Journal of Big Data

METHODOLOGY Open Access

Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data

Insha Ullah* and Kerrie Mengersen

*Correspondence:
inshahullah03@yahoo.com
Science and Engineering Faculty, Queensland University of Technology, 2 George St, Brisbane, QLD 4000, Australia

Abstract

Due to their conceptual simplicity and flexibility, non-parametric mixture models are widely used to identify latent clusters in data. However, when it comes to Big Data, such as Landsat imagery, such model fitting is computationally prohibitive. To overcome this issue, we fit Bayesian non-parametric models to pre-smoothed data, thereby reducing the computational time from days to minutes, while disregarding little of the useful information. Tree based clustering is used to partition the clusters into smaller and smaller clusters in order to identify clusters of high, medium and low interest. The tree-based clustering method is applied to Landsat images from the Brisbane region, which were the actual sources of motivation for development of the method. The images are taken as a part of the red imported fire-ant eradication program that was launched in September 2001 and which is funded by all Australian states and territories, along with the federal government. To satisfy budgetary constraints, modelling is performed to estimate the risk of fire-ant incursion in each cluster so that the eradication program focuses on high risk clusters. The likelihood of containment is successfully derived by combining the fieldwork survey data with the results obtained from the proposed method.

Keywords: Dirichlet process mixture models, *k*-means clustering, Tree-based clustering, Satellite imagery data, Fire-ants habitat, One-class support vector machine

Introduction

Red imported fire-ants 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 eradication 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 [1]. However, the field data are presence-only data [2]: information on unobserved fire-ant presence is missing from the data. Supervised learning models such as logistic regression to predict occurrence probability are too arbitrary for the presence-only data and are seldom justifiable [3].

The most appealing method for the presence-only data here is to divide the whole region into smaller clusters based on satellite imagery data and determine the possibility of fire-ant containment in each cluster. However, the clustering methods usually result in a small number of large clusters, some of which can be further partitioned. This requires a tree-like implementation of a clustering algorithm which in turn requires model selection (the pre-specification of the number of clusters, K) at each node of the tree.

The most appealing method for the presence-only data here is to divide the whole region into smaller clusters based on satellite imagery data and determine the possibility of fire-ant containment in each cluster. However, the clustering methods usually result in a small number of large clusters, some of which can be further partitioned. This requires a tree-like implementation of a clustering algorithm. One could choose a computationally faster method, such as k-means clustering [4], but a tree-like implementation of k-means clustering requires model selection (the pre-specification of the number of clusters, K) at each node of the tree. In addition, k-means clustering has been criticised for its susceptibility to converge to a local optimum. The mean-shift algorithm [5] do not have the model selection problem, however, it is computationally expensive and may not be suitable for clustering very large datasets.

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., [6]). 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., [7]) and parallel programming approaches (see, e.g., [8–10]). Relatively less computationally demanding methods for fitting the mixture models include approximate Bayesian inference techniques such as variational inference [11–14] and approximate Bayesian computation [15, 16]. Huang and Gelman [17] 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. Another strategy to speed up computations is to improve inference about the parameters of the component of interest in the mixture model. This is adopted by [18], where an initial sub-sample is analysed to

guide selection from targeted components in a sequential manner using Sequential Monte Carlo sampling. To make it work, an adequate representation of the component of interest is important 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.

Often, in massive datasets, most of the data provide similar information. Consider, for example, satellite imagery where observations from the parks (and playgrounds) in an urban area will look similar except for some noise (anything that makes a park different from other parks). Similarly, large water-bodies may contribute millions of repeated observations. The sampling based approaches tends to oversample large bodies with similar visual attributes (probably of less interest) and are likely to miss some smaller clusters of interest such as disturb earth in our application. This eventually will produce results that are biased towards a small number of larger clusters, which may in turn lead to lower quality clusters [19]. It is sensible to cluster similar observations and reduce them to a quantized value (average observations in each cluster) representative of all the values in a cluster.

In this article, we adopt the strategy of data filtering and smoothing through averaging similar observations. This, on the one hand substantially reduces the size of the data, while on the other hand it suppresses noise. We achieve this through k-means clustering by deliberately over-clustering (choose a very large number of clusters) in the first level; therefore, sidestep the main drawbacks of k-means clustering algorithm. The mixture models are then fitted to a reduced dataset in the second level. This two-step process is applied in a tree-like structure to partition the clusters into smaller and smaller clusters in order to identify clusters of high, medium, and low interest. 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 accounts for correlations between variables, but also learns K in a data-driven fashion that makes it suitable for tree-based algorithms. Our method is explained in "Methods" section and applied to a case study in "Results and discussion" section, where it is also compared to an alternative SVM approach. Finally, the conclusions are presented in "Conclusions" section.

Methods

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. Denote the p-dimensional Gaussian density by $\mathcal{N}_p(\cdot)$ then a mixture of K Gaussian components takes the form

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

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 [20], 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) 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 [21–23]. By integrating out G, the predictive distribution for a component parameter follows a Pólya urn scheme [24]

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

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

Simpler and more efficient methods have been developed to fit posterior of 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 [25]. 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 [26]. 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 $x_i|z_i = k, \theta_k \sim \mathcal{N}(x|\theta_k)$.

Blocked Gibbs sampling scheme to fit DPGMM

A blocked Gibbs sampler [26] 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

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

   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

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

Data preprocessing: turning big into small

In massive datasets, when much of the data provides similar information, a sensible strategy would be to group similar observations together to get an adequate representation from each group. This may, however, lead to substantial loss of information, which can be reduced by introducing a reasonably large number of clusters. The term 'reasonably large' is used to emphasise the underlying trade-off between the number of clusters and the loss of information that may be incurred; a smaller number of clusters leads to a larger amount of information loss. This first-level clustering is followed by a quantization step (rather than sampling) that involves mapping a larger set of values to a smaller set by suppressing the noise.

We achieve the above with k-means clustering, a popular clustering algorithm, because of its scalability and efficiency in large data sets. The algorithm employs a proximity matrix (Euclidean distance) whereby the sum of the squared distances from the observations in each cluster to their cluster centres is minimized [4]. Several algorithms have been proposed to derive a solution to the k-means problem. However, the algorithm in [27] is known to perform well.

In order to maintain the quantized set as closely to the original dataset as possible, we use a large number of clusters. In this way, we sidestep the two well-known drawbacks (the model selection and convergence to a local optimum) of the k-means clustering. In Fig. 1a simulated dataset of 5000 observations from a 5 component Gaussian mixture is plotted overlaid by 500 quantized values obtained via k-means clustering. Note that we use k-means as a preliminary dimension reduction step to alleviate the computational burden for the more flexible and sophisticated mixture models, which allows incorporation of additional available information and also takes into account the correlation between variables.

Big data implementation of DPGMM

As mentioned above, the posterior inference of DPGMM does not scale well to Big data. Here we propose a multi-step process. The first step involves reducing the of size N_0, say, to a informative smaller dataset of size N_1, say, via a quantization method such as k-means. The second step is the usual DPGMM implemented with the quantised values. This reduces the number of clusters from N_1 to $K_1 \ll N_1$. The process can be stopped here if the resultant number of clusters meets a pre-specified criteria. Example criteria may be, if the clusters are adequately interpretable; if the number of observations in a cluster reaches a minimum size; the DPGMM fits only one cluster; or if a cluster of interest is identified. In the case study considered here this would entail a small number of clusters encapsulating fire ant presence. In practice, however, it is often preferable to further partition large clusters obtained at the first layer of DPGMM. To proceed, we track back to the original data for each cluster of interest (leaving out the components of non-interest) and repeat the above process (k-means clustering, quantization, and DPGMM) until no more partitioning is required.

