Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

A HIERARCHICAL DEPENDENT DIRICHLET PROCESS PRIOR FOR MODELLING BIRD MIGRATION PATTERNS IN THE UK

By Alex Diana[*], Eleni Matechou[*], Jim Griffin[†], and Alison Johnston[‡]

University of Kent [*], *University College London* [†] *and Cornell University*[‡]

Environmental changes in recent years have been linked to phenological shifts, which in turn are linked to the survival of species. The work in this paper is motivated by capture-recapture data on blackcaps collected by the British Trust for Ornithology as part of the Constant Effort Sites monitoring scheme. Blackcaps overwinter abroad and migrate to the UK annually for breeding purposes. We propose a novel Bayesian nonparametric approach for expressing the bivariate density of individual arrival and departure times at different sites across a number of years as a mixture model. The new model combines the ideas of the hierarchical and the dependent Dirichlet process, allowing the estimation of site-specific weights and year-specific mixture locations, which are modelled as functions of environmental covariates using a multivariate extension of the Gaussian process. The proposed modelling framework is extremely general and can be used in any context where multivariate density estimation is performed jointly across different groups and in the presence of a continuous covariate.

1. Introduction.

Describing abundance, distribution and phenology of wild animals is key to understanding the drivers of populations and therefore to designing effective conservation strategies. During this period of rapid environmental change and degradation of the natural world, it is important to develop statistical methods that utilise currently available data to provide increased understanding of species dynamics and the impact of climate change on species. The annual cycle of migratory species makes them particularly sensitive to impacts of climate change, but also makes them challenging to study. In this paper we study the phenology and abundance of migratory birds in Great Britain, in order to better understand their populations and the impacts of climate. Phenology has been linked to the survival of species, with populations that did not show a phenological response to climate change declining, as birds fail to breed at the time of maximal food abundance (Both et al., 2006; Møller et al., 2008).

Capture-recapture (CR) is one of the most commonly employed protocols in ecology to estimate the main demographic parameters of a wildlife population. CR is performed by visiting a site several times and capturing and marking a subset of the individuals before releasing them back into the population. The work in this paper is motivated by CR data on birds, collected by the British Trust for Ornithology (BTO) at different sites since 1983 as part of the Constant Effort Sites (CES) monitoring scheme, described in Peach et al. (1996). Specifically, we consider data on blackcaps, that are known to breed in the UK but overwinter in Africa. The CES scheme has been already adopted across Europe. For instance, Eglington et al. (2015) used data from constant effort ringing protocols in Western Europe to assess the productivity of several bird species, while Johnston et al. (2016) estimated annual survival from similar data.

Individuals of the same species are expected to share many of their migratory behaviours even if breeding at different sites. This led us to adopt a joint modelling approach for their migration pattern across different sites. Such a modelling approach is also motivated from the fact that fewer than 15 birds were captured at least once and fewer than 5 were captured more than once in 80% of the sites. Such small sample sizes prohibit us from studying phenology or estimating population sizes at these sites when modelling data at each site separately, as for example using the approach of Matechou et al. (2017) (MC17). Instead, a joint modelling approach enables us to study migration patterns across the UK without being limited to only using sites where large numbers of individuals are caught. There is also considerable interest in determining the effect of changes in environmental conditions due to climate change on the migration patterns of animals, including birds. In order to link phenological changes to environmental conditions, we introduce a year-specific weather covariate, specifically the average North-Atlantic Oscillation (NAO), in modelling phenology, expressed through the arrival and departure density of individuals at the different sites.

In ecological applications, parametric models often entail assumptions on the population studied that are difficult to assess in practice. In particular, as wildlife populations typically present considerable heterogeneity, the use of parametric models in ecology can be prone to model misspecification. As a result, Bayesian nonparametric models have recently been more frequently adopted. The most popular nonparametric prior employed in these applications is the Dirichlet Process (DP) prior of Ferguson (1973). The DP is a prior for densities that can be centered around any continuous distribution. However, as samples from the DP are always discrete distributions, the DP is often convolved with a continuous kernel when used as a prior for continuous distributions. The result of this convolution is called a DP mixture and gives rise to a mixture distribution with an *a-priori* infinite number of mixture components. Thanks to this flexibility, this prior has been adopted in several ecological applications. First, Dorazio et al. (2008) extended the N-mixture model of Royle (2004a) with the DP mixture of normals to allow for a variable number of mixture components in the prior distribution of population sizes. Ford et al. (2015) used a DP mixture to model heterogeneity in capture and survival probabilities in a closed population of whales. Manrique-Vallier (2016) used a DP mixture of product-Bernoulli distributions to estimate the size of a closed population in multiple CR data. Finally, MC17 used the Gamma process (Kingman, 1993), which can be expressed in terms of the DP, to model the arrival intensity of a population given a CR dataset.

Our model extends MC17 by borrowing ideas from two other popular nonparametric priors, the Hierarchical Dirichlet process (HDP) of Teh et al. (2006) and the single-p Dependent Dirichlet process (DDP) of MacEachern (1999). The former is an extension of the DP for data collected in several groups, while the latter is an extension of the DP that allows the introduction of covariates. Combining these two models, we define the Hierarchical Dependent Dirichlet process (HDDP), which can be used as the mixing measure of a continuous kernel to estimate densities as functions of continuous and categorical covariates.

As a result, our model is completely flexible in the sense that it assumes a mixture distribution with an *a-priori* infinite number of mixture components for the arrival and departure distribution at each specific site and in each year. Moreover, as a result of the clustering properties of the model, these mixture components can be shared across different sites. The ecological interpretation is that birds at different sites can belong to the same cohort, sharing similar migration behaviour, which in the model equates to one of the mixture components. Thus, even if there is no information available on the number of cohorts of birds with similar migratory behaviour, the model can naturally adapt to any number of cohorts, by varying the number of mixture components in each site-specific density.

The paper is organized as follows. In Section 2 we describe the existing model of MC17. In Section 3 we introduce the mathematical concepts necessary to define the model presented in this paper. In Section 4 we define the new model proposed. The results of fitting the model to simulated data and to the BTO data are presented in Section 5. Section 6 concludes the paper and introduces some potential future directions. The details of the sampler are presented in the appendix.

2. The existing model.

The model of MC17 performs inference from a single CR dataset. As mentioned in the introduction, CR data are collected by capturing individuals present at the site during K repeated sampling occasions. The data can be summarised in the form of a matrix H, with individual capture histories of the D caught individuals represented in the rows and the K capture occasions represented in the columns of the matrix. The capture history of individual i, H_i, corresponding to the i-th row of H, has k-th element equal to 1 if the individual was caught at the k-th sampling occasion, and equal to 0 otherwise.

The probability of capturing an individual that is present, p, is assumed to be constant across sampling occasions and common between individuals. The population size, which corresponds to the overall number of individuals that visited the site, is denoted by N.

Moreover, the model assumes that birds can enter the site at any continuous time, ζ, called the arrival time, and stay for a time δ, referred to as length of stay. The arrival time of each individual is sampled from a Poisson process with intensity $\nu(\zeta)$, which is taken to be a mixture of normal distributions $\nu(\zeta) = \int_{-\infty}^{\infty} \int_0^{\infty} \mathrm{N}(\zeta|\mu, \sigma^2)\ G(d\mu, d\sigma^2)$, where G is a Gamma process with shape αG_0 and scale τ, where $\alpha, \tau > 0$ and G_0 is a distribution function. The Gamma process is a completely random measure (Kingman, 1967), whose Levy intensity is given by $\nu(ds, dx) = \exp\left(-\frac{s}{\tau}\right) s^{-1} ds\ \alpha G_0(dx)$. It is closely related to the more popular DP, as the latter arises as normalisation of the Gamma process (Ferguson, 1973; Kingman, 1993), since the normalised random measure $P(\cdot) = \frac{G(\cdot)}{G(\Omega)}$, where Ω is the sample space, is distributed as a DP. Thanks to this property, the Gamma process can be decomposed as $G = \omega P$, where P is distributed as a DP with concentration parameter α and corresponds to the normalized density of the process and $\omega \sim \mathrm{Gamma}(\alpha, \tau)$ is the overall intensity of the process. The intensity $\nu(\zeta)$ can be expressed as

$$\nu(\zeta) = \omega \underbrace{\int_{-\infty}^{\infty} \int_0^{\infty} \mathrm{N}(\zeta|\mu, \sigma^2)\ P(d\mu, d\sigma^2)}_{f_X} \tag{2.1}$$

Given G, the sample size N is distributed as a Poisson(ω) and the arrival times $\zeta_1, \dots, \zeta_N$ are i.i.d. from f_X. The previous representation motivates the use of the intensity function, as it allows us to sample the population size and the arrival times conditionally independent on each other, as used for example in Wolpert and Ickstadt (1998).

The length of stay is modelled by a survival function with piecewise constant hazard rate f_Y. The model can be expressed through latent variables in a hierarchical form as

$$\begin{cases} H_{ik}|\zeta_i, \delta_i, p \sim \text{Bernoulli}(pz_{ik}) & i = 1, \dots, N \quad k = 1, \dots, K \\ \zeta_i \overset{i.i.d.}{\sim} f_X & i = 1, \dots, N \\ \delta_i \overset{i.i.d.}{\sim} f_Y & i = 1, \dots, N \\ N|\omega \sim \text{Poisson}(\omega) \\ \omega|\alpha, \tau \sim \text{Gamma}(\alpha, \tau) \end{cases} \tag{2.2}$$

where z_{ik} is 1 if individual i is available at sampling occasion k (if $\zeta_i < t_k < \zeta_i + \delta_i$) and 0 otherwise.

