Bayesian arrival model for Atlantic salmon smolt counts powered by environmental covariates and expert knowledge

Henni Pulkkinen, Panu Orell, Jaakko Erkinaro, Samu Mäntyniemi

H. Pulkkinen.* Natural Resources Institute Finland, Paavo Havaksen tie 3 P.O. Box 413, 90014-Oulu. Finland.
henni.pulkkinen@luke.fi

P. Orell. Natural Resources Institute Finland, Paavo Havaksen tie 3 P.O. Box 413, 90014-Oulu. Finland.
panu.orell@luke.fi

J. Erkinaro. Natural Resources Institute Finland, Paavo Havaksen tie 3 P.O. Box 413, 90014-Oulu. Finland.
jaakko.erkinaro@luke.fi

S. Mäntyniemi. Natural Resources Institute Finland, Latokartanonkaari 9, 00790 Helsinki. Finland.
samu.mantyniemi@luke.fi

Corresponding author: Henni Pulkkinen (phone: +358 40 485 0088)

Abstract

Annual run size and timing of Atlantic salmon smolt migration was estimated using Bayesian model framework and data from six years of a video monitoring survey. The model has a modular structure. It separates sub-processes of departing, traveling and observing, of which the first two together define the arrival distribution. The sub-processes utilize biological background and expert knowledge about the migratory behavior of smolts and about the probability to observe them from the video footage under varying environmental conditions. Daily mean temperature and discharge were used as environmental covariates. The model framework does not require assuming a simple distributional shape for the arrival dynamics and thus also allows for multimodal arrival distributions. Results indicate that 25 % - 75% of smolts pass the Utsjoki monitoring site unobserved. Predictive studies were made to estimate daily run size in cases with missing counts either at the beginning or in the middle of the run, indicating good predictive performance.

Keywords: salmon, arrival model, passage count, Bayesian models, environmental covariates, expert knowledge, biological realism, missing data, prediction

Introduction

Migratory fish species are often monitored along their migration routes at different life stages to collect fishery-independent data for stock assessment. Monitoring methods include e.g. counting fences, video, sonar and snorkeling counts, and common to all is that they rarely provide perfect information about the number of individuals passing the system (Dempson et al. 1991; Romakkaniemi et al. 2000; Orell et al. 2007; 2011). Part of the fish run may be missed because of difficult environmental conditions or device failures, there may be double counting, or the data

achieved may be otherwise partial or biased (e.g. Holmes et al. 2006). Thus, the assumptions made when interpreting the data can have a great influence on the estimated stock abundance.

Increasing the information beyond the actual monitoring data collected can improve the trustworthiness of stock assessments (Kuparinen et al. 2012): such information can be based on the resources available in other fields of biological research such as ecology or life history theory, but also on the physical and technical details of the observation processes. Methods of Bayesian inference enable both estimation of uncertainty and combination of various sources of information. Statistical methods are often used, however, without considering the biological realism and just focusing on statistical data analysis. Such a procedure easily leads to awkward model assumptions that cannot be biologically interpreted and the potential effects of those modelling decisions may pass undiscussed.

During the course of time, various methods have been used to estimate run dynamics and passage counts at point census stations. The simplest methods are based on “connect-the-dots” type of linear interpolation (Gewin and WanHatten 2005, Johnson et al. 2007) but such approach requires an observation before and after the missing time point and thus missing tails cannot be estimated. If it is assumed that a proportion passing the site on a given date is constant over the years then estimation of missing tails is possible using the methods proposed by, e.g, Van Alen (2000), but considering the variability, for example, in annual environmental conditions (e.g. Erkinaro et al. 1998; Orell et al. 2007; Otero et al. 2014), such assumption is problematic.

Hilborn et al. (1999) used a maximum likelihood method to estimate number of salmon with ground-based stream survey, accounting also for estimate of uncertainty. Su et al. (2001) continued their work with hierarchical Bayesian approach enabling learning from years with more data to those with missing data. Furthermore, Sethi and Bradley (2016) introduced a Bayesian approach to estimate missing passage at weirs with run curve model to account for arrival dynamics and process variation model to describe the observed data. While these studies attempt to account for uncertainty, they

do not discuss their model assumptions against biological knowledge. This together with consideration of difference between the observation and model prediction as “noise” instead of proportion unobserved (as e.g. in Sethi and Bradley 2016), may cast some doubt on the biological meaningfulness of the results.

In our study, a Bayesian model framework was introduced for estimating the annual number of Atlantic salmon (*Salmo salar*) smolts passing the video monitoring site at a subarctic River Utsjoki in northern Finland. The framework consists of sub-processes of departing, travelling and observing. When combined, the processes of departing and travelling form the arrival distribution that describes the timing of the smolt run. Environmental covariates were used as a source of information in the sub-processes and expert knowledge is utilized in providing biologically meaningful model structures and prior distributions. Hierarchical structure was assumed over the study years making it possible to learn from the processes, borrowing strength from data rich years of study to those with missing data. Predictive studies illustrate the model performance in cases with missing data either at the beginning or in the middle of the run. We show that a biologically meaningful passage model can be built without assuming any specific shape for the arrival distribution. Furthermore, we claim that expert knowledge is a key source of information in estimating the total uncertainty in the annual total passage especially when environmental conditions have a great influence on the probability to observe individuals passing the survey site.

Materials and Methods

Data

River Utsjoki is a tributary of the large River Teno system at the northernmost border between Finland and Norway. Each spring, the monitoring system of 8 video cameras is set at the river bed under a bridge close to the tributary outlet (see Orell et al. 2011; Fig 1). The video footage provides

data on both smolts descending and adults ascending the river during the course of summer. In this study we analyze the smolt count data from years 2005-2009 and 2014. These years were chosen to demonstrate the highly variable environmental conditions, especially the strongly varying discharge (Figure 2). The smolt data are aggregated in daily counts of individuals over 61 days of June-July.

Data on environmental covariates contain air temperature and discharge measurements. These datasets are aggregated into daily averages. Both covariates are measured near the video site, but conditions may differ between the video site and upper parts of the river from where some of the smolts begin their migration. These datasets are, however, considered as reasonable proxies for the environmental conditions affecting the migration behavior of smolts.

Data from the first 23 days of June in 2005 are missing because the high level of water delayed the setup of the monitoring system in that year. Thus, the missing counts are predicted for those dates. Predictive studies were performed by treating data in 2007 and 2014 as partly missing: For 2007 the first 17% of the run strength and for 2014 the peak +-2 days count was removed from the analysis. To evaluate the predictive performance of the model, the model predicted counts for those dates were compared with the observed data.

Expert elicitation as a source of informative priors

Model structure and informative priors were elicited from an expert that is most familiar with the behavior of salmon smolts and the video monitoring system at Utsjoki. The expert was asked to base his views on his background knowledge about the processes instead of considering the particular data that will be included in the study. Elicitation was carried out iteratively, first by asking preliminary questions on, e.g. biological process in question, and then providing graphical illustration of the parameters. Prior distributions were adjusted until the expert agreed that the illustration is in line with his views. Some examples of questions elicited from the expert are listed in table 1.

Probability model of the problem

The model is constructed by thinking of four steps, described in detail in the following sections:

1) Process of departing, in which the smolts make the decision to depart while daily temperature affects the probability to depart in a given day;
2) Process of travelling, in which the smolt's travel time (in days) from the place of departure to the video site is modeled with daily discharge as a covariate;
3) Arrival distribution, which links together the processes of departing and travelling into a timing of the smolt run;
4) Observation process, in which the discharge influences visibility and the probability that an individual smolt passing the site is observed from the video footage.

Process of departing

An individual smolt's probability to begin the migration in day i of year y given that the smolt has not departed yet is considered to depend on the temperature on that day. We assume a logit-normally linear relationship for the smolt's probability to depart $(p_{i,y})$ and temperature $t_{i,y}$:

(1) $logit(p_{i,y}) \sim N(\alpha^P + \beta^P \, t_{i,y} \; , \; s^P)$

Informative prior distributions are given to parameters α^P, β^P and s^P according to expert view (see tables 1 and 2).

