Modelling Uncertainty in Satellite Derived Land Cover Maps

Edward Cripps [*] Anthony O'Hagan[†] Tristan Quaife [‡]
Clive Anderson [†]

January 14, 2008

Abstract

The purpose of this paper is to quantify uncertainty associated with land cover maps derived from satellite data. Satellites record energy reflected from the earth's surface and this information may be used to derive maps of land cover. Typically these maps are thematically organised in terms of the land surface vegetation. The final product is reported as recordings of distinct vegetation classes at pixels across a spatial grid. The recordings are possibly misclassified and the accuracy of these land cover maps is often assessed by a confusion matrix derived from a comparison of the map to a field survey at a sample of pixels across the region. The data in the confusion matrix is small relative to the entire map and does not provide spatial information regarding map accuracy. We propose a Bayesian method to address these two issues. The model describes multinomial recordings with misclassification probabilities and incorporates a spatial correlation structure that is suited to the case where little spatial information exists. Our method allows us to estimate the posterior distributions of the land cover proportions for individual sites as well for the entire region, features previously unavailable in accuracy assessment techniques. We present the results of our method applied to a recently developed satellite derived land cover map, the Land Cover Map 2000, for the region of England and Wales.

[*]Australian Institute of Marine Science, Townsville
[†]Department of Probability and Statistics, University of Sheffield
[‡]Department of Geography, University College London.

Keywords: multinomial distribution, misclassification probabilities, spatial correlation, remote sensing, land cover maps, confusion matrix

1 Introduction

Since the early 1990's a variety of satellite derived land cover maps have been made available to, and are now becoming established within, the scientific community. Satellite derived, or remote sensing, land cover maps (henceforth referred to as RS maps) produce estimates of land cover, for example vegetation type, via image classification of satellite data. They are generally considered the most important source of information on contemporary land cover, primarily due to the high spatial resolution maps they produce of wide geographical areas. However, RS maps are not a record of the true vegetation cover - satellite sensors record energy incident at the sensor, situated several hundreds of kilometres above the surface of the earth, from which the land surface information is inferred. Hence, a RS map is an estimate of the truth and the accuracy of it needs to be sensibly quantified. This is evident from the fact that many RS maps show poor agreement, whilst reporting the same physical quantities of interest, for the same region (see Giri et al. (2005) and Hansen and Reed (2000)). Furthermore, users of RS maps are frequently required to manipulate the original product to suit their own needs and this may result in additional classification inaccuracies. A comprehensive review of possible uncertainties of RS maps is beyond the scope of this paper and the reader is referred to Jung et al. (2006) and Foody (2002) for further details.

A significant question in the remote sensing literature then is how to assess the accuracy of a RS map. Typically, accuracy is determined by comparing the RS map with some form of reference data. At randomly sampled locations (often stratified) over the region of interest the vegetation cover is either physically reported or assessed by aerial photography and interpreted by trained individuals. A comparison of the RS land cover map at these points with the reference measurements yields an error or confusion matrix which is used as a criterion to assess accuracy (see Stehman et al. (2000), Mayaux et al. (2006) and Foody (2007)). The accuracy measures derived from the confusion matrix fall into three categories:

1. The probability of correctly classifying pixels across the entire map, known as the overall or map accuracy.

2. The probability that a pixel is classified as vegetation class i, say, by a RS map given that the true vegetation class is j, say. We refer to these probabilities as the forward probabilities, known in the remote sensing literature as user's accuracy.

3. The probability that a pixel has true vegetation class i given that the pixel is classified as vegetation class j by the RS map. We refer to these probabilities as the backward probabilities, known in the remote sensing literature as producer's accuracy.

Green and Strawderman (1994) describe a Bayesian model to assess the above accuracy measures by treating the backward probabilities as the parameters of a multinomial sample contained in the confusion matrix. The RS map provides a known marginal probability of classification from which, together with the backward probabilities, the forward probabilities and map accuracy can be calculated. All accuracy measures are reported as region-wide accuracies and the issue of spatial variation is not addressed.

In this paper we construct a Bayesian model to quantify uncertainty in the true vegetation cover over the region of interest. By uncertainty, as opposed to accuracy, we mean specifying a model that yields a posterior distribution of the true vegetation classes. Our approach allows us to quantify the overall map uncertainty as well as the spatial distribution of uncertainty within the map. To the best of our knowledge we are the first to take such an approach. The model is suited to the case where there exist large amounts of (possibly misclassified) recordings across a spatial region and a small sample with which to calibrate them. Since users of RS maps often require the original high resolution product to be uspcaled to coarser resolutions the model focuses on inference for an aggregation of pixels.

We introduce the model by first describing inference for a single pixel. This stage involves identifying the backward probabilities for a pixel. In constrast to Green and Strawderman (1994), we model the region-wide forward probabilities as the parameters of a multinomial sample contained in the confusion matrix and, combined with a pixel-specific prior for the true vegetation classes, derive the pixel-specific backward probabilities. Next, we extend the above model to include inference for the true proportions of vegetation classes within some area of land comprising a number of pixels. We refer to such an aggregate area as a site. For a given site, the true proportions are distributed as an average of multinomial distributions with probabilities the relevant backward probabilities for that site. For inference

at the aggregation level it is important to incorporate the spatial correlation of the proportions. Since we have essentially no information regarding this correlation we describe a simple approach that induces spatial correlation of the various proportions by modelling their variance.

We present estimates, obtained by Monte Carlo simulation, of the posterior distributions of the true vegetation proportions at each site and for the entire region for a recently developed RS map - the Land Cover Map 2000 (LCM2000). The motivation for analysing the LCM2000 is the authors' involvement with the Center for Terrestrial Carbon Dynamics (CTCD). The aim of the CTCD is to provide estimates of the terrestrial carbon balances. To this end Woodward and Lomas (2004) developed the Sheffield Dynamic Vegetation Model (SDGVM). One input required by the SDGVM is the type and quantity of vegetation present and the data used in this article is a processed version of the LCM2000 that has been transformed for use by the SDGVM.

