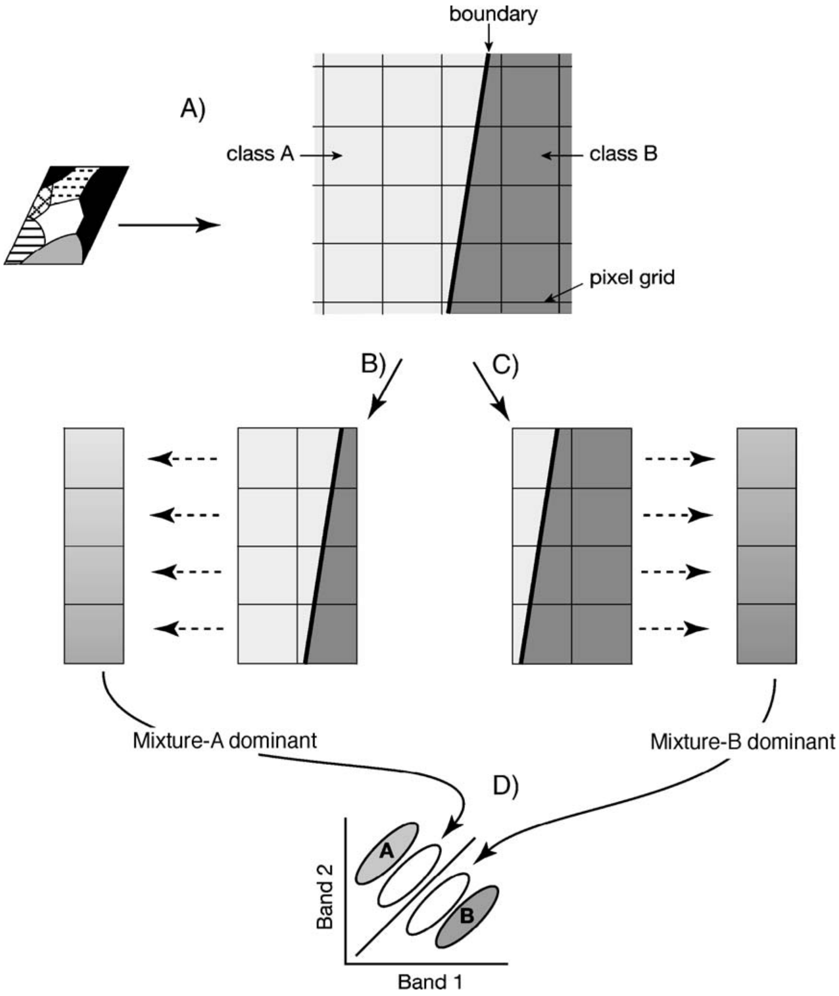

Fig. 4. Summary of the method based on mixed spectral responses. (A) An appropriate field boundary separating two classes, A and B, is identified in the image. A transect of mixed pixels that runs along the boundary is located. (B) For each pixel along this transect, its DN is averaged with its neighbouring pixel in the field containing class A. (C) For each pixel along the transect, its DN is averaged with its neighbouring pixel in the field containing class B. (D) From (B) and (C) two sets of mixed spectral response have been derived for the pair of classes. In the example shown there are four cases for each mixture. One set of mixed responses is dominated by the response of class A while the other is dominated by that of class B and these lie closer to the expected location of the hyperplane to separate the classes than the pure class distributions. The cases in are labelled by their dominant class and used to train the SVM.

Table 1
The SVM parameter settings used

Training set	C	γ
Conventional, 30p	8.0	2^{-5}
Mixed spectral response	8.0	2^{-8}
Conventional, same size	32.0	2^{-4}
Conventional, same fields	8.0	2^{-3}

each other. In this research, however, 5 fields were used to provide the required boundary information (Fig. 3). The boundaries selected varied in length from 12 to 28 pixels. The total boundary length involving winter wheat was 41 pixels while that for spring barley and sugar beet was 40 and 25 pixels, respectively. Given that the mixed spectral responses were derived by averaging the DN of the transect pixels with those from neigbouring pixels drawn from inside the fields, the total number of pixels associated with each class was 82, 80 and 50 for the winter wheat, spring barley and sugar beet classes, respectively (note that the mixed pixels along the transect were associated with two classes simultaneously). The data set derived from this process was referred to as the mixed spectral response training set.

