Integration of multisource image data at different resolutions and time points. A mathematical framework for EOtools

SAMBA/25/2002

Geir Storvik
Roger Fjørtoft
Anne Solberg

September 4, 2002

Integration of multisource image data at different resolutions and time points. A mathematical framework for EOtools

Geir Storvik
Roger Fjørtoft
Anne Solberg

August 2002

Abstract

In this note we describe a full Bayesian model for integration multisource image data which can be collected either at different resolutions or at different time points. The main approach is based on specifying a prior distribution for the class-configuration at a fixed reference resolution. Data at other resolutions is then connected to the reference resolution through a full specified statistical model.

Based on the model, algorithms both for segmentation and parameter estimation are proposed. Several alternatives are possible, and choices between these will have to be made based on future experiments.

***Keywords*—Image analysis, multisource, multiresolution, multitemporal, statistical modelling**

1 Introduction

Earth observation is currently developing more rapidly than ever before. During most of the last three decades satellite-based remote sensing has typically been accomplished by a few satellites with a single sensor on each satellite and low temporal coverage. The last few years and the near future show a big difference. Satellite platforms with a large number of sensors are coming (Terra and ENVISAT), and many satellites with one or a few sensors emerge. The coverage of the Earth in space, time and electromagnetic spectrum is increasing correspondingly fast. This development opens for a significant change in the approach of analysis of earth observation data. Traditionally, analysis of such data has been performed by processing single satellite images separately. Higher coverage in space, time and spectrum opens for analysis of time series of data, combination of images from different sensor types and with different resolution, and better integration with ancillary data and models. Classical methodology, like Maximum Likelihood or Maximum A Posteriori (MAP) classification, will in general not be able to extract all the interesting information from this kind of data sets.

The EOtools project will develop a common mathematical framework for multi-data analysis. The EOtools project is divided into six Work Packages: WP 1: Mathematical foundation; WP 2: Multi-sensor analysis; WP 3: Hyperspectral analysis; WP 4: Multi-scale analysis; WP 5: Multi-temporal analysis; and WP 6: Validation. This report describes the basis of the mathematical framework as developed in WP 1. The framework will be tested and possibly further developed in WPs 2-6. Thus, this report includes a presentation and discussion of the mathematical modelling, but no validation results.

The mathematical framework should be general, and it must include the following elements:

- **Multi-sensor data**
- **Multi-temporal data**
- **Multi-resolution data**

The model should also work for hyperspectral data (> 50 spectral bands). In principle, hyperspectral data can be treated just like multispectral data using multivariate probability distributions, but high-dimensional spaces pose certain problems. Robust parameter estimation or dimensionality reduction must be applied. A special work package (WP 3) is devoted to analysis of hyperspectral data, which will not be further discussed in this report.

2 Main approach

Based on remotely sensed image data (and other available information), we want to predict or estimate a physical parameter or process. The elements of the process can either be a categorical (e.g. one of a given set of ground-cover classes), or a continuous (e.g. the estimated biomass of a forest area).

Our main approach will be based on Bayesian hierarchical modelling. The unknown "physical" process $\mathbf{z} = \{\mathbf{z}(\mathbf{u}, t)\}$ will be modelled through a prior model $p(\mathbf{z})$, while data $\mathbf{y}$ will be modelled conditionally on $\mathbf{z}$ through a distribution function $p(\mathbf{y}|\mathbf{z})$. Inference on $\mathbf{z}$ given the data is based on the posterior distribution

$$p(\mathbf{z}|\mathbf{y}) \propto p(\mathbf{z})p(\mathbf{y}|\mathbf{z}) \tag{1}$$

When modelling the physical process $\mathbf{z}$, we have the choice between regarding $\mathbf{z}$ as a process defined continuously in space through geostatistical modelling (as in the FOREMMS project, cf. Høst et al. (2001)) or to discretise the space to a lattice of pixels characterised by class memberships.

The Markov random field (MRF) approach is more flexible than the "geostatistical approach" in the sense that it is easier to include constraints on what kind of classes that can be neighbours in space and time. One should perhaps also take into account that MRF models are more commonly recognised in the image analysis literature.

For these reasons, we have decided to go for MRF models in this project. In order to take multi-resolution data into account, the MRF model is specified on a *reference resolution*, which is the resolution of the desired output. Data at all resolutions are then modelled related to the class structure at this reference resolution.

In section 3, the model for the underlying structure is described. We discuss how to combine this with the different types of data in section 4. Segmentation is treated in section 5, while parameter estimation is considered in section 6.

3 An MRF framework for the Z process

This section describes prior distributions for the ground truth. We will assume that the ground truth can be described by characteristics of this process inside each pixel at a user-defined reference resolution. Such characteristics can consist of categorical information such as class-membership, but also continuous variables describing the characteristics in more detail (i.e. biomass in forest areas). In this section we will only consider class-membership. Extension of this model to include other characteristics will be discussed in section 7.

The choice of the reference resolution will depend on applications and data available. In most cases, however, the resolution will be too coarse to assume that each pixel only consists of one class. Our analysis will therefore be based on a mixed pixel approach. Define $z_k(i,t)$ to be the proportion of class k in pixel i at time t, Then $z_k(i,t) \geq 0$ and $\sum_{k=1}^{K} z_k(i,t) = 1$, where K is the number of possible classes. Note that the assumption that a pixel belongs to one single class can be incorporated by adding the constraint $z_k(i,t) \in \{0,1\}$.

Much work has been done on specifying models for mixed pixels on a pixel-by-pixel basis (see Pedrycz (1990) and the references therein). In Salzenstein and Pieczynski (1997) and Caillol et al. (1993) the possibility of both hard (i.e. no mixing) and soft (mixed) pixels are considered.

Kent and Mardia (1988) introduced a spatial model for mixed pixels. Their model does however allow ratios to be outside the interval $[0,1]$, correcting for this by some ad hoc procedure. Caillol et al. (1997) incorporated contextual information by local adjustment of a non-contextual segmentation result. More formal models for spatial dependence connected to mixed pixels are given in Choi et al. (1991) and Salzenstein and Pieczynski (1997), the latter allowing for a mix of hard and soft pixels.

