1

Application of Bayesian mixture models to satellite images and estimating the risk of fire-ant incursion in the identified geographical cluster

Insha Ullah and Kerrie Mengersen

School of Mathematical Sciences, ARC Centre of Mathematical and Statistical Frontiers, Science and Engineering Faculty, Queensland University of Technology

Summary. Bayesian non-parametric mixture models have found great success in the statistical practice of identifying latent clusters in data. However, fitting such models can be computationally intensive and of less practical use when it comes to tall datasets, such as Landsat imagery. To overcome this issue, we propose to obtain multiple samples from data using stratified random sampling to enforce adequate representation in each sample from sub-populations that may exist in data. The non-parametric model is then fitted to each sample dataset independently to obtain posterior estimates. Label correspondence across multiple estimates is achieved using multivariate component densities of a chosen reference partition followed by pooling multiple posterior estimates to form a consensus posterior inference. The labels for pixels in the entire image are inferred using the conditional posterior distribution given pooled estimates, thereby substantially reducing the computational time and memory requirement.

The method is tested on Landsat images from the Brisbane region in Australia, which were compiled as a part of the national program for the eradication of the imported red fire-ant that was launched in September 2001 and which continues to the present date. The aim is to estimate the risk of fire-ant incursion in each of the identified geographical cluster so that the eradication program focuses on high risk areas.

1.1 Introduction

Imported red fire-ant have been a cause for concern in Brisbane, Australia. They are an invasive species and their spread could have serious social, environmental and economic impacts throughout Australia. They were first discovered in February 2001 in surrounding areas of the Port of Brisbane but are believed to have been imported a couple of decades prior to 2001. Despite the eradication program, which was launched in September 2001, spread from the initial Brisbane infestation has led to infestations around the greater Brisbane area. Isolated incursions have been found even beyond the greater Brisbane area.

In order to prioritize the use of the surveillance budget and to promote better decision making, modelling is performed to estimate the risk of fire-ant incursion in each area so that the eradication program focuses on high risk areas. As part of the surveillance program the colony locations were recorded prior to their eradication. The analysis of imagery data in combination with the location observations helps identify the preferred habitats of fire-ants

(Spring and Cacho, 2015). However, the field data are presence-only data (Guillera-Arroita et al, 2015): information on observed absences is not available and it is not reasonable to assume that areas where the pest has not been observed are absences since they are known to have very wide potential habitat. Hence supervised learning models such as logistic regression to predict occurrence probability are too arbitrary for the presence-only data and are not justifiable in this situation (Hastie and Fithian, 2013).

In light of the above, unsupervised clustering methods are more appealing for these presence-only data. These methods involves dividing the whole region into smaller clusters based on available covariate data and determining the possibility of presence in each cluster. In the context of our case study, the covariates are obtained from the satellite imagery and the presence if interest is fire-ants. However, this requires model selection that is the pre-specification of the number of clusters, K.

Dirichlet process Gaussian mixture models (DPGMMs) have been widely adopted as a data-driven cluster analysis technique. The main attraction of these models lies in sidestepping model selection by assuming that data are generated from a distribution that has a potentially infinite number of components. However, for a limited amount of data, only a finite number of components is detected and an appropriate value for the number of components has to be determined directly from data in a Bayesian manner (hence the term, 'data-driven'). These infinite, non-parametric representations allow the models to grow in size to accommodate the complexity of the data dynamically. However, they are computationally demanding and do not scale well to the satellite imagery data, each image of which is usually made up of millions of pixels. This is because they need to iterate through the full dataset at each iteration of the MCMC algorithm (see, e.g., Bardenet et al, 2017). The computational time per iteration increases with the increasing sizes of the datasets.

How to scale Bayesian mixture models up to massive data comprises a significant proportion of contemporary statistical research. One way to speed up computations is to use graphics processing units (see, e.g., Lee et al, 2010; Suchard et al, 2010) and parallel programming approaches (see, e.g., Guha et al, 2012; Chang and Fisher III, 2013; Williamson et al, 2013). Relatively less computationally demanding methods for fitting the mixture models include approximate Bayesian inference techniques such as variational inference (McGrory and Titterington, 2007; Ormerod and Wand, 2010; Hoffman et al, 2013; Blei et al, 2017) and approximate Bayesian computation (Marin et al, 2012; Moores et al, 2015). Other strategies to speed up computations are the sampling based approaches. This is adopted by Huang and Gelman (2005) who partition the data at random and perform MCMC independently on each subset to draw samples from the posterior given the data subset. They suggested methods based on normal approximation and importance re-sampling to make consensus posteriors. A similar idea has been proposed in Scott et al (2016) with a different rule for combining posterior draws. Manolopoulou et al (2010) improve inference about the parameters of the component of interest in the mixture model. An initial sub-sample is analysed to guide selection from targeted components in a sequential manner using Sequential Monte Carlo sampling. This approach depends critically on an adequate representation of the component of interest in the initial random sample. However, in a massive dataset, a low probability component of interest is likely to escape the initial random sample, which will lead to unreliable inference.

In satellite imagery, most of the data are replications. For example, all water pixels should appear similar while pixels from the land covered with the same crop should produce similar observations. Thus, inference based on a stratified random sample of the data should be representative of the whole image. This is possible in the case of supervised learning where the training data is labelled a priori. In the case of unsupervised learning one could use a computationally faster method such as k-means clustering to first label the data. These labels

could then be used to obtain a stratified random sample (hence enforcing representation from each sub-population). A much more reliable inference based on a stratified random sample can be obtained using more flexible and sophisticated mixture models which allow incorporation of additional available information and also take into account the correlation between variables rather than imposing a simple model, such as k-means clustering, just because of computational problems.

In this article, we fit a Bayesian mixture model to stratified samples that have been selected from pre-clustered images. Importantly, we make use of the strengths of two clustering methods: the computationally less demanding method of k-means clustering and the more sophisticated DPGMMs, which not only account for correlations between variables, but also learn K in a data-driven fashion. Our method is explained in Sections 1.2 and 1.3 and applied to a case study in Section 1.4 and 1.5. Conclusions are presented in Section 1.6.

