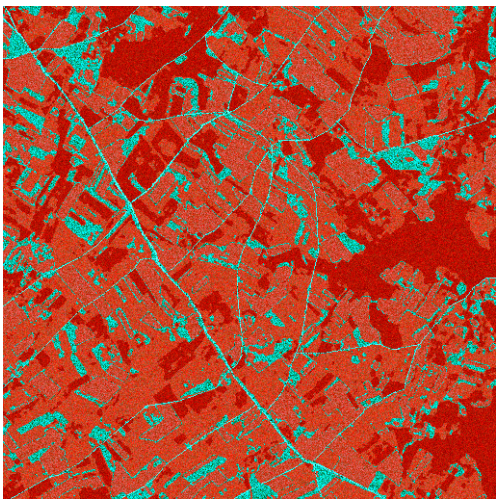
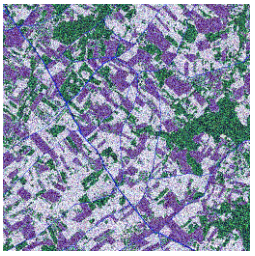
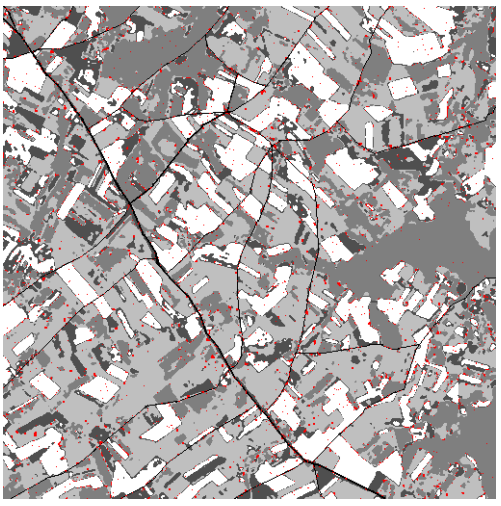
Parameter estimation and classification of multi-scale remote sensing data

Norsk Regnesentral
ANVENDT DATAFORSKNING
Norwegian Computing Center/Applied Research and Development

NOTAT/NOTE







SAMBA/06/2003

Geir Storvik
Roger Fjørtoft
Anne Solberg

May 2003

© Copyright Norsk Regnesentral

Norsk Regnesentral
ANVENDT DATAFORSKNING

NR-notat/NR Note

Tittel/Title:

Parameter estimation and classification of multi-scale remote sensing data

Dato/Date: May
År/Year: 2003
Notat nr/
Note no: SAMBA/06/2003

Forfatter/Author:

Geir Storvik, Roger Fjørtoft, Anne H. S. Solberg

Sammendrag/Abstract:

The rapidly growing number of earth observation satellites provide a much better coverage in space, time and electromagnetic spectrum than in the past. Analysis of compound data sets therefore steadily gains importance. One of the key challenges in multi-sensor image fusion is the combination of images with different resolution to obtain geophysical information of higher quality.

In this report we describe a Bayesian model for integration of multi-scale image data. The approach is based on the concept of a reference resolution. Data at this and lower resolutions are connected to the reference resolution through a fully specified statistical model. Algorithms for parameter estimation and classification based on the multi-scale model are proposed, and results and comparisons with single-scale classification are presented.

Emneord/Keywords: Data fusion, image analysis, multi-scale classification, remote sensing,

Tilgjengelighet/Availability: Open

Prosjekt/Project: EOtools

Prosjektnr./Project no.: 830 110

Satsningsfelt/Research field: Remote sensing

Antall sider/No. of pages: 36

Norsk Regnesentral / Norwegian Computing Center
Gaustadalléen 23, Postboks 114 Blindern, 0314 Oslo, Norway
Telefon 22 85 25 00, telefax 22 69 76 60

© Copyright Norsk Regnesentral

Parameter estimation and classification of multi-scale remote sensing data

Geir Storvik, Roger Fjørtoft and Anne H. S. Solberg
Norwegian Computing Center
Oslo, Norway

May 2003

Contents

1 Introduction

The rapidly growing number of earth observation satellites provide a much better coverage in space, time and electromagnetic spectrum than in the past. Analysis of compound data sets therefore steadily gains importance. One of the key challenges in multi-sensor image fusion is how to combine images with different resolution to obtain more precise results.

For all multi-sensor image analysis tasks, image co-registration is an important prerequisite. For multi-scale data sets it seems necessary to require sub-pixel accuracy with respect to the finest resolution. However, we consider multi-scale image registration to be outside the scope of this study.

In the literature, there are many approaches to multi-sensor image analysis, but few of them treat the multi-scale image analysis properly. The common approach is to resample the different images to be fused to a common pixel resolution, often corresponding to the lowest resolution among the images. In this report, we present a multi-scale model, which allows preserving the details found in the highest resolution image, while exploiting the spectral information at lower resolution.

The remote sensing literature contains many examples of multi-scale data visualization, e.g. by merging a panchromatic image with high resolution with lower-resolution multispectral images (see e.g. Price [1999]). The purpose is visualization, and the multispectral data is overlaid on the panchromatic image using different colors. Wavelet models are common for this kind of applications. However, little work has been done on true multi-scale classification models.

Some work has been done on fusing different data sources when the aim is to estimate an underlying *continuous variable*. Some of these methods are based on transforming all data to the same scale [Núñez et al., 1999]. Daniel and Willsky [1997] considered estimation of hydraulic conductivity for characterizing groundwater flow, fusion of different data sources using multi-resolution methodology. Hirst et al. [2003] proposed estimation of environmental components using data at different resolutions through hierarchical Gaussian models. Both assume data on coarser scales are linear combinations of data at finer scales.

Puyou-Lascassies et al. [1994] and Zhukov et al. [1999] considered unmixing of low-resolution data by using class-information obtained from classification of high-resolution data. The unmixing is performed through several sequential steps, but no formal model for the complete data set is derived. Price [1999] proposed unmixing by relating correlation between low-resolution data and high-resolution data resampled to low-resolution to correlation between high-resolution data and low-resolution data resampled to high resolution. The possibility of mixed pixels was not taken into account.

In Crawford et al. [1999], separate classifications were performed based on data from each resolution. The resulting resolution-dependent probabilities were averaged over the resolutions.

Multi-resolution tree models are sometimes used for multi-scale analysis (see e.g. Luettgen et al. [1994]). Such models yield a multi-scale representation through a quad

tree, in which each pixel at a given resolution is decomposed into four child pixels at higher resolution, which are correlated. This gives a model with very strange correlation structures.

The multi-scale model presented in this document is based on the concept of a *reference resolution* and is developed in a Bayesian framework [Besag, 1989]. We let the reference resolution correspond to the highest resolution present in the data set. For each pixel of the input image at the reference resolution we assume that there is an underlying discrete class. The observed pixel values are modeled conditionally on the classes. The properties of the class label image are described through an *a priori* model. Markov random fields and Potts model have been selected for this purpose. Data at coarser resolutions are modeled as mixed pixels, i.e., the observations are allowed to include contributions from several distinct classes. In this way it is e.g. possible to exploit spectrally richer images at lower resolution to obtain more accurate classification results at the reference level, without smoothing the results as much as if we simply oversampled the low resolution data to the reference resolution prior to the analysis. Another advantage using mixed pixel models is that we simultaneously obtain an "unmixing" of the low-resolution data to the reference resolution, which can also be used for other purposes.

The classification is based on maximization of the *a posteriori* probability (MAP) using the iterative conditional modes (ICM) algorithm. The starting point is a maximum likelihood (ML) classification of the data at the reference level, where a training set is available. Inside the ICM loop, the parameters of Potts model are computed with Newton-Raphson, and the radiometric parameters are estimated through expectation maximization (EM). We have so far used multivariate Gaussian distributions to describe the data given the class.

The results of single- and multi-scale classification results obtained on simulated and real satellite images are presented and discussed.

2 Model

The model consists of two parts, the prior model for the class image (label image) and the likelihood for data which is a model conditional on the class image.

2.1 Model for class image

Define a *reference resolution* for which

- Each pixel contain only one class
- All observations are at this or a coarser resolution

Assume the class image is described by $\mathbf{z} = (z_1, ..., z_n)$ where z_i defines the class in pixel i at the reference resolution. Consider the model

$$\pi(\mathbf{z}) = \frac{1}{C(\boldsymbol{\theta}_1)} \exp\{\sum_i \alpha_{z_i} - \beta \sum_{i \sim i'} \phi_{i,i'}(z_i, z_{i'})\} \quad (1)$$

where $\phi_{i,i'}(z_i, z_j)$ is a potential function modeling spatial interaction, $\boldsymbol{\theta}_1 = (\alpha_1, ..., \alpha_K, \beta)$ and $C(\boldsymbol{\theta}_1)$ is a normalization constant. We have here used the notation $i \sim i'$ to denote that pixels i and i' are neighbors. This is an ordinary Markov Random Field (MRF) model used widely for this purpose [Geman and Geman, 1984, Besag, 1986].

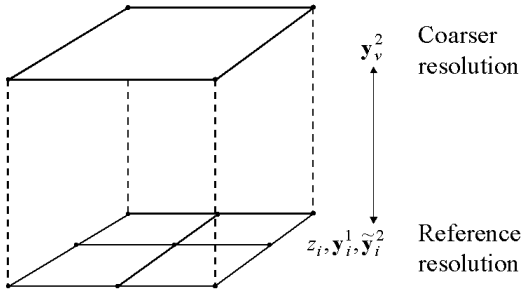


Figure 1: Illustration of the multi-scale model in the case of two resolution levels.

2.2 Model for data

Assume that observations $\mathbf{y} = (\mathbf{y}^1, ..., \mathbf{y}^p)$ are available where $\mathbf{y}^j$ is a (possibly multi-spectral) image at resolution r_j.

We will assume that data at different resolutions are conditionally independent, i.e.

$$p(\mathbf{y}|\mathbf{z}) = \prod_{j=1}^{p} p(\mathbf{y}^j|\mathbf{z}).$$

This is a reasonable assumption for sensors with different spectral properties. A similar assumption is often used for fusion of multi-temporal data, see e.g.Solberg [1999].

For simplicity, we assume that the pixel dimensions at lower resolutions are entire multiples of the pixel dimensions at the reference resolution (level 1), and that the images are perfectly overlapping.[1]

In order to specify $p(\mathbf{y}^j|\mathbf{z})$, we introduce the following notation:

$s_j(v)$ =The set of pixels at the reference resolution which is contained in pixel v at resolution r_j.

m_j =The number of pixels in $s_j(v)$.

$v_j(i)$ =The pixel at resolution r_j containing pixel i at the reference resolution.

The data at coarser resolutions are modeled as mixed pixels, i.e. we introduce hidden variables $\tilde{\mathbf{y}}_i^j$ at the reference level that sums up to the observed pixel values $\mathbf{y}_v^j$ at the coarser resolution level j:

$$\mathbf{y}_v^j = \frac{1}{m_j} \sum_{i \in s_j(v)} \tilde{\mathbf{y}}_i^j. \tag{2}$$

[1] An extension to partly overlapping pixels and images is possible, but it implies a more complicated notation and implementation.

We can consider $\tilde{\mathbf{y}}_i^j$ as the observation that *would* be obtained if a sensor existed that had the radiometric properties of the sensor at level j, but acquired images at the reference resolution (level 1). A similar assumption has been used in i.e. Price [1999] and Zhukov et al. [1999].

We further assume that

$$\tilde{\mathbf{y}}_i^j|\mathbf{z} \sim N(\boldsymbol{\mu}_{z_i}^j, \boldsymbol{\Sigma}_{z_i}^j). \tag{3}$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Normal distribution with expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In the following we will use

$$\boldsymbol{\theta}_2 = \{(\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j), j = 1, ..., p, k = 1, ..., K\}$$

i.e. the set of parameters involved in the likelihood functions.

3 Classification

Assuming parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are known, classification is based on the posterior distribution

$$p(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}) \propto p(\mathbf{z}; \boldsymbol{\theta}_1) \prod_{j=1}^{p} p(\mathbf{y}^j|\mathbf{z}; \boldsymbol{\theta}_2) \tag{4}$$

The maximum a posteriori (MAP) is obtained by maximizing (4) with respect to the complete vector $\mathbf{z}$. Explicit MAP solutions have been obtained in Storvik and Dahl [2000] for problems with all observations given at the same resolution. Although such algorithms (based on integer programming and Lagrangian based methods) should be possible also to develop for the problem at hand, we will here consider a simpler version based on the ICM algorithm [Besag, 1986].

