Assignment 2: Bayesian State-Space Models – Serengeti Wildebeest

Team C

Introduction

The wildebeest (*Connochaetes taurinus*) of the Kenyan and Tanzanian Serengeti ecosystem represent 'one of the largest migratory ungulate populations in the world' (*Hilborn and Mangel, 1997).* The study and management of these populations is intrinsic to understanding and managing the wider ecosystem, which is of national and international importance. Population abundance data have been collected since the early 1960s ups to the late 1980s. Illegal harvesting, for which we have estimated rates, has affected the population and the rinderpest virus is sustained within it. Biologists and managers are particularly interested in understanding the relationship between population growth rate and rainfall, a proxy for water availability.

In this report, a discrete time model with density-independent exponential growth is developed which includes estimates of harvest. The model begins in 1960, even though there is no abundance data for this year, and it extends for five years beyond the end of the abundance data in order to make predictions for the years 1990-94.

Method

The wildebeest data contain 30 years' observations with 18 years missing data. As such, the sample size of this dataset was too small to produce valuable results using classical analysis. Therefore, Bayesian state-space model and Markov chain Monte Carlo (MCMC) simulation (*Plummer et al., 2006*) was used which was informed by prior knowledge of wildebeest ecology. The simulation allows a synthetic increase in the sample size allowing production of distributions for the parameters of interest.

The analysis was processed in R-software (*R Core Team, 2021*) with the Just Another Gibbs Sampler (JAGS) package (*Depaoli et al., 2016*) for Bayesian analysis Using Gibbs Sampling (BUGS) language (*Lunn et al., 2000*). Our model was built on three fundamental bases - priors and constraints, state likelihood process and observation likelihood process. Firstly, the priors and constraints section described the four parameters of initial population (N), prior for the intercept (β_0), prior for the slope (β_1) and standard deviation for the population (σ_N). Diffuse priors were used for all the parameters (Table 1). Initial population was defined by the population count from the data with unfilled values for five years' projection. The first year (1960) value was randomly picked from a Uniform distribution between zero and two (abundance is measured in millions) as the minimum and the maximum from our data were included in this range. The priors of intercept (β_0), and slope (β_1) were set as Normally distributed with a mean of zero and standard deviation of 100. The prior of standard deviation for the population was randomly picked from a Uniform distribution between zero and one to obtain all the possible values for the population size.

Prior
N1 ~ uniform(0, 2)
beta0 ~ normal(0, 100)
beta1 ~ normal(0, 100)
sig.n ~ uniform(0, 1)

Table 1. Prior parameterisations

Model equations:

1. Growth rate and rainfall relationship:

   $$\log(\mathrm{r_t}) = \beta_0 + \beta_1 \mathrm{Rain_t}$$

2. State process likelihood:

   $$N_t | N_{(t-1)} \sim \mathcal{N}(r_{(t-1)} N_{(t-1)} - C_{(t-1)},\ \sigma_N^2)$$

3. Observation process likelihood:

   $$y_t | N_t \sim \mathcal{N}(N_t, \sigma_y^2)$$

where r_t is the growth rate, N_t is the abundance at timestep t, y_t represents the observation model, C represents the harvest rate, σ represents the standard devations, and $\mathcal{N}$ represents a Normal distribution.

Our model divided the likelihood into two processes - state process (N) and observation process (y). In the state process, we used the available years in our data (from 1960 to 1989) plus five projected

years (1990 to 1994) in the loop's simulation. The years with missing data were backfilled using a "last observation carried forward" (LOCF) function which filled the missing data with previous the year's value. The relationship of the growth rate with rainfall was described under this process and log-link function was used to fit the equation in our model. The next year's population was expected to depend on a Normal distribution with the mean of current year growth rate, population size and illegal harvesting. The standard deviation for the population was generated from the prior and converted into precision. Moreover, the observation process only considered the years that had data recorded, so all the years with missing data were excluded. The observations simulated were dependent on a Normal distribution with mean of the initial population data and standard deviation calculated from the data's standard error.