3.1 Multi-resolution aspects

In the literature, various approaches to multi-resolution Markov field models are presented. Krishnamachari and Chellappa (1997) introduced a multi-resolution Gauss-MRF model for efficient segmentation of a single resolution image. In their model, coarser resolution sample fields are obtained by sub-sampling the fine resolution field. Under these conditions, the coarse resolution images are non-Markov, but they prove that a Gauss-MRF approximation of the probability density function can be found by minimisation of the Kullback-Leibler distance.

A common model for multi-resolution MRF models is the quadtree model (Li et al. 2000). In this model, the hidden class process is modelled on different grids from fine to coarse resolution. Class labels are defined at each resolution, but the simultaneous model of neighbouring pixels will depend on the way parent nodes are defined. With a rigid block structure, child blocks descended from different parent blocks are treated conditionally independent. By modelling the class process at all resolutions, not all pixels will consist of pure classes (at higher resolutions), and a mixture class must be introduces. The coarse-resolution fields will not be strictly Markovian as mentioned above, and Markov approximations of the probability densities must be found. These issues constitute a major drawback of quadtree MRF models. The major advantage of quadtree models is the potential for fast image classification by doing the optimisation at coarse resolution and then iterating to finer resolutions. Lakshmanan and Derin (1993) presents a Gaussian MRF model for multi-resolution data for fast segmentation of single-resolution image data. They give a thorough discussion of the fact that the coarse resolution sampled data are not strictly Markovian, and propose Markov approximations of the probability density functions.

Another approach is presented in Comer and Delp (1999). They model the data, which were originally observed at a single resolution, using a multi-resolution Gaussian autoregressive model, and the multi-resolution class process using a 3D neighbourhood structure which is different from the traditional quadtree approach. The neighbourhood system is defined across resolutions so that a pixel has neighbours on the same resolution (siblings), at the coarser resolution (parent) and at a finer resolution (children). They introduced a multi-scale version of the MPM (Maximum Posterior Marginals) algorithm for segmentation purposes.

Our approach is based on the assumption that *data* will be available at different resolutions, but that segmentation is only to be performed at one resolution. We assume therefore that the classes are modelled only at this resolution, the *reference resolution*. When choosing this resolution, the following aspects should be taking into account:

- The assumption of a MRF is only valid at the chosen resolution.
- Computation complexity and need for high resolution segmentation.

3.2 Spatial modelling

In this section we will introduce a spatial model for the distribution of the mixed pixels. We will follow the approach by Salzenstein and Pieczynski (1997), allowing for both "hard" and "soft" pixels. They specify the distribution for the mixels through a density with respect to the measure ν^N where N is the number of pixels while

$$\nu = \delta + \mu.$$

Here δ is the Dirac measure on $H = \{\mathbf{z}_i; \max_k z_{i,k} = 1\}$ while μ is the Lebesgue measure on the simplex $S = \{\mathbf{z}_i; z_{i,k} \in (0,1), \sum_k z_{i,k} = 1, \max_k z_{i,k} < 1\}$. (Note that Salzenstein and Pieczynski (1997) only considered the case of two classes.) We will say that $\mathbf{z}_i$ is *hard* if $\mathbf{z}_i \in H$ while it is *soft* if $\mathbf{z}_i \in S$. The density for $\mathbf{z}$ is given by

$$\pi(\mathbf{z}) \propto \exp\{-U(\mathbf{z})\}$$

where $U(\mathbf{z})$ is then energy function, specified by

$$U(\mathbf{z}) = \sum_C \phi_C(\mathbf{z}_C)$$

where the sum is over all cliques. We will only consider second order neighbourhoods corresponding to that C is either equal to $\{i\}$ or to $\{i,j\}$ where $|i-j| = 1$. Define

$$\phi_i(\mathbf{z}_i) = \begin{cases} \alpha_k & \text{if } z_{i,k} = 1 \\ g(\mathbf{z}_i) & \text{if } \max_k z_{i,k} < 1 \end{cases}$$

Both Salzenstein and Pieczynski (1997) and Choi et al. (1991) assumed g uniform on S, but other choices, emphasising that most mixels should contain mainly one or two classes should be considered. Some possibilities are

$$g(\mathbf{z}_i) = \begin{cases} \sum_{k=1}^K \alpha_k \log(z_{i,k}), & \text{or} \\ \sum_{k=1}^K \alpha_k z_{i,k}, & \text{or} \\ \sum_{k=1}^K \alpha_k z_{i,k}^2. \end{cases}$$

With no spatial dependence, the first choice corresponds to a Dirichlet distribution for $\mathbf{z}_i = (z_{i,1}, ..., z_{i,K})$ which is a commonly used model in statistics for proportions. The two other choices are closer to MRFs that are typically used when pixels only are allowed to contain one class.

Assume all α_k's have the same value corresponding to all classes being equally probable. The first choice then gives highest probability (or density value, to be more specific) to configurations where all classes have equal proportion, the second choice give equal probability to all possible configurations, while the last one give highest probability to configurations having proportions close to zero or one. The last situation is perhaps most realistic for ground cover, but simulations from all models should be performed in order to decide which potential to use.

Turning to cliques of neighbour pixels, $\phi_{i,j}(\mathbf{z}_i, \mathbf{z}_j)$ defines the spatial structure. Assume

$$\phi_{i,j}(\mathbf{z}_i, \mathbf{z}_j) = \exp\{\sum_k \beta_k ||z_{i,k} - z_{j,k}||^p\}$$

$p = 2$ was used in (Choi et al. 1991), while Salzenstein and Pieczynski (1997) used $p = 1$. Neither of these choices take into account that some combinations of classes can be more probable than others. Modifications of the potentials are possible in order to incorporate such structures.