The ICM algorithm consist of sequential optimization of the components of $\mathbf{z}$ by maximizing $p(z_i|z_{i'}, i' \neq i, \mathbf{y})$ Using (1) and (4),

$$p(z_i|z_{i'}, i' \neq i, \mathbf{y}) \propto \exp\{\alpha_{z_i} - \beta \sum_{i':i' \sim i} \phi_{i,i'}(z_i, z_{i'})\} \prod_{j=1}^{p} p(\mathbf{y}_{v_j(i)}^j|\mathbf{z}). \tag{5}$$

The ICM steps are defined through calculating the quantities above for $z_i = 1, ..., K$ and choose the class with highest probability. From (2) and (3), we get

$$\mathbf{y}_v^j|\mathbf{z} \sim N(\frac{1}{m_j} \sum_{i \in s_j(v)} \boldsymbol{\mu}_{z_i}^j, \frac{1}{m_j} \sum_{i \in s_j(v)} \boldsymbol{\Sigma}_{z_i}^j) \tag{6}$$

making (5) relatively easy to calculate.

4 Parameter estimation

The model described in section 2 contains two sets of parameters

$$\begin{aligned}\boldsymbol{\theta}_1 &= \text{Parameters in the prior model for } \mathbf{z}\\ \boldsymbol{\theta}_2 &= \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, ..., K\}\\ &= \text{Parameters in likelihood for data conditional on } \mathbf{z}\end{aligned}$$

These parameters must either be estimated or specified by other means [Dubes and Jain, 1989]. In particular, assuming a training set $\mathcal{T}$ consisting of pixels with known classes are available, i.e. $\mathcal{T} = \{i; z_i \text{ known}\}$, the parameters in $\boldsymbol{\theta}_2$ can easily be estimated by maximum likelihood methods. Parameters in the prior models are possible to assign some reasonable values (i.e. all $\alpha_k = 0$ and $\beta \in [1, 2]$).

When training sets are small, utilizing observations from pixels with unknown class-labels is possible. A maximum likelihood estimate for the complete parameter vector $\boldsymbol{\theta}$ would then be to maximize

$$\begin{aligned}L(\boldsymbol{\theta}) &= p(\mathbf{y}|\mathcal{T}; \boldsymbol{\theta})\\ &= \sum_{\mathbf{z}\cap\mathcal{T}} p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_2) p(\mathbf{z}; \boldsymbol{\theta}_1)\end{aligned}$$

where $\mathbf{z} \cap \mathcal{T}$ means that $\mathbf{z}$ must satisfy the constraints given by the training set $\mathcal{T}$. In practice, numerical methods, such as the EM algorithm must be applied for this task.

Here, we consider a simpler approach, mainly because the presence of data at several resolutions makes the computational challenges even harder than in the standard case. We follow Besag [1986] by applying the following procedure:

1. Obtain an initial guess $\widehat{\mathbf{z}}$ of the true scene $\mathbf{z}^*$ with initial guesses for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

2. Estimate $\boldsymbol{\theta}_1$ by maximizing the *pseudo-likelihood*

$$\tilde{p}(\mathbf{z}|\boldsymbol{\theta}_1) = \prod_i p(z_i|z_{\partial i}; \boldsymbol{\theta}_1). \tag{7}$$

   on the current $\widehat{\mathbf{z}}$ to obtain a new $\widehat{\boldsymbol{\theta}}_1$. Here ∂i is a neighborhood of pixel i, typically the four nearest neighbors.

3. Estimate $\boldsymbol{\theta}_2$ by the value which maximizes $l(\mathbf{y}|\widehat{\mathbf{z}}, \boldsymbol{\theta}_2)$.

4. Carry out a *single* cycle of ICM based on the current $\widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$.

5. Return to 2 for fixed number of cycles or until approximative convergence.

This procedure can be considered as an algorithm for optimizing

$$p(\mathbf{z}, \mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = p(\mathbf{z}|\boldsymbol{\theta}_1) p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \tag{8}$$

with respect to both $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *and* $\mathbf{z}$. This is in contrast to ordinary maximum likelihood estimation where $\mathbf{z}$ is marginalized out. Basing the estimation on maximization of (8) *can* give biased estimates. For large amounts of data, this bias is hopefully small.

A further approximation is introduced in step 2 by estimating $\boldsymbol{\theta}_1$ based on the pseudo-likelihood for $\boldsymbol{\theta}_1$ based on $\mathbf{z}$ rather than the actual likelihood function. This approximation is introduced because of the difficulty in maximizing $p(\mathbf{z}|\boldsymbol{\theta}_1)$ with respect to $\boldsymbol{\theta}_1$.

Some remarks on the different steps:

Step 1 We follow the general suggestion from Besag [1986] by ignoring spatial dependence for the initial guess of $\widehat{\mathbf{z}}$. This makes an initial guess for $\boldsymbol{\theta}_1$ unnecessary. An alternative is to use some reasonable initial values for $\boldsymbol{\theta}_1$.

Initial guesses for $\boldsymbol{\theta}_2$ can be based on training data. Those data sources that do not have any training data on the corresponding resolution can be ignored in the initial step.

Step 2 The pseudo-likelihood is performed through Newton-Raphson. The details are given in appendix A.

Step 3 This part consist of ordinary maximum likelihood estimation conditioned on a known class-configuration. A complicated step is however that data at coarser resolution will be a mixture of contributions from different classes. This requires specifically tuned algorithms for this part. We propose the use of the EM algorithm, treating the contributions from different classes in the mixture observation as missing, for this purpose. The details are given in the following subsection.

Step 4 This is equivalent to classification based on the ICM algorithm as described in section 4.

4.1 Estimation of $\boldsymbol{\theta}_2$

Let us assume that $\widehat{\mathbf{z}}$ is known. We want to maximize the likelihood $p(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$. Now

$$p(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_2) = \prod_{j=1}^{p} \prod_{v} p(\mathbf{y}_v^j|\mathbf{z};\boldsymbol{\theta}_2)$$

and from (6) $p(\mathbf{y}_v^j|\mathbf{z};\boldsymbol{\theta}_2)$ is known. If all data were given at the reference resolution, this step would correspond to ordinary maximum likelihood estimation and analytical expressions for the optimal value of $\boldsymbol{\theta}_2$ is available. In the general case with data at different resolutions, this is no longer the case, and optimization will have to be performed through some sort of numerical optimizer. Gradient based optimizers (such as Newton-Raphson) can be applied, but we will consider an alternative where the variables $\tilde{\mathbf{y}}_i^j$ defined in (2) are considered as missing data. Estimation of $\boldsymbol{\theta}_2$ can then

be performed through the EM algorithm. Define $\tilde{\mathbf{y}} = \{\tilde{\mathbf{y}}_i^j\}$ to be the complete data. Note that $\mathbf{y}$ is a (deterministic) function of $\tilde{\mathbf{y}}$. The EM algorithm at iteration $s+1$ is defined through the follow two steps:

E-step Compute

$$Q(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^s) = E^{p(\tilde{\mathbf{y}}|\mathbf{y},\mathbf{z};\boldsymbol{\theta}_2^s)} \log[p(\tilde{\mathbf{y}}|\mathbf{z};\boldsymbol{\theta}_2)]$$

M-step Maximize $Q(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^s)$ with respect to $\boldsymbol{\theta}_2$ to obtain $\boldsymbol{\theta}_2^{s+1}$.

Now

$$\log[p(\tilde{\mathbf{y}}|\mathbf{z};\boldsymbol{\theta}_2)] = \sum_{j=1}^{p}\sum_{k=1}^{K} \log[p(\tilde{\mathbf{y}}^j(k)|\mathbf{z};\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j)]$$

where $\tilde{\mathbf{y}}^j(k)$ is the set of complete data corresponding to pixels where the class is k, i.e. $\tilde{\mathbf{y}}^j(k) = \{\tilde{\mathbf{y}}_i^j; z_i = k\}$. This shows that

$$Q(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^s) = \sum_{j=1}^{p}\sum_{k=1}^{K} Q_k^j(\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j, \boldsymbol{\theta}_2^s)$$

where

$$Q_k^j(\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j, \boldsymbol{\theta}_2^s) = E^{p(\tilde{\mathbf{y}}|\mathbf{y},\mathbf{z};\boldsymbol{\theta}_2^s)} \log[p(\tilde{\mathbf{y}}^j(k)|\mathbf{z};\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j)]$$

Since $(\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j)$ is only contained in $Q_k^j(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^s)$, maximization can be performed on each source and each class separately. Further

$$\begin{aligned}
&\log[p(\tilde{\mathbf{y}}^j(k)|\mathbf{z};\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j)] \\
&\quad = \sum_i I(z_i = k) \log[p(\tilde{\mathbf{y}}_i^j|z_i;\boldsymbol{\mu}_k^j, \boldsymbol{\Sigma}_k^j)] \\
&\quad = C - \frac{n_k}{2}\log(|\boldsymbol{\Sigma}_k^j|) - \frac{1}{2}\sum_i I(z_i = k)(\tilde{\mathbf{y}}_i^j - \boldsymbol{\mu}_k^j)^T[\boldsymbol{\Sigma}_k^j]^{-1}(\tilde{\mathbf{y}}_i^j - \boldsymbol{\mu}_k^j)
\end{aligned}$$

where C is a constant not depending on the parameters in $\boldsymbol{\theta}_2$ or on the observations, while n_k is the number of pixels from class k. Now $\tilde{\mathbf{y}}_i^j|\mathbf{y},\mathbf{z} \sim N(\boldsymbol{\eta}_i^j, \mathbf{S}_i^j)$ where a procedure for calculating $\boldsymbol{\eta}_i^j$ and $\mathbf{S}_i^j$ is given in appendix B. Then

$$\begin{aligned}
&\mathsf{E}^{p(\tilde{\mathbf{y}}|\mathbf{y},\mathbf{z};\boldsymbol{\theta}_2^s)}[(\tilde{\mathbf{y}}_i^j - \boldsymbol{\mu}_k^j)^T[\boldsymbol{\Sigma}_k^j]^{-1}(\tilde{\mathbf{y}}_i^j - \boldsymbol{\mu}_k^j)] \\
&\quad = \mathsf{E}^{p(\tilde{\mathbf{y}}|\mathbf{y},\mathbf{z};\boldsymbol{\theta}_2)}\mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}(\tilde{\mathbf{y}}_i^j - \boldsymbol{\eta}_i^j + \boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)(\tilde{\mathbf{y}}_i^j - \boldsymbol{\eta}_i^j + \boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T) \\
&\quad = \mathsf{E}^{p(\tilde{\mathbf{y}}|\mathbf{y},\mathbf{z};\boldsymbol{\theta}_2)}\mathsf{tr}\left([\boldsymbol{\Sigma}_k^j]^{-1}((\tilde{\mathbf{y}}_i^j - \boldsymbol{\eta}_i^j)(\tilde{\mathbf{y}}_i^j - \boldsymbol{\eta}_i^j)^T + (\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T)\right. \\
&\qquad\qquad \left. -2(\tilde{\mathbf{y}}_i^j - \boldsymbol{\eta}_i^j)(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T\right) \\
&\quad = \mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}(\mathbf{S}_i^j + (\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T)) \\
&\quad = \mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}\mathbf{S}_i^j) + (\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T[\boldsymbol{\Sigma}_k^j]^{-1}(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)
\end{aligned}$$

Further [Mardia et al., 1979, page 97]

$$\begin{aligned}
Q_j^k(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^s) =& C_j^k - \frac{n_k}{2}\log(|\boldsymbol{\Sigma}_k^j|) - \frac{1}{2}\sum_i I(z_i = k)\mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}\mathbf{S}_i^j) \\
& - \frac{1}{2}\sum_i I(z_i = k)(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j)^T[\boldsymbol{\Sigma}_k^j]^{-1}(\boldsymbol{\eta}_i^j - \boldsymbol{\mu}_k^j) \\
=& C_j^k - \frac{n_k}{2}\log(|\boldsymbol{\Sigma}_k^j|) - \frac{1}{2}\mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}\overline{\mathbf{S}}_k^j) - \frac{n_k}{2}\mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}\mathbf{V}_k^j) \\
& - \frac{n_k}{2}\mathsf{tr}([\boldsymbol{\Sigma}_k^j]^{-1}(\overline{\boldsymbol{\eta}}_k^j - \boldsymbol{\mu}_k^j)^T))[\boldsymbol{\Sigma}_k^j]^{-1}(\overline{\boldsymbol{\eta}}_k^j - \boldsymbol{\mu}_k^j)
\end{aligned}$$

where

$$\begin{aligned}
\overline{\boldsymbol{\eta}}_k^j =& \frac{1}{n_k}\sum_i I(z_i = k)\boldsymbol{\eta}_i^j \\
\overline{\mathbf{S}}_k^j =& \frac{1}{n_k}\sum_i I(z_i = k)\mathbf{S}_i^j \\
\mathbf{V}_k^j =& \frac{1}{n_k}\sum_i I(z_i = k)(\boldsymbol{\eta}_i^j - \overline{\boldsymbol{\eta}}_k^j)(\boldsymbol{\eta}_i^j - \overline{\boldsymbol{\eta}}_k^j)^T
\end{aligned}$$

From this we see that the optimal value for $\boldsymbol{\mu}_k^j$ and $\boldsymbol{\Sigma}_k^j$ are

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_k^j =& \overline{\boldsymbol{\eta}}_k^j, \\
\widehat{\boldsymbol{\Sigma}_k^j} =& \overline{\mathbf{S}}_k^j + \mathbf{V}_k^j.
\end{aligned}$$

Remark 1 For sources that are observed at the reference resolution, $\tilde{\mathbf{y}}_j = \mathbf{y}_j$ resulting in that $\boldsymbol{\eta}_i^j = \mathbf{y}_i^j$ and $\mathbf{S}_i^j = \mathbf{0}$.

