**Chapter 2**

**Literature Review**

**2.1 Introduction to Population Biology**

Population biology is the study of biological populations (Hastings, 2013). The goal of the study of population biology is to identify why a species populations fluctuate the way that they do. Central to this goal is the creation of theoretical models used to model the fluctuations of a species population (Royama, 2012). In the context of species conservation, accurate population models are vital. These models, if correctly engineered, can be used to analyse historical data, and extrapolated to future dates.

There are 4 primary factors that contribute to the change in population for most species. These include births and immigration, which increase a species population, and deaths and emigration, which decrease a species population. These 4 primary factors contribute directly to the intrinsic growth rate of the population. There are then further, secondary, factors that influence the four primary factors. These include population density, distribution, sex ratio and age structure. Furthermore, the environments carrying capacity also influences the 4 primary factors. All these factors, directly or indirectly, affect the population dynamics of the species exposed to them. The ways in which species interact with their ecosystem can be very complex and it’s almost impossible to account for all the variables that influence a species population dynamic.

Nevertheless, population models are still useful, albeit not all encompassing. There are many examples of positive outcomes due to population models. One example is how population dynamics modelling played an important role in reversing the decline in the endangered Loggerhead Sea Turtle in the US (Crouse et al., 1987; Crowder et al., 1994). Initial conservation efforts, in the 1980s, had been focused on protecting the turtle nests and hatchlings. However, it was later discovered, using population models with a focus on species age, that the declining growth rate of these turtles were attributed to subadult and adult turtles at sea. These findings lead to the use of turtle excluder devices (TEDs) in shrimp trawls in turtle habitats and has since had a substantially positive impact on the growth rate of the Loggerhead turtles.

Population models have also been used to manage decisions regarding over-abundant species in an ecosystem. The American bullfrog is an alien species on Vancouver Island. These bullfrogs have been rapidly increasing and negatively affecting biodiversity in its ecosystem. A matrix model investigation conducted by Govindarajulu et al, (2005) found that the current strategy of removing tadpoles only lead to higher tadpole survival due to reduced competition in a density-dependent context. Instead, they reported that the culling of tadpoles in autumn would provide the most effective approach in managing the over-abundant bullfrogs.

As is the case with all models, “all models are wrong, but some are useful”. The complexity of species interactions with each other and their environment make it impossible for any population modeller to account for all variables in their model. However, as proved above, models can still be very useful in species conservation when the correct variables are accounted for, and the correct model used.

**2.2 State Space Models**

State space models (SSMs) have become a popular tool used by ecologists to analyse ecological time-series. There is enough evidence to argue that SSMs should be the default tool in analysing time series for ecologists (Auger-Méthé et al., 2021). SSMs have been used in various ecological studies such as fisheries stock assessment (Aeberhard et al., n.d.), movement ecology (Patterson et al., 2008) and, most commonly, modelling population dynamics (Newman et al., 2014). These are only a few examples of where SSMs are used. Their application in ecology is vast.

**// description of why SSMs are so widely used (the whole separating latent and observation error thing)**

SSMs are able to model temporal autocorrelation in such a way that it separates observation error from process error (Auger-Méthé et al., 2021). Herein lies the value of SSMs. SSMs are hierarchical models that essentially model two different time series. One being the state process (latent process, hidden process) which is the unobserved time series that represents the true underlying population dynamics. The other being the observation process that represents the observed counts of that state process. For example, in a situation where you are analysing the population dynamics of bird species, the actual bird population would be the state process and the bird counts would be the observation process. The bird counts would, of course, be subject to observation error. Therefore, separating the observation process from the state process provides a more accurate estimation of the true population. The beauty in this methodology is that variation attributed to population size, population growth rate, environmental stochasticity and any other forms of stochasticity can be estimated and programmed into the model (Kéry & Schaub, 2011). This allows the ecologist to program the SSM time series such that it accounts for known biological factors that affect the population dynamics. This includes birth rates, death rates, fecundity, immigration, emigration, etc. Figure 1 shows a visual illustration of this separation between state and observation processes.

When using SSMs one must take a few assumptions into consideration:

1. SSMs can only produce unbiased estimates of population size if false negative (detection probability) and false positive (double counting) observations cancel out on average.
2. SSMs can only produce unbiased estimates if detection probability is less than 1 and has no discernible pattern over time.
3. Unbiased estimates cannot be produced if there are temporal patterns in detection probability or false positive rates.