The paper proceeds as follows. Section 2 outlines the statistical model at the pixel level and section 3 extends this to inference at the site level. Section 4 provides the details of the LCM2000 map and the final corresponding product required by the SDGVM. Section 5 displays the results as maps of simulated posterior means and standard deviations of the proportions of vegetation classes in England and Wales and also the posterior distributions for the region as a whole. Section 6 concludes the article and presents some directions for future research.

2 Inference at the pixel level

Inference at the pixel level involves identifying the probabilities of the true vegetation classes given the RS map classification (the backward probabilities) and is informed by the confusion matrix. Pixel-specific inference for the backward probabilities is not available from the confusion matrix alone since this information is only provided at the map level. Instead, we assume there is some region-wide characteristic of the remote sensing mechanism itself that contributes to the observation errors (i.e. the RS map misallocating a vegetation class) and model the region-wide forward probabilities as the parameters of a multinomial sample contained in the confusion matrix. The backward probabilities can then be derived from Bayes theorem using the forward probabilities and a prior distribution for the true vegetation classes.

Green and Strawderman (1994) note that confusion matrices are often constructed from a stratified sample, and argue that the matrix therefore comprises multinomial samples from the backward, but not forward, probabilities. While this may be true if the stratification is according to the RS map's vegetation classes, this is not necessarily the case in practice. A confusion matrix may be obtained from a random sample of pixels or by stratifying regionally. We believe that it is the forward probabilities that describe the intrinsic error characteristics of the remote sensing and the algorithm used to deduce vegetation class from the remote sensing data. Bayes' theorem then implies that the backward probabilities will depend on the mix of true vegetation classes (which determines the prior probabilities).

Region-wide forward probabilities

Suppose we have a RS map that reports k vegetation classes at a number of pixels across a region. Also, data is available that compares the RS map classifications with a ground survey at a sampled number of pixels within the region. This data is represented by the confusion matrix, the elements of which, $c_{t't}$, are the counts of sampled pixels over the region whose RS map classification is t' and whose true vegetation class is t. We write

$$\boldsymbol{c}_t = (c_{1't}, c_{2't}, \ldots, c_{k't})^T, \text{ for } t = 1, 2, \ldots, k$$

and

$$\mathbf{C} = (\boldsymbol{c}_1^T, \boldsymbol{c}_2^T, \ldots, \boldsymbol{c}_k^T)^T.$$

We treat $\mathbf{C}$ as a multinomial sample from the region with parameters the forward probabilities. The forward probabilities are written as

$$\lambda_{t't} = P(x = t' | \xi = t) \quad \text{for} \quad t, t' = 1, 2, \ldots, k$$

where ξ is the true vegetation cover and x is the RS map classification. For clarity below we write

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_k^T)^T,$$

where each

$$\boldsymbol{\lambda}_t = (\lambda_{1't}, \lambda_{2't}, \ldots, \lambda_{k't})^T$$

is a vector of probabilities of the RS map classifications given that the true classification is t. The likelihood of $\mathbf{C}$, conditioned on $\boldsymbol{\lambda}$, can be written as

$$P(\mathbf{C} \,|\boldsymbol{\lambda}) \propto \prod_{t=1}^{k} \prod_{t'=1}^{k} (\lambda_{t't})^{c_{t't}}. \tag{1}$$

We assume the $\boldsymbol{\lambda}_t$s are independent of each other and assign each a Dirichlet prior density with parameters $\boldsymbol{\alpha}_t = (\alpha_{1't}, \alpha_{2't}, \ldots, \alpha_{k't})^T$ for $t = 1, 2, \ldots, k$ such that

$$P(\boldsymbol{\lambda}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_k) = \prod_{t=1}^{k} P(\boldsymbol{\lambda}_t|\boldsymbol{\alpha}_t) \tag{2}$$

where

$$P(\boldsymbol{\lambda}_t|\boldsymbol{\alpha}_t) \propto \prod_{t'=1}^{k} \lambda_{t't}^{\alpha_{t't}-1} \quad \text{for } t = 1, 2, \ldots, k. \tag{3}$$

From (1), (2) and (3) the posterior distribution of $\boldsymbol{\lambda}$ is

$$\begin{aligned} P(\boldsymbol{\lambda}|\mathbf{C}) &\propto P(\mathbf{C}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_k) \\ &\propto \prod_{t=1}^{k}\prod_{t'=1}^{k} \left\{(\lambda_{t't})^{c_{t't}+\alpha_{t't}-1}\right\}. \end{aligned}$$

Therefore, conditioned on the data contained in the confusion matrix, the posterior distributions of the region-wide forward probability vectors are independent and Dirichlet with parameters $(\boldsymbol{c}_t + \boldsymbol{\alpha}_t)$ and denoted by

$$\boldsymbol{\lambda}_t|\boldsymbol{c}_t, \boldsymbol{\alpha}_t \sim D_i(\boldsymbol{c}_t + \boldsymbol{\alpha}_t). \tag{4}$$

For the moment assume we are only interested in inference for individual pixels. To quantify the uncertainty of the true vegetation classes given the RS map classification the backward probabilities,

$$\kappa_{tt'} = P(\xi = t|x = t') \quad \text{for} \quad t, t' = 1, 2, \ldots, k,$$

could be modelled as mathematically implied by the forward probabilities and the prior probabilities of the true vegetation class for the entire region,

$$\pi_t = P(\xi = t), \quad \text{for} \quad t = 1, 2, \ldots, k,$$

since together these imply the joint distribution. Whereas Green and Strawderman (1994) use the RS map to obtain the known classification probabilities we use the RS map to derive prior probabilities for the true vegetation classes. We now extend the notion of incorporating a prior distribution for the true vegetation classes to the pixel level and hence model spatial variability within the backward probabilities.