1.2 Dirichlet process Gaussian mixture models

Assume that we are interested in clustering real-valued observations contained in $X = (x_1, \ldots, x_n)$, where x_i is a p-dimensional sample realization made independently over n objects. Denoting the p-dimensional Gaussian density by $\mathcal{N}(\cdot)$, a mixture of K Gaussian components takes the form

$$f(x|\theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\theta_k), \tag{1.1}$$

where $\theta_k = \{\mu_k, \Sigma_k\}$ contains the unknown mean vector μ_k and the covariance matrix Σ_k is associated with component k. The parameters $\pi = \{\pi_1, \ldots, \pi_K\}$ are the unknown mixing proportion, which satisfies $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

In Dirichlet process Gaussian mixture models (Rasmussen, 2000), the number of components K is an unknown parameter without any upper bound and inference algorithms are used to facilitate learning K from the observed data. Therefore, with every new data observation, there is a chance for the emergence of an additional component.

Define a latent indicator z_i, $i = 1, \ldots, n$, such that the prior probability of assigning a particular observation x_i to a cluster k is $p(z_i = k|\pi) = \pi_k$. Given the cluster assignment indicator z_i and the prior distribution G on the component parameters, the model in (1.1) can be expressed as:

$$\begin{aligned} x|z_i = k, \theta_k &\sim \mathcal{N}(x|\theta_k), \\ \theta_k|G &\sim G, \\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0), \end{aligned}$$

where G_0 is the base distribution for the Dirichlet process prior such that $E(G) = G_0$ and α is the concentration parameter. Integrating out the infinite dimensional G from the posterior allows the application of Gibbs sampling to DPGMM (Escobar, 1994; MacEachern, 1994; Escobar and West, 1995). By integrating out G, the predictive distribution for a component parameter follows a Pólya urn scheme (Blackwell and MacQueen, 1973)

$$\theta_k|\theta_1, \ldots, \theta_{k-1} \sim \frac{\alpha}{k-1+\alpha} G_0 + \frac{1}{k-1+\alpha} \sum_{i=1}^{k-1} \delta_{\theta_i}(\cdot).$$

Specifying a Gamma prior over the Dirichlet concentration parameter α, $\alpha \sim Ga(\eta_1, \eta_2)$, allows the drawing of posterior inference about the number of components, K.

Simpler and more efficient methods have been developed to fit the DPGMM. Consider two independent random variables $V_k \sim Beta(1,\alpha)$ and $\theta_k \sim G_0$, for $k = \{1,2,\ldots\}$. The stick-breaking process formulation of G is such that

$$\pi_k = \begin{cases} V_k & (k=1) \\ V_k \prod_{i=1}^{k-1}(1-V_i) & (k>1) \end{cases},$$

and

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot),$$

where $\delta_{\theta_i}(\cdot)$ is a discrete measure concentrated at θ_k (Sethuraman, 1994). In practice, however, the Dirichlet process is truncated by fixing K to a large number such that the number of active clusters remains far less than K (Ishwaran and James, 2002). A truncated Dirichlet process is achieved by letting $V_K = 1$, which also ensures that $\sum_{k=1}^{K} \pi_k = 1$. The base distribution G_0 is specified as a bivariate normal-inverse Wishart

$$G_0(\mu_k, \Sigma_k) = \mathcal{N}(\mu_k|\mu_0, a_0\Sigma_k) IW(\Sigma_k|s_0, S_0),$$

where μ_0 is the prior mean, a_0 is a scaling constant to control variability of μ around μ_0, s_0 denotes the degrees of freedom and S_0 represent our prior belief about the covariances among variables. The data generating process can be described as follows:

1. For $k = 1,\ldots,K$: draw $V_k|\alpha \sim Beta(1,\alpha)$ and $\theta_k|G_0 \sim G_0$.
2. For the nth data point: draw $z_i|V_1,\ldots,V_k \sim Mult(\pi)$ and draw $x_i|z_i = k, \theta_k \sim \mathcal{N}(x|\theta_k)$

1.2.1 Blocked Gibbs sampling scheme to fit DPGMM

A blocked Gibbs sampler (Ishwaran and James, 2002) avoids marginalization over the prior G, thus allowing G to be directly involved in the Gibbs sampling scheme. The algorithm is described as follows:

1. Update z by multinomial sampling with probabilities

$$p(z_i = k|x, \pi, \theta) \propto \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

2. Update the stick breaking variable V by independently sampling from a beta distribution

$$p(V|x) \sim Beta\left(1 + n_k, \alpha + \sum_{i=k+1}^{K} n_i\right),$$

   where $V_k = 1$ and n_k is the number of observations in component k. Obtain π by setting $\pi_1 = V_1$ and $\pi_k = V_k \prod_{i=1}^{k-1}(1-V_i)$ for $k > 1$.
3. Update α by sampling independently from

$$p(\alpha|V) \sim Ga\left(\eta_1 + K - 1, \eta_2 - \sum_{i=1}^{K-1} \log(1-V_i)\right),$$

4. Update Σ_k by sampling from

$$p(\Sigma_k|x,z) \sim IW(\Sigma_k|s_k,S_k),$$

where

$$s_k = s_0 + n_k,$$
$$S_k = S_0 + \sum_{z_i=k}(x_i-\bar{x}_k)(x_i-\bar{x}_k)^t + \frac{n_k}{1+n_k a_0}(\bar{x}_k-\mu_0)(\bar{x}_k-\mu_0)^t$$

and

$$\bar{x}_k = \frac{1}{n_k}\sum_{z_i=k} x_i.$$

5. Update μ_k by sampling from

$$p(\mu_k|x,z,\Sigma_k) \sim \mathcal{N}(\mu_k|m_k,a_k\Sigma_k),$$

where

$$m_k = \frac{a_0\mu_0 + n_k\bar{x}_k}{a_0+n_k}$$

and

$$a_k = \frac{a_0}{1+a_0 n_k}.$$

1.3 The method

As noted earlier, in satellite imagery, many of the pixels are exact replicates or at least provide similar information (different up to a level of noise). To reduce computational time and memory storage requirements, it is sufficient to obtain an adequate representation from each group of similar observations rather than analysing data that include tens of thousands of duplicate copies of observations. The full dataset can be mapped onto the cluster obtained based on sample data. Many authors have resorted to this option; for example, Kulkarni and Callan (2010) used a 0.1% sample of 500 million documents and extended results to cluster the rest of the documents. However, if a sample is selected at random, it is likely that some smaller clusters of interest are not sampled. This eventually will produce results that are biased towards a small number of larger clusters, which may in turn lead to lower quality clusters (De Vries et al, 2015).

