Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Land Cover Classification of Cloud-Contaminated Multi-Temporal High-Resolution Images

Note no SAMBA/45/09
Author Arnt-Børre Salberg
Date December 7, 2009

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in the areas of information and communication technology and applied statistical modeling. The clients are a broad range of industrial, commercial and public service organizations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have for us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Land Cover Classification of Cloud-Contaminated Multi-Temporal High-Resolution Images**
Author	**Arnt-Børre Salberg**
Date	December 7, 2009
Publication number	SAMBA/45/09

Abstract

We show how methods proposed in the statistical community dealing with missing data may be applied for land cover classification, were optical observations are missing due to clouds and snow. The proposed method may be divided into two stages; (i) cloud/snow-classification and (ii) training and classification.

The purpose of the cloud/snow classification stage is to determine which pixels are missing due to clouds and snow. All pixels in each optical image are classified into the classes cloud, snow, water and vegetation using a suitable classifier. The pixels classified as cloud or snow are labeled as missing, and this information is used in the training and classification stages which deal with the land cover classification. In the training stage the unknown parameters of the statistical distribution modeling the land cover classes are estimated using the Expectation Maximization algorithm. If a non-parametric classifier is applied, the training is omitted. In the classification stage the pixels are classified into various land cover classes. Here we apply the Maximum Likelihood (assuming normal distributions), k-NN and Parzen classifiers, all modified to handle missing features.

The classifiers are evaluated on Landsat (both TM and ETM+) images covering a scene at about 900 m a.s.l. in the Hardangervidda mountain-plateau in South Norway, where 4 869 in situ samples of the land cover classes water, ridge, leeside, snow-bed, mire, forest and rock are obtained. The results show that proper modeling of the missing pixels improves the classification rate by 5-10%, and by using multiple images we increase the chance of observing the land cover type substantially. The nonparametric classifiers applied ignores the missing-data mechanism, and are therefore particularly suitable for remote sensing applications where the pixels covered by snow and cloud may depend on the land cover type.

Keywords	Missing data, land cover classification, cloud, snow
Target group	Remote sensing
Availability	Open
Project	GB-JO
Project number	220412-WP5
Research field	Remote sensing
Number of pages	23

1 Introduction

Land cover classification based on a single satellite image can be challenging due to bad weather conditions and limited spatial and spectral resolutions. To increase the classification performance, the use of multi-temporal satellite images has been applied (see e.g. Agrawal et al., 2003; Aurdal et al., 2005; Jing et al., 2009). By using multi-temporal satellite images for land cover classification, the ground vegetation may be observed at different phenological states. The vegetation species are therefore more easily distinguished since we observe how the spectral signatures vary through the growth season.

Often, information from a set of multi-temporal images is utilized after pixel level fusion of the images. In pixel-level fusion the images are merged by stacking the multi-temporal pixels into a vector of measurements (Solberg, 2007). The resulting feature vector will then be of higher dimension, and often possess stronger discrimination power since some classes may not be separable in a single image (Jing et al., 2009).

The use of high-resolution images in a multi-temporal classification context has been limited, mainly because high resolution images have longer revisit time than medium or low resolution images (Solberg, 2007). In Northern Europe, clouds and snow may prevent us from observing all pixels in optical remote sensing, and we are missing parts of the observed data. Moreover, the number of training points are often a small subset of the whole image, and due to the missingness of the data we may experience that we have only a few complete training vectors available. This may cause rank deficient covariance matrices which makes the task of designing classifiers more challenging.

When faced with data that contain missing observations, ad hoc solutions such as case deletion or imputation are often applied to convert the data into a complete-data format (Dixon, 1979). However, using such methods often introduces biased parameter estimates and distorted covariance matrices (Schafer, 1997). Statistical analysis with missing or incomplete data has been well documented (Little and Rubin, 1987; Schafer, 1997). In particular, parameter estimation with missing data has been studied extensively, and the Expectation Maximization (EM) algorithm being one of the most popular methods for estimating unknown parameters of a probability density function (PDF) (Little and Rubin, 1987).

Pattern recognition with missing data has also received some attention (see e.g. DePasquale and Polikar, 2007; Dixon, 1979; Marlin, 2008; Mojirsheibani and Montazeri, 2007a,b; Morris et al., 1998; Pelckmans et al., 2005; Twala et al., 2008). Mojirsheibani and Montazeri (2007b) proposed representations for the best (Bayes) classifier when some of the covariates can be missing without imposing any assumptions on the underlying missing data mechanism. Furthermore, the proposed classifiers (both parametric and non-parametric) are not based on any data imputation techniques. Marlin (2008) focused on problems of collaborative prediction with non-random missing data and classification with missing features. Several procedures for classification with missing features using generative classifiers, the combination of standard discriminative classifiers with single and multiple imputation, classification in subspace, and an approach by modifying the classifier input representation to include response indicators.

In remote sensing applications, both statistical and non-statistical methods to handle missing or incomplete data have received little attention. Aksoy et al. (2009) described how decision-tree classifiers can be trained with alternative decision nodes for handling missing data in multi-source information fusion, and showed the superiority of the classifier for land cover classification for aerial, Ikonos and DEM-based data sets.

Digital elevation models (DEMs) have often been included as ancillary data in landscape classification tasks. Since a DEM has 100% coverage of the scene, it may play an important role of

supporting training and classification when the optical observations are missing due to clouds or snow. By establishing a model describing the interaction between the DEM-based features and the observed optical features, we may utilize this knowledge when classifying data with missing covariates.

In this paper we show and discuss how to handle missing observations of remote sensing data for pixel classification of multi-temporal high-resolution satellite images, and in particular we idenity the underlying missing data mechanism related snow and clouds (Sec. 3.1).

We propose a two-stage classifier for classifying cloud and snow contaminated pixel-level fused multi-temporal images, where the first stage consists of identifying pixels containing snow and clouds, and the second stage performs the land cover classification with proper modeling of the missing observations (Sec. 3.3 - 3.5).

Recovering the original scene from cloud contaminated satellite scenes has been studied extensively (Holben, 1986; Melgani, 2006), and Melgani (2006) proposed a method to reconstruct areas of missing land cover observations from multi-temporal multi-spectral images using a contextual prediction system. The proposed scheme also facilities estimation of the missing feature for Gaussian and mixture Gaussian data assumptions, and may be used to reconstruct areas covered by clouds and snow (Sec. 3.6).

2 Experimental data set

2.1 Remote sensing and DEM data

The optical data applied in this paper is Landsat ETM+ images (path/row 199/18) covering Hardangervidda mountain plateau in South Norway (about 900m a.s.l) from late May to mid September (2004-05-31, 2000-07-23, 2002-08-14, and 2002-09-15). The four Landsat images of the scene are shown in Fig. 1, and the color images are created from band 3, 4 and 5. Blue pixels correspond to ice/snow, white pixels correspond to clouds, and green/brown pixels correspond to vegetation/soil. The images contain six spectral layers each, these were the standard Landsat spectra except the thermal IR bands (bands 6-1 and 6-2) and the panchromatic band (band 8). The images were radiometrically and geometrically corrected using the standard terrain correction (Level 1T) which incorporates ground control and a DEM to obtain topographic accuracy.