Pixel-specific backward and prior probabilities

Let π_t^p be the prior probability that the true vegetation class is t at pixel p. Since there is no explicit prior knowledge about the π_t^ps we employ a simplistic approach and use the RS map to provide information about the pixel-specific priors. We equate each π_t^p to be the proportion of pixels observed to be in vegetation class t by the RS map in pixel p and in some neighbourhood of pixels around p. We acknowledge that strictly these are not prior probabilities but observed frequencies. However, we believe that this is a good approximation to a more realistic (but computationally much more complex) model in which the prior probabilities were given a random field distribution with nearest neighbour covariance structure. RS maps typically provide pixels across a regular grid and we define a δ nearest neighbour approach such that the neighbourhood forms a square around pixel p and contains a total number of pixels

$$n^p = (2\delta + 1)^2.$$

Then, for each pixel

$$\pi_t^p = \frac{n_t^p}{n^p} \quad t = 1, 2, \ldots, k.$$

where n_t^p is the number of pixels in that neighbourhood classified as vegetation class t. Then the pixel-specific backward probabilities are modelled as

$$\kappa_{tt'}^p = \frac{\lambda_{t't}\pi_t^p}{\sum_{t=1}^k \lambda_{t't}\pi_t^p} \quad \text{for} \quad t, t' = 1, 2, \ldots, k. \tag{5}$$

The posterior distribution of these backward probabilities is then determined by the distribution of the forward probabilities $\lambda_{t't}$. To compute any required inferences it is a simple matter to simulate the forward probabilities from their Dirichlet distributions (4) and to evaluate (5) for each set of simulated $\lambda_{t't}$s.

We have outlined a model for the pixel-specific backward vectors that uses the confusion matrix to model the forward probabilities, as opposed to the backward probabilities, and incorporates spatial variability in the pixel-specific backward probabilities via a 'surrogate' prior evaluated at each pixel from the RS map. The next section describes inference for aggregating the pixel level to the site level.

3 Aggregation to the site level

Although RS maps are able to provide pixels at high resolutions, users of maps often require a coarser resolution. For example, the pixels in the RS map presented and analysed in sections 4 and 5 are required by the CTCD to be aggregated up to a grid size compatible with the grid size on which the SDGVM process model operates. We call these larger grid sizes sites and now describe inference about proportions of the true vegetation classes in a site.

Inference at the site level can only be partially addressed by aggregating the pixel inferences within that site. That is, the expected number of pixels of a certain category in a site is obtainable as the sum of the corresponding pixel-specific backward probabilities. However, we cannot derive an appropriate measure of uncertainty from the single pixels alone since this would ignore any correlation across the site-specific proportions.

It is easy to recognise that, in addition to a region-wide process contributing to the observation errors, some spatial correlation should also exist. The causes of error in RS vegetation classification lie in the detailed characteristics of the vegetation and the conditions (such as weather and incidence angle) of the RS observation. If we consider two neighbouring pixels with the same true vegetation class, then they are likely to have similar vegetation characteristics and the RS observations are likely to have been made under similar conditions. Hence, if one of them is misclassified by the RS map as vegetation class $t' \neq t$ it is more probable that the other will also be misclassified as t'.

Site-specific forward probabilities

To model spatial correlation in these forward probabilities explicitly would be complex and any such model would introduce a number of parameters for which we have essentially no data. Instead we represent the correlation implicitly. A consequence of correlation will be that the proportions of pixels at sites s that are correctly or incorrectly classified in particular ways will be more variable than if they were independent. Without correlation, if we had enough pixels in a given site, then the proportions of all the pixels in that site having true vegetation class t that are observed to be in vegetation class t' will almost certainly be very close to $\lambda_{t't}$; there will be essentially no uncertainty. At the other extreme, if all the pixels in the site with vegetation class t are so highly correlated that they are always observed to be in the same vegetation class, then the resulting proportions of correct and incorrect

classifications will be highly uncertain, being driven by the observation of just one site.

We model this by letting the pixels still be independent, but allowing the $\boldsymbol{\lambda}_t^s$s to vary from the countrywide vector $\boldsymbol{\lambda}_t$. We therefore model the site-specific observation errors as

$$\boldsymbol{\lambda}_t^s | \boldsymbol{\lambda}_t, d \sim D_i(d\boldsymbol{\lambda}_t)$$

for each site in the region. This introduces a single parameter, d, to control the degree of increased variability, and hence implicitly the degree of correlation at the pixel level. Notice that we do not allow for correlation in observations between sites. The assumption is that sites are large enough for it to be reasonable to assume independence, although uncertainty about the common $\boldsymbol{\lambda}_t$ vectors induces some correlation. Then, by redefining the prior probibilities to be site-specific we construct site-specific backward vectors as

$$\kappa_{tt'}^s = \frac{\lambda_{t't}^s \pi_t^s}{\sum_{t=1}^k \lambda_{t't}^s \pi_t^s} \quad \text{for} \quad t, t' = 1, 2, \ldots, k.$$

Before proceeding further we provide a better understanding of the role of d in this model. Consider a site s in which there are n_t^s pixels that are truly in vegetation class t, and let $p_{t't}^s$ be the proportion of these that are observed to be in vegetation class t'. If the pixels are all classified independently with a common forward vector $\boldsymbol{\lambda}_t$, then the number $n_t^s p_{t't}^s$ which are classified as t' will be binomially distributed with mean $n_t^s \lambda_{t't}$ and variance $n_t^s \lambda_{t't}(1-\lambda_{t't})$. Hence $p_{t't}^s$ has mean $\lambda_{t't}$ and variance $\lambda_{t't}(1-\lambda_{t't})/n_t^s$, and for large n_t^s this underlies the statement previously that the proportion will almost certainly be very close to $\lambda_{t't}$. If they are not classified independently, then the variance will be larger. However, we have opted not to model dependence in the classification directly.

