Submitted to the Annals of Applied Statistics
2005 Vol. 0, No. 0, 1–8
arXiv: arXiv:0000.0000

SUPPLEMENTARY MATERIAL FOR: MODELLING INDIVIDUAL MIGRATION PATTERNS USING A BAYESIAN NONPARAMETRIC APPROACH FOR CAPTURE-RECAPTURE DATA

By Eleni Matechou and François Caron

University of Kent and University of Oxford

CONTENTS

1. Convergence diagnostics for the analysis of the great crested newt data set. We run three chains, (discarded 50000 iterations, thining = 300) with starting values for the parameters randomly generated from their parameter space.

For each parameter (N, ω, α, τ, β_1 and β_2 which are, respectively, the intercept and slope of the logistic regression model for p with the number of traps in the pond as the covariate, γ_1 and γ_2 which are, respectively, the intercept and slope of the logistic regression model for ϕ with standardised calendar time as the covariate) we show the three trace plots the resulting posterior densities and finally the Gelman-Rubin diagnostic plot produced using the R-package coda.

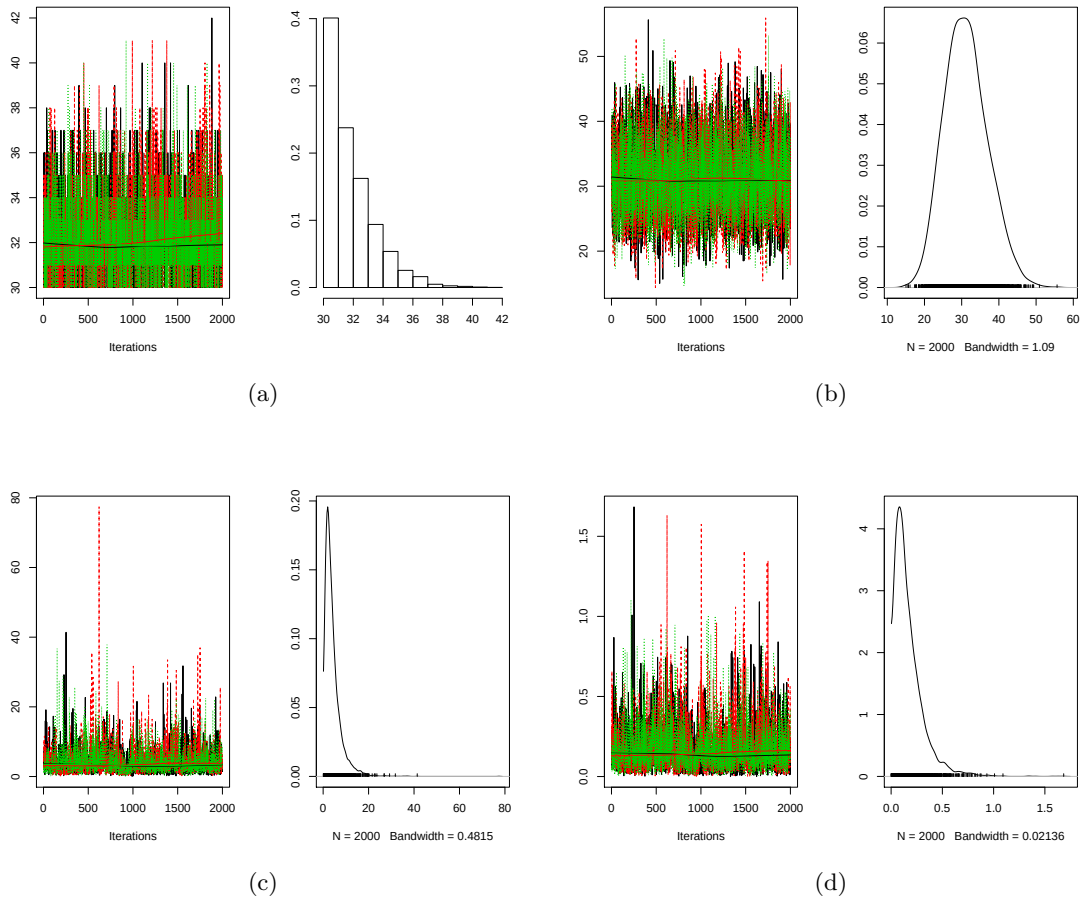
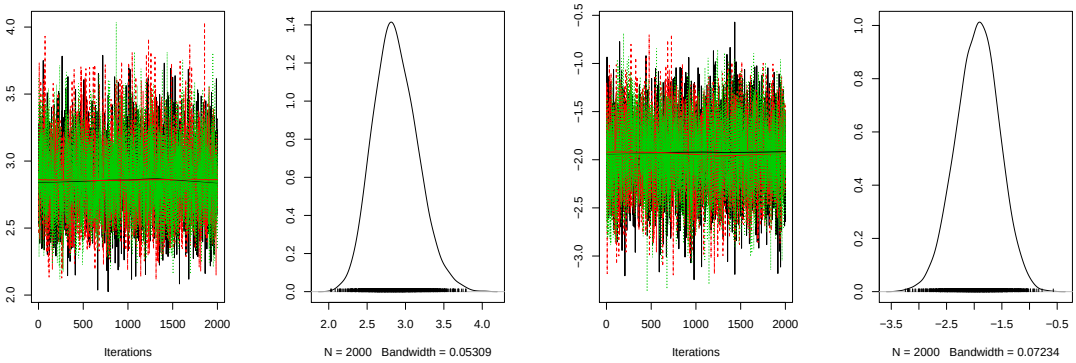
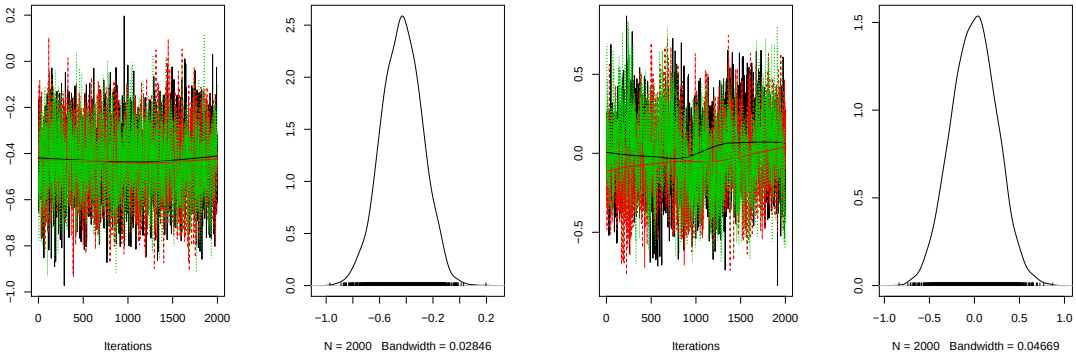


FIG. 1: Trace and posterior density plots for (a) N, (b) ω, (c) α, (d) τ.

