Spatial and Temporal Models in Contextual Classification

Bo Ranneby

Abstract. Different kinds of spatial and temporal models and their use for contextual classification of satellite images are discussed. Models describing the ground cover and its changes over time and models of the spectral signatures in multitemporal images are studied. Isotropic autocorrelation models as well as mosaic models are used for quantification of the variation in space and time. For the different models parameter estimation, robustness properties, and classification performance are examined. The use of appropriate estimation techniques is crucial for the result of the classification. The use of multitemporal images makes it easier to construct robust estimation methods, and also improves the classification results.

INTRODUCTION

The purpose of this paper is to discuss some models and estimation problems which appear in the classification of remote sensing images. The true image on a two-dimensional regular lattice $D$ will be denoted by $\{C(s), s \in D\}$. The measurement of pixel $s$ is given by $Z(s)$. Unfortunately, there is no one to one correspondence between the $C$- and $Z$-process, but they are contaminated with noise. The classification algorithms referred to in the present paper relies on Bayesian methods. Let $\pi(c)$ denote the prior distribution of the $C$-process and let $f(Z | C)$ denote the conditional distribution of $Z$ given $C$. Then the posterior distribution of $C$ given $Z$ is $p(C | Z) \propto \pi(c) f(Z | C)$. The maximum a posteriori estimator is the one that maximizes $p(C | Z)$. The crucial steps in the derivation of the classification algorithm are modelling the different processes.

MOSAIC-MODELS

For many of the applications we have in mind mosaic-models or models related to mosaic models will give a realistic description of true classes of the image and the development over time of these classes. As a starting point to the

---

1 Department of Forest Resource Management and Geomatics
Swedish University of Agricultural Sciences
S-901 83 UMEÅ
Sweden
study of mosaic-models we consider stochastic methods of dividing the plane into contiguous subsets called cells.

A simple (to handle mathematically) mosaic is the following line-mosaic, see Pielou (1964) and Switzer (1965). The plane is partitioned into convex cells by locating straight lines at random according to the following procedure. Choose points \((\alpha,\theta)\), in the infinite strip \(0 \leq \theta < \pi\), according to a Poisson process with intensity \(\gamma\) points per area unit and associate with each point \((\alpha,\theta)\) the line

\[ y \cos \theta + x \sin \theta - \alpha = 0 \]

Each cell is then independently assigned its color according to a probability law, which selects class \(i\) with probability \(p_i\). If there is only two colors to choose between we get the Pielou-Switzer model.

This model has also been used by Owen (1984) in contextual classification. If the intensity parameter of the Poisson field is small relative to the size of the pixels then:

(i) only three types of patterns \(X, L,\) and \(T\) are possible in the first order neighborhoods \((s, Ns, Ss, Es, Ws)\).

(ii) each neighborhood has the same probability of intersecting one boundary line, and small probability of intersecting two or more boundaries.

(iii) the portion of a boundary within a neighborhood is a line segment, so that a single boundary can induce only an \(L\)- or a \(T\)-pattern.

To this model the variation in time is naturally incorporated by independently assigning a stochastic process \(\xi(t)\) to each cell. The processes give the color of the cells at different times.

As the size of forest stands is large compared to the pixel size this kind of models will be useful in contextual classification of forested areas.

Let us now consider another way of generating the cell partition. Consider points (centers) located in the plane according to a Poisson process with density \(\lambda\). Attach to each cell its Dirichlet cell consisting of those points that are closer to this center than to any other. The Dirichlet cells of a realization of a Poisson process constitute a random division of the plane into polygons varying in size and shape. The variance of the area of a cell is \(0.28018 / \lambda^2\), see Matérn (1979). The mosaic, and its variation over time, is then obtained by independently assigning a stochastic process to each cell giving the color of the cell at different times.
For space-time models of this kind it holds that the correlation in space and time is the product of the spatial correlation and the correlation in time given by the correlation of the process $\xi(t)$. As shown in Ranneby (1982) this holds for a rather large class of models. If we have a partition into cells so the cell-model is isotropic and we to each cell, independently of the cell partition and of other cells, have attached stochastic processes $\xi_1(t), \xi_2(t),...,\xi_n(t)$ to the cells, then the correlation function of the variation in space and time is given by the product of the correlation function of the variation in space and the correlation function of the variation in time. It is crucial that the $\xi$-processes have the same distribution. The result can be extended to models where neighboring cells may be dependent.