To facilitate comparison between the classification trained with the mixed spectral responses and those trained with more conventional approaches, two further training sets were generated. These training sets were generated to provide classifications which could be compared directly, albeit from different perspectives, against that based on the mixed spectral responses. The key feature of these additional training sets was that their size, in terms of ground data needed, could be considered the same as that used in generating the mixed spectral response training set. First, a training set containing the same number of cases per-class as in that used in the derivation of the training set based on mixed spectral responses was generated. This contained 82, 80 and 50 randomly selected pure cases of winter wheat, spring barley and sugar beet. In keeping with conventional practice, all of the sampled pixels were drawn from large areas of homogeneous cover. The training set was, therefore, similar to that used to generate the benchmark classification but with a smaller sample size, which matched that associated with the training set based on the mixed spectral responses. In this way it would be possible to compare directly a conventional approach to training with that based on mixed responses in which the sample sizes used were the same. The data set derived is referred to as the conventional (same size) training set.

The final training set derived was based on the spectral response of the 5 fields used to generate the mixed spectral response data. The aim was to use the ground data on the 5 selected fields in a conventional manner. For this, the boundary region was masked out and only the pure pixels that were separated from the boundary by at least 2 pixels were acquired for training purposes. In this way, a standard set of pure pixels extracted from the same fields used to generate the mixed spectral response data set were acquired for training. Thus, the same total amount of ground data, in terms of fields of known class membership, was available to the analysis as for the classification based on the mixed spectral responses. This training set comprised 225, 203 and 100 pixels of winter wheat, spring barley and sugar beet, respectively. This is referred to as the conventional (same fields) training set.

The C and γ parameters for each SVM classification were selected with the aid of a five-fold cross-validation analysis undertaken. The values used in each classification are presented in Table 1.

The accuracy of each classification was assessed using a testing set comprising 90 randomly selected pure pixels of each class. Classification accuracy was expressed as the percentage of correctly allocated cases but the confusion matrices are also presented below to allow other measures of accuracy to be derived if desired. The classification accuracy statements were compared to determine if the classifications differed significantly in terms of accuracy. Since the same testing set was used to evaluate all of the classifications, this comparison was based on a McNemar test (Foody, 2004). With this approach, the evaluation of the significance of a difference in estimated accuracy is based on the magnitude of a computed z score. The difference between two accuracy statements is viewed as being statistically significant at the 95% level of confidence if $z > |1.96|$.

4. Results and discussion

The classification trained with 90 cases of each class in accordance with the widely used 30p heuristic yielded a very accurate classification. The overall accuracy of this classification was 92.59% with, as expected from the crop canopy covers evident at the time of image acquisition, most confusion between the winter wheat and spring barley classes (Table 2). This benchmark classification derived using a conventional approach to training showed that the classes may be classified to a high accuracy.

The classification trained with the mixed spectral responses also yielded a highly accurate classification (Table 3). The accuracy of this classification was 91.11%, marginally but insignificantly less than that derived from the benchmark analysis ($z = 1.15$). The pattern of class allocations in the two classifications was also very similar (Tables 2 and 3). Thus, a classification trained on mixed spectral responses yielded a similar classification that was of comparable accuracy to that derived from the adoption of the conventional approach. Moreover, the classification trained on mixed spectral responses required substantially less ground data, just class membership

Table 2
Confusion matrix from the SVM classification derived with the use of the conventional, 30p, training set

Actual class	Predicted class			
	ww	SpB	sb	Σ
Winter wheat (ww)	84	6	0	90
Spring barley (SpB)	14	76	0	90
Sugar beet (sb)	0	0	90	90
Σ	98	82	90	270

The overall accuracy was 92.59%.

Table 3
Confusion matrix from the SVM classification derived with the use of the mixed spectral response training set

Actual class	Predicted class			
	ww	SpB	sb	Σ
Winter wheat (ww)	83	7	0	90
Spring barley (SpB)	17	73	0	90
Sugar beet (sb)	0	0	90	90
Σ	100	80	90	270

The overall accuracy was 91.11%.

information on 5 fields whereas the conventional approach needed information on 90 pixels drawn from numerous fields across the test site.