For any of the choices of potentials, the model corresponds to

$$\begin{aligned}\mathbf{Pr}(\mathbf{z}_i|\mathbf{z}_j; j \neq i) &= \mathbf{Pr}(\mathbf{z}_i|\mathbf{z}_j; |j-i| = 1) \\ &\propto \exp\{\phi_i(\mathbf{z}_i) - \sum_{|j-i|=1} \phi_{i,j}(\mathbf{z}_i, \mathbf{z}_j)\} \end{aligned} \quad (2)$$

i.e. given the classes of the nearest neighbours of pixel i, the classes of other pixels do not give any extra information. Such models can be considered as local smooth models, since only local information is incorporated. Such models have been widely used in image processing. From the Hammersley-Clifford theorem, these conditional probabilities completely specify the distribution of $\mathbf{z}$. The marginal probabilities are however not tractable.

3.3 Temporal modelling

The $\mathbf{z}$ process may change with time. In principle we could extend the Markov field model with another dimension. However, because much information about how classes change over time is available, a more beneficial approach would be to directly model the time dynamics. In particular we want to model

$$\mathbf{Pr}(\mathbf{z}^t|\mathbf{z}^{t-1}, \mathbf{z}^{t-2}, ...).$$

Furthermore, most pixels will not change state from one time-point to another.

A simplifying assumption is

$$\mathbf{Pr}(\mathbf{z}^t|\mathbf{z}^{t-1}, \mathbf{z}^{t-2}, ...) = \mathbf{Pr}(\mathbf{z}^t|\mathbf{z}^{t-1}),$$

corresponding to assuming that $\{\mathbf{z}^t\}$ is a first order Markov chain in time. Assuming the temporal scale is not too fine, such a model would be reasonable. In this case we will have to distinguish between several possibilities:

- No change is made from time $t-1$ to time t. A Dirac measure for this possibility should be made.
- A change when $\mathbf{z}_i^{t-1}$ is hard. In that case, both a change to a hard and a soft $\mathbf{z}_i^t$ should be considered, giving a mixture of a Dirac and Lebesgue measure.
- A change when $\mathbf{z}_i^{t-1}$ is soft. In this case, only a Lebesgue measure should be sufficient

Specification of the distribution can again be performed through potentials. Spatial smoothness of changes should however also be taken into account. We will consider transition densities of the form

$$\mathbf{Pr}(\mathbf{z}^t|\mathbf{z}^{t-1}) \propto \exp\{\sum_i \gamma_i(\mathbf{z}_i^t; \mathbf{z}_i^{t-1}) + \sum_{|i-j|=1} \gamma_{i,j}(\mathbf{z}_i^t, \mathbf{z}_j^t; \mathbf{z}_i^{t-1}, \mathbf{z}_j^{t-1})\} \quad (3)$$

where γ_i describes changes in one pixel while $\gamma_{i,j}$ describes the spatial structure of changes. To our knowledge, specification of such transition probabilities in a mixed pixel setting has never been considered. Much of the same ideas as for the specification of spatial structure can however be applied. Specific choices will be made at a later stage, using experience from simulation studies.

4 Models for data

Data will be modelled conditionally on the ground truth $\mathbf{z}$. We will first consider observations at the reference resolution (section 4.1, while observations at other resolutions will be considered in 4.2. Multi-sensor and multi-temporal data will be considered in sections 4.3 and 4.4, respectively. Specific distributions to be used is considered in 4.5.

4.1 Observations at the reference resolution

Assume $\mathbf{y}$ is a (possibly multi-spectral) image observed at time t. The resolution of $\mathbf{y}$ is assumed to be equal to the reference resolution. In that case, we make the standard assumption that observations from neighbouring pixels are conditionally independent, i.e.

$$p(\mathbf{y}|\mathbf{z}) = \prod_i f(\mathbf{y}_i|\mathbf{z}_i). \tag{4}$$

An extension of this simple model could be to assume conditional spatial dependence between the $\mathbf{y}$'s, but this will not be considered here.

4.2 Multi-resolution data

Assume now that the image $\tilde{\mathbf{y}}$ is observed with $\tilde{\mathbf{y}}_i$ being observed over an area covering pixels $i_1, ..., i_m$ on the reference resolution. Our basic assumption is that

$$\tilde{\mathbf{y}}_i = \frac{1}{m_j} \sum_v \mathbf{y}_{i_v} \tag{5}$$

where $\mathbf{y}_{i_v}$ is the observation that would have been obtained if observation at the reference resolution was available. This is how optical sensors work, and is also reasonable for radar images in single look complex (SLC) mode. Modelling of $\tilde{\mathbf{y}_i}$ goes through modelling of $\{\mathbf{y}_{i_v}, v = 1, ..., m\}$. In particular, we assume that the $\mathbf{y}_{i_v}$'s follow the basic conditional independence assumption (4)**. The full conditional distribution of $\mathbf{y}_i$ is then given by**

$$f(\tilde{\mathbf{y}_i}|\mathbf{z}) = \int_{m^{-1}\sum_v \mathbf{y}_{i_v} = \tilde{\mathbf{y}_i}} \prod_{i=1}^{m} f_{\mathbf{z}_{i_v}}(\mathbf{y}_{i_v}) d\mathbf{y}_{i_v} \tag{6}$$

In general this will be a difficult distribution to evaluate. By including $\mathbf{y}_{i_v}$ as missing variables, however, efficient segmentation and estimation procedures can still be constructed (see sections 5 and 6). If all $\mathbf{y}_{i_v}$'s are Gaussian, so will $\mathbf{y}_i$ be, and in that case the distribution (6) **can be directly calculated.**

4.3 Multi-sensor images

In the general data fusion literature, information can be combined at three different levels (see, e.g. Abidi and Gonzalez (1992)): data level fusion, feature level fusion, and decision level fusion. Since our focus is on the Z process, we only consider decision-level fusion here.

Assume we have data $\mathbf{y}^1, ..., \mathbf{y}^p$ obtained from different sensors. The different data sources are all assumed to be "measurements" of an "underlying truth". A possible model is then to assume that $\mathbf{y}^1, ..., \mathbf{y}^p$ are conditionally independent given $\mathbf{z}$, i.e.