Remark 2 Since this EM algorithm is part of an outer maximization algorithm, only a small number of iterations should be sufficient, at least at an early stage of the outer algorithm.

The full procedure is then given by the following algorithm

Algorithm 1 Repeat the following steps a sufficient number of iterations:

1. For each pixel v at resolution j, compute

$$\begin{aligned}
\overline{\boldsymbol{\mu}}_2 =& \frac{1}{m_j}\sum_{i \in s_j(v)} \boldsymbol{\mu}_{z_i}^j, \\
\overline{\boldsymbol{\Sigma}}_{22} =& \frac{1}{m_j^2}\sum_{i \in s_j(v)} \boldsymbol{\Sigma}_{z_i}^j, \\
\mathbf{A}_2 =& \overline{\boldsymbol{\Sigma}}_{22}^{-1}(\mathbf{y}_v^j - \overline{\boldsymbol{\mu}}_2).
\end{aligned}$$

(a) For each pixel i at reference resolution inside pixel v at resolution j, compute

$$\boldsymbol{\eta}_i^j = \mathsf{E}[\tilde{\mathbf{y}}_i^j | \mathbf{y}_v^j, \mathbf{z}] = \boldsymbol{\mu}_{z_i} + \frac{1}{m_j} \boldsymbol{\Sigma}_{z_i}^j \mathbf{A}_2$$
$$\mathbf{S}_i^j = \mathsf{Var}[\tilde{\mathbf{y}}_i^j | \mathbf{y}_v^j, \mathbf{z}] = \boldsymbol{\Sigma}_{z_i} - \frac{1}{m_j^2} \boldsymbol{\Sigma}_{z_i} \overline{\boldsymbol{\Sigma}}_{22}^{-1} \boldsymbol{\Sigma}_{z_i}$$

2. For $k = 1, ..., K$, compute

$$\overline{\boldsymbol{\eta}}_k^j = \frac{1}{n_k} \sum_i I(z_i = k) \boldsymbol{\eta}_i^j$$
$$\overline{\mathbf{S}}_k^j = \frac{1}{n_k} \sum_i I(z_i = k) \mathbf{S}_i^j$$
$$\mathbf{V}_k^j = \frac{1}{n_k} \sum_i I(z_i = k) (\boldsymbol{\eta}_i^j - \overline{\boldsymbol{\eta}}_k^j)(\boldsymbol{\eta}_i^j - \overline{\boldsymbol{\eta}}_k^j)^T$$

where $n_k = \sum_i I(z_i = k)$, and

$$\widehat{\boldsymbol{\mu}}_k^j = \overline{\boldsymbol{\eta}}_k^j,$$
$$\widehat{\boldsymbol{\Sigma}}_k^j = \overline{\mathbf{S}}_k^j + \mathbf{V}_k^j.$$

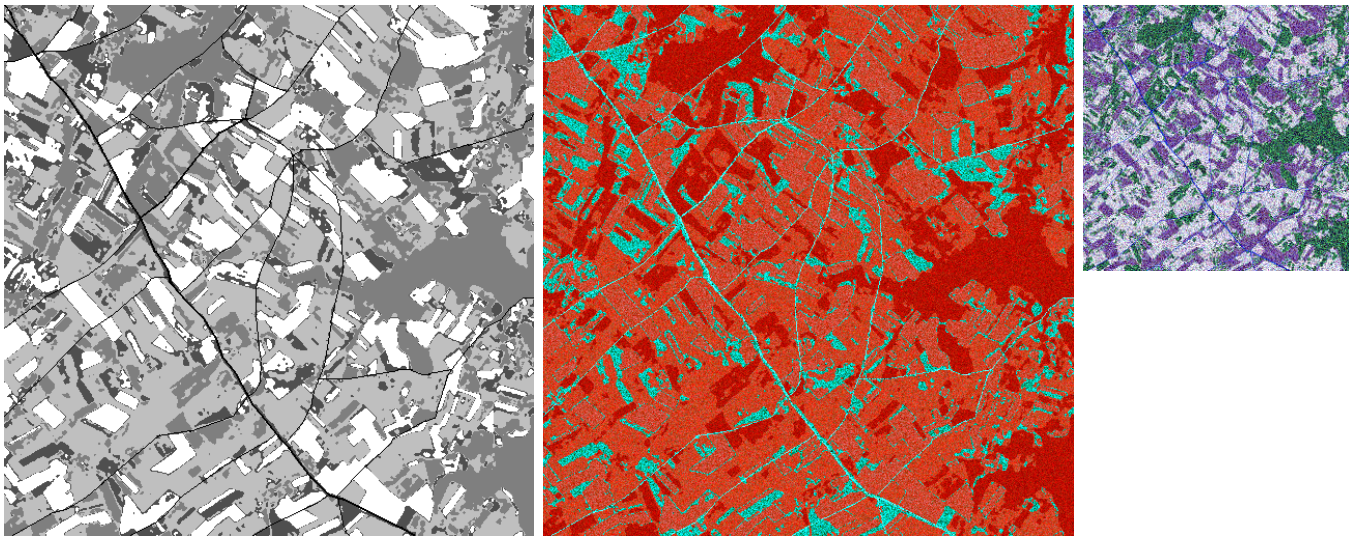
5 Results

5.1 Simulated images

The proposed multi-scale classification scheme has been tested on a simulated data set, where the class properties are derived from real SPOT XS and Landsat TM images. The reason why we use simulated images is that the new method only can be expected to bring an improvement in the presence of fine details, i.e. fine structures such as roads and transitions corresponding to region boundaries. Sufficiently detailed ground truth is rarely available for such zones, as the ground truth samples generally are placed well inside homogeneous regions. This was also the case for the real images that were available to us.

The class label image that we used is shown in Fig. 2 (a). It represents an agricultural scene and is derived from a class image used in a previous study [Fjørtoft et al., 2003], by manually adding roads, so that the number of classes becomes 5. The rest of the image is derived from a real satellite image. It should be noted that the class image not is a realization of a Markov random field.

Covariance matrices and mean vectors for roads (class 1), urban areas (class 2), forests (class 3) and two kinds of agricultural crops (class 4 and class 5) were estimated from a SPOT XS image of Bourges, France, and the associated ground truth. The 3-band image shown in Fig. 2 (b) was obtained by simulating multivariate Gaussian noise according to the estimated class properties, and by attributing the signal vectors to the pixels of the ideal class image.



(a) (b) (c)

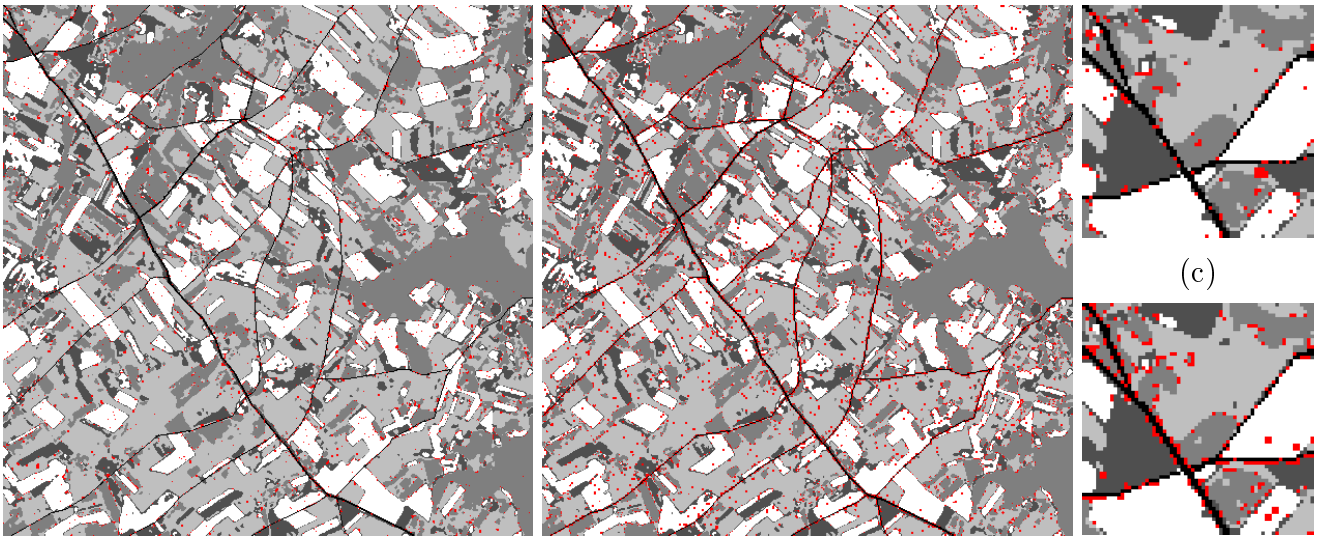
Figure 2: Simulated images for multi-scale classification: (a) 512×512 class label image with 5 classes. (b) 3-band 512×512 simulated image with class characteristics derived from SPOT XS data. (c) 6-band 256×256 simulated image with class characteristics derived from Landsat TM data (3 bands shown).

Likewise, a 6-band image was simulated from the class properties of a Landsat TM image of Kjeller, Norway, and the label image shown in Fig. 2 (a). However, the estimated covariance matrices were first multiplied by a factor of 4, then 2×2 pixel blocks of the simulated image were averaged, and the entire image was down-sampled by 2×2. Hence we obtained the image shown in Fig. 2 (c), which has 2×2 times lower resolution than the initial class image, and contains mixed pixels at region boundaries, in agreement with our multi-scale model. The class properties of the simulated image are very close to those of the original Landsat TM image.

The simulated images looks somewhat noisier than real SPOT XS and Landsat TM images. This is due to the fact that there often are slow spatial variations in the class properties of real images that contribute to the overall class variance. In a simulated image generated as described above, the total class variance can be observed between neighboring pixels.

A 100×100 pixels square in the ideal class image shown in Fig. 2 (a), containing a sufficient number of samples of all 5 classes, was defined as training set, and the rest of the ideal class image constituted the test set. To make the results as representative as possible, we decided to separate the estimation and classification phases, by first running the ICM loop until convergence with the training set fixed to estimate the radiometric and spatial parameters, and then, by running the ICM until convergence without fixing the test set and without parameter estimation, in order to obtain the final classification. For practical applications, it would be sufficient to run the ICM until convergence only once, with fixed training set, estimating the parameters and simultaneously iterating towards the final classification.

The result of the multi-scale classification scheme is shown in Fig. 3 (a) and the con-



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

Figure 3: Classification result of (a) the multi-scale classification scheme and (b) the corresponding single-scale classification scheme applied after resampling of the second image to the reference resolution. Correctly classified pixels are shown in gray levels and misclassified pixels are shown in red. The upper left corners are magnified to show some representative details of (c) the multi-scale result and (d) the single-scale result.

Table 1: Confusion matrix for the multi-scale result in Fig. 3 (a)

$p(\hat{z}\|z)$	$\hat{z}=1$	$\hat{z}=2$	$\hat{z}=3$	$\hat{z}=4$	$\hat{z}=5$
$z=1$	0.893	0.107	0.000	0.000	0.000
$z=2$	0.015	0.985	0.000	0.000	0.000
$z=3$	0.000	0.000	0.993	0.006	0.001
$z=4$	0.000	0.000	0.007	0.966	0.027
$z=5$	0.000	0.000	0.008	0.006	0.986

Table 2: Confusion matrix for the single-scale result in Fig. 3 (b)

$p(\hat{z}\|z)$	$\hat{z}=1$	$\hat{z}=2$	$\hat{z}=3$	$\hat{z}=4$	$\hat{z}=5$
$z=1$	0.944	0.056	0.000	0.000	0.000
$z=2$	0.022	0.978	0.000	0.000	0.000
$z=3$	0.010	0.000	0.957	0.031	0.002
$z=4$	0.010	0.000	0.005	0.940	0.044
$z=5$	0.008	0.000	0.002	0.037	0.952

fusion matrix (rows=reference, columns=result) is given in Table 1. The convergence criterion was reached after 7 iterations with the training set fixed, and the computing time on a Linux PC with a Pentium III 733 MHz processor was 8 m 33 s. For the additional ICM loop without parameter estimation and fixed training set (which can

be omitted in practical applications), convergence was reached after 6 iterations and the processing time was 3 m 42 s. The overall probability of correct classification, computed on the test set, was 97.8 %.

