Bayesian arrival model for passage counts powered by environmental covariates and expert knowledge

##### Henni Pulkkinen, Panu Orell, Samu Mäntyniemi

# Abstract

Not more than 175 words

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

# Introduction

An introduction generally need not exceed 375–500 words.

Migratory fish species are often monitored along their migration routes to collect fishery independent data for the needs of stock assessment. Monitoring surveys can be various (e.g. ground based, weirs, traps, video or sonar count) and common to all is they rarely provide perfect information about the number of individuals passing the system. Part of the run may be missed because of difficult environmental conditions or device failures, there may be double counting or the data achieved may be some other way partial or biased. Thus, the assumptions made when the data are interpreted can have a great influence on the estimated stock abundance.

Increasing the information beyond the 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. Moreover, methods of Bayesian inference enable both estimation of uncertainty and combination of various sources of information. Often statistical methods are used, however, without considering the biological realism and just focusing on the data analysis. Such procedure easily results in making awkward model assumptions that cannot be biologically interpreted and the potential effects of those decisions may pass undiscussed.

During the course of time, various methods have been used to estimate run dynamics and passage counts. 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 a passage observation before and after missing datum and thus missing tails cannot be estimated. By assuming constant proportion of run passing on a given date enables estimation of missing tails using expectation-maximization algorithms (Van Alen 2000), but considering the variability, for example, in environmental conditions, such assumption is problematic.

Hilborn et al. (1999) implemented maximum likelihood method for estimating number of salmon with ground-based stream survey, accounting also for estimate of uncertainty. Su et al. (2001) extended the method with hierarchical Bayesian approach enabling learning from years with more data to those with missing data. Furthermore, Sethi and Bradley (2016) introduced 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 account for uncertainty, they do not justify or 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, casts doubt on the meaningfulness of the estimated uncertainty.

In this study, we introduce a model framework that is based on biological theory. The framework estimates annual number of Atlantic salmon smolts (*Salmo salar*) passing the video monitoring site in river Utsjoki. We assume no specific shape for the arrival distribution. The model assumptions focus on the underlying (biological) processes that are mathematically constructed utilizing expert knowledge. Environmental covariates have an essential role in the model formulation, describing how altering conditions in nature affect the key processes. - Bayesian structure is assumed over the study years making it possible to learn from the processes and borrowing strength from data rich datasets to those with missing data. The model is built in pieces considering the elements of:

1. Smolts making the decision to depart while daily temperature affects the probability to depart in a given day;
2. Time it takes in days for an average smolt to arrive at the video site after migration decision has been made, daily flow velocity affecting the travel time;
3. Observation process in which flow (/level of water) influences visibility and the probability that an individual smolt passing the site is observed;

# Materials and Methods

Limit the information on materials and methods to what is needed in judging whether the findings are valid. To facilitate assessment, give all the information in one section when possible. Refer to the literature concerning descriptions of equipment or techniques already published, detailing only adaptations. Often, it helps to begin statements on procedures with a phrase indicating the purpose, such as “To determine … we …”. If the section is long, consider using subheadings corresponding to headings for the findings.

Data

River Utsjoki is a tributary of river Teno in the northernmost border of Finland and Norway. Each spring, the monitoring system of 8 video cameras is set at the river bed under a bridge in Utsjoki village (Fig 1). The video footage provides data on both smolts decending and adults ascending the river during the course of summer. In this study we analyze data from years 2005, 2006, 2008 and 2014. These years were chosen for illustration since environmental conditions, especially flow velocity, had large variation in those years (Figure 2). The data is aggregated in daily counts over 61 days of June-July. In 2005, data from the first 23 days of June was missed because the high level of water prevented the setup of the monitoring system. Thus, prediction of the missing counts for those dates is one of the key issues in the study.

Data on environmental covariates contains daily air temperature and flow velocity measurements. These datasets are aggregated into daily averages. Both covariates are measured near the video site (although conditions may differ between the video site and upper parts of the stream from where the smolts begin their migration). However, these datasets are considered as reasonable proxies for the environmental conditions affecting the migration behavior of the smolts.

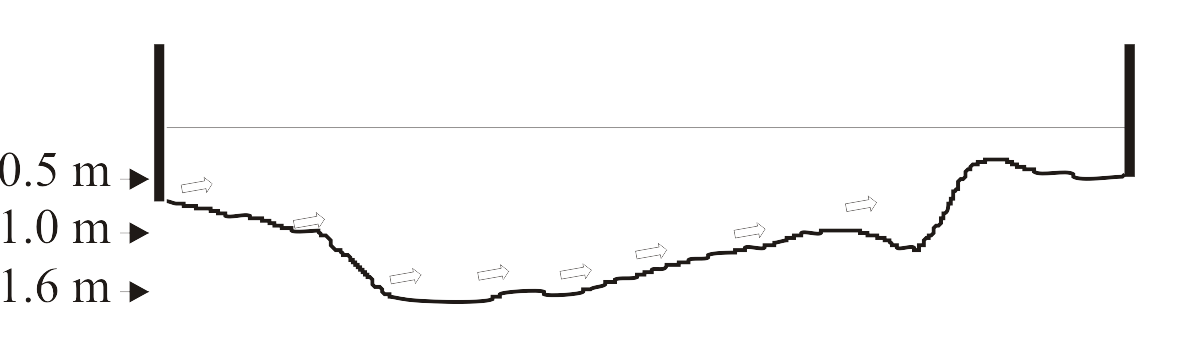
Figure 1: Setup of the video cameras at the river bed in Utsjoki.

