Stat 532 Project: Statistical arrival models to estimate missed passage counts at fish weirs

By Suresh Andrew Sethi and Catherine Bradley, Project by Leslie Gains-Germain

Fall 2015

Contents

1 Introduction 2

2 Model 2

3 Simulation of artificial data 4

4 Model fitting and results 8

5 Posterior predictive checks 8

6 My opinions and what I learned 8

7 Recommendations 8

8 References 8

9 R Code Appendix 8

1 Introduction

The goal of this paper is to provide a new method for imputing passage counts on missed dates when enumerating fish passing by a weir. A weir is a device, pictured below, designed to span an entire stream and count every single fish that passes by.

Weirs do not always operate as intended, however. High water can destroy a weir, bears can interfere with the weir or the weir operator, and ice can delay the installation or require early removal of a weir.

Sethi and Bradley explain that the current method used to impute passage counts on missing dates "typically involve a "connect-the-dots" linear interpolation scheme (e.g. Gewin et al. 2005; Johnson et al. 2007)". They explain that while this method is easy to implement, it has two main drawbacks. First, it does not provide uncertainty in the missing passage estimates, and second it cannot estimate missing dates at the beginning and end of the run.

The methods proposed by Sethi and Bradley in this paper solve both of these drawbacks. They use a parametric curve to describe the passage of a run of fish at a weir (referred to as the arrival model), and a separate probability model to describe the variability of fish counts around the run curve. The following quote is a nice summary of their reasons for using Bayesian implementation.

> Models are fit in a Bayesian framework, providing a straightforward means to summarize uncertainty about total run size estimates, arrival model characteristics (e.g. peak run date), and estimates of predicted passage counts on missing-observation dates (Sethi and Bradley 5).

The main research goal is to provide a new method for estimating total run size and missed passage counts. The secondary research goal is to estimate parameters of the arrival model that describe the passage of fish over time.

2 Model

First, they specify the run curves, or arrival models, that describe the passage of fish over time at a weir. They use two common cumulative distribution functions to describe the run curves, the normal distribution and the skew normal distribution. It is important to note that they could have used any parametric curve for this purpose, but they chose common cumulative distribution functions for convenience. The following quote provides a nice concise explanation.

> To avoid confusion, it is worth emphasizing that the cumulative distribution functions used to specify arrival dynamics do not represent probability models for weir passage counts; they are merely convenient mathematical functions to describe the shape of the arrival curve of fish at a weir (Sethi and Bradley 8).

Below, I use the same notation they use in the paper because I think it's really clear.

Normal Run Curve $p_t = F_N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-(\nu-\mu)^2}{2\sigma^2}} d\nu$

Skew Normal Run Curve $p_t = F_{SN}(t|\xi, \omega, \alpha) = \frac{2}{\omega}\phi(\frac{t-\xi}{\omega})\Phi(\frac{\alpha(t-\xi)}{\omega})$

where p_t is the cumulative proportion of the run that has passed the weir at time step t. A time step t is defined as a 24 hour day because daily counts are commonly recorded as weir data.

F_N and F_{SN} are the cumulative distribution functions for the normal and skew normal distributions, respectively. μ and σ are the location and scale parameters for the normal model. For the skew normal model, ξ is the location parameter, ω is the scale parameter, and α is the shape parameter. ϕ is the standard normal density function and Φ is the standard normal cumulative distribution function.

Daily passage counts are calculated as follows.

$$c_t = S(p_t - p_{t-1})$$

where c_t is a daily passage count for a given run curve and S is the total run size scalar.

Below are the priors for the parameters of the normal run curve:

$$\mu \sim Unif(150, 300) \qquad f(\mu) = \frac{1}{150} I(\mu)_{(150,300)}$$

$$\sigma \sim Unif(1, 50) \qquad f(\sigma) = \frac{1}{49} I(\sigma)_{(1,50)}$$

$$log(S) \sim N(7.5, 0.25) \qquad f(log(S)) = \frac{1}{\sqrt{0.125\pi}} e^{-\frac{(log(S)-7.5)}{0.125}}$$

Below are the priors for the parameters of the skew-normal run curve:

$$\xi \sim Unif(150, 300) \qquad f(\xi) = \frac{1}{150} I(\xi)_{(150,300)}$$

$$\omega \sim Unif(1, 50) \qquad f(\omega) = \frac{1}{49} I(\omega)_{(1,50)}$$

$$\alpha \sim Unif(-10, 10) \qquad f(\alpha) = \frac{1}{20} I(\alpha)_{(-10,10)}$$

$$log(S) \sim N(7.5, 0.25) \qquad f(log(S)) = \frac{1}{\sqrt{0.125\pi}} e^{-\frac{(log(S)-7.5)}{0.125}}$$

Next, a probability model is used to describe the amount of variation in the passage counts around the run curve.