The daily probabilities of departing (given those individuals have not departed yet; $p_{i,y}$) can be transformed into expected proportions of the total smolt run that depart each day ($P_{i,y}$) as

(2) $P_{1,y} = p_{1,y}$

(3) $P_{i,y} = (1 - \sum_{k=1}^{i-1} p_{k,y}) \, p_{i,y}$

(4) $P_{61,y} = 1 - \sum_{i=1}^{60} p_{i,y}$.

Process of travelling

According to the expert view, each smolt should arrive at the video site within 14 days after departing. Thus we assume that the smolts that depart on day i pass the video site in one of the next 14 days according to discretized distribution function of lognormal distribution (Schwarz & Dempson 1994, Mäntyniemi & Romakkaniemi 2002):

$$(5)\ Q^D_{i,j,y} = \phi\left(\frac{\log(j-i+0.5-M_{i,y})}{S}\right) - I_{i,j}\phi\left(\frac{\log(j-i+0.5-M_{i,y})}{S}\right).$$

Here ϕ is the cumulative density function of the standard normal distribution, $M_{i,y}$ is the mean of log(travel time) to video site of smolts that depart on day i , and S is the standard deviation of log(travel time). Thus, $Q^D_{i,j,y}$ is the conditional probability that a smolt that departed on day i will arrive on day j, when j goes from i to $i + 13$.

To ensure that all smolts arrive within 14 days of departing, $Q^D_{i,j,y}$'s are standardized:

$$(6)\ q^D_{i,j,y} = Q^D_{i,j,y} / \sum_{j=i}^{(i+13)} (Q^D_{i,j,y}),$$

and thus $\sum_{j=i}^{(i+13)} (q^D_{i,j,y}) = 1$.

It is further considered that increasing discharge increases the speed at which the smolts travel, shortening their travel time. The expected travel time on real scale $\mu^D_{j,y} = \exp\,(M_{i,y} + 0.5\,S^2)$ is assumed to follow a lognormal distribution with expected value $\gamma_{i,y}$ and coefficient of variation cv^D and it is assumed to depend on the discharge ($f_{i,y}$):

$$(7)\ \gamma_{i,y} = \exp\,(\alpha^D + \beta^D \cdot f_{i,y}).$$

Again, informative prior distributions are given for parameters α^D, β^D and $cv^{\mu D}$ according to expert view (see tables 1 and 2).

Arrival distribution

By combining the processes of departing and travelling, we achieve joint probabilities that a smolt departs on day i and passes the video site on day j:

$$(8)\ q^J_{i,j,y} = q^D_{i,j,y}\, P_{i,y}.$$

The expected proportions of smolts (out of the total run size) passing the video site each day can then be found by marginalizing over the day of departure:

$$(9)\ \mu^N_{j,y} = \begin{cases} \sum q^J_{j,1:j,y}\ ,\ j \leq 14 \\ \sum q^J_{j,(i-13):j,y}\ ,j\ > 14. \end{cases}$$

Furthermore, we give a uniform prior distribution on log scale for the total annual size of the smolt run passing the video site:

$$(10)\qquad \log\,(N^{tot}_y) \sim Unif(7,15).$$

Parameterisation on the log-scale allows for a wide, minimally informative prior distribution for the annual total abundance. This annual run size is assumed to be Dirichlet-multinomially distributed over the 61 days of June-July. For computational simplicity, we follow Mäntyniemi et al. (2015) and approximate multinomial distribution as

$$(11)\qquad N_{j,y} = q^N_{j,y}\, N^{tot}_y,$$

where $q_{j,y}^{N}$ is the proportion of smolts passing the video site in day j and $q_{j,y}^{N}$'s follow a Dirichlet distribution parameterized with the expected proportions of smolts passing the video site each day $\mu_{j,y}^{N}$ and overdispersion η^{N}:

$$(12) \qquad q_{1:61,y}^{N} \sim Dirich(\mu_{1,y}^{N}\eta^{N}, \ldots, \mu_{61,y}^{N}\eta^{N}).$$

To ease the computation we approximate the Dirichlet-distribution with a set of lognormal distributions (see appendix A for details).

The annual proportional prior distributions for the the smolt run over 61 days of June-July are illustrated in figure 3. These priors are based on the model structure and data on environmental covariates (daily mean air temperature and discharge).

Observation process

In previous sections we have covered the processes assumed to affect the timing of the smolt passage at the video site. To estimate the total uncertainty in the smolt run size, the process of observing smolts from the video footage needs to be included. It seems natural to assume that some of the individuals pass the site unobserved, and thus this proportion must be accounted for. Next, we introduce two versions for the observation process:

The first version (model M1) is simple yet not very realistic. It assumes that the number of smolts observed at the video site on day i in year y ($N_{i,y}^{Obs}$) follows a Beta-binomial distribution:

$$(13) \qquad N_{j,y}^{Obs} \sim Beta - bin(100,10, N_{i,y})$$

Here $N_{j,y}$ is the total daily number of smolts passing the video site and numerical constants indicate that the probability that an individual smolt is observed varies between days according to $Beta(100,10)$ distribution with fixed mean 0.91 and standard deviation 0.03. Thus, 9% of smolts are assumed to pass the video site unobserved on average, with 95% PI of [5%, 14%]. Albeit an observation model with randomly varying probability for observing is not realistic, such a model can be useful for illustrating how the processes of departing and travelling affect the model results when considering assuming close to perfect process of observing.

The second version (model M2) for an observation process aims at higher realism, accounting for expert views on how high the probability to observe a smolt may be in excellent vs. poor circumstances and how the discharge would affect this probability. Again the number of smolts observed follows a Beta-binomial distribution:

(14) $$N_{j,y}^{Obs} \sim Beta - bin(\mu_{j,y}^{B}\,\eta^{B},\ (1-\mu_{j,y}^{B})\eta^{B},\ N_{j,y})$$

Now the probability to observe a smolt follows a $Beta(\mu_{j,y}^{B}\,\eta^{B},\ (1-\mu_{j,y}^{B})\eta^{B})$ distribution, where $\mu_{j,y}^{B}$ is the expected probability for a smolt to be observed and η^{B} is the overdispersion parameter. Overdispersion is thought to arise from the schooling behavior of smolts. Groups of smolts can be either detected or missed, which increases the variance of the observation process (Mäntyniemi & Romakkaniemi 2002, Linden & Mäntyniemi 2011, Mäntyniemi et al. 2015). The expected value of the process is linked to discharge following the expert view that in very good visibility (low discharge) at maximum 90% of the smolts are observed. As the discharge increases, the visibility decreases, and the volume of water to be covered by the video monitoring increases. Therefore, according to the expert, observation probability decreases gradually to a minimum of 30%. The expected probability $\mu_{i,y}^{B}$ is considered to follow logit-normally linear relationship on the interval [0.3,0.9]:

(15) $$\mu_{j,y}^{B} = 0.6\left(\frac{\exp(BB_{j,y})}{1+\exp(BB_{j,y})}\right) + 0.3$$

(16) $$BB_{j,y} \sim N\left(\alpha^{B} - \beta^{B} f_{j,y},\ sd_{BB}^{2}\right),$$

where α^B, β^B and sd^2_{BB} have prior distributions according to the expert view (Tables 1 and 2).

Models M1 and M2 were fitted using MCMC sampling with JAGS 4.2.0 software (Just Another Gibbs Sampler; Plummer 2003) and R runjags package (Denwood 2016). Both models were run with 2 chains for 1 500 000 iterations this taking around 3 days per model. Convergence was diagnosed using Gelman-Rubin statistics (Gelman and Rubin 1992). The complete JAGS code for the model structure can be found from appendix B.