Instead, consider the effect of the model which allocates site s forward probabilities, $\boldsymbol{\lambda}_t^s$, the Dirichlet distribution $Di(d\boldsymbol{\lambda}_t)$. Now, conditional on the region-wide forward probabilities, $\lambda_{t't}^s$ has a Beta density with mean $\lambda_{t't}$ and variance $\lambda_{t't}(1-\lambda_{t't})/(d+1)$. Hence, the variance of $p_{t't}^s$ conditioned on

$\lambda_{t't}$ is

$$\begin{aligned}
\mathrm{var}(p^s_{t't} \mid \lambda_{t't}) &= E\left\{\mathrm{var}(p^s_{t't} \mid \lambda^s_{t't}, \lambda_{t't}) \mid \lambda_{t't}\right\} + \mathrm{var}\left\{E(p_{t't} \mid \lambda^s_{t't}, \lambda_{t't}) \mid \lambda_{t't}\right\} \\
&= E\left\{\lambda^s_{t't}(1-\lambda^s_{t't})/n^s_t \mid \lambda_{t't}\right\} + \mathrm{var}\left\{\lambda^s_{t't} \mid \lambda_{t't}\right\} \\
&= \frac{1}{n^s_t}\frac{d}{d+1}\lambda_{t't}(1-\lambda_{t't}) + \frac{1}{d+1}\lambda_{t't}(1-\lambda_{t't}) \\
&= \frac{1}{n^s_t}\frac{d+n^s_t}{d+1}\lambda_{t't}(1-\lambda_{t't}) \ .
\end{aligned}$$

For large n^s_t, this is effectively $\lambda_{t't}(1-\lambda_{t't})/(d+1)$, and so the variance of $p^s_{t't}$ under this model approximates to the variance that one would have if just $d+1$ pixels were independently misclassified. So this model can be thought of as inducing dependence between the pixels in a site so that they divide into $d+1$ independently classified clumps, but that all the pixels in a clump are classified the same. This provides an interpretation of d in terms of the idea of locally dependent classification. The smaller d is, the more variability we introduce between the $\boldsymbol{\lambda}^s_t$ vectors of different pixels, and the greater local dependence is implied.

Site-specific true vegetation class proportions

We now deal with the inference of an aggregation of pixels to the site level and model the true proportions of vegetation classes within a site. Let the true proportion of classification t at site s be γ^s_t, for $t = 1, 2, \ldots, k$ and $s = 1, 2, \ldots, S$. It is the uncertainty in the γ^s_t s that we wish to characterise. Let the proportion of vegetation class t' at site s according to the RS map be $g^s_{t'}$. Letting n^s be the total number of the RS map pixels in the site, this corresponds to

$$n^s_{t'} = n^s g^s_{t'}$$

actual pixels observed to be in vegetation class t' by the RS map.

For site s the vector of proportions of true vegetation classes is written as

$$\boldsymbol{\gamma}^s = (\gamma^s_1, \gamma^s_2, \ldots, \gamma^s_k)^T,$$

and the site-specific backward vectors as

$$\boldsymbol{\kappa}^s_{t'} = (\kappa^s_{1t'}, \kappa_{2t'} \ldots, \kappa^s_{kt'})^T, \ \text{ for } \ t' = 1, 2, \ldots, k.$$

If we know each $\boldsymbol{\kappa}^s_{t'}$ then it is natural to suppose that the individual pixels in site s are independently observed by the RS map, and hence that the

true vegetation class for each pixel can be independently inferred using the relevant $\boldsymbol{\kappa}^s_{t'}$ vector. Then, conditional on the relevant $\boldsymbol{\kappa}^s_{t'}$s, $\boldsymbol{\gamma}^s$ is a linear combination of k independent multinomial distributions:

$$\boldsymbol{\gamma}^s|\boldsymbol{\kappa}^s_1, \boldsymbol{\kappa}^s_2, \ldots, \boldsymbol{\kappa}^s_k \sim \frac{1}{n^s}\sum_{t'=1}^{k} M(n^s_{t'}, \boldsymbol{\kappa}^s_{t'}),$$

where $n^s_{tt'}$ is the number of pixels in site s observed by the RS map to be vegetation class t' but are actually vegetation class t and $M(n^s_{t'}, \boldsymbol{\kappa}^s_{t'})$ is the multinomial distribution over $n^s_{t'}$ observations with probability vector $\boldsymbol{\kappa}^s_{t'}$. Thus,

$$P(\boldsymbol{\gamma}^s|\boldsymbol{\kappa}^s_1, \boldsymbol{\kappa}^s_2, \ldots, \boldsymbol{\kappa}^s_k) = \frac{1}{n^s}\sum_{t'=1}^{k}\left\{\frac{n^s_{t'}!}{\prod_{t=1}^{k} n^s_{tt'}!}\prod_{t=1}^{k}(\kappa^s_{tt'})^{n^s_{tt'}}\right\}.$$

Aggregation across sites

The range over which we would expect correlations to exist in the forward probabilities is limited, and for aggregation over a large area it becomes difficult to assess a suitable value for d if this is regarded as a single site. We therefore propose defining sites of such a size that we expect correlation within sites but correlation between sites will be weak enough to ignore. Thus, the region of interest is divided into sites of this size, and the aggregation treats these as independent given the $\lambda_{t't}$s. This is the approach used in section 4, where we are interested in the whole of England and Wales, but work with sites of size one-sixth of a degree.

The results presented in section 5 below contain the estimated posterior means and standard deviations of each $\boldsymbol{\gamma}^s$ for individual sites and for the whole region. To compute these quantities we perform the following Monte Carlo simulation scheme:

For $g = 1, 2, \ldots, 10000$

(1) Draw $\boldsymbol{\lambda_t} \sim D(\mathbf{c}_t + \boldsymbol{\alpha})$ for $t = 1, 2, \ldots, k$

For $s = 1, 2, \ldots, S$

(2) Draw $\boldsymbol{\lambda}^s_t \sim D(d\boldsymbol{\lambda}_t)$ for $t = 1, 2, \ldots, k$

(3) Calculate $\boldsymbol{\kappa}^s_{t'}$, for $t' = 1, 2, \ldots, k$ from $\boldsymbol{\pi}^s$ and $\boldsymbol{\lambda}^s_t$.

(4) Draw $\boldsymbol{\gamma}^s \sim \frac{1}{n^s}\sum_{t'=1}^{k} M(n^s_{t'}, \boldsymbol{\kappa}^s_{t'})$

(5) Sum γ^ss over the whole region.

end

end.