Direct comparison between the two approaches to training is difficult as the size of the training sets used differed. However, a classification trained using the same number of training cases for each class as associated with the analysis based on mixed spectral responses yielded a similar classification (Table 4) to the other analyses (Tables 2 and 3). The differences in accuracy between these classifications was insignificant ($z=0.23$ and $z=1.13$ for the comparison of the classifications derived with the conventional (same size) training set against that from the use of the mixed spectral responses and the 30p training sets respectively). However, the use of just the pure pixels from the 5 fields that provided the mixed spectral response data resulted in a classification of markedly, and significantly at 95% level of confidence, lower accuracy than the others (Table 5). This was evident in comparison against the classification trained with mixed spectral responses ($z=3.40$) and the training set formed through the use of the 30p heuristic ($z=3.53$). Thus, although the approach based on mixed spectral responses required only class membership for a few fields and resulted in an accurate classification the use of the pure pixels from these fields yielded a classification of substantially lower accuracy.

The results demonstrate that it is possible to use a small training set based on mixed spectral responses to derive a classification of similar accuracy to one trained using a larger number of pure pixels acquired in a conventional manner. These results indicate considerable promise for adopting a very different approach to the training of supervised classifications than that widely promoted. The approach also offers the potential for substantial savings in training data acquisition as the number of training cases and the precision with which they must be defined spatially is less than with the conventional approach. This is most apparent perhaps in relation to the effort required to acquire the training sample. With the conventional approach adopted it is necessary to strictly adhere to the constraints imposed by a random sampling design while the approach based on mixed spectral responses is compatible with an easier to implement judgemental sampling approach. The conventional approach may, for example, require ground data to be collected for sites distributed widely across the entire study area in locations that might be difficult to access and which may need to be located with high precision. With the alternative approach attention is focused on boundary regions that typically are easy to observe, in both the field and imagery, as well as selected on the basis of analyst's judgement which may include issues such as convenience and practicality. Thus judgemental sampling, which is often viewed unfavorably and avoided but which is easy to implement, may have a role to play in training image classifications.

Table 4
Confusion matrix from the SVM classification derived with the use of the conventional, same size, training set

Actual class	Predicted class			
	ww	SpB	sb	Σ
Winter wheat (ww)	79	11	0	90
Spring barley (SpB)	12	78	0	90
Sugar beet (sb)	0	0	90	90
Σ	91	89	90	270

The overall accuracy was 91.48%.

Table 5
Confusion matrix from the SVM classification derived with the use of the conventional, same fields, training set

Actual class	Predicted class			
	ww	SpB	sb	Σ
Winter wheat (ww)	83	1	6	90
Spring barley (SpB)	33	57	0	90
Sugar beet (sb)	0	0	90	90
Σ	116	58	96	270

The overall accuracy was 85.18%.

Clearly, the work reported here represents only a first step and there are many issues to be addressed. The example used was, for instance, of a very simple classification problem to illustrate the potential of the approach. More complex situations exist, particularly with greater spectral overlap in the class distributions. Since the classes here were relatively highly separable the precise location of the hyperplane to classify the data was not critical. It would be interesting to evaluate in detail the location of the hyperplane used to derive the classification relative to the optimal separating hyperplane derived from a large sample in the normal way. Ideally, the hyperplane fitted using the mixed spectral response data would lie close to or correspond exactly with the real optimal separating hyperplane to discriminate the classes. Other issues also require further evaluation. For example, the orientation of the boundary with respect to the image's raster grid will impact on the class composition of the mixed pixels. If, for example, the boundary is parallel with the rows or columns of the grid the mixed pixels will have the same class composition. If, however, the boundary was to cut diagonally through the pixel grid then the composition of the mixed pixels would vary. The implications of this in terms of the class composition contributing to the mixed responses and its impact on the location of the separating hyperplane in particular should be evaluated. Additionally, for example, the approach presented will be impacted by variations in boundary properties and pixel size. If, for example, the boundary between the fields was a wide hedge, perhaps wider than the dimensions of a pixel, the boundary region focused on in this article would not convey much information about the mixing of the crop spectral responses. In such circumstances the approach presented here would result in the boundary itself being characterised and so not necessarily help in locating an appropriate hyperplane. Finally, as the training stage of the classification should be tailored to the nature of the classifier to be used, the applicability of the general approach to other classifiers should be evaluated. Further work is, therefore, required to understand the potentials of the approach and raise awareness of its limitations. Current work is, for example, focused on the potential of using just mixed pixels to convey the information required for a classification.