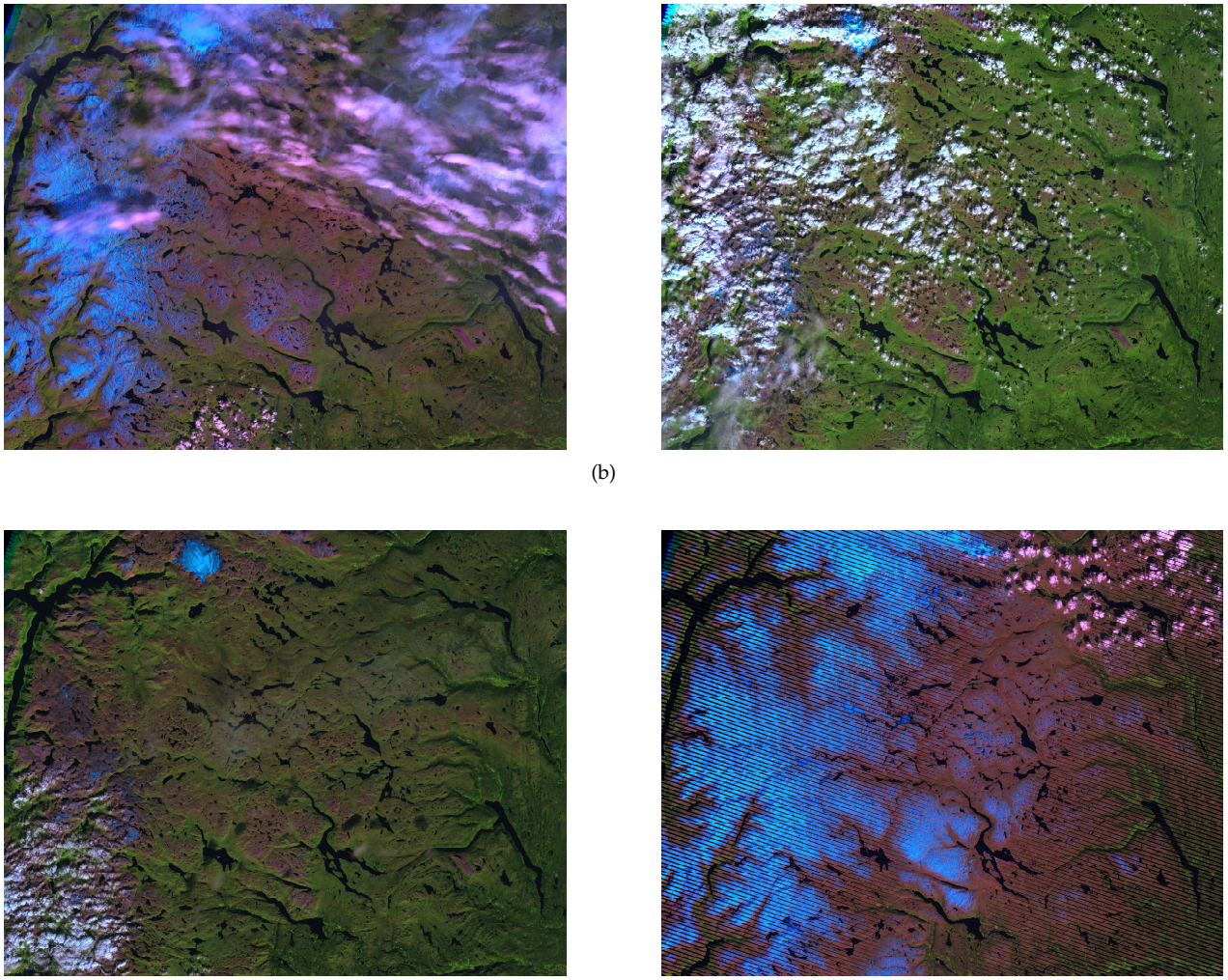
The DEM has a spatial resolution of $25 \times 25\text{m}^2$. In order to match the Landsat pixel resolution of $30 \times 30\text{m}^2$, the DEM has been resampled using cubic interpolation. From the DEM the terrain slope was calculated, and elevation and slope were used as ancillary data (Fig. 2).

2.2 In situ data

A total of 4861 pixels were labeled according to the ground truth vegetation type obtained from field measurements. The vegetation and landscape features were divided into the following classes: *water, ridge, leeside, snowbed, mire, forest* and *rock* (see Tab. 1). Please note from the table that the portion of acquired sample points do not correspond to the priori land cover probabilities of Hardangervidda. For testing of the classifiers, the sample points were divided into two sets (equal size and randomly selected), one for training and one for testing. The proportion of observed (non-missing) sample points is as low as 8.7% for snowbed vegetation in the May 31, 2004 image (see Tab. 1), and full coverage for ridge, snowbed, mire and water for the Sep. 15, 2002 image. Estimation of the priori probabilities for the land cover classes is based on the work by Gaare et al. (2005) with necessary adjustments to differences in class definitions.

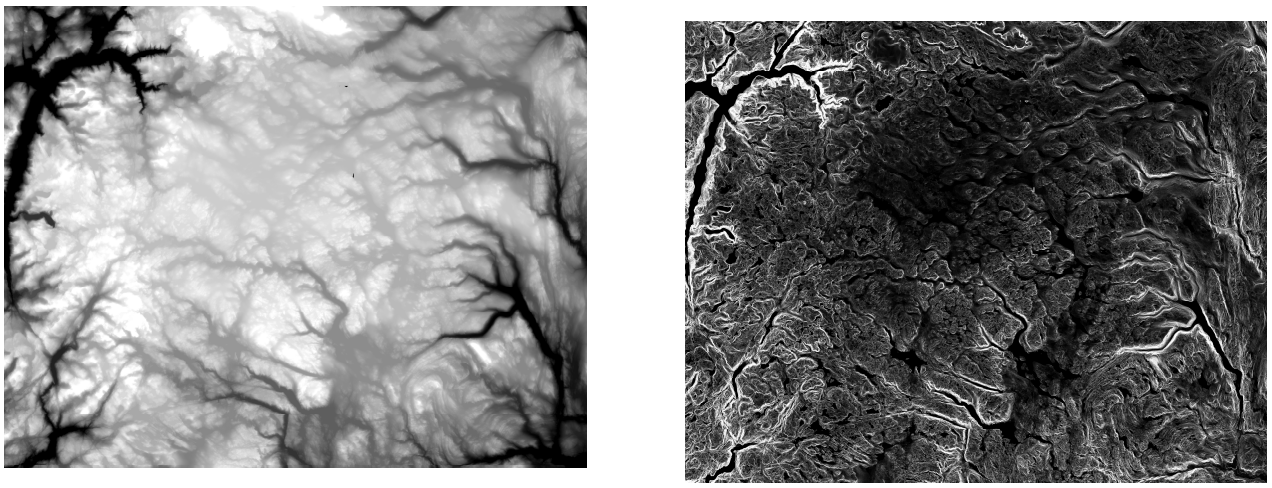
Class	Number of sample points	Observed sample points 2000-07-23	2002-08-14	2002-09-15	2004-05-31	Priori prob.
Water	1 395	71%	68.1%	99.9%	51.2%	9.3%
Ridge	1 345	47%	84%	100%	57.4%	27%
Leeside	360	75%	87.2%	98%	69.2%	21.9%
Snowbed	275	49.4%	88%	99.9%	8.7%	18.3%
Mire	554	67.3%	73.4%	100%	53.8%	9.2%
Forest	443	87.6%	78.1%	98%	86%	4.8%
Rock	489	58.3%	59%	92%	19.6%	9.7%
Total	4 861	63.3%	75.9%	98.8%	52.1%	100%

Table 1. Number of sample points in each land cover class of the Landsat images (labeled by acquisition date), the proportion of cloud- and snow-free samples points in each class, and the corresponding priori probabilities.



(a) (b) (c) (d)

Figure 1. Landsat ETM+ images acquired at time instants (a) 23 Jul. 2000, (b) 14 Aug. 2002, (c) 15 Sep. 2002, (d) 31 May 2004



(a) (b)

Figure 2. Acillary data used. (a) Elevation and (b) slope

Image 1												
Image 2												
Image 3												
Pixel number	1	2	3	4	5	6	7	8	9	10	11	12

Figure 3. Illustration of a typical missing data pattern for 12 different pixels. The feature vector is constructed by stacking the features of the 3 images. Gray coloring corresponds to a missing observation.

3 Methods for land cover classification with missing data

3.1 Modeling missing data in remote sensing

Given a set of geo-referenced high-resolution multi-spectral images, we construct a feature vector for a given pixel by stacking all spectral data into a single vector $\mathbf{x}$. Let the value for a given feature at a given pixel be modeled as

$$x = rv + (1 - r)w, \tag{1}$$

where $r = 1$ if the pixel value corresponds to a land cover observation (not snow) v, and $r = 0$ if the pixel correspond to a cloud or snow observation w. Here we are interested in the land cover observations v, and if v is not observed (i.e. $r = 0$) we consider v as missing. Hence, r is a response indicator of the missing-data mechanism. A typical missing data scenario is shown in Fig. 3, where we show a missing data pattern for 12 different pixels. Here we have constructed the feature vectors by stacking 3 images each with 3 features. Gray coloring corresponds to missing observations.

A graphical illustration of the causal dependencies for the missing data mechanisms is shown in Fig. 4, where the data distribution of the land cover features $\mathbf{X}$ is modeled as $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu})$, $\boldsymbol{\mu}$ is a parameter vector and $\mathbf{Z}$ denotes the latent variables (if any). Now, let $\mathbf{R}$ denote the response indicators for all features and all pixels. When modeling missing data, the missing mechanism

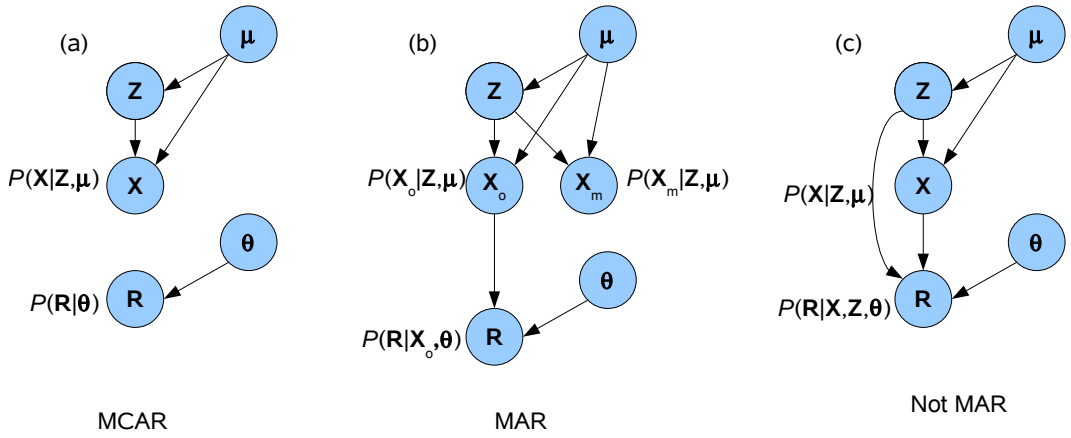


Figure 4. Graphical illustration of the causal dependencies for MCAR (a), MAR (b) and not MAR (c).