The mosaic models above describe the variation on the ground. An extremely simple model describing probabilities of different class configuration are obtained by assuming conditional independence. The conditional probability of getting class configurations $a,b,c,d$ given class $k$ in the centre pixel is given by:

$$g(a, b, c, d \mid k) = p(a \mid k) p(b \mid k) p(c \mid k) p(d \mid k)$$

This model has been used by Haslett (1985) for contextual classification. It is easily extended to images from several occasions, see Flygare (1993).

MODELS FOR SPECTRAL REFLECTANCES

For modelling the spectral reflectances autocorrelation models have shown to be useful. In most cases these models are assumed to be multivariate Gaussian processes. The vector $X(s)$ of observed reflectances for pixel $s$ at time $t$ is assumed to be a sum of two independent variables $Y(s)$ and $\varepsilon(s)$. The $\varepsilon$-process is autocorrelated noise, while the $Y$-process gives the true signal. Usually we have images from one or two occasions. When we have two images $t$ equals 1 or 2. When we have only one image the index $t$ is omitted. The following assumptions will be made for the autocorrelation models.

$A_1$: The random fields $\{Y(s), s \in D\}$ and $\{\varepsilon(s), s \in D\}$ are independent

$A_2$: The random field $\{\varepsilon(s), s \in D\}$ is Gaussian with expectation zero and $\text{Cov}(\varepsilon(s), \varepsilon^T(s)) = \rho^d \theta \Sigma$, where $d = \|s - s_1\|$ is the distance between $s$ and $s_1$.

$A_3$: The random field $\{Y(s), s \in D\}$ is Gaussian with expectation zero and covariance matrix $(1 - \theta)\Sigma$ and $\text{Cov}(Y(s), Y^T(s_1)) = 0$, $s \neq s_1$.

The parameter $\theta$, $0 \leq \theta \leq 1$, is a proportion parameter telling us the amount of variation in $X(s)$ coming from $\varepsilon(s)$ vs $Y(s)$ and $\rho$, $0 \leq \rho < 1$, is the spatial
autocorrelation between noise vectors \( \varepsilon(s) \) and \( \varepsilon(s_1) \) for pixels a unit distance apart.

The assumptions \( A_1-A_3 \) are enough for one image models. For two image models the following additional assumptions are made:

\[ A_4: \] The three random fields \( \{ (Y(s,2), Y(s,1)), s \in D \}, \{ \varepsilon(s,2), s \in D \} \) and \( \{ \varepsilon(s,1), s \in D \} \) are independent.

\[ A_5: \] The random fields \( \{ \varepsilon(s,2), s \in D \} \) and \( \{ \varepsilon(s,1), s \in D \} \) are Gaussian with expectation zero and with the same covariance matrix \( \theta \Sigma \).

\[ A_6: \] \( (Y(s,2), Y(s,1)) \) and \( (Y(s_1,2), Y(s_1,1)) \) are independent for \( s \neq s_1 \).

\[ A_7: \] \( (Y(s,2), Y(s,1)) \) have Gaussian distributions with zero expectation and covariance matrix \( (1 - \theta)\Sigma \) for \( Y(s,2) \) and \( Y(s,1) \) and with cross covariance function \( \text{Cov}(Y(s,2), Y^T(s,1)) = \tau(1 - \theta)\Sigma \).

To make these models operational the different parameters have to be estimated.

**ESTIMATION PROBLEMS**

There are interesting estimation problems associated with several of the models described above. Usually the training sets for the different classes are selected by subjective methods - areas which are typical for the different classes are chosen. This implies that the means may be unbiasedly estimated but the variation within the classes may be heavily underestimated. If so, too many pixels will not be “typical” for any class, which may result in an unnecessary large “doubt”-class. If the training sets are chosen by objective methods this problem is avoided.