Results

When compared with the observed smolt counts, annual posterior estimates of daily run size (based on M2) indicate that 25-75% of smolts pass the video site unobserved (Figure 4). Considering the estimated uncertainty among the years without missing data, the proportion of unobserved is 90% at most (in 2008 upper limit for 90% PI is 16000 smolts vs. observed data of 8400 smolts). Posterior estimates of daily arrivals are illustrated in figures 5 (M1) and 6 (M2).

Table 3 shows the posteriors for the time periods with missing counts as well as the observed annual counts for comparison when available. M2 predicts 24-33% higher number of smolts compared to M1 as a result of more realistic observation model. Scenarios to study predictive performance of the model contain 'first 17% of missing' -scenario for 2007 and 'peak day +- 2 days missing' –scenario for 2014. In both cases, all daily counts fall in the predicted 90% PI indicating good predictive performance (Table 3). The percentage of predicted out of the estimated annual total is close to identical between models in all 3 cases (Table 3).

Figures 7 and 8 illustrate the prior and posterior distributions from M2 for expected processes of departing and travelling. Posterior distributions from M1 are nearly identical and thus not illustrated here. The expected probability to begin the migration at given temperature updates quite heavily from the prior distributions, supporting 4-5 degrees higher temperature at which the migration sets off (Figure 7). The posterior cumulative distribution of the expected travel time to the video site

(Figure 8) supports somewhat longer travel times than expected a priori both for low ($10m^3/s$) and high ($100m^3/s$) discharge. Note that posterior cumulative distributions at these specific levels of discharge are chosen for illustrative purposes only and that in practice the models allow any positive value for the discharge.

In M2, also the observation probability depends on the discharge. Figure 9 illustrates the prior and posterior distributions of the observation process given the discharge. Priors and posteriors are close the same, indicating that in this case the combination of priors and data did not result in further learning about the process of observing.

Discussion

In our study, an arrival model for Atlantic salmon smolts was designed by separating the processes of departing, travelling and observing. The processes of departing and travelling together form the arrival distribution, i.e. the timing of the smolt run. This modular structure enabled us to concentrate on biologically meaningful mathematical assumptions related to each sub-process instead of assuming that fish count data as such follows a specific distributional shape. The model framework aimed at high biological realism by accounting for expert knowledge in the model formulation, firstly by including choices for model structures and prior distributions, and secondly by linking environmental covariates to provide information about the timing of the migration. The estimated total uncertainty accounted for the uncertainty in the sub-processes, including the observation process in which variability in discharge is known to have a large impact on the probability to observe passing smolts.

One of the most important advantages of our approach is that the need to assume a specific parametric shape for the arrival distribution is avoided. Past studies (Hilborn et al. 1999, Su et al. 2001, Sethi and Bradley 2016) have assumed unimodal shape (e.g. normal, skew-normal, Student's t) for the arrival curve. However, when taking a look at the shape of (any) arrival time series it seems

clear that such assumption is a tolerable simplification at best (see e.g. figure 5 in this paper, figure 2 in Sethi and Bradley 2016).

Earlier studies have paid little attention to the role of observation models in estimating the total uncertainty. Hillborn et al. (1999) distinguished process and observation errors, assuming observer efficiency either as known or unchanging inside the study period. Su et al. (2001) discussed the issue of error in the observation process and mainly concluded that such error can cause convergence problems in sampling. Sethi and Bradly (2016) incorporated uncertainty only on the estimates of missing passage counts, treating the observed counts as true total number of passing without uncertainty. As they assumed normal or negative Binomial models for the observed passage counts, they in fact allowed model predicted passage counts to be either greater or smaller than the observed counts. This in turn can lead to underestimation of the missing counts. We claim that it is necessary to consider the observed count as a minimum estimate for the true number especially if the back-and-forth movement of individuals is negligible. Along with this notion the need for a realistic observation process becomes obvious. Depending on the approach, the estimate of total annual passage can be very different. As Hilborn et al. (1999) put it, 'when we admit uncertainty in observer efficiency we become much less certain about the actual escapement'.

Sethi and Bradley (2016) fitted datasets in their study using several combinations of arrival and process error models. They utilized information-theoretic model selection (DIC, Spiegelhalter et al. 2002) to find out the best fitting model for each dataset. However, no discussion took place whether some of these models would be biologically more reasonable than another. We argue that the model selection should primarily be based on biological realism and in case biological model uncertainty is needed to be tackled, Bayesian model averaging (BMA, Hoeting et al. 1999) should be preferred over DIC or other information criteria. The purpose of an information criterion is to select a statistical model that would be able to replicate the observed data with least amount of

parameters. Such a goal does not necessitate an ecological model: generic machine learning models are probably going to perform much better in this regard.

Su et al. (2001) used hierarchical structure in their study, assuming that years in which timing of the run is well known can be used for passing information to years with missing data. We argue that learning should rather take place over the (biological) process, assuming exchangeability over year specific parameters of those processes. Our model structure can also be considered as hierarchical, although we simply assume that the underlying processes are the same from year to year allowing between- and within-years variation. If there were a need for a meta-analysis of similar studies over several stocks of the same species, exchangeability in biological parameters could be considered between the stocks. Such approach could also allow meta-analysis of data from different types of monitoring systems if detailed observation processes were tailored for each system and exchangeability was considered as reasonable assumption for the biological processes.

Our model framework could be fine-tuned and developed further. In an ideal situation, environmental covariates would be available from upstream locations from where the smolts originate. Data on air temperature would be replaced with water temperature to better explain the physiological processes the smolts encounter. To increase the realism even more, the schooling behavior of smolts could be accounted for in more detail than done here. Salmon smolts have a tendency to school during their migration (e.g. Bakshtanskiy et al. 1988, Davidsen et al. 2005) and this creates more variation in the monitoring data compared to independent movement (Mäntyniemi and Romakkaniemi 2002, Mäntyniemi et al. 2015). Schooling behaviour could also be linked with the process of travelling as the passing schools of fish have been suggested to stimulate migratory behavior in the remaining smolts (e.g. Hansen and Jonsson, 1985). Schooling also affects the observation process as, on one hand, schools are easier to spot than individual smolts, and on the other hand, it can be difficult to have an exact count of individuals from a large school. For

Utsjoki video survey, daily average school sizes of smolts are recorded and incorporating those into the model framework is an interesting topic for future research.

Despite the above possibilities for further development of the modeling framework, we claim that our model already contains the elements of realism that are required to provide credible estimates of the total annual smolt run. This approach can be applied to similar monitoring surveys where point-checking the numbers of migrating fish (Trépanier et al. 1996), by e.g. sonar, trap, or weir is used for estimating run sizes. As far as we know, environmental covariates have not been included so far to estimate both the dynamics of passage and the observation process. Our model framework also indicated a good predictive performance for missing dates both at the beginning and in the middle of the run.

Acknowledgements

We thank the numerous field workers enabling the yearly monitoring programme and smolt data collection at the River Utsjoki. We also acknowledge the crucial technical knowhow by Jorma Kuusela in running the annual monitoring and the instrumental role of Dr. Martin Svenning and Anders Lamberg in early phases of the River Utsjoki monitoring programme. Funding for this study was provided by the Academy of Finland (project No. 286334).

References

Bakshtanskiy, E., Nesterov, V., and Neklyodov, M. 1988. Development of schooling behaviour in juvenile Atlantic salmon, Salmo salar, during seaward migration. J. Ichthyol. 28: 91–101.

Davidsen, J., Svenning, M. A., Orell, P., Yoccoz, N., Dempson, J. B., Niemelä, E., Klemetsen, A., Lamberg, A., and Erkinaro, J. 2005. Spatial and temporal migration of wild Atlantic salmon smolts determined from a video camera array in the sub-Arctic River Tana. Fisheries Research 74: 210-222.

Dempson, J. B, and Stansbury, D. E. 1991. Using partial counting fences and atwo-sample stratified design for mark–recapture estimation of an Atlantic salmonsmolt population. N. Am. J. Fish. Manage. 11: 27–37.

Denwood, M.J. 2016. runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. J. Stat. Softw. 71(9): 1-25. doi:10.18637/jss.v071.i09.