that creates $\mathbf{R}$ plays a crucial role for designing algorithms for estimating unknown parameters of the distribution of the land cover features and for designing the classifier. In particular, it is important to quantify the interaction between $\mathbf{R}$ and the observed and missing features of $\mathbf{X}$ and any latent variables. The first category of missing data is called "missing completely at random" (MCAR). Data is MCAR when the response indicator variables $\mathbf{R}$ are independent of the data variables $\mathbf{X}$ and any latent variables $\mathbf{Z}$. Hence, we may express the distribution for $\mathbf{R}$ as $P(\mathbf{R}|\mathbf{X},\mathbf{Z},\boldsymbol{\theta}) = P(\mathbf{R}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector (Fig. 4(a)). Another category of missing data is "missing at random" (MAR). The MAR condition may be expressed as $P(\mathbf{R} = \mathbf{r}|\mathbf{X} = \mathbf{x},\mathbf{Z} = \mathbf{z},\boldsymbol{\theta}) = P(\mathbf{R} = \mathbf{r}|\mathbf{X}_o = \mathbf{x}_o|\boldsymbol{\theta})$, where $\mathbf{X}_o$ denotes the observed features. Hence, the missing data only depend on the observed features of $\mathbf{X}$, not the missing features (Fig. 4(b)). Many algorithms for estimating unknown parameter assumes that the missing data is MAR. In order to perform maximum likelihood estimation of $\boldsymbol{\mu}$ or to perform likelihood ratio tests concerning $\boldsymbol{\mu}$ without regard for the missing data mechanism, we also need to assume distinctness of the parameters in addition to MAR (Schafer, 1997). I.e., we need to assume that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are distinct. If both MAR and distinctness hold, the missing data mechanism is said to be *ignorable* (Little and Rubin, 1987; Schafer, 1997).

In optical remote sensing the missing observations (at least due to snow) may indeed correlate with land cover class. However, within a land cover class we may assume that the missing data mechanism of $\mathbf{R}$ is independent of the optical features of $\mathbf{X}$. When using ancillary data based on digital elevation models, these features are always observable and will therefore never be missing. Thus, even if the missing data mechanism of the optical features is dependent of the DEM features within a vegetation class, the missing data mechanism may be considered as MAR.

Another missing data source is sensor failure. On May 31, 2003 the Scan Line Corrector (SLC) in the Landsat ETM+ instrument failed, resulting that approximately one-fourth of the data in a Landsat 7 image is missing when acquired without a functional SLC. Pixels missing due to SLC failure will be regarded as MCAR since the missing mechanism is not related to the observed, missing or any other variable.

Even for cases where the missing data are not precisely MAR, a general procedure ignoring the missing data mechanism still tend to be better than ad hoc procedures such as case deletion and zero-insertion since by ignorable missing data procedures remove all of the nonresponse bias explained by the observed data, whereas the ad hoc procedures may not (Schafer, 1997).

3.2 Learning with missing data

Parameter estimation with missing data is a well explored topic in statistics, and one of the most common algorithms applied to cope with missing data is the Expectation-Maximization (EM) algorithm (Little and Rubin, 1987). We will here briefly review the EM algorithm.

When we have incomplete data, the maximum likelihood (ML) estimates of the parameters of any probability density function (PDF) may be computed using the EM algorithm. The EM algorithm ignores the mechanism causing the missing data, since it assumes that the probability that a value is missing does not depend on the missing data value itself.

3.2.1 Gaussian distributions

For the case of missing values in Gaussian distributed data, estimation of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ has been studied extensively (Little and Rubin, 1987; Schneider, 2001), and we will briefly review the algorithm

At the tth iteration of the EM algorithm let $\widehat{\boldsymbol{\mu}}(t)$ and $\widehat{\boldsymbol{\Sigma}}(t)$ denote the current estimate of the mean vector and covariance matrix, respectively. Further, let $\mathcal{X}_o$ the observed data. The E step of the algorithm consists of calculating the sufficient statistics (Little and Rubin, 1987; Schneider, 2001)

$$\mathrm{E}\left\{\mathbf{X}_i|\mathcal{X}_o,\widehat{\boldsymbol{\mu}}(t),\widehat{\boldsymbol{\Sigma}}(t)\right\}=\sum_{i=1}^{N}\widehat{\mathbf{x}}_i(t) \tag{2}$$

$$\mathrm{E}\left\{\mathbf{X}_i\mathbf{X}_i^T|\mathcal{X}_o,\widehat{\boldsymbol{\mu}}(t),\widehat{\boldsymbol{\Sigma}}(t)\right\}=\sum_{i=1}^{N}\left(\widehat{\mathbf{x}}_i(t)[\widehat{\mathbf{x}}_i(t)]^T+\widehat{\mathbf{C}}_i(t)\right) \tag{3}$$

where $\widehat{\mathbf{x}}_i(t)$ is an estimate of the feature vector $\mathbf{X}_i$ at iteration t and $\widehat{\mathbf{C}}_i(t)$ is defined as (Little and Rubin, 1987)

$$\widehat{\mathbf{C}}_i(t)=\begin{cases}\widehat{\boldsymbol{\Sigma}}_{m(i)}(t)-\widehat{\boldsymbol{\Sigma}}_{m(i)o(i)}(t)\widehat{\boldsymbol{\Sigma}}_{o(i)}^{-1}(t)\widehat{\boldsymbol{\Sigma}}_{o(i)m(i)}(t), & \text{for elements corresponding to the missing part of } \mathbf{X}_i\\ \mathbf{0} & \text{for elements corresponding to the observed part of } \mathbf{X}_i\end{cases} \tag{4}$$

Let $\widehat{\mathbf{x}}_{o,i}(t)$ and $\widehat{\mathbf{x}}_{m,i}(t)$ denote the observed and missing part of $\widehat{\mathbf{x}}_i(t)$. Then we have that

$$\widehat{\mathbf{x}}_{o,i}(t)=\mathbf{X}_{o,i}\quad\text{and} \tag{5}$$

$$\widehat{\mathbf{x}}_{m,i}(t)=\widehat{\boldsymbol{\mu}}_{m(i)}(t)+\widehat{\boldsymbol{\Sigma}}_{m(i)o(i)}(t)\widehat{\boldsymbol{\Sigma}}_{o(i)}^{-1}(t)\left[\mathbf{x}_{o,i}(t)-\widehat{\boldsymbol{\mu}}_{o(i)}(t)\right]. \tag{6}$$

with missing data where $\widehat{\boldsymbol{\mu}}_{o(i)}(t)$ and $\widehat{\boldsymbol{\mu}}_{m(i)}(t)$ denotes the part of $\widehat{\boldsymbol{\mu}}_i$ corresponding to the observed and missing part of $\mathbf{X}_i$, respectively. Thus, the missing values are replaced by the estimated conditional mean values of $\mathbf{X}_{m,i}$ given the set observed values $\mathbf{X}_{o,i}$ of $\mathbf{X}_i$.

The M step of the EM algorithm is straightforward. Updated estimates of the mean vector and covariance matrix are computed from the estimated complete data sufficient statistics (Little and Rubin, 1987; Schneider, 2001)

$$\widehat{\boldsymbol{\mu}}(t+1)=\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i(t)\quad\text{and} \tag{7}$$

$$\widehat{\boldsymbol{\Sigma}}(t+1)=\frac{1}{N}\sum_{i=1}^{N}\left[(\widehat{\mathbf{x}}_i(t)-\widehat{\boldsymbol{\mu}}(t+1))\left(\widehat{\mathbf{x}}_i(t)-\widehat{\boldsymbol{\mu}}^T(t+1)\right)^T+\widehat{\mathbf{C}}_i(t)\right]. \tag{8}$$

Please note that Eq. (6) provides an estimate for the missing feature of the ith pixel, given the observed features and the estimated mean and covariance matrix. This estimate may be used for image restoration of areas covered by clouds or snow.