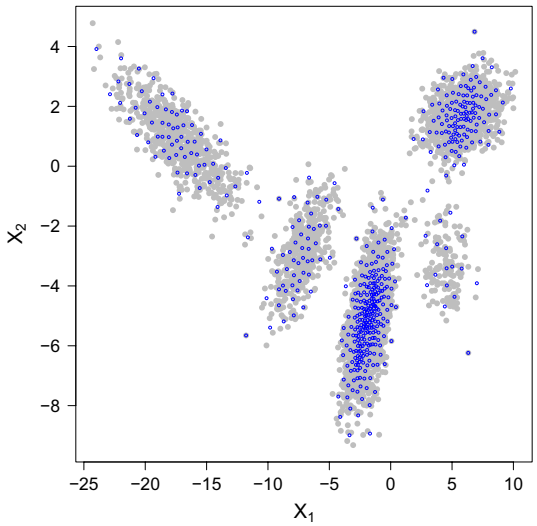

Fig. 1 Quantization via *k*-means. A dataset of 5000 observations (plotted in solid gray points) is drawn from a Gaussian mixture of 5 components and reduced to 500 quantized values (plotted in blue circles)

The method is summarized in the following steps:

1. Start with the observed data of size N_0 and obtain $N_1 \ll N_0$ clusters using *k*-means clustering, with $k = N_1$.
2. Obtain the means of the N_1 clusters as the quantized set of values.
3. Apply DPGMM to the N_1 quantized values obtained in Step 2. This will reduce the number of clusters from $N - 1$ to a much smaller number, K_1.
4. Identify the components of interest. Stop the process or go to Step 5 if further partitioning is desirable.
5. Drop all the clusters of non-interest and repeat Steps 1–4 separately for each component of interest or a pre-specified stopping rule is reached.

Results and discussion

Effect of quantization on posterior inference

Here we demonstrate the effect of quantization on the posterior estimates using a simulated dataset. We generate 5000 observations from a 2-dimensional Gaussian mixture model with 5 components. The dataset is plotted in Fig. 1 overlaid by 500 quantized values obtained via *k*-means clustering.

The posterior estimates are obtained using DPGMM, described in "Methods" section, first conditional on the full dataset (all 5000 observation) and then conditional on the 500 quantized values. The results are shown in Fig. 2. The estimates for the components means based on the quantized values are comparable in accuracy to the estimates based on the full dataset. However, as one can expect, the estimates based on the quantized values are less precise than the estimates based on the full dataset. Although an increase in the number of quantized values usually improves the

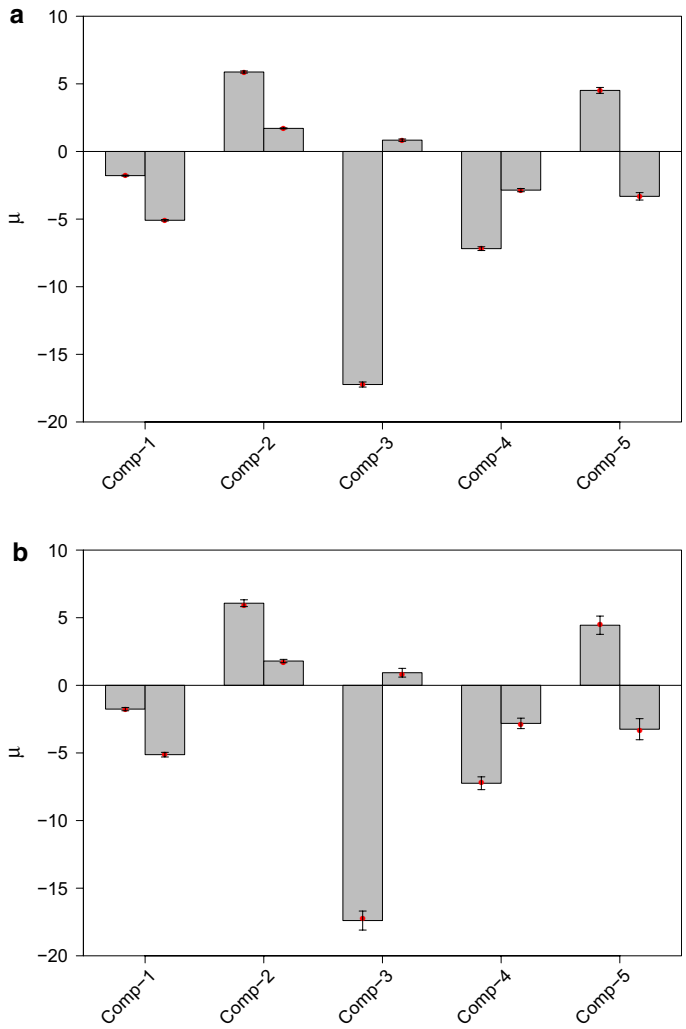

Fig. 2 Posterior estimates of the components mean vectors, for the data shown in Fig. 1, conditional on **a** the full 5000 observations and **b** the 500 quantized values. The error bars indicate ± 2SD and the solid red points show the true means

estimates so long as resources allow, a slight loss in accuracy and efficiency may be acceptable given the fact that one can explore very large datasets on, for example, a laptop.

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 15,107 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 Fig. 3. The proportion of colonies identified for each year are provided in Fig. 4. 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, but possible reasons for it still require further investigation.

Fig. 3 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

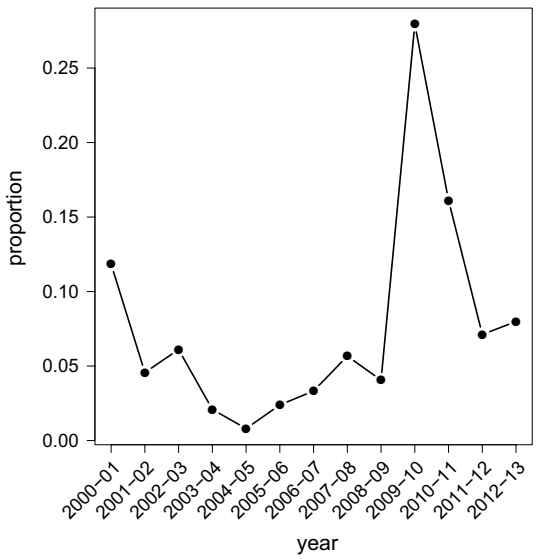

Fig. 4 Proportions of fire ant colonies detected each year from 2000–2013

A Landsat image is also available for each year of the study period. 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. Since a part of the image was required to cover the study area, the images were first cropped to limit them to the study region (the urban area). This resulted in a set of 13 well-aligned images. The cropped images were converted into workable data files using the 'raster' package [28] in R. Note that we use 6 Landsat spectral bands: 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. Figure 5 shows the densities of all 6 variables from 2005 image, with black lines for pixels containing colonies and red lines for pixels where containment is not recorded. A clear shift in densities for the known colony sites can be seen for most of the variables. This indicates that the imagery data does provide some insight into the attributes of a preferred habitat for the establishment of fire-ant colonies.