Erkinaro, J., Julkunen, M. and Niemelä, E. 1998. Migration of juvenile Atlantic salmon Salmo salar in small tributaries of the subarctic River Teno, northern Finland. Aquaculture 168: 105-119.

Gelman, A., and Rubin, D.B. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7: 457–511.

Hilborn, R., Bue, B.G., and Sharr, S. 1999. Estimating spawning escapements from periodic counts: a comparison of methods. Can. J. Fish. Aquat. Sci. 56(5): 888–896. doi:10.1139/f99-013.

Hansen, L.P., and Jonsson, B. 1985. Downstream migration of hatchery-reared smolts of Atlantic salmon (Salmo salar L.) in the River Imsa. Aquaculture, 45: 237–248.

Holmes, J.A, Cronkite, G.M.W., Enzenhofer, H.J., and Mulligan, T.J. 2006. Accuracy and precision of fish-count data from a “dual-frequency identification sonar” (DIDSON) imaging system. ICES J. Mar. Sci. 63: 543–555. doi:10.1016/j.icesjms.2005.08.015

Kuparinen, A., Mäntyniemi, S., Hutchings, J.A., and Kuikka, S. 2012. Increasing biological realism of fisheries stock assessment: towards hierarchical Bayesian methods. Environ. Rev. 20: 135–151. doi:10.1139/A2012-006.

Mäntyniemi, S., and Romakkaniemi, A. 2002. Bayesian mark-recapture estimation with an application to a salmonid smolt population. Can. J. Fish. Aquat. Sci. 59: 1748-1758. doi:10.1139/F02-146.

Mäntyniemi, S.H.P., Whitlock, R.E., Perälä, T.A., Blomstedt, P.A., Vanhatalo, J.P., Rincon, M.M., Kuparinen, A.K., Pulkkinen, H.P., and Kuikka, O.S. 2015. General state-space population dynamics model for Bayesian stock assessment. ICES J. Mar. Sci. 72: 2209-2222. doi:10.1093/icesjms/fsv117

Orell, P., Erkinaro, J., Svenning, M., Davidsen, J., and Niemelä, E. 2007. Synchrony in the downstream migration of smolts and upstream migration of adult Atlantic salmon in the sub-Arctic River Utsjoki. J. Fish. Biol. 71: 1735–1750.

Orell, P., Erkinaro, J., and Karppinen, P. 2011. Accuracy of snorkelling counts in assessing spawning stock of Atlantic salmon, Salmo salar, as verified by radio-tagged fish and underwater video monitoring. Fisheries. Manag. Ecol. 18: 392–399.

Otero, J., L'Abée-Lund, J.H., Castro-Santos, T., Leonardsson, K., Storvik, G.O., Jonsson, B., Dempson, B., Russell, I.C., Jensen, A.J., Baglinière, J.-L., Dionne, M., Armstrong, J.D., Romakkaniemi, A., Letcher,

B.H., Kocik, J.F., Erkinaro, J., Poole, R., Rogan, G., Lundqvist, H., MacLean, J.C., Jokikokko, E., Arnekleiv, J.V., Kennedy, R.J., Niemelä, E., Caballero, P., Music, P.A., Antonsson, T., Gudjonsson, S., Veselov, A.E. Lamberg, A., Groom, S., Taylor, B.H., Taberner, M., Dillane, M., Arnason, F., Horton, G., Hvidsten, N.A., Jonsson, I.R., Jonsson, N., McKelvey, S., Næsje, T., Skaala, Ø., Smith, G.W., Sægrov, H., Stenseth, N.C. and Vøllestad, L.A. 2014. Basin-scale phenology and effects of climate variability on global timing of initial seaward migration of Atlantic salmon (Salmo salar). Glob. Change. Biol. 20: 61-75.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In* proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), 20-22 March, Vienna, Austria. ISSN 1609-395X.

Romakkaniemi, A., Lilja, J., Nykänen, M., Marjomäki, T.J. and Jurvelius, J. 2000. Spawning run of Atlantic Salmon (Salmo salar) in the River Tornionjoki monitored by horizontal split-beam echosounding. Aquat. Living. Resour. 13: 349-354.

Schwarz, C., and Dempson, J. 1994. Mark-recapture estimation of a salmon smolt population. Biometrics. 50: 98-108.

Sethi, A.S., and Bradley, C. 2016. Statistical arrival models to estimate missed passage counts at fish weirs. Can. J. Fish. Aquat. Sci. 73: 1251–1260. doi:10.1139/cjfas-2015-0318.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. 2002. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. Ser. B. 64: 583–639.

Su, Z., Adkison, M.D., and Van Alen, B.W. 2001. A hierarchical Bayesian model for estimating historical salmon escapement and escapement timing. Can. J. Fish. Aquat. Sci. 58: 1648–1662. doi:10.1139/f01-099.

Trépanier, S., Rodríquez, M. A. and Magnan, P. 1996. Spawning migrations in landlocked Atlantic salmon: time series modelling of river discharge and water temperature effects. J. Fish Biol. 48: 925–936.

423 Table 1. Examples of questions elicited from the expert

Question	Expert view
Process of departing	
- At what range of temperature (air temperature at Utsjoki logger) first smolts are considered to begin their migration?	6-8 °C
- What level of temperature would be considered so high that the process of migration not be influenced by further increase in temperature (i.e. logit curve levels out)?	12 °C
Process of travelling	
- What is the the maximum travel time (in days) in which a smolt can be considered to arrive at the video site?	14 days
- What is the minimum travel time in which a smolt can be considered to arrive at the video site?	1-2 days
Process of observing	
- What is the upper (/lower) limit for the expected probability at which an average smolt is considered to be observed in best (/worst) possible environmental conditions, i.e. at minimum (/maximum) discharge?	90% (30%)
- What is the level of discharge at which the visibility can be considered optimal?	20 m^3/s
- What is the level of discharge at which the visibility is considered to be poorest?	50-60 m^3/s

424

425

426

Table 2: List of symbols, descriptions and prior distributions

Symbol	Description	Prior distribution
$logN_y^{tot}$	Total annual run size on log-scale	Uniform(7,15)
N_y^{tot}	Total annual run size on real scale	-
$N_{j,y}$	Daily number of smolts passing the video site	-
$q_{j,y}^N$	Daily proportion of smolts passing the video site	-
$\mu_{j,y}^N$	Expected proportion of smolts passing the video site	-
η^N	Overdispersion of the arrival process	Uniform(0.001,100000)
$p_{i,y}$	Probability to depart (if haven't departed yet)	-
$P_{i,y}$	Proportion of smolts departing each day	-
α^P	Intercept for probability to depart given temperature	N(-20,1)
β^P	Regression coefficient for probability to depart given temperature	logN(0.6,0.1)
s^P	Standard deviation of probability to depart	logN(0,1)
$Q_{i,j,y}^D$	Unstandardised proportion for travel time	-
$q_{i,j,y}^D$	Standardised proportion for travel time	-
$M_{i,y}$	Mean of log(travel time)	-
S	Standard deviation of log(travel time)	
$\mu_{i,y}^D$	Expected travel time	-
$\gamma_{i,y}$	Expected value of expected travel time	-
cv^D	Coefficient of variation of log(travel time)	Uniform(0.001,2)
α^D	Intercept for the expected value of expected travel time given flow	logN(0.52,0.07)
β^D	Regression coefficient for the expected value of expected	logN(-4.6,0.04)

	travel time given flow	
$cv^{\mu D}$	Coefficient of variation of expected travel time	Uniform(0.001,1)
$q^{J}_{i,j,y}$	Probability that a smolt departs on day i and passes the video site on day j	-
$\mu^{B}_{j,y}$	Expected probability to observe a smolt (M2)	-
η^{B}	Overdispersion of the observation process	Uniform(5,1000)
α^{B}	Intercept for probability to be observed	N(2.9,0.02)
β^{B}	Regression coefficient for probability to be observed	logN(-2.6,0.001)
sd^{2}_{BB}	Standard deviation of expected probability to be observed	logN(-0.23, 0.005)
Data		
$t_{i,y}$	Mean daily air temperature	-
$f_{j,y}$	Mean daily discharge	-
$N^{Obs}_{j,y}$	Daily smolt count	-

Table 3. Results from the predictive study in 2005 (real missing data), 2007 ('first 17% missed' -scenario) and 2014 ('peak of the migration +-2 days missing' -scenario). For posterior distributions, median values and 90% probability intervals are illustrated.