In this paper, we jointly model arrival and lengths of stay non-parametrically and extend the work of MC17 by defining the Hierarchical Dependent Dirichlet process, which allows us to jointly model data collected

- at different sites, while sharing information between sites, using the properties of the HDP, and
- across different years, accounting for the effect of a continuous covariate on migration patterns, with correlation over time modelled using a multivariate Gaussian process.

3. Theory.

3.1. *Hierarchical Dependent Dirichlet Process mixtures.* Before introducing the Hierarchical Dependent Dirichlet Process (HDDP) we present some standard models from the Bayesian nonparametrics literature.

The Dirichlet Process (DP), already mentioned in the introduction, is a random measure F with two parameters: a distribution G_0, called the base measure, and a positive real number α, called the concentration parameter, which tunes the variability of F around the base measure. It is denoted by DP(α, G_0) and it can be represented as $\sum_{i=1}^{\infty} \phi_i \delta_{\theta_i}$, with $\theta_i \sim G_0$ and the ϕ_is generated according to the stick-breaking process (Sethuraman, 1994). According to this process, given a sequence of variables $v_i \sim$ Beta$(1, \alpha)$, the weights are generated as $\phi_i = \left(\prod_{j=1}^{i-1} v_j \right) v_i$. The θ_i are often referred to as cluster locations, while the ϕ_i are called weights.

A popular extension of the DP, designed to work with data collected in different groups, is the Hierarchical Dirichlet process of Teh et al. (2006). In order to model data from different groups, the HDP assumes a random measure, F_j, for the j-th group, and a global random probability measure F_0. The global measure is assumed to have a DP prior $F_0 \sim$ DP(γ, G_0), while the group-specific random measures have independent DP prior $F_j \sim \mathrm{DP}(\alpha, F_0)$. Parameter γ tunes the variability of F_0 around G_0 and α tunes the variability of F_j around F_0. According to the stick-breaking representation, $F_0 = \sum_{i=1}^{\infty} \phi_i \delta_{\theta_i}$ and $F_j = \sum_{i=1}^{\infty} \pi_{ij} \delta_{\theta_i}$, and the distribution of the weights $\pi_{\cdot j}$ can be obtained in closed form as $\pi_{kj} = \left(\prod_{i=1}^{k-1} v_{ij}\right) v_{kj}$ where $v_{kj} \mid (\alpha, \phi_1, \ldots, \phi_k) \sim \mathrm{Beta}\left(\alpha\phi_k, \alpha\left(1 - \sum_{l=1}^{k} \phi_l\right)\right)$. Hence, every F_j is essentially obtained by keeping the same atoms of F_0 but redistributing the weights. No variation is induced in the cluster locations of the group-specific DPs.

The HDP is often conveniently described via the Chinese restaurant franchise (CRF) representation. According to the CRF representation, every observation in a group corresponds to a customer in a restaurant. In addition, the cluster locations of F_0 $\theta_1, \ldots, \theta_K \overset{i.i.d.}{\sim} G_0$, represent the dishes that can be served in the restaurant. To link the customers to the dishes, customer i in restaurant j is assigned to a table t_{ij}, while table t in restaurant j is assigned to dish k_{jt}. As a consequence, the dish served to customer i in restaurant j is $k_{jt_{ij}}$, which we define as c_{ij}. In addition, following the notation established in the literature, n_{jt} denotes the number of customers sitting at table t in restaurant j, m_k the number of tables serving dish k and M the total number of tables.

Thanks to the CRF representation, we can express the distribution of the allocations c_{ij} of customers to dishes by first defining the distribution of allocations t_{ij} of customers to tables and then the distribution of the allocations k_{jt} of tables to dishes. We can generate a sample from the CRF by sampling iteratively according to the following scheme. A new customer is assigned to an

$$\begin{cases} \text{existing table } t & \text{with probability} \quad \frac{n_{jt}}{n_{jt}+\alpha} \\ \text{new table } t^\star \text{ serving existing dish } \theta_k & \text{with probability} \quad \frac{\alpha}{n_{jt}+\alpha} \; \frac{m_k}{M+\gamma} \\ \text{new table } t^\star \text{ with new dish } \theta_{k^\star} \sim G_0 & \text{with probability} \quad \frac{\alpha}{n_{jt}+\alpha} \; \frac{\gamma}{M+\gamma} \end{cases}$$

Likewise, a table is assigned to an

$$\begin{cases} \text{existing dish } \theta_k & \text{with probability} \quad \frac{m_k}{M+\gamma} \\ \text{new dish } \theta_{k^\star} \sim G_0 & \text{with probability} \quad \frac{\gamma}{M+\gamma} \end{cases}$$

The implied distribution on the c_{ij} is defined as $\mathrm{CRF}(\alpha, \gamma)$.

In the application to the BTO dataset, the birds are represented by the customers and the dishes correspond to the same migratory behaviour. Thanks to the CRF, groups of birds belonging to different sites can still share the same migratory behaviour if they are assigned to tables serving the same dish.

Another extension of the DP, designed to work with general covariates, is the Dependent Dirichlet Process (DDP) of MacEachern (1999). The DDP is a random measure F_x that can be written as

$$F_x = \sum_{k=1}^{\infty} \phi_k \delta_{\theta_i(x)}$$

where the cluster locations $\theta_i(x)$ are drawn independently from a stochastic process G_x, allowing F_x to depend on continuous covariates, if a continuous process, such as a GP, is assumed for G_x. The weights ϕ_i are drawn from the stick-breaking process as in the standard DP. More information on other nonparametric priors can be found in Hjort et al. (2010).

In this paper, we perform density estimation conditionally on general covariates in a context where we have several groups. To achieve this, we combine the idea of the HDP and the DDP defining the Hierarchical Dependent Dirichlet Process (HDDP) as a HDP where the DP F_0 in the top level is replaced by a DDP.

Definition 3.0.1. *Let G_x be a stochastic process. The measures F_{jx} are said to follow a HDDP prior if, for each group j and each value x of the covariate*

$$\begin{cases} F_x = \sum_{k=1}^{\infty} \phi_k \delta_{\theta_k(x)} & \theta_k(x) \sim G_x \\ F_{jx} = \sum_{k=1}^{\infty} \pi_{kj} \delta_{\theta_k(x)} \end{cases} \tag{3.1}$$

where the weights ϕ_k and π_{kj} follow the same distribution as the weights of the HDP.

As we can see, the covariate, x, is introduced in the top-level and not in the group-specific DPs, which implies that the effect of the covariate is assumed to be the same across groups. However, as the DDP is assumed as a prior distribution for the group-specific measures, the weights are constant for each value of the covariate.

As opposed to a standard dataset analysed in Teh et al. (2006), our data have an additional third dimension, given by the covariate x. However, as mentioned above, the covariate only affects the cluster locations. As a result, the CRF representation of the HDP can be used to describe the HDDP, since the covariate does not play a role when assigning the observations to clusters.

To conclude, we term as Hierarchical Dependent Dirichlet process mixtures the process obtained when the HDDP is used as the mixing measure of the parameters of a continuous kernel.

3.2. *Multivariate Gaussian Process (MGP).*

Before introducing the MGP we start by describing the univariate version. A GP is a prior distribution on a function $f : \mathbb{R}^q \to \mathbb{R}$, defined by the distribution of f evaluated on any finite collection of points $(x_1, \dots, x_n)$. Specifically, we write $f \sim \text{GP}(0, k)$ if, for any $(x_1, \dots, x_n) : x_i \in \mathbb{R}^q$

$$(f(x_1), \dots, f(x_n)) \sim \text{N}(0, K((x_1, \dots, x_n), (x_1, \dots, x_n))$$

where $\{K((x_1, \dots, x_n), (x_1, \dots, x_n))\}_{ij} = \sigma^2 k(x_i, x_j)$ and k is a correlation function. In our case, we consider the Gaussian radial basis function $k(x, x') = \exp\left(-\frac{|x-x'|^2}{l^2}\right)$, with $l > 0$. For more information on Gaussian processes, see Rasmussen (2006).

In the case of multivariate data, that is, if f is a function from $\mathbb{R}^q$ to $\mathbb{R}^p$, the MGP prior is defined based on the matrix normal distribution. A variable X is said to follow a matrix normal distribution $\text{MN}(M, U, V)$ if $vec(X) \sim \text{N}(vec(M), V \otimes U)$, where U is called the among row covariance matrix, V is called the among column covariance matrix and $\otimes$ is the Kronecker product.

The MPG prior on f is defined in the following way.

Definition 3.0.2. *Let Σ be a $p \times p$ positive definite matrix and μ an $n \times p$ matrix. We say that $f = (f_1, \dots, f_p) \sim MGP(\mu, K, \Sigma)$ if*

$$((f_1(x_1), \dots, f_p(x_1)), \dots, (f_1(x_n), \dots, f_p(x_n))) \sim MN(\mu, K((x_1, \dots, x_n), (x_1, \dots, x_n)), \Sigma).$$

This construction of the multivariate Gaussian process is also presented in Chen et al. (2017). By defining the MGP in terms of the matrix normal distribution, we have implicitly assumed that the cross-covariance matrix of the vector $((f_1(x_1), \dots, f_p(x_1)), \dots, (f_1(x_n), \dots, f_p(x_n)))$ is separable, that is, it can be factorised as $\Sigma \otimes K((x_1, \dots, x_n), (x_1, \dots, x_n))$.