Normal Process Variation $\qquad c_t^0 \sim N(c_t, \tau)$

Negative Binomial Process Variation $\qquad c_t^0 \sim NegBinom(\lambda = \frac{\theta}{\theta + c_t}, \theta)$

where c_t^0 is an observed passage count, τ is the variance of the normal model, and θ is the dispersion parameter of the negative binomial model.

Below are the priors for the parameters of the normal process variation model:

$$\tau \sim Unif(0.1, 1000) \qquad f(\tau) = \frac{1}{999.9} I(\tau)_{(0.1,1000)}$$

Below are the priors for the parameters of the negative binomial process variation model:

$$\theta \sim Gam(0.1, 0.1) \qquad f(\theta) = \frac{\theta^{-0.9} 0.1^{0.1} e^{-0.1\theta}}{\Gamma(0.1)}$$

3 Simulation of artificial data

I started by writing my own code to simulate data. During the process, I realized that code for simulating data is included in the paper. I compared my code to the code they used for simulations, and I decided to show their code here (cleaned up a little bit by me) and include a thorough description of what they did. It was pretty similar to the code I wrote, but sometimes I learn more by going through someone else's code than I do from writing my own code, especially having first written it myself.

They started by defining the skew normal run curve. The location parameter of the skew normal run curve (ξ) was set at 185 because sets the run peak to occur around the 4th of July. The scale parameter (ω) was set at 7.5 because this produces a run of about two months long. This makes sense, because salmon usually run up the river in the summer, and the run usually lasts about two months. The skew parameter (α) was set to 0 in the simulations. The total run size was set at 10000 fish.

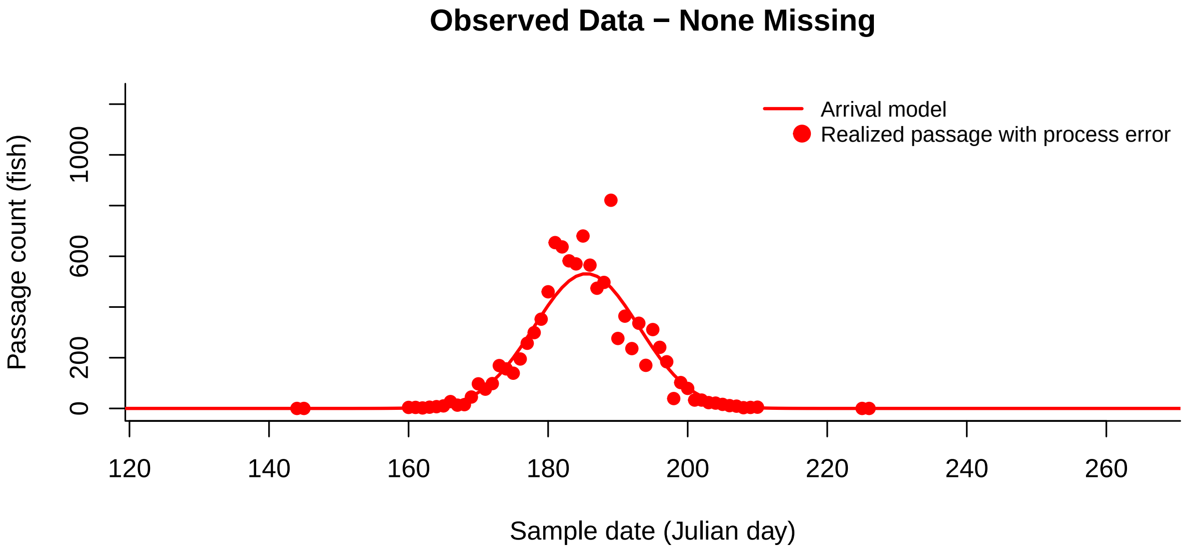
They used the psn function in the sn package to define the skew normal run curve. After defining the passage counts for each day according to the run curve, they simulated observed passage counts by taking random draws from a negative binomial distribution centered at the run curve with dispersion parameter equal to 7. The overdispersion parameter was set to 7 and "was chosen to reflect variation observed in actual Yukon River Pacific Salmon data (Sethi and Bradley 12)."

```
#Simulate "true" data modeled after Pacific Salmon runs: Normal arrival curve model,
#run size = 10,000 fish,
#negative binomial process error
set.seed(205) # set random seed if desired

# Normal-curve shaped arrival model parameters
true.day.v <- 1:365 # model time steps, modeled as 24 hour days, e.g. Julian days
skew <- 0 # force skew to zero
run.mu <- 185 # results in a mode on approximately 4th of July peak run
run.sd <- 7.5 # produces a run of about 40 days long
S <- 10000 # 10k fish run

#Negative Binomial variation parameter
theta <- 7.0 #overdispersion based on Pacific salmon data

# simulate run arrival under the arrival model
# for convenience the sn package is used, defining a skewnormal with shape = 0,
#i.e. Normal distribution
require(sn)
true.run.v <- S*(psn(true.day.v, xi = run.mu, omega = run.sd, alpha = skew,
                     engine = "biv.nt.prob")
                 - psn(true.day.v - 1, xi = run.mu, omega = run.sd, alpha = skew,
                       engine = "biv.nt.prob"))

# simulate observed data under process error
real.dat.v <- rnbinom(n = length(true.run.v), mu = true.run.v, size = theta)

# note, ignore NA warnings here as these are due to rounding issues producing
#very small negative counts in tails
```