Year	Sum of missing counts	% of missing counts out of total observed	model M1		model M2	
			Number of predicted	% of predicted out of total	Number of predicted	% of predicted out of total
2005	NA	NA	1620 (340,3930)	10% (2%,22%)	2020 (450,4850)	10% (3%,22%)
2007	2539	17%	2500 (510,6220)	14% (4%,31%)	3220 (700,7990)	15% (4%,32%)
2014	4085	21%	3790 (2030,6040)	18% (11%,26%)	4700 (2620,7390)	18% (11%,26%)

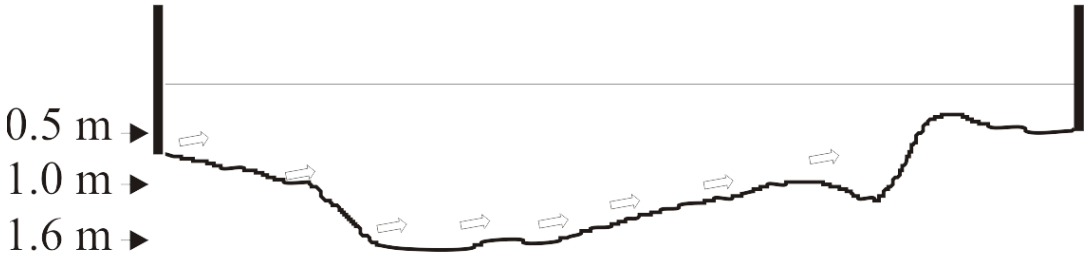

Figure 1: Setup of the video monitoring system at the river bed in Utsjoki. Arrows illustrate the position and direction of each of the 8 cameras.

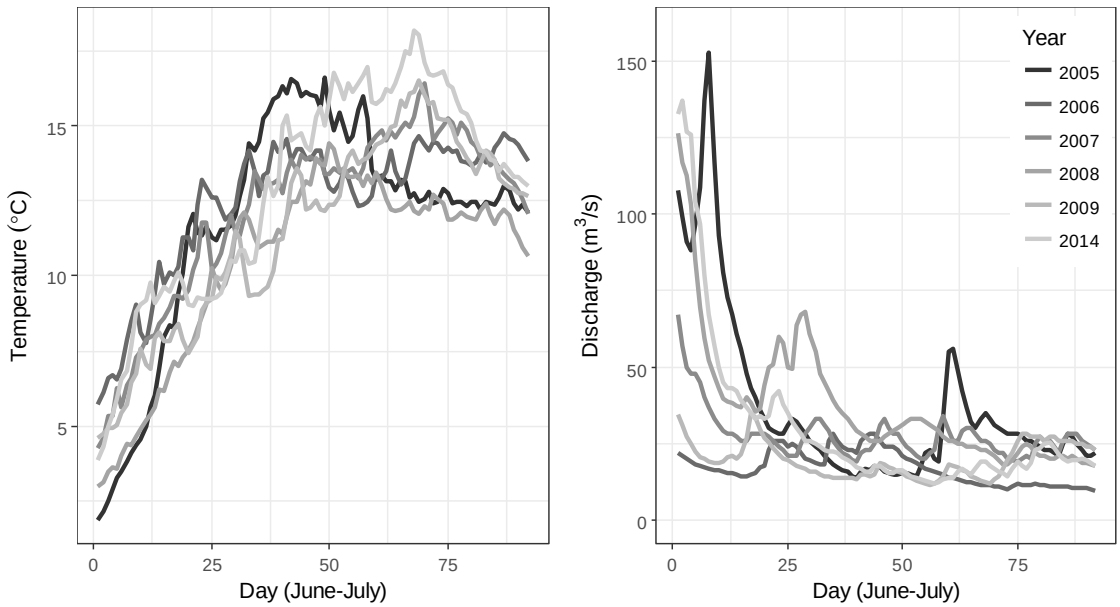

Figure 2. Data on daily mean air temperature and mean discharge at Utsjoki on study years 2005-2009 and 2014.

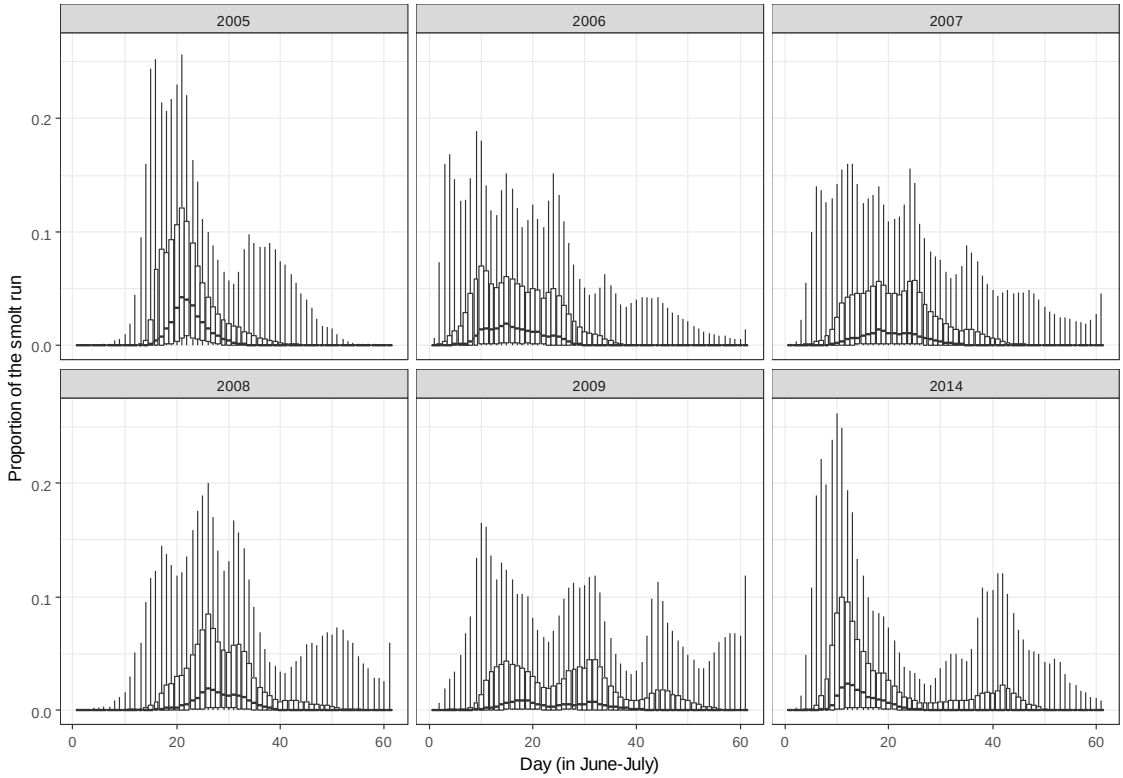

Figure 3. Annual prior distributions for the timing of the smolt run given the environmental covariates (daily mean air temperature and discharge).