To assess the improvement brought by the new multi-scale approach, we compared it with the corresponding single-scale analysis of the same data set. First we resampled the image with Landsat TM properties to the reference resolution by replacing each pixel by a block of 2×2 identical pixels, and concatenated the resulting image bands with those of the image with SPOT XS properties, so that we got a 512×512 with 9 bands. ICM loops for parameter estimation and classification were carried out until convergence. The only difference from the multi-scale algorithm is that the computations related specifically to the multi-scale model are let out. The result is shown in Fig. 3 (b) and the confusion matrix is given in Table 2. Convergence was reached after 13 iterations and 1 m 42 s for the combined estimation and classification ICM loop, and after 4 iterations and 16 s for the final (optional) classification loop. The overall probability of correct classification was here 95.2 %. For comparison, we reached 90.9 % using only the 3-band image with SPOT XS properties, and 78.3 % when using only the resampled 6-band image with Landsat TM properties.

Table 3: Performance on simulated data set

Method	Data	Accuracy	Time
Multi-scale ICM	XS + TM	97.8 %	8 m 33 s
Single-scale ICM	XS + TM	95.2 %	1 m 42 s
Single-scale ICM	XS only	90.9 %	45 s
Single-scale ICM	TM only	78.3 %	54 s

The most important performance measures are summarized in Table 3. The computing time excludes the optional second ICM loop, that can be omitted in practise. We see that there are only 9.1 % of misclassified pixels when applying the single-scale ICM scheme to the image with SPOT XS properties only. The same method applied to a concatenation of this image and a resampled version of the image with Landsat TM properties further reduces the proportion of misclassified pixels by 4.3 %, and another 2.6 % are eliminated with the new multi-scale method, so that only 2.2 % of misclassified pixels remain. However, the last improvement is obtained at the cost of 5 times higher computing time.

As can be seen from Fig. 3, the multi-scale approach, as could be expected, mainly improves the classification accuracy near fine structures and boundaries between regions. Careful study of the confusion matrices in Tables 1 and 2 reveals that the multi-scale scheme for this data set does not improve the number of correctly classified pixels of class 1 (roads), as the spectral characteristics of this class are quite distinct, but it significantly reduces the confusion with this class, i.e. that it performs better in the neighborhood of these fine structures. It also performs significantly better at the boundaries between larger regions.

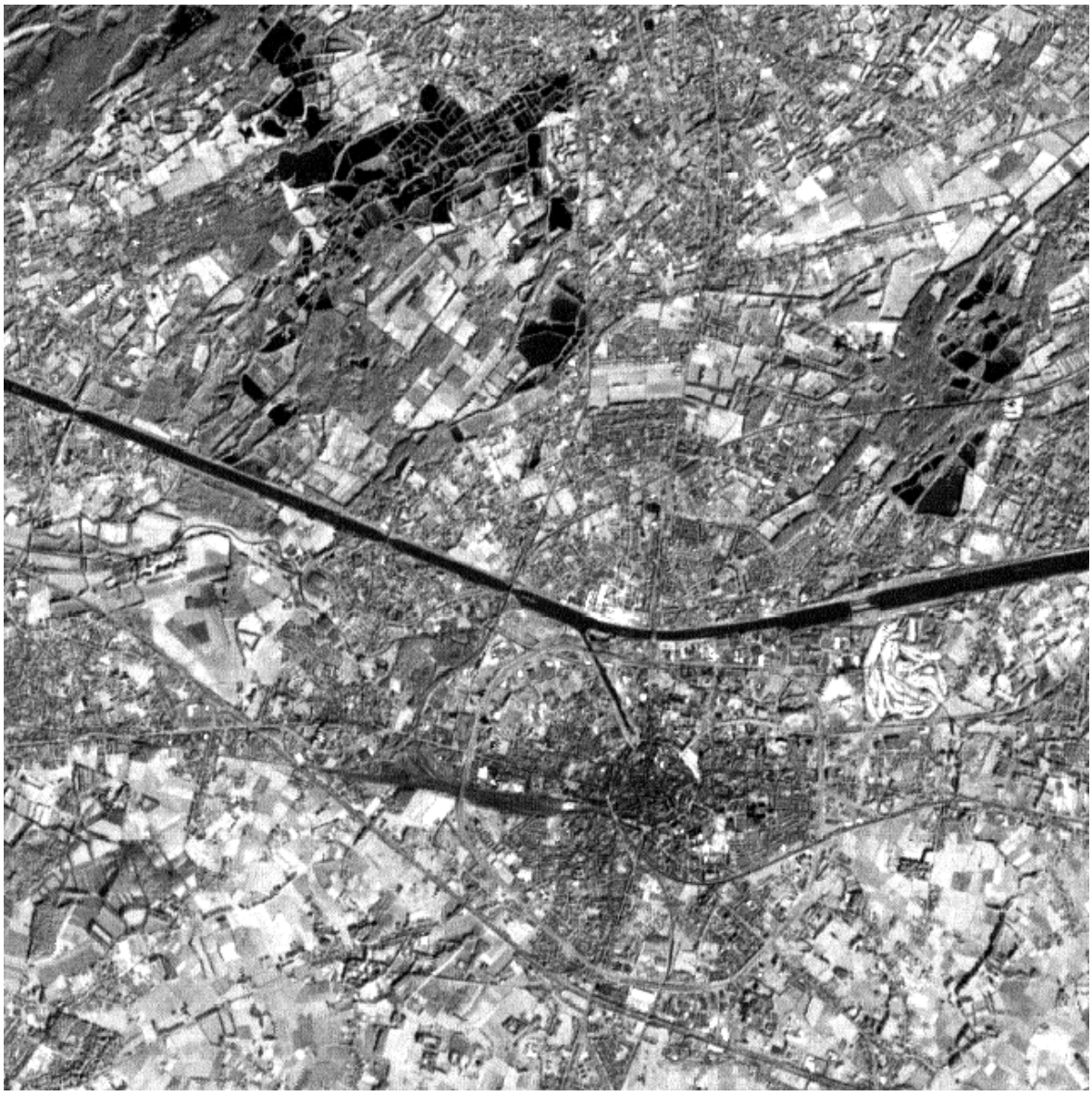
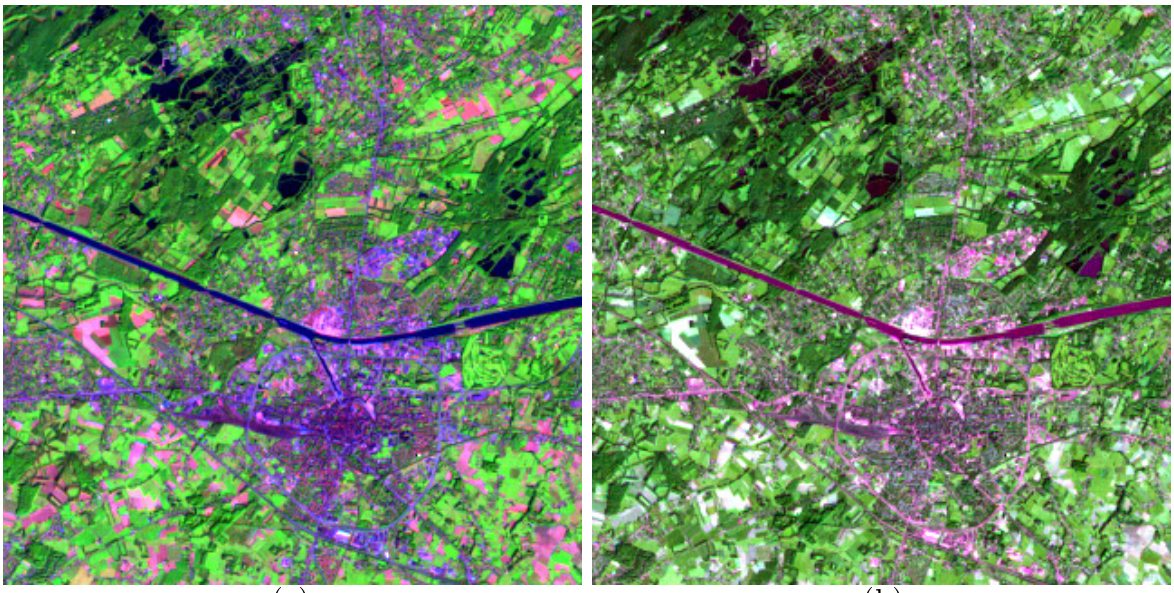


Figure 4: 640×640 extract of the panchromatic band with 15 m resolution of a Landsat TM image of Hasselt, Belgium. Landsat 7 image courtesy ESA 1999 – distribution Eurimage.

5.2 SPOT and Landsat images of Hasselt, Belgium

The multi-scale classification scheme has also been tested on real data. Fig. 4 shows a 640×640 extract of the panchromatic band of a Landsat TM image. It has 15 m resolution and covers Hasselt, Belgium. The six spectral bands with 30 m resolution of the Landsat TM image are displayed as color composites in Fig. 5. The data set also includes the SPOT XS image in Fig. 6 that has been co-registered with the panchromatic band of the Landsat TM image and resampled to 15 m resolution. The images are from the benchmark data sets "grss_dfc_0002" (Landsat) and "grss_dfc_0003" (SPOT)



(a) (b)

Figure 5: Spectral bands with 30 m resolution of a Landsat TM image of Hasselt, Belgium. (a) 320 × 320 color composite of bands 7, 4 and 1. (b) 320 × 320 color composite of bands 2, 5 and 3. Landsat 7 image courtesy ESA 1999 – distribution Eurimage.

of the Data Fusion Committee of the IEEE Geoscience and Remote Sensing Society. The training and test sets are illustrated in Figs. 7 and 8, respectively. The six classes that have been defined (and the colors used to represent them) are:

1. Water (blue)
2. Urban area (red)
3. Forest 1 (dark green)
4. Forest 2 (green)
5. Agriculture 1 (yellow)
6. Agriculture 2 (cyan)

The regions in the training and test sets were selected by photo-interpretation and digitized and represented as label images using ENVI's region of interest (ROI) tool. Regions that are not in the same state in both images were avoided.

As usual, the ground truth samples are placed well inside homogeneous regions, so we cannot expect the quantitative performance measures to reflect what happens near fine structures and region borders. It should be noted that the Landsat and SPOT images are not acquired at the same date, so that there are differences in landcover between them. In particular, this concerns the classes agriculture 1 (vegetation) and

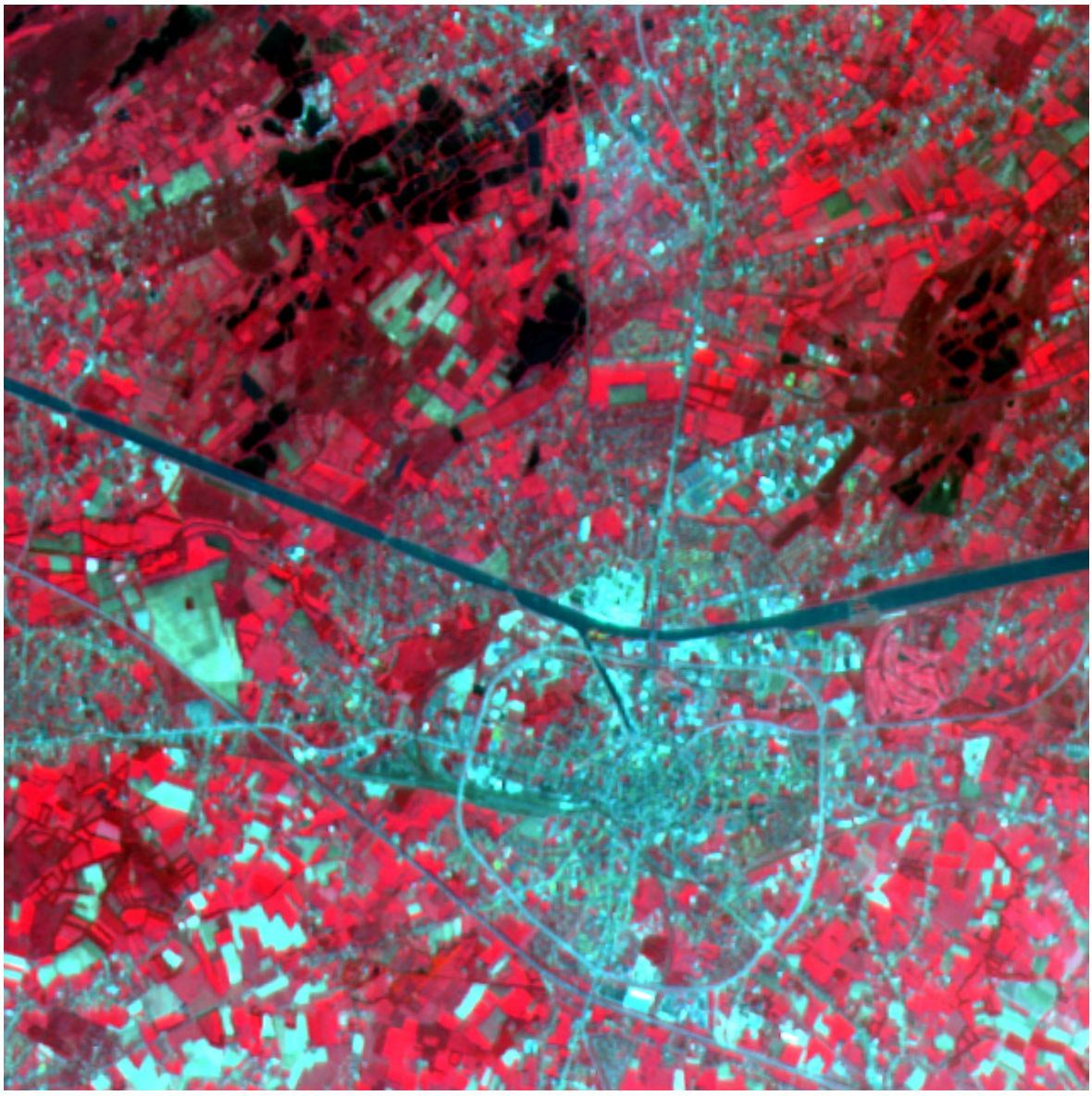


Figure 6: Color composite of bands 3, 2 and 1 of a SPOT XS image of Hasselt, Belgium, which has been co-registered with the Landsat TM image in Fig. 4 and resampled to 15 m resolution. ©Copyright CNES 1995 – distribution SPOT Image.

agriculture 2 (bare soil). When creating the training and test sets, we have therefore only included areas that seem to be in the same state in both images.