4 LCM2000 and SDGVM

The data analysed in this paper is the LCM2000 converted into a format suitable for use by the SDGVM process model. The LCM2000 is a RS map produced by the UK Centre for Ecology and Hydrology (CEH) (Haines-Young et al., 2000). The base spatial resolution of the LCM2000 is a 25m $\times$ 25m spatial grid covering Great Britain, each pixel of which contains one of 26 subclasses of vegetation types classified from spectral information collected from various orbital sensors. The data we analyse is the LCM2000 for England and Wales and converted into a format required by the SDGVM. To be compatible with the SDGVM the data have been manipulated in two ways.

First, a simplified vegetation class structure is required for SDGVM and the 26 vegetation types are grouped into 5 plant functional types (PFTs) under the assumption that a species assigned to a given PFT will exhibit the same behaviour in terrestrial carbon flux processes. The PFTs represented within the SDGVM for England and Wales are deciduous broadleaf (DcBl), evergreen needleleaf (EvNl), Grassland and Crop, with an additional noncontributing class representing urban areas and bare ground that are considered to take no part in the natural carbon cycle. This PFT is denoted by Bare. DcBl consists of both trees and shrubs that lose their foliage every year. EvNl consists predominantly of coniferous trees. The Grassland and Crop PFTs differ in that crops are harvested at the end of the year, thus having an artificially shortened leaf lifespan, and the harvested biomass is removed from the system. The aggregation from 26 vegetation classes to 5 PFTs not only reduces the computational load (an important consideration given the high computing demands of SDGVM simulations) but can also ease interpretation of the calculations. Although aggregation from the original land cover scheme to the PFT level negates some errors in the data set (those mislassifications between land cover types that belong to the same PFT), some confusion between the resulting PFTs still persists. By way of example, areas of shrubland which should contribute to the DcBl PFT are often erroneously classified as one of the classes that makes up the Grassland PFT. For details of the classification legend from vegetation types

SDGVM PFT	LCM Land cover class
DcBl	broadleaf woodland, dwarf shrub heath, open dwarf shrub heath
Crop	arable cereals, arable horticulture, non-rotational horticulture
Grassland	improved grassland, set aside grassland, natural grass, calcareous grass, acid grass, bracken, fen, marsh, swamp, bog, montane habitats, salt marsh
EvNl	Coniferous Woodland
Bare	sea, inland water, inland bare ground, estuary, suburb and rural development, littoral rock, littoral sediment, continuous urban, supra-littoral rock, supra-littoral sediment

Table 1: Classification legend from PFTs (left column) to vegetation classes (right column).

to PFTs see Table 1.

Second, the spatial resolution of the LCM2000 measurements has been altered. LCM2000 is upscaled from the 25m $\times$ 25m resolution to a $1/6^{th}$ degree resolution and reported as proportions. The original measurements, in the British Ordnance Survey grid, were reprojected into geographic longitudes and latitudes and then aggregated to $1/6^{th}$ degree resolution sites in order to match the grid size of the climate database used in SDGVM. The proportions of the various PFT types were preserved in this process and a record of the total number of pixels (or counts) contributing to each $1/6^{th}$ degree grid cell was kept. The area of a grid cell measured in degrees varies as a function of latitude, so the count per grid cell is not constant.

The final result of the LCM2000 product used in our analysis is 707 sites representing $1/6^{th}$ degree resolution for England and Wales with the LCM2000 recordings of proportions of the 5 PFTs in each site (and the total number of pixels in each site). For clarity, we only report the results in detail for the PFTs DcBl, Crop, Grassland and EvNl, although Bare is included in the ac-

tual model. Figure 1 displays maps of proportions of DcBl, Crop, Grassland and EvNl as classified by LCM2000 for SDGVM sites across England and Wales. Figure 1 shows that DcBl and EvNl occupy a relatively little proportion of land cover across England and Wales, while Crops and Grasslands dominate the region with Crops predominant in the east and Grasslands in the west. Spatial correlation is obvious in all four maps. Note, there is an area in the south-east where none of DcBl, Crop, Grassland and EvNl have a large proportion. This is the greater London metropolitan region and is dominated by Bare.

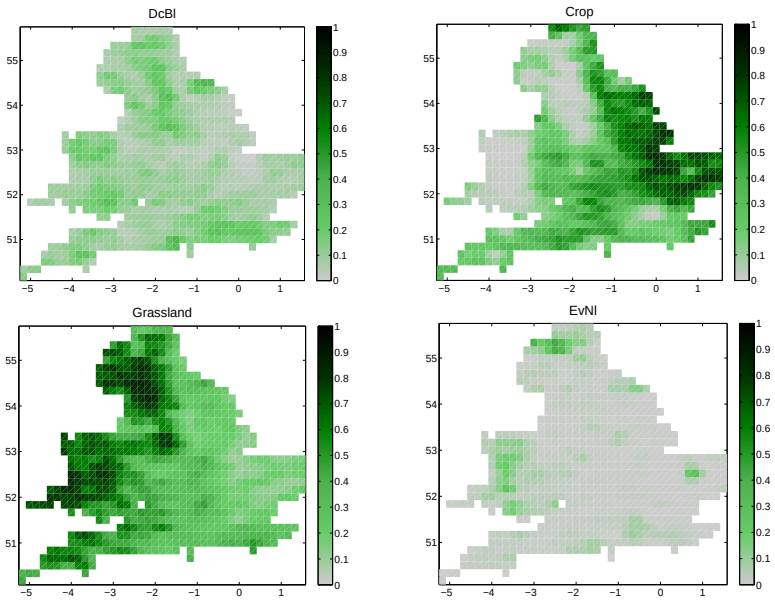

Figure 1: Maps of the proportions of PFTs DcBl, Crop, Grassland and EvNl for England and Wales as recorded by LCM2000.

Finally, we describe the confusion matrix. Fuller et al. (2002) report a confusion matrix of the 26 vegetation classes between LCM2000 and the Country Side field survey (CS2000) in the year 2000 for England and Wales. To obtain the confusion matrix Fuller et al. (2002) stratify England and Wales into regions of vegetation classes and individual confusion matrices are calculated for each vegetation stratum. These confusion matrices are combined to give the matrix reported in Fuller et al. (2002), each matrix contributing a weighting according to its extent in England and Wales. We acknowledge that due to the stratified sampling procedure the data contained in the confusion matrix is not strictly a multinomial sample across England and

Wales. Further manipulation is required since we require the measurements to be aggregated to PFTs as opposed to vegetation classes. Table 2 reports the aggregated confusion matrix calculated from that in Fuller et al. (2002) into PFT form.