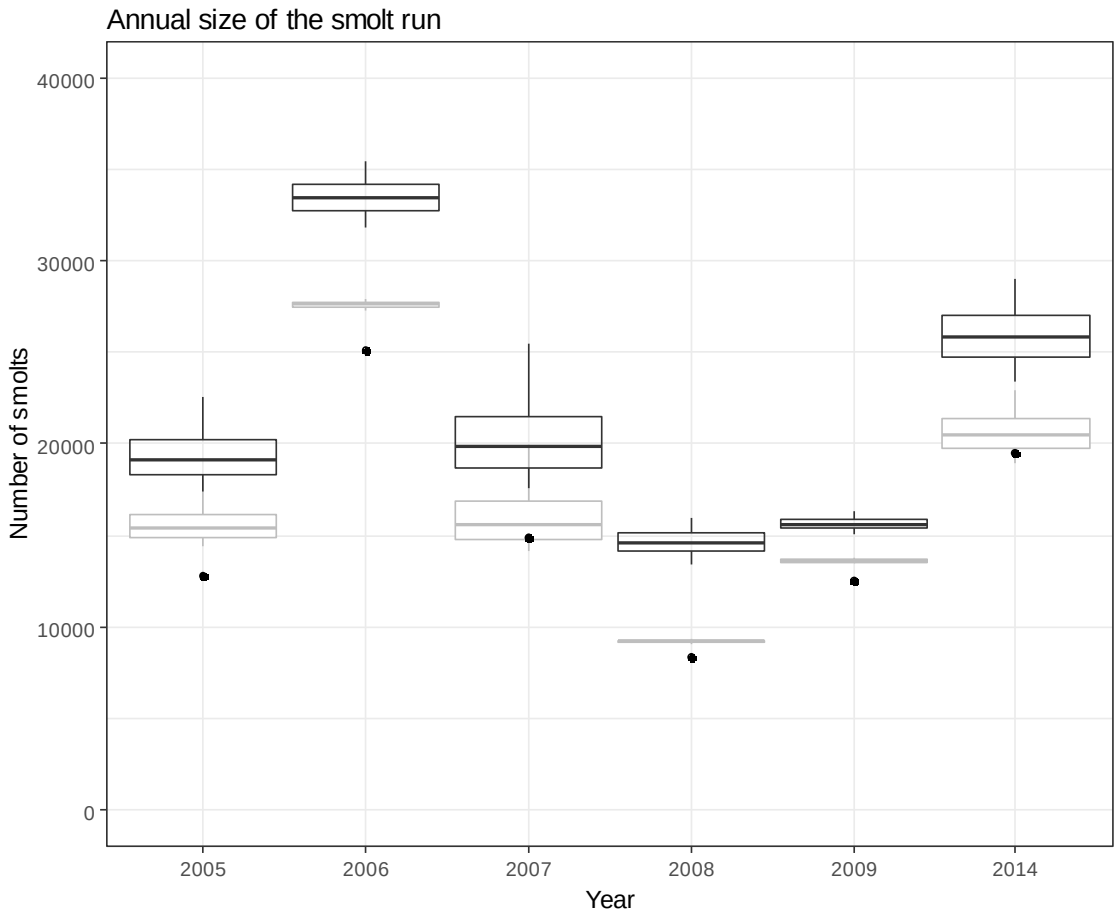

Figure 4. Annual number of smolts passing the video site. Posterior distributions of M1 (grey boxplots), M2 (black boxplots) and count data (black dots).

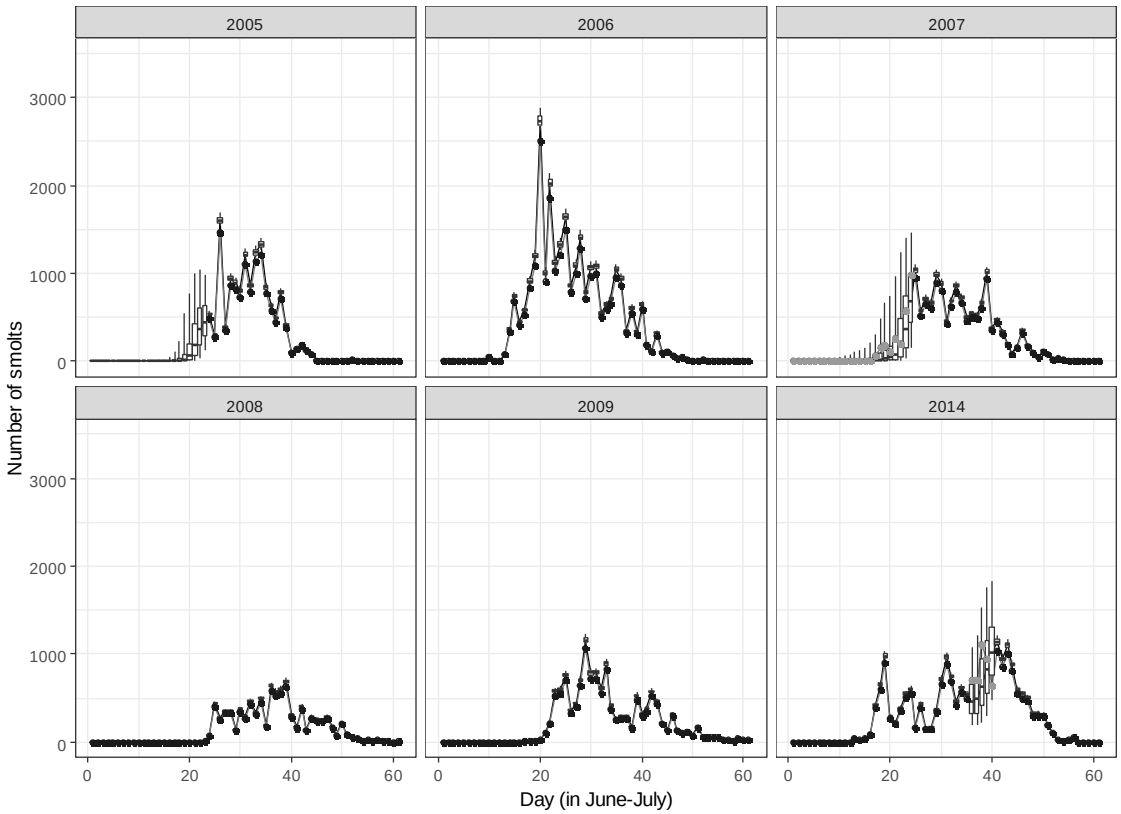

Figure 5. Daily number of smolts that pass the video site in 2005-2009 and 2014. Posterior distributions of M1 (boxplots) and count data (dots). Data points included in the study are marked with black and the ones excluded (counts treated as missing) marked with grey.