We also used R for the substantive statistical analysis. To solve the *k*-means problem, we used the algorithm in [27], 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. Larger number of starting points, however, increases computation cost which is due to multiple runs of the algorithm. We avoided this by using parallel processing facility in R provided by *foreach* loop from the 'foreach' package. Since our use of *k*-means clustering is to reduce the dimension of the data to a set of quantised values (rather than final clustering), we did not find noticeable difference in terms of visual interpretation while using a single random starting point. Note also that very large number of N_1 also increases computational time and memory requirement, especially when N_0 is large. To fit DPGMM, we translated Matlab codes, available at http://ftp.stat.duke.edu/WorkingPapers/09-26.html, into R codes (for details about Matlab codes, see, [18]). We used 30,000 iterations of blocked Gibbs sampler including 2000 burn-in iterations at each node of the tree. The overall computation time averaged over the 13 images considered in this study was 10 h and 58 min when $N_1 = 3000$. This computation time reduced to 7 h and 33 min for $N_1 = 2000$ and increased to 16 hours and 26 minutes when we set $N_1 = 4000$. Note that we used the high performance computing facility at the Queensland University of Technology for our computations which has 2.6 Ghz processors with 251 Gb memory. The computational time can be further reduced by using R package 'Rcpp' [29], which interfaces C and C++ code in R.

Analysis and results

To learn about the attributes of fire-ants' preferred habitats, we classified satellite imagery data. Each of the 13 images was converted to a data matrix of 3,216,582 rows (pixels) and six columns (spectral bands). As a preprocessing step, we reduced the dimension of the data using *k*-means clustering from $N_0 = 3,216,582$ to $N_1 = 3000$ quantized values. The DPGMM was then fitted iteratively in a tree-like structure (as described in "Methods" section) to the quantized values. This was done independently for each image from the year 2001 to 2013. We tested a range of values of N_1 and found that the number of components and their structures did not change (in terms of visual interpretation) as we increased the value of N_1 beyond 3000. Therefore, we set $N_1 = 3000$ for all the results shown here (even for the classification of sub-classes).

The classification based on the images from years 2002, 2007 and 2010 are shown, respectively, in Figs. 6, 7 and 8. The proportion of fire-ant identified in each cluster are presented in Tables 1, 2 and 3. Note that each of these tables is based on a single year

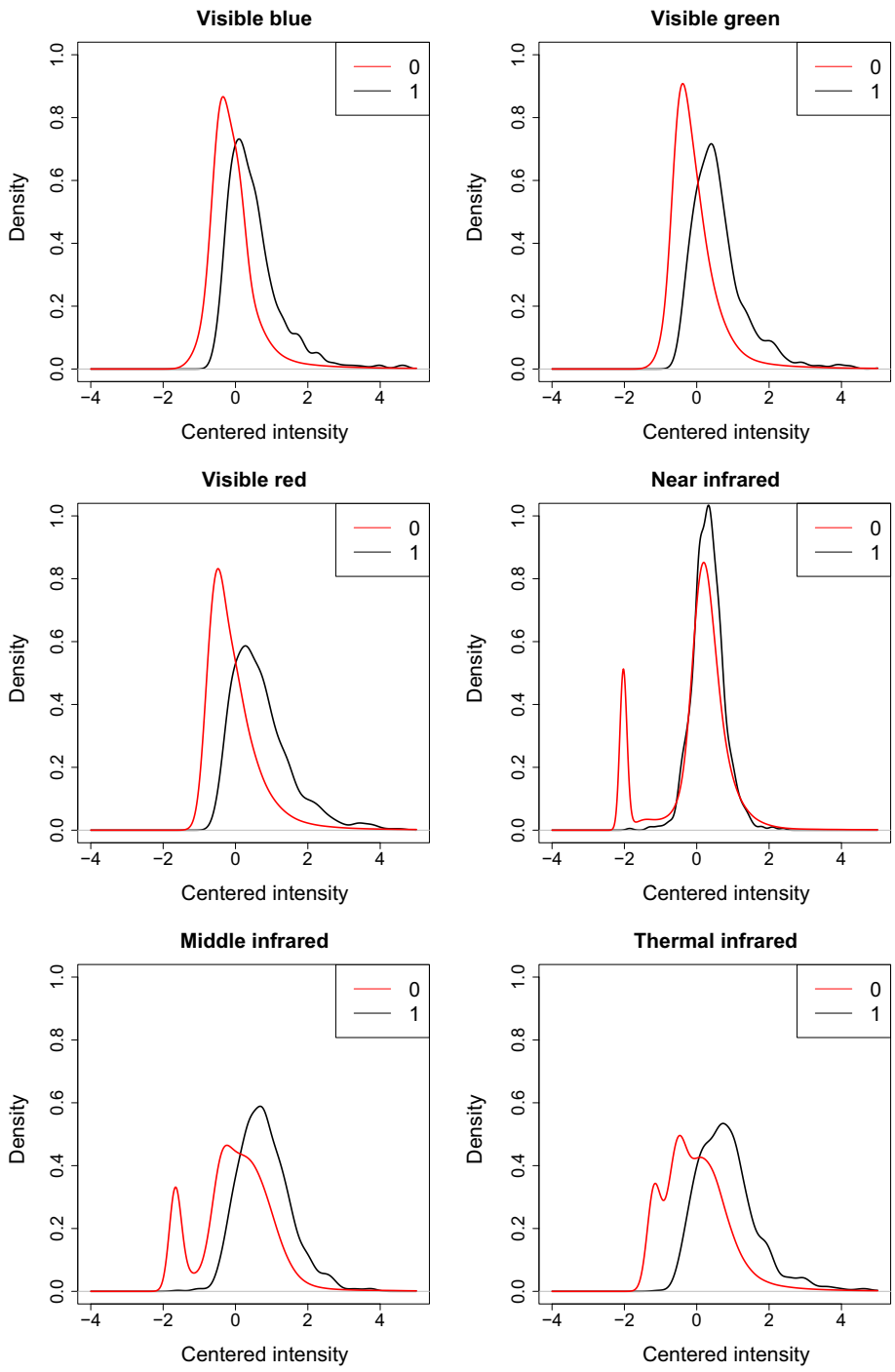

Fig. 5 Densities of 6 Landsat bands from 2005 image with 1 (in black) indicates colonies are identified and 0 (in red) indicates otherwise

image, however, the proportions of the observed fire-ant 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 Additional file 1 due to the compatibility of the results across different years.