In the MCMC simulation, three chains were used to run the analysis with one thinning sample. In order to achieve better convergence for our simulation, burn-in of some iterations was necessary. We used zero iterations for burn-in to initiate our model and determined that about ten thousand iterations for burn-in was enough. Gelman-Rubin statistic was used to compare the intra-chain and inter-chain variances in our model and resulting the value approached one which confirmed that our simulation was converged (Figure 1.). From the simulation, a hundred thousand iterations were generated from each MCMC chain, resulting in a total of three hundred thousand iterations.

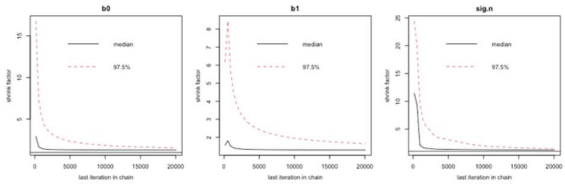

Figure 1. Graphical output of the Gelman-Rubin calculation after running the model with zero burn-in and twenty thousand iterations. The lines are shown approaching an asymptote to one before ten thousand iterations.

Results (and discussion)

MCMC simulation based on the abundance model specified above obtains robust estimates for the parameters in which we are most interested: the intercept (β_0) and slope coefficients (β_1) which inform our rainfall model of discrete annual growth rate (r_t) and the standard deviation of the abundance estimate (σ_N). In addition, we obtain standard deviations and 95% confidence bounds for these estimates from the MCMC simulation.

The parameter trace plots assess the chain mixing after the specified burn-in of five thousand iterations. The trace of each chain in all parameters (Figure 2.) show good mixing as the superimposed signals appeared as characteristic 'fat hairy caterpillars', with no individual chain's signal deviating noticeably from the others. The smooth, unimodal posterior density plots provide visual reassurance of the convergence of the MCMC chains (Figure 3). Smooth and Normally distributed density plots were obtained for both intercept (β_0) and slope (β_1) parameters, which indicated good mixing of MCMC chains in our model. The σ_N parameter's density plot also had a smooth curve but right-skewed distribution, which is not unexpected, since this parameter is constrained to positive values only. According to the results from trace plots and density plots, the MCMC explored the parameter space efficiently, leading to informative parameter samples. Furthermore, the potential scale reduction factor ($\hat{R}$) values of all three parameters from the posterior summary were 1.00 or 1.01, crucially less than 1.1, confirming the convergence of our model (Table 2). Consequently, from the $\hat{R}$ value, trace plots and density plots, we were confident in the reliability of our parameter results.

Parameter	Mean	Sd	Lower	Medium	Upper	Rhat	n.eff
b0	0.092	0.071	-0.070	0.099	0.214	1.00	1260
b1	-0.016	0.041	-0.091	-0.020	0.076	1.00	1351
sig.n	0.043	0.041	0.001	0.032	0.147	1.01	475

Table 2. Bayesian posterior summary of three parameters which are intercept (β_0), slope (β_1) and standard deviation for the population.

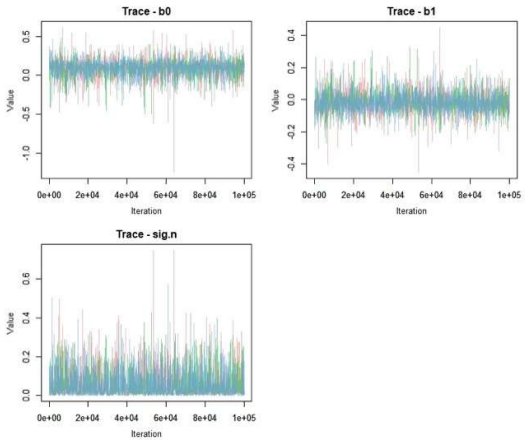

Figure 2. MCMC trace plots of three parameters which are intercept (β_0), slope (β_1) and standard deviation for the population (σ_N) with three chains and 100,000 iterations each.