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

4.4 Multi-source and multi-temporal data

Our model specification on data has so far only been related to one time point. Extensions to multi-temporal data is done similarly as for multi-sensor data. Assume now $\mathbf{y}^{t,j}$ contains all data available at time t from source j. Our assumption is then that

$$p(\mathbf{y}|\mathbf{z}) = \prod_{j=1}^{p} \prod_{s=1}^{q} p(\mathbf{y}^{t_s,j}|\mathbf{z}^{t_s})$$

i.e. we assume both conditional independence and that observations at time t_s are only influenced by $\mathbf{z}$ through $\mathbf{z}^{t_s}$.

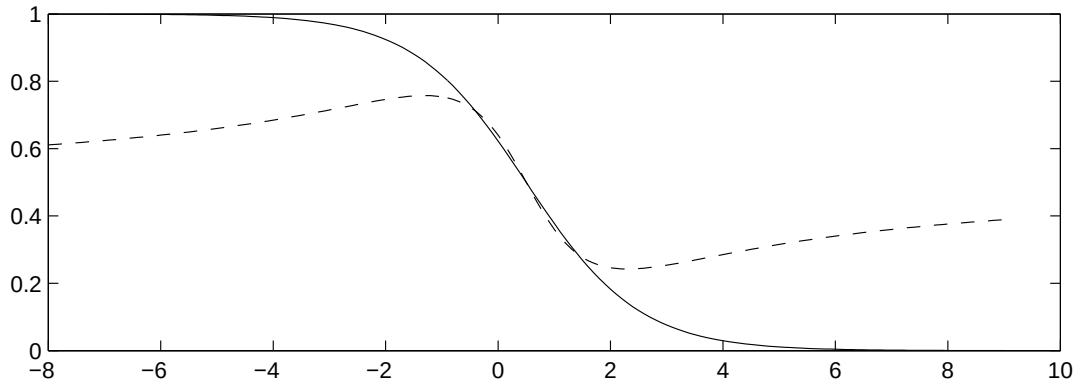

Figure 1: $\mathbf{Pr}(z_{i,1} = 1|y_i)$ *as a function of* y_i *for Gaussian distribution (solid line) and a T-distribution with 3 degrees of freedom (dashed line). In both cases* $E[y_i|z_{i,1} = 1] = 0$, $E[y_i|z_{i,2} = 1] = 1$, *while* $\mathsf{var}[y_1|z_{i,k} = 1] = 1$ *for* $k = 1, 2$.

4.5 Distributions for data

Gaussian distributions are relatively easy to handle mathematically and computationally and are the most commonly used model for compound remote sensing data. Such models are reasonable for optical imagery if the area of which the pixel cover is homogeneous. If the area is heterogeneous, which is the typical situation, more heavy-tailed distributions (e.g. T-distributions) might be preferable. When used in segmentation, heavy-tailed distributions also give less extreme class probabilities, which is an undesirable property of the multivariate Gaussian distribution. This is illustrated in Figure 1 where $\mathrm{Pr}(z_{i,1} = 1|y_i)$ is plotted with two possible classes, $E[y_i|z_{i,1} = 1] = 0$, $E[y_i|z_{i,2} = 1] = 1$, and $\mathsf{var}[y_1|z_{i,k} = 1] = 1$ for both classes. Furthermore, only hard pixels are considered and the prior probability for each class is 0.5. The solid line is the posterior probability as a function of y_i if a Gaussian density is used, while the dashed line is for a T-distribution. We see that although the probabilities are similar in the range $[0 - 0.4, 1.5]$, they behave quite differently outside this interval. For the Gaussian distribution, more extreme values of y_i give more extreme class probabilities, while for the T-distribution these probabilities converge towards 0.5.

Statistical modelling of radar images is more complicated due to *speckle*, which can be modelled as a strong, multiplicative, noise-like effect. There are several possible settings:

1. **The complex amplitudes of SLC radar images are well described by circular complex Gaussian distributions (Novak and Burl 1990). The full resolution SLC images carry a maximum of information. Moreover, the dependencies between pixels and images are easy to incorporate in the Gaussian case. We here suppose that the speckle is fully developed (Goodman 1984), which is generally the case for natural surfaces.**
2. **The real pixel values of single look intensity images are exponentially distributed. Assuming independent looks, the pixels of multi-look intensity images follow a Gamma distribution (Goodman 1984). In this case, it is extremely difficult to take spatial and inter-image correlations into account. In fact, the Gamma distribution appears because intensity images are transformations of SLC images.**
3. **When the equivalent number of independent looks is 5–10 or higher, the intensity distributions become approximately Gaussian, which facilitates the joint exploitation of several image channels (e.g. several optical bands combined with radar images). It should, however, be noted that a higher number of looks (lower speckle variance) is obtained through a smoothing operation that inevitably reduces the spatial resolution.**
4. **We can also work on the logarithm of the intensity, which is Fisher-Tippett distributed (Arsenault and April 1976) in the single look case and which tends faster towards Gaussianity than the intensity when the number of looks increases. The logarithmic transformation also makes the speckle additive. This format is clearly sub-optimal for estimation of the mean radar reflectivity (Zrnic 1975, Fjørtoft and Lopès 2001), but it is well suited for texture estimation**

(Oliver and Quegan 1998).

Due to the speckle, the values of single pixels in radar images are not very meaningful, so we generally have to consider several pixels, assumed to belong to the same class, in order to interpret the data. Classification and estimation methods for radar images must therefore in general be contextual. The distributions cited above are valid both for individual pixels and for areas (classes) with constant radar reflectivity (assuming calibrated data), i.e. areas where the fluctuations are due to speckle only.

If the radar reflectivity has texture, there will be an extra variability in the pixel values. In the case of SLC images this extra variability can be included by assuming that the variance in the complex Gaussian distribution is stochastic. If the variance is inverse chi-squared, the result will be a T-distribution. For transformed images, other distributions than Gaussian or T-distributions may be preferable. For intensity images, for example, the radar reflectivity of a class is often supposed to be Gamma distributed, which together with Gamma distributed speckle yields K-distributed intensities. In such cases, it is very difficult to relate observations on one resolution level to observations on other resolution levels. A possible solution is to make this relation through the original SLC data.