We use a similar sampling based strategy but select a sample of size n in a way that potentially ensures representation from very small clusters that may exist in the data. This is made possible by first arbitrarily clustering the N pixels of the whole image into a large number, say C, of smaller clusters (C is much larger than the actual number of clusters one can expect in the whole image) using computationally faster k-means clustering (MacQueen et al, 1967), which is a popular clustering algorithm because of its scalability and efficiency in large data sets (Jain, 2010). The pre-clustered image is then sampled using stratified sampling with proportional allocation; that is, a sample of size $n_i = n(N_i/N)$ is chosen from the ith cluster, where $i = 1,\ldots,C$ and N_i denotes the number of pixels in ith cluster. Note that the total sample size, $n = n_1 + \ldots + n_C$, should be large enough to contain a reasonable number of observations from the smallest cluster obtained via k-means clustering. Another way to ensure adequate representation from the smallest cluster is, for example, by increasing each n_i by the size of smallest cluster, say n_s, that is $n_i = n(N_i/N) + n_s$ or by a fraction of n_s if n_s is large. The sample of size n thus obtained is clustered using DPGMM.

To control for sampling variation we obtain M samples each of size n using the above process and apply DPGMM independently to each sample. The label correspondence across mixture components from the multiple samples is created using multivariate component densities of a chosen reference partition and the mean vectors from the rest of M partitions as data. This is followed by pooling posterior estimates based on multiple samples to form consensus posterior estimates. Denote the mth stratified random sample obtained from X by $X_{(m)}$, $m = 1,\ldots,M$, the respective sample-data posterior by $p(\theta|X_{(m)})$ and the sample-posterior estimate of the kth component parameter by $\hat{\theta}_{k(m)} = \{\hat{\mu}_{k(m)}, \widehat{\Sigma}_{k(m)}\}$. Then the pooled estimates of the parameters of the kth component are obtained using the following identities (Huang and Gelman, 2005):

$$\hat{\mu}_k = \widehat{\Sigma}_k \left(\sum_{m=1}^{M} \widehat{\Sigma}_{k(m)}^{-1} \hat{\mu}_{k(m)} \right)$$

and

$$\widehat{\Sigma}_k = \left(\sum_{m=1}^{M} \widehat{\Sigma}_{k(m)}^{-1} \right)^{-1}.$$

The labels for the N pixels in the entire image are inferred using the conditional posterior distribution given the pooled estimates.

1.4 The Data

Since the launch of the fire-ant eradication program in September 2001, data have been collected on the location of each colony that has been found. The dataset used in this case study comprises 17717 locations where nests of fire-ants were identified during the years 2001-2013. These locations are indicated on a Google image snap-shot provided in Figure 1.1. The proportion of colonies identified for each year are provided in Figure 1.2. A sudden rise in the number of identified nests during 2009-2010 and then a drop back to normal in the following years is surprising. There may be a number of factors responsible for this phenomenon, such as flooding events, changes in surveillance processes or major developmental projects, but definitive reasons for it still require further investigation.

A Landsat image is also available for each year of the study. These were acquired on days of low cloud coverage, generally in the period between May and September, most commonly in July. These images were chosen as being typical winter images, and sufficiently near to the date required to be included in the winter planning period for summer surveillance. The images were converted into workable data files using the 'raster' package (Hijmans et al, 2016) in R. Note that we use 6 Landsat spectral bands (variables): visible blue, visible green, visible red, near infrared, middle infrared, and thermal infrared. The Landsat variables were centred at mean zero and scaled to a unit variance.

We also used R for the substantive statistical analysis. To solve the k-means problem, we used the algorithm in Hartigan and Wong (1979), which is a default option in the R function *kmeans()*, available from the 'stats' package. Since it is recommended to make repeated runs with different random starting points and choose the run that gives the minimum within-class variance, we used 8 random starting points in our analysis. Note that the function *kmeans()* also allows to specify multiple random starting points. A larger number of starting points, however, increases computation cost, particularly when the number of clusters is larger, which is due to multiple runs of the algorithm. We avoided this by using the parallel processing facility in R provided by *foreach* loop from the 'foreach' package. Since the k-means clustering is

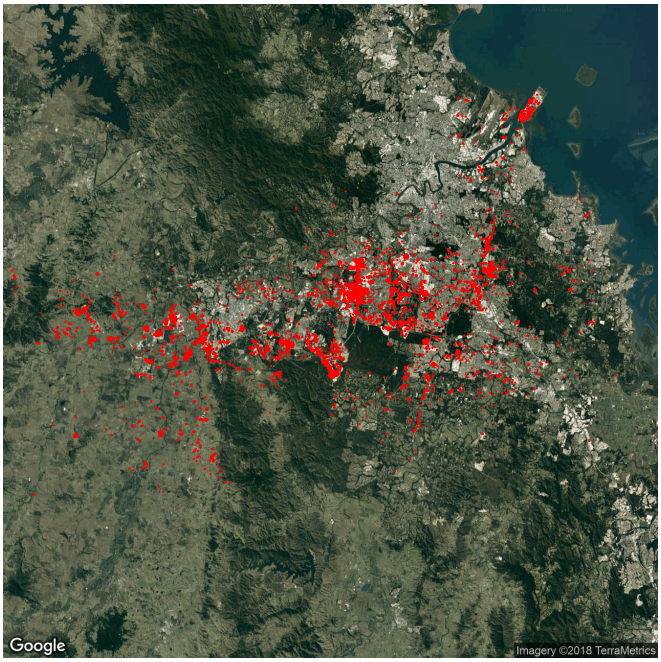

Fig. 1.1. Google image snapshot of the study area and the observed location of fire-ant colonies (indicated by red dots) over the study period 2001-2013.

intended to include small strata in order to acquire a representative sample (rather than final clustering), we did not find noticeable differences in terms of visual interpretation when used a single random starting point.

