BAYESIAN CONSTRUCTION OF THEMATIC MAPS FROM SATELLITE IMAGERY

E.J. Green, Rutgers, A.F.M. Smith, Imperial College, W.E. Strawderman, Rutgers
Edwin J. Green, Dept of Natural Resources, Cook College, Rutgers University, PO Box 231,
New Brunswick, NJ 08903-0231, USA email: green@ocean.rutgers.edu

Abstract

We examine the problem of constructing thematic maps which show land-use patterns, given satellite data and deterministic prior information. The latter came from a road network. We make the simplifying assumption that the satellite observations are conditionally independent, given the scene. We specify a normal likelihood, the usual normal-inverse Wishart prior for the distributional parameters, and a Markov random field prior for the image together with the road information. Our results indicate that the Markov random field prior and road network information help improve classification, despite evidence that the assumption of normal likelihoods is suspect.

keywords: Gibbs sampling, Markov random fields.

1. I. Introduction

It is commonplace to develop thematic maps from digitized satellite imagery. Here we define a thematic map to be a two-dimensional representation of the earth's surface, in which each pixel is associated with one and only one land-use classification. In practice, most maps are developed using maximum likelihood. However, it has been shown (*e.g.*, see Besag 1986, Besag et al. 1991; Klein and Press 1992; Wilson and Green 1993) that Bayesian models often result in maps of far superior quality. In this paper, we report the results of an experiment in which thematic maps were developed from multidimensional satellite data. In one instance we employed a simple Bayesian model which just specified a Markov random field prior. In another, we incorporated deterministic information from a road network into the Bayesian model (an idea first explored by Frigessi and Stander 1994). All maps are compared with one produced via maximum likelihood and with a map produced from digitized low-level aerial photography. However it should be noted that the latter map is from 1984, while the satellite data is from 1986, so that while the aerial photography map can be considered to be a suggestion of the "true map," it cannot be considered to be equivalent to the truth. This complicates somewhat our later judgments as to how well methods are performing.

2. Maximum Likelihood method

To facilitate discussion, we introduce the following notation, borrowed largely from Besag (1986): Let S represent the two-dimensional scene to be mapped. Suppose there are n pixels in the map and each pixel is to be assigned to one and only one of c classes. Let $X = (X_1, X_2, ..., X_n)'$ denote a random vector where X_i assigns a class to pixel i. Denote the land use classifications by the integers 1, 2, ..., c, and let x_i be a realization of X_i, *i.e.*, x_i takes on one of the unordered values (1, 2,..., c), $i = 1, 2, ..., n$, and let $x = (x_1, x_2, ..., x_n)'$. Furthermore, let y_i be the p-dimensional observation at pixel i, and $y = (y_1', y_2', ..., y_n')'$. Normally, the first step in the mapping process is to acquire a data set, known as *training data*, for which the classification of each pixel is *known*, and for which there is satellite coverage.

The maximum likelihood (ML) approach bases inference on the assumed probability densities for the observations in each land-use class. It is usual to assume that $(y_j|x_j = k) \sim N(\mu_k, \Sigma_k)$, $j = 1, 2, ..., n$, for both the pixels in S and the training data (we will examine the normality assumption later). The y_j's are assumed to be conditionally independent. Next, the training data are sorted into classes, and μ_k and Σ_k estimated from the training data pixels in class k. Each pixel is assigned to the class with the greatest likelihood. This procedure is carried out pixel by pixel, and it is implicitly assumed that the x_j's are independent. However, once the map is constructed, it is common to observe an isolated pixel classified into one category while all the surrounding pixels are classified into another. This "salt and pepper" appearance is usually considered an anomaly, and some sort of ad-hoc, majority-vote smoother is passed over the map in order to eliminate these situations.

3. Data

In this study, we use data from the thematic mapper (TM) satellite. This satellite records data in 7 bands, and each observation represents a $30m\times 30m$ pixel on the ground. The digitized reflectances in each band are discrete variables with 256 possible values.

Our data are from a summer 1986 image of the Hightstown US Geological Survey quadrangle in New Jersey. The quadrangle is located in the central portion of the state, at the transition between the Piedmont and Coastal Plane physiographic regions. The area which historically has been used for agricultural production is undergoing conversion to more intensive land use activities and exhibits a complex landscape pattern. More details about the data can be obtained from Airola and Vogel (1988).

We chose a cloud-free 200 × 50 pixel area (10,000 pixels) for detailed study. Additionally, we randomly selected 50 observations in each of the five land-use categories from the remaining map area to serve as training data.