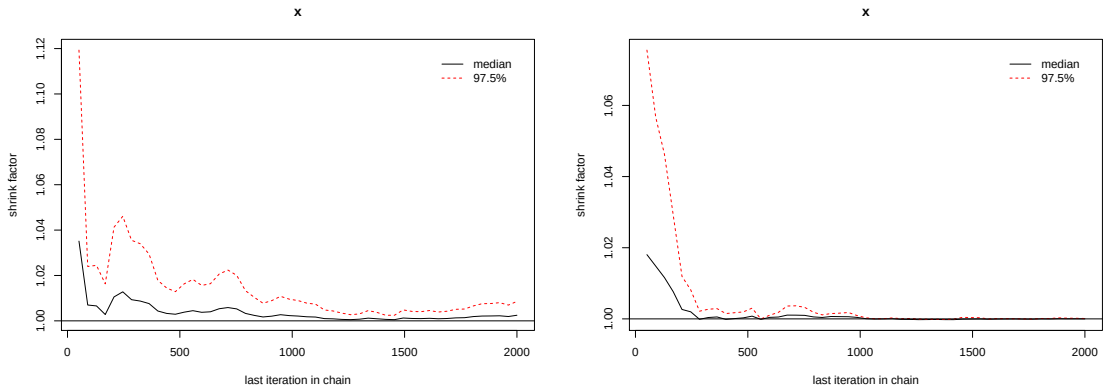


(a) (b)



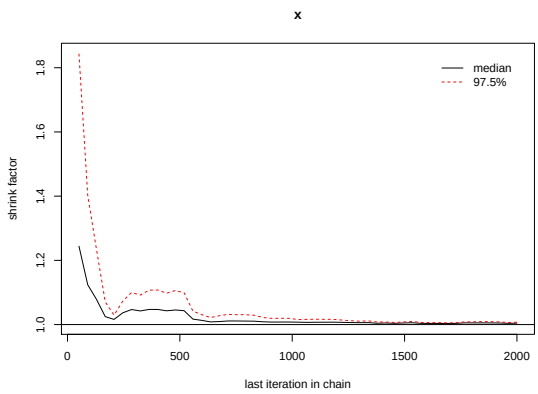
(c) (d)

FIG. 2: Trace and posterior density plots for (a) γ_1, (b) γ_2, (c) β_1 and (d) β_2.

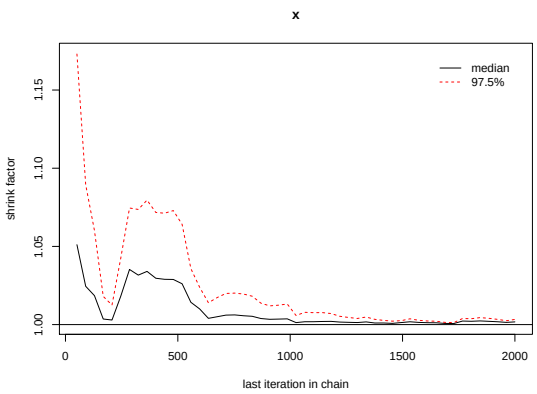


(a)

(b)

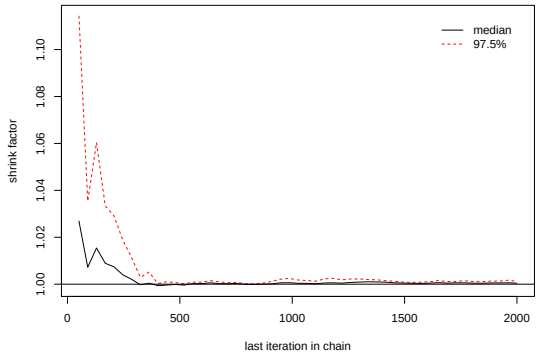


(c)

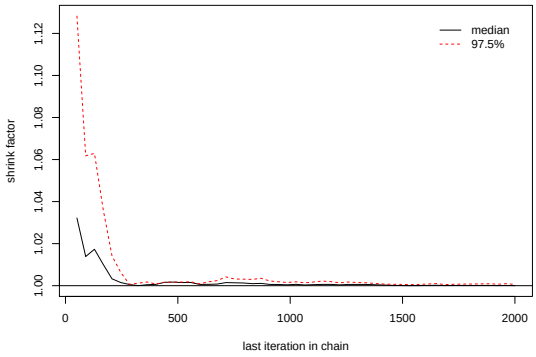


(d)

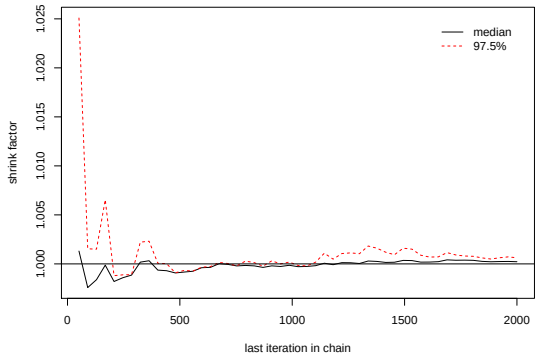
FIG. 3: Gelman-Rubin diagnostics plots for (a) N, (b) ω, (c) α, (d) τ.



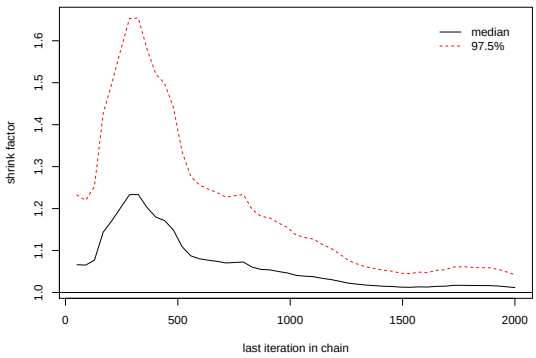
(e)



(f)



(g)



(h)

FIG. 3: Gelman-Rubin diagnostics plots for (a) γ_1, (b) γ_2, (c) β_1 and (d) β_2.