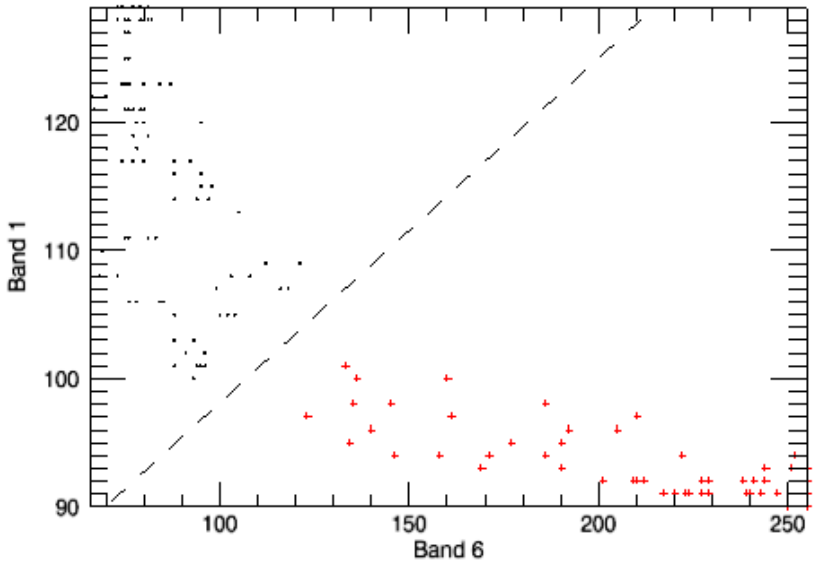


Figure 5. Illustration of a typical scatter plot of pixel observations (digital number) of Band 1 and Band 6. Crosses and dots correspond to cloud and land cover observations, respectively. The line indicates the decision boundary between the vegetation and cloud class.

3.2.2 Gaussian mixture models

A possible disadvantage of using a Gaussian distribution to model the data, is the lack of flexibility when compared to non-parametric methods. However, by using mixture models this drawback may be circumvented since finite mixture models are capable of modeling a wide range of densities. The density of the Gaussian finite mixture model is defined as

$$f(\mathbf{x}) = \sum_{i=1}^{G} p_i \phi(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad \sum_{i=1}^{G} p_i = 1, \quad p_i \geq 0, \forall i, \tag{9}$$

where $\phi(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the density of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Gaussian mixture models may also be trained from incomplete data using the EM algorithm for both estimation of the mixture components and for coping with missing data (Ghahramani and Jordan, 1994; Lin et al., 2006).

3.3 Cloud and snow classification

In order to determine if clouds or snow occupy a given pixel, we need to perform cloud/snow detection of the pixels in the acquired images. To illustrate the cloud/snow pixel detection principle we consider a scatter plot of pixel observations of Band 1 and Band 6 (Fig. 5). In the figure

the crosses and dots correspond to and cloud and land cover pixel observations, respectively. By choosing a decision boundary according to the indicated line, we may perfectly separate the cloud pixels from the land cover pixels, and the missing data mechanism may be considered as MAR. Note that this is conceptually different from censoring of the data (which is not MAR), since the cloud/snow detection does not change the distribution of the land cover observations. However, if the cloud/snow pixels are not easily separated (i.e. the are erroneous classified pixels) from the land cover pixels the MAR assumption is not satisfied.

3.4 Classification with missing data

In the Bayesian approach, the optimal classifier is obtained by determining the classifier that minimizes the expected loss or Bayes risk. Using a zero-one loss function the optimal Bayes classifier corresponds to the minimum error classifier (Duda et al., 2001).

Let $\mathbf{x}^{(\ell)}$ be the part of $\mathbf{x}$ corresponding to the missing data indicator vector $\mathbf{r}_k$, $f(\mathbf{x}^{(k)}|\omega_i)$ denote the PDF of the subvector $\mathbf{x}^{(\ell)}$ given class ω_i, and let $\boldsymbol{\rho}$ denote a binary vector with 0 at element j if the the jth element of the feature vector $\mathbf{x}$ is missing, and 1 otherwise. The probability of selecting class ω_i given $\mathbf{x}^{(\ell)}$ is then

$$P(\omega_i|\mathbf{x}^{(\ell)}) = \frac{P(\omega_i)f(\mathbf{x}^{(\ell)}|\omega_i)}{\sum_{j=1}^{C} f(\mathbf{x}^{(\ell)}|\omega_j)P(\omega_j)}. \tag{10}$$

where $P(\omega_j)$ is the priori distribution of class ω_i. Now, let $P(\boldsymbol{\rho} = \mathbf{r}_\ell|\mathbf{x}^{(\ell)}, \omega_i)$ denote the conditional probability that the missing data pattern is equal to $\mathbf{r}_\ell$, given $\mathbf{x}^{(\ell)}$ and class ω_i. The optimal classifier may then be formulated as (Mojirsheibani and Montazeri, 2007b)

$$\mathcal{G}_m(\mathbf{x}, \boldsymbol{\rho}) = \arg\max_{\omega_i} \sum_{\ell=1}^{L} I(\boldsymbol{\rho} = \mathbf{r}_\ell) P(\omega_i|\mathbf{x}^{(\ell)}, \boldsymbol{\rho} = \mathbf{r}_\ell) \tag{11}$$

$$= \arg\max_{\omega_i} \sum_{\ell=1}^{L} I(\boldsymbol{\rho} = \mathbf{r}_\ell) P\{\boldsymbol{\rho} = \mathbf{r}_\ell|\mathbf{x}^{(\ell)}, \omega_i\} f(\mathbf{x}^{(\ell)}|\omega_i) P(\omega_i) \tag{12}$$

$$\tag{13}$$

where L is the number of different indicator vectors $\mathbf{r}_\ell$. For parametric classifier the PDF of $\mathbf{x}^{(\ell)}$ for class ω_i is modelled as $f(\mathbf{x}^{(\ell)}|\omega_i, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a parameter vector.

Note that the missing data mechanism introduces an additional probability $P\{\boldsymbol{\rho} = \mathbf{r}_\ell|\mathbf{x}^{(\ell)}, \omega_i\}$ which is conditioned on the vector $\mathbf{x}^{(\ell)}$ and class ω_i. Further, by the inclusion of $P\{\boldsymbol{\rho} = \mathbf{r}_\ell|\mathbf{x}^{(\ell)}, \omega_i\}$ the classifier takes into account that the missing data mechanism may be different for different land cover classes. In general, since the missing pattern depends on all elements of $\mathbf{x}^{(\ell)}$ (and not only the observed ones), the missing mechanism is not MAR. However, as we argued in Sec. 3.1, within a given class the missing mechanism only depend on the observed values (for DEM based ancillary data). If the $P\{\boldsymbol{\rho} = \mathbf{r}_k|\mathbf{x}^{(\ell)}, \omega_i\} = P\{\boldsymbol{\rho} = \mathbf{r}_\ell\}$ the missing mechanism is MCAR, and the optimal classifier is simply the marginal distribution where the missing features have been integrated out.

3.4.1 Nonparametric classification

Nonparametric kernel classification rules derived from incomplete (missing) data were discussed in (Mojirsheibani and Montazeri, 2007b; Pawlak, 1993). Assume that we have the following training data set available

$$\mathcal{D} = \{(\mathbf{X}_1, \boldsymbol{\rho}_1), (\mathbf{X}_2, \boldsymbol{\rho}_2), \ldots, (\mathbf{X}_N, \boldsymbol{\rho}_N)\}, \tag{14}$$

where $\boldsymbol{\rho}_i$ is an indicator vector of the observations in $\mathbf{X}_i$.

k-NN classifier The k-NN classifier relies on the distances between the observed feature vector and the training vectors. Pattern recognition using k-NN in a missing data environment has been considered by (Dixon, 1979; Mojirsheibani and Montazeri, 2007b; Morin and Reaside, 1981). Mojirsheibani and Montazeri (2007b) proposed a nearest-neighbour approach that does not depend on any MAR assumptions. Let $\mathcal{I}_m$ be the index set corresponding to the elements in $\mathcal{D}$ with class label ω_m, and define

$$\phi^{\text{k-NN}}_{\mathcal{I}_m,\ell}(\mathbf{x}^{(\ell)}) = \sum_{i\in\mathcal{I}_m} \mathcal{W}_{\mathcal{I}_m,i}(\mathbf{x}^{(\ell)}, \mathbf{r}_\ell) \tag{15}$$

where

$$\mathcal{W}_{\mathcal{I}_m,i}(\mathbf{x}^{(\ell)}, \mathbf{r}_\ell) = \begin{cases} 1, & \text{if } \mathbf{X}_i^{(\ell)} \text{ is one of the } k \text{ NNs of } \mathbf{x}^{(\ell)} \text{ among all} \\ & \quad \text{those } \mathbf{X}_j^{(\ell)}\text{s, } j \in \mathcal{I}_m \text{ for which } \boldsymbol{\rho}_j = \mathbf{r}_\ell \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Then the k-NN discriminant function for the mth class may be formulated as

$$\mathcal{G}_m(\mathbf{x}, \boldsymbol{\rho}) = \sum_{\ell=1}^{L} I(\boldsymbol{\rho}, \mathbf{r}_\ell)\phi^{\text{k-NN}}_{\mathcal{I}_m,\ell}(\mathbf{x}^{(\ell)}), \tag{17}$$

where L denotes the number of possible indicator vectors $\mathbf{r}_k$. As a classification rule we choose the class corresponding to the maximum discriminant function. The k-NN classifier works on the selection of samples among the training data that has the exact same missing data pattern as the test vector, and perform the k-NN rule among these samples. The discriminant function in Eq. (17) applies an estimate of the class prior probability equal to $\widehat{P}(\omega_m, \boldsymbol{\rho}) = |\mathcal{I}_{m,\boldsymbol{\rho}}|/N_{\boldsymbol{\rho}}$ (van der Heijden et al., 2004), where $\mathcal{I}_{m,\boldsymbol{\rho}}$ denotes the set of training vectors of class m where $\boldsymbol{\rho}_i$ is equal to $\boldsymbol{\rho}$, and $N_{\boldsymbol{\rho}}$ is the number of training vector with missing data pattern $\boldsymbol{\rho}$. Scaling the rule in Eq. (17) with $P(\omega_i)/|\mathcal{I}_{m,\boldsymbol{\rho}}|$ in order to incorporate the priori class probability does not make sense for k-NN with small values of k. For the nearest neighbor rule, methods based on a weighted-distance measure (Brown and Koplowitz, 1979) is preferred to include priori weighting.