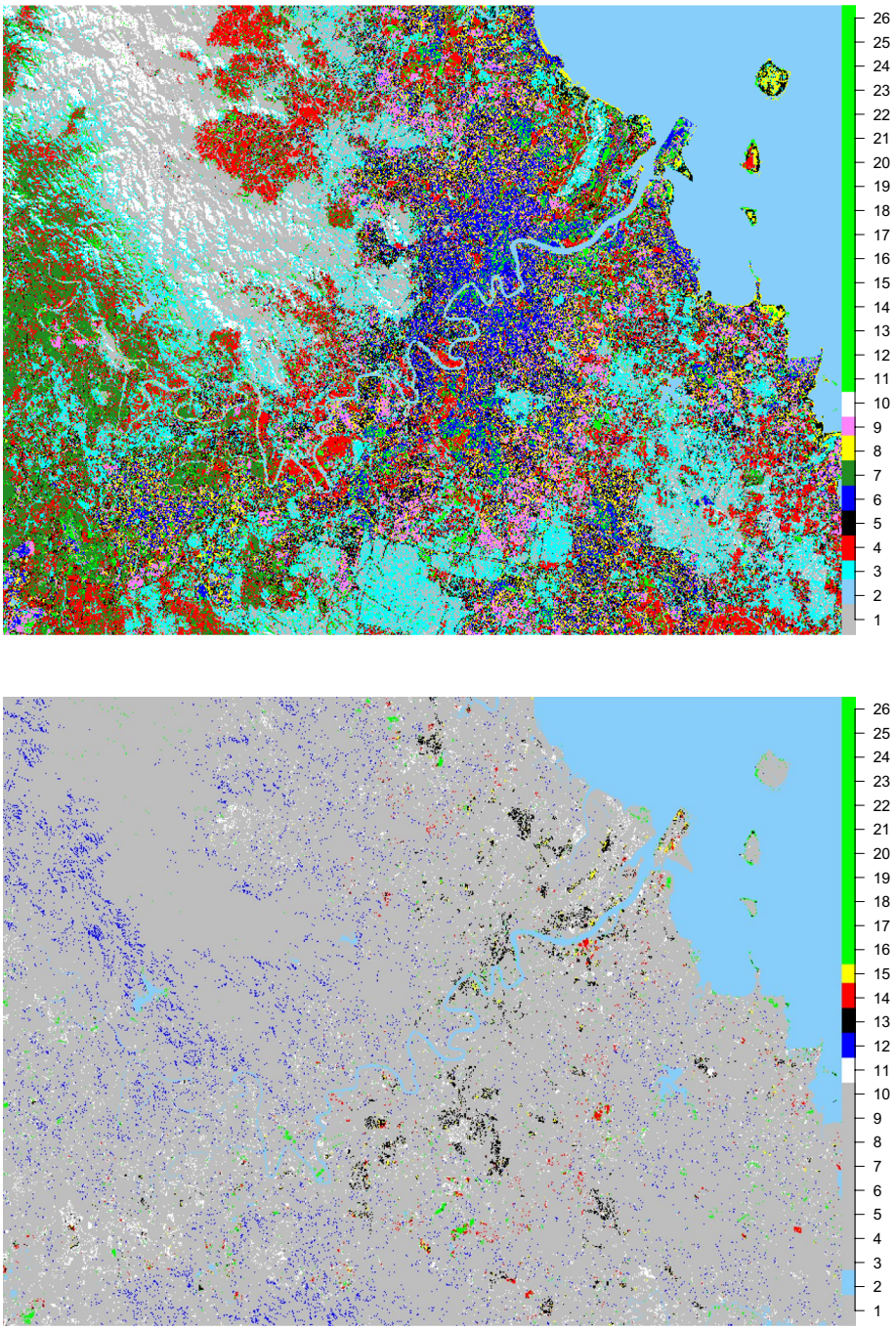

Fig. 6 Cluster analysis of satellite image of the Brisbane area taken in 2002. For clarity, some of the clusters are merged together in bright-green colour and the results are presented in two plots: (left penal) 1: mountains and forest, 2: water, 3: forest 4: mix of parks, playgrounds and grassland, 5: old residential areas including some roads, 6: old residential areas, 7: scrub-land, 8: Bright surfaces including seashore, 9: new residential areas, and 10: mountains and forest; (right penal) 11: parks and playgrounds, 12: mountains and forest, 13: commercial buildings, 14: disturbed earth (recent deforestation), 15: impervious surfaces

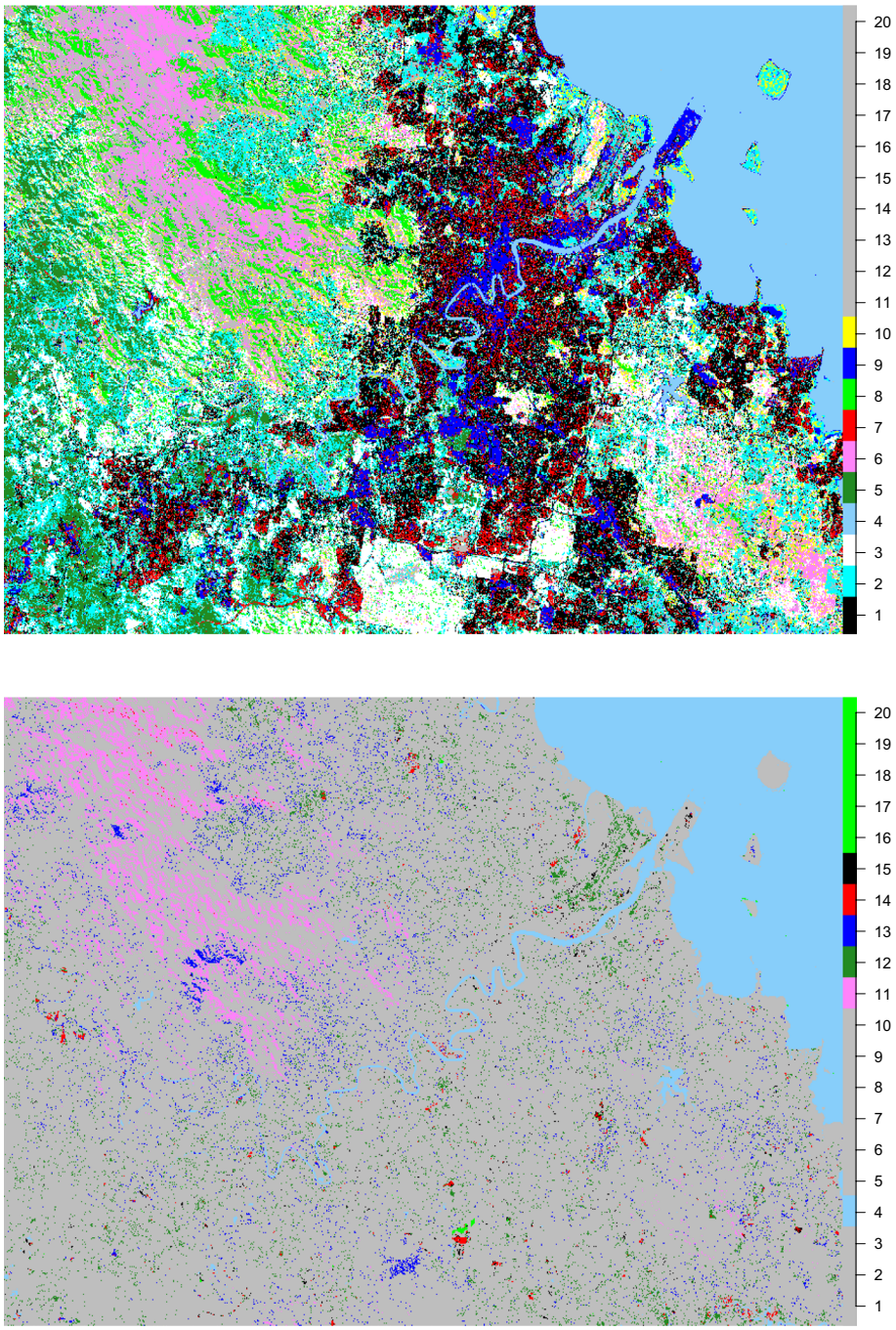

Fig. 7 Cluster analysis of satellite image of the Brisbane area taken in 2007. For clarity, some of the clusters are merged together and the results are presented in two plots: (left penal) 1: old residential areas, 2: parks, playgrounds, and grasslands, 3: forest, 4: water, 5: scrub-land, 6: mountains, 7: mostly new residential areas, 8: forest, 9: commercial buildings, and 10: forest; (right penal) 4: water, 11: mountains and forest, 12: bare ground, 13: forest, 14: disturbed earth (recent deforestation), 15: disturbed earth (recent deforestation)