The advantage of this construction is that if we assume that the observations $\boldsymbol{y}_i = (y_{i1}, \dots, y_{ip})$ are generated according to

$$\boldsymbol{y}_i \sim \mathrm{N}((f_1(x_i), \dots, f_p(x_i))^T, \Sigma)$$

the posterior predictive distribution of new observations is available in closed form. Assuming we have available observations $(x, \boldsymbol{y})$, the posterior predictive distribution for new observations with covariates $x^\star$ is:

$$\boldsymbol{y}^\star \sim \mathrm{N}(vec(\mu_2) + (K_\star(K+I)^{-1} \otimes I)(\boldsymbol{y} - vec(\mu_1), ((K(x^\star, x^\star) + I) - K_\star(K+I)^{-1}{K_\star}^T), \Sigma)$$

where $K := K(x, x)$, $K_\star := K(x, x^\star)$.

In addition, we can account for the effect of covariates on the mean. If we have R covariates arranged in an $n \times R$ matrix X and coefficients $\beta = \begin{bmatrix} \beta_1^1 & \dots & \beta_1^p \\ \vdots & & \vdots \\ \beta_R^1 & \dots & \beta_R^p \end{bmatrix}$, we define the MGP as:

$$((f_1(x_1), \dots, f_p(x_1)), \dots, (f_1(x_n), \dots, f_p(x_n))) \sim \mathrm{MN}(X\beta, K((x_1, \dots, x_n), (x_1, \dots, x_n)), \Sigma)$$

A useful property of this construction is that, if a prior distribution $\mathrm{MN}(b, B, \Sigma)$ is assumed for β, the marginal distribution of f is still a MGP prior of the form $\mathrm{MGP}(K^\star(K^{-1}XB^\star)B^{-1}b, K^\star, \Sigma)$, with $K^\star = (K + XBX^T)^{-1}$ and $B^\star = K^{-1}X(X^TK^{-1}X + B^{-1})^{-1}$. The calculations can be found in the supplementary material (Diana et al., 2018).

4. Bayesian nonparametric model for CR data collected at multiple sites and multiple years. The data can be expressed in the form H_{ijy}, where H_{ijy} is the capture history, defined in Section 2, of individual i at site j in year y, and we perform sampling at J sites in Y different years. At site j and year y, captures take place on C_{jy} sampling occasions at times $t_1^{jy}, \dots, t_{C_{jy}}^{jy}$. Sampling times and the number of sampling occasions may differ across sites and years. We denote by x_y the value of the year-specific environmental covariate associated with year y. The site and year specific covariate associated with capture probability at site j and year y is denoted by λ_{jy}.

4.1. *Sampling scheme.* Capture probabilities are modelled using a logistic mixed effects model, where the site-specific intercept is assumed to be constant across years in the same group and all intercepts share a common prior distribution. The model for capture probability at site j in year y can be written as

$$\begin{cases} \text{logit}(p_{jy}) = \alpha_j^p + \lambda_{jy}\beta^p \\ \beta^p \sim \text{N}(0, B^p) \\ \alpha_j^p \sim \text{N}(a_0^p, A_0^p) \end{cases}$$

where B^p is the prior variance of β^p and a_0^p, A_0^p are chosen according to expert knowledge.

The choice of a mixed effects model is motivated by the study design of the CES scheme, according to which, sampling at the different sites is performed with the same effort. However, additional site characteristics, such as habitat and structure of the site, present an additional source of variation affecting capture probability that is not explained by the covariate, but instead modelled by the site-varying intercepts.

4.2. *Arrival and Departure Process.* We denote by ζ_{ijy} and δ_{ijy} respectively the arrival time and length of stay of individual i at site j in year y. We do not work directly with arrival and departure times because these two quantities do not lie in $\mathbb{R}^2$ (departure is obviously always later than arrival) and this would imply the need to work with a bivariate truncated normal, for which conjugate schemes are not available, resulting in computationally intensive inference. Instead, we choose to work with arrival times and a transformation of the length of stay, $\eta := h(\delta)$, in order to make the latter lie in $\mathbb{R}$. Although the logarithm is the common choice, it would lead to a lognormal behaviour in the right tail once we assign a normal prior distribution to $h(\delta)$, as the tails of the DP mixture behave approximately as the tails of the kernel. In order to have a normal behaviour also in the right tail, we choose $h(x) = \begin{cases} \log(x) & x \leq 1 \\ x - 1 & x > 1 \end{cases}$.

Borrowing ideas from MC17, we assume that for each site, arrival times and transformed lengths of stay are drawn from a Poisson process with non-homogeneous intensity ν_{jy}, modelled as

$$\nu_{jy}(\zeta, \eta) = \omega_j \int \text{N}(\zeta, \eta | \mu_{x_y}, \Sigma) dP_{jy}(\mu_{x_y}, \Sigma) \tag{4.1}$$

where P_{jy} is the year and site-specific mixing measure of the parameters μ_{x_y} and Σ of the normal distribution, and ω_j is the site-specific intensity. The link with MC17 is clear if we compare (4.1) with (2.1). The bivariate density ν_{jy} of arrival times and lengths of stay is allowed to be site and year dependent, by replacing the DP with a HDDP, unlike MC17, who use a univariate DP mixture.

To achieve this, we define $\theta = (\mu, \Sigma, \beta)$, where μ is the $Y \times 2$ matrix of all the means μ_{x_y} of arrival and departure times for each covariate value, Σ is the 2×2 covariance matrix, β is an $R \times 2$ matrix expressing the trend of the means across the years and R is the dimension of the year-specific covariate (including the intercept). The prior distributions for these quantities are

$$\begin{cases} \mu \sim \mathrm{MGP}(X\beta, K(\boldsymbol{y}, \boldsymbol{y}), \Sigma) \\ \Sigma \sim \mathrm{IW}(\nu_0, \Sigma_0) \\ \beta \sim \mathrm{MN}(b, B, \Sigma) \end{cases} \tag{4.2}$$

where IW is the inverse-Wishart distribution, ν_0 is the number of degrees of freedom, $\mathbb{E}[\mathrm{IW}(\nu_0, \Sigma_0)] = \frac{1}{\nu_0 - 3}\Sigma_0$, $X = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_Y \end{bmatrix}^T$, $\boldsymbol{y} = (1, \dots, Y)$, b is an $R \times 2$ matrix and B is an $R \times R$ matrix.

The measure $P_{jy}(\theta)$ is allowed to be year and site dependent by assuming the HDDP prior defined in (3.1), where P_{jy} has the same prior as the F_{jx}. As shown in the appendix, the choice of such prior distribution for θ will allow us to make straightforward posterior inference when the measure F_{jx} is convolved with a bivariate Gaussian kernel, as in our case. Keeping in mind the explicit expression of the DP, the resulting model for a specific year and site can be written as

$$f_{jy}(\zeta, \eta) = \sum_{k=1}^{\infty} \pi_{kj} N(\zeta, \eta | (\mu_k)_y, \Sigma_k)$$

where the (μ_k, Σ_k) are shared between groups.

Every cluster has its own regression coefficient β with a common prior distribution $\mathrm{MN}(b, B, \Sigma)$. However, in order to estimate the overall trend across all clusters, we assign an additional hyperprior distribution $b \sim \mathrm{MN}(b_0, B_0, \Sigma_b)$. The posterior distribution for b will give the overall trend of arrival and length of stay for all groups across the years.

For the overall intensity of the process, ω_j, we keep the same prior distribution as in the Gamma process case but in order to share information between sites and years, we assume that intensities now have a prior distribution $\omega_j | \alpha, \tau_j \sim \mathrm{Gamma}(\alpha, \frac{\alpha}{\tau_j})$ where α is the standard shape and τ_j is the mean of the Gamma distribution. The parameters τ_j, α and γ are assumed to have Gamma prior distributions, which is a standard choice.

The model can be summarised with the introduction of latent variables.

$$
\left\{
\begin{array}{ll}
H_{ijyl} \mid \zeta_{ijy}, \eta_{ijy}, p_{ij} \sim \text{Bernoulli}(p_{ij} z_{ijyl}) & i = 1, \dots, N_{jy} \quad j = 1, \dots, J \quad y = 1, \dots, Y \quad l = 1, \dots, C_{jy} \\
z_{ijyl} = \begin{cases} 1 & \text{if } \zeta_{ijy} < t_l^{jy} < \zeta_{ijy} + \delta_{ijy} \\ 0 & \text{otherwise} \end{cases} & i = 1, \dots, N_{jy} \quad j = 1, \dots, J \quad y = 1, \dots, Y \quad l = 1, \dots, C_{jy} \\
(\zeta_{ijy}, \eta_{ijy}) \mid c_{ijy}, \{\mu_k\}, \{\Sigma_k\} \sim \text{N}((\mu_{c_{ijy}})_{x_y}, \Sigma_{c_{ijy}}) & i = 1, \dots, N_{jy} \quad j = 1, \dots, J \quad y = 1, \dots, Y \\
\mu_k \sim \text{MGP}(X\beta_k, K(x, x), \Sigma_k) & k = 1, \dots, K \\
\Sigma_k \sim \text{IW}(\nu_0, \Sigma_0) & k = 1, \dots, K \\
\beta_k \sim \text{MN}(b, B, \Sigma_k) & k = 1, \dots, K \\
b \sim \text{MN}(b_0, B_0, \Sigma_b) & \\
c_{ijy} \sim \text{CRF}(\alpha, \gamma) & i = 1, \dots, \sum_{y=1}^{Y} N_{jy} \quad j = 1, \dots, J \quad y = 1, \dots, Y \\
N_{jy} | \omega_j \sim \text{Poisson}(\omega_j) & j = 1, \dots, J \qquad\qquad y = 1, \dots, Y \\
\omega_j | \alpha, \tau_j \sim \text{Gamma}\left(\alpha, \frac{\alpha}{\tau_j}\right) & j = 1, \dots, J
\end{array}
\right.
$$