Our deterministic information consisted of a digitized road network for the Hightstown quadrangle. Regrettably, the road network is circa 1990. The local history of the Hightstown quadrangle is such that there was a period of rapid conversion of agricultural land to housing developments in the mid 80's. This rapid phase of conversion ended in the late 80's, when real estate values began to decline. As the TM data were from 1986, we requested that a remote sensing expert familiar with the Hightstown area "mask" out all roads which were thought to *possibly* post-date the 1986 TM image. The resulting road network is shown for the 200 × 50 pixel area under study in Figure 1.

We also have access to a 1984 image of the Hightstown quadrangle. This image was produced by the State of New Jersey's Department of Environmental Protection and Energy (DEPE) and was derived from intensive classification of low-level aerial photography. We regard this map as a pseudo-benchmark against which we may judge maps constructed from TM data. However, due to the discrepancy in the dates (1984 vs. 1986) and the fact that we *know* that this area underwent development during the two year interval, we can only regard the DEPE as an approximation of the truth. Note that we used the DEPE classification to select training data. Hence it is possible that some of the pixels in the training data are incorrectly classified, leading to poorer estimates of μ_k and Σ_k, $k = 1, 2, \ldots, c$, than would otherwise be expected.

4. Bayes Model

Our basic model is similar to the ones proposed in the classic papers by Geman and Geman (1984) and Besag (1986). We assume the prior density to be a Gibbs distribution, or equivalently, a Markov random field (MRF). We also assume the fields are locally dependent with second order neighborhoods, i.e., $p(x_i|x_{-i}) = p(x_i|x_{\delta_i})$, where x_{-i} represents the classification of all pixels in S *except* pixel i; x_{δ_i} represents the classification of all pixels in the neighborhood of pixel i; and the neighborhood of pixel i consists of the 8 nearest pixels (*i.e.*, the pixels above, below, to either side, and diagonally adjacent). We assume that x_i is not a member of x_{δ_i}.

Let $u_i(k)$ and $v_i(k)$ be the number of first order neighbors (above, below, and to either side) and number of second order neighbors (diagonal), respectively, of pixel i that are members of class k in an arbitrary realization of X. Then we write the MRF as:

$$p(x_i = k|x_{\delta_i}) \propto \exp\left\{\beta_k \left[u_i(k) + v_i(k)/\sqrt{2}\right]\right\}.$$

Here β_k is a positive constant which encourages clumping of pixels of the same land-use. As β_k, $k = 1, 2, \ldots, c$ is unknown, a formally complete Bayesian model would include a prior density for this parameter. However, it is mathematically intractable to solve for the posterior distribution $p(x_j = k|y_j)$ or the conditional distributions needed in Markov chain Monte Carlo methods if a prior density is assigned to β_k (see, *e.g.*, Besag 1986). Besag (1986) suggested a pseudo-maximum likelihood method for estimating β_k. In an earlier, unpublished simulation study using data from the same quadrangle, we found that the pseudo-maximum likelihood estimates were close to 1.0 for all categories, and that the resulting map was insensitive to moderate perturbation of the values for this parameter. Hence in the present study we have set $\beta_k = 1.0\ \forall k$. The diagonal neighbors are down-weighted by $(1/\sqrt{2})$ to reduce rotational variability (Frigessi and Stander 1994).

With regard to μ_k and Σ_k, we use the usual conjugate normal and inverse Wishart prior distributions (see, *e.g.*, Gelfand et al. 1990). The complete Bayes model is then

$$(y_j|x_j = k, \mu_k, \Sigma_k) \sim N(\mu_k, \Sigma_k),\ j = 1, 2, \ldots, n, \tag{1}$$

$$\begin{aligned} p(x_i &= k \mid x_{\delta_i}) \propto \exp\left\{\beta_k \left[\frac{u_i(k) + v_i(k)}{\sqrt{2}}\right]\right\} \\ i &= 1, 2, \ldots, n, \end{aligned} \tag{2}$$

$$(\mu_i|\eta, \Gamma) \sim N(\eta, \Gamma),\ i = 1, 2, \ldots, c, \tag{3}$$

$$(\Sigma_i^{-1}|\rho, R) \sim W(\rho, (\rho R)^{-1}),\ i = 1, 2, \ldots, c. \tag{4}$$