4.6 Relating observations to mixed pixels

As described in section 3.2, a pixel at the reference resolution may contain several classes. The observations have to related to this in some way. Our basic assumption is to assume that an observation from a pixel i can be considered as an average of "observations" at point resolution, i.e. (assuming the pixel has size 1)

$$\mathbf{y}_i = \int_{\mathbf{x}\in i} \mathbf{y}(\mathbf{x}) d\mathbf{x}. \tag{7}$$

Now define $z(\mathbf{x})$ to be the class at point $\mathbf{x}$ and perform the following rewriting:

$$\begin{aligned}\mathbf{y}_i &= \int_{\mathbf{x}\in i} \mathbf{y}(\mathbf{x}) \sum_k I(z(\mathbf{x}) = k) d\mathbf{x} \\ &= \sum_k \int_{\mathbf{x}\in i} \mathbf{y}(\mathbf{x}) I(z(\mathbf{x}) = k) d\mathbf{x} \\ &= \sum_k z_{i,k}\mathbf{y}_{i,k}\end{aligned} \tag{8}$$

where

$$\begin{aligned}z_{i,k} &= \int_{\mathbf{x}\in i} I(z(\mathbf{x}) = k) d\mathbf{x} \\ \mathbf{y}_{i,k} &= \frac{\int_{\mathbf{x}\in i} \mathbf{y}(\mathbf{x}) I(z(\mathbf{x}) = k) d\mathbf{x}}{\int_{\mathbf{x}\in i} I(z(\mathbf{x}) = k) d\mathbf{x}}\end{aligned}$$

i.e. $z_{i,k}$ is the proportion of class k in pixel i (equal to the previous definition used in section 3.2) while $\mathbf{y}_{i,k}$ is the part of $\mathbf{y}_i$ coming from the proportion of class k in pixel i. Note that we here have made the assumption that only one class is present at each point. The relation (8) can however be made valid also in the more general case with a modified interpretation of $z_{i,k}$ and $\mathbf{y}_{i,k}$.

The distribution of $\mathbf{y}_i$ can now be defined through specifications of the distributions for $\mathbf{y}_{i,k}$. In general the distribution of $\mathbf{y}_i$ can be complicated, but for Gaussian distributions, simplifications are possible. Assume

$$\mathbf{y}_{i,k} \sim N(\mu_k, z_{i,k}^{-1}\Sigma_k).$$

The factor $z_{i,k}^{-1}$ makes the definition consistent with the assumption (7). Assume further that all $\mathbf{y}_{i,k}, k = 1, ..., K$ are independent (given $z_{i,k}, k = 1, ..., k$. Then

$$\mathbf{y}_i \sim N(\sum_k z_{i,k}\mu_k, \sum_k z_{i,k}\Sigma_k).$$

These distributions are simple to calculate for given $z_{i,k}, k = 1, ..., K$.

Remark: Note that the assumption of conditional independence between the $\mathbf{y}_{i,k}$'s is somewhat weaker than the assumption that all $\mathbf{y}(\mathbf{x})$ are conditionally independent (which would lead to the same result). In reality $\mathbf{y}(\mathbf{x})$'s from close sites would be positively correlation. If the correlation range is small compared to the size of the (largest ones of the) $z_{i,k}$'s, neglecting this dependence should have small impact.

5 Image segmentation

5.1 Procedure

The Bayesian paradigm is to base segmentation on the posterior distribution

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

Depending on the type of loss function, different estimates for $\mathbf{z}$ emerge. Using a global $0-1$ loss function,

$$L(\widehat{\mathbf{z}}, \mathbf{z}) = I(\widehat{\mathbf{z}} = \mathbf{z}),$$

the optimal estimate for $\mathbf{z}$ is the Maximum A Posteriori (MAP) solution:

$$\widehat{\mathbf{z}}^{\mathbf{MAP}} = \underset{\mathbf{z}}{\mathbf{argmax}}\, p(\mathbf{z}|\mathbf{y}^1, ..., \mathbf{y}^p).$$

Alternatively, if

$$L(\widehat{\mathbf{z}}, \mathbf{z}) = \sum_i I(\widehat{\mathbf{z}}_i, \mathbf{z}_i),$$

a Maximum Posterior Marginal (MPM) type of estimator given by

$$\widehat{\mathbf{z}}_i^{\mathbf{MPM}} = \underset{\mathbf{z}_i}{\mathbf{argmax}}\, p(\mathbf{z}_i|\mathbf{y}^1, ..., \mathbf{y}^p)$$

is optimal.

In image analysis, minimisation of the energy function, defined by

$$\begin{aligned} U(\mathbf{z}) &= -\log(p(\mathbf{z}|\mathbf{y}^1, ..., \mathbf{y}^p)) \\ &= -\log(p(\mathbf{z})) - \sum_{j=1}^{p} \log(p(\mathbf{y}^j|\mathbf{z})) \end{aligned}$$

is often considered. This minimisation corresponds to maximising the posterior distribution. Writing

$$\begin{aligned} U_z(\mathbf{z}) &= -\log(p(\mathbf{z})) \\ U_d^j(\mathbf{z}) &= -\log(p(\mathbf{y}^j|\mathbf{z})) \end{aligned}$$

the energy becomes

$$U(\mathbf{z}) = U_z(\mathbf{z}) + \sum_{j=1}^{p} U_d^j(\mathbf{z})$$

In the image processing literature, such energy functions are sometimes modified to

$$U(\mathbf{z}) = \sum_{j=1}^{p} \phi_j [U_z(\mathbf{z}) + U_d^j(\mathbf{z})] = U_z(\mathbf{z}) + \sum_{j=1}^{p} \phi_j U_d^j(\mathbf{z})$$

where we without loss of generality have assumed $\sum_j \phi_j = 1$. In a modelling setting this would correspond to changing $p(\mathbf{y}^j|\mathbf{z})$ to $p(\mathbf{y}^j|\mathbf{z})^{\phi_j}$, indicating that the distributions are not properly specified. For Gaussian distributions, this would mean modification of the covariance structure. However, specification of the ϕ_j's are usually made by a criterion corresponding to minimisation of error rate while maximum likelihood estimation is used for specification of parameters inside the U_d^j's. A comparison of these different parameter specification procedures will be considered in future work.