Figure 2. Data on daily mean air temperature and mean flow velocity at Utsjoki on study years

Statistical models

Next, the main pieces of model framework are described. Some nuisance parameters are not included to avoid cluttering the paper. However, the complete JAGS code for the model structure can be found from appendix B.

Process of departing

**An individual smolt’s probability to begin the migration in day (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 and temperature :**

**Informative prior distributions are given to parameters , and according to expert view (see tables 1 and 2).**

**The daily probabilities of departing (given those individuals have not departed yet)() can be transformed into proportions of the total smolt run that depart each day () as**

1. **.**

Process of travelling

According to the expert view, it shouldn’t take more than 14 days for smolts to arrive at the video site after departing. Thus, we assume that the smolts that depart on day pass the video site in one of the next 14 days according to cumulative distribution function of lognormal distribution:

1. .

Here is the cumulative density function of the standard normal distribution, is the mean of log(travel time) to video site of smolts that depart on day , and is the standard deviation of log(travel time) (Schwarz & Dempson 1994, Mäntyniemi & Romakkaniemi 2002). Thus, is the proportion of smolts that depart on day and arrive on day , when goes from to

To ensure that all smolts arrive within 14 days of departing, ’s are standardized:

1. ,

and thus

It is further considered that increasing flow velocity increases the speed of the smolts and shortens their travel time. The expected travelling time on real scale is assumed to follow a lognormal distribution with expected value and coefficient of variation and it is assumed to depend on the flow velocity ():

1. **).**

Again, informative prior distributions are given for parameters **, and according to expert view (see tables 1 and 2).**

Arrival distribution

**Total annual size of the smolt run passing the video site on is assumed to follow a uniform distribution on a log scale:**

1. **.**

**Parameterisation at the log-scale allows for a wide, minimally informative prior distribution for the annual total abundance. The annual run size is further considered to be distributed over 61 days of June-July according to Dirichlet-multinomial distribution. Dirichlet-multinomial is a multivariate version of a beta-binomial distribution, meaning that instead of two potential outcomes (as with binomial distribution) there can be any finite number of potential outcomes, this being 61 in our case. For computational simplicity, we approximate multinomial distribution as**



**Here ’s are the** proportions of smolts passing the video site each day, assumed to follow a Dirichlet distribution:

1. **,**

**where is the expected proportion of smolts passing the video site in day and is the overdispersion parameter. To ease the computation, however, we use lognormal** approximation for Dirichlet-distribution **(see appendix A for details).**

**When combining the processes of departing and travelling described earlier in this chapter, we can achieve the following joint distribution:**

1. **.**

**With this set of parameters (**) **we can finally formulate the arrival distribution, i.e. the expected proportion of smolts passing the video site each day:**

Observation process

In previous sections we have covered the processes assumed to affect the timing of the smolt passage on the video site. To combine the observed number of smolts at the video monitoring experiment, we need yet to include the process of observing. It seems natural to assume that there are always some individuals that pass the site unobserved, and thus this proportion must be acknowledged when total run size is estimated. Next, we introduce two versions for observation process.

First version is simple yet not very realistic. We assume that the number of smolts observed at the video site on day in year ( follows Beta-binomial distribution:

1. .

Here is the total number of smolts passing the video site and numerical constants indicate that the probability that an individual smolt is observed follows distribution having mean 0.91, and standard deviation of 0.03. Thus, 9% of smolts would pass the video site unobserved, with 95% PI [5%, 14%]. Albeit an observation model with fixed probability for observing is not realistic, it can be used to illustrate how the processes of departing and travelling affect the model results yet acknowledging that the process of observing is not perfect.

The second version for observation process aims at higher realism, accounting for expert views on how great the probability to observe a smolt may be in excellent vs. poor circumstances and how the level of water in the stream could affect this probability. Let’s consider again that the number observed follows a Beta-binomial distribution:

1. .

The probability to observe a smolt follows now a distribution, where is the expected probability to be observed and is the overdispersion parameter. The expected value of the process is once more linked to an environmental covariate, flow, following the expert view that in very good visibility (low flow) at maximum 90% of the smolts can be observed. As the flow increases, the visibility decreases and again, according to the expert, observation probability decreases gradually at minimum of 30%. Between these limits of 30% and 90%, the expected probability is considered to follow logit-normally linear relationship:

1. and

where , and have