2. Convergence diagnostics for the analysis of the reed warbler data set. We run three chains, (discarded 50000 iterations, thining = 300) with starting values for the parameters randomly generated from their parameter space.

For each parameter (N, ω, α, τ, the logistic transformation of capture probability p, logit(p), the logistic transformation of survival probability, ϕ, logit(ϕ)) we show the three trace plots the resulting posterior densities and the Gelman-Rubin diagnostic plot produced using the R-package coda.

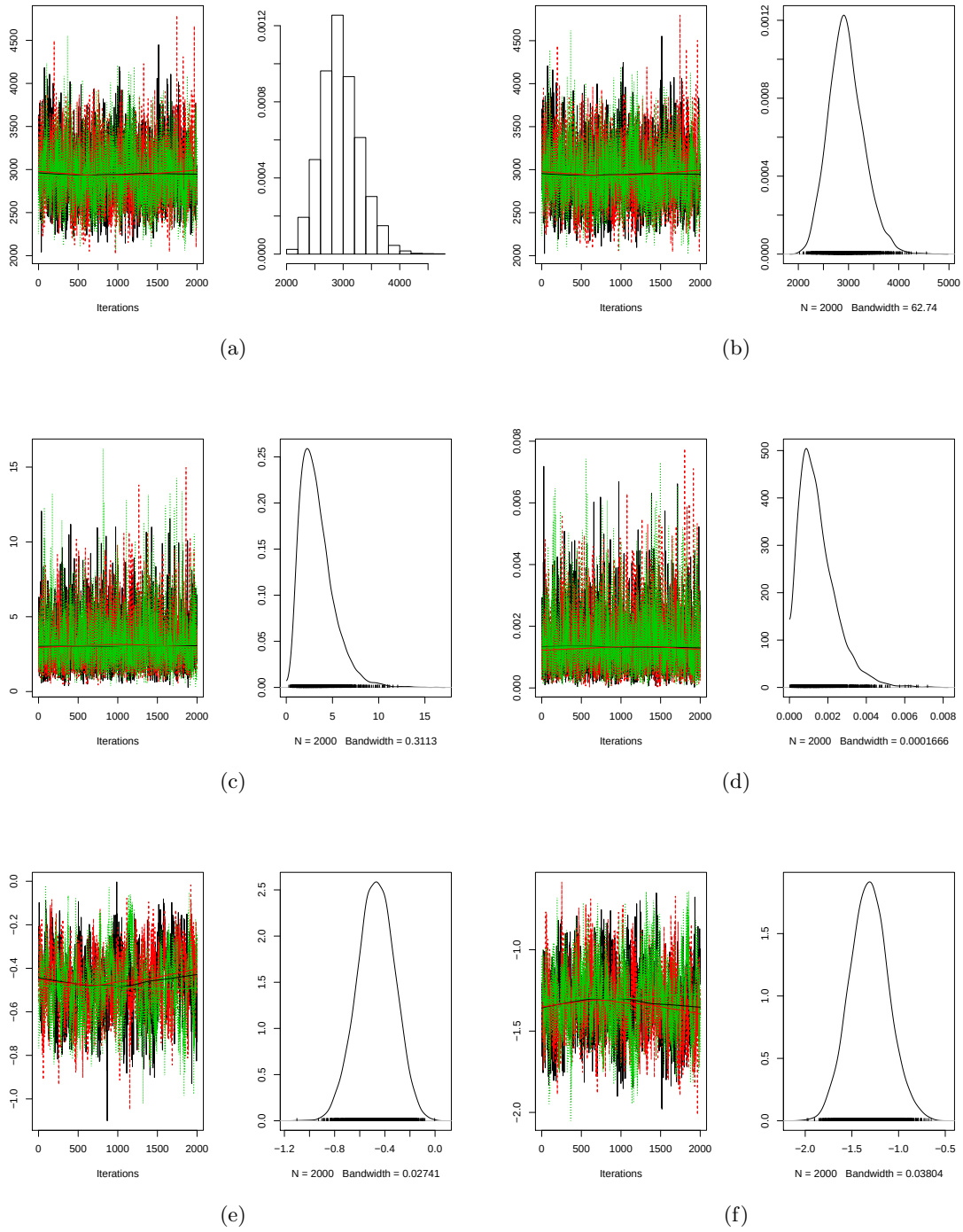


FIG. 4: Trace and posterior density plots for (a) N, (b) ω, (c) α, (d) τ, (e) logit(ϕ) and (f) logit(p).