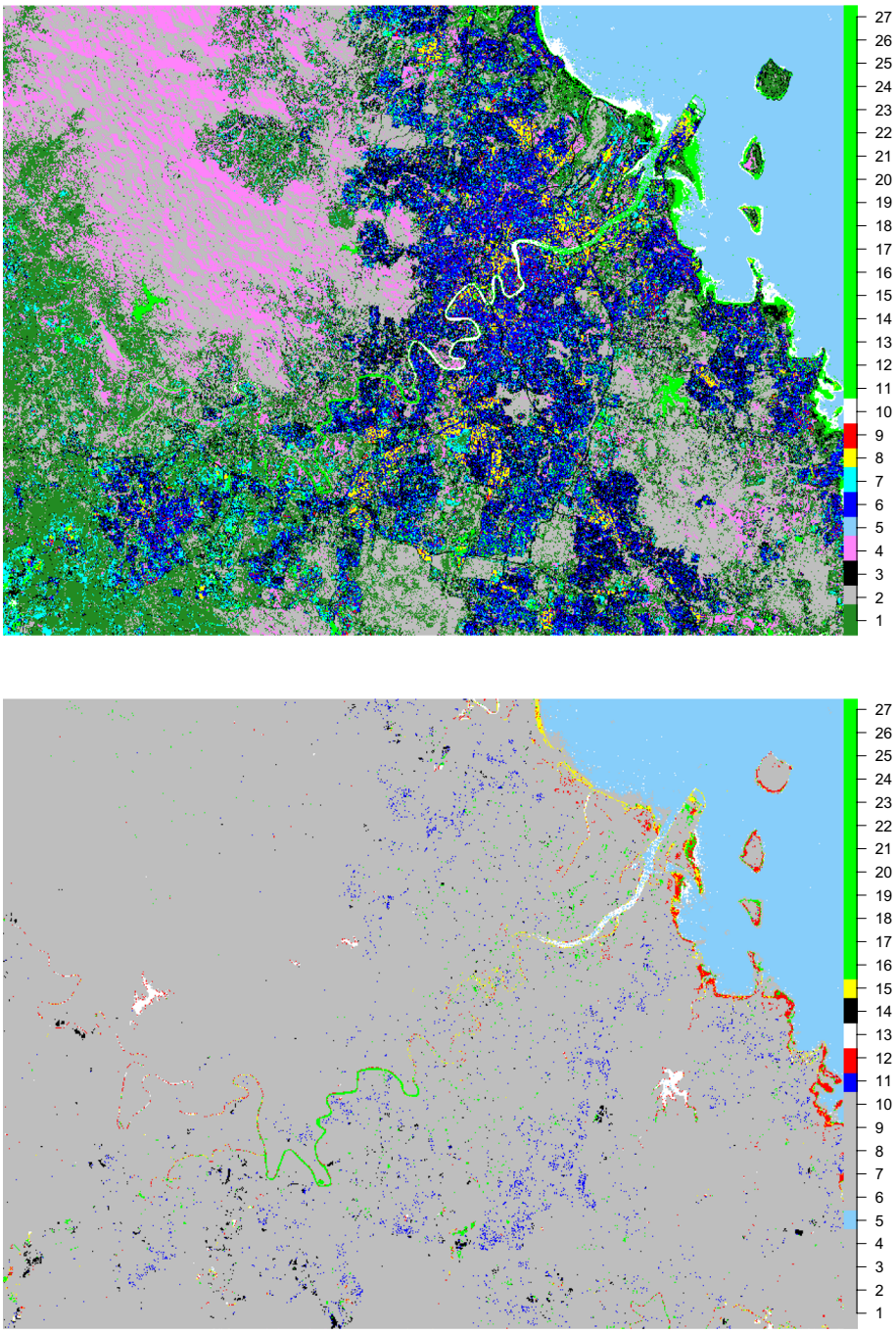

Fig. 8 Cluster analysis of satellite image of the Brisbane area taken in 2010. For clarity, some of the clusters are merged together in bright-green clolour and the results are presented in two plots: (left penal) 1: scrub-land or bare ground, 2: forest, 3: roads and old residential areas, 4: mountain and forest mixed with playgrounds, 5: water, 6: residential areas, 7: bare ground and impervious surfaces, 8: commertial buildings, 9: residential area (with bright steel roofing) mixed with commercial buildings, 10: seashore; (right penal) 5: water, 11: new residential areas, 12: water and seashore, 13: water, 14: disturbed earth, 15: water and seashore

The final number of components per image varies across different years but remains at between 20 and 42. Some of these variations can be possibly attributed to the time of the day the image is acquired. For example, the mountainous and forest area, which is

Table 1 The percentages of fire-ant colonies identified in each of the spatial components (shown in Fig. 6) over the period of 13 years conditional on the image acquired in 2002 (highlighted in italic)

C. No	C. Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1	16.3	1.1	*1.2*	1.8	1.6	2.5	2.2	0.6	2.0	3.9	3.3	3.5	6.8	8.5
2	13.5	0.1	*0.0*	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
3	13.3	3.1	*2.0*	2.8	2.2	6.6	4.5	2.6	1.1	3.4	6.3	20.3	16.9	14.1
4	12.8	9.3	*10.0*	19.7	16.8	14.9	32.8	14.6	17.5	19.8	50.1	31.3	18.4	26.0
5	9.7	19.6	*16.4*	15.9	10.8	11.6	14.6	7.4	7.6	9.7	5.1	17.5	12.2	13.0
6	7.0	18.6	*13.4*	11.8	7.9	8.3	12.0	13.8	12.4	4.1	0.7	1.8	3.6	3.6
7	6.2	0.9	*1.9*	3.0	19.6	3.3	9.5	34.5	27.8	25.0	22.7	5.8	16.7	12.5
8	6.0	8.3	*9.0*	9.0	4.4	3.3	5.9	4.0	6.3	3.1	1.4	6.5	4.0	3.9
9	5.5	22.0	*31.8*	24.3	11.7	24.8	7.6	13.0	13.9	11.5	3.4	5.6	8.5	6.2
10	3.5	0.1	*0.1*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
11	2.2	6.6	*6.5*	5.4	20.3	9.9	6.4	2.8	9.0	13.6	1.8	3.1	6.1	5.8
12	1.9	0.8	*0.3*	0.3	0.3	0.8	1.7	0.4	0.5	0.8	0.8	1.3	2.4	2.5
13	0.9	4.8	*5.5*	1.7	0.9	0.8	0.0	2.2	0.4	0.5	0.1	0.4	0.2	1.1
14	0.4	2.9	*1.2*	2.4	1.9	7.4	0.8	3.0	1.1	1.6	2.5	1.8	0.2	1.2
15	0.2	1.6	*0.3*	0.5	0.0	1.7	0.0	0.0	0.0	0.2	0.6	0.5	0.0	0.3
16	0.2	0.3	*0.1*	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.5
17	0.1	0.0	*0.1*	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.2
18	0.1	0.0	*0.0*	0.5	1.6	4.1	0.8	1.2	0.0	1.3	0.1	0.7	3.5	0.3
19	0.0	0.0	*0.0*	0.0	0.0	0.0	0.0	0.0	0.2	1.5	1.1	0.0	0.0	0.2
20	0.0	0.0	*0.0*	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1
21	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	0.0
22	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	0.0
23	0.0	0.0	*0.0*	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
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	0.0
25	0.0	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
26	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	0.0
Total incursions		1788	*690*	923	316	121	361	502	856	613	4220	2436	1075	1206

The C. No indicates component numbers corresponding to the component numbers in Fig. 6. 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

Table 2 The percentages of fire-ant colonies identified in each of the spatial components (shown in Fig. 7) over the period of 13 years conditional on the image acquired in 2007 (highlighted in italic)