Then, they set starting and ending observation days at the weir, day 160 and 210, corresponding to early June and late July, spanning the two month run. They did this to prevent numerical difficulties that would occur from small predicted estimates in the tails of the run. To bracket the beginning and ends of the run, they inserted known zero passage days before and after the run. These zero passage days reflect prior knowledge about when the run occurs. They chose to incorporate this prior knowledge because preliminary analyses indicated that including these known zero passage days sped up convergence, particularly when data were missing in the tails of the run (Sethi and Bradley 11). In the paper, they discuss sensitivity of the results to the placement of the zero passage dates (Sethi and Bradley 47). In the simulation shown here, they inserted known zero passage days before and after the run on days $144, 145, 225$ and 226.

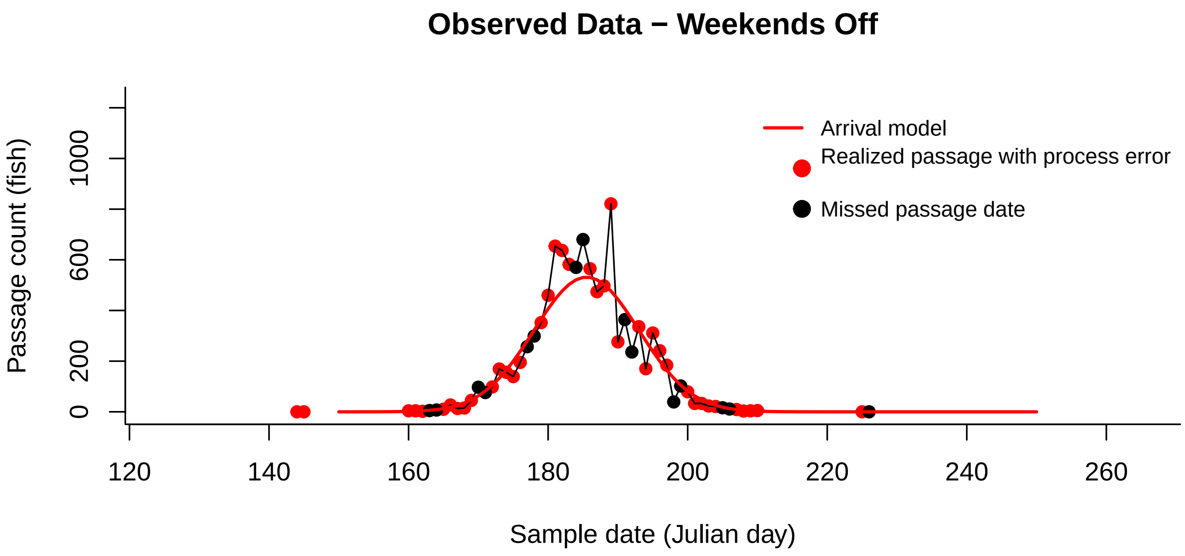
```
#Simulate observations at the weir with missing data following the "Weekends Off" scenario
#(observe 5, miss 2 days, ...)
obs.start <- 160 # first day to observe at weir, e.g. June 9th by this example
obs.end <- 210 # last day to observe at weir, e.g. July 29th by this example
obs.dat.v <- data.frame(Date = true.day.v[obs.start:obs.end],
                        Count = real.dat.v[obs.start:obs.end])

# include known-zero passage days
# Here, zero-passage dates before observation are equivalent to May 24 and 25
#post-observation zero-passage dates are equivalent to August 13 and 14
obs.dat.v <- rbind(data.frame(Date=144:145, Count=0),
                   obs.dat.v, data.frame(Date=225:226, Count=0))
```

The simulated observed data, before data was removed to represent missing data, is shown below.