Using the same settings as for the simulated data set presented in section 5.1, we performed iterative parameter estimation and classification both with the multi-scale scheme and the corresponding single-scale scheme. The classified images are shown in Figs. 9 and 10, and the confusion matrices are presented in Tables 4 and 5 respectively.

The overall probability of correct classification in the tests set is 97.6 % for the multi-scale scheme and 97.3 % for the single-scale scheme. Despite the insignificant difference in measured performance, some differences can be noticed visually. The

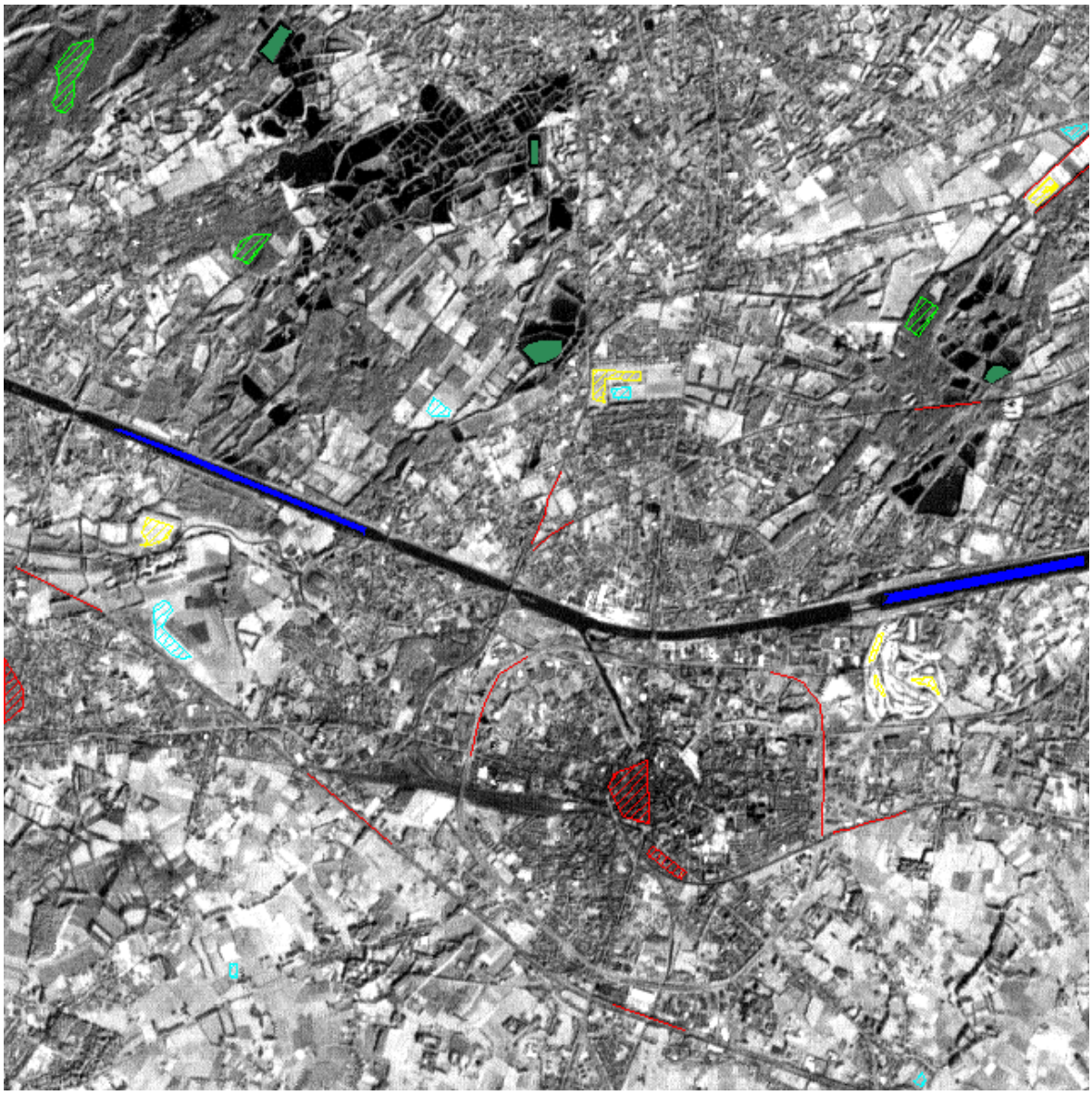


Figure 7: Training set with 6 classes overlayed on the 640×640 panchromatic Landsat TM band with 15 m resolution.

region borders are much smoother for the multi-scale approach, and some structures are better preserved, e.g between the two forest classes in the upper left quarter of the image. We also see that the decision limits between the classes Agriculture 1 (vegetation), Agriculture 2 (bare soil) and Urban have become different for the two approaches through the iterative parameter estimation phase. As mentioned above, there are some differences in the landcover between the Landsat and SPOT images due to different acquisition dates and agricultural activity. The multi-scale result here seems close to the situation depicted by the SPOT image (15 m), whereas the result of the single-scale result is closer to the situation in the Landsat image, where the majority of the bands have lower resolution (30 m).

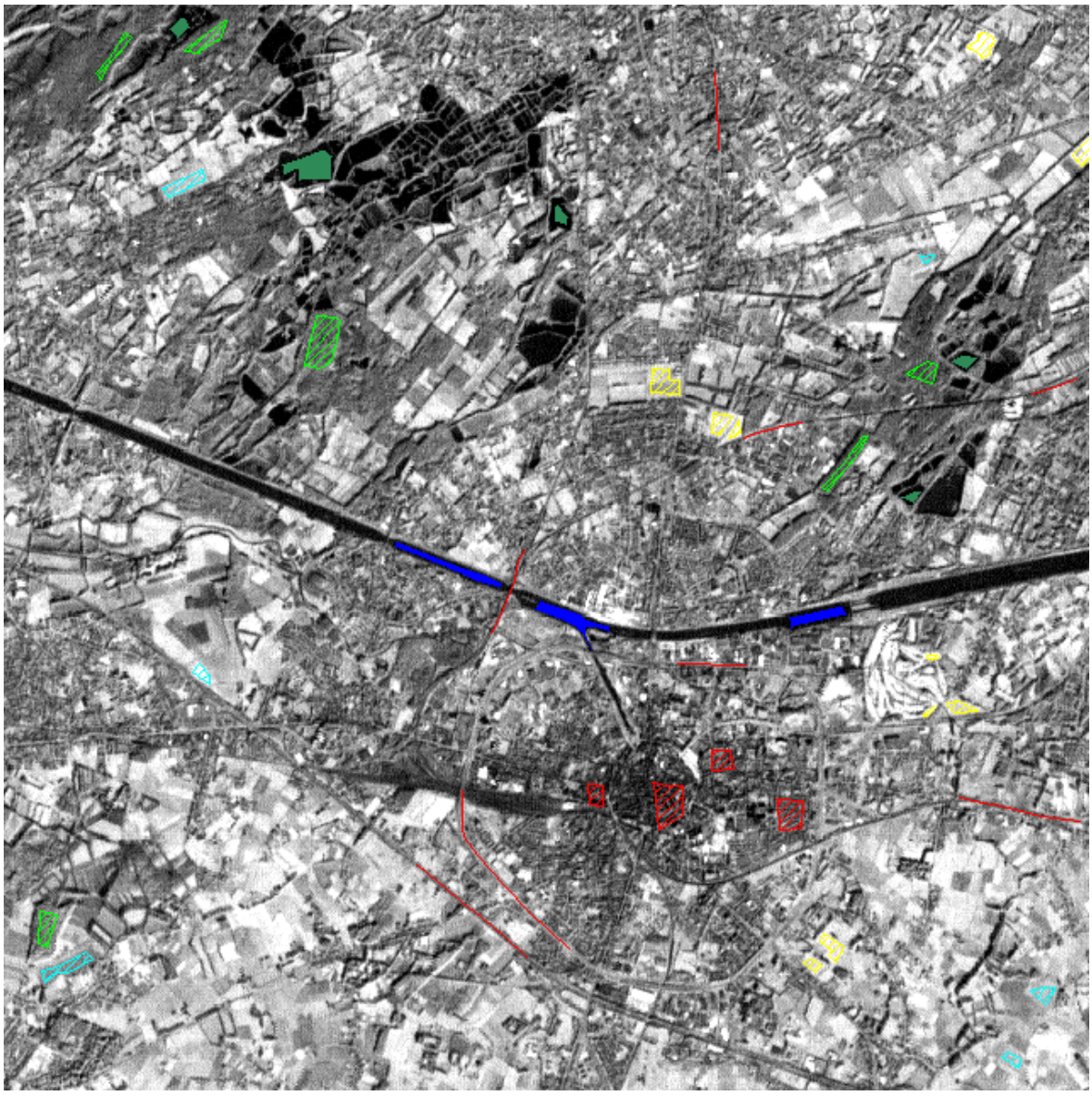


Figure 8: Test set with 6 classes overlayed on the 640×640 panchromatic Landsat TM band with 15 m resolution.

The differences between the single-scale and the multi-scale classification results are smaller for this real data set than for the simulated images described in section 5.1, probably because the real data set does not fit the multi-scale model equally well:

- The 15 and 30 m Landsat bands are not independent (same spectral domain).
- The SPOT image is resampled to the same geometry as the Landsat image and from 20 to 15 m resolution (implying mixed pixels at the reference resolution).
- There is significant intra-class variation (can be interpreted as a mixture of pixels from different classes at coarser resolution levels and lead to increased confusion).

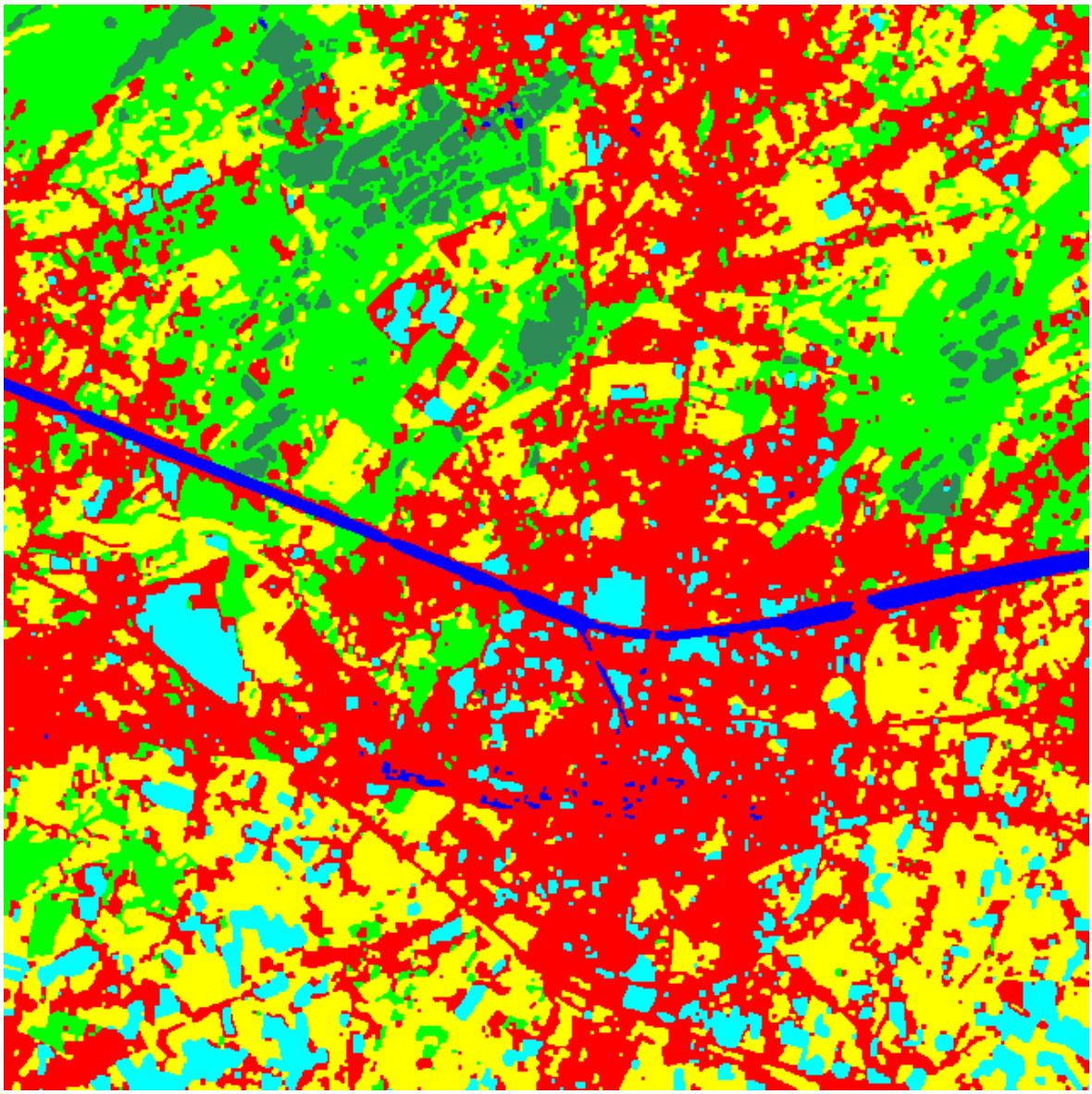


Figure 9: Result of multi-scale classification of the images of Hasselt, Belgium.

