Trend analysis for USFWS species status assessment for bull trout (*Salvelinus confluentus*)

Mark D. Scheuerell[1,*]

[1] U.S. Geological Survey Washington Cooperative Fish and Wildlife Research Unit, School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA

[*] Correspondence: Mark D. Scheuerell <scheuerl@uw.edu>

Background

Bull trout (*Salvelinus confluentus*) in the western U.S. were listed as threatened under the U.S. Endangered Species Act in 1998. The purpose of this analysis is to estimate trends in the abundance (counts) of bull trout within predefined core areas spread across Oregon, Washington, Idaho, and Montana, as part of the current Species Status Assessment (SSA).

Data

I provided a MS Excel template for the desired data in a "tidy" format, which consisted of the following 8 fields (columns):

- `dataset` (i.e., integer value for unique ID)
- `recovery unit` (e.g., Mid Columbia)
- `core area` (e.g., South Fork Clearwater)
- `popn/stream` (e.g., Crooked River)
- `metric` (e.g., abundance)
- `method` (e.g., redd survey)
- `year`
- `value` (i.e., counts)

The data coordinators also provided me with some metadata indicating which of the data specific to a location were generally for adults versus juveniles. Data files were subsequently provided to me by data coordinators from each of the four states:

- Oregon: Stephanie Gunckel (ODFW)
- Washington: Marie Winkowski (WDFW)
- Idaho: Brett Bowersox (IDFG)
- Montana: Dan Brewer (USFWS)

The data from Montana came via biologists with the USGS (Clint Muhlfeld, Tim Cline), and did not conform to the template file I had provided. Thus, those data were subjected to additional cleaning prior to their inclusion with the data from other states (see below).

Modeling framework

Population model

I fit discrete-time versions of exponential models for population growth (decline), such that the abundance of bull trout (N) is a function of the initial population size N_0, time (t), the population growth rate (u), and a time-dependent stochastic effect of the environment (w). Specifically, in continuous time the model is

$$N(t) = N_0 \exp(u) \exp(wt). \tag{1}$$

In discrete time, with a time step of 1 unit (e.g., a year), the model becomes

$$N_t = N_{t-1} \exp(u + w_t). \tag{2}$$

If we take the logarithm of both sides and define $x_t = \log(N_t)$, we have

$$x_t = x_{t-1} + u + w_t. \tag{3}$$

Further defining $w_t \sim \mathrm{N}(0, q)$ leads us to a so-called "biased random walk" model, where u is the tendency for the population to increase or decrease each time step (i.e., the bias), and w_t is some unknown stochastic aspect of the environment that partially drives population dynamics.

Observation model

The data available to us rarely come from complete censuses, and instead are typically derived from partial counts. Furthermore, mistakes may occur when counting individuals or redds. Thus, we should account for these possible sampling or observation errors with a so-called "data model."

In this case, we assume that the data in hand are a somewhat distorted view of the "true state of nature," such that the logarithm of the observed count at time t (y_t) equals that of the true count plus or minus some error. Specifically, we can write this as

$$y_t = x_t + a + v_t \tag{4}$$

where a is an offset to account for partial sampling, and $v_t \sim \mathrm{N}(0, r)$.

State-space model

We can combine equations (3) and (4), along with a definition for the initial state (x_0), into a so-called "state-space model," where

$$\begin{aligned} y_t &= x_t + a + v_t \\ x_t &= x_{t-1} + u + w_t \\ x_0 &\sim \mathrm{N}(\mu, \sigma) \end{aligned} \tag{5}$$

Multiple populations

Here we want to estimate the annual change in population size for each of the many different core areas across the four states. Furthermore, some core areas comprise several different populations/locations, so we need to frame our state-space model in a multivariate context.

Observation model

If we have n different populations within a core area, then our observation model becomes

$$y_{i,t} = x_{i,t} + a_i + v_{i,t} \tag{6}$$

where $y_{i,t}$ is the log-count for population i and year t, a_i is an offset to account for partial sampling in population i, and $v_{i,t} \sim \mathrm{N}(0, r_i)$[1]. We can combine each of the population specific observation models into a matrix form, such that

$$\begin{aligned} y_{1,t} &= x_{1,t} + a_1 + v_{1,t} \\ y_{2,t} &= x_{2,t} + a_2 + v_{2,t} \\ &\vdots \\ y_{n,t} &= x_{n,t} + a_3 + v_{n,t} \end{aligned} \tag{7}$$

becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_t + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}_t , \tag{8}$$

or more compactly in matrix notation as

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{a} + \mathbf{v}_t. \tag{9}$$

[1] Here the variance of the observation errors is assumed to be population specific, but it might be reasonable to assume that each survey/census type might have the same variance, such that $v_{i,t} \sim \mathrm{N}(0, r)$.

where $\mathbf{y}_t$, $\mathbf{x}_t$, $\mathbf{a}$, $\mathbf{v}_t$ are all $n \times 1$ vectors, and $\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$.