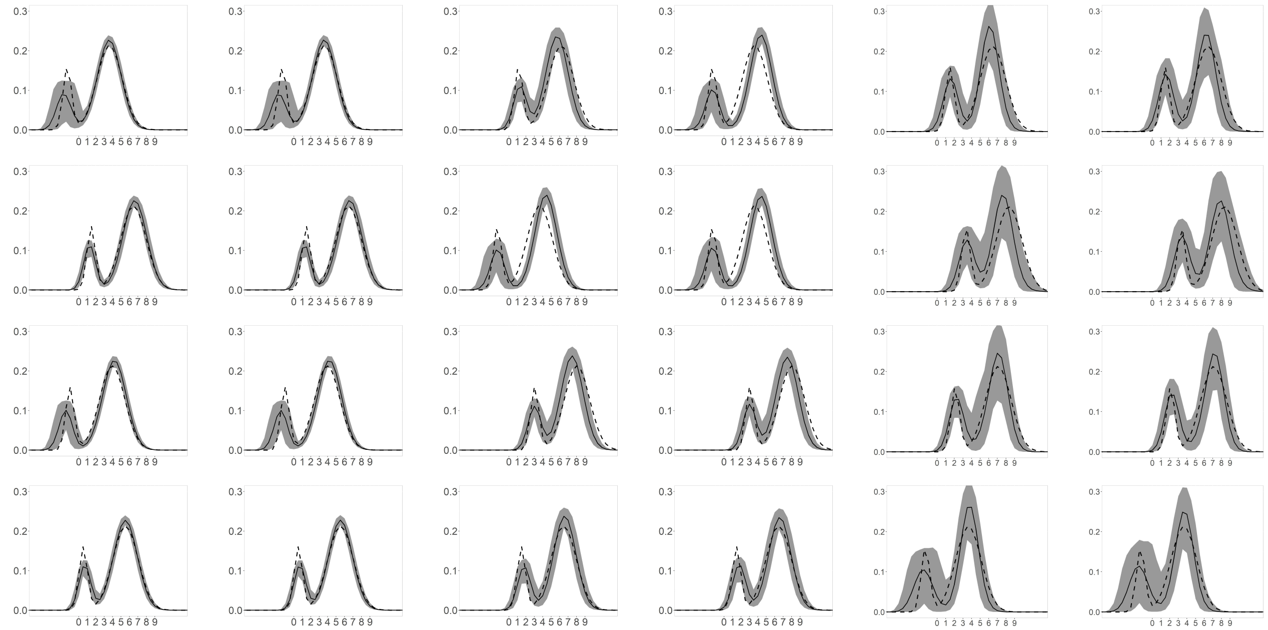
where K is the total number of clusters and in the CRF assignments of c_{ijy} the variable j indexes the groups and i and y index the observations.

5. Application.

5.1. *Simulations.* In order to assess the performance of the model, we have simulated several sets of data and compared the posterior distributions of the main quantities of interest with the true values used to simulate the data. The simulated data consist of 2 sites and 16 years, with 10 sampling occasions in each year. In order to have population sizes similar to the ones in the CES data, the site-specific intensities ω_j of the prior distribution of the population sizes are sampled from a Gamma distribution with mean 60 and variance 200, population sizes for each year are then sampled from a Poisson with the intensity ω_j sampled above. Arrival times and lengths of stay are sampled keeping in mind the CES data, which consist of a mixture of individuals with different patterns of arrival and stay. In particular, it is known (Johnston et al., 2016) that there are two groups of birds that use the sites; "residents" that breed at the sites and may return in subsequent years, and "transients" that pass through the site on the way to breeding grounds further north, or wintering grounds further south. To model this behaviour, we sample from the following mixture distribution

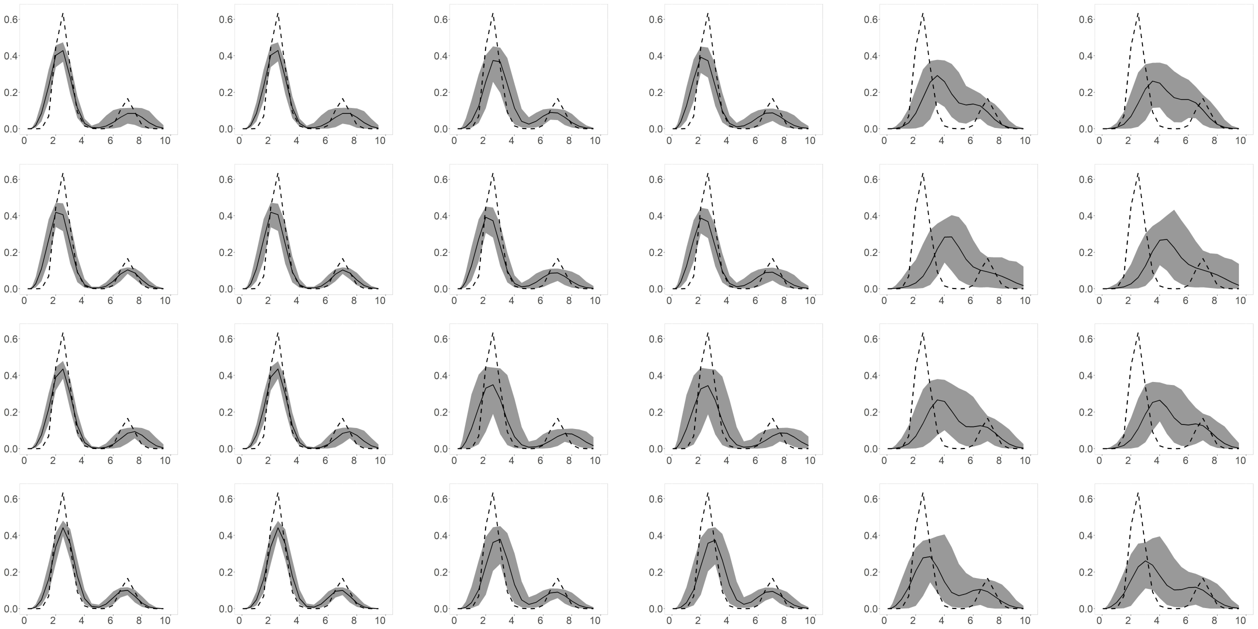
$$
\begin{bmatrix} \zeta_{ijy} \\ \delta_{ijy} \end{bmatrix} \sim 0.8 \begin{bmatrix} \text{N}(6 + 1\ x_y, 1.5) \\ \text{Gamma}(25, 10) \end{bmatrix} + 0.2 \begin{bmatrix} \text{N}(1 + 1\ x_y, .5) \\ \text{Gamma}(210, 30) \end{bmatrix}
$$

where y indexes the year. The values x_y of the covariate are sampled from a $\text{N}(0, 1)$.



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig 1: Arrival times - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true distribution used to simulate the data and the grey area represents the 95% posterior credible interval (PCI).



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig 2: Lengths of stay - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true distribution used to simulate the data and the grey area represents the 95% PCI.

We performed three sets of simulations, each of them with different values of the capture probabilities. We sample values from a logistic-normal with scale 0.1 and location equal to, respectively 1, 0 and -1 for the three sets of simulations, which corresponds to capture probabilities centred around, respectively, 0.73, 0.5 and 0.26.

In order to choose the value of the length scale parameter l of the MGP, we have performed a sensitivity analysis considering the values 0.1, 0.3 and 0.5, obtaining practically identical results. Thus we fixed the value to 0.3, as values outside the range considered would give a correlation between close points which is either too large or too small for our application.

The posterior distributions of the arrival densities and lengths of stay for the three sets of simulations are shown (for a subset of 4 years), respectively, in Fig. 1 and 2. In the case of the arrival densities, the posterior mean densities closely resemble the true densities. As capture probability decreases, the estimates present, as expected, more variance and the model splits one of the modes in two separate clusters. In the case of the lengths of stay, for all simulated data the posterior mean density is smoother than the true distribution, a fact that becomes progressively more evident as capture probability decreases.

The posterior distributions of the regression coefficients b_{21} and b_{22} are shown in Fig. 3. The estimates of the posterior means are similar and close to the true values, but the cases with lower capture