5.2 Computational issues

We will discuss three alternatives for segmentation. The first (Markov Chain Monte Carlo) can be used for obtaining MPM estimates. The second (the EM algorithm) aims at calculating the MAP solution. The last (the ICM algorithm) is generally considered as an approximation to the MAP solution, but can also be interpreted as something between MPM and MAP.

For simplicity, we here neglect temporal dynamics and assume that all observation are made at the same point of time. A central part of the segmentation process is the calculation of the posterior probability of class membership for a given pixel, given all available data *and* all other class memberships. Now

$$p(\mathbf{z}_i|\mathbf{y}, \mathbf{z}_j, j \neq i) \propto p(\mathbf{z}_i|\mathbf{z}_j, |j-i|=1) f(\mathbf{y}_{(i)}|\mathbf{z}) \tag{9}$$

where $\mathbf{y}_{(i)}^t$ is the collection of all observations obtained from areas covering pixel i. The quantity $p(\mathbf{z}_i = k|\mathbf{z}_j, |j-i|=1)$ is directly available from the prior model. The other quantity $f(\mathbf{y}_{(i)}|\mathbf{z})$ will contain multiplicative terms of the form (6) which can be difficult to obtain if data are given on other resolutions than the reference resolution. A way around this problem is to treat the $\mathbf{y}_{i_v}$'s as additional unknown stochastic variables. Assume $\tilde{\mathbf{y}}$ contains data on all resolution levels. Define $\mathbf{y}^c = \{\mathbf{y}_{j,i_v}, v = 1, ..., m\}$ to be the "complete" set of observations. Note that $\tilde{\mathbf{y}}$ is a function of $\mathbf{y}^c$, because of equation (5).

5.3 Markov Chain Monte Carlo

A general procedure that can work, but that will be slow, is to use Markov chain Monte Carlo (MCMC) simulations with both $\mathbf{z}$ and all $\mathbf{y}_{j,i_v}$'s treated as stochastic variables. Such a procedure would give the full posterior distribution for $\mathbf{z}$ given the available data. If processing don't need to be performed very frequently, a high computational cost can be allowed. An MCMC algorithm for simulation from $p(\mathbf{z}, \mathbf{y}^c|\mathbf{y})$ will typically simulate small blocks of $(\mathbf{z}, \mathbf{y}^c)$ at each iteration step. A complicating factor will be that $\mathbf{z}_i$ has a state space that is a combination of a discrete space δ and a continuous one μ. In order to perform simulations, reversible jump MCMC algorithms (Green 1995) must be applied.

The advantage of including $\mathbf{y}^c$ as missing variables in the MCMC step is to avoid the difficult calculations of (6). If observations are assumed Gaussian, direct calculation of (9) is possible, and simulation of $\mathbf{y}^c$ can be avoided.

5.4 The EM algorithm

Another alternative is to use the Expectation-Maximisation (EM) algorithm treating $\mathbf{z}$ as the parameter of interest while $\mathbf{y}^c$ is treated as missing data. The EM algorithm then alternates between two steps:

Expectation step : Calculate $Q = E[\log p(\mathbf{z}|\mathbf{y}, \mathbf{y}_c)]$ where the expectation is over $\mathbf{y}_c$ given $\mathbf{z}$ and $\mathbf{y}$.

Maximisation step : Maximise Q with respect to $\mathbf{z}$.

The maximisation step can be simplified, as it is enough that only a better $\mathbf{z}$ configuration than the previous one is found, for instance through a few iterations of the ICM procedure. This algorithm is guaranteed to increase the posterior probability at each step.

5.5 ICM

Another alternative is to to construct some kind of ICM procedure (Besag 1986), if possible .

Possible procedure: For each pixel i, find the optimal value of $\mathbf{z}_i$ given $\mathbf{z}_j$ for $j \neq i$ and the data $\mathbf{y}$.

5.6 Discussion

The MCMC approach can be used to obtain marginal posterior modes while the EM- and ICM-based algorithms will converge towards local modes of the full posterior. In addition, the MCMC approach can be used for obtaining summary statistics and uncertainty measures thereof.

The MCMC approach will probably be too computer intensive, but it should be implemented as an option. It will further be an important part of estimation procedures discussed in the next section.

Which of the EM- and ICM-based algorithms that is preferable, is difficult to say before experimenting on real data. Both procedures should therefore be implemented and tried out.

More advanced methods aiming at finding the actual MAP solution, based on the framework presented in Storvik and Dahl (2000), could be of interest, but should perhaps not be given priority at this stage.

6 Estimation of parameters

There are two sets of parameters. The first set is those involved in the prior distribution for the class process. Although in principle these parameters can be estimated from data, training data sets are not always representative with respect to the distribution of classes (and in particular transitions between classes) that will be seen in real images. When estimation is performed purely based on training data, we therefore prefer to specify this through prior knowledge about the images to be observed. Also in the case of estimation based on unclassified data, one should consider the possibility of specifying these parameters, or incorporate information about these parameters into a prior on the parameters.

The other set involves parameters connected to the distributions for the observed data (mainly observed images but also other data sources).

We will in the following define θ to be the vector of all parameters involved, with θ_0 being the subpart containing parameters from the prior of $\mathbf{z}$.

6.1 Estimation based on training sets

If training sets with true class labels are available, estimates can be obtained through maximum likelihood estimation, i.e. through maximisation of the likelihood function

$$L(\theta) = \prod_j p(\mathbf{y}^j|\mathbf{z};\theta).$$

Typically, parameters are only present in one of the distributions at the right hand side, i.e.

$$L(\theta) = \prod_j p(\mathbf{y}^j|\mathbf{z};\theta_j).$$

In that case, we may maximise each

$$L_j(\theta^j) = p(\mathbf{y}^j|\mathbf{z};\theta_j)$$

separately.

6.2 Estimation based on unclassified data

A problem with estimates purely based on training sets is that such data can be sparse, in particular for data at coarse scales. In such cases, utilising the large amount of data with unknown class labels can be beneficial.

Two main approaches approaches are possible for estimation in such cases. The first approach is a Bayesian one based on MCMC. It is discussed in section 6.2.1. The second approach is based on maximum likelihood estimation using the EM algorithm, and is discussed in section 6.2.2.