Population model

Just as we did for the observation model, we can write the models for population dynamics as

$$x_{i,t} = x_{i,t-1} + u_i + w_{i,t} \tag{10}$$

where u_i is the bias, which is unique to each population[2], and $w_{i,t} \sim \text{N}(0, q_i)$[3]. We can again express all of the population models in matrix form, such that

$$\begin{aligned} x_{1,t} &= x_{1,t-1} + u_1 + w_{1,t} \\ x_{2,t} &= x_{2,t-1} + u_2 + w_{2,t} \\ &\vdots \\ x_{n,t} &= x_{n,t-1} + u_3 + w_{n,t} \end{aligned} \tag{11}$$

becomes

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_t + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_t, \tag{12}$$

or more compactly in matrix notation as

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t \tag{13}$$

where $\mathbf{x}_t$, $\mathbf{x}_{t-1}$, $\mathbf{u}$, $\mathbf{w}_t$ are all $n \times 1$ vectors, and $\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$.

State-space forms

At this point, however, we are assuming that the monitoring data for each population is telling us something about only the specific population itself, rather than contributing information to the population trend at the larger scale of their core area, which is the really

[2] It might be reasonable to assume that some/all of the populations have the same bias, given their membership within a core area.

[3] Here the variance of the process errors is assumed to be population specific, but it might be reasonable to assume that they all have the same variance, given their membership within a core area, such that $w_{i,t} \sim \text{N}(0, q)$.

the scale of interest here. Thus, we need to modify our equations to accommodate this hierarchical framework.

For example, assume that we have $p = 2$ core areas (call them A and B), each with data from 2 representative populations. In this case, $n = 4$, but the number of states (i.e., the number of rows in $\mathbf{x}_t$) is 2, so we need a way to "map" each of the observed time series onto its respective core area. We begin by writing out the equations for the observations in long matrix form akin to equation (8), such that

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}_t = \begin{bmatrix} x_A \\ x_A \\ x_B \\ x_B \end{bmatrix}_t + \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}_t , \tag{14}$$

Because both x_A and x_B appear twice in equation (14), we can use a 4×2 matrix of 1's and 0's as our map. Specifically, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}_t = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_A \\ x_B \end{bmatrix}_t + \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}_t , \tag{15}$$

We can write equation (15) more compactly in matrix notation as

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{v}_t. \tag{16}$$

where $\mathbf{y}_t$, $\mathbf{a}$, and $\mathbf{v}_t$ are all $n \times 1$ vectors, $\mathbf{Z}$ is an $n \times k$ matrix, and $\mathbf{x}_t$ is a $k \times 1$ vector.

The equation for the population dynamics in each of the 2 core areas then becomes

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix}_t = \begin{bmatrix} x_A \\ x_B \end{bmatrix}_{t-1} + \begin{bmatrix} u_A \\ u_B \end{bmatrix} + \begin{bmatrix} w_A \\ w_B \end{bmatrix}_t , \tag{17}$$

which can be written more compactly in matrix notation as

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \tag{18}$$

and combined with equation (16) to form the full multivariate state-space model

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{v}_t \\ \mathbf{x}_t &= \mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t. \end{aligned} \tag{19}$$

Thus, by simply altering the dimensions of $\mathbf{Z}$, and the locations of 1's and 0's within it, we can evaluate any number of different hypotheses about how the population dynamics are structured spatially. For example, if we set $\mathbf{Z}$ equal to an $n \times n$ identity matrix, where

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{20}$$

then each of the time series of data is assumed to represent a unique state of nature. If, on the other hand, we set $\mathbf{Z}$ equal to an $n \times 1$ column vector of 1's, such that

$$\mathbf{Z} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{21}$$

then each of the time series of data is assumed to represent a sample from a single state of nature.

Variance specification

The multivariate state-space model allows us to be quite specific about how the observation errors ($\mathbf{v}_t$) and process errors ($\mathbf{w}_t$) are related to one another, if at all. In the most simple case, the errors could be independent and identically distributed (IID), such that (for the observation variance)

$$\mathbf{R} = \begin{bmatrix} r & 0 & \cdots & 0 \\ 0 & r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r \end{bmatrix}. \tag{22}$$

Alternatively, the errors might be independent, but not identically distributed

$$\mathbf{R} = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n \end{bmatrix}, \tag{23}$$

or identically distributed, but not independent

$$\mathbf{R} = \begin{bmatrix} r & c & \cdots & c \\ c & r & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & r \end{bmatrix}. \tag{24}$$

Model fitting

All models were fit using the `{MARSS}` package (Holmes *et al.* 2012; Holmes *et al.* 2020) for the **R** computing software (R Core Team 2020). All of the data and code necessary to reproduce the results of the analysis can be found online at https://github.com/mdscheuerell/bulltrout.

References

Holmes E, Ward E, Scheuerell M, and Wills K. 2020. MARSS: Multivariate autoregressive state-space modeling.

Holmes EE, Ward EJ, and Wills K. 2012. MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal* **4**: 30.

R Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.