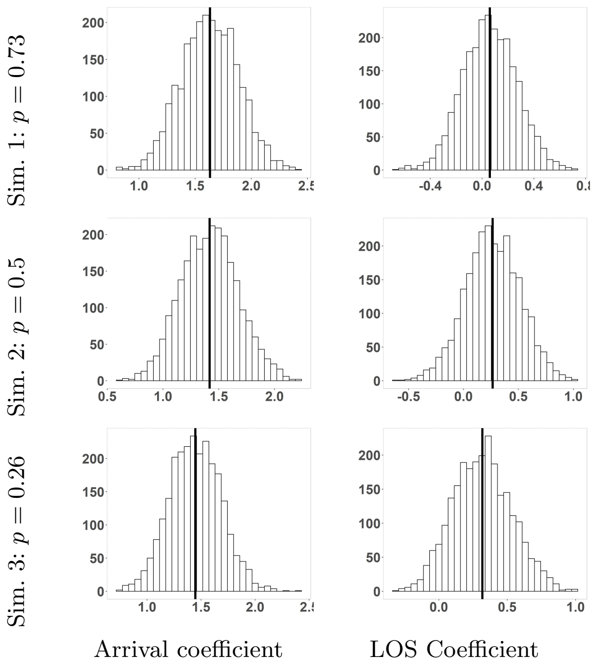
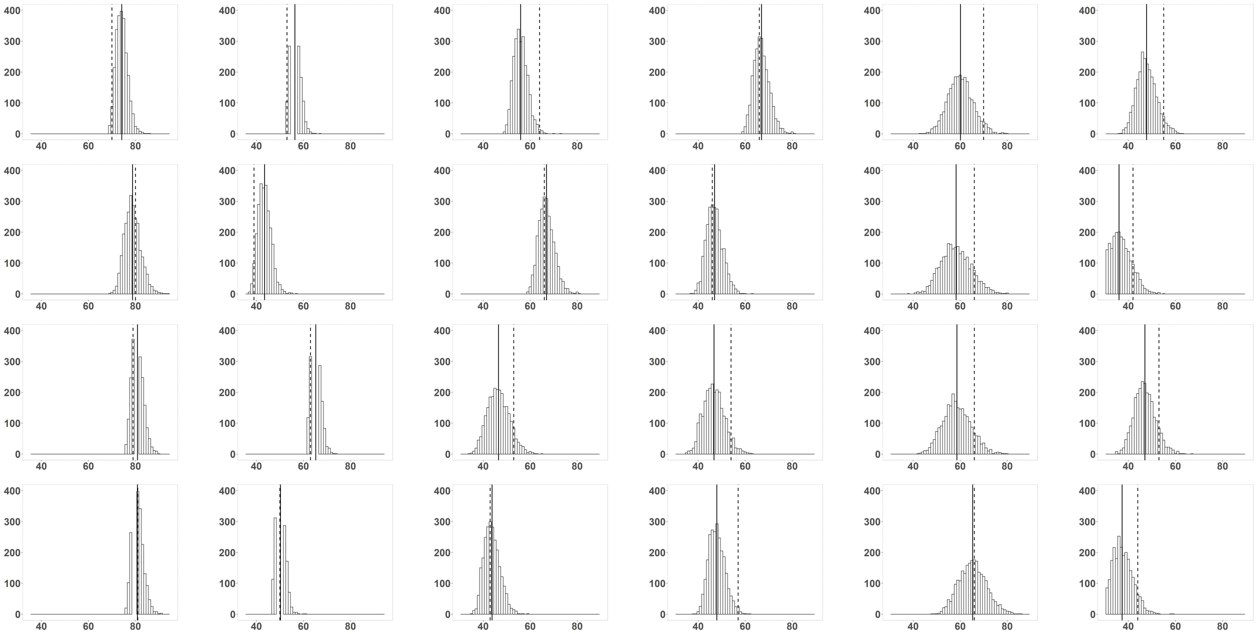
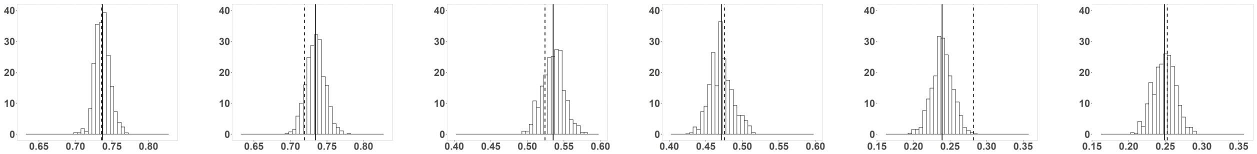


Fig 3: Posterior distributions of the regression coefficients of arrival times, b_{21}, and length of stay, b_{22}, for the three sets of simulations, with the solid line representing the posterior mean. The true value is fixed at 1.5 for the arrival times and 0 for the lengths of stay.



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig 4: Population sizes - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true population size.



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig 5: Capture probabilities - Posterior densities for the three sets of simulations. The solid line represents the posterior mean, the dashed line represents the true capture probability.

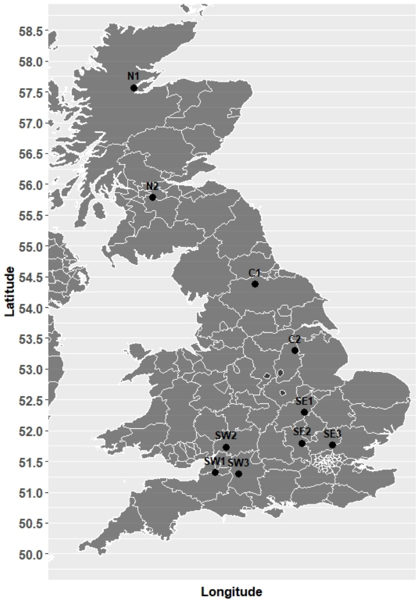


Fig 6: Map of the CES sites used in the analysis, with site ID shown above the sites.

probabilities exhibit more variance in the estimates. The posterior distributions of the population sizes are shown in Fig. 4, where it can be seen that, aside from the case with lowest capture probability, the posterior mean is generally close to the true value, which is always included in the corresponding 95% PCI. Clearly, population size is consistently either over-estimated or under-estimated at some site. This is due to the model assuming that mean population size at each site is constant over time. The posterior distributions of the capture probabilities are shown in Fig. 5. As was the case when inferring population size, the posterior variance increases as capture probability decreases.

5.2. *BTO's Constant Effort Sampling Scheme Data.* We apply the model to CR data of blackcaps collected by the BTO at several breeding and stopover sites across the UK. We discarded all the juvenile birds as, being born at the site in the same year they are captured, they do not provide any information on the arrival density. Even though the complete data consist of more than 100 sites for more than 20 years, we work on a subset of 10 sites across 16 consecutive years, from 1998 to 2013, with a total of 3401 birds caught, as working with the entire data would not be feasible in terms of computational time. We selected these 10 sites by choosing the subset where sampling occurred for the highest number of consecutive years, because we are interested in estimating the regression coefficient for the year-continuous covariate. The locations of the sites are indicated on the map shown in Fig. 6.

The prior specification is based on previous studies (Peach et al., 1996; Johnston et al., 2016). Arrival times and lengths of stay are modelled in weeks, and their prior distribution is chosen to have

95% of the mass of the arrival distribution from three weeks before the start of the sampling period up to the end of it, and 95% of the mass of the departure distribution from the start of the sampling period, up to three weeks after the end. The prior on the capture probability and the prior on the mean τ_j of the intensity of the population size are shown in the supplementary material (Diana et al., 2018).

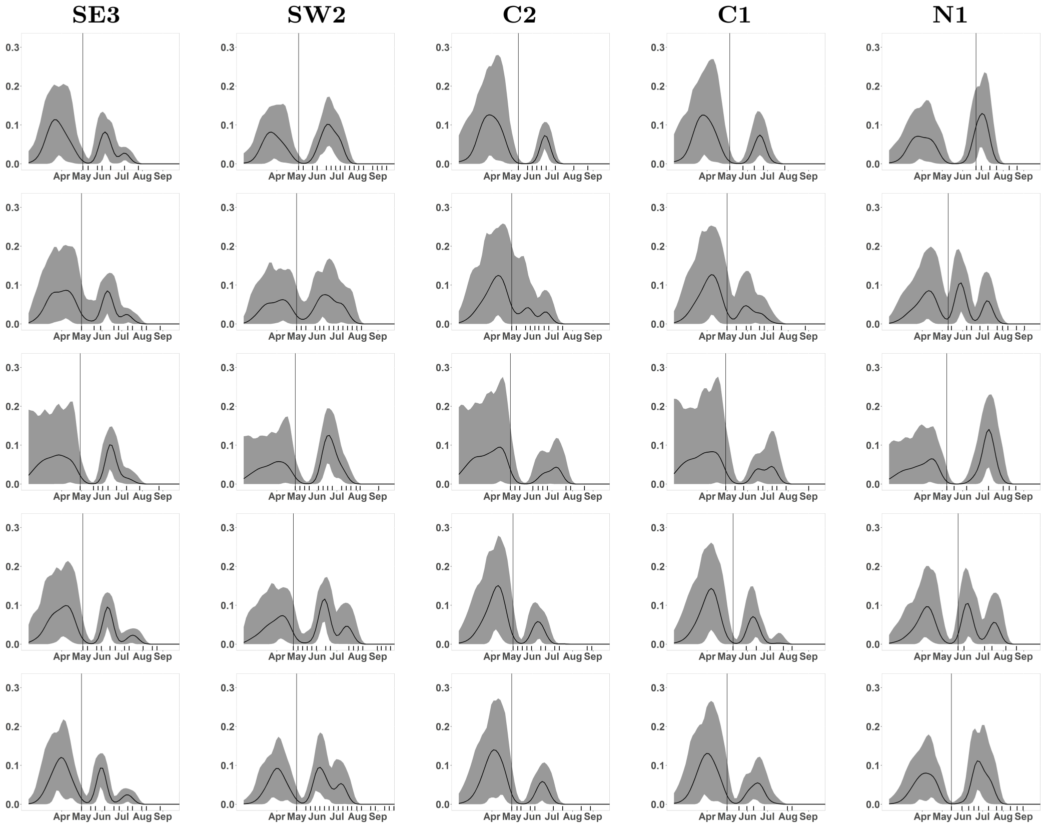


Fig 7: Arrival times - Posterior distribution for a subset of 5 sites, with site names given on top of each column, and 5 years. The black line shows the posterior mean density and the grey area shows the 95% PCI. The sampling occasions are shown in bold on the x-axis and the black line shows the first sampling occasion.

As a year-specific covariate, we use the average North-Atlantic Oscillation (NAO) in the months from January to April, as these are the months preceding the sampling period. This choice is motivated by the fact that the NAO is thought to represent the overall trend of global temperatures. The covariate λ_{jy} used to model the capture probability is the length of the net placed at each site.