6.2.1 Estimation based on MCMC

The Bayesian approach is to treat the unknown parameters θ as stochastic variables. A prior distribution $\pi(\theta)$, reflecting the prior information about the unknown parameters, is specified. Inference about θ is based on the posterior distribution

$$p(\theta|\mathbf{y}) \propto \pi(\theta)p(\mathbf{y}|\theta). \tag{10}$$

A common estimator for θ is the posterior mean,

$$\mathsf{E}^{p(\theta,|\mathbf{y})}[\theta] = \int_{\theta} \theta p(\theta|\mathbf{y}) d\theta.$$

Direct calculation of such integrals are typically not possible for complex models, as in our case. A common approach in modern statistics is to approximate the above expectation by simulation:

$$\mathsf{E}^{p}(\theta,|\mathbf{y})[\theta] \approx \frac{1}{M} \sum_{m=1}^{M} \theta^m$$

where $\theta^m, m = 1, ..., M$ are (approximate) samples from the distribution $p(\theta|\mathbf{y})$.

Simulation from $p(\theta|\mathbf{y})$ is far from simple, mainly due to the complicated distribution $p(\mathbf{y}|\theta)$. A simplifying approach is to extend the simulation to include the "missing" variables $\mathbf{z}$ and $\mathbf{y}^c$. The combined posterior for all these parameters is given by

$$p(\theta, \mathbf{z}, \mathbf{y}^c|\mathbf{y}) \propto \pi(\theta) p(\mathbf{z}|\theta) p(\mathbf{y}^c|\mathbf{z}; \theta)$$

utilising that $\mathbf{y}$ is a subset of $\mathbf{y}^c$. If $(\theta, \mathbf{z}, \mathbf{y}^c)$ is a simulated sample from $p(\theta, \mathbf{z}, \mathbf{y}^c|\mathbf{y})$, then θ is a simulated sample from $p(\theta|\mathbf{y})$, which is what we need.

Simulation from $p(\theta, \mathbf{z}, \mathbf{y}^c|\mathbf{y})$ can be performed by MCMC algorithms. Simulation of $(\mathbf{z}, \mathbf{y}^c)$ given θ can be done as discussed in 5.3. The distribution of θ given the other variables will depend on the prior distribution $\pi(\theta)$. A complicated part in this case is the unknown normalisation constant in the distribution $p(\mathbf{z}|\theta)$ which will depend on θ. This complication will disappear when parameters in the distribution for $p(\mathbf{z})$ are specified by hand. For parameters involved in the distributions of the data, conjugate priors can be used, making direct sampling possible.

6.2.2 Maximum likelihood estimation

An alternative approach is to maximise the likelihood

$$L(\theta) = p(\mathbf{y}|\theta)$$

with respect to θ. This approach treat θ as fixed parameters and do not need to specify a prior for them.

The likelihood has the same complicated structure as the posterior distribution, and runs into the same problems as distribution (10) due to the unknown $\mathbf{z}$ and $\mathbf{y}^c$. In this case, a possible numerical procedure is to treat $\mathbf{z}$ and $\mathbf{y}^c$ as missing data in an EM context. The likelihood for the complete data is then given by

$$L_c(\theta; \mathbf{z}, \mathbf{y}^c) = p(\mathbf{z}) \prod_{j=1}^{p} p(\mathbf{y}_j^c|\mathbf{z}; \theta)$$

while the log-likelihood is equal to

$$l_c(\theta; \mathbf{z}, \mathbf{y}^c) = \log[p(\mathbf{z})] + \sum_{j=1}^{p} \log[p(\mathbf{y}_j^c|\mathbf{z}; \theta)]$$

The EM algorithm is based on two steps:

Expectation step : Calculate

$$Q(\theta, \theta^s) = E_{\theta^s}[l_c(\theta; \mathbf{z}, \mathbf{y}^c)|\mathbf{y}]$$

where expectation is over $\mathbf{z}$ and $\mathbf{y}_j^c$.

Maximisation step : Maximise $Q(\theta, \theta^{s-1})$ with respect to θ to obtain θ^{s+1}.

General theory guarantees that $l(\theta^s; \mathbf{y})$ is a nondecreasing function in s, making $\{\theta^s\}$ converge to a (local) maximum.

A problem with this algorithm is that the expectation step will be difficult, due to the fact that both $\mathbf{z}$ and $\mathbf{y}_j^c$ are missing. The common way of performing the expectation step in such cases is simulation. In this case we need to simulate from the distribution of $(\mathbf{z}, \mathbf{y}^c)$ given $\mathbf{y}$ and θ. Compared to the Bayesian approach discussed previously, simulation is somewhat simpler, as we don't need to simulate θ. However, the distribution is complex, and MCMC procedures are necessary. These can be constructed similarly to the posterior simulation.

6.2.3 Estimation during ICM

Besag (1986) proposed an approximate procedure for parameter estimation based on a "pseudo-likelihood" criterion. Let θ_0 be the parameters involved in the prior distribution of $\mathbf{z}$, while θ_j is the parameters involved in the likelihood of the data from source j. The procedure is defined as follows:

1. Obtain an initial estimate $\widehat{\mathbf{z}}$ of the true scene $\mathbf{z}$ with guesses for θ_0 and $\theta_j, j = 1, ..., p$.
2. Estimate θ_0 by maximum pseudo-likelihood on the current $\widehat{\mathbf{z}}$ to obtain a new $\widehat{\theta}_0$, that is, choose $\widehat{\theta}_0$ to maximise

$$\prod_i p(\widehat{z}_i | \widehat{z}_j, j \neq i).$$

3. Estimate θ_j by the value $\widehat{\theta}_j$ which maximises

$$p(\mathbf{y}_j | \widehat{\mathbf{z}}; \theta_j) = p(\mathbf{y}_j | \widehat{\mathbf{z}}; \theta_j). \tag{11}$$

4. Carry out a *single* cycle of ICM based on the current $\widehat{\mathbf{z}}, \widehat{\theta}$.
5. Return to 2 until convergence.