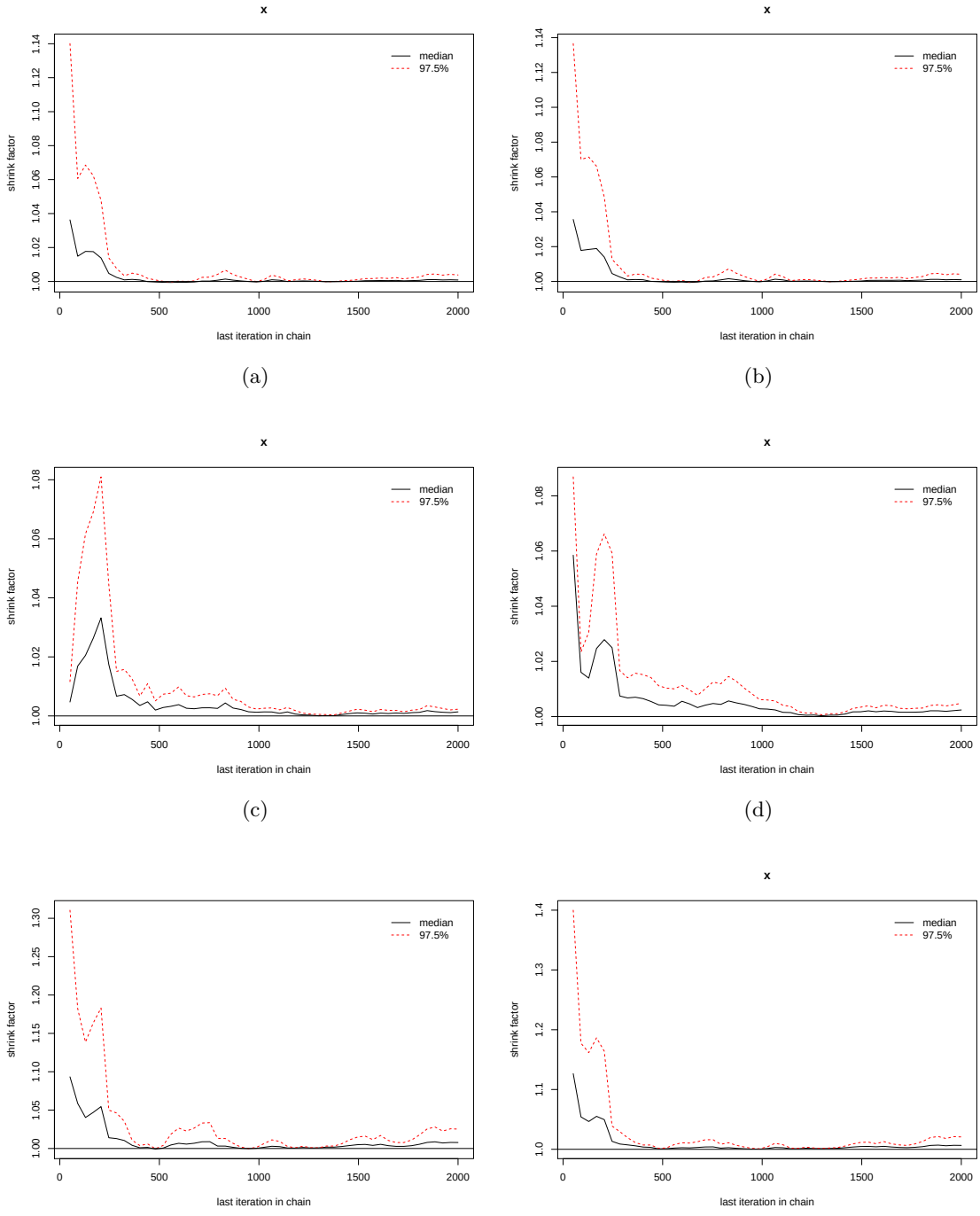


FIG. 5: Gelman-Rubin diagnostics plots for (a) N, (b) ω, (c) α, (d) τ, (e) logit(ϕ) and (f) logit(p).

3. Comparison with an existing model. We compare the results for both case studies presented in section 3 to those obtained by an existing JS-type model, and specifically to the Pledger et al. (2009) model.

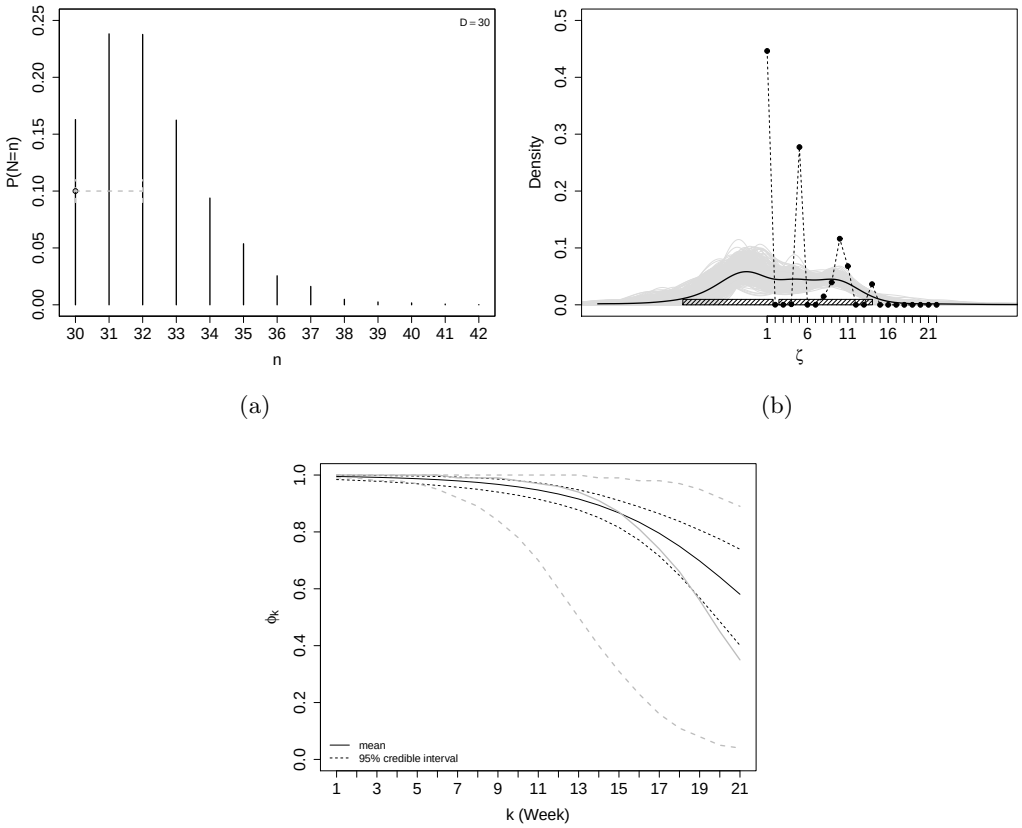
3.1. *Great crested newts.* In Fig. 6 (a), we plot the posterior distribution of N obtained by our proposed model while the horizontal gray dashed bar and the empty circle represent the 95% asymptotic confidence interval and the point estimate, respectively, obtained for N by the Pledger et al. (2009) model. This is [30, 32] individuals and it encompasses roughly 64% of the posterior mass for N obtained by our algorithm. In this case, even though the time between samples (one week) was longer compared to the study of reed warblers, the number of individuals that could have departed without ever becoming available for detection is negligible since great crested newts tend to have long stopover at breeding sites that span a number of weeks.

In Fig. 6 (b), the density estimates for ζ obtained at 500 randomly selected iterations of our proposed algorithm are shown by the gray lines, with the black line showing the mean density and the tick marks on the x-axis indicating sampling occasions. The position of the boxes on the x-axis indicates the values of ζ that fall in the 95% HPD interval while their height is equal to the lowest density value in the interval. The points represent the point estimates obtained by the Pledger et al. (2009) model of the proportion of the "super-population" size that were new arrivals at each sampling occasions. These are connected by the dashed lines for ease of comparison. The arrival pattern suggested by the Pledger et al. (2009) suggests more abrupt changes to the population compared to our results, which, due to the use of the normal mixture model, suggest a much smoother migration to the ponds. However, both approaches agree that roughly 40% of the individuals were already present at the start of the study and they also agree on the position of the modes of arrival times, i.e. week 5 and weeks 9-10.