To fit a DPGMM, we translated Matlab code, available at `http://ftp.stat.duke.edu/WorkingPapers/09-26.html`, into R code (for details about Matlab codes, see, Manolopoulou et al, 2010). Due to having no formal convergence guarantee, we did experiments with different images (considered in this study) to decide on the total number of iterations of the blocked Gibbs sampling algorithm including the burn-in iterations. In our experiments we found that the algorithm provided visually interpretable solutions after 1000 iterations (our final results can be visualized and checked with Google maps) and we did not see any noticeable difference when used with a larger number of iterations (30000 iterations excluding 5000 burn-in iterations). Therefore, we used 10,000 iterations of a blocked Gibbs sampler excluding the first 2,000 burn-in iterations in all the analyses whose results are shown here. The overall computation time averaged over the 13 images considered in this study was 8 hours and 11 minutes when we set $n = 100000$ and $M = 10$. This computation time increased to 14 hours and 5 minutes for $n = 200000$. Note that we used the high performance computing facility at the Queensland University of Technology for our computations which has 2.6Ghz processors with 251Gb memory.

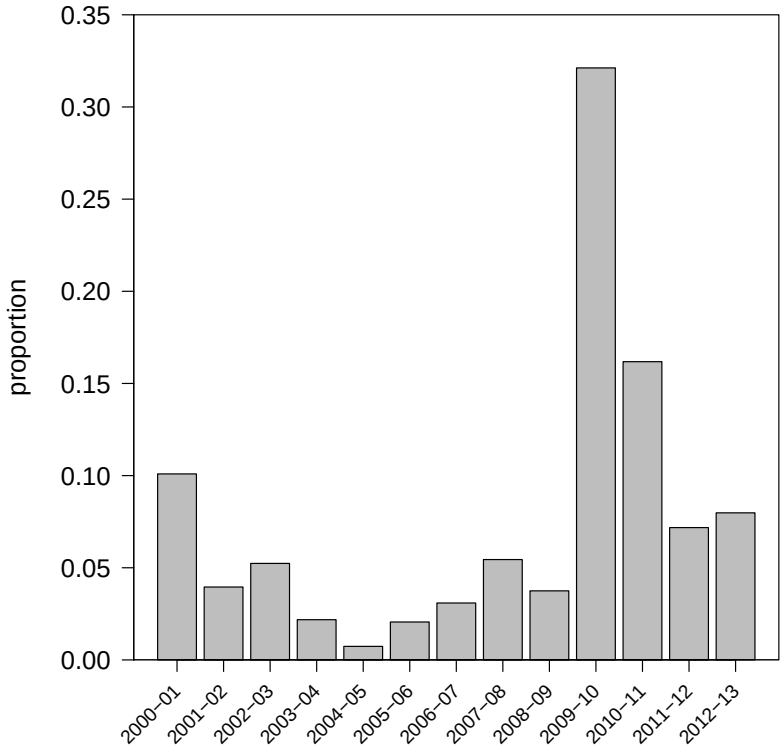

Fig. 1.2. Proportions of fire-ant colonies detected each year from 2000-2013.

1.5 Analysis and results

The aim of the analyses was to find out about the potential characteristics of fire-ants' preferred habitats by classifying the satellite images. The images were first clustered arbitrarily into large number of clusters using k-means clustering. We tried $C = 50, 100, 150, 200, 250, 300$ and show the results for $C = 100$, since we did not notice significant improvement for larger values of C in terms of visual interpretation. Ten stratified samples ($M = 10$) each of size $n = 100000$ were selected using proportional allocations. The DPGMM was then fitted to each sample independently in parallel and the pooled estimates were obtained by combining the posterior estimates across the multiple samples. The results based on different samples were very consistent apart from the labels correspondence issue. For example, the component-1 represented water in the partitioning based on the sample-1 but it represented forest areas in the partitioning based on the sample-2. We dealt with this problem by using density of a chosen reference partition (the one that gave the maximum number of components) and considering the mean vectors from the rest of the partitions as data. In this way, the water component in all partitions had a high probability to correspond with the water component

in the reference partition; therefore, we re-labelled them the same across different partitions. In our analysis we used $M = 10$ because of the availability of high performance computing facility. However, we did not notice any visually interpretable change when we used a smaller value of $M = 5$ on a personal 8-core Intel platform with processor speed 3.4 GHz and 16 GB of memory. The whole process for this experiment took 5 hours with 5000 iterations of a blocked Gibbs sampler excluding 1000 burn-in iterations. The reference partition was chosen, among M partitions, as the one with the largest number of components. The labels for the whole image are inferred using obtained posterior distribution given the pooled estimates. This process was performed independently for each image from the year 2001 to 2013. We tested a range of values of n (between 10000 and 300000, inclusive) and found that the number of components and their structure did not change (in terms of visual interpretation) as we increased the value of n beyond 100000. Therefore, we set $n = 100000$ for all the results shown here.

The classification based on the images from years 2003 and 2010 are shown, respectively, in Figures 1.3 and 1.4. The proportion of observed fire-ants identified in each cluster are presented in Tables 1.1 and 1.2. Note that each of these tables is based on a single year image; however, the proportions of the observed fire-ants for the rest of the study period that falls in a particular class are also provided for prediction purpose. The figures for other years and their respective tables are diverted to the supplementary material due to the compatibility of the results across different years.