Now, if the number of features missing is high and the number of training data vectors is low, we may lack sample points where $\boldsymbol{\rho}_j = \mathbf{r}_\ell$, and the feature vector $\mathbf{x}$ cannot be classified. To handle such cases we propose an alternative indicator function;

$$\widetilde{I}(\boldsymbol{\rho}, \mathbf{r}_\ell) = \begin{cases} 1, & \text{if } \mathbf{1}^T(\boldsymbol{\rho} \odot \mathbf{r}_\ell - \boldsymbol{\rho}) = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

where $\odot$ denotes elementwise multiplication. This indicator function ignores the values of $\mathbf{r}_\ell$ at elements where $\boldsymbol{\rho}$ is equal to zero and considers all $\mathbf{r}_\ell$ that has 1s on corresponding locations as $\boldsymbol{\rho}$. However, the classifier is now biased with respect to the missing data mechanism.

Parzen classifier The Parzen classifier applies the Parzen density estimate as a means for classification (Fukunaga, 1990). Given a set of independent and identically distributed samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from the true density f, the Parzen window estimator for this distribution is defined as (Fukunaga, 1990)

$$\widehat{p}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{X}_i) \tag{19}$$

Here, W_{σ^2} is the Parzen window, or kernel, and σ^2 controls the width of the kernel. The Parzen window must integrate to one, and is typically chosen to be a PDF itself, such as the Gaussian kernel. Hence,

$$W_{\sigma^2}(\mathbf{x}, \mathbf{X}_i) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x} - \mathbf{X}_i\|^2}{2\sigma^2}\right). \tag{20}$$

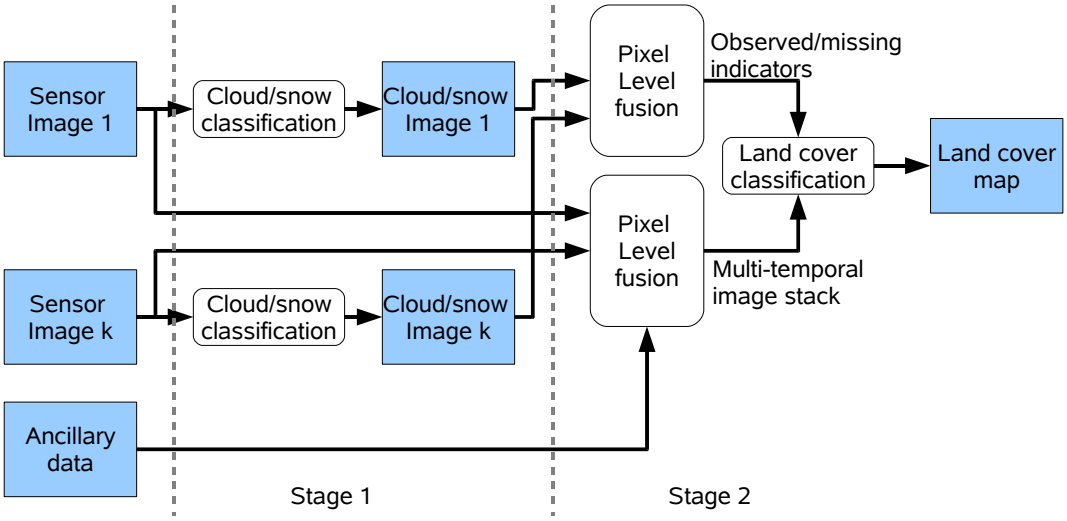


Figure 6. Two-stage classifier for pixel classification of cloud/snow contaminated multi-temporal images.

Note also that the width of the Parzen window affects the density estimate much more than the actual form of the window function (Scott, 1992; Wand and Jones, 1995).

Mojirsheibani and Montazeri (2007b) proposed the following Parzen window estimator for missing data (without any MAR assumption);

$$\widehat{p}(\mathbf{x},\boldsymbol{\rho}|\omega_m) = \sum_{\ell=1}^{L} \frac{I(\boldsymbol{\rho},\mathbf{r}_\ell)}{(2\pi\sigma^2)^{|\mathcal{P}_\ell|/2} N_{m,\ell}} \left[\sum_{i\in\mathcal{I}_m} I(\boldsymbol{\rho}_i,\mathbf{r}_\ell) \exp\left(-\frac{\|\mathbf{x}^{(\ell)} - \mathbf{X}_i^{(\ell)}\|^2}{2\sigma^2} \right) \right] \tag{21}$$

where $|\mathcal{P}_\ell|$ is the number of elements in $\mathbf{r}_\ell$ equal to one, and $N_{m,\ell}$ is the number of elements in $\mathcal{I}_m$ where $I(\boldsymbol{\rho},\mathbf{p}_\ell)$ is equal to 1.

The discriminant function for the Parzen classifier may be written as

$$\mathcal{G}_m(\mathbf{x},\boldsymbol{\rho}) = P(\omega_m)\widehat{p}(\mathbf{x},\boldsymbol{\rho}|\omega_m), \tag{22}$$

and we apply Silverman's mean integrated squared error method (Silverman, 1986) to estimate the smoothing parameter σ^2.

Also here, we may encounter data shortness if we have too few training points, and for such cases we therefore propose to consider the indicator function $\widetilde{I}(\boldsymbol{\rho}_i,\mathbf{r}_\ell)$.

3.5 Two-stage classifier

To classify the cloud contaminated multi-temporal high-resolution images we propose a two-stage classification approach (Fig. 6). In the first stage, the cloud/snow contaminated pixels of each input image are identified by the pre-classification procedure (Sec. 3.3). Since cloud/snow pixels have totally different spectral signatures from land cover vegetation and rocks the performance of the pre-classifier is expected to be high. In Stage 2, the sensor images, the ancillary data (if any) and the corresponding cloud/snow maps are pixel-level fused by vector stacking the features of each pixel. The pixel-level fused image and the corresponding observed/missing map are then sent to the classifier for land cover pixel classification. The classifier may be any classifier capable of dealing with missing data (for instance, any of the classifier described in Sec. 3.4). The output of the classifier is then the land cover class map.

3.6 Image reconstruction of cloud/snow contaminated images

The knowledge of the class-dependent mean vectors and covariance matrices of the optical features may be used to estimate the missing observations by using the observed features of a given pixel. To do so, first we perform a land cover classification of the scene as described in the previous section to determine the land cover class of each pixel in the scene. Then, assuming that the data is Gaussian, we may apply the minimum mean-squared error estimator (Scharf, 1991)

$$\widehat{\mathbf{x}}_m = \widehat{\boldsymbol{\mu}}_m^c + \widehat{\boldsymbol{\Sigma}}_{mo}^c [\widehat{\boldsymbol{\Sigma}}_o^c]^{-1} (\mathbf{x}_o - \widehat{\boldsymbol{\mu}}_o^c) \tag{23}$$

to estimate the missing features of the pixel-level fused multi-temporal images. Here c denotes the land cover class obtained from the land cover classification. If we use a Gaussian mixture model, an estimate of the missing data may be obtained as

$$\widehat{\mathbf{x}}_m = \sum_{i=1}^{G} \widehat{w}_i \left[\widehat{\boldsymbol{\mu}}_{m,i}^c + \widehat{\boldsymbol{\Sigma}}_{mo,i}^c [\widehat{\boldsymbol{\Sigma}}_{o,i}^c]^{-1} (\mathbf{x}_o - \widehat{\boldsymbol{\mu}}_{o,i}^c) \right] \tag{24}$$

where $\widehat{w}_i$ is the posterior probability of $\mathbf{x}$ belonging to the ith sub-class, $i = 1, \ldots, G$ (Lin et al., 2006).

From the equations above we note that by using the knowledge about the detected land cover class for given pixel, and how the features correlates within the corresponding class, we may estimate the missing parts of the feature vector. As a final step, the estimated features are assigned to its corresponding image, and a reconstructed image is obtained.