(Kéry & Schaub, 2011)

**// different SSMs used in time series (density dependant, linear, exponential, etc)**

There are many ways to develop an SSM. The key aspects to determine the structure of the model is based on your data and the requirements of the research. This section explores and scrutinizes some of the SSM models that can be used in the context of avian population dynamics.

**Simple Exponential Model**

The first model to consider is the simplest, to aide understanding. SSMs, when applied to population dynamics, exhibits a Markovian process. The population size at time t+1 is dependent on the population size at time t. A simple model to describe the SSM is the exponential population dynamics model:

 (1)

Where refers to the population growth rate at time t to T-1, T being the number of years in the time series. Nt refers to the population size at time t. Kery and Shaub (2011) use this model as the basis to understand SSMs.

The fundamental goal of population dynamics analysis is to calculate the population growth rate and determine the factors affecting it. This is a straightforward task, given that the true population is known each year. With a true population size the ecologist can determine factors such as the strength of density dependence and the influence of the environment on the population. However, knowing the true population of a species is not possible. Instead the ecologist only has raw counts which are highly susceptible to observation error. This observation error must be accounted for in the model. If not, the ecologist could misinterpret observation error for some form of biological variation.

SSMs deal with observation error as it separates the state process from the observation process. Equation 1 represents the state process:

 (1)

 (2)

In the case of the SSM, the population growth rate is represented as a Normal random process. The mean,  , is the long-term growth rate of the population. The variance, , represents the stochasticity of the growth rate caused by environmental variability. The first year, N1, isn’t defined in equation 1 but modelled as a prior distribution in a Bayesian framework or it can be fixed to the initial count.

The next set of equations needed is the observations. The observations are dependent on the state process and are defined as follows:

 (4)

 (5)

Here yt represents the counts observed from the true population Nt. εt refers to the observation error with mean 0 and variance  (observation error). The model assumes that λt and εt are serially independent. That is they are independent from each other and both are identically distributed.

These equations display the flexibility of the SSM. The state process can be modified to include processes such as survival rate, birth rate, immigration, etc. Or one could use a logarithmic model instead of an exponential one. While the observation process can be modified to account for different types of observation errors such as binomial or poisson errors.

**SSM for migrant and non-migrant avian species**

The simple model, described above, is very useful as it implicitly accounts for various biological and environmental variation. Given the flexibility of the model, one can modify the model to account for known biological processes that affect population variation. This works well when modelling migratory avian species (Barshep, 2017; Camp et al., 2016; Wilson et al., 2011). Barshep (2017) uses SSMs on data contained in the CWAC dataset. The aim of the study was to model population changes of 29 migrant species and 25 non-migrant species then identify which life-history traits was most responsible for these fluctuations. Two different SSMs were used. Both were logarithmic models. Model one, for resident, “sedentary residents with limited or no seasonal movement between sites”. Model 2, for migrants, “Palaearctic migrants and intra-African migrants”.

Model 1 is a univariate model that has summer/winter as a covariate. The observation process is modelled as followed:

 (6)

Where ln(yt) is the logarithm of the count, mu\_t is the true population, gamma\_t is the summer/winter covariate and alpha\_t is the observation error.

The state process was modelled as a linear trend model as the resident species tend to remain in the same place all year round. The state process is represented as follows:

ln(relative population size at time t):  (7)

where Beta\_t is the rate of change of the population over time. Gamma\_t and Beta\_t is modelled as follows:

ln(population change):  (8)

ln(seasonal effect):  (9)

The variables alpha\_t, w\_t, epsilon\_t and zeta\_t are all serially independent. This concludes model one that is applied to resident species. See figure 2 for a visual illustration of model 1.

Model 2 is applied to migrant species and considers their migratory movement. The migrant species in South African wetlands tend to breed in the Northern Hemisphere (during the southern winter) then migrate south for the southern summer. The young species tend to remain in the south during the subsequent southern winter period. Model 2 was thus formulated as a bivariate time series model with the summer population in the state process as the main focus. The proportion of species that stay over for the subsequent southern winter was modelled as the second covariate. Figure 3 shows a visual description of this model. The mathematical representation of model 2 is as follows:

*Observation process*

ln(summer count):  (10)

ln(winter count):  (11)

where λt is the proportion of migrant species that stay over from summer to winter.

*State process*

ln(summer population change):  (12)

ln(population change):  (13)

ln(winter to summer ratio):  (14)

Here, again, alpha\_t, e\_t, w\_t, zeta\_t and epsilon\_t are all serially independent.