The final number of components per image varied across different years but stayed below 36. The variation in the number of components was mainly due to a number of very small clusters that each contained less than 1% of the total pixels. However, the number of components that consisted of more than 1% of the pixels were quite consistent across different years and remained around 20. Some of the variation in the number of clusters across different years could possibly be attributed to the time of the day the image was acquired. For example, the mountainous area was broken into a various number of components in images from different years possibly because of shadows (see components 10 and 12 in Figure 1.3 and components 10, 13 and 18 in Figure 1.4). In the image from 2001, clouds over the mountains were well separated (image not shown here). Other variations are because of the changes in the landscape over time. For example, Wyaralong Dam cannot be seen in Figure 1.3 but can be seen in Figure 1.4 since it was built in 2009-2010.

The large components were materially similar across different years and were visually interpretable into different land cover classes, namely, hills, forest, water, residential areas, warehouses, roads, parks and play grounds, plain areas with natural non-forest vegetation (scrub-land) and some impervious surfaces, and new development sites or land with recent deforestation. Other smaller clusters (each consisting of less than 1% of the pixels and visually not interpretable) are found to be of less interest and are therefore merged together in the figures.

The water component in the image was always well separated from the rest of the components and was often partitioned into shallow and deep water (see components 6, 18 and 20 in Figure 1.3 and components 12 and 14 in Figure 1.4). Although this component is not of interest to us, it helps in identifying and interpreting other components. The parks and playgrounds were found to be consistently at risk of infestation over time (see components 17 and 9, respectively in Tables 1.1 and Table 1.2). The components that represent the scrub-land with thinner forest and the land with natural vegetation are generally the largest by area and are found to be consistently at risk of fire-ant incursion (see components 1 and 2 in Tables 1.1; components 1, 3, and 4 in Table 1.2); in particular, the incursion in component 3 of Table 1.2 has increased over time.

The residential area (see component 3 in Table 1.1 and components 7 and 19 in Tables 1.2) including the areas with commercial buildings (see component 21 in Table 1.1 and component 15 in Tables 1.2) were found to be at high risk in the initial years when the eradication program started. However, the risk of incursion declined soon after the launch of eradication program in this class, which probably shows that the eradication program has been more effective in the residential areas. A potential reason could be swift reporting once the incursion has been observed. The risk of incursion increased in the components that represent agricultural fields and in the components that represents impervious surfaces and scrub-land (see, respectively, component 16 and 7 in Table 1.1).

The components with forest areas were found to be consistently at low risk of fire-ant incursion (see component 4 in Tables 1.1 and component 2 1.2). Similarly, the mountainous areas were also found to be at low risk (see components 10 and 12 in Tables 1.1 and components 10, 13 and 18 in Tables 1.2).

As mentioned above, Tables 1.1 and 1.2 also present the proportions of fire-ant nests observed in the years other than the one in which the analysed image was acquired. In general, the classes with high proportions of fire-ant nests in the image year calibrate well with the proportions in a few years that follow. For example, in Table 1.1 areas in component 1 were at risk of fire-ant incursions in 2003 (contained 16.2% of the observed nests) remained at similar risk in the following year (component 1 contained 16.7% of the observed nests in 2004). The risk of infestation in component 3 of Table 1.1 is consistent for a few years following 2003 (2004 is an exception that contained 23.4% of the observed nests) with a gradual decreasing trend in the later years. Similarly, in Table 1.2, which is based on classification of image from 2011, component 3 was found to be at highest risk in 2011 (contained 49.2% of the observed nests) and remained at high risk in the following two years (contained 29.8% in 2012 and 30.9% in 2013). The risk of incursion in component 7 was almost doubled in the following years (contained 10.1% of the observed nests in 2011, 23.6% of the observed nests in 2012 and 20.5% of the observed nests in 2013). The component 15 contained 6.5% in 2011 and 9.2% in 2012. Some of the potential factors for anomalous changes could possibly be attributed climatic events such as floods or drought.

The above results indicate that image classification provides useful information for operational projects. The classification can be produced routinely at a low cost, which when combined with the observed data helps in learning about the high risk areas. These high risk areas could be prioritized in order to satisfy budgetary constraints. For example, as mentioned above the infestation of fire-ants has declined in residential areas over a period of 13 years and probably reflects the success of the eradication program but has increased in other components such as scrub-land and agricultural fields that needs to be prioritized in future.