We present results for 5 sites, out of the 10 shown in Fig. 6, for years 2003, 2005, 2007, 2009 and 2011. Additional plots can be found in the supplementary material (Diana et al., 2018). Between these

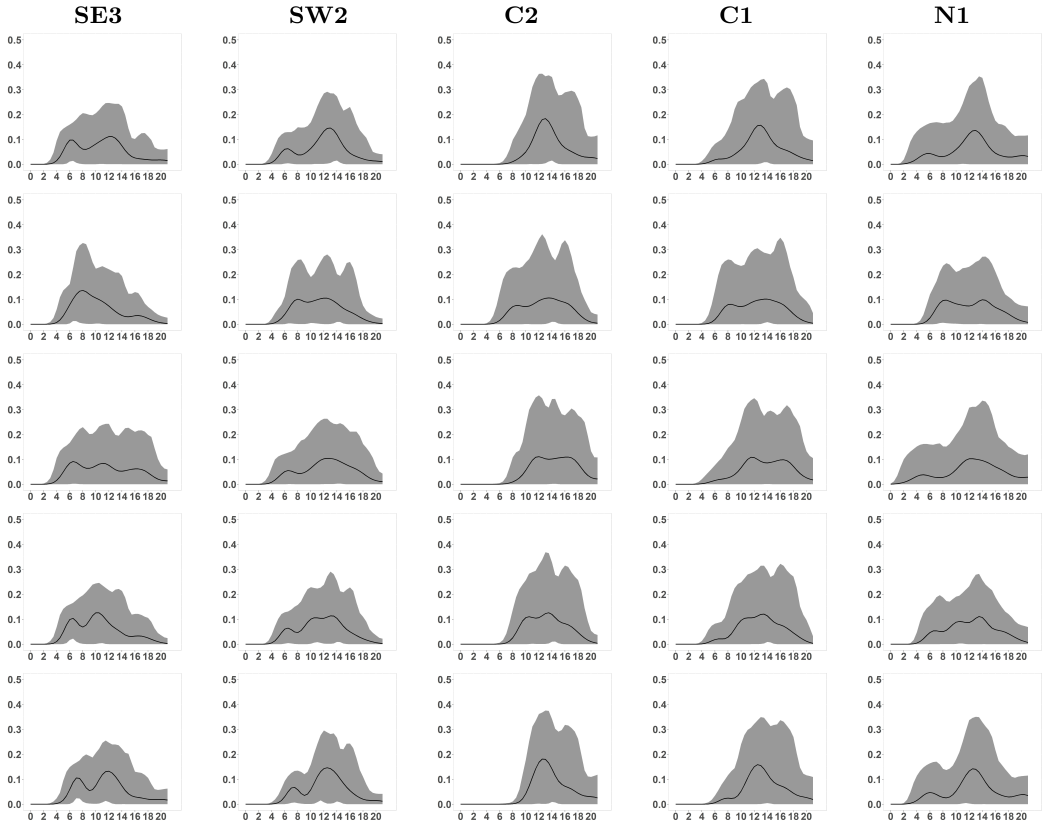


Fig 8: Lengths of stay - Posterior distribution for a subset of 5 sites, with site names given on top of each column, and 5 years. The black line shows the posterior mean density and the grey area shows the 95% PCI.

sites, we chose two sites in the South, two in the center and one in the North, in order to highlight differences in the densities for sites at several latitudes. We first focus on the arrival distributions, shown in Fig. 7. All of the distributions present a mode before the first sampling occasion, which can be interpreted as the result of the many individuals arriving before sampling has begun. In fact, all of the data show a high number of captures in the first and second sampling occasions, while the number decreases in the middle of the sampling period. The remainder of the peaks are likely to correspond to the transient birds arriving at the sites later in the season only for feeding. It can also be noticed that northern sites (e.g. $N1$ and $N2$) present a higher number of birds arriving later in the season, suggesting that the birds arriving in the UK stop first at the southern sites before reaching the sites in the north. The length of stay densities, presented in Fig. 8, also exhibit several peaks because of the presence of the breeding birds and transient birds. However, due to the lack of data in some of the sites, the two modes are likely to merge in some cases.

Population sizes for the same sites and years as those considered in Fig. 7 are presented in Fig. 9. Comparison with the posterior densities of capture probabilities in Fig. 10 shows that, as expected, smaller estimates of the capture probability are generally associated with greater variance in population size estimates. Moreover, northern sites present overall lower population sizes than southern sites.

The posterior distributions of the coefficients are shown in Fig. 11. The 95% PCIs of the arrival time and length of stay components of the regression coefficient b include 0, suggesting that the NAO has no effect on the patterns of arrival and length of stay, which agrees with previous findings (Robson and Barriocanal, 2011; Gienapp et al., 2007). However, this does not necessarily imply that the arrival and length of stay distributions in the clusters do not exhibit trends across the years, but it might be that some clusters have positive shifting trends while others have negative shifting trends, which would imply an overall posterior close to 0.

6. Discussion. In this paper we have developed a model to estimate arrival and departure distributions in a multi-site and multi-year capture-recapture data set with annual environmental covariates and site-specific variation, and applied this model to real data. Moreover, we have performed a simulation study to assess the validity of the model on simulated data with similar features, obtaining encouraging results even when capture probability is low.

The dataset used in our application consists of a mixture of breeding and transient birds. Although breeding birds tend to return to the same site in different years, transient birds change the site they visit from year to year. As a result, changes in population sizes at each given site across the years do not reflect an actual change in the number of birds of the population. For this reason, in section 4.2, we chose not to adopt a model for the evolution of population sizes over time but we only assume that population sizes are sampled from the same common distribution. Because of the lack of site-fidelity of blackcaps, changes in the populations' behaviour are not evaluated by analyzing the evolution of population sizes but instead by observing the changes in phenology, summarised in the arrival and departure distribution for each site and year, in relation to an indicator of global temperature, as the NAO in our case.

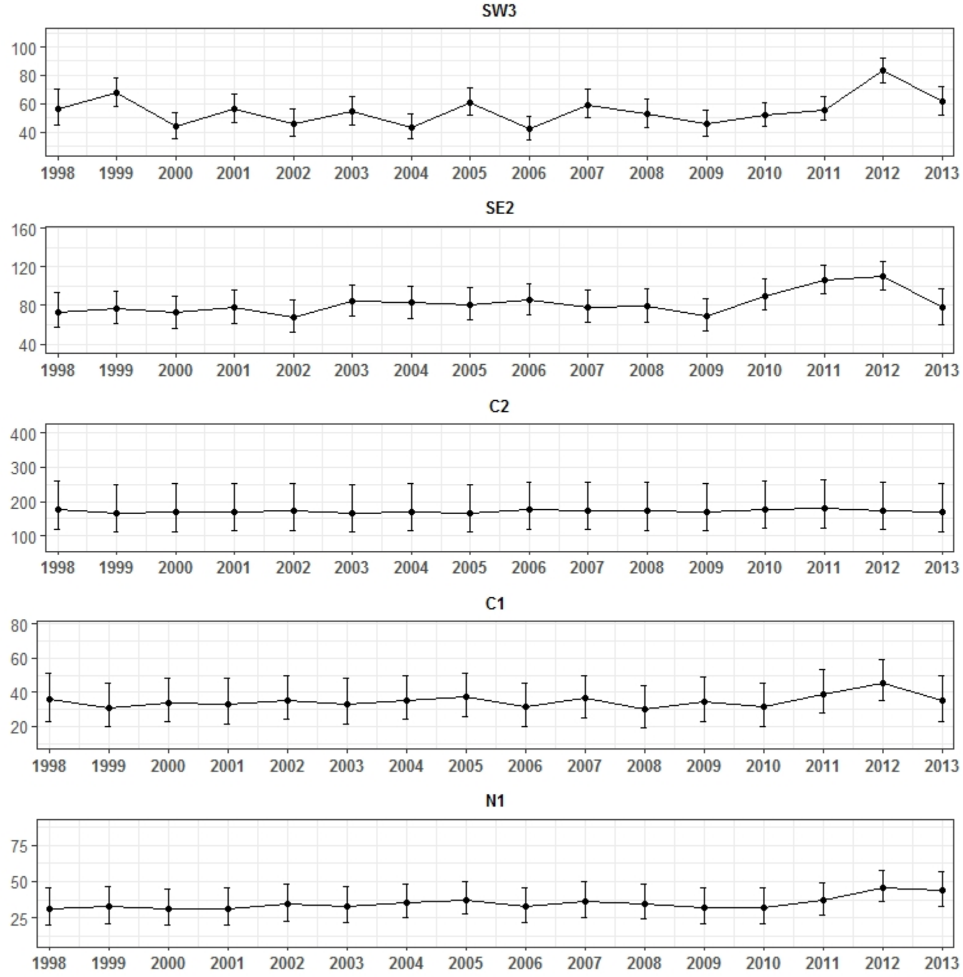


Fig 9: Posterior densities of the population sizes for a subset of 5 sites, with site names given on top of each plot. The bars show the 95% credible intervals, while the dots show the posterior means.

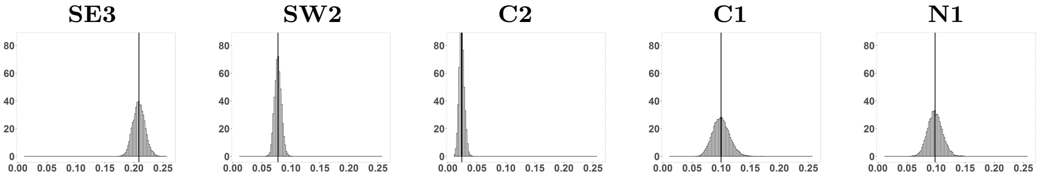


Fig 10: Posterior densities of the capture probabilities for a subset of 5 sites, with site names given on top of each column. The black vertical line shows the posterior mean.