For good performance of classification methods based on the autocorrelation model it is necessary to have almost unbiased estimates of the parameters. If the noise-process had been observable the autocorrelation parameters had been estimated by replacing the theoretical moments by their empirical correspondence. Since we can only observe the X-process there is a risk that the autocorrelations in the noise-process are contaminated with the autocorrelations for the true spectral reflectances. To avoid this homogeneous pixelcrosses (the pixel and its N-, S-, W-, and E-neighbors) should be used. If the spectral reflectances are the same for all pixels, then as we want, it is only the noise-process, which causes the autocorrelations. This is also why Hjort et al (1984) recommends homogeneous pixel crosses. This is the best we can do if only one image is available.

If images from two occasions, close in time, are available it is possible to base our estimates on the difference of the two images. Taking the differences over areas with slow changes we get under reasonable assumptions, the difference between the
autocorrelated noise for the two images. This has been studied in Flygare (1996), where she discusses some robustness properties and shows strong consistency of the autocorrelation estimates.

Let us have a closer look at the robustness properties for the two methods. Suppose that assumption $A_3$ in the previous section is not satisfied but that $\text{Cov}(Y(s), Y(s_1)) = \eta(d)(1 - \theta)\Sigma$, where $\eta(d)$ is an unknown positive correlation function with $\eta(0) = 1$. If $A_3$ holds then $\eta(d) = 0$ for $s \neq s_1$.

For the one image method we get that

$$E \left[ \frac{1}{n_k} \sum (X^k(s) - \mu(k))(X^k(s_1) - \mu(k))^T \right] \rightarrow \eta(d)(1 - \theta)\Sigma + \rho^d \theta \Sigma,$$

As $\eta(d)$ in most applications is larger than zero this results in an overestimation.

For the difference method the situation is somewhat different. Here we get the following expected value for our estimate:

$$2(1 - \tau)\eta(d)(1 - \theta)\Sigma + 2\rho^d \theta \Sigma.$$

If $\tau$ is close to 1 we see that the expected value is approximately $2\rho^d \theta \Sigma$, which is what we get when all assumptions are fulfilled. As a consequence the difference method is less sensitive to errors in the model.

**MIXTURES OF MULTINORMAL DISTRIBUTIONS**

In Taxt et al. (1991) mixtures of normal distributions are used for classification of images. They have rather frequently observed bimodal and multimodal empirical distributions for feature vectors from the same class, when working with remote sensing images. If the training sets are objectively selected there are reasons to believe that empirical distributions of this kind will be more common. Taxt et al. (1991) used linear combinations of two multinormal distributions to model the distributions for the different classes. A typical model for the density function $f(x)$ of the feature vector is given by

$$f(x) = (1 - \pi)N_d(\mu_1, K)(x) + \pi N_d(\mu_2, M)(x)$$

The estimation of the parameters in such distributions is known to be a difficult problem, both from statistical and algorithmic points of view.

Example. Let $\xi_1, \xi_2, \ldots$ be i.i.d. observations from a mixture of two bivariate normal distributions. The density function $f(x,y,\theta)$ is given by

$$f(x,y,\theta) = p \ h(x,y,\mu_1,\mu_2,\sigma_1,\sigma_2,\rho_1) + (1-p) \ h(x,y,\mu_3,\mu_4,\sigma_3,\sigma_4,\rho_2),$$

455
where \( h \) is the density function for a bivariate normal distribution with parameters indicated by the notation. Say that we have observations \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\). If we put \( \mu_1 = x_1, \mu_2 = y_1 \) and let \( \sigma_1, (or \sigma_2) \) go to zero, then the likelihood function tends to infinity. Consequently, the ML-method is not suitable.

The same problem appears also for univariate distributions. In Ranneby (1984) a new estimation method, the Maximum Spacing (MSP) Method, was introduced which gives consistent estimators also for distributions of this kind. The method is closely related to the Maximum Likelihood Method - both can be derived as approximations of the Kullback-Leibler information. The MSP-method is constructed so that each contribution is bounded from above for any parameter combination.

The same estimation method has been used by Cheng & Amin (1983) for three-parameter Weibull and gamma models. The method is also discussed in Titterington (1985), where he states that "in principle, of course it would be possible to treat multivariate data by grouping: although definition of the multinomial cells would be more awkward."

Since we do not have any natural order relation in \( \mathbb{R}^d \) when \( d > 1 \), it is not at all obvious how the generalization shall be carried out. In the univariate case we attach to each observation \( \xi_i \), the variable \( \eta_i (n) = (n + 1) \) "distance to the nearest observation to the right of \( \xi_i \)."