5. Summary and conclusion

The accuracy of a supervised image classification is a function of the training data set used. Many studies have shown that variables such as the size, composition and nature of the training cases used can have a substantial impact on image classification accuracy. However, most guidance on training set design urges the use of a large number of pure pixels. Emphasis is also placed on describing the classes rather than on the provision of information to separate them. It has been argued here that the basis of the conventional approach to the design of the training stage of a supervised classification is not always appropriate and an alternative suggested.

In remote sensing, the aim of the training stage has typically been seen as being the production of descriptive statistics for each class which may then be used in the determination of class membership by the selected classifier. The typical approach used involves the acquisition of a sample of pixels of known class membership from the image. The ruling paradigm for training a classifier seems based implicitly on a desire to provide an accurate characterization of the classes. Emphasis in such an approach is on the use of a large number of pure pixels selected at random. Such an approach may not always be appropriate and the training stage should be designed with regard to the classifier selected for the analysis. Here, we have proposed an approach for the design of the training stage that is essentially the opposite, the use of a small number of purposefully selected mixed spectral responses to provide information to discriminate the classes with a SVM classifier. The approach, in effect, has focused on the boundary between classes in geographical space for the provision of the information needed to separate the classes in spectral feature space. These geographical locations are, in the conventional practice, often regarded as the most undesirable samples and often deliberately masked-out of analyses.

The new approach to the design of the training stage was grounded on the basis of the proposed classifier, an SVM. It or a derivative of it may be applicable to other classifiers as the aim is simply to use the training stage to provide information that helps separate the classes rather than describe them. The approach proposed may, however, be inappropriate for use with some classifiers. For example, the approach proposed would not be expected to yield training data that gave a representative description of the classes and so be inappropriate for many basic statistical classifiers. It is important, therefore, that the training data acquisition programme is informed of the type of classifier to be used. The way that a SVM operates provides considerable potential to refine the design of training data collection activities. In the example shown above, a small amount of ground data (the class membership for 5 fields) was required to accurately classify three classes with a SVM. For example, the use of a training set comprising mixed spectral responses yielded a classification with an accuracy of 91.11%, nearly identical to that that derived from the use of a conventionally defined training set containing the same number of cases which had an accuracy of 91.48%. The new approach is, therefore, an attractive alternative, but not necessarily a replacement, of the conventional approach, especially in circumstances when training data collection is costly.

Although the classification was very simplistic the work reported represents the first steps in trying to design a new approach to classifier training that is tailored to the way in which the specific classifier operates which also has operational advantages (principally the potential to use a small and inexpensive sample to form the training set). A central focus is to acquire training data to separate the classes and not to describe them as well as to free the training data collection programme from the constraints often imposed by the need to adhere to the demands of random sampling by allowing judgemental sampling.

Acknowledgments

The authors are grateful to the NERC Earth Observation Data Centre for the provision of the SPOT HRV data and the European commission for the crop map made available through the AgriSAR campaign. We are also grateful to the British Council for providing a Commonwealth Scholarship to AM to undertake postgraduate research at the University of Southampton while on leave from the Punjab Remote Sensing Centre. The support vector machines were constructed with BSVM (version 2.01) software developed by Chih-Wei Hsu and Chih-Jen Lin, National Taiwan University, Taipei. We are also grateful to Dr. Jadunandan Dash for assistance with the research. Finally, we are grateful to the three referees for their helpful comments on the original submission.

References

Airkan, M. (2004). Parcel-based crop mapping through multi-temporal masking classification of Landsat 7 images in Karacabey, Turkey. *Proceedings of the ISPRS symposium, Istanbul International Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, *Vol. 34*. (pp.).

Aria, K. (1992). A supervised Thematic Mapper classification with a purification of training samples. *International Journal of Remote Sensing*, *13*, 2039−2049.