We also assume conditional independence of the μ_i's and Σ_i's. It is implicit in model [1] that the pixels are conditionally independent, given the scene. This is usually felt to be an inappropriate assumption (*e.g.*, see Hjort and Mohn 1987). It seems intuitive that pixels of the same class which are close together will be more alike than pixels of the same class which are far apart. Green and Wilson (1992) have modeled this spatial correlation. However, their model includes an additional parameter to measure the strength of the correlation. A formally complete Bayesian model including a prior distribution for this parameter seems intractable for the same reasons as for β_k. The alternative is to specify a value for the parameter, or employ a pseudo-maximum likelihood technique. Inclusion of this parameter also complicates the full conditional distributions of μ_k and Σ_k, making Markov chain Monte Carlo difficult. Hence, in the current report we make the simplifying assumption of conditional independence. Work on a more realistic model is ongoing.

Let $y^{(k)}$ indicate the subset of y consisting of all training pixels which are members of class k. From [1], [3], and [4], we derive the densities $p(\mu_k|\Sigma_k^{-1}, y^{(k)})$ and $p(\Sigma_k^{-1}|\mu_k, y^{(k)})$. These are well known (*e.g.*, see Gelfand et al. 1990) to be: $(\mu_k|\Sigma_k^{-1}, y^{(k)}, \eta, \Gamma) \sim N(\nu_k, \Psi_k)$ and $(\Sigma_k^{-1}|\mu_k, y^{(k)}, \rho, R) \sim W(q_k, Q_k)$, where: $\nu_k = \Psi_k \left\{ \left(m_k \Sigma_k^{-1} \bar{y}^{(k)} \right) + \Gamma^{-1}\eta \right\}$; $\Psi_k = \left(m_k \Sigma_k^{-1} + \Gamma^{-1} \right)^{-1}$; $\bar{y}^{(k)} = \sum_{i=1}^{m_k} y_i^{(k)} / m_k$; $Q_k = \left[\sum_{i=1}^{m_k} (y_i^{(k)} - \mu_k)(y_i^{(k)} - \mu_k)' + \rho R \right]^{-1}$; $q_k = m_k + \rho$; $y_i^{(k)}$ is the observation at pixel i in $y^{(k)}$; and m_k is the number of training pixels in category k. These conditional densities represent the prior densities of μ_k and Σ_k before the map is constructed. All that remains to complete the model is to specify values for η, Γ, ρ, and R. We let $\Gamma^{-1} = 0$, indicating vague prior information on μ_k and obviating the need to specify η. Similarly we let $\rho = 0$, obviating the need to specify R. We view this as a way of indicating that we know nothing about the distribution of Σ_k^{-1} before viewing the training data.

The object of interest is the posterior distribution $p(x_j = k|y_j)$, $j = 1, 2, ..., n$. However, analytic solution is impossible and we proceed by using a Gibbs sampling variant of Markov chain Monte Carlo to sample from the posterior distributions, and hence to estimate any function of the posteriors to any desired degree of accuracy. For more details on Gibbs sampling in the context of models of the type used here, consult, among others, Frigessi and Stander (1994), Gelfand and Smith (1990), Gelfand et al. (1990), or Smith and Roberts (1993).

4.1 Deterministic Information

The above represents our basic Bayesian model. As mentioned earlier, we also have access to deterministic prior information: the road network. We incorporated this information by modifying our prior distribution to: $p(x_i = k|x_{\delta_i}, d_i) \propto \exp\left\{\beta_k \left[u_i(k) + v_i(k)/\sqrt{2}\right]\right\} \cdot p(d_i|x_i = k, x_{\delta_i})$. Here d_i is the distance to nearest road for pixel i. Values for $p(d_i|x_i = k, x_{\delta_i})$ were estimated as follows: The 200 × 50 section to be mapped was removed from the quadrangle. Then four neighbor classes were established for the remaining pixels. These classes were 0-2, 3-5, 6-7, and 8 neighbors of the same land use type. The pixels were sorted by land use class and neighbor class. Next the pixels in each land use-neighbor class were sorted into classes depending on distance to the nearest road. Six such classes were established. These were: *i.* $d_i \leq 30$, *ii.* $30 < d_i \leq 60$, *iii.* $60 < d_i \leq 120$, *iv.* $120 < d_i \leq 240$, *v.* $240 < d_i \leq 300$, and *vi.* $300 < d_i.d_i \leq 30$. The number of pixels in each land use-neighbor-road distance class was divided by the number of pixels in the appropriate land use-neighbor class, resulting in a multinomial distribution for $p(d_i|x_i = k, x_{\delta_i})$.