In \( \mathbb{R}^d, d > 1 \), we cannot use \( \eta_i (n) \). However, to each observation \( \xi_i \), we can attach its Dirichlet cell. The Dirichlet cell \( D_n(\xi_i) \) surrounding \( \xi_i \) consist of all points \( x \in \mathbb{R}^d \) which are closer to \( \xi_i \) than to any other observation.

Let \( \xi_1, \xi_2, \ldots \) be i.i.d. random vectors in \( \mathbb{R}^d \) with true density function \( g(x) \) and suppose we assign a model with density functions \( \{f(x, \theta); \theta \in \Theta\} \), where \( \Theta \subset \mathbb{R}^d \). Let \( A \) be a measurable subset of \( \mathbb{R}^d \). Then \( P_\theta (A) \) denotes the probability that \( \xi_i \) belongs to the set \( A \) calculated under the assumption that the density function is given by \( f(x, \theta) \) (i.e. \( P_\theta (A) = \int f(x, \theta) dx \)). The true probability distribution is denoted by \( P_0 (\cdot) \).

The Dirichlet cells \( D_n(\xi_i), i = 1,2, \ldots, n \) split \( \mathbb{R}^d \) into \( n \) identically distributed random sets. Since \( \sum_{i=1}^n P_0 (D_n(\xi_i)) = 1 \), symmetry arguments give

\[
E[P_0 (D_n(\xi_i))] = \frac{1}{n} \quad \text{or} \quad P_0 (D_n(\xi_i)) \equiv \frac{1}{n}.
\]

Since it also holds that

\[
P_0 (D_n(\xi_i)) \equiv |D_n(\xi_i)| g(\xi_i),
\]

we get that \( g(\xi_i) \equiv 1/n |D_n(\xi_i)| \). Here \( |D_n(\xi_i)| \) denotes the "volume of the set \( D_n(\xi_i)\)".
Consequently, the random variable $n D_n(\xi_i) \equiv n D_n(\xi_i) / f_0(\xi_i) / g(\xi_i)$ may be interpreted as a non-parametric estimate of $1 / g(\xi_i)$. Since

$$nP_\theta(D_n(\xi_i)) \equiv n D_n(\xi_i) f_0(\xi_i) \approx f_\theta(\xi_i) / g(\xi_i),$$

it is intuitively obvious that

$$\frac{1}{n} \sum_{i=1}^{n} \ln n P_\theta(D_n(\xi_i))$$

will be an approximation of the Kullback-Leibler distance, $I(g, f_\theta)$, defined by

$$I(g, f_\theta) = \int g(x) \ln g(x) - \ln f_\theta(x) dx.$$

The approximation above is analogous to the univariate approximation given in Ranneby (1984) and the Maximum Spacing Estimate of $\theta$ is obtained by maximizing approximation (1) above.

For each class we estimate the mixture proportion, the mean vectors and the covariance matrices. Remote sensing images have the property that the interclass distances are rather small so the different parts of the mixture distributions are not easy to separate. This makes it difficult to find good starting values which are crucial for other estimation methods. Although the multivariate MSP- method theoretically give an attractive estimation method much work remains until all practical problems are solved.

Taxt et al (1991) found a substantial increase in the correct classification rates when using mixture models instead of multivariate normal distributions.

**CLASSIFICATION RESULTS**

In Flygare (1995) the performance of some contextual classification methods is evaluated on Landsat TM data. Haslett’s model was used for the conditional class distribution and in three of the methods she used models with autocorrelated spectral reflectances. Methods utilizing images from one and two occasions were compared. There was no surprise that an autocorrelation model utilizing information from two occasions gave the best classification results. Mean Absolute Deviation was used to compare the classification results with the evaluation data. A 21% improvement was obtained with the two image autocorrelation method compared to the best one image method.
BIOGRAPHICAL SKETCH

Bo Ranneby is professor at the Swedish University of Agricultural Sciences. He took a Ph.D. in mathematical statistics at Umeå University in 1975. In 1992 he was elected as an ordinary member of the International Statistical Institute. He is head of the Department of Forest Resource Management and Geomatics.

REFERENCES