Table 4: Confusion matrix for the multi-scale result in Fig. 9

$p(\hat{z}\|z)$	$\hat{z}=1$	$\hat{z}=2$	$\hat{z}=3$	$\hat{z}=4$	$\hat{z}=5$	$\hat{z}=6$
$z=1$	1.000	0.000	0.000	0.000	0.000	0.000
$z=2$	0.019	0.941	0.000	0.020	0.004	0.016
$z=3$	0.000	0.000	1.000	0.000	0.000	0.000
$z=4$	0.000	0.008	0.000	0.992	0.000	0.000
$z=5$	0.000	0.000	0.000	0.000	1.000	0.000
$z=6$	0.000	0.062	0.000	0.000	0.004	0.934

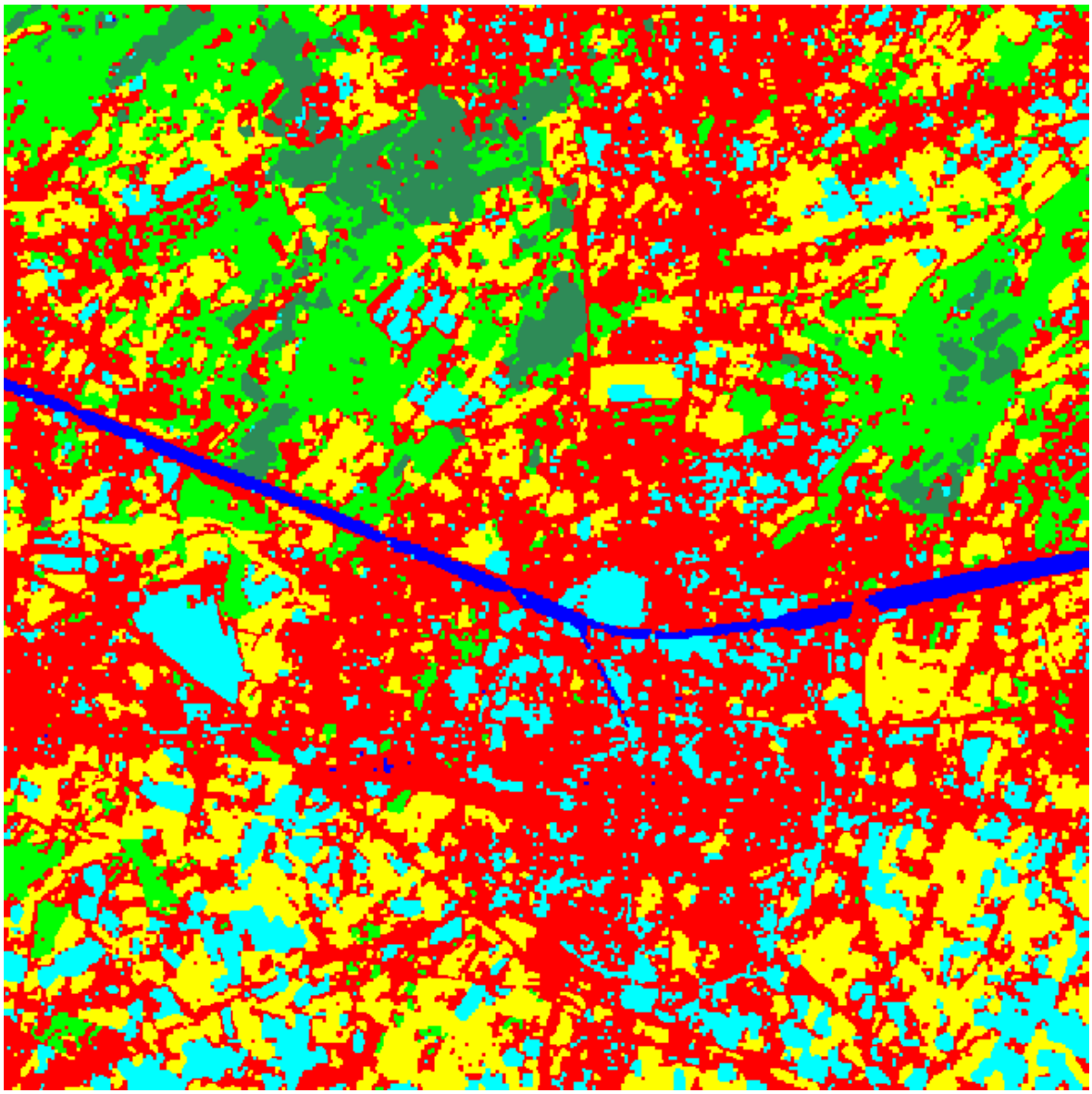


Figure 10: Result of single-scale classification of the images of Hasselt, Belgium.

Table 5: Confusion matrix for the single-scale result in Fig. 10

$p(\hat{z}\|z)$	$\hat{z}=1$	$\hat{z}=2$	$\hat{z}=3$	$\hat{z}=4$	$\hat{z}=5$	$\hat{z}=6$
$z=1$	0.995	0.005	0.000	0.000	0.000	0.000
$z=2$	0.004	0.918	0.000	0.016	0.004	0.058
$z=3$	0.000	0.000	1.000	0.000	0.000	0.000
$z=4$	0.000	0.014	0.000	0.986	0.000	0.000
$z=5$	0.000	0.005	0.000	0.000	0.995	0.000
$z=6$	0.000	0.010	0.000	0.000	0.000	0.990

Figure 11: Panchromatic band (2.5 m) of SPOT 5 image from Østfold, Norway. ©Copyright CNES 2002 - distribution SPOT Image.

5.3 SPOT 5 image from Østfold, Norway

Comparison of the single-scale and the multi-scale classification has also been carried out on a SPOT 5 image consisting of a panchromatic band with 2.5 m resolution and four XS spectral bands with 10 m resolution. In order to reduce the computing time, we only processed the 1024×1024 extract of the panchromatic band shown in Fig. 11 and the corresponding 256×256 extract of the XS spectral bands shown in Fig. 12.

A training set with 10 classes was created based on visual inspection and is shown in Fig. 13. In an operational system, training data could be derived from existing maps. The classes (and color coding, using ENVI terminology) are:

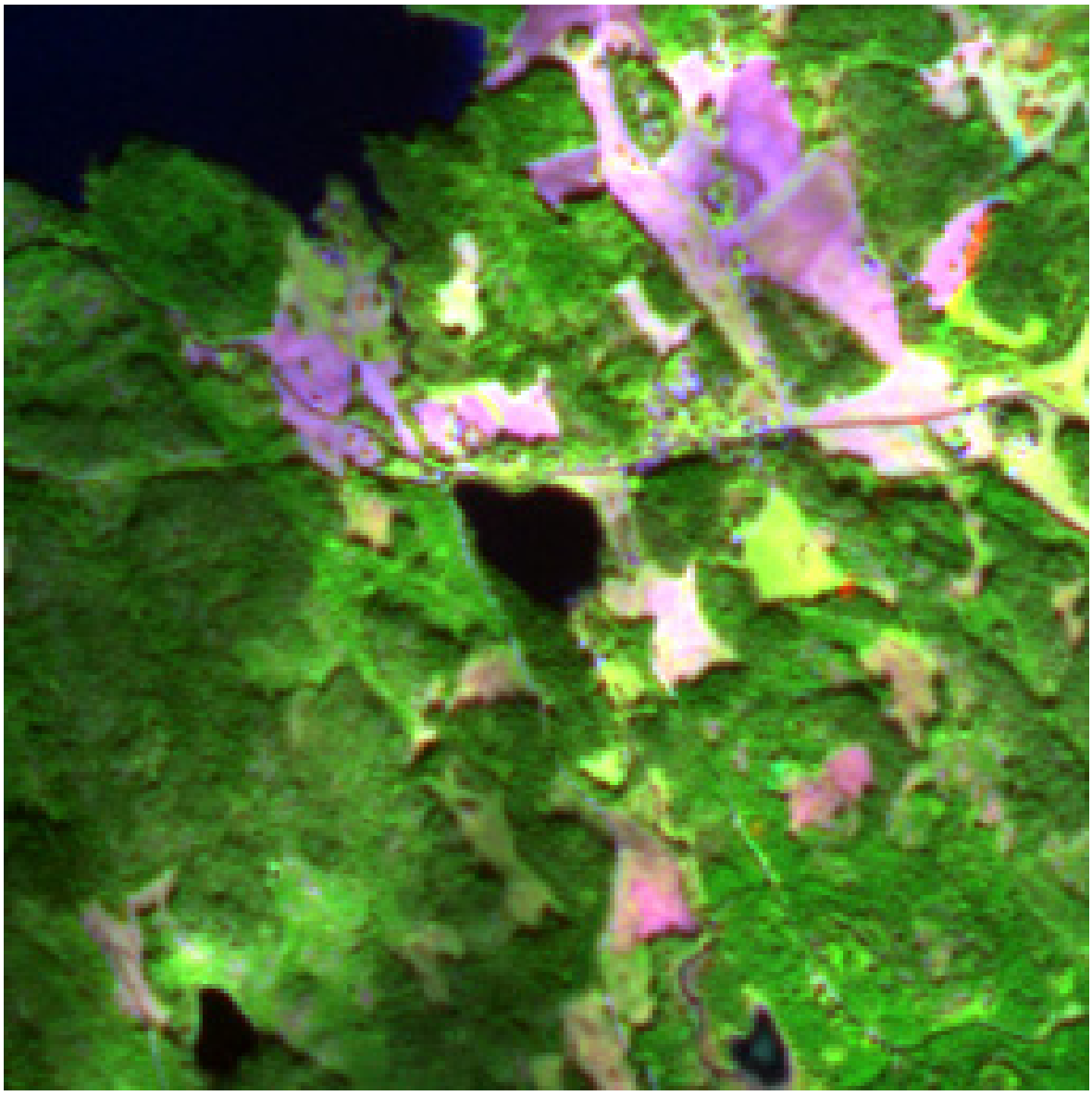


Figure 12: 4×4 magnification of RGB composite of XS bands (10 m) of SPOT 5 image from Østfold, Norway. ©Copyright CNES 2002 - distribution SPOT Image.