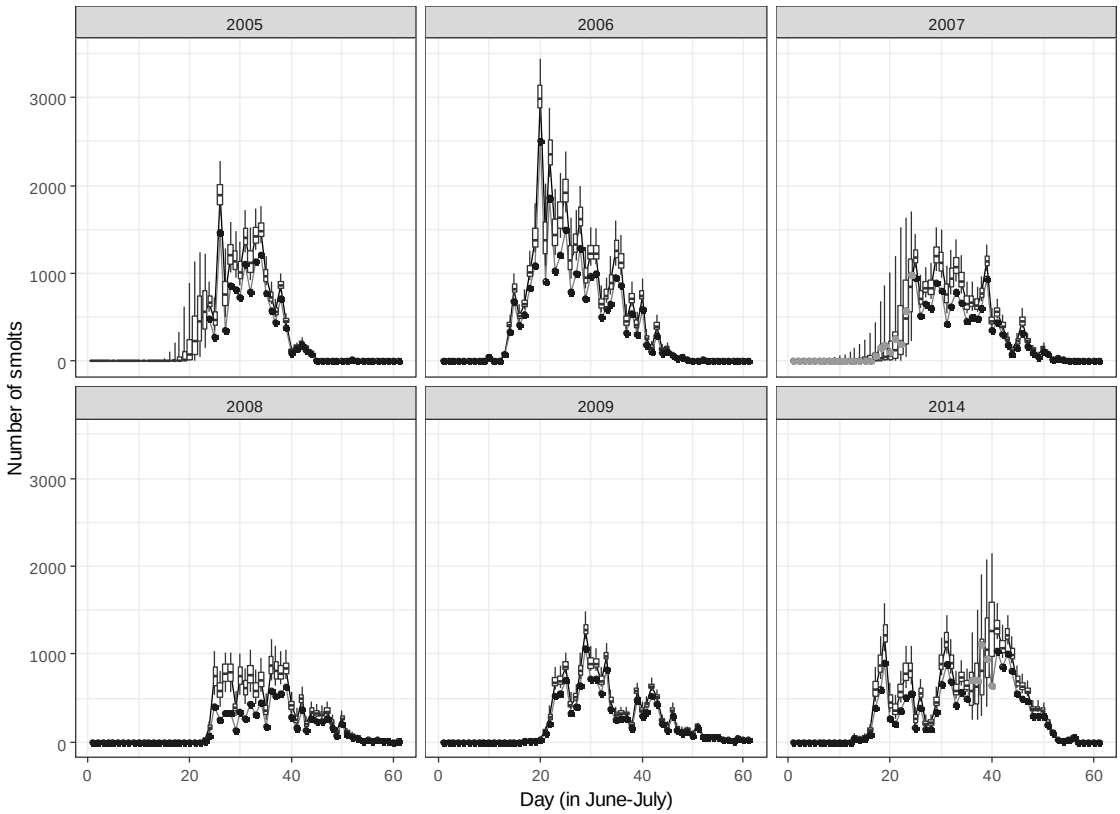

Figure 6. Daily number of smolts that pass the video site in 2005-2009 and 2014. Posterior distributions of M2 (boxplots) and count data (dots). Data points included in the study are marked with black and the ones excluded (counts treated as missing) marked with grey.

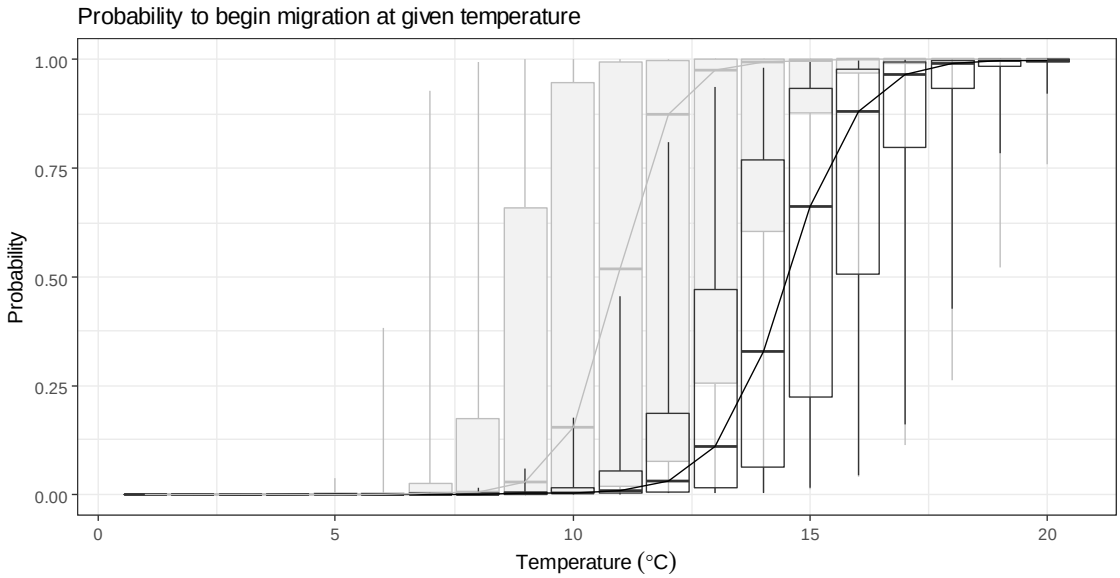

Figure 7. Probability that a smolt begins the migration at given temperature. Prior (grey boxplots) and posterior distributions (black boxplots) from M2.

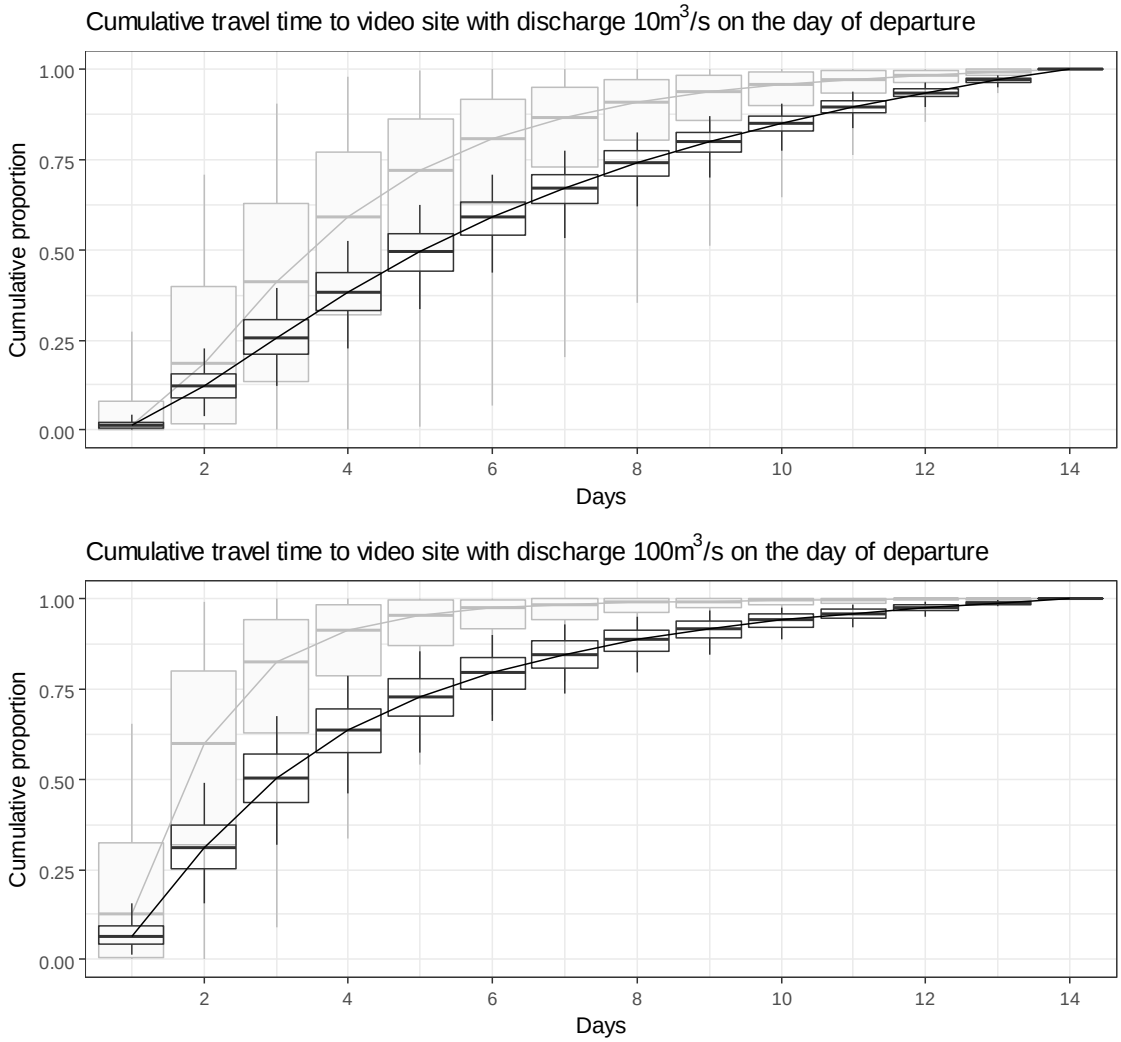

Figure 8. Cumulative distribution of expected travel time in days from the point of departure to the video site. Above: Cumulative distribution at low discharge (10 m^3/s). Below: Cumulative distribution at high discharge (100 m^3/s). Prior (grey boxplots) and posterior distributions (black boxplots) from M2.