In Fig. 6 (c), the point estimates obtained by the Pledger et al. (2009) model for ϕ as a function of week number, shown by the gray line, agree with our posterior means, but the 95% asymptotic confidence intervals, shown by the gray dashed lines, are considerably wider, especially towards the end of the study period when fewer individuals are present.

Finally, the Pledger et al. (2009) model estimates p equal to 0.42 with a 95% asymptotic confidence interval = (0.33, 0.52) when the number of traps is 8. The corresponding values for p when the number of traps is 6 are 0.39, (0.30, 0.49). Both sets of values agree with our results as presented in section 3.1.

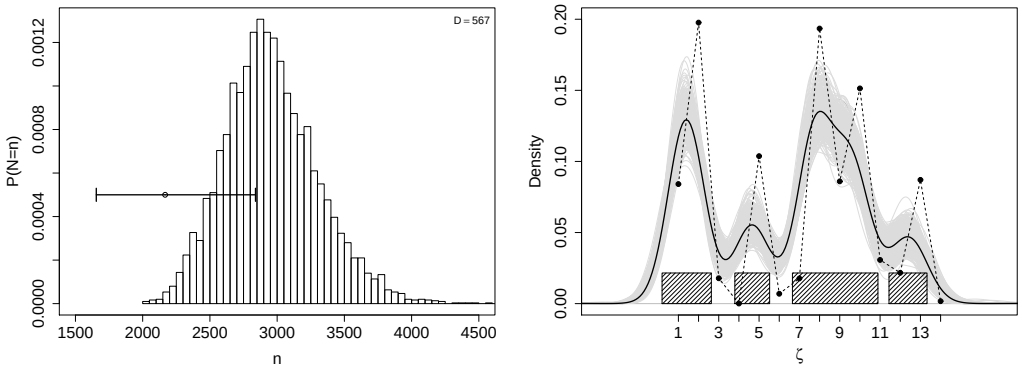
3.2. *Reed Warblers.* In Fig. 7 (a), the histogram represents the posterior distribution of N obtained by our proposed model while the vertical bar and the empty circle represent the 95% asymptotic confidence interval and the point estimate, respectively, obtained for N by the Pledger et al. (2009) model. The Pledger et al. (2009) model gives a lower estimate for N because it only accounts for individuals that became available for detection at least once while our approach also accounts for individuals that might have departed



(a) (b)

(c)

FIG. 6: Great crested newt data. Comparison of the results obtained by our proposed model and the Pledger et al. (2009) model.



(a) (b)

FIG. 7: Reed warbler data. Comparison of the results obtained by our proposed model and the Pledger et al. (2009) model.

before any sampling occasion took place. The difference is considerable, although the confidence/credible bands are overlapping, because the time between samples (5 days) was long relative to the average stopover duration of individuals in the population (roughly 8 days).

In Fig. 7 (b), the density estimates for ζ obtained at 500 randomly selected iterations of our proposed algorithm are shown by the gray lines, with the black line showing the mean density and the tick marks on the x-axis indicating sampling occasions. The position of the boxes on the x-axis indicates the values of ζ that fall in the 95% HPD interval while their height is equal to the lowest density value in the interval. The points represent the point estimates obtained by the Pledger et al. (2009) model of the proportion of the "super-population" size that were new arrivals at each sampling occasions. These are connected by the dashed lines for ease of comparison. The arrival pattern suggested by the two approaches is similar, with our approach providing a smoother representation of the spikier arrival pattern obtained by the Pledger et al. (2009) model. We note here that with our approach we estimate the arrival time of each individual, as opposed to the proportion of individuals that were new arrivals at each occasion.

Finally, the Pledger et al. (2009) model estimates ϕ equal to 0.43 with a 95% asymptotic confidence interval = (0.35, 0.52). The corresponding values for p are 0.17, (0.12, 0.24). Both sets of values agree with our results as presented in section 3.2.

4. Sensitivity analysis.

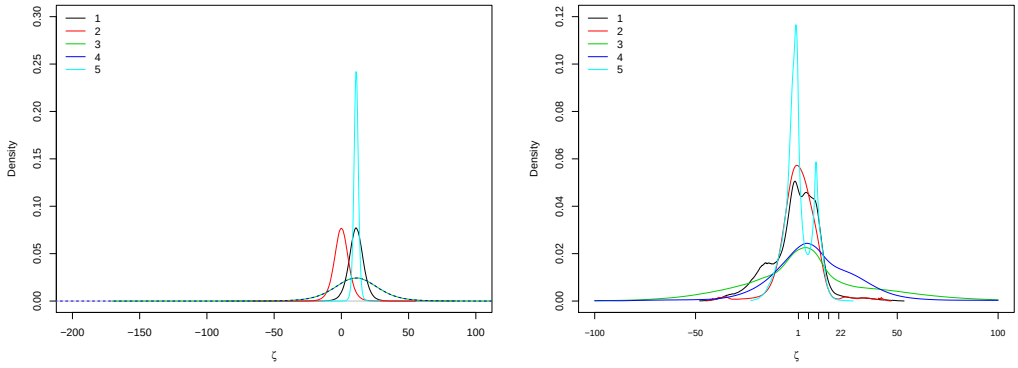
4.1. *Specifying different hyperparameters for G_0.* For both data sets considered in the paper, we specified G_0 to reflect our prior expectation regarding the arrival pattern at the site: most individuals will arrive during the study period and arrivals are centred around the middle of the study period. For example, for the data set of great crested newts we set $\mu_0 = K/2 = 11$, $\nu_0 = 4$, $\lambda_0 = 1$ and $\kappa_0 = 0.01$. In this section we explore the effect on the posterior density for N and on the mean normalised intensity of arrivals for the data set of great crested newts for different specifications for G_0. We list the five scenarios below, in each case noting the way in which they differ to the set up used for the analysis presented in the paper and commenting on the results.

The densities of arrival times sampled from the corresponding priors are plotted in Fig. 8 (a) while the mean normalised intensity of arrivals obtained in each case using our model is plotted in Fig. 8 (b). The posterior distribution and cumulative posterior distribution obtained for N in each case are plotted in Fig. 9 while the posterior distribution and cumulative posterior distribution of the number of individuals with arrival times *before* the end of the study period, which we denote by M, in Fig. 10.