1. Water (blue)
2. Forest (sea green)
3. Forest 2 (green)
4. Agriculture 1 (yellow)
5. Agriculture 2 (thistle 1)
6. Agriculture 3 (orchid)

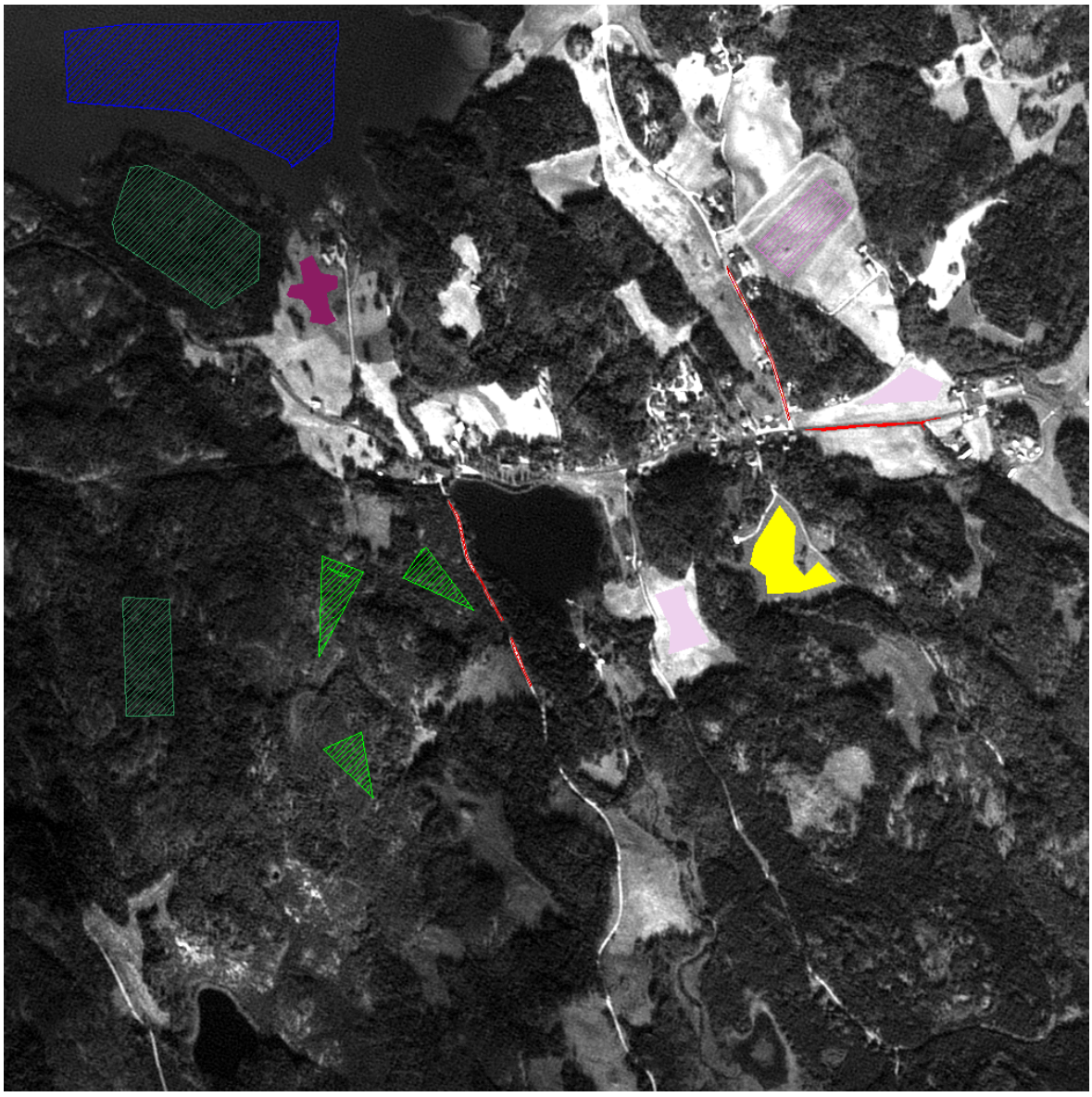


Figure 13: Training set with 10 classes shown with the panchromatic band of the SPOT 5 image as background.

7. Agriculture 4 (maroon 3)
8. Road 1 (red)
9. Road 2 (red 1)
10. Road 3 (red 2)

It may not be necessary to distinguish between e.g. different kinds of roads in the end result, but it was here necessary to define several road classes for the automatic analysis, due to great variability in spectral properties.

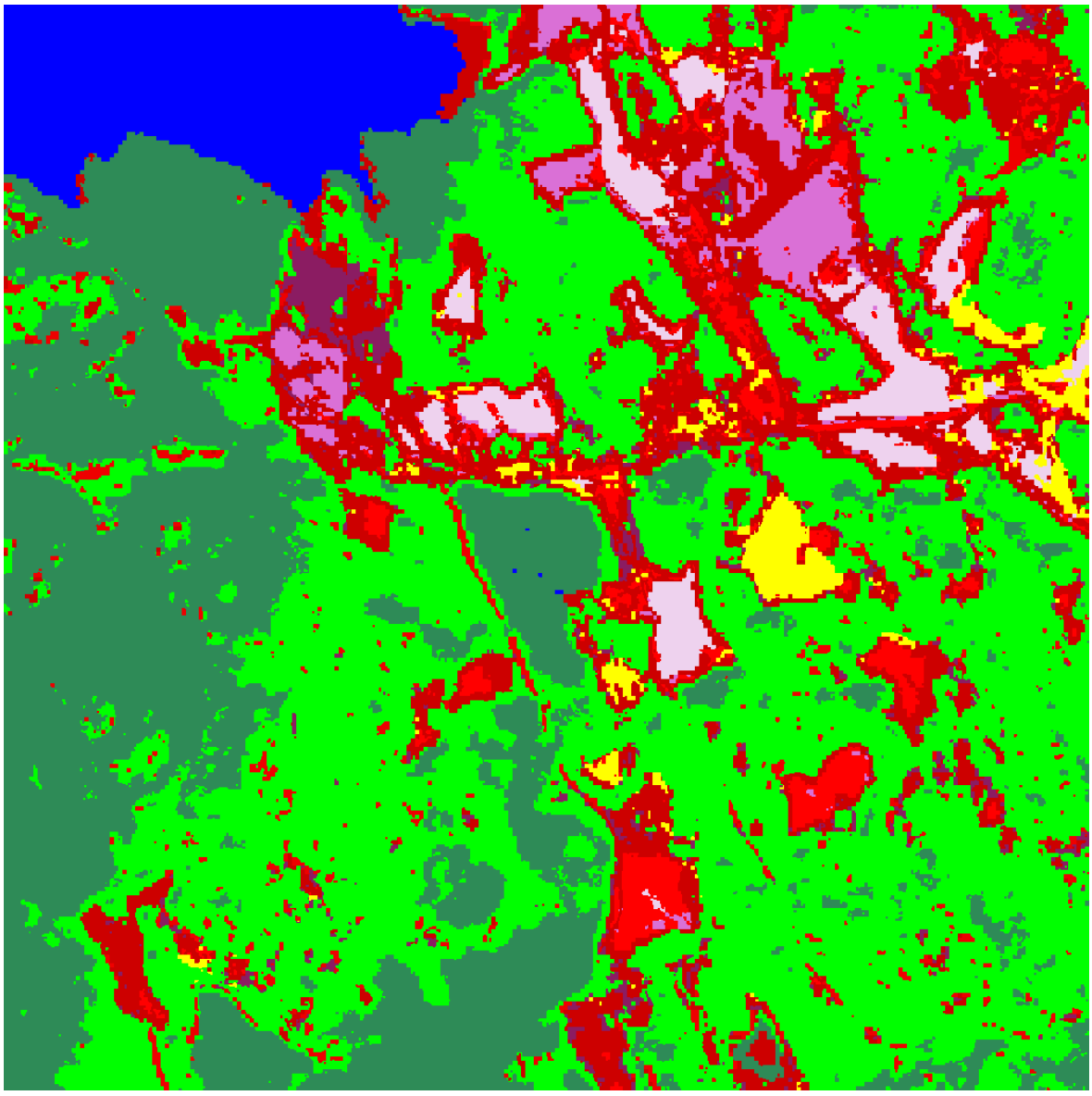


Figure 14: Result of single-scale ML parameter estimation and ICM classification, with fixed Potts parameters.

For this image we used ML estimation of the class parameters based on the training set only, rather than estimation within the ICM loop. The parameters of Potts model were fixed to 1.0 for the regularity parameter and 0.0 for the others. The result obtained with the single-scale classification scheme, applied to the panchromatic band and the XS bands resampled to the same resolution, is shown in Fig. 14. We see that some major roads are found, but they are smeared out. Roads that are partly hidden by tree crowns or shadows are generally not detected. In addition, many agricultural fields (possibly bare soil) are erroneously classified as roads.

With our new multi-scale classification scheme there is much less confusion with other classes, and finer structures such as roads are far better preserved, as can be seen

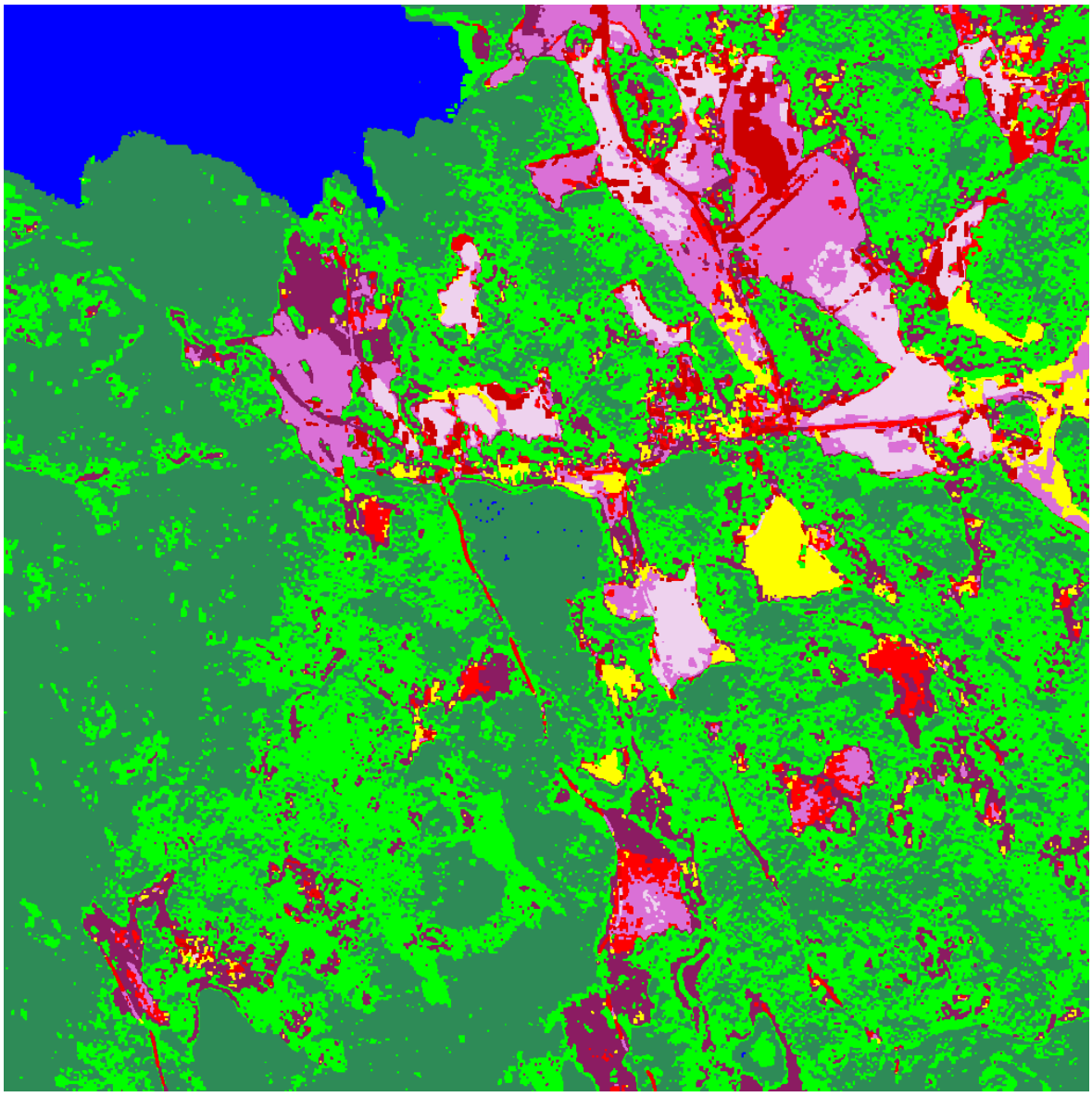


Figure 15: Result of multi-scale ML parameter estimation and ICM classification, with fixed Potts parameters.

in Fig. 15. However, road segments covered by trees or in shadow still pose problems, and some small roads visible by eye are not detected.

As we do not have reliable ground truth, no quantitative comparison has been carried out for this image. However, the general impression based on visual inspection is that the multi-scale approach gives a far better result than the single-scale approach for this SPOT 5 data set, and that the weaknesses observed for the data set composed of Landsat and SPOT images, described in section 5.2, are less pronounced here.

6 Conclusion

This report describes a Bayesian model for integration of multi-scale image data. The approach is based on the concept of a reference resolution. Data at this and lower resolutions are connected to the reference resolution through a fully specified statistical model. Algorithms for parameter estimation and classification based on the multi-scale model are proposed, and results and comparisons with single-scale classification are presented for simulated and real satellite images.

The results obtained on a simulated data set that fits the multi-scale model clearly demonstrate that the multi-scale scheme produce higher classification accuracy near fine structures and region boundaries, as compared to single-scale analysis where the images with coarser resolution have first been resampled to the reference solution.

A similar comparison was carried out on a real data set, containing a Landsat TM panchromatic band and the three spectral band of a SPOT XS image at the reference resolution, and 6 Landsat TM spectral bands with 2×2 times lower resolution. Even though some details are in favor of the multi-scale approach, the differences are here much smaller than for the simulated data set, most probably because the real data set does not fit the multi-scale model equally well. In particular, the assumption that the pixels at the reference levels are not mixed is quite far from reality for the SPOT XS image, as it has been co-registered with the panchromatic Landsat TM band and resampled from 20 to 15 m resolution. Moreover, the panchromatic and spectral bands of the Landsat TM image are not really independent. Thereto comes the general problem of intra-class variability for real images, which is here aggravated by the fact that the images are not acquired simultaneously (actually not even the same year).