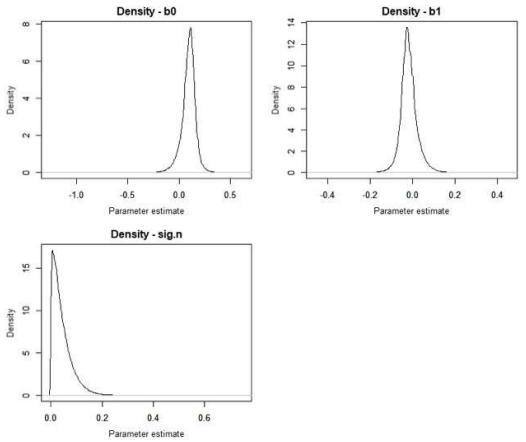

Figure 3. MCMC density plots of three parameters which are intercept (β_0), slope (β_1) and standard deviation for the population (σ_N).

The estimate for the intercept in the regression equation was given as 0.0924, implying a hypothetical yearly growth rate of $exp(0.0924) \approx 1$, regardless of rainfall (Table 2). However, caution should be exercised in interpreting the intercept, as taken on its own it interpolates the data to unreasonable scenarios (*i.e.* in a year with severe drought and zero rainfall, we would not expect a positive growth rate in a large mammal that requires a lot of water to survive, let alone to reproduce). The rainfall's covariate (β_1) is $exp(-0.016)$, such that, for example if there were 1 dm (decimetre) of rainfall, then $r_t = exp(0.0924 - 0.016) = 1.08$. *i.e.* the population would increases by 8% (ignoring the stochastic element of the model). The population standard deviation over the period 1960 to 1994 was estimated to be 0.0448. However, the accuracy of this estimate might be lower due to the much smaller sample size reported as compared with the other two parameters.

Once the model is satisfactorily parametrised, we can predict population projections for the five years following the period of data collection. This was incorporated into the MCMC evaluation to assess confidence intervals for the predictions, which have necessarily wider bounds than the preceding observed data. For the years 1990-1994, the harvesting rate is set as constant, and the rainfall was assumed to be the average over the years 1960-1989. For a more complex model, rainfall itself could also be modelled, since it features high variability and is rarely consistent form year to year. Nevertheless, an averaged rate is adequate for these purposes. The model predicts (Figure 4 and Table 3) a consistent increase in mean wildebeest abundance from 1.63 million in 1990 to 1.97 million in 1994, a growth rate of approximately 4% per year. The 95% confidence bounds in 1994 range from 1.3 to 2.5 million wildebeest, representing a change from the 1989 observations of -0.38 to + 0.83 million individuals, or -30% to +48% change from the 1989 observations. Although the mean abundance is expected to increase, the 95% bounds signify the expected ranges of abundance values, and we would caution future users of the model to use the ranges rather than just the mean figures in reporting results.

The standard deviations associated with the predictions also increased for successive years, however, as more uncertainty arises from extrapolation further beyond the observatory data. As such, the confidence intervals for the predictions grow wider for each successive year. It should be noted here that the model is not density-dependent (by design) and that previous work on the wildebeest data (*c.f.* Assignment 1) suggests that we are nowhere near the carrying capacity. However, the biologists/managers should be aware that there will be confounding factors beyond the scope of our model which will affect abundance.