4.2 Gibbs sampler

Here we give a brief description of our application of the Gibbs sampler. Given initial guesses for each parameter, we randomly draw a value for μ_k from $p(\mu_k|x, y, \Sigma_k^{-1})$. Then we draw a value from $p(\Sigma_k^{-1}|x, y, \mu_k)$ (where μ_k was set to the values just obtained). Finally we cycle through the map, drawing a value from $p(x_j = k|y_j, x_{-j}, \mu_k, \Sigma_k^{-1})$ for each pixel, given the values just obtained for μ_k and Σ_k^{-1} and the current classifications of the pixels in its neighborhood. Each randomly drawn value for x_j becomes the current value for that pixel. After we cycle through the map, we draw a new value for μ_k and repeated the process. After a suitably large number of iterations, the randomly drawn values for x_j, μ_k, and Σ_k^{-1} are regarded as having arisen from their joint posterior distribution, and the values of x_j as having arisen from the marginal posterior distribution for pixel j, $j = 1, 2, ..., n$.

In order to investigate the effects of our prior assumptions, we also fitted the Bayes model without prior distributions on μ_k and Σ_k (i.e., these parameters were held fixed at the mle's from the training data) but with the road distance information, and *also* without the prior distributions on μ_k

and Σ_k *or* the road distance information. Hence in the former case we included the MRF and the deterministic road information in the prior, and in the latter, we only included the MRF.

Determining when the Gibbs sampler has converged to the target posterior distributions was problematic. Due to the vast number of parameters, convergence diagnostics based on the observed stream of values for each parameter were impractical. In this study, we use the following crude technique: after each iteration, we computed the number of pixels in each land use category. When these numbers appeared to have settled down from iteration to iteration, we concluded that the sampler had converged.

After we were satisfied that convergence had been attained, we ran the sampler for an additional t cycles. For each pixel, we collected the number of times it was classified into each category during the t cycles. These values formed a histogram, and the pixel ultimately was assigned to the category corresponding to the mode of the histogram.

We used the MLE's from the training data as starting values for μ_k and Σ_k and the maximum likelihood classifications as starting values for the $x_i^{\prime}s$. For all models, it appeared that the Gibbs sampler converged quickly (< 100 iterations). Nevertheless, we ran the Gibbs sampler for 1000 iterations, and discarded the values from the first 500 iterations. Hence our Gibbs sampling estimates are based on samples of size 500 from the marginal posterior distribution for each pixel.

5. Results

The DEPE classification is presented in Figure 2, and the maximum likelihood solution is given in Figure 3. In these maps, as in all subsequent ones, we have labelled the land-use categorizes follows: 1 = developed, 2 = agriculture, 3 = forest, 4 = water, and 5 = transitional areas. The shading in the figures assigns the lightest color to category 1 (developed) and the darkest to category 5 (transitional areas). Figure 3 displays the maximum likelihood map, which is clearly unsatisfactory. The map resulting from specifying priors for μ_k and Σ_k and a MRF prior, but without using the road distance information is displayed in Figure 4 while the map from assuming a MRF prior and using the road distance information with no priors for μ_k and Σ_k is displayed in Figure 5. The other two Bayes maps (MRF, priors μ_k and Σ_k, road distance information used; and MRF only) were similar to Figure 5.

The Bayes map in Figure 5 is generally qualitatively superior to the maximum likelihood map in Figure 3. It seems to capture most of the structure evident in the DEPE classification (Figure 2). The map in Figure 5 is extraordinary; although it appears to capture most of the general detail in the DEPE image, large areas are classified incorrectly. In particular, many of the developed areas are classified as water. Recall that this map results from a assuming a Markov random field, and priors on μ_k and Σ_k, but without using the road information. We believe that the reason for the disappointing results of this model are due to misspecification of the likelihood; the data are not multivariate normal. To support this claim we have drawn a random sample of size 1000 without replacement from the data in each land-use category, and constructed normal Q-Q plots for the data from each band. For nearly all band-category combination, the Q-Q plots showed indications of non-normality. As an example, a Q-Q plot from band 7 for the developed category is displayed in Figure 6. Attempts to transform the data to more closely approximate normal data using standard transformations were unsuccessful. There are at lęast 3 possible reasons for the non-normality evident in the Q-Q plots. First, it may be due to the discrepancy between the dates of the DEPE image and the TM data. The data in the Q-Q plots were selected based on their DEPE classifications, but the data is the TM data. Secondly, the classes may be too broad. For instance, the category "developed" includes housing developments, industrial areas, and parks. Perhaps the data would more closely approximate a normal distribution if finer classes were used. Finally, it may be that the data are truly not normally distributed.