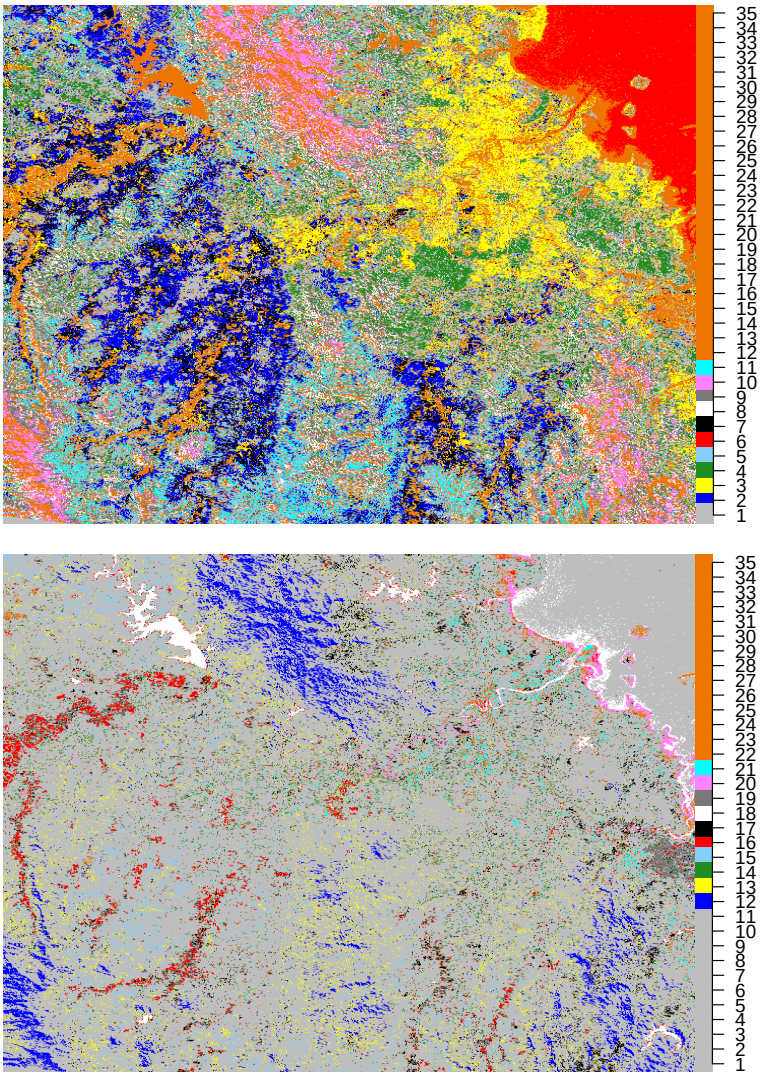

Fig. 1.3. Cluster analysis of satellite image of the Brisbane area taken in 2004. For clarity, some of the clusters are merged together, in dark-orange (top) and gray colours (bottom), and the results are presented in two plots: (top panel) 1: scrub-land with thinner forest, 2: scrub-land with natural vegetation, 3: residential area 4: Dense forest, 5: mountainous areas with scarce forest, 6: water, 7: mix of impervious surfaces and scrub-land, 8: mountainous areas, 9: mountainous areas, 10: hills and forest, 11: mountainous areas with scars forest ; (bottom panel) 12: hilly areas, 13: hard for visual interpretation, 14: residential area, 15: hard for visual interpretation, 16: fields, 17: mix of parks, playgrounds and grassland, 18: shallow water, 19: fields with crops, 20: seashore and shallow water, and 21: commercial buildings. Cluster 22 to cluster 35 are too small to be visually interpreted.

Table 1.1. The percentages of fire-ant colonies identified in each of the spatial components (shown in Figure 1.3) over the period of 13 years conditional on the image acquired in 2004 (highlighted in gray). The C.No indicates component numbers corresponding to the component numbers in Figure 1.3. The C.Size (in %) indicates the size of a cluster relative to image. The clusters are sorted in descending order with respect to their sizes.

C.No	C.Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1	17.48	13.9	11.2	16.2	16.7	20.6	16.8	5.5	9.5	10.2	11.8	10.8	18.5	19.9
2	9.77	1.1	1.9	3.3	10.7	3.1	0.5	4.7	9.8	11.5	36.1	4.3	10.7	8.7
3	8.75	53.1	47.7	50.8	23.4	53.3	56.4	35.9	33.8	26.3	9.6	30.2	18.6	14.4
4	8.06	1.0	1.9	1.1	1.0	0.8	1.6	0.7	1.2	3.9	2.2	6.2	12.9	13.0
5	6.02	1.7	1.3	1.4	5.7	1.5	0.8	0.7	2.0	4.0	5.4	2.6	2.8	3.9
6	5.58	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	5.49	1.3	2.1	2.9	27.1	0.0	0.3	14.4	18.6	19.6	21.4	12.8	11.2	7.4
8	5.37	0.0	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.3	1.4	1.5
9	4.65	0.1	0.0	0.3	0.3	0.8	0.0	0.0	0.0	0.2	0.0	0.1	0.5	1.3
10	4.16	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4
11	4.04	0.3	0.6	1.0	2.7	0.0	0.0	0.0	0.2	0.8	0.3	0.2	0.6	1.1
12	3.78	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
13	2.94	0.2	1.0	1.1	0.3	0.0	0.0	0.0	0.0	0.2	0.5	1.0	1.4	1.4
14	2.37	9.9	10.7	10.1	4.4	5.4	5.9	3.1	3.4	6.5	4.1	4.2	4.3	6.9
15	2.01	0.0	0.0	0.1	0.5	0.8	0.0	0.0	0.0	0.2	0.6	0.3	0.9	0.9
16	1.99	1.1	3.6	1.5	1.0	0.8	7.6	25.9	15.0	6.6	5.0	22.3	8.7	12.0
17	1.86	0.7	2.7	2.0	1.6	3.1	2.2	0.5	0.9	1.7	0.5	1.0	2.2	2.0
18	1.78	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	1.09	0.7	0.3	0.6	0.6	2.3	2.2	0.5	1.0	4.5	0.2	0.7	1.5	0.9
20	0.64	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2
21	0.61	9.5	10.9	3.7	2.1	4.6	1.6	1.9	1.0	1.1	1.5	1.7	1.3	1.8
22	0.48	4.6	3.1	2.3	0.5	0.0	2.2	3.8	2.5	0.6	0.1	0.1	0.9	0.6
23	0.36	0.1	0.3	0.3	0.0	3.1	1.5	1.8	0.3	1.6	0.1	0.7	1.2	1.0
24	0.34	0.2	0.1	0.3	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.1	0.0	0.1
25	0.23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.2	0.2
26	0.06	0.0	0.0	0.0	0.5	0.0	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
27	0.03	0.3	0.3	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1
28	0.03	0.2	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
31	0.01	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.1
32	0.00	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
33	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total incursions		1788	701	928	387	130	365	547	965	664	5690	2866	1272	1414