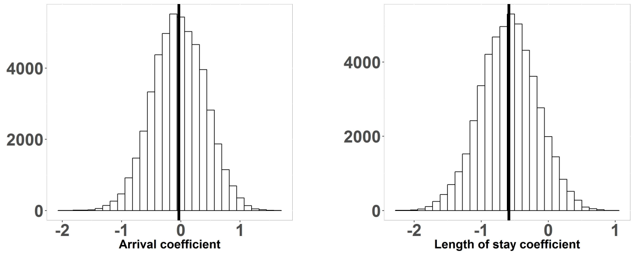


Fig 11: Posterior distribution of the arrival time and length of stay components of the regression coefficient b. The black line represents the posterior mean.

In this model, we did not track the same individuals across the years. The choice is motivated by the fact that the number of recaptures of the same individuals in different years is too low to motivate such a modelling approach. However, the model could be further extended in the cases of species exhibiting higher longevity and site-fidelity than the blackcaps.

We have followed the approach of MC17, using a Bayesian nonparametric approach to estimate the arrival and departure densities. However, MC17 only allow the estimation of the arrival density for a single site and year. In our work, we first added an additional level of complexity modelling nonparametrically both the arrival and departure density. Then, as our goal was to perform density estimation in several sites conditional on year-specific covariates, we extended their model to account for these additional effects. The starting point to achieve a joint modelling of data collected at different sites is the use of the Hierarchical Dirichlet Process (HDP) of Teh et al. (2006) in place of the DP. However, since this model does not allow the introduction of continuous covariates, we further extend the HDP defining the Hierarchical Dependent Dirichet Process. Lastly, to introduce a correlation structure over time, we started from the idea of the Gaussian process (GP) and, as we are modelling arrival and length of stay jointly and the GP can only give univariate outputs, we adopted an extension of the GP to the case of p-dimensional outputs. Another interesting definition of the GP with multiple outputs can be found in Álvarez and Lawrence (2011). However, the advantage of our construction is that we still maintain the useful conjugacy properties of the GP, which allows us to straightforwardly use the sampling schemes available for the HDP.

The Bayesian nonparametric model defined here is extremely general and can be used in a generic context where multivariate density estimation is to be performed jointly across different groups and in the presence of a continuous covariate, which extends the model presented in De Iorio et al. (2004).

As it is clear from equation (3.1), the model can account for covariates only in the cluster locations, however Griffin and Leisen (2017) have defined a nonparametric model, known as compound random measures, which can account for covariates in the cluster weights. However, in the case of compound random measures, inference is more difficult as the sampling scheme based on the CRF cannot be used anymore. Moreover, our model allows the introduction of covariates only across time, while in some scenarios it could be useful to adopt spatial covariates, for example the latitude of the site, in order to account for differences in arrival patterns according to site-specific covariates. Even if our model accounts for additional random effects from site to site, explaining the variation through covariates would require a change to the structure of the model.

In Section 5.2, we mention that we choose a subset of the data in order to be able to run the algorithm in a feasible computational time. In fact, given the large number of observations, one of the challenges of the model is the computational complexity, which scales linearly with the number of observations. This is a common drawback of all algorithms based on the Chinese Restaurant representation, as the sampler requires to update the cluster allocations of each individual by computing the probability of belonging to each cluster, which is a computationally expensive operation. Sampling from the posterior of DP mixtures without having to update the cluster allocations as in Escobar and West (1995) is still an open problem, and it goes beyond the scope of this paper. A potentially faster algorithm to sample from a DP mixture model, based on parallel computation, has been proposed by Ge et al. (2015). Moreover, inference for the HDDP mixtures is performed on the space of the latent arrival times and lengths of stay, which further slows down the mixing, making necessary to run the MCMC for more iterations. An alternative algorithm to speed up the mixing has been proposed by Jain and Neal (2004).

We note that we have not used any spatial information on the sites and, as a result, sites are assumed to be exchangeable, in the sense that permuting the site labels has no effect on our inference. This is generally the case when data are collected at a number of sites but the models employed are not spatially-explicit. See for example the occupancy model by MacKenzie et al. (2002) and extensions as well as the N-mixture model by Royle (2004b) and extensions. Since there is only a small number of sites, which are not in close proximity to one another, any spatial autocorrelation in our application is expected to be low.

The code used to generate results has been written in R (R Core Team, 2014), while the most computationally expensive part of the algorithm, such as the Gibbs sampler for the clusters allocation, has been written integrating C++ and R using the Rcpp package of Eddelbuettel et al. (2011). The code is available upon request.

SUPPLEMENTARY MATERIAL

Supplementary material

(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide the complete expressions of the posterior distributions and additional plots of the prior and posterior distributions.

References.

Álvarez, M. A. and Lawrence, N. D. 2011. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12(May):1459–1500.

Both, C., Bouwhuis, S., Lessells, C., and Visser, M. E. 2006. Climate change and population declines in a long-distance migratory bird. *Nature*, 441(7089):81.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. 2018. Distribution theory for hierarchical processes. *Annals of Statistics*.

Chen, Z., Wang, B., and Gorban, A. N. 2017. Multivariate Gaussian and Student-t process regression for multi-output prediction. *arXiv preprint arXiv:1703.04455*.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. 2004. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215.

Diana, A., Griffin, J., and Matechou, E. 2018. Supplementary material: A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., and Jordan, F. 2008. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64(2):635–644.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. 2011. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eglington, S. M., Julliard, R., Gargallo, G., van der Jeugd, H. P., Pearce-Higgins, J. W., Baillie, S. R., and Robinson, R. A. 2015. Latitudinal gradients in the productivity of e uropean migrant warblers have not shifted northwards during a period of climate change. *Global Ecology and Biogeography*, 24(4):427–436.

Escobar, M. D. and West, M. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, pages 209–230.

Ford, J. H., Patterson, T. A., and Bravington, M. V. 2015. Modelling latent individual heterogeneity in mark-recapture data with Dirichlet process priors. *arXiv preprint arXiv:1511.07103*.

Ge, H., Chen, Y., Wan, M., and Ghahramani, Z. 2015. Distributed inference for dirichlet process mixture models. In *International Conference on Machine Learning*, pages 2276–2284.

Gienapp, P., Leimu, R., and Merilä, J. 2007. Responses to climate change in avian migration timemicroevolution versus phenotypic plasticity. *Climate Research*, 35(1/2):25–35.

Griffin, J. E. and Leisen, F. 2017. Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):525–545.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. 2010. *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Jain, S. and Neal, R. M. 2004. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of computational and Graphical Statistics*, 13(1):158–182.

Johnston, A., Robinson, R. A., Gargallo, G., Julliard, R., Jeugd, H., and Baillie, S. R. 2016. Survival of afro-palaearctic passerine migrants in western Europe and the impacts of seasonal weather variables. *Ibis*, 158(3):465–480.

Kingman, J. 1967. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kingman, J. F. C. 1993. *Poisson processes*, volume 3. Clarendon Press.

MacEachern, S. N. 1999. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, pages 50–5. Alexandria, Va: American Statistical Association.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Manrique-Vallier, D. 2016. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254.

Matechou, E., Caron, F., et al. 2017. Modelling individual migration patterns using a Bayesian nonparametric approach for capture–recapture data. *The Annals of Applied Statistics*, 11(1):21–40.

Møller, A. P., Rubolini, D., and Lehikoinen, E. 2008. Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proceedings of the National Academy of Sciences*.

Peach, W., Buckland, S., and Baillie, S. 1996. The use of constant effort mist-netting to measure between-year changes in the abundance and productivity of common passerines. *Bird Study*, 43(2):142–156.

R Core Team 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. 2006. Gaussian processes in machine learning. MIT Press.

Robson, D. and Barriocanal, C. 2011. Ecological conditions in wintering and passage areas as determinants of timing of spring migration in trans-saharan migratory birds. *Journal of Animal Ecology*, 80(2):320–331.

Royle, J. A. 2004a. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A. 2004b. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–81.

Wolpert, R. L. and Ickstadt, K. 1998. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267.

APPENDIX

MCMC Algorithm. The vector of unknown parameters is:

$$(\{\zeta_{ijy}\}, \{\delta_{ijy}\}, \{c_{ijy}\}, \mu_k, \Sigma_k, b, \{N_{jy}\}, \{p_{jy}\}, \alpha, \gamma, \{\tau_j\}, \{\omega_j\})\,.$$

Clearly the posterior distribution is intractable and we obtain samples from it using the following steps in a Gibbs sampler: cluster allocations $\{c_{ijy}\}$ are sampled using the update in Teh et al. (2006) that makes use of the CRF representation, while cluster locations $\{\mu_k, \Sigma_k\}$ are updated conditional on the allocations. The population sizes N_{jy} are updated using the rejection algorithm employed in MC17. The arrival times and lengths of stay $(\zeta_{ijy}, \delta_{ijy})$ are sampled jointly using a simple MH update. To update b, first we sample the β_k and then we sample b from its full conditional. Finally, capture probabilities are updated using a MH step. For the remaining hyperparameters α, γ, τ_j and ω_j, we can sample directly from the full conditional.

A detailed description of each Gibbs sampler can be found below.