Belousov, A. I., Verzakov, S. A., & von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: support vector machines. *Chemometrics and Intelligent Laboratory Systems*, *64*, 15−25.

Brown, M., Lewis, H. G., & Gunn, S. R. (2000). Linear spectral mixture models and support vector machines for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 2346−2360.

Buttner, G., Hajos, T., & Korandi, M. (1989). Improvements to the effectiveness of supervised training procedures. *International Journal of Remote Sensing*, *10*, 1005−1013.

Chi, M., & Bruzzone, L. (2005). A semilabeled-sample-driven bagging technique for ill-posed classification problems. *IEEE Geoscience and Remote Sensing Letters*, *2*, 69−73.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273−297.

Ediriwickrema, J., & Khorram, S. (1997). Hierarchical maximum-likelihood classification for improved accuracies. *IEEE Transactions on Geoscience and Remote Sensing*, *35*, 810−816.

Fardanesh, M. T., & Ersoy, O. K. (1998). Classification accuracy improvement of neural network classifiers by using unlabeled data. *IEEE Transactions on Geoscience and Remote Sensing*, *36*, 1020−1025.

Foody, G. M. (1999). The significance of border training patterns in classification by a feedforward neural network using backpropagation learning. *International Journal of Remote Sensing*, *20*, 3549−3562.

Foody, G. M. (2004). Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, *70*, 627−633.

Foody, G. M., & Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, *18*, 799−810.

Foody, G. M., & Mathur, A. (2004a). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*, 1335−1343.

Foody, G. M., & Mathur, A. (2004b). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment*, *93*, 107−117.

Foody, G. M., McCulloch, M. B., & Yates, W. B. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, *16*, 1707−1723.

Hsu, C. -W., & Lin, C. -J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*, 415−425.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*, 725−749.

Kuo, B. C., & Landgrebe, D. A. (2002). A covariance estimator for small sample size classification problems and its application to feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, *40*, 814−819.

Mather, P. M. (2004). *Computer Processing of Remotely-Sensed Images*, 3rd edition. Chichester: Wiley.

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*, 1778−1790.

Nelson, R. F., Latty, R. S., & Mott, G. (1984). Classifying northern forests using Thematic Mapper simulator data. *Photogrammetric Engineering and Remote Sensing*, *50*, 607−617.

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*, 1007−1011.

Piper, J. (1992). Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, *13*, 685−692.

Song, C., Woodcock, C. E., Seto, K. C., Pax Lenney, M., & Macomber, S. A. (2001). Classification and Change Detection Using Landsat TM Data. When and how to correct for atmospheric effects? *Remote Sensing of Environment*, *75*, 230−244.

Staufer, P., & Fischer, M. M. (1997). Spectral pattern recognition by a two-layer perceptron: effects of training set size. In I. Kanellopoulos, G. G. Wilkinson, F. Roli, & J. Austin (Eds.), *Neurocomoputation in Remote Sensing Data Analysis* (pp. 105−116). Berlin: Springer.

Sun, W. X., Heidt, V., Gong, P., & Xu, G. (2003). Information fusion for rural land-use classification with high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *41*, 883−890.

Tadjudin, S., & Landgrebe, D. A. (2000). Robust parameter estimation for mixture model. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 439−445.

Todd, W. J., Ustin, S., & Haman, J. F. (1980). Landsat wildland mapping accuracy. *Photogrammetric Engineering and Remote Sensing*, *46*, 509−520.

Tsai, F., & Philpot, W. D. (2002). A derivative-aided hyperspectral image analysis system for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, *40*, 416−425.

Van Niel, T. G., McVicar, T. R., & Datt, B. (2005). On the relationship between training sample size and data dimensionality of broadband multi-temporal classification. *Remote Sensing of Environment*, *98*, 468−480.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.

Wang, J. G., Neskovic, P., & Cooper, L. N. (2005). Training data selection for support vector machines. *Lecture Notes in Computer Science*, *3610*, 554−564.

Zhuang, X., Engel, B. A., Lozano-Garcia, D. F., Fernandez, R. N., & Johannsen, C. J. (1994). Optimisation of training data required for neuro-classification. *International Journal of Remote Sensing*, *15*, 3271−3277.