We believe the above problem with the misspecified likelihood arose only in Figure 4 because for this map, we specified prior distributions for μ_k and Σ_k, but did not use the deterministic road information. Since the likelihood was apparently misspecified, there is no reason to believe the conjugate normal-inverse Wishart prior on μ_k and Σ_k is appropriate. For the Bayes model which included priors on μ_k and Σ_k and the deterministic road information, the latter apparently compensated for the misspecification of the likelihood and the priors on μ_k and Σ_k.

In terms of percent error, when judged against the DEPE classification in Figure 2, 28.5% of the pixels in the maximum likelihood map were incorrect. For the Bayes maps in Figures 4 and 5, the percentages incorrect were 36.5%, and 21.4%, respectively. The benchmark for acceptance of a thematic map is 80% accuracy. Hence while none of the maps created here exceed the usual accuracy requirement, the map in Figure 5 is quite close. Given that the truth is unknown, it is possible that these maps are within the required 80% accuracy level.

6. Conclusions

It seems the Markov random field and deterministic road distance information are helpful in creating thematic maps. However, the usual distributional assumptions for the satellite data are questionable.

There is clearly scope for additional research on incorporating deterministic information such as road networks in the image prior. Other source of information, such as elevation, *and/or* hydrological information might be used. Also, rather than incorporating this prior information by means of a multinomial distribution as done here, it may be preferable to characterize this information by means of continuous functions as done in Frigessi and Stander (1994). Of course in the latter approach, estimation of the function parameters may be problematic.

REFERENCES

Airola, T.M. and J. Vogel. 1988. Use of thematic mapper digital data for updating the New Jersey land cover component of the 1987 National Resources Inventory. *Journal of Soil and Water Conservation* 43:425-428.

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

Besag, J., J. York, and A. Mollié. 1991. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43:1-59.

Frigessi, A. and J. Stander. 1994. Informative priors for the Bayesian classification of satellite images. *Journal of the American Statistical Association* 89:703-709.

Gelfand, A.E., S.E. Hills, A. Racine-Poon, and A.F.M. Smith. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 85:972-985.

Geman, S. and Geman D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721-741.

Hjort, N.L. and E. Mohn. 1987. Topics in the statistical analysis of remotely sensed data. *Bulletin of the International Statistics Institute* 52:23-44.

Klein, R. and S.J. Press. 1992. Adaptive Bayesian classification of spatial data. *Journal of the American Statistical Association* 87:844-851.

Smith, A.F.M. and G.O. Roberts. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* **B**, 55:3-23.

Wilson, J.D. and P.J. Green. 1993. A Bayesian analysis of remotely sensed data using a hierarchical model. University of Bristol Mathematics Research Report no S-93-02.