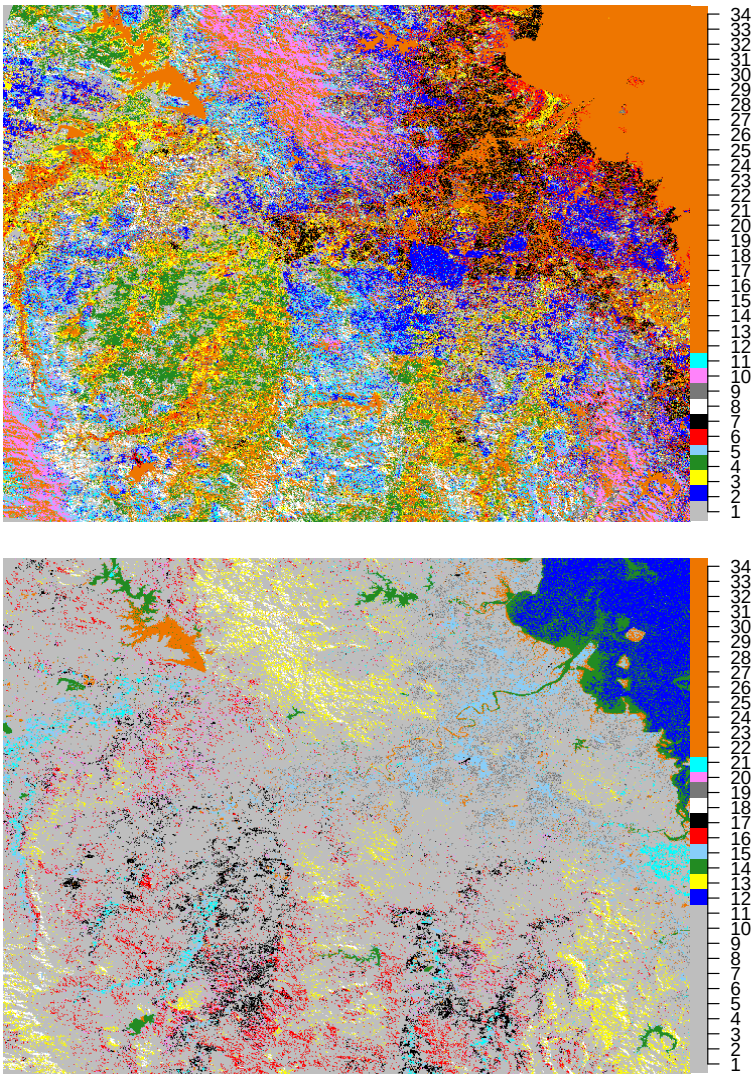

Fig. 1.4. Cluster analysis of satellite image of the Brisbane area taken in 2011. For clarity, some of the clusters are merged together, in dark-orange (top) and gray colours (bottom), and the results are presented in two plots: (top panel) 1: scrub-land with thinner forest, 2: dense forest, 3: impervious surfaces and scrub-land, 4: scrub-land with natural vegetation, 5: mountainous areas with thinner forest, 6: forest, 7: residential area, 8: mountainous areas scarce forest, 9: parks and playgrounds, 10: hills with dense forest, 11: mountainous areas with scarce forest ; (bottom panel) 12: deep water, 13: hilly areas with dense forest, 14: shallow water, 15: mix of commercial buildings and fields without crops, 16: hard for visual interpretation, 17: hard for visual interpretation, 18: hills, 19: residential areas, 20: hard for visual interpretation, and 21: fields with crops. Cluster 22 to cluster 34 are too small to be visually interpreted.

Table 1.2. The percentages of fire-ant colonies identified in each of the spatial components (shown in Figure 1.4) over the period of 13 years conditional on the image acquired in 2011 (highlighted in gray). The C.No indicates component numbers corresponding to the component numbers in Figure 1.4. The C.Size (in %) indicates the size of a cluster relative to image. The clusters are sorted in descending order with respect to their sizes.

C.No	C.Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1	12.05	2.5	3.9	6.0	14.8	6.9	8.9	3.6	3.8	2.4	14.7	3.2	1.8	4.5
2	9.59	1.4	1.6	2.4	4.5	3.1	1.4	0.0	0.2	0.0	0.1	0.1	0.3	0.5
3	7.93	9.0	10.2	9.0	30.1	21.3	7.4	27.9	24.2	40.1	24.2	49.2	29.8	30.9
4	7.40	0.2	0.4	0.1	3.8	0.0	0.0	4.7	0.5	4.1	27.7	11.3	0.9	3.9
5	6.01	0.1	0.3	0.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8
6	5.28	10.3	9.8	11.0	4.8	7.7	7.9	1.9	5.6	7.0	2.3	2.0	1.2	2.5
7	5.27	27.6	27.4	28.1	12.7	13.9	23.4	13.6	16.6	14.2	5.0	10.1	23.6	20.5
8	4.85	0.2	0.3	0.1	0.8	0.0	2.1	1.6	0.4	0.2	0.5	0.5	0.6	0.7
9	4.79	6.7	7.4	9.3	7.2	15.3	23.1	3.2	4.3	13.5	5.3	5.1	6.3	6.3
10	4.72	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	4.36	0.1	0.1	0.1	0.3	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.1	0.1
12	4.10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	3.65	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	3.51	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	3.38	31.6	29.6	21.0	10.5	14.3	18.7	27.2	31.5	6.2	3.1	6.5	9.2	9.4
16	3.12	0.2	0.0	0.4	1.3	1.5	0.0	0.7	0.0	0.2	1.6	0.9	0.1	0.6
17	2.81	0.1	1.1	0.2	4.5	0.0	0.0	4.5	1.5	3.8	11.1	2.2	3.4	1.5
18	1.66	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	1.43	8.7	5.9	9.6	1.8	7.7	5.7	1.9	1.4	3.2	0.9	2.2	2.2	1.7
20	1.25	0.1	0.1	0.2	0.0	0.8	0.8	0.4	0.6	1.5	2.3	0.9	0.2	0.6
21	0.99	0.3	0.4	0.1	0.8	0.0	0.0	0.0	0.7	0.5	0.1	0.5	0.7	0.9
22	0.48	0.2	0.0	0.1	0.0	0.0	0.0	0.0	1.0	0.0	0.1	0.0	0.0	0.0
23	0.44	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.27	0.2	0.6	0.4	1.0	3.8	0.0	6.6	2.4	2.4	0.4	3.5	10.9	9.9
25	0.24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.24	0.2	0.6	0.3	0.0	3.7	0.5	2.2	4.8	0.9	0.2	1.6	8.5	4.6
27	0.12	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	0.04	0.5	0.4	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
29	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.2	0.1	0.0	0.1
30	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
33	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total incursions		1788	701	928	387	130	365	547	965	664	5690	2866	1272	1414

1.6 Conclusions and recommendations

DPGMM are computationally prohibitive for tall datasets such as satellite imagery data. We used computationally faster k-means clustering to pre-cluster the data into a large number of clusters and obtain stratified samples of suitable sizes to ensure representation from very

small clusters. These samples are partitioned using DPGMM and the posterior estimates are pooled across multiple samples. The labels for all the pixels in the image are predicted using the posterior distribution given the pooled estimates of components parameters. The proposed method enables classification of a dataset with millions of observations in a matter of hours and minutes.