	CS2000				
LCM2000	DcBl	EvNl	Grassland	Crop	Bare
DcBl	66	3	19	4	5
EvNl	8	20	1	0	0
Grassland	31	5	356	22	15
Crop	7	1	41	289	9
Bare	2	0	3	8	81

Table 2: Confusion matrix between LCM2000 and CS2000 for England and Wales.

5 Results

We now present the results for the LCM2000 described in section 4 and used for the SDGVM. In this case, we have 5 PFT or vegetation classes reported, for $s = 1, 2, \ldots, 707$ sites across England and Wales. To run the model it is first necessary to specify values for the hyperpior parameters of the region-wide forward vectors, the neighbourhood structure for site-specific prior probabilities and the parameter, d, that governs the degree of variability across the site-specific forward vectors.

Since we have negligible prior knowledge of the region-wide forward vectors we allocate them an uninformative prior and set

$$\boldsymbol{\alpha}_t = (1, 1, 1, 1, 1)^T, \quad \text{for} \quad t = 1, 2, \ldots, 5.$$

We choose $d = 1$, which corresponds to thinking that all the pixels of a given PFT in a site are classified as just two clumps. This represents a belief in rather strong dependence in the classification over a site, so that $d = 1$ is perhaps too small. However, we prefer to err on the side of assuming too much uncertainty than too little.

Finally, the nearest neighbour parameter is set to $\delta = 5$ sites. We note that there is a small discrepancy introduced between inference about individual pixels and inference about sites. The neighbourhood that we define for a

site is not identical to the neighbourhood that we would have for individual pixels within the site, even if the pixel neighbourhood corresponds to δ sites. Hence the individual pixels would have prior probabilities that vary over the site and differ from the prior probabilities used for the site. However, we are using sites covering many pixels and will ignore this small discrepancy.

Figure 2 displays the plots of π_t^s for PFTs DcBl, Grassland, Crop and EvNl. Figure 2 captures the general pattern of the LCM2000 maps in Figure 1, but suggests the priors should not dominate our inference.

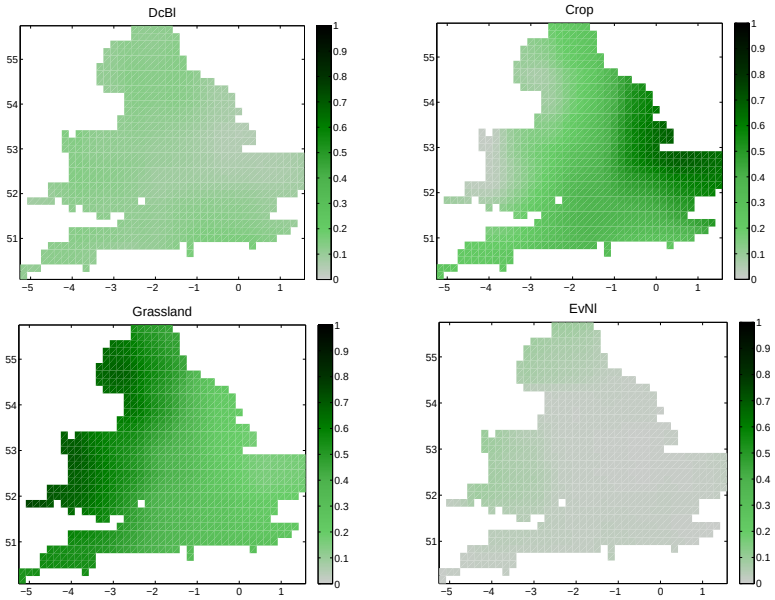

Figure 2: Maps of the priors of PFTs DcBl, Crop, Grassland and EvNl (clockwise from top left respectively).

The results reported in this section are the simulated posterior means and standard deviations (henceforth referred to simply as the means and standard deviations) of the true PFT proportions for all sites, $\boldsymbol{\gamma}^s$. Also, we report histograms of the estimated posterior distributions of the average of the true proportions over all sites:

$$\bar{\boldsymbol{\gamma}} = \frac{1}{707}\sum_{s=1}^{707}\boldsymbol{\gamma}^s.$$

As in the section 4, for clarity we only provide a detailed discussion of the

results relating to the PFTs DcBl, Grassland, Crop and EvNl although we provide a brief summary of the results for Bare at the end of this section.

Figure 3 displays the plots of the means of PFTs DcBl, Crop, Grassland and EvNl for England and Wales. The means for DcBl, EvNl and Crop appear similar to the LCM2000 maps in Figure 1. In particular, the LCM2000 measurements of DcBl are at their highest in regions of the north and south of Wales, areas in the north-west and central regions of England (corresponding to the Lake and Peak Districts), and in south-eastern England and are in agreement with the LCM2000 recordings. The LCM2000 recordings and the means also agree that EvNl is more abundant in the Lake District and south Wales. Additionally, Thetford forest in East Anglia (the area east of the meridian and north of 52 dgrees), the largest lowland pine forest in the UK, is clearly identified as EvNl. The means of Crop closely follow the LCM2000 map which records Crop as prevalent in the north-east of England, the mid-east coast (East Anglia) and south England although the means appear to be smaller compared to the LCM2000 map in north-east England and the Midlands. The means of PFT Grassland follow the same pattern as the LCM2000 map and indicate that Wales and the Peak and Lake Districts are mainly populated by Grassland. However, Figure 3 implies the proportions of Grasslands in these regions are also smaller than the LCM2000 measurements. In general, our model appears to estimate values of DcBl and EvNl higher than reported by the LCM2000 and estimates values of Grassland and Crop lower than reported by the LCM2000. We return to this issue and its reason when discussing the results relating to the region-wide PFTs below.