These models may be highly simplified and could still explicitly account for factors such as birth rate, death rate, density dependence, etc. However, it is difficult to obtain this data and to account for environmental factors for each wetland. The SSM model is beneficial to use in this case as all the unknown ecological variability is captured in the error terms in the state process.

**Inferential approaches to SSM**

SSMs are flexible in that it allows for may different types of biological processes to be coded into the model. SSMs are also flexible in terms of how it’s fit. There is 2 fundamentally different ways to fit SSMs. That is a frequentist approach or a Bayesian approach (Auger-Méthé et al., 2021).

**Frequentist approach**

Brief description

The frequentist, or classical, approach is common in most statistical analysis. The frequentist approach focuses on long run averages and p-values to determine the probability of obtaining a realised outcome, or something more extreme, given the null hypothesis (Bolker, 2008). If the outcome probability is significantly small you reject the null hypothesis, otherwise you cannot reject it. According to Bolker (2008), “the classical framework says you can never prove that something is true, you can only fail to prove it is false.”

Pros/cons

The frequentist approach is widely used in statistics but has a few drawbacks. Firstly, the framework relies on hypothetical probabilities of outcomes based on the experiment. The p-value is also a contentious issue as it is often misused in ecological research (Yoccoz, 1991).

When fitting SSMs according to the frequentist framework, we essentially perform maximum likelihood estimation on the joint likelihood of the parameters, θ, and the states, z. The mathematical structure can be illustrated as follows:

 (15)

Where θo and θp represent the observation process parameters and the state process parameters respectively. And T representing the length of the time series. To simplify the maximum likelihood process, one would first maximise the marginal likelihood and integrate out the hidden states (z). Thus, only having to maximise the marginal likelihood with respect to the parameters (θ). To estimate the hidden states, one would use the predicted parameters and calculate the conditional distribution of the hidden states given the observations (y) and the estimated parameters (θ^hat). Calculating the maximum marginal likelihood with respect to the parameters is a computationally expensive problem. This is the common pitfall for SSMs in general, whether one uses the frequentist or Bayesian approach.

Applications in the SSM context (Kalman filter, laplace approximations, sequential monte carlo)

There are two common methods one can use to evaluate the maximum marginal likelihood. Namely the Kalman Filter and the Laplace approximation method (which approximates the maximum marginal likelihood as opposed to evaluating it).

**Bayesian approach**

Brief description

While the frequentist approach is based on hypothetical probabilities, the Bayesian approach is based on prior distributions. Where the frequentist method defines unknown model parameters as long-term averages, the Bayesian approach defines unknown model parameters as probability distributions. These distributions are based on expert knowledge.

The Bayesian approach resolves the conceptual issues brought about in the frequentist approach. Instead of hypothetical probabilities, the Bayesian approach focuses on the data at hand. This way the statistician can make direct statements about the probability of hypotheses or parameter values without the use of contentious p-values (Bolker, 2008).

Bayesian approaches can also encapsulate expert knowledge in the formulation of its priors. This could be both a good thing and a bad thing. These priors affect the outcome of the model and has been used as a criticism of the Bayesian framework (Dennis, 1996). However, if the expert knowledge is sound, it would greatly benefit the model and will lead to favourable and more biologically sound outcomes. If incorrect, the ecologist may be adding erroneous biological or ecological beliefs into the model. Thus, producing an outcome that does not relate to reality.

Given the criticism of subjective priors (Lele & Dennis, 2009), the least controversial way for choosing priors is, therefore, to use non-informative priors as far as possible. These non-informative priors refer to normal or uniform priors. However, it is also believed that using only non-informative priors limits the benefits the Bayesian approach has to offer (Banner et al., 2020).

Besides the use of priors being a contentious topic, the Bayesian approach still holds a vast array of benefits. Bayesian models are conceptually easier to understand when applied to hierarchical models (Kéry & Schaub, 2011) and their ability to predict the true state process is invaluable in ecology (Banner et al., 2020). In population dynamics, there is often a true population and observations based on that population. These observations very seldom reflect the true population and are often prone to observation error. Given this population dynamics structure, the Bayesian approach is a more intuitive method to use.

When applying SSMs in a Bayesian context, the main goal is to determine the posterior distribution for the states and the parameters based on their prior distributions. The mathematical representation of this approach is as follows:

 (16)

Where  is the joint likelihood and π(θ|λ) is the prior distribution(s). The denominator of equation 16 represents the probability of the data for all possible values of the states and parameters. This equation forms the basic structure of the bayes theorem:

 (17)

Where y refers to the observed data and θ refers to the state and the parameters.