Although Besag (1986) state that little is known of the convergence properties of the above procedure, experiences have been encouraging. The advantage of this approach is that maximisation of the pseudo-likelihood (11) is much simpler than maximisation of the full likelihood of $\mathbf{z}$, which involves a complicated normalisation constant.

In our case with "missing" data also among the observations $\mathbf{y}$, the maximisation of $p(\mathbf{y}_j | \widehat{\mathbf{z}}; \theta_j)$ can be complicated. One possibility is to use an internal EM algorithm on this part. Possibly, only one or a few iterations in this integral EM algorithm is needed.

7 Extensions for class properties and sensing condition properties

7.1 Class properties and ancillary processes

So far the underlying truth is defined through class membership. One might be interested in a more specific property which can conditioned on the class, e.g. the biomass of a forested region. Furthermore, other processes such as weather (temperature, rainfall, presence of snow and so on) can also affect the observations. In order to take this into account, we introduce two new processes, $\mathbf{P}$ and $\mathbf{W}$.

$\mathbf{P}$ is a process describing the properties of a class in a more detailed way. We divide $\mathbf{P}$ into $\{\mathbf{P}_k, k = 1, ..., K\}$ where $\mathbf{P}_k$ are properties related to class k. The processes for the different classes are modelled independently. Typically $\mathbf{P}_k$ is some continuous valued process which can be modelled as a Gaussian (or log-Gaussian) process. Note that $\mathbf{P}_k$ is defined at all spatial locations, also at those for which $\mathbf{z}$ is different from k. Our interest will however only be in $\mathbf{P}_k$ where the underlying class *is* k. Defining $\mathbf{P}_k$ everywhere makes it possible to specify correlation structures through "holes" in the spatial domain with different classes.

$\mathbf{W}$ are processes totally unrelated to the $\mathbf{Z}$ process (typically weather variables) which can affect the observations. Such processes can be modelled either through MRF models, if the variable is categorical (i.e. presence of snow), or through (log-) Gaussian fields for continuous variables.

Models for data can now be made conditional on $\mathbf{z}$, $\mathbf{P}$ and $\mathbf{W}$. Inference on $\mathbf{z}$ can be made by considering the simultaneous posterior for $(\mathbf{z}, \mathbf{P}, \mathbf{W})$:

$$\mathrm{Pr}(\mathbf{z}, \mathbf{P}, \mathbf{W} | \mathbf{y}) \propto \mathrm{Pr}(\mathbf{z})\, \mathrm{Pr}(\mathbf{P} | \mathbf{z})\, \mathrm{Pr}(\mathbf{W}) p(\mathbf{y} | \mathbf{z}, \mathbf{P}, \mathbf{W}).$$

Because of the existence of $\mathbf{P}_k$ at all sites, a reasonable assumption can be $\mathrm{Pr}(\mathbf{P} | \mathbf{z}) = \mathrm{Pr}(\mathbf{P})$.

References

M. A. Abidi and R. C. Gonzalez. *Data Fusion in Robotics and Machine Intelligence*. Academic Press, Inc., 1992.

H. H. Arsenault and G. April. Properties of speckle integrated with a finite aperture and logarithmically transformed. *Journal of the Optical Society of America*, 66(11):1160–1163, 1976.

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

H. Caillol, A. Hillion, and W Pieczynski. Fuzzy random fields and unsupervised bayesian segmentation of images. *IEEE Trans. Geoscience and Remote Sensing*, 31:891–810, 1993.

H. Caillol, W Pieczynski, and A. Hillion. Estimation of fuzzy gaussian mixture and unsupervised statistical image segmentation. *IEEE Trans. Image Processing*, 6(3):425–440, 1997.

H. S. Choi, D. R. Haynor, and Y. M. Kim. Partial volume tissue classification of multichannel magnetic-resonance images - a mixel model. *IEEE Trans. Medical Imaging*, 10(3):395–407, 1991.

M. L. Comer and E. J. Delp. Segmentation of textured images using a multiresolution gaussian autoregressive model. *IEEE Trans. Image Processing*, 8:408–420, 1999.

R. Fjørtoft and A. Lopès. Estimation of the mean radar reflectivity from a finite number of correlated samples. *IEEE Trans. Geoscience and Remote Sensing (TGARS)*, 39(1):196–199, January 2001.

J. W. Goodman. Statistical properties of laser speckle patterns. In J. C. Dainty, editor, *Laser Speckle and Related Phenomena*. Springer-Verlag, New York, second edition, 1984.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

G. Høst, G. H. Steinbakk, I. F. Tvete, Ø. Skare, and C. Varin. A generalised linear mixed model for multiple scale remote sensing data. Manuscript under preparation, 2001.

J. T. Kent and K. V. Mardia. Spatial classification using fuzzy membership models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(5):659–671, 1988.

S Krishnamachari and R. Chellappa. Multiresolution gauss-markov random field models for texture segmentation. *IEEE Trans. Image Processing*, 6(2):251–267, 1997.

S. Lakshmanan and H. Derin. Valid parameter space for 2-d gaussian markov random-fields. *IEEE Transactions on Information Theory*, 39:703–709, 1993.

J. Li, R. Gray, and R. Olshen. Multiresolution image classification by hierarchical modeling with two-dimensional hidden markov models. *IEEE Transactions on Information Theory*, 46:1826–1841, 2000.

L. M. Novak and M. C. Burl. Optimal speckle reduction in polarimetric SAR imagery. *IEEE Trans. Aerospace Electronics Systems*, 26(2):293–305, 1990.

C. J. Oliver and S. Quegan. *Understanding Synthetic Aperture Radar Images*. Artech House, Norwood, MA, 1998.

W. Pedrycz. Fuzzy sets in pattern recognition: Methodology and methods. *Pattern Recognition*, 23 (4):121–146, 1990.

F Salzenstein and W Pieczynski. Parameter estimation in hidden fuzzy markov random fields and image segmentation. *Graphical models and image processing*, 59(4):205–220, 1997.

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

D. S. Zrnic. Moments of estimated input power for finite sample averages of radar receiver outputs. *IEEE Trans. Aerospace Electronics Systems*, 11(1):109–113, January 1975.