We clustered satellite images to identify the land cover classes that are at high, medium, and low risk of infestation of fire-ants. Residential areas were found to be at a high risk of infestation in the initial years of the eradication program that started in 2001. However, the risk of incursion in the residential area declined within a few years after the start of eradication program. The scrub-land with natural vegetation and the classes that represent agricultural fields have seen high incursions in the later half of the study period. Parks, playgrounds and some impervious surfaces were also found to be at risk of infestation.

The overall analyses show that clustering satellite images could be very useful to make rational decisions about where the eradication program needs to focus next. For example, the eradication program is found to be successful in the residential areas perhaps due to prompt response from the residents and businesses. However, as mentioned earlier, other clusters such as scrub-land with natural vegetation, agricultural fields, parks, playgrounds and roads have seen high incursions in the later years of the study periods. Since having fire-ants at home or at the commercial places are more threatening as compared to having encountered them at the park or on a road, people are more likely to report them when they are posing a threat to their personal comfort. This makes it important to create awareness among the public about these high risk areas, in order to better support the collective effort to detect and manage this pest.

Note that this is was an initial exploratory study and has some limitations pertaining to it. First, the information from both the images and fire-ant incursions data have spatio-temporal structures, we fitted a separate model for each year and calculated the proportions of presence-only data observed in that year in the clusters found by the model. A more principled way would be to embed the presence-only data in the fitted model. This would require a hierarchical model that in one level performs the clustering based on the spectral bands and in the other level uses the clusters as predictors in a model for the presence-only data. One need to account for spatial dependence in such model, which could potentially play an important role in the problem being tackled. Second, we did not account for the temporal effects in our model and calculated the proportions of the observed presence-only data for other years assuming no significant temporal changes in the land-cover over a period of few years. A more sophisticated model that take into account the temporal variation in the land cover would be required. We leave these extensions for future research.

Acknowledgement. This research was supported by an ARC Australian Laureate Fellowship for project, Bayesian Learning for Decision Making in the Big Data Era under Grant no. FL150100150. The authors also acknowledge the support of the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and the support of QUT's high-performance computing and Research Support (HPC) group.

References

Bardenet R, Doucet A, Holmes C (2017) On markov chain monte carlo methods for tall data. The Journal of Machine Learning Research 18(1):1515–1557

Blackwell D, MacQueen JB (1973) Ferguson distributions via pólya urn schemes. The annals of statistics pp 353–355

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. Journal of the American Statistical Association 112(518):859–877

Chang J, Fisher III JW (2013) Parallel sampling of dp mixture models using sub-cluster splits. In: Advances in Neural Information Processing Systems, pp 620–628

De Vries CM, De Vine L, Geva S, Nayak R (2015) Parallel streaming signature em-tree: A clustering algorithm for web scale applications. In: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp 216–226

Escobar MD (1994) Estimating normal means with a dirichlet process prior. Journal of the American Statistical Association 89(425):268–277

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. Journal of the american statistical association 90(430):577–588

Guha S, Hafen R, Rounds J, Xia J, Li J, Xi B, Cleveland WS (2012) Large complex data: divide and recombine (d&r) with rhipe. Stat 1(1):53–67

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA (2015) Is my species distribution model fit for purpose? matching data and models to applications. Global Ecology and Biogeography 24(3):276–292

Hartigan JA, Wong MA (1979) Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 28(1):100–108

Hastie T, Fithian W (2013) Inference from presence-only data; the ongoing controversy. Ecography 36(8):864–867

Hijmans RJ, van Etten J, Cheng J, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, Bevan A, Racine EB, Shortridge A, et al (2016) Package 'raster'. R package https://cran r-project org/web/packages/raster/index html (accessed 1 October 2016)

Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. The Journal of Machine Learning Research 14(1):1303–1347

Huang Z, Gelman A (2005) Sampling for bayesian computation with large datasets. Technical Report

Ishwaran H, James LF (2002) Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. Journal of computational and graphical statistics 11(3):508–532

Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern recognition letters 31(8):651–666

Kulkarni A, Callan J (2010) Document allocation policies for selective searching of distributed indexes. In: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, pp 449–458

Lee A, Yau C, Giles MB, Doucet A, Holmes CC (2010) On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. Journal of computational and graphical statistics 19(4):769–789

MacEachern SN (1994) Estimating normal means with a conjugate style dirichlet process prior. Communications in Statistics-Simulation and Computation 23(3):727–741

MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA., vol 1, pp 281–297

Manolopoulou I, Chan C, West M (2010) Selection sampling from large data sets for targeted inference in mixture modeling. Bayesian analysis (Online) 5(3):1

Marin JM, Pudlo P, Robert CP, Ryder RJ (2012) Approximate bayesian computational methods. Statistics and Computing pp 1–14

McGrory CA, Titterington D (2007) Variational approximations in bayesian model selection for finite mixture distributions. Computational Statistics & Data Analysis 51(11):5352–5367

Moores MT, Drovandi CC, Mengersen K, Robert CP (2015) Pre-processing for approximate bayesian computation in image analysis. Statistics and Computing 25(1):23–33

Ormerod JT, Wand MP (2010) Explaining variational approximations. The American Statistician 64(2):140–153

Rasmussen CE (2000) The infinite gaussian mixture model. In: Advances in neural information processing systems, pp 554–560

Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE (2016) Bayes and big data: The consensus monte carlo algorithm. International Journal of Management Science and Engineering Management 11(2):78–88

Sethuraman J (1994) A constructive definition of dirichlet priors. Statistica sinica pp 639–650

Spring D, Cacho OJ (2015) Estimating eradication probabilities and trade-offs for decision analysis in invasive species eradication programs. Biological invasions 17(1):191–204

Suchard MA, Wang Q, Chan C, Frelinger J, Cron A, West M (2010) Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. Journal of computational and graphical statistics 19(2):419–438

Williamson S, Dubey A, Xing EP (2013) Parallel markov chain monte carlo for nonparametric mixture models. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp 98–106