The posterior distribution is the focus in the Bayesian approach. This is done by obtaining random draws from the right-hand side of equation 16. This is a computationally expensive task that is impossible to do by hand. There are various algorithms that approximate the posterior distribution. Namely, Metropolis Hastings samplers, Hamiltonian Monte Carlo and the Gibbs sampler, amongst many. These algorithms all fall under the broader Markov Chain Monte Carlo (MCMC) algorithm class. MCMC are methods that approximate the posterior distribution by sampling from a probabilistic distribution (b, 2017).

Another important aspect in the Bayesian approach is to apply convergence tests on the MCMC outcome, regardless of the sampler used. Testing convergence ensures that the sampler used, after approximating the probabilistic distribution many times, has converged to a common interval across all MCMC chains. This can often be viewed on a traceplot of MCMC outcomes (Auger-Méthé et al., 2021). A more formal approach would be the Gelman-Rubin metric (Gelman & Rubin, 1992), which is a popular metric to measure convergence in an MCMC context.

**Model Selection**

A major drawback for SSMs is the issue of model selection. Model selection for SSM is still an area of active research (Jonsen et al., 2013). SSM model selection becomes difficult as the computation power needed to run an SSM is so large. Often ecologists end up using only one SSM model in the study because of limited computational power.

Model comparison methods will differ depending on whether the SSM is conducted in a Frequentist framework or a Bayesian framework. The most used model selection method for a frequentist SSM is the Akaike Information Criterion (AIC). The AIC is a popular model selection method in many statistical models. The AIC measures how well the model fits the data and penalizes models with a higher number of parameters. AIC, however, doesn’t work well for SSMs, even in a frequentist framework, for various reasons (Auger-Méthé et al., 2021).

For SSMs in a Bayesian context, AIC is not well defined (Hooten & Hobbs, 2015). There are 3 notable model selection methods used for Bayesian models. These are the WAIC, DIC and BIC. Many sources of literature have proven DIC and BIC to not be very good at all in Bayesian model selection (Auger-Méthé et al., 2021; Gelman et al., 2014; Hooten & Hobbs, 2015). Sources have found WAIC to be more effective than DIC and BIC, even though it still fails in many cases (Gelman et al., 2014). The method most of the literature agreed upon was a cross validation method. In this cross-validation method, the SSM would be fit with a training set of the data, while an unused subset of the data (the test set) is used to determine the effectiveness of the SSMs forecasting capacity. The SSM that forecasts values closest to the corresponding test set values is then considered the “best” model. This method works well but has its limitations when data is sparse. In the case of sparsity, splitting the data into a train and test set becomes difficult as the train set may not be sufficient to produce an accurate model. Furthermore, the cross-validation method for SSMs is computationally expensive.

Yoccoz (1991) and Cox et al. (1977), states that, at times, ecologists tend to overemphasize the use of statistical measures for model selection and fail to prioritise biological or ecological significance. Given the contentious issues regarding model selection for SSMs, it may be of benefit to choose the SSM that is based on sound ecological fundamentals rather than contentious model selection statistics. More time and effort should be focused on the modelling process and ensuring the correct ecological processes are modelled into the SSM.

Pros

Cons

Applications in the SSM context (metropolis hastings sampler, Hamiltonian monte carlo, convergence, priors)

// inferential approaches to fitting SSMs (discuss MCMC, Hastings metropolis and Kalman filter algorithms here, use table 2 in a guide to state space modelling.pdf)

// model selection (why it’s hard, AIC, WAIC, DIC and CV, then using biology understanding to choose models instead (proposed by yoccoz)) – a draw back for SSMs

**References**

Aeberhard, W. H., Cantoni, E., Field, C., Künsch, H. R., Mills Flemming, J., & Xu, X. (n.d.). Robust estimation for discrete-time state space models. *Scandinavian Journal of Statistics*, *n/a*(n/a). https://doi.org/10.1111/sjos.12482

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. A., Leos-Barajas, V., Flemming, J. M., Nielsen, A., Petris, G., & Thomas, L. (2021). A guide to state-space modeling of ecological time series. *ArXiv:2002.02001 [q-Bio, Stat]*. http://arxiv.org/abs/2002.02001

b. (2017, December 24). *A Zero-Math Introduction to Markov Chain Monte Carlo Methods*. Medium. https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50

Banner, K. M., Irvine, K. M., & Rodhouse, T. J. (2020). The use of Bayesian priors in Ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, *11*(8), 882–889. https://doi.org/10.1111/2041-210X.13407

Barshep, Y. (2017). Identifying ecological and life-history drivers of population dynamics of wetland birds in South Africa. *Global Ecology and Conservation*, 12.