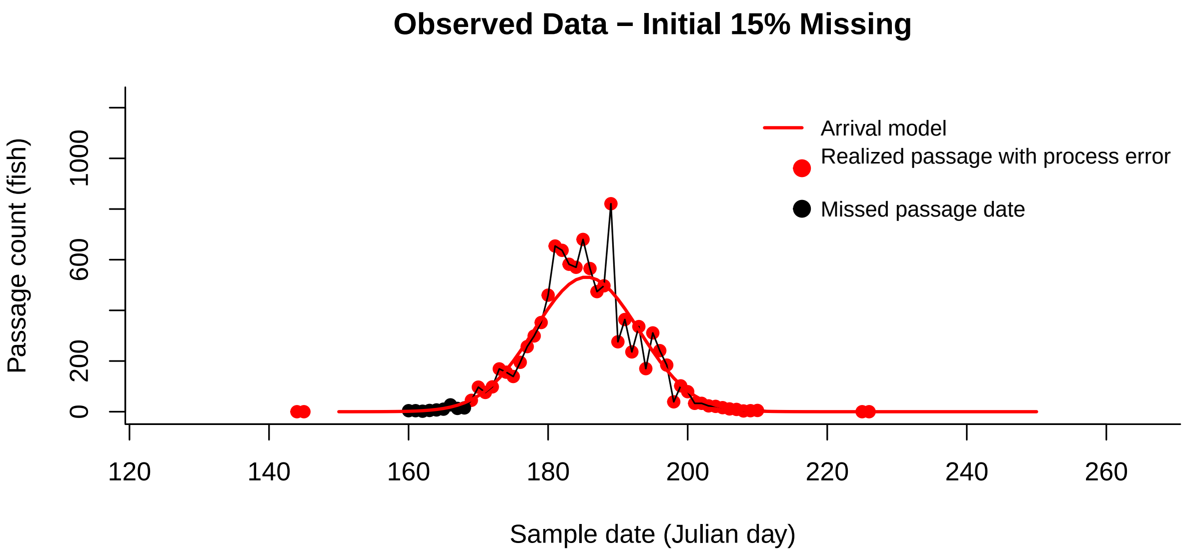
The following code removes some of the data to represent a “weekends off” scenario where data are missing from the weekends. The plot below shows the data after the missing dates have been removed.

```
# remove "weekend" data, i.e. sample for five days, take two off, sample five days...
`%notin%`<- Negate(`%in%`) # helper function
knockout.date.v <- obs.dat.v[1:nrow(obs.dat.v) %in%
                               c(seq(from=6,to=nrow(obs.dat.v),by=7),
                                 seq(from=7,to=nrow(obs.dat.v),by=7)), "Date"] #missed weekends
knockout.passage.v <- obs.dat.v[obs.dat.v$Date %in% knockout.date.v, "Count"] # missed passages
obs.dat.v <- obs.dat.v[obs.dat.v$Date %notin% knockout.date.v, ] #remove weekend passages from dataset
```



The following code removes data in the early tail of the run to represent a scenario where the weir couldn't be installed until after the run had began. The plot below shows the data with the initial 15% removed.

```
# remove first 15% of data
`%notin%`<- Negate(`%in%`) # helper function
knockout.date.v <- obs.dat.v[3:11, "Date"] #missed early tail
knockout.passage.v <- obs.dat.v[obs.dat.v$Date %in% knockout.date.v, "Count"] # missed passages (initia
obs.dat.v <- obs.dat.v[obs.dat.v$Date %notin% knockout.date.v, ] #remove intial 15%
```



4 Model fitting and results

5 Posterior predictive checks

6 My opinions and what I learned

7 Recommendations

8 References

Gewin, C.S., and VanHatten, G.K. 2005. Abundance and run timing of adult Pacific Salmon in the East Fork Andreafsky River, Yukon Delta National Wildlife Refuge, Alaska, 2003. U.S. Fish and Wildlife Service Data Series Report 2005-10, Anchorage, Alaska.

Johnson, D.H., Shrier, B.M., O'Neal, J.S., Knutzen, J.A., Augerot, X., O'Neil, T.A., and Pearsons, T.A. (Eds.) 2007. Salmon field protocols handbook. American Fisheries Society, Bethesda, Maryland.

Suresh, Sethi, and Catherine Bradley. "Statistical Arrival Models to Estimate Missed Passage Counts at Fish Weirs." Canadian Journal of Fisheries and Aquatic Sciences. Draft.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

9 R Code Appendix