4 Experimental results

4.1 Land cover classification of Hardangervidda

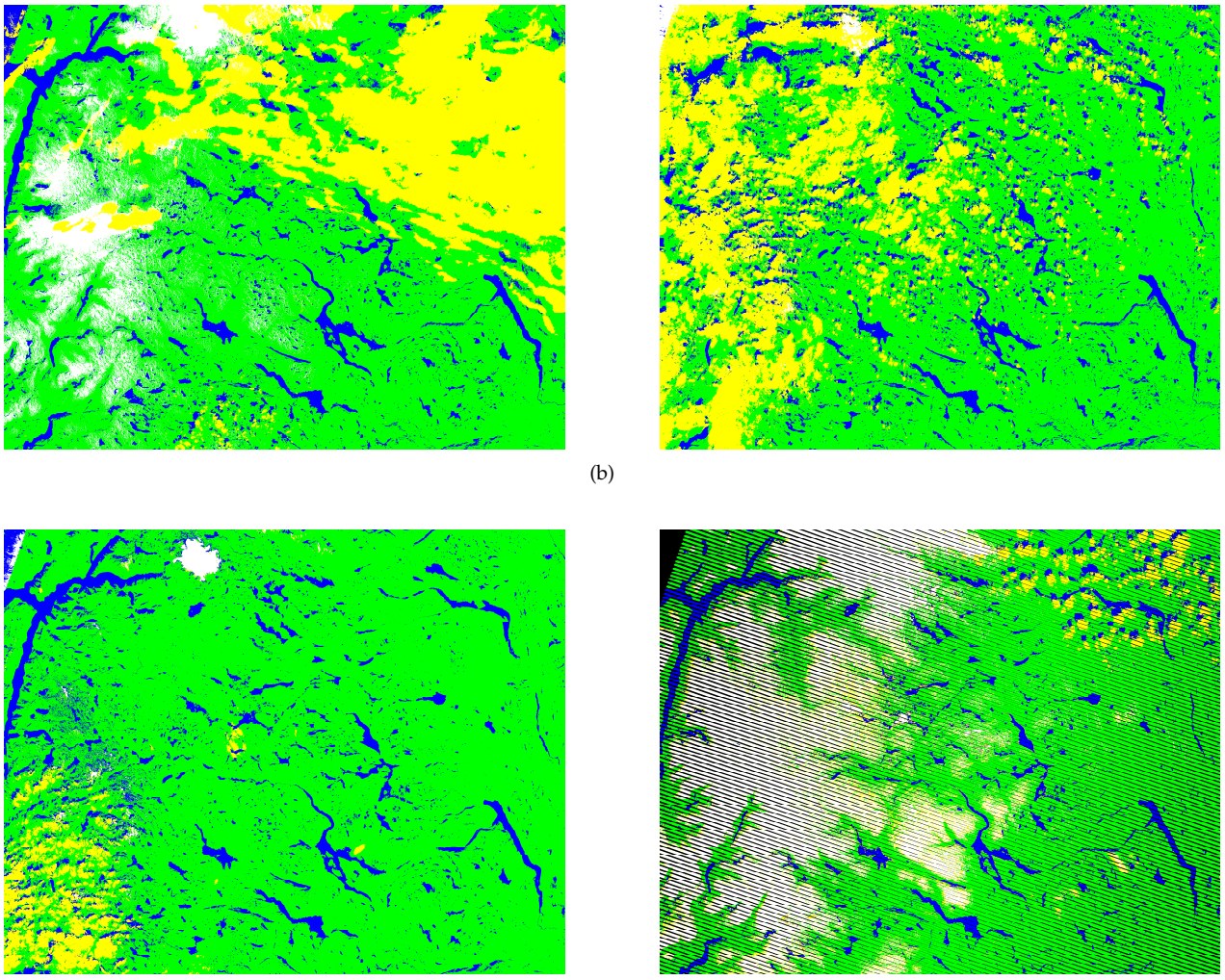
To test the missing feature classification method in a remote sensing context we applied the two-stage classifier to a sequence of three Landsat ETM+ images (2000-07-23, 2002-08-14 and 2002-09-15). The feature vector was constructed by stacking the six multi-spectral bands into an 18-element feature vector. In some experiments elevation and slope were also applied as features, and the feature vector was in those cases augmented to a 20-element vector.

4.1.1 Stage 1 - Cloud/snow classification

Before we classify the land cover we perform cloud/snow classification of the three Landsat images. Since cloud and snow were easily observable in the high-resolution multi-spectral image, we first performed a visual inspection of the image to determine if there were any clouds or snow present. If, so training data corresponding to cloud, snow, water and vegetation observations were manually labeled, and the image was classified to the labels cloud, snow, water and vegetation. The classifier was the built-in SVM classifier in Environment for Visualizing Images (ENVI), with Gaussian kernel function. Fig. 7(a-c) show the results of the cloud/snow classification where the Landsat images are classified into the four classes *snow/ice* (white), *clouds* (yellow), *water* (blue), and *vegetation* (green). From these image, pixels that were classified to clouds or snow/ice were labeled as *missing*, and pixels classified as water or vegetation were labeled as *observed*. The proportions of pixels classified as non-missing (vegetation and water) for 2000-07-23, 2002-08-14, and 2002-09-15 images were 63.3%, 75.9%, and 98.8%, respectively.

4.1.2 Stage 2 - Land cover classification

To classify the land cover of the Hardangervidda scene into the classes *water, ridge, leeside, snowbed, mire, forest* and *rock*, we first constructed the response indicator matrix $\mathbf{R}$ from the cloud/snow



(a) (b)

(c) (d)

Figure 7. Cloud/snow classification of the four Landsat ETM+ images acquired at time instants (a) July 23, 2000, (b) August 14, 2002, (c) September 15, 2002, and (d) May 31, 2004. The color corresponds to *snow/ice* (white), *clouds* (yellow), *water* (blue), *vegetation* (green), and sensor failure (black stripes).

classification (see Sec. 4.1.1). In order to evaluate the classifiers we randomly divided the data sets into two halves, one for training and the other for testing, 100 times. Since the prior probabilities of the land cover classes and the acquired portion of sample points differed substantially (see Tab. 1), prior class probabilities were not applied, and we assumed that each class had an equal prior probability (expect for the k-NN-rule where prior probabilities was not included). For the Gaussian classifier, the EM-algorithm was stopped after 50 iterations. For the Parzen and kNN classifier, the data were rescaled to zero mean and unit variance in all feature directions. The mean values and standard deviations needed to perform the scaling were estimated from the training data only. Euclidean distance measures was used for the k-NN classifier. The smoothing parameter provided by Silverman's method and used in the Parzen classifier was too high, and adjusted to $\sigma/8$ to obtain higher classification performance.

For all methods the classification performance improved when using pixel-level image fusion (Tab. 2, column "Accuracy excl. missing data"). For the Gaussian and k-NN classifiers, the performance increased from 62.8% and 63.4% to 74.9% and 75.2%, respectively. As expected the portion of pixels that were classified also increased, which resulted in an even higher classification performance when considering all pixels (column "Accuracy incl. missing data"). For all classifiers we observed that the performance improved when including the DEM-based ancillary data (elevation and slope). The best classifier was the 1-NN classifier based on three Landsat scenes (2000-07-23, 2002-08-14 and 2002-09-15) and elevation and slope (Tab. 2) with a classification accuracy of 80.9%. From the confusion matrix of this classifier we see that the classifier tended to mix snowbed and ridge vegetation (Tab. 3). Comparing the confusion matrix of the 1-NN classifier (Tab. 3) with the one of the Gaussian classifier (Tab. 4) we see that the Gaussian classifier more often classified rock as water and ridge as mire. However, it classified snowbed and mire more correctly than the 1-NN classifier.

The land cover map, the output of the two-stage classifier, shows that we were able to classify all pixels in the scene using the images 2000-07-23, 2002-08-14, and 2002-09-15 and slope even with this severe amount of cloud-contaminated pixels (Fig. 8). Note that we have not included elevation as a feature for creating the thematic map of the land cover since we did not have training samples supporting the range of variation of the elevation in the scene.

4.2 Image reconstruction of cloud/snow contaminated images

Using the proposed image reconstruction algorithm we pixel-level fused the Landsat images (2000-07-23, 2002-08-14 and 2002-09-15) and the slope image into an image of 19 features. The missing features were then estimated and stored in its corresponding Landsat image. From the subset of the original 2002-08-14 image we successfully reconstructed the vegetation cover by clouds (Fig. 9). However, note that the cloud shadows still remain in the reconstructed image.

To evaluate the image reconstruction algorithm further we pixel-level fused the 2004-05-31 ETM+ SLC-OFF image (Fig. 1(d)) and its corresponding cloud-snow map (Fig. 7(d)) with the 2000-07-23 and 2002-09-15 ETM+scenes, and terrain slope. Using the image reconstruction algorithm we were able to reconstruct and remove the stripes of the ETM+ SLC-OFF image (Fig. 10). From the figure we observe that the stripes in the reconstructed image are barely visible, and the algorithm also seemed to capture the optical features of the hidden land cover. Please also note that this scene is of a completely different phenological state than the other scenes.