1. As in the paper and described above. Even though the prior for arrival times suggests that most individuals arrived during the study period, the posterior suggests that about 40% of the individuals where already present at the start of the study.
2. Setting $\mu_0 = 0$ i.e. expecting a priori that about half of the individuals will be there at the start of the study. The prior of arrival times essentially agrees with the data and the posterior of arrival times matches the one obtained in scenario 1. The same applies to the posterior obtained for N and for M. Conclusion: the results are robust to specifying μ_0 considerably smaller than in the original analysis (but see comments about scenario 3 below)
3. Setting $\lambda_0 = 10$ i.e. expecting a priori that the arrival times of the different clusters are more variable than in scenario 1 with 95% of the mass of the prior distribution of σ (approximately) in (1, 3) instead of (0.3, 0.9) which was the case in scenario 1. In this case the prior of arrival times is much flatter and supports arrival times before the start (like in scenario 2) but more importantly after the end of the study. This is an important issue since there are no data available after the end of the study and hence the posterior will be dominated and completely determined by the prior. As a result, the mean normalised intensity does not drop to 0 for values of $\zeta > K$ as quickly and more individuals contribute to N than in scenario 1. These are individuals that arrived after the end of the study so in a sense almost fictual since they have completely been generated by the prior. Hence, unless there is in fact prior information that arrivals continue after the end of the study, *we advise against the use of such flat priors in this case.*
   We note here that in scenario 2 the prior supported the arrival of individuals before the start of the study but that did not have any considerable effect on our inference

compared to scenario 1, even though there are of course no data available for that period either. However, there is an important difference between the two periods, before the start and after the end of the study. For the former, individuals that arrived have to either depart before the study commences or they have to stay until the data collection process begins. Since for the study of great crested newts individuals are estimated to remain until at least the middle of the study, they would have to remain and be available for detection and hence contribute to the data. Hence, in this case, prior support for $\zeta < 1$ is in fact penalised by the data for the reason given above while the same does not apply for prior support for $\zeta > K$. The posterior obtained for M is not considerably different -although still shifted slightly to the right- to that obtained under scenarios 1 and 2, further supporting this point. We note that this is not generally the case and for example if apparent survival was low at the start of the study then the prior of scenario 2 would also lead to higher values of N and M.

4. Setting $\kappa_0 = 0.001$ i.e. expecting a priori that the standard deviation of the mean arrival times of the different clusters is greater than in scenario 1 with 95% of the prior distribution for μ (approximately) in (33, 95) as opposed to (0.3, 0.9), which was the case in scenario 1. The density of samples of arrival times from this prior is practically identical to that of scenario 3 but using a different set of hyperparameters. As a result, inference is practically identical to that of scenario 3.
5. Setting $\lambda_0 = 0.0.1$ i.e. expecting a priori that the arrival times of the different clusters are less variable than in scenario 1 with 95% of the mass of the prior distribution of σ (approximately) in (0.01, 0.03) instead of (0.3, 0.9) which was the case in scenario 1. In this case, the mean normalised intensity of arrival times obtained is less smooth than in scenario 1, which results from the fact that the left tail is much shorter and hence all arrivals that occured before the start of the study are estimated to have happened right before of the study. Again this is simply the effect of the prior which suggests that arrivals happen in short abrupt bursts rather than being more spread out over a number of weeks. Nevertheless, the arrival pattern obtained is similar to that of scenario 1 and this is also the case for the posterior distribution for N and M.



(a) (b)

FIG. 8: Sensitivity analysis for the great crested newt data. Density of arrival times sampled from the five different specifications of G_0, (a), and resulting mean normalised intensities, (b).