Bolker, B. M. (2008). Ecological Models and Data in R. In *Ecological Models and Data in R*. Princeton University Press. https://doi.org/10.1515/9781400840908

Bozeman Science. (2015, October 1). *Population Ecology*. https://www.youtube.com/watch?v=PQ-CQ3CQE3g

Camp, R. J., Brinck, K. W., Gorresen, P. M., & Paxton, E. H. (2016). Evaluating abundance and trends in a Hawaiian avian community using state-space analysis. *Bird Conservation International*, *26*(2), 225–242. https://doi.org/10.1017/S0959270915000088

Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J. F., Barndorff-Nielsen, O., & Keuls, M. (1977). The Role of Significance Tests [with Discussion and Reply]. *Scandinavian Journal of Statistics*, *4*(2), 49–70.

Crouse, D. T., Crowder, L. B., & Caswell, H. (1987). A Stage-Based Population Model for Loggerhead Sea Turtles and Implications for Conservation. *Ecology*, *68*(5), 1412–1423. https://doi.org/10.2307/1939225

Crowder, L. B., Crouse, D. T., Heppell, S. S., & Martin, T. H. (1994). Predicting the Impact of Turtle Excluder Devices on Loggerhead Sea Turtle Populations. *Ecological Applications*, *4*(3), 437–445. https://doi.org/10.2307/1941948

Dennis, B. (1996). Discussion: Should Ecologists Become Bayesians? *Ecological Applications*, *6*(4), 1095–1103. https://doi.org/10.2307/2269594

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Govindarajulu, P., Altwegg, R., & Anholt, B. R. (2005). Matrix Model Investigation of Invasive Species Control: Bullfrogs on Vancouver Island. *Ecological Applications*, *15*(6), 2161–2170. https://doi.org/10.1890/05-0486

Hastings, A. (2013). *Population Biology: Concepts and Models*. Springer Science & Business Media.

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28. https://doi.org/10.1890/14-0661.1

Jonsen, I. D., Basson, M., Bestley, S., Bravington, M. V., Patterson, T. A., Pedersen, M. W., Thomson, R., Thygesen, U. H., & Wotherspoon, S. J. (2013). State-space models for bio-loggers: A methodological road map. *Deep Sea Research Part II: Topical Studies in Oceanography*, *88–89*, 34–46. https://doi.org/10.1016/j.dsr2.2012.07.008

Kéry, M., & Schaub, M. (2011). *Bayesian Population Analysis using WinBUGS ‐*. 444.

Lele, S. R., & Dennis, B. (2009). Bayesian Methods for Hierarchical Models: Are Ecologists Making a Faustian Bargain? *Ecological Applications*, *19*(3), 581–584.

Newman, K. B., Buckland, S. T., Morgan, B. J. T., King, R., Borchers, D. L., Cole, D. J., Besbeas, P., Gimenez, O., & Thomas, L. (2014). State-Space Models. In K. B. Newman, S. T. Buckland, B. J. T. Morgan, R. King, D. L. Borchers, D. J. Cole, P. Besbeas, O. Gimenez, & L. Thomas (Eds.), *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods* (pp. 39–50). Springer. https://doi.org/10.1007/978-1-4939-0977-3\_3

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23*(2), 87–94. https://doi.org/10.1016/j.tree.2007.10.009

*Read “Using Science to Improve the BLM Wild Horse and Burro Program: A Way Forward” at NAP.edu*. (n.d.). https://doi.org/10.17226/13511

Royama, T. (2012). *Analytical Population Dynamics*. Springer Science & Business Media.

Wilson, H. B., Kendall, B. E., Fuller, R. A., Milton, D. A., & Possingham, H. P. (2011). Analyzing Variability and the Rate of Decline of Migratory Shorebirds in Moreton Bay, Australia. *Conservation Biology*, *25*(4), 758–766. https://doi.org/10.1111/j.1523-1739.2011.01670.x

Yoccoz, N. G. (1991). Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. *Bulletin of the Ecological Society of America*, *72*(2), 106–111.