C. No	C. Size	2001	2002	2003	2004	2005	2006	*2007*	2008	2009	2010	2011	2012	2013
1	16.4	37.8	28.6	34.7	18.7	27.3	18.8	*14.0*	18.3	21.9	5.4	9.9	14.3	14.4
2	15.6	11.8	15.7	17.1	27.8	26.4	26.3	*20.6*	16.3	26.5	26.6	15.3	25.3	26.6
3	14.2	4.4	5.7	5.2	5.7	6.6	2.2	*0.2*	1.2	1.6	2.7	4.4	8.8	8.7
4	13.2	0.3	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	0.0	0.0	0.1
5	7.2	2.2	5.8	6.1	19.3	2.5	8.4	*39.9*	30.8	9.0	52.2	53.5	19.0	25.4
6	6.4	0.0	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	0.0	0.1	0.2
7	5.7	15.3	18.0	15.9	9.8	18.2	19.0	*16.4*	16.6	25.2	7.3	9.8	18.8	13.1
8	5.3	0.2	0.3	0.3	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	0.1	0.4	0.9
9	4.1	23.0	19.0	14.3	7.0	6.6	9.8	*1.6*	5.8	6.5	0.9	1.4	2.0	4.2
10	3.7	0.2	0.1	1.0	1.3	0.8	2.8	*0.0*	0.2	0.3	0.1	0.2	0.3	0.1
11	3.5	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
12	2.4	3.8	4.1	4.4	10.4	9.1	5.9	*5.6*	10.6	3.6	1.5	2.5	6.0	2.9
13	1.8	0.3	0.6	0.3	0.0	2.5	3.9	*0.2*	0.0	2.3	0.1	1.2	2.3	0.9
14	0.3	0.3	0.3	0.0	0.0	0.0	1.4	*1.4*	0.1	0.3	1.4	0.7	1.4	0.7
15	0.2	0.5	1.5	0.4	0.0	0.0	1.4	*0.2*	0.1	1.5	1.1	1.0	1.2	1.8
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	1.1	0.6	0.0	0.0	0.0
17	0.0	0.0	0.3	0.2	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0
18	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	0.0
19	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	0.0
20	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	0.0
Total incursions		1788	690	923	316	121	361	*502*	856	613	4220	2436	1075	1206

The C. No indicates component numbers corresponding to the component numbers in Fig. 7. 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

Table 3 The percentages of fire-ant colonies identified in each of the spatial components (shown in Fig. 8) over the period of 13 years conditional on the image acquired in 2010 (highlighted in italic)

C. No	C. Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	*2010*	2011	2012	2013
1	21.9	11.6	12.8	17.0	25.3	16.5	26.9	25.1	15.6	19.6	*67.2*	40.5	25.7	26.0
2	21.9	1.7	2.0	2.5	4.4	0.8	2.2	0.0	0.5	0.3	*0.5*	0.1	2.0	1.5
3	11.8	29.4	24.4	24.2	27.2	25.6	19.3	14.2	13.9	14.2	*7.2*	9.9	12.9	13.5
4	11.2	2.0	2.5	3.9	3.5	8.3	5.9	2.0	3.2	11.1	*0.6*	0.9	2.5	4.6
5	10.8	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
6	10.7	30.6	25.1	27.1	15.2	8.3	23.0	15.8	16.3	10.6	*3.9*	6.9	13.5	10.7
7	5.1	9.4	13.5	10.5	12.3	25.6	13.2	23.8	18.4	34.7	*18.9*	34.7	30.2	28.7
8	1.5	7.0	9.6	6.9	7.3	3.3	1.1	1.6	3.7	0.8	*0.0*	0.6	1.6	1.2
9	1.1	2.9	4.1	1.6	3.2	3.3	2.0	0.6	2.1	2.3	*0.7*	1.7	1.6	1.6
10	0.9	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
11	0.7	2.9	2.8	3.6	1.3	0.0	2.8	3.0	1.2	0.7	*0.6*	0.9	2.7	2.5
12	0.6	0.0	0.0	0.2	0.0	0.0	0.0	0.2	1.4	0.2	*0.0*	0.0	0.0	0.0
13	0.5	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
14	0.4	1.2	1.5	1.8	0.0	7.4	1.1	10.0	22.6	5.2	*0.3*	3.2	7.0	9.4
15	0.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	*0.0*	0.0	0.0	0.1
16	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	*0.0*	0.0	0.0	0.0
17	0.2	0.9	1.2	0.4	0.3	0.0	2.5	0.0	0.2	0.2	*0.1*	0.5	0.3	0.2
18	0.1	0.0	0.0	0.1	0.0	0.0	0.0	2.0	0.0	0.0	*0.0*	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	*0.0*	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.5	0.0	*0.0*	0.0	0.0	0.0
21	0.0	0.0	0.4	0.1	0.0	0.8	0.0	0.2	0.0	0.0	*0.0*	0.0	0.0	0.0
22	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	0.0
23	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.1	0.1
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	0.0
25	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	0.0
26	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	0.0

Table 3 (Continued)

C. No	C. Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	*2010*	2011	2012	2013
27	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	0.0
Total incursions		1788	690	923	316	121	361	502	856	613	*4220*	2436	1075	1206

The C. No indicates component numbers corresponding to the component numbers in Fig. 8. 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

broken into three components in Figs. 6 and 7, makes two components in Fig. 8, possibly because of shadows. In some images roads are relatively well separated (Figs. 6 and 8) but this is not always the case (Fig. 7). Other variations are because of the changes in the landscape over time. However, the number of components that consist of more than 1% of the pixels remains below 15 for most of the images. These large components are materially similar across different years and are visually interpretable into different land cover classes, namely, mountains, 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 is always well separated from the rest of the components. Although this component is not of interest to us, it helps in identifying and interpreting other components. The components that represent the mountains and forest are the largest by area and is found to be consistently at low risk of fire-ant incursion (see components 1 and 3 in Table 1; components 3, 6, 8, 11, and 13 in Table 2; and components 2 and 4 in Table 3). The scrub-land is found to be at high risk of infestation (see components 7, 5 and 1, respectively, in Tables 1, 2 and 3) followed by parks and playgrounds (see components 4 and 11 in Table 1 and component 2 in Table 2). These two types of land cover classes are well separated in most of the images (see Figs. 6 and 7). The old residential zones (see components 5 and 6 in Table 1, component 1 in Tables 2 and component 6 in 3) including the areas with commercial buildings (see component 13 in Table 1, component 9 in Tables 2 and component 8 in 3) are 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 new residential zones have seen occasional high incursions even in later years (see component 9 in Table 1, component 7 in Table 2). Roads, and new development zones (disturbed earth, recent deforestation) are also found to be at moderate risk consistently thorough the study period (see component 5 in Table 1 and component 3 in Table 3 for roads and component 14 in Tables 1, 2 and 3 for disturbed earth). The potential factors may include moving soil and other materials to and from the development sites.