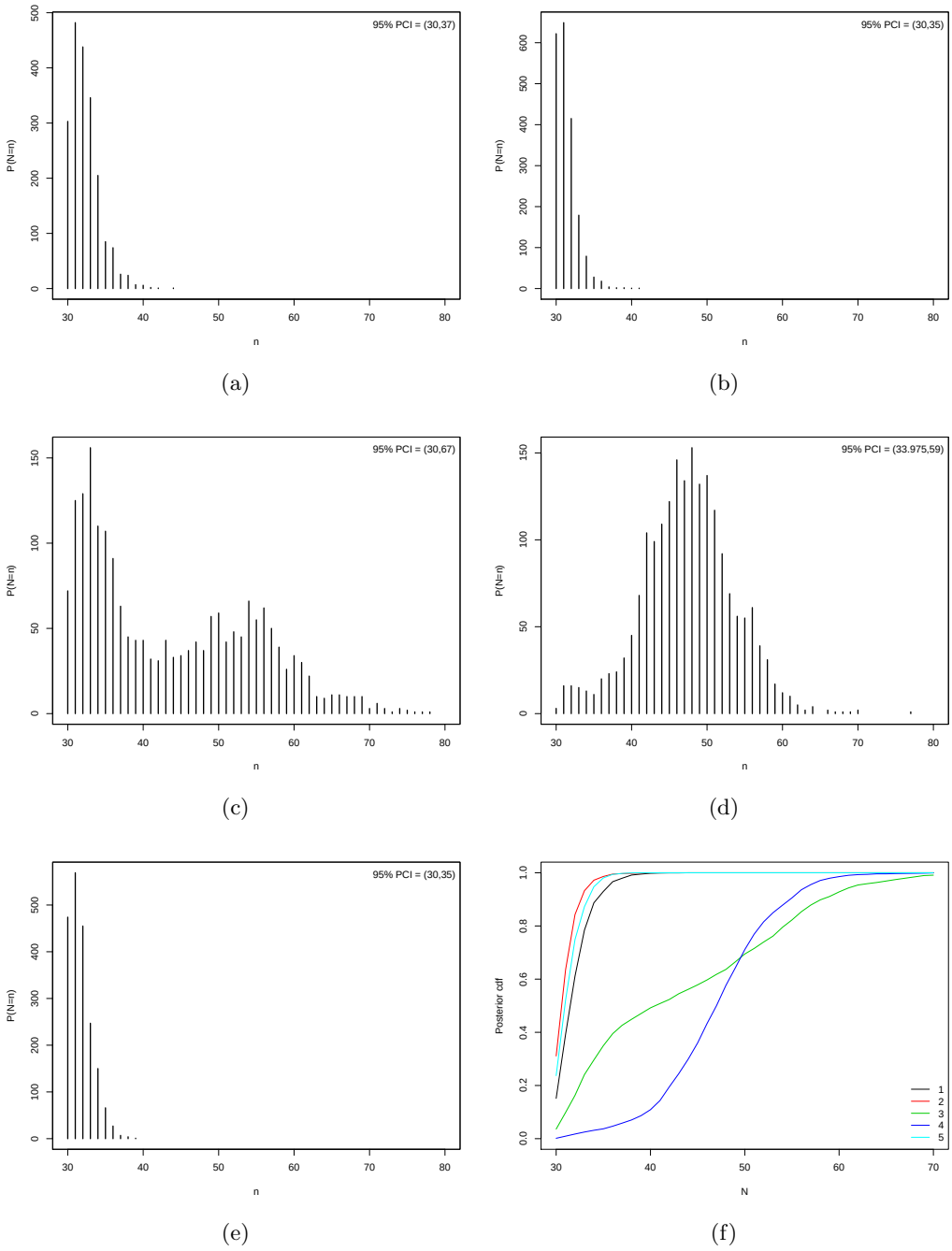
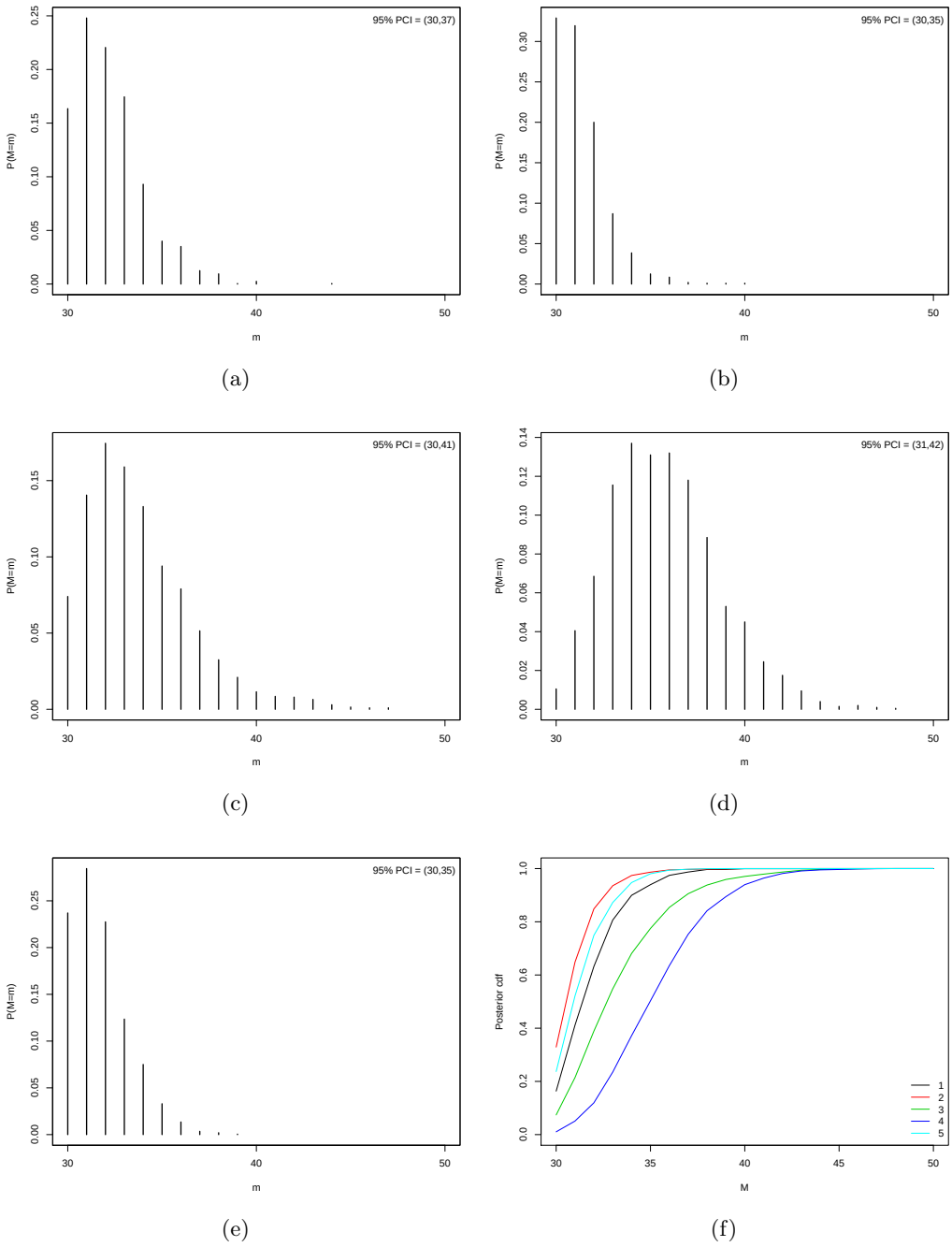


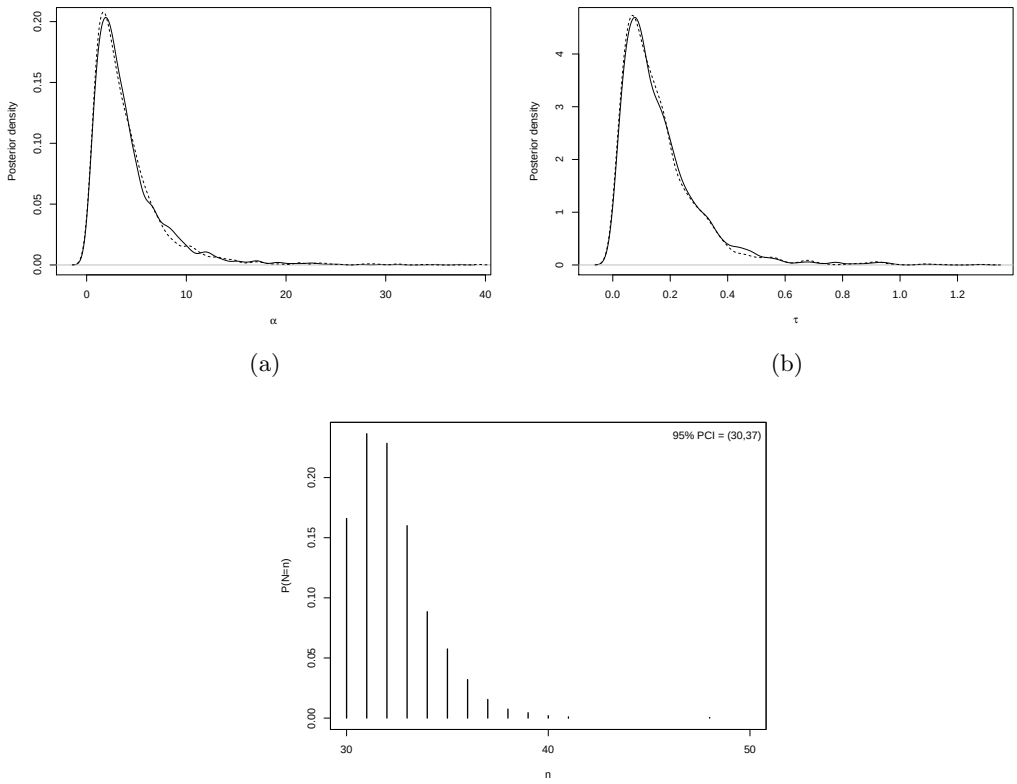
FIG. 9: Sensitivity analysis for the great crested newt data. Posterior distribution of N and 95% posterior credible intervals shown in the top right corner under the six different specifications of G_0, (a):1, (b):2, (c):3, (d):4, (e):5, (f): posterior cdf for all five scenarios.