Method	ETM+ 2000-07-23	ETM+ 2002-08-14	ETM+ 2002-09-15	ETM+ (SLC-OFF) 2004-05-31	DEM	Accuracy excl. missing data	Accuracy incl. missing data	Portion classified
Gaussian	X					69.0 ± 1.00	52.0 ± 0.73	63.2
		X				62.8 ± 0.96	57.0 ± 0.74	76.0
			X			66.5 ± 0.62	65.8 ± 0.63	98.9
				X		74.5 ± 1.16	52.9 ± 0.72	52.1
	X	X	X			74.9 ± 0.78	74.3 ± 0.80	99.1
	X	X	X		X	77.9 ± 0.74	77.9 ± 0.74	100
k-NN (k=1)	X					67.9 ± 0.97	51.2 ± 0.72	63.2
		X				63.4 ± 0.85	57.4 ± 0.70	76.0
			X			66.8 ± 0.73	66.1 ± 0.74	98.9
				X		76.2 ± 1.0	53.8 ± 0.62	52.1
	X	X	X			77.4 ± 0.64	76.7 ± 0.66	99.1
	X	X	X		X	80.9 ± 0.59	80.9 ± 0.59	100
k-NN (k=3)	X	X	X		X	79.4 ± 0.62	79.4 ± 0.62	100
k-NN (k=5)	X	X	X		X	78.0 ± 0.66	78.0 ± 0.66	100
Parzen (smooth param=0.079)	X	X	X			76.8 ± 0.71	76.1 ± 0.72	99.1
Parzen (smooth param=0.082)	X	X	X		X	80.3 ± 0.57	80.3 ± 0.57	100

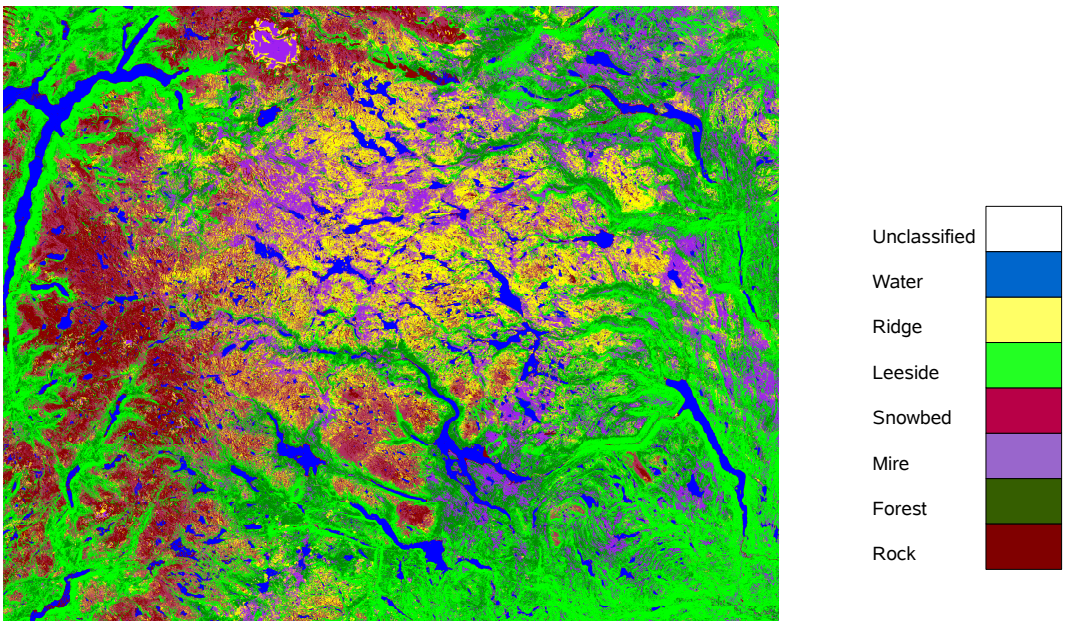
Table 2. Summary of classification results.

	Water	Ridge	Leeside	Snowbed	Mire	Forest	Rock
Water	97	0	0	0	0	0	8
Ridge	0	83	22	38	21	2	3
Leeside	0	5	45	6	5	14	0
Snowbed	0	5	5	47	2	0	2
Mire	0	6	8	6	70	2	1
Forest	0	1	19	1	1	81	0
Rock	3	1	1	3	0	0	85

Table 3. Confusion matrix for the 1-NN classifier using the Landsat ETM+ images 2000-07-23, 2002-08-14, and 2002-09-15 and elevation and slope.

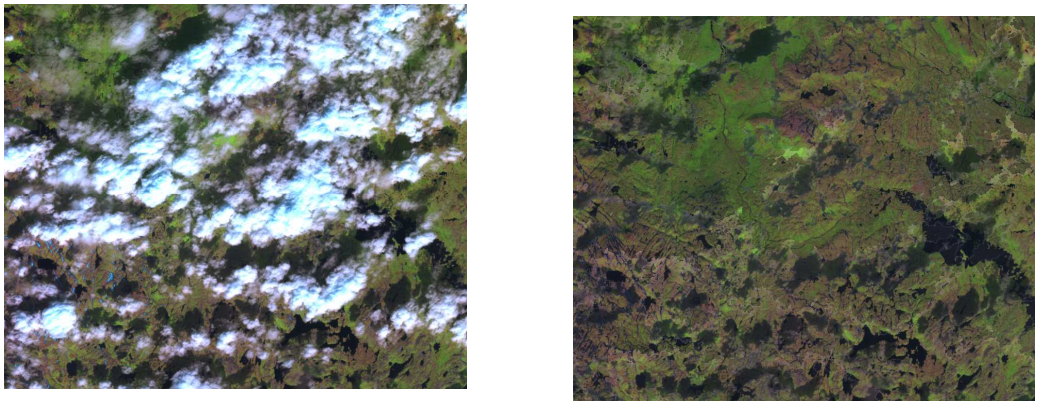
	Water	Ridge	Leeside	Snowbed	Mire	Forest	Rock
Water	96	0	0	0	0	0	16
Ridge	0	68	15	22	9	1	1
Leeside	0	6	45	9	5	12	1
Snowbed	0	9	7	61	3	0	6
Mire	0	15	7	5	81	1	0
Forest	0	1	26	1	2	86	0
Rock	4	1	0	2	0	0	76

Table 4. Confusion matrix for the Gaussian classifier using the Landsat ETM+ images 2000-07-23, 2002-08-14, 2002-09-15 and elevation and slope.



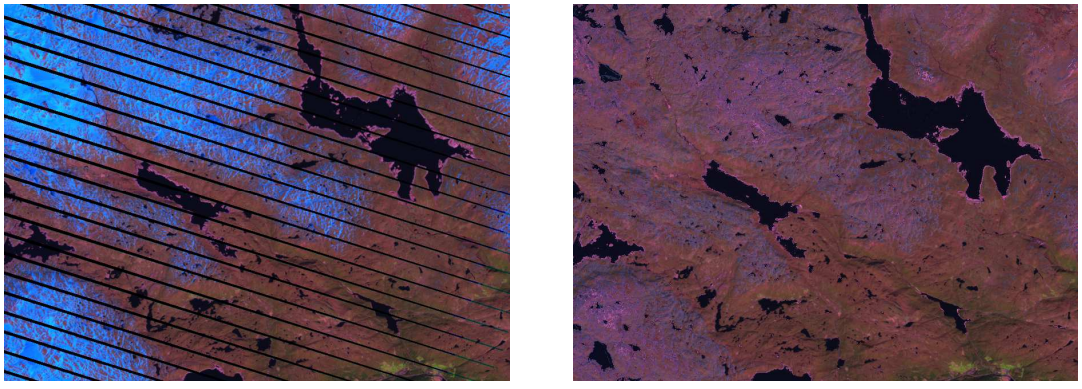
(a) (b)

Figure 8. The land cover map obtained from the Landsat ETM+ images 2000-07-23, 2002-08-14, and 2002-09-15 and slope using a Gaussian classifier.



(a) (b)

Figure 9. (a) Sub-section of the 2002-08-14 image. (b) Reconstructed version of the 2002-08-14 sub-section.



(a) (b)

Figure 10. (a) Sub-section of the 2004-05-31 image. (b) Reconstructed version of the 2004-05-31 sub-section.

5 Discussion and conclusions

In this paper, we performed land cover classification of multi-temporal optical remote sensing images using statistical methods to handle missing data due to clouds and snow. The proposed method is a two-stage approach, where we first classify the pixels that contain clouds or snow, and label these pixels as missing. Then, using the obtained missing data indicators the pixels are classified to the land cover classes of interest using learning and classification algorithms suitable for handling missing observations. The results showed that by processing the missing observations properly we may obtain increased classification power by pixel-level fusion of cloud and snow contaminated satellite images.

If too many images are applied in pixel-level fusion we risk running into the curse of dimensionality when the training data set is limited (Duda et al., 2001), and nearest neighbor classifiers suffer from bias in high dimensions (Hastie and Tibshirani, 1996). Further improvement of the classification performance may be obtained by careful selection of features using feature selection and/or feature extraction (van der Heijden et al., 2004), by considering methods such as the discriminant adaptive nearest neighbor (DANN) classification (Hastie and Tibshirani, 1996), or by using bootstrap methods such as Bagging (Duda et al., 2001). Note that it is also possible to apply an automatic procedure without labeled training data to classify snow and clouds pixels in the Landsat scenes (Hollingsworth et al., 1996; Irish, 2000).