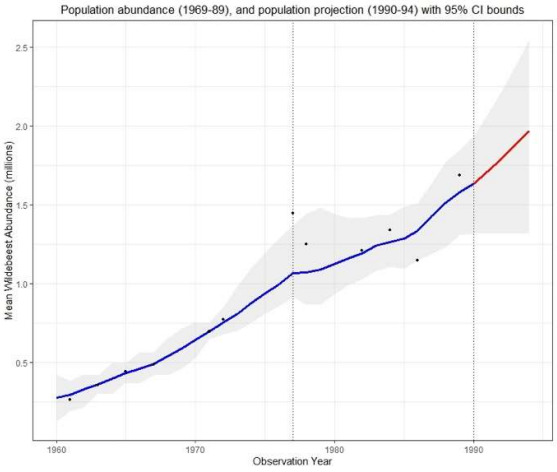

Figure 4. Modelled wildebeest abundance (millions) 1960-1989, with mean abundance in blue and 95% CI in grey, and projected abundance figures for 1990-1994, with mean abundance in red and 95% CI in grey. The dotted line at 1977 represents the introduction of the harvesting parameter into the model, and the dotted line at 1990 represents the beginning of the projected data.

Year	Mean	Sd	Lower	Medium	Upper
1990	1.635	0.158	1.316	1.638	1.934
1991	1.710	0.192	1.315	1.715	2.068
1992	1.791	0.229	1.315	1.798	2.214
1993	1.878	0.269	1.314	1.887	2.375
1994	1.971	0.312	1.316	1.983	2.549

Table 3. Predicted mean wildebeest abundance (millions) projected for the years 1990-1994, including standard deviation and confidence bounds.

There are certain aspects of the results described above which behove reflection. The estimated coefficient β_1 is negative when r_t is modelled as $log(r_t)$, although non-negative when exponentiated, with a value of nearly one, which is a strong positive slope. The fitted model describes the observed data well, supported by the fact that majority of observed data points lie within the 95% confidence bound region for the state populations. There is, however, one data point which is not contained within this region, represented by the observation for the year 1977. An explanation for this might be found in the consideration of illegal harvesting. The data documenting wildebeest capture by harvesting began in 1977, whereas observations of wildebeest abundance began in 1960. Consequently, there are essentially two different population trajectories projected by the model – before 1977 and after 1977. Due to this, the 1977 data point itself is not well fitted by model which incorporates both pre- and post-harvesting observations. From the plot of the model fit it can also be seen that the first six observations are fitted extremely well, as they lie very close to the red fitted line, since these observations are associated with low standard deviations, whereas the subsequent observations show greater deviation from this line. The greater variance in the latter years of the observation combined with the introduction of harvesting into the model in 1977 may explain why the model fits poorly to the 1977 data. The beginning of harvesting documentation may have some bearing in the explanation of the widening confidence region, as well as the relative lack of funding received by the data collectors in later years, leading to less precise observations.

The model as it stands can be used, with caveats noted above, to predict future abundance ranges, although the managers of the region may wish to commission further work to incorporate density dependence, and/or to spend some resources to collect more survey data. With enough observations, we could conduct validation of the model by only using the first 75% (for example) of data points to parameterise the model, and subsequently assess how well the future predictions fit the remaining 25% of the data.

Contribution statement

Everyone contributed equally to writing the code, interpreting the results, and planning the structure of the report. AR concentrated on interpreting the results, writing them up, and editing the other sections. AM concentrated on model fitting, coding, plotting, report writing and editing, WEE concentrated on report writing and editing, coding, developing the plots and tables. KA concentrated on developing the methodology section and the Gelbin-Rubin implementation, as well as writing and editing the report. All group members attended meetings and debated and contributed to the contents of the report. Overall, the contributions were deemed to be equal and workload was distributed fairly among the group.

References

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS): Flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics, 41*(6), 628–649.

Hilborn, R. and Mangel, M. (1997). The Ecological Detective: Confronting Models with Data. Monographs in Population Biology 28. Princeton University Press. ISBN: 9780691034973

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325–337.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, vol 6, 7-11

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL: https://www.R-project.org/.

Code Supplement

Please see separately attached R file “Team C Assignment 2.R” for fully commented code for our model.