A major contribution of our model is to measure the uncertainty in the true proportions of PFT for England and Wales. Figure 4 displays the standard deviations, or uncertainty levels, of DcBl, Grassland, Crop and EvNl. The highest standard deviations for DcBl and EvNl are 0.1024 and 0.0986, both at longitude and latitude (55.25,-2.583), in the far north-east of England, suggesting a high level of uncertainty in classifying these two PFTs at this site. The highest standard deviation for Grassland is 0.1296 at longitude and latitude (53.25,-1.917), in the Peak District, and the highest standard deviation of Crop is 0.1418 at longitude and latitude (55.75,-2.083), in the far north of England. Figure 4 also suggests that higher standard deviations follow the general pattern of higher means for DcBl, Grassland and EvNl but not for Crop. For EvNl both the means and standard deviations are relatively high in the far north of England and Wales. DcBl typically has higher standard deviations in Wales and the Peak and Lake Districts. Similarly for

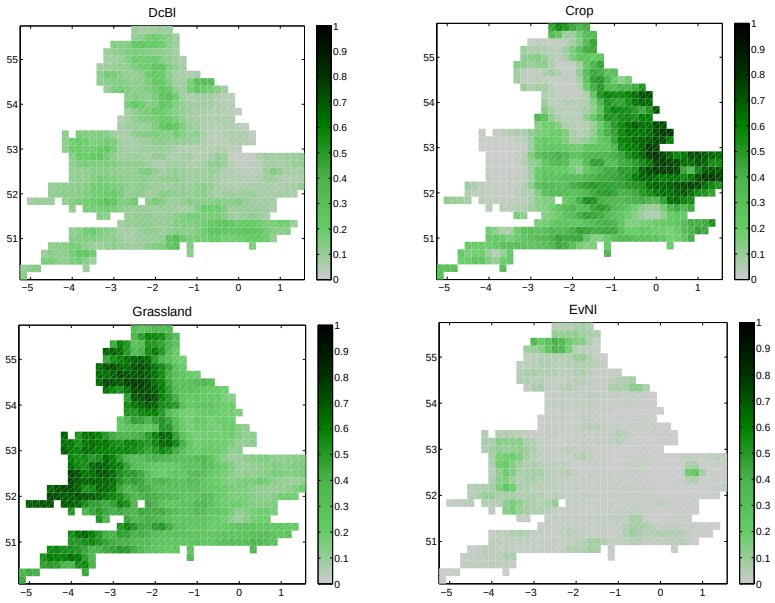

Figure 3: Maps of the simulated posterior means of PFTs DcBl, Crop, Grassland and EvNl in England and Wales.

Grassland, higher standard deviations and means appear to coincide in the regions of Wales and the Peak and Lake Districts. The larger uncertainties in the proportions in these regions may arise due to the difficulty in assigning vegetation classes to either DcBl or Grassland as mentioned in section 4. Crop, however, has comparatively similar standard deviations in East Anglia (where its means are highest), central England and into the south of England. The Midlands, and the region due south of the Midlands, is an area where LCM2000 allocates higher levels of Crop and comparable levels of DcBl, Grassland and EvNl, although Grassland is recorded by LCM2000 as dominating the bordering region to the west of the Midlands. Figure 4 suggests that the allocation of classes to PFTs Crop and Grasslands in this region have high uncertainty. Figure 4 also highlights uncertainty in the allocation of PFT Crop close to London.

Figure 5 reports histograms of the estimated posterior distributions of the region-wide proportions, $\bar{\gamma}$, of the average proportions of DcBl, Crop, Grassland and EvNl over England and Wales. Figure 5 also shows the values of the average proportions over England and Wales as recorded in LCM2000, which are 0.1102, 0.3209, 0.4216, 0.0316 for DcBl, Crop, Grassland and EvNl respectively. For each PFT the posterior distribution shows a shift away from

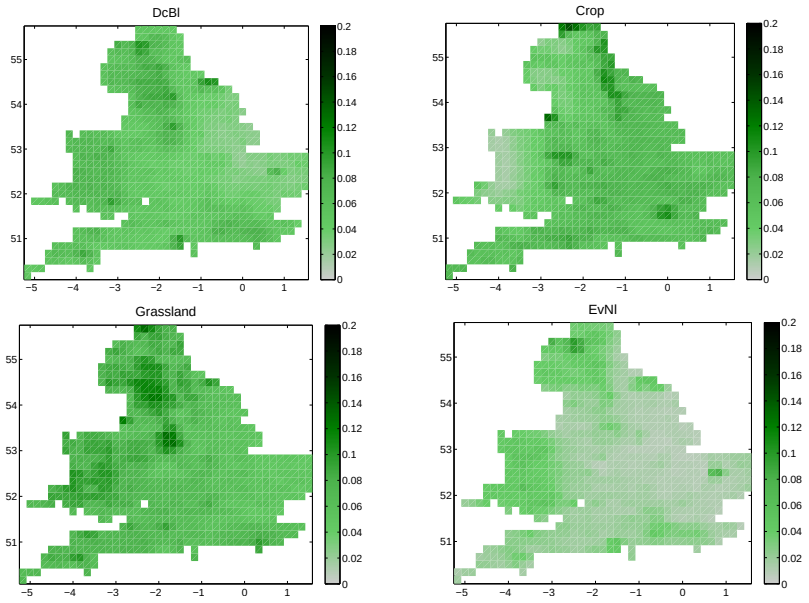

Figure 4: Maps of the simulated posterior standard deviations of PFTs DcBl, Crop, Grassland and EvNl in England and Wales.

the LCM2000 recordings. For Grassland and Crop the shift is negative and for DcBl and EvNl it is positive. This suggests that LCM2000 allocates Grassland and Crop more often than exist in reality and allocates DcBl and EvNl less often than should exist in reality. Table 3 contains the reason for this result. Table 3 reports the marginal counts for LCM2000 for all PFTs and the marginal counts for the ground survey data from CS2000. For Grassland and Crop, LCM2000 over allocates the number of sites for these PFTs in England and Wales compared to the CS2000. For DcBl LCM2000 under allocates the number of sites compared to CS2000. LCM2000 and CS2000 report the same values of EvNl. Another way to see this is to note the asymmetries in the confusion matrix. For instance, Grassland is quite rarely allocated erroneously by LCM2000 to DcBl, whereas there is a much higher probability for DcBl to be incorrectly allocated to Grassland.