For a SPOT 5 data set consisting of a panchromatic band with 2.5 m resolution and four XS spectral bands with 10 m resolution, the multi-scale classification scheme yields less confusion with other classes, and finer structures such as roads are far better preserved than with the single-scale method.

The proposed multi-scale estimation and classification method provides a better way of exploiting spectrally rich images at lower resolution together with images at the reference resolution, when the goal is to obtain accurate and spatially detailed classification result at the reference resolution. However, deviations from the underlying assumptions of the multi-scale model can reduce the advantages compared to single-scale analysis. Critical factors are:

- The co-registration of the images should have sub-pixel accuracy with respect to the pixel size at the reference resolution.

- The images at the reference resolution should ideally not be mixed pixels. The co-registration and resampling should therefore be done in a way that does not imply significant smoothing.

- The multi-scale scheme may, due to a higher degree of freedom, be more sensitive to intra-class variability and inter-class separability than the single-scale scheme.

Acknowledgments

We thank the Data Fusion Committee of the IEEE Geoscience and Remote Sensing Society for making the Landsat and SPOT images of Hasselt available. We also acknowledge the Norwegian Mapping Authority, for whom we carried out the analysis of the SPOT 5 image of Østfold.

A Maximization of the pseudo-likelihood for $\boldsymbol{\theta}$

The pseudo-likelihood function is given by

$$p_{pl}(\boldsymbol{\theta}_1) = \prod_i \frac{\exp\{\sum_k \alpha_k I(z_i = k) - \beta \sum_{j\in\partial i} \phi(z_i, z_j)\}}{\sum_l \exp\{\alpha_l - \beta \sum_{j\in\partial} \phi(l, z_j)\}}$$

and the pseudo-loglikelihood is

$$l_{pl}(\boldsymbol{\theta}_1) = \sum_i \Bigg[\sum_k \alpha_k I(z_i = k) - \beta \sum_{j\in\partial i} \phi(z_i, z_j) - \log(\sum_l \exp\{\alpha_l - \beta \sum_{j\in\partial i} \phi(l, z_j)\}) \Bigg]$$

Maximization of this expression directly is not possible, but can be performed by Newton-Raphson. Simple derivations show that

$$\frac{\partial}{\partial \alpha_k} l_{pl}(\boldsymbol{\theta}_1) = \sum_i I(z_i = k) - \sum_i \frac{A^i_{k,0}}{\sum_l A^i_{l,0}}$$
$$\frac{\partial}{\partial \beta} l_{pl}(\boldsymbol{\theta}_1) = -\sum_i \sum_{j\in\partial i} \phi(z_i, z_j) + \sum_i \frac{\sum_l A^i_{l,1}}{\sum_l A^i_{l,0}}$$

where

$$A^i_{l,v} = \exp\{\alpha_l - \beta \sum_{j\in\partial i} \phi(l, z_j)\} [\sum_{j\in\partial i} \phi(l, z_j)]^v.$$

Similarly, we get

$$\frac{\partial^2}{\partial \alpha_k^2} l_{pl}(\boldsymbol{\theta}_1) = \sum_i \frac{A^i_{k,0} A^i_{k,0} - A^i_{k,0} \sum_l A^i_{l,0}}{[\sum_l A^i_{l,0}]^2}$$
$$\frac{\partial^2}{\partial \alpha_k \partial \alpha_{k'}} l_{pl}(\boldsymbol{\theta}_1) = \sum_i \frac{A^i_{k,0} A^i_{k',0}}{[\sum_l A^i_{l,0}]^2}, \qquad k' \neq k$$
$$\frac{\partial^2}{\partial \alpha_k \partial \beta} l_{pl}(\boldsymbol{\theta}_1) = \sum_i \frac{A^i_{k,1} \sum_l A^i_{l,0} - A^i_{k,0} \sum_l A^i_{l,1}}{[\sum_l A^i_{l,0}]^2}$$
$$\frac{\partial^2}{\partial \beta^2} l_{pl}(\boldsymbol{\theta}_1) = \sum_i \frac{[\sum_l A^i_{l,1}][\sum_l A_{l,1}] - [\sum_l A^i_{l,2}][\sum_l A_{l,0}]}{[\sum_l A^i_{l,0}]^2}$$

Remark Similar to the estimation of $\boldsymbol{\theta}_2$ only a few iterations should be necessary. Note that the Hessian matrix $\mathbf{H}(\boldsymbol{\theta}_1)$ defined by

$$\mathbf{H}(\boldsymbol{\theta}_1) = \begin{pmatrix} \frac{\partial^2}{\partial\alpha_1\partial\alpha_1} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_1\partial\alpha_2} l_{pl}(\boldsymbol{\theta}_1) & \cdots & \frac{\partial^2}{\partial\alpha_2\partial\alpha_K} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_2\partial\beta} l_{pl}(\boldsymbol{\theta}_1) \\ \frac{\partial^2}{\partial\alpha_2\partial\alpha_1} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_2\partial\alpha_2} l_{pl}(\boldsymbol{\theta}_1) & \cdots & \frac{\partial^2}{\partial\alpha_2\partial\alpha_K} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_2\partial\beta} l_{pl}(\boldsymbol{\theta}_1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial^2}{\partial\alpha_K\partial\alpha_1} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_K\partial\alpha_2} l_{pl}(\boldsymbol{\theta}_1) & \cdots & \frac{\partial^2}{\partial\alpha_K\partial\alpha_K} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_K\partial\beta} l_{pl}(\boldsymbol{\theta}_1) \\ \frac{\partial^2}{\partial\alpha_1\partial\beta} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\alpha_2\partial\beta} l_{pl}(\boldsymbol{\theta}_1) & \cdots & \frac{\partial^2}{\partial\alpha_K\partial\beta} l_{pl}(\boldsymbol{\theta}_1) & \frac{\partial^2}{\partial\beta\partial\beta} l_{pl}(\boldsymbol{\theta}_1) \end{pmatrix}$$

is always negative definite (not true??), simplifying the maximization. Further, the Newton-Raphson algorithm is given by

$$\boldsymbol{\theta}_1^{r+1} = \boldsymbol{\theta}_1^r - \mathbf{H}(\boldsymbol{\theta}_1^r)^{-1}\mathbf{s}(\boldsymbol{\theta}_1^r)$$

where

$$\mathbf{s}(\boldsymbol{\theta}_1) = \begin{pmatrix} \frac{\partial}{\partial\alpha_1} l_{pl}(\boldsymbol{\theta}_1) \\ \frac{\partial}{\partial\alpha_2} l_{pl}(\boldsymbol{\theta}_1) \\ \vdots \\ \frac{\partial}{\partial\alpha_K} l_{pl}(\boldsymbol{\theta}_1) \\ \frac{\partial}{\partial\beta} l_{pl}(\boldsymbol{\theta}_1) \end{pmatrix}$$

B Distribution of $\tilde{\mathbf{y}}$

We will consider the expectation and covariance matrix of $\{\tilde{\mathbf{y}}_i^j\}$, conditioned on that

$$\frac{1}{m_j} \sum_{i' \in s_j(v_j(i))} \tilde{\mathbf{y}}_{i'}^j = \mathbf{y}_{v_j(i)}^j$$

Since $\{\tilde{\mathbf{y}}_i^j\}$ follows a Gaussian distribution, also the conditional distribution will be Gaussian. In the following we will use $v = v_j(i)$ to simplify the notation. Define $\tilde{\mathbf{y}}_v^j$ to be the vector of $\{\tilde{\mathbf{y}}_i^j, i \in s_j(v)\}$. Then

$$\mathsf{E}\left[\begin{array}{c} \tilde{\mathbf{y}}_v^j \\ \mathbf{y}_v^j \end{array} \middle| \mathbf{z} \right] = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \overline{\boldsymbol{\mu}}_2 \end{pmatrix}$$

where $\boldsymbol{\mu}_1$ is the vector of $\{\boldsymbol{\mu}_{z_i}^j, i \in s_j(v)\}$ while

$$\overline{\boldsymbol{\mu}}_2 = \frac{1}{m_j} \sum_{i \in s_j(v)} \boldsymbol{\mu}_{z_i}^j.$$

Further

$$\mathsf{Var}\left[\begin{array}{c} \tilde{\mathbf{y}}_i^j \\ \mathbf{y}_i^j \end{array} | \mathbf{z} \right] = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \overline{\boldsymbol{\Sigma}}_{22} \end{pmatrix}$$

where $\mathbf{\Sigma}_{11}$ is a block diagonal matrix containing $\{\mathbf{\Sigma}^j_{z_i}, i \in s_j(v)\}$ as block diagonals, $\mathbf{\Sigma}_{12}$ is a column vector of matrices with elements $\{\frac{1}{m_j}\mathbf{\Sigma}^j_{z_i}, i \in s_j(v)\}$ and $\overline{\mathbf{\Sigma}}_{22} = \frac{1}{m_j^2}\sum_{i\in s_j(v)}\mathbf{\Sigma}^j_{z_i}$. Then [Mardia et al., 1979, Page 63]

$$\begin{aligned}\boldsymbol{\eta}^j_v &= \mathsf{E}[\tilde{\mathbf{y}}^j_i|\mathbf{y}^j_v, \mathbf{z}] = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\overline{\mathbf{\Sigma}}^{-1}_{22}(\mathbf{y}^j_v - \overline{\boldsymbol{\mu}}_2)\\ \mathbf{S}^j_v &= \mathsf{Var}[\tilde{\mathbf{y}}^j_i|\mathbf{y}^j_v, \mathbf{z}] = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\overline{\mathbf{\Sigma}}^{-1}_{22}\mathbf{\Sigma}_{21}.\end{aligned}$$

From this we get that $\tilde{\mathbf{y}}^j_i|\mathbf{y} \sim N(\boldsymbol{\eta}^j_i, \mathbf{S}^j_i)$ where $\boldsymbol{\eta}^j_i$ is the (vector) component of $\boldsymbol{\eta}^j_v$ corresponding to pixel i while $\mathbf{S}^j_i$ is the block diagonal of $\mathbf{S}^j_v$ corresponding to pixel i.

The first time these quantities need to be calculated, some initial values for $\boldsymbol{\mu}^j_k$ and $\mathbf{\Sigma}^j_k$. A possible choice is

$$\boldsymbol{\mu}^j_k = \mathbf{0}, \quad \mathbf{\Sigma}^j_k = \boldsymbol{I}$$

This results in that

$$\boldsymbol{\eta}^j_i = \mathbf{y}^j_v, \quad \mathbf{S}^j_i = \mathbf{0}$$

in the first iteration of the inner EM algorithm

References

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48(3):259–302, 1986.

J. Besag. Towards Bayesian image analysis. *Journal of Applied Statistics*, 16(3):395–407, 1989.

M. M. Crawford, S. Kumar, M. R. Ricard, J. C. Gibeaut, and A. Neuenshwander. Fusion of airborne polarimetric and interferometric SAR for classification of coastal enviromnents. *IEEE Trans. Geoscience and Remote Sensing*, 37(3):1306–1315, 1999.

M. M. Daniel and A. S. Willsky. A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing. *Proceedings of the IEEE*, 85(1):164–180, 1997.

R. C. Dubes and A. K. Jain. Random field models in image analysis. *Journal of Applied Statistics*, 16(2):131–164, 1989.

R. Fjørtoft, Y. Delignon, W. Pieczynski, M. Sigelle, and F. Tupin. Unsupervised classification of radar images using hidden Markov chains and hidden Markov random fields. *IEEE Trans. Geoscience and Remote Sensing*, 41(3), 2003.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(6): 721–741, 1984.

D. Hirst, G. Storvik, and A. R. Syversveen. A hierarchical modelling approach to combining environmental data at different scales. *jrssc*, 2003. To be published.

M. R. Luettgen, W. Clem Karl, and A. S. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Trans. Image Processing*, 3(1):41–63, 1994.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

J. Núñez, X. Otazu, O. Fors, A. Prades, V. Palà, and A. Román. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Trans. Geoscience and Remote Sensing*, 37(3):1204–1210, 1999.

J. C. Price. Combining multispectral data of differing spatial resolution. *IEEE Trans. Geoscience and Remote Sensing*, 37(3):1199–1203, 1999.

P. Puyou-Lascassies, A. Podaire, and M. Gay. Extracting crop radiometric responses from simulated low and high spatial resolution satellite data using a linear mixing model. *int. J. Remote Sensing*, 15(18):3767–3784, 1994.

A. H. S. Solberg. Contextual data fusion applied to forest map revision. *IEEE Trans. Geoscience and Remote Sensing*, 37:1234–1243, 1999.

G. Storvik and G. Dahl. Lagrangian based methods for finding MAP solutions for MRF models. *IEEE Trans. Image Processing*, 9(3):469–479, 2000.

B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhäckel. Unmixing-based multisensor multiresolution image fusion. *IEEE Trans. Geoscience and Remote Sensing*, 37(3):1212–1226, 1999.