Fig 1. Road Network

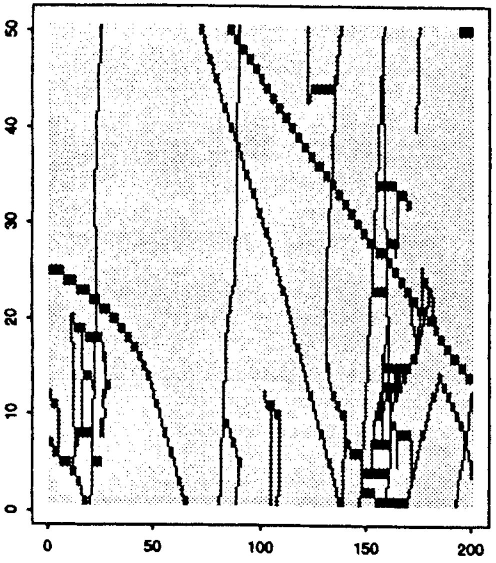


Fig 2. DEPE Classification

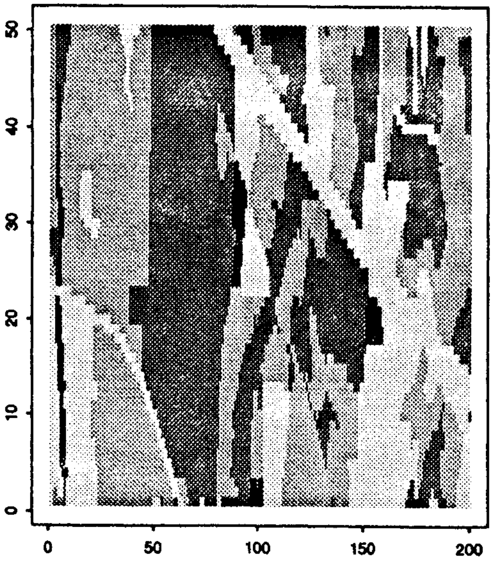


Fig 3. Maximum Likelihood

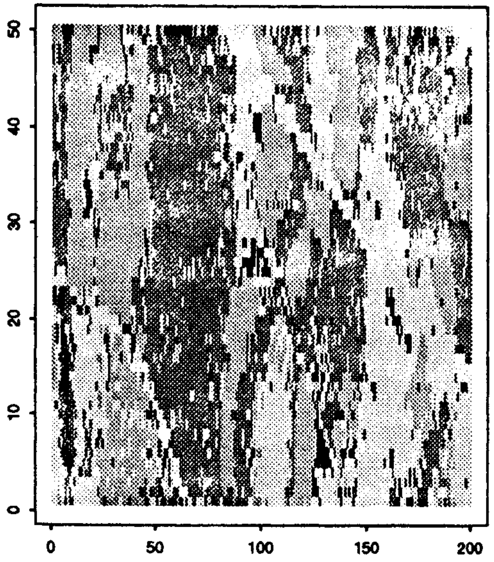


Fig 4. Bayes, no road info.

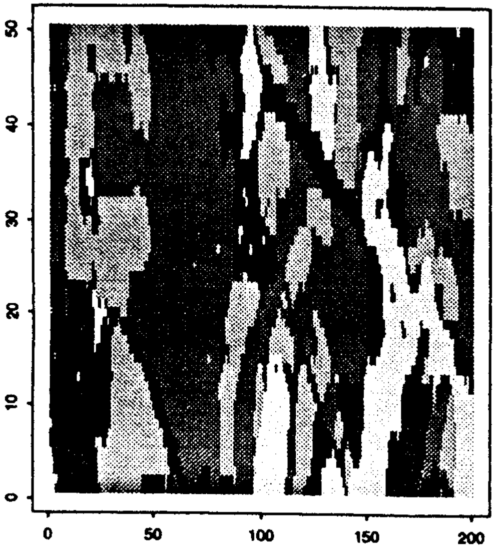


Fig 5. Bayes, no hyperpriors

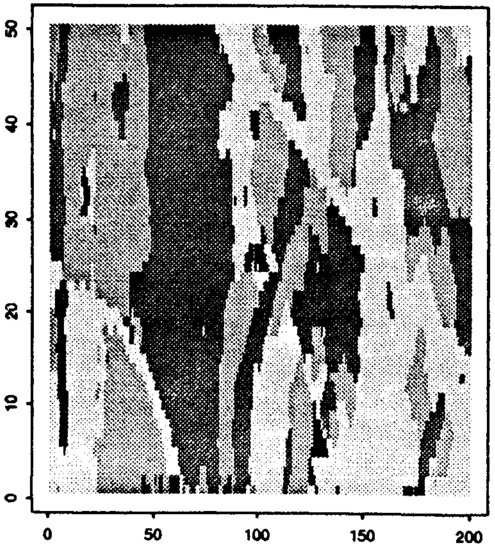
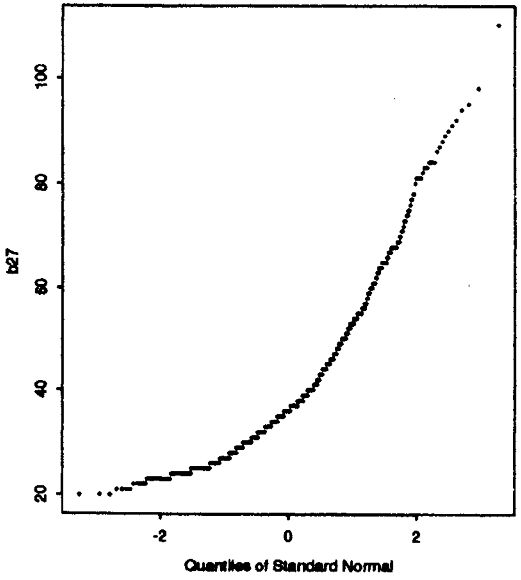


Fig 6. QQ plot, band 7, developed


b27
Quantiles of Standard Normal