As mentioned above the Tables 1, 2 and 3 also presents 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 the year that follows. For example, consider Table 1 in which component 5 (contained 16.4% of the observed nests) and component 9 (contained 31.8% of the observed nests) were at high risk of fire-ant incursions in 2002 remained at high risk in 2003 (component 5 contained 15.9% of the observed nests and component 9 contained 24.3% of the observed nests). Similarly, in Table 3, which is based on classification of image from 2010, component 1 and component 7 together contains 86.1% of the observed nests in 2010. The component 1 was at highest risk in 2010 (contained 67.2% of the observed nests), which was also at highest risk in 2011 (contained 40.5% of

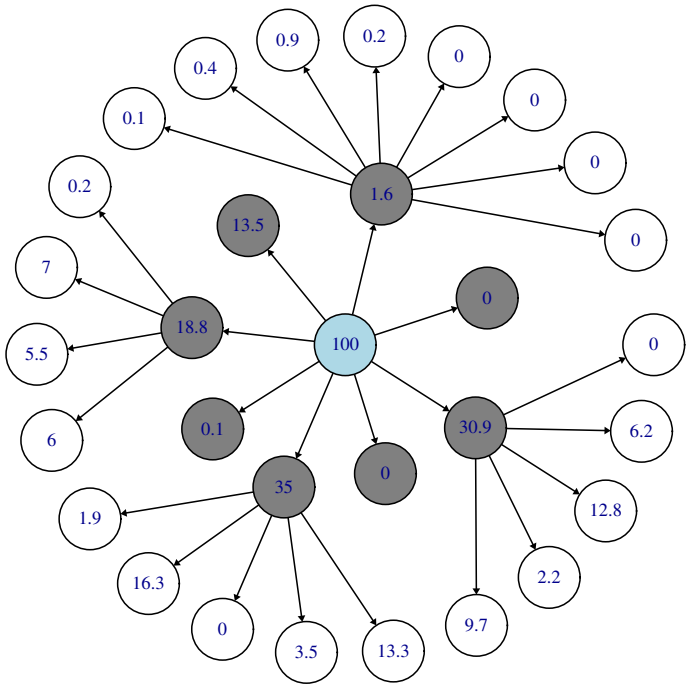

Fig. 9 Tree diagram for the image shown in Fig. 6. The number in each bubble indicates the percentage of pixels contained in the corresponding cluster

the observed nests). The component 5 contained 18.9% in 2010 and 34.7% in 2011. Some of the potential factor for anomalous changes could 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 constants. For example, the component 2 in Table 3, which covered 21.9% of the study area contained 67.2% of the fire-ant incursions and could be targeted in the fire-ant eradication program in the following year.

The trees generated in the classification of images from the years 2002, 2007, and 2010 are diagrammed, respectively, in Figs. 9, 10 and 11. In all the three cases, the stopping criteria (the node is not of interest or cannot be clustered any more or too small to split it further) met at the second level where the tree stops growing any further. In most of the cases larger clusters are classified into visually interpretable smaller clusters. See, for example, Fig. 9 where a node that is made up of 30.9% of pixels is broken into five clusters containing 12.8%, 9.7%, 6.2%, 2.2 and 0% of the pixels: the first of these clusters represents mix of parks, playgrounds and grassland (component-4 in Table 1); the second of these clusters represents old residential area including roads (component-5 in Table 1); the third cluster represents scrubland; the forth cluster represents parks and playgrounds; and the fifth cluster is too small to be visually interpreted. In other cases a small cluster is further partitioned into a few clusters in which case some have distinct characteristics. For example, in Fig. 9, a component contains 1.6% of the pixels is

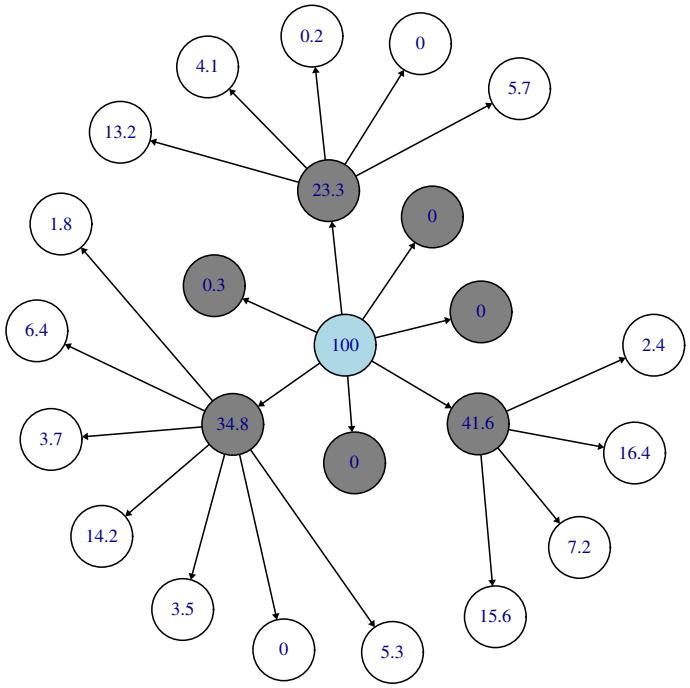

Fig. 10 Tree diagram for the image shown in Fig. 7. The number in each bubble indicates the percentage of pixels contained in the corresponding cluster

partitioned into eight clusters. Three clusters out of these six clusters consists of 0.9% (component 13 in Fig. 6 which represents steel roofs of the commercial buildings), 0.4% (component-14 in Fig. 6 which represents disturbed earth), and 0.2% (component-15 in Fig. 6 which represents some impervious surfaces) of the pixels and represent different bright surfaces. The rest of these eight clusters are too small for visual interpretation.

One-class support vector machine (OCSVM)

Another technique that seems suitable for the presence-only data is the one-class support vector machine [30]. It has been used for anomaly and outlier detection. This technique first attempts to learn the decision boundary based on the training dataset while incorporating a soft margin classifier in order to account for outliers in the training dataset. For each test data point it is then determined if it falls in an anomalous class that is outside the decision boundary. This technique is computationally faster and is available through the R function *svm()* from the 'e1071' package [31].

We use of OCSVM to determine if the pixels with fire-ant nests share some attributes, hence fall in the same class. We trained the OCSVM anomaly detector considering all the pixels that contained fire-ant nests as a training dataset. The rest of the pixels that do not contain fire-ant nests were used as a test dataset. The land-cover classification results based on the image from 2010 are shown in Fig. 12. Out of the 1108 observation in the training dataset (including those with multiple nests), 116 observations were found to be in outlying class. The results from test data suggest that the area that needs to be targeted in the fire-ant eradication program consists of 39.66% of the pixels. This exclude some of the clusters that were identified as at risk of infestation of fire-ants using

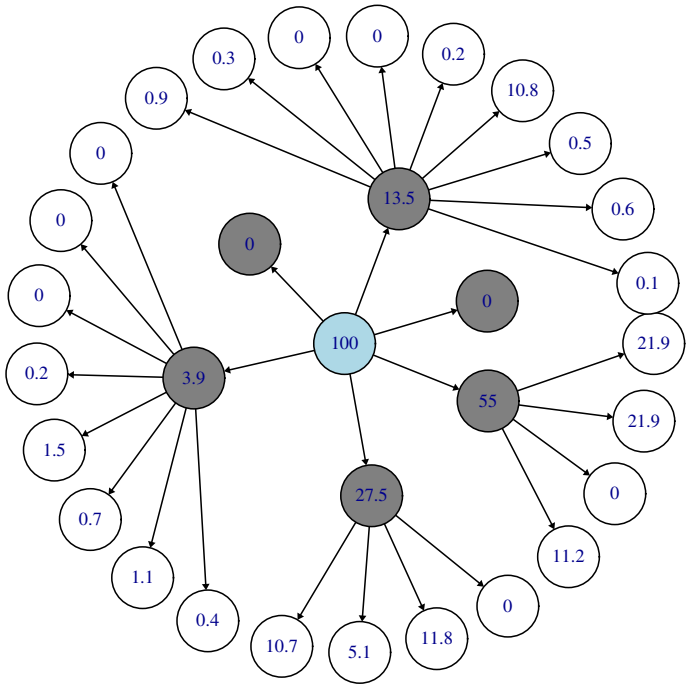

Fig. 11 Tree diagram for the image shown in Fig. 8. The number in each bubble indicates the percentage of pixels contained in the corresponding cluster