(a) (b) (c) (d) (e) (f)

FIG. 10: Sensitivity analysis for the great crested newt data. Posterior distribution of M which we define here as the number of individuals with $\zeta \leq K$ and 95% posterior credible intervals shown in the top right corner under the six different specifications of G_0, (a):1, (b):2, (c):3, (d):4, (e):5, (f): posterior cdf for all scenarios.

4.2. *Specifying different priors for α and τ.* In this section we assess the effect of the prior distributions chosen for α and τ on their posterior distributions and also on the posterior distribution for N by changing the improper Gamma priors used for the analysis considered in the paper to proper Gamma(1,1) priors for both parameters.



(a) (b)

(c)

FIG. 11: Sensitivity analysis for the great crested newt data. Posterior distributions obtained for parameters α, (a) and τ, (b) when a Gamma(1,1) prior is chosen (posterior density shown by the dotted line) for both α and τ instead of the improper Gamma priors that were considered for the analysis presented in the paper (posterior density shown by the solid line). The posterior distribution obtained for N is shown in (c) and is practically indistinguishable to the one obtained in the analysis presented in the paper, which is shown in FIG. 9 (a).

4.3. *Specifying different prior variances for γ and β.* In this section we assess the effect of the prior variances chosen for the regression coefficients γ_1, γ_2, β_1 and β_2 when these are increased to 10^2 and 100^2 compared to the value of 1 which was chosen for the analysis presented in the paper.

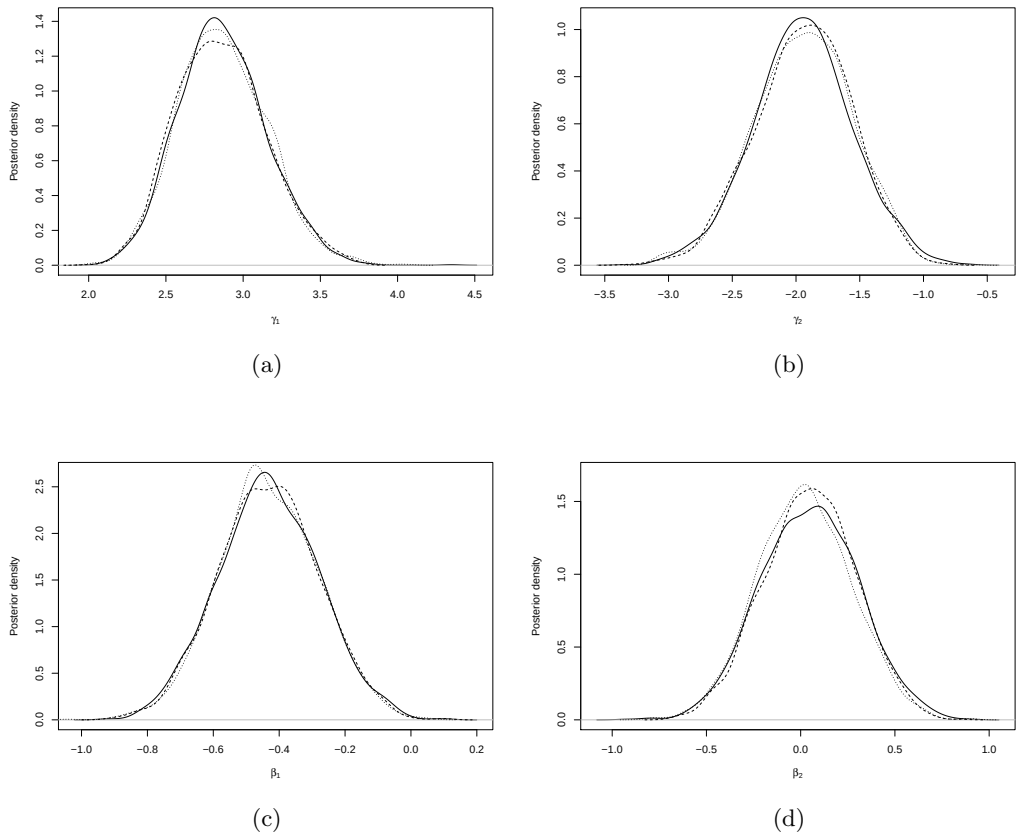


FIG. 12: Sensitivity analysis for the great crested newt data. Posterior distributions obtained for the regression coefficients for modelling ϕ and p when setting the prior variances to 1 (solid line), 10^2 (dashed line) and 100^2 (dotted line).

References.

Pledger, S., Efford, M., Pollock, K. H., Collazo, J. A., and Lyons, J. E. 2009. Stopover duration analysis with departure probability dependent on unknown time since arrival. *Environmental and Ecological Statistics (Edited by D.L.Thomson, E.G.Cooch and M.J. Conroy)*, 3:349–363.

E-mail: E.Matechou@kent.ac.uk

School of Mathematics, Statistics & Actuarial Science,
University of Kent, Canterbury, UK,
E-mail: E.Matechou@kent.ac.uk

E-mail: Caron@stats.ox.ac.uk

Department of Statistics,
University of Oxford, Oxford, UK
E-mail: Caron@stats.ox.ac.uk