When using terrain elevation as ancillary data we should not classify pixels at altitudes different than what is spanned by the training data. If so the elevation feature would act as an outlier, and may affect the classification results substantially. Other ancillary data than the elevation and slope may be applied. The topographical wetness index based on the upslope contributing area (Tarboton, 1997) is one particular interesting choice. Topographical correction of the intensity values using e.g. c-correction may also be applied to increase the performance, particularly for north-facing steep slopes which are often in shadow. Furthermore, we may also consider features extracted from synthetic aperture radar (SAR) images. Full polarimetric SAR images have successfully been applied for vegetation classification (e.g. Doulgeris et al., 2008) and may therefore possess a valuable contribution to both land cover classification and image reconstruction of cloud-contaminated high-resolution images.

Even if the clouds are 100% correctly detected in the cloud/snow detection stage, cloud shadows are expected to degrade the classification performance. Due to the low brightness of the shadows, the shadowed pixels are often classified as water. However, by including proper ancillary data, such as slope, the classification performance for pixels in the shadow increases since the slope of water bodies is zero.

The optimal discriminant function given in Eq. (13) contains the probability $P\{\boldsymbol{\rho} = \mathbf{r}_\ell | \mathbf{x}^{(\ell)}, \omega_i\}$, which is an important difference from discriminant functions for the non-missing data case (Duda et al., 2001), but also other classifiers proposed for missing data (Duda et al., 2001; Morris et al., 1998). Even for the MAR missing data mechanisms this probability is in general conditioned on the class ω_i, i.e. the optimal classifier takes into account that the missing data mechanism may depend on the class ω_i. For remote sensing this is definitely an important issue, since the degree of snow covering a given area depends on the land cover class or vegetation. For ancillary data, such as elevation and slope, this is definitely the case (snow cover tends to depend on the elevation and slope). However, as pointed out in Sec. 3.1, elevation and slope are always observable and therefore not dependent on the missing data mechanism. Obtaining an estimate for $P\{\boldsymbol{\rho} = \mathbf{r}_k | \mathbf{x}^{(k)}, \omega_i\}$ is not easy (Mojirsheibani and Montazeri, 2007b) since the probability changes through the season (e.g. nearly all data will be missing during the winter months).

Censoring of intensity values due to camera saturation also causes the "true observations" to

be missing. However, in this case the missing data mechanism is not MAR (Little and Rubin, 1987). Fortunately, the non-parametric classification methods given in Sec. 3.4.1 do not require any MAR assumptions in order to be valid. For the case of parameter estimation, the censored missing data requires careful modeling (Little and Rubin, 1987).

We also expect, as shown by (Besag, 1986), that by including spatial contextual modeling the classification performance will improve. Markov random field based models optimized using iterated conditional modes may be performed on the initial classification results for any classifier. We do not have training data for evaluation of such models, and have therefore not applied any contextual information, however, it is expected that the land cover map will be smoother.

Proper modeling of the missing data mechanism is the desirable for land cover classification using cloud- and snow-contaminated multi-temporal images. In particular, nonparametric classifiers are suitable since their design does not depend on the missing data mechanisms, and is therefore suitable for remote sensing applications where the pixels covered by snow and cloud may depend on the land cover type.

Acknowledgements

The author thank Norwegian Institute for Nature Research (NINA) for kindling providing us with the in situ data of land cover classes covering Hardangervidda. The author likes in particular to thank Tobias Falldorf, NINA, for sharing his thoughts and experience with the data set. Many thanks goes also to Magne Aldrin, Rune Soberg, Øivind Due Trier, Ragnar Bang Huseby (all Norwegian Computing Center) for discussions on missing data and proofreading, and , and Sigrid Grepstad (Norwegian University of Science and Technology) for discussion on missing data related to Gaussian mixture distributions.

References

Agrawal, S., Joshi, P. K., Shukla, Y., and Roy, P. S. (2003). SPOT VEGETATION multi temporal data for classifying vegetation in south central asia. *Current Science*, 84(11):1440–1448.

Aksoy, S., Koperski, K., and Tusk, C. (2009). Land cover classification with multi-sensor fusion of partly missing data. *Photogrammetric Engineering & Remote Sensing*, 75(5):577–593.

Aurdal, L., Huseby, R. B., Eikvil, L., Solberg, R., Vikhamar, D., and Solberg, A. (2005). Use of hidden markov models and phenology for multitemporal satellite image classification: Applications to mountain vegetation classification. In *Proc. Int. Workshop Analysis Multi-Temporal Remote Sensing Images*, pages 220–224, Biloxi, Miss.

Besag, J. (1986). On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B*, 48(3):259–302.

Brown, T. A. and Koplowitz, J. (1979). The weighted nearest neighbor rule for class dependent sample sizes. *IEEE Trans. Inform. Theory*, IT-25(2):617–619.

DePasquale, J. and Polikar, R. (2007). Random feature subset selection for analysis of data with missing features. In *Proc. Int. Joint Conf. Neural Networks*, pages 2378–2383, Orlando, Fl.

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Trans. Syst. Man Cybern.*, SMC-9:617–621.

Doulgeris, A. P., Anfinsen, S. N., and Eltoft, T. (2008). Classificatio with a non-gaussian model for PolSAR data. *IEEE Trans. Geosci. Remote Sens.*, 46(10):2999–3008.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classificaition*. Wiley, New York, 2nd edition.

Fukunaga, K. (1990). *Statistical Pattern Recognition*. Academic Press, Boston, Mass.

Gaare, E., Tømmervik, H., and Hoem, S. A. (2005). Reinens beiter påHardangervidda. Utviklingen fra 1988 til 2004 (in Norwegian)-NINA Rapport 53. Technical report, Norwegian Institute for Nature Research (NINA), Trondheim/Tromsø.

Ghahramani, Z. and Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, San Fransico, Calif.

Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(6).

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *Int. J. Remote Sens.*, 7:1417–1434.

Hollingsworth, B., Chen, L. Q., Reichenibach, S. E., and Irish, R. (1996). Automated cloud cover assessment for landsat tm images. In *Proc. Soc. Photo-Optical Instrumentation Engineers (SPIE)*, volume 2819 of *Conference on Imaging Spectrometry II*, pages 170–179, Denver, Co.

Irish, P. (2000). Landsat 7 automatic cloud cover assessment. In *Proc. Soc. Photo-Optical Instrumentation Engineers (SPIE)*, volume 4049 of *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI*, pages 348–355, Orlando, Fla.

Jing, X., Wang, J., Huang, W., and Liu, L. (2009). *Study on Forest Vegetation Classification Based on Multitemporal Remote Sensing Images*, volume 293, pages 115–123. Boston, Mass.

Lin, T. I., Lee, J. C., and Ho, H. J. (2006). On fast supervised learning for normal mixture models with missing information. *Pattern Recognition*, 39:1177–1187.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Marlin, B. M. (2008). *Missing Data Problems in Machine Learning*. Doctor of philosophy, Graduate Department Computer Science, University of Toronto, Toronto, Canada.

Melgani, F. (2006). Contextual reconstruction of cloud-contaminated multitemporal multispectral images. *IEEE Trans. Geosci. Remote Sens.*, 44(2):442–455.

Mojirsheibani, M. and Montazeri, Z. (2007a). On nonparametric classification with missing covariates. *J. Multivariate Analysis*, 98:1051–1071.

Mojirsheibani, M. and Montazeri, Z. (2007b). Statistical classification with missing covariates. *J. R. Statist. Soc. B*, 69(Part 5):839–857.

Morin, R. L. and Reaside, D. E. (1981). A reappraisal of distance-weighted *k*-nearest neighbor classification for pattern recognition with missing data. *IEEE Trans. Syst., Man. Cybern.*, SMC-11(3):241–243.

Morris, A. C., Cooke, M. P., and Green, P. D. (1998). Some solution to the missing feature problem in data classification, with application to noise robust ASR. In *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, volume 2, pages 737–740, Seatle, WA, USA.

Pawlak, M. (1993). Kernel classification rules from missing data. *IEEE Trans. Inform. Theory*, 39(3):979–988.

Pelckmans, K., De Brabanter, J., Suykens, J. A. K., and De Moor, B. (2005). Handling missing data values in support vector machines classifiers. *Neural Networks*, 18:684–692.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapmann & Hall, London.

Scharf, L. L. (1991). *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley, Reading, Mass.

Schneider, T. (2001). Analysis of incomplete data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, 14:853–871.

Scott, D. W. (1992). *Multivariate Density Estimation*. John Wiley & Sons, New York.

Silverman, B. W. (1986). *Density esimation for statistics and data analysis*. Chapman and Hall, London.

Solberg, A. H. S. (2007). *Data Fusion for Remote-Sensing Applications*, chapter Signal and Image Processing for Remote Sensing, pages 515–537. Taylor & Francis, Boca Raton, Fla.

Tarboton, D. G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, 33(2):309–319.

Twala, B. E. T. H., Jones, M. C., and Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Lett.*, 29:950–956.

van der Heijden, F., Duin, R. P. W., de Ridder, D., and Tax, D. M. J. (2004). *Classification, Parameter Estimation and State Estimation*. Wiley, Chichester, England.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.