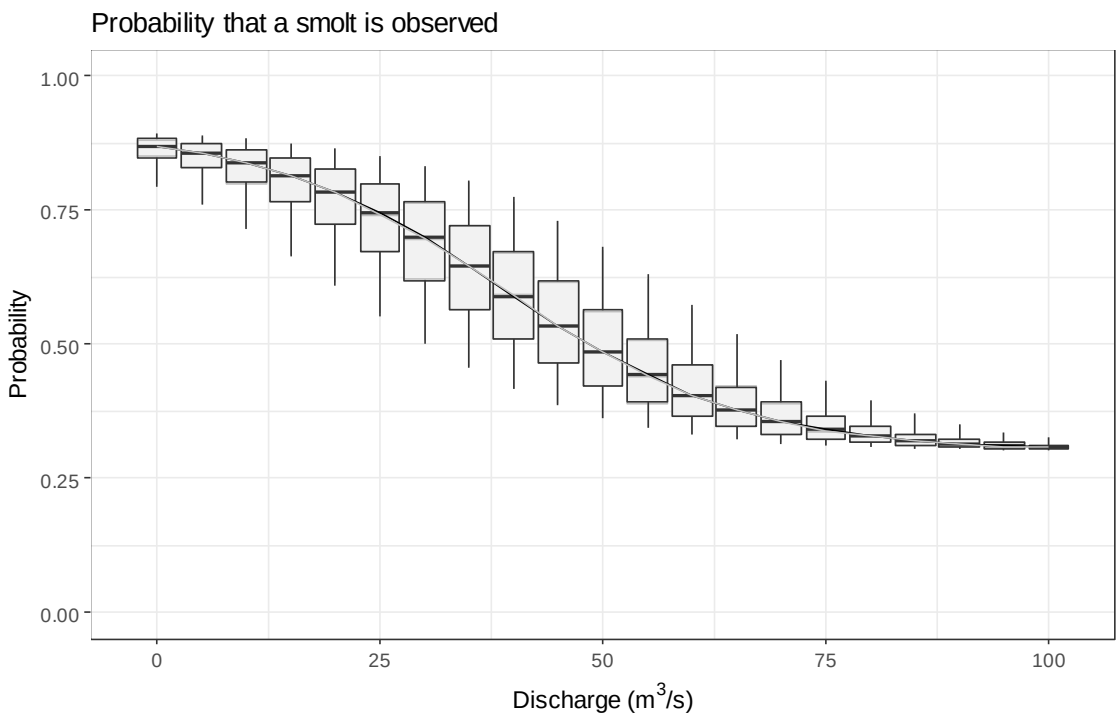

Figure 9. The expected observation probability at given discharge. Prior (grey boxplots) and posterior distributions (black boxplots) from M2.

Appendix A: Lognormal approximation for Dirichlet-distribution

Let's consider a set of Dirichlet-distributed parameters:

(A.1) $$q_{1:n} \sim Dirich(\mu_1\eta, \dots, \mu_n\eta),$$

Where μ_i is the expected probability for slot i and η is the overdispersion parameter. Such Dirichlet-distribution can be approximated with a set of lognormally distributed z_i by assuming

(A.2) $$q_{1:n} = z_i / \sum_i z_i$$

(A.3) $$z_i \sim logN(M_i, T_i)$$

(A.4) $$M_i = \log(\mu_i) - \frac{0.5}{T_i}$$ and

(A.5) $$T_i = \log\left(\frac{1}{\mu_i \eta} + 1\right).$$

Appendix B: JAGS code for the arrival model

```
model{

 # Process of departing
 # ==========================
 for(y in 1:nYears){
   for(i in 1:nDays){
     # p: probability to start migration at day t, if haven't done so earlier
     # t: daily mean temperature
     logit(p[i,y])<-p2[i,y]
     p2[i,y]~dnorm(aP+bP*t[i,y],1/pow(sdP,2))
   }
 }
 # P: expected proportion of smolts departing at day i
 for(y in 1:nYears){
   P[1,y]<-p[1,y]
   for(i in 2:(nDays-1)){
     P[i,y]<-(1-sum(P[1:(i-1),y]))*p[i,y]
```

```
    }
    P[nDays,y]<-1-sum(P[1:60,y])
  }

  aP~dnorm(-20,1)
  bP~dlnorm(0.6,10)
  sdP~dlnorm(0,1)

  # Process of travelling
  # ==============
  # qD: probability that a smolt that departs at day i passes the video site at day j
  for(y in 1:nYears){
    for(i in 1:nDays){
      # i: day of departure
      # j: day of passing the video site
      # f: daily mean discharge
      # j==i
      qDx[i,i,y]<-phi((log(0.5)-MD[i,y])/SD)
      # j>i
      for(j in (i+1):(i+13)){
        qDx[i,j,y]<-phi((log(j-i+0.5)-MD[i,y])/SD)-phi((log(j-i-0.5)-MD[i,y])/SD)
      }
      for(j in i:(i+13)){
        qD[i,j,y]<-qDx[i,j,y]/(sum(qDx[i,i:(i+13),y])+0.0001)
      }
      MD[i,y]<-log(muD[i,y])-0.5/TD
      muD[i,y]~dlnorm(log(exp(aD-bD*f[i,y]))-0.5/TmuD, TmuD)
    }
  }
```

```
SD<-1/sqrt(TD)
TmuD<-1/log(cvmuD*cvmuD+1)
TD<-1/log(cvD*cvD+1)

aD~dlnorm(0.52,14)
bD~dlnorm(-4.6,25)
cvmuD~dunif(0.001,1)
cvD~dunif(0.001,2)

# Arrival distribution
# ==========================================
# Joint distribution of qD and P: prob that a smolt departing day i arrives day j
for(y in 1:nYears){
  for(i in 1:nDays){ # day of departure
    for(j in i:(i+13)){ # day of passing the video site
      qJ[j,i,y]<-qD[i,j,y]*P[i,y]
    }
  }

  # Expected proportion of smolts passing the video site each day
  for(j in 1:14){
    muqN[j,y]<-sum(qJ[j,1:j,y])+0.0001 # small constant added for numerical stability
  }
  for(j in 15:nDays){
    muqN[j,y]<-sum(qJ[j,(j-13):j,y])+0.0001
  }
}

for(y in 1:nYears){
```

```
    Ntot[y]<-exp(LNtot[y]) # total run size in year y
    LNtot[y]~dunif(7,15)

    for(j in 1:(nDays-1)){
      N[j,y]<-round(qN[j,y]*Ntot[y]) # daily number of fish passing
    }
    N[nDays,y]<-round(Ntot[y]*(1-sum(qN[1:(nDays-1),y])))

    # qN: daily proportion of smolts
    # dirichlet-distribution approximated with lognormal
    qN[1:nDays,y]<-zN[1:nDays,y]/sum(zN[1:nDays,y])

    for(i in 1:nDays){
      zN[j,y]~dlnorm(MN[j,y], TauN[j,y])
    }

    alphaN[1:nDays,y]<-muqN[1:nDays,y]*eta_alphaN
    MN[1:nDays,y]<-log(muqN[1:nDays,y])-0.5/TauN[1:nDays,y]
    TauN[1:nDays,y]<-1/log((1/alphaN[1:nDays,y])+1)
  }

  eta_alphaN~dunif(0.001,100000)

  # Observation process
  # ====================
  for(y in 1:nYears){
    for(j in 1:nDays){
      # Nobs: Observed number of smolts
      Nobs[j,y]~dbetabin(100,10,N[j,y])  # M1
```

```
      Nobs[j,y]~dbetabin(muB[j,y]*etaB,(1-muB[j,y])*etaB,N[j,y])  #M2

      muB[j,y]<-0.6*(exp(BB[j,y])/(1+exp(BB[j,y])))+0.3
      BB[j,y]~dnorm(aB-bB*f[j,y],1/pow(sdBB,2))

    }
  }
  # priors for observation process M2
  aB~dnorm(2.9,60)
  bB~dlnorm(-2.6,984)
  sdBB~dlnorm(-0.23,210)
  etaB~dunif(5,1000)
}
```