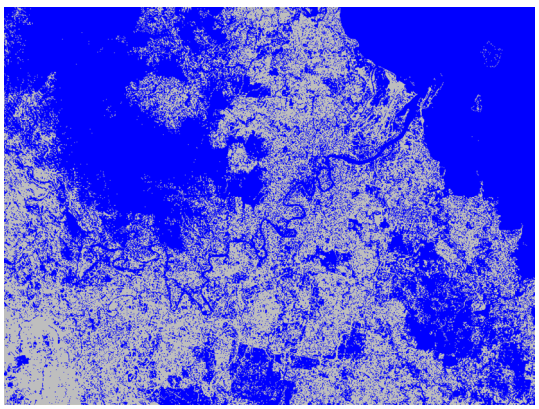

Fig. 12 Classification of image from 2010 using OCSVM. The preferred habitat of fire-ant is shown in gray

our multi-step method, for example, buildings with steel roof top such as area with commercial buildings and disturbed earth. These are mainly the clusters whose representative pixels in the training dataset stood out as outliers. Moreover, the OCSVM provided less details as compared to our method; for example, it does not distinguish high and low risk classes and where the eradication program has been more successful. The two approaches are, however, in agreement for some of the high-risk clusters, for example, residential area and scrubland.

Conclusions

DPGMM are computationally prohibitive for large datasets, their implementation in tree-based clustering algorithm dramatically increase the computational time even for intermediate size dataset. We used *k*-means clustering to reduce the size of dataset to a smaller set of quantized values. This led to one of the key achievements of this work, which is the scaling of DPGMM to large datasets and its tree-based implementation to identify the components of interest. The proposed method enables to classify a dataset with millions of observation in a matter of minutes.

We used the method to classify satellite imagery data in order to identify the land cover classes that are at high, medium, and low risk of infestation of fire-ants. The plain areas with non-forest natural vegetation (scrub-land) and parks and playgrounds are found to be at high risk of infestation. Roads and new development zones are also among the preferred habitats (although at a moderate risk through the study period). Residential areas are also found to be at a high risk of infestation in the initial years of the study period. However, the risk has declined soon after the start of eradication program, perhaps showing the effectiveness of the program in residential zones.

Note that the main objective of this study was to scale Bayesian mixture models to Big data. We achieved this by using an algorithm that is parallelizable and reaches the final fine clusters in a tree-like structure. We used the algorithm to cluster satellite images and connected the presence-only observed data to the clusters thus obtained to describe the proportions of the observed presences in each cluster. We also calculated the proportions of the observed presences for the years other than the year in which the image was acquired assuming no significant temporal changes in the land-cover over a period of few years. A more principled way, however, 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 too, which could potentially play an important role in the problem being tackled. A more sophisticated model that take into account both the spatial and temporal dependence would be required. We leave these extensions for future research.

Additional file

Additional file 1. Additional figures and tables presenting the results for the rest of the study period not shown in the main text.

Abbreviations

MCMC: Markov Chain Monte Carlo; DPGMM: Dirichlet process Gaussian mixture models; OCSVM: one-class support vector machine.

Authors' contributions

Insha Ullah did the literature review, contributed to the methodology development, implemented and evaluated the method and drafted the manuscript. Kerrie Mengersen contributed to the the methodology development, refined the concepts and revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We are thankful to Clair Alston-Knox for providing the data and R codes to read satellite imagery data.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
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 (ARC) Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

Publisher's Note

Received: 7 December 2018 Accepted: 25 February 2019
Published online: 22 March 2019

References

1. Spring D, Cacho OJ. Estimating eradication probabilities and trade-offs for decision analysis in invasive species eradication programs. Biol Invasions. 2015;17(1):191–204.
2. Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA. Is my species distribution model fit for purpose? Matching data and models to applications. Glob Ecol Biogeogr. 2015;24(3):276–92.
3. Hastie T, Fithian W. Inference from presence-only data; the ongoing controversy. Ecography. 2013;36(8):864–7.
4. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967; 1:281–297.
5. Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inform Theory. 1975;21(1):32–40.
6. Bardenet R, Doucet A, Holmes C. On markov chain monte carlo methods for tall data. 2015. arXiv preprint arXiv :1505.02827.
7. Lee A, Yau C, Giles MB, Doucet A, Holmes CC. On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. J Comput Graph Stat. 2010;19(4):769–89.
8. Guha S, Hafen R, Rounds J, Xia J, Li J, Xi B, Cleveland WS. Large complex data: divide and recombine (d&r) with rhipe. Statistics. 2012;1(1):53–67.
9. Chang J, Fisher III JW. Parallel sampling of dp mixture models using sub-cluster splits. In: Advances in Neural Information Processing Systems, 2013; 620–628.
10. Williamson S, Dubey A, Xing EP. Parallel markov chain monte carlo for nonparametric mixture models. In: Proceedings of the 30th international conference on machine learning (ICML-13). 2013. p. 98–106.
11. McGrory CA, Titterington D. Variational approximations in Bayesian model selection for finite mixture distributions. Comput Stat Data Analy. 2007;51(11):5352–67.
12. Ormerod JT, Wand MP. Explaining variational approximations. Am Stat. 2010;64(2):140–53.
13. Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. J Mach Learn Res. 2013;14(1):1303–47.
14. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. J Am Stat Assoc. 2017;112(518):859–77.
15. Marin J-M, Pudlo P, Robert CP, Ryder RJ. Approximate bayesian computational methods. Stat Comput. 2012;22:1167–80.
16. Moores MT, Drovandi CC, Mengersen K, Robert CP. Pre-processing for approximate Bayesian computation in image analysis. Stat Comput. 2015;25(1):23–33.
17. Huang Z, Gelman A. Sampling for bayesian computation with large datasets. 2005.
18. Manolopoulou I, Chan C, West M. Selection sampling from large data sets for targeted inference in mixture modeling. Bayesian Anal. 2010;5(3):1.
19. De Vries CM, De Vine L, Geva S, Nayak R. Parallel streaming signature em-tree: a clustering algorithm for web scale applications. In: Proceedings of the 24th international conference on World Wide Web. 2015; 216–226. International World Wide Web Conferences Steering Committee.
20. Rasmussen CE. The infinite gaussian mixture model. In: Advances in neural information processing systems. 2000. p. 554–560.
21. Escobar MD. Estimating normal means with a dirichlet process prior. J Am Stat Assoc. 1994;89(425):268–77.
22. MacEachern SN. Estimating normal means with a conjugate style dirichlet process prior. Commun Stat Simul Comput. 1994;23(3):727–41.
23. Escobar MD, West M. Bayesian density estimation and inference using mixtures. J Am Stat Assoc. 1995;90(430):577–88.
24. Blackwell D, MacQueen JB. Ferguson distributions via polya urn schemes. Ann Stat. 1973;1:353–5.
25. Sethuraman J. A constructive definition of dirichlet priors. Statistica Sinica. 1994;4:639–50.
26. Ishwaran H, James LF. Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. J Comput Graph Stat. 2002;11(3):508–32.

27. Hartigan JA, Wong MA. Algorithm as 136: A k-means clustering algorithm. J R Stat Soc. 1979;28(1):100–8.
28. Hijmans RJ, van Etten J, Cheng J, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, Bevan A, Racine EB, Shortridge A, et al. Package 'raster'. R package. 2016. https://cran.r-project.org/web/packages/raster/index.html (accessed 1 October 2016)
29. Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J, Bates D. Rcpp: Seamless r and c++ integration. J Stat Softw. 2011;40(8):1–18.
30. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71.
31. Meyer D. Support vector machines: The interface to libsvm in package e1071. 2004.