Although we have not shown maps for PFT Bare, it was included in the model and we give a brief summary of the results now. The means of Bare closely followed the pattern of the LCM2000 recordings, which suggest Bare is relatively sparse across England and Wales except around major urban areas, most notably London, Birmingham and Manchester. The standard deviations tend to be relative high in similar areas. Over all of England and

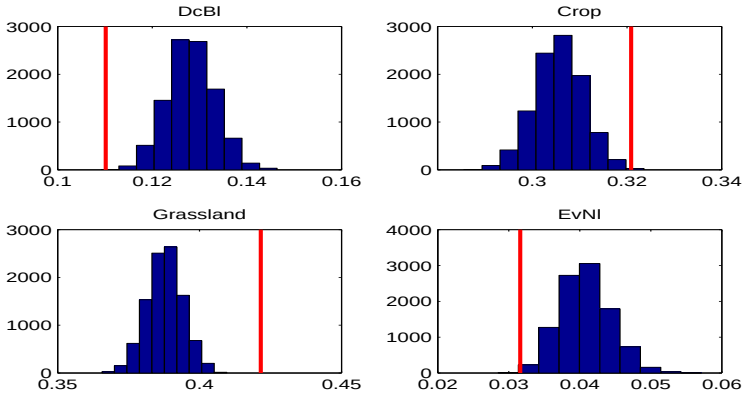

Figure 5: Histograms of the simulated posterior distributions of the overall PFTs DcBl, Crop, Grassland and EvNl for England and Wales. The vertical line shows the overall LCM2000 recording.

	DcBl	Crop	Grassland	EvNl	Bare
CS2000	114	323	420	29	110
LCM2000	97	347	429	29	94

Table 3: Marginal counts for England and Wales of the PFTs DcBl, EvNl, Grassland, Crop and Bare as recorded by CS2000 (top row) and LCM2000 (bottom row).

Wales, our model allocates more Bare than is suggested by the LCM2000. An examination at Table 3 explains why this is the case with total counts from the CS2000 survey higher than the total counts of LCM2000.

6 Conclusion and directions for future research

In this article we have provided a general framework to model uncertainty in satellite derived land cover maps, when the final product is reported as site-specific proportions and have data contained in a confusion matrix with which to calibrate the map. The method describes a multinomial model with misclassified probabilities that includes a simple spatial correlation structure suited to the case where little data is available regarding spatial information. Our method allows us to quantify uncertainty for the true site-specific land

cover types, and for the entire region, and this is new to the remote sensing community.

For illustrative purposes we have applied our method to a land cover map (the LCM2000) that is of recent importance to the scientific community. The results identify specific sites where the LCM2000 has difficulty in classifying specific vegetation types. In particular, the LCM2000 exhibits relatively high uncertainty between the DcBl and Grasland PFTs, especially in Wales and the Peak and Lake Districts. The problematic nature of discriminating between land cover types that are spectrally similarly (in this case shrubs and grasslands) in satellite data has been noted for several of the major remote sensing land cover products (McCallum et al., 2006). Being able to provide well-founded estimates of site-specific uncertainty with such land cover classifications increases their value not only to an end user but also to the remote sensing community itself. The results also show that in general the LCM2000 overestimates the proportions of Grasslands and Crops and underestimates the proportions of DcBl, EvNl and Bare across England and Wales.

Future directions for applications and research include the following. First, satellite derived land cover maps are also available at global scales. The method in this article has been applied to England and Wales, a substantially smaller region. Application and analysis of global maps may be of interest to the remote sensing community and various users of such maps. Modelling the spatial variation and correlation of land cover proportions across the globe may require more information than has been made available to us and incorporating external data or multiple confusion matrices for this purpose may be of interest. Second, we assumed the data in the confusion matrix were a multinomial sample over the region. This assumption was not strictly true, and an interesting direction for future research would be to model the information in the confusion matrix more accurately. Finally, we may elicit expert prior information from the remote sensing community better construct the priors for the observation errors of the RS map.

References

Foody, G. M. (2002), "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, 80, 185–201.

— (2007), "The evaluation and comparison of thematic maps derived from

remote sensing," *7th Internations Symposium on Spatial Accuracy Assessment and Environmental Sciences*.

Fuller, R. M., Smith, G. M., Sanderson, J. M., H. R. A., and Thomson, A. G. (2002), "Countryside Survery 2000, Module 7. Land cover map 2000. Final Report," *Technical report*.

Giri, C., Zhu, Z., and Reed, B. (2005), "A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets," *Remote Sensing of Environment*, 94, 123–132.

Green, E. J. and Strawderman, W. E. (1994), "Determining the accuracy of thematic maps," *The Statistician*, 43.

Hansen, M. and Reed, B. (2000), "A comparison of the IGBP DISCover and University of Maryland 1 km global land cover products," *International Journal of Remote Sensing*, 21, 1365–1373.

Jung, M., Henkel, K., Herold, M., and Churkina, G. (2006), "Exploiting synergies ofglobal land cover products for carbon cycle modelling," *Remote Sensing of Environment*, 101.

Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., Naumov, S., De Miranda, E. E., Di Bella, C. M., Ordoyne, C., Kopin, Y., and Roy, P. (2006), "Validation of the Global Land Cover 2000 Map," *IEEE Transactions on Geoscience and Remote Sensing*, 44.

McCallum, I., Obersteiner, M., Nilsson, S., and Shivdenko, A. (2006), "A spatial comparison of four satellite derived 1 km global land cover datasets," *Applied Earth Observation and Geoinformation*, 8, 246–255.

Stehman, S. V., Czaplewski, R. L., Nusser, L. Y., and Zhu, Z. (2000), "Combining accuracy assesement of land cover maps with environmental modelling," *Environmental Monitoring and Assessment*, 64, 115–126.

Woodward, F. and Lomas, M. R. (2004), "Vegetation dynamics - simulating responses to climatic change," *Biological Reviews*, 79.

Acknowledgements This project was funded by the NERC Center for Terrestrial Carbon Dynamics.