1. **Sample the cluster means and covariance matrices (μ_j, Σ_j):**
   For each cluster $k = 1, \ldots, K$, we sample (μ_k, Σ_k) from the posterior distribution:

$$p\left(\mu_k, \Sigma_k | \{\zeta_{ijy}\}, \{\delta_{ijy}\}, \{c_{ijy}\}, b, B, \nu_0, \Sigma_0\right) \propto$$
$$p\left(\mu_k | \{\zeta_{ijy}\}, \{\delta_{ijy}\}, \{c_{ijy}\}, \Sigma_k, b, B\right) p(\Sigma_j | \{\zeta_{ijy}\}, \{\delta_{ijy}\}, \{c_{ijy}\}, B, \nu_0, \Sigma_0)$$

   As shown in the supplementary material (Diana et al., 2018), the posterior distribution for μ_k is still a MN distribution, while the posterior distribution for Σ_k is still an inverse-Wishart. In our application, to efficiently compute the posterior distributions, we rely on the fact that the covariate, being year-specific, takes only as many values as the number of years, Y. Thus, instead of building the covariance matrix of the MGP using all the individuals in the cluster, we calculate the covariance computed using only the value of the covariates at the observed points. Moreover, as these points are fixed in the model, the inverse of the covariance matrix of the GP can be precomputed.
2. **Sample the allocation $\{c_{ijy}\}$ of individuals to the different clusters:**
   Following Teh et al. (2006) and the notation of section 3.1, the variables c_{ijy} are updated using the CRF representation defined in Section 3.1, by first sampling the allocations t_{ij} of the customers to the tables and then the allocations k_{jt} of the tables to the dishes.
   We use the superscript $^{-ij}$ to indicate that the quantities are computed removing customer i from restaurant j and $^{-jt}$ when removing table t from restaurant j.
   At each step of the Gibbs sampler, individual i in group j having covariate y is assigned to either an existing table in the current restaurant, a new table serving an existing dish or a new table serving a new dish

$$\begin{cases} \text{existing table } t & \text{with probability} \quad \frac{n_{jt}^{-ij}}{n_{jt}^{-ij}+\alpha}\ p((\zeta_{ijy},\delta_{ijy})|x_y,\mu_{k_{jt}},\Sigma_{k_{jt}}) \\ \text{new table } t^\star \text{ with existing dish } k & \text{with probability} \quad \frac{\alpha}{n_{jt}^{-ij}+\alpha}\frac{m_k^{-ij}}{M^{-ij}+\gamma}\ p((\zeta_{ijy},\delta_{ijy})|x_y,\mu_k,\Sigma_k) \\ \text{new table } t^\star \text{ with new dish } k^\star & \text{with probability} \quad \frac{\alpha}{n_{jt}^{-ij}+\alpha}\frac{\gamma}{M^{-ij}+\gamma}\ p((\zeta_{ijy},\delta_{ijy})|x_y,b,\Sigma_0,\nu_0) \end{cases}$$

Similarly, tables are allocated to

$$\begin{cases} \text{existing dish } k & \text{with probability} \quad \frac{m_k^{-jt}}{M^{-jt}+\gamma}\ p(\{(\zeta_{ijy},\delta_{ijy})\}_{t_{ij}=t}|\{x_y\},\mu_{k_{jt}},\Sigma_{k_{jt}}) \\ \text{new dish } k^\star & \text{with probability} \quad \frac{\gamma}{M^{-jt}+\gamma}\ p(\{(\zeta_{ijy},\delta_{ijy})\}_{t_{ij}=t}|x_y,b,\Sigma_0,\nu_0) \end{cases}$$

As opposed to the original algorithm of Teh et al. (2006), instead of computing the posterior distribution of $(\zeta_{ijy},\delta_{ijy})$ conditional on the current elements in the cluster, which would excessively slow down the algorithm if repeated for each point, we compute the update conditional on the cluster locations (μ_k,Σ_k) computed in the previous step.

3. **Sample the population sizes $\{N_{jy}\}$:**
Following MC17, for each site j and year y, conditional on the measure P_{jy}, the arrival times and length of stay
$\zeta_{(D_{jy}+1):N_{jy},j,y},\delta_{(D_{jy}+1):N_{jy},j,y}$ of the uncaptured birds are distributed from a non-homogeneous Poisson process with intensity

$$\nu_0(\zeta,\delta)=\nu(\zeta,\delta)\ \mathbb{P}(H=(0,\dots,0)|\zeta,\delta,p)$$

It follows that samples from the posterior distribution of

$$\left(N_{jy},c_{(D_{jy}+1):N_{jy},j,y},\zeta_{(D_{jy}+1):N_{jy},j,y},\delta_{(D_{jy}+1):N_{jy},j,y}\right)$$

can be obtained with a rejection algorithm in the following way. First, sample $N_0\sim\text{Poisson}(\omega_j)$, then, for $i=1,\dots,N_0$ sample:

$$\{c_{ijy}\}_{i=1,\dots,N_0}\mid \text{CRF}(\alpha,\gamma)$$
$$(\zeta_{ijy},\eta_{ijy})\sim \text{N}(\mu_{c_{ijy}},\Sigma_{c_{ijy}})$$
$$H_{ijy}\sim Pr(\zeta_{ijy},\eta_{ijy},p_{jy})$$

and accept the individual if capture history H_{ijy} has no non-zero entries. The new population size is given by $D_{jy}+\tilde{N}_0$ where D_{jy} is the number of captured individuals at site j in year y and $\tilde{N}_0$ is the number of accepted individuals.

4. **Sample the hyperparameters** $(\alpha, \gamma, \{\tau_j\}, \{\omega_j\})$**:**
   τ_j and ω_j are updated as

$$\omega_j \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{Y} N_{ji}, \frac{\alpha}{\tau_j} + Y\right)$$

$$p(\tau_j|\alpha, \omega_j) \propto p(\tau_j)p(\omega_j|\alpha, \tau_j) \propto \text{Gamma}(\alpha_\tau, \beta_\tau)\tau_j^{-\alpha} e^{-\alpha\frac{\omega_j}{\tau_j}}$$

   The posterior distributions for α and γ are found by adapting the update for the concentration parameter of the DP presented in Escobar and West (1995). Details are presented in the supplementary material (Diana et al., 2018). An exact expression of the likelihood of α and γ given an allocation of individuals to the cluster can be found in Camerlenghi et al. (2018).
5. **Sample the mean** b **of the prior distribution of the cluster-specific regression coefficients:**
   In order to improve the mixing for the posterior distribution of b, we introduce the variables $\delta_k := \beta_k - b$. After sampling the δ_k from their posterior distribution (which can be found in the supplementary material (Diana et al., 2018)), the posterior distribution of b given δ_k can be computed as

$$p(b|\{\beta_k\}_{k=1,\dots,K}, B, b_0, B_0, \{\Sigma_k\}_{k=1,\dots,J}) \propto$$

$$p(\{\beta_k\}_{k=1,\dots,K}|b, B)p(b|b_0, B_0) \propto \prod_{k=1}^{K} \text{MN}(\beta_k|b, B, \Sigma_k)\text{MN}(b|b_0, B_0, \Sigma_0) \propto$$

   The complete formulas can be found in the supplementary material.
6. **Sample the latent arrival times and length of stay** $\{\zeta_{ijy}\}, \{\delta_{ijy}\}$**:**
   Given the continuous arrival time and length of stay of each individual, if we consider the partition defined by the sampling occasions $t_1^{jy}, \dots, t_{C_{jy}}^{jy}$, we can define as b_{ijy} and d_{ijy} the intervals where individual ijy respectively arrives and departs. Given these intervals, the posterior distribution for $(\zeta_{ijy}, \delta_{ijy})$ is:

$$p(\zeta_{ijy}, \delta_{ijy}|\mu_{c_{ijy}}, \Sigma_{c_{ijy}}, H_{jy}, p_{jy}) \propto p(\zeta_{ijy}, \delta_{ijy}|\mu_{c_{ijy}}, \Sigma_{c_{ijy}})p(H_{jy}|p_{jy}, \zeta_{ijy}, \delta_{ijy}) =$$

$$\text{N}(\zeta_{ijy}, \eta_{ijy}|\mu_{c_{ijy}}, \Sigma_{c_{ijy}})\ p_{jy}^{\sum_{k=1}^{C_{jy}} H_{ijyk}}(1 - p_{jy})^{d_{ijy} - b_{ijy} - \sum_{k=1}^{C_{jy}} H_{ijyk}}$$

7. **Sample the coefficient** $\{\alpha_{jy}^p\}$ **and** β^p **of the capture probabilities** $\{p_{jy}\}$**:**
   Although not available in analytic form, the posterior distribution can be computed as

$$p(\{\alpha_j^p\}, \beta^p|\{\zeta_{ijy}\}, \{\delta_{ijy}\}, \{x_{jy}\}) \propto p(\beta^p|b_0^p, B)p(\{\alpha_j^p\}|a_0^p, A_0^p)p(\{H_{jy}\}|\{\zeta_{ijy}\}, \{\delta_{ijy}\}, p_{jy}) =$$

$$\text{N}(\beta^p|b_0^p, B)\prod_j \text{N}(\alpha_j^p|a_0^p, A_0^p)\ \prod_{j,y}\prod_{i=1}^{N_{jy}} p_{jy}^{\sum_{k=1}^{C_{jy}} H_{ijyk}}(1 - p_{jy})^{d_{ijy} - b_{ijy} - \sum_{k=1}^{C_{jy}} H_{ijyk}}$$

A. DIANA
E. MATECHOU
SCHOOL OF MATHEMATICS,
STATISTICS AND ACTUARIAL SCIENCE
UNIVERSITY OF KENT, CANTERBURY, UK
E-MAIL: ad603@kent.ac.uk
e.matechou@kent.ac.uk

J. GRIFFIN
DEPARTMENT OF STATISTICAL SCIENCE,
UNIVERSITY COLLEGE LONDON, LONDON, UK
E-MAIL: j.griffin@ucl.ac.uk

A. JOHNSTON
BRITISH TRUST FOR ORNTIHOLOGY, THE NUNNERY,
THETFORD, NORFOLK, IP24 2PU, UK
CORNELL LAB OF ORNITHOLOGY, 159 SAPSUCKER WOODS ROAD,
ITHACA, NY 14850, USA
CONSERVATION SCIENCE GROUP, DEPT OF ZOOLOGY,
UNIVERSITY OF CAMBRIDGE, DOWNING STREET, CAMBRIDGE, CB2 3EJ, UK
E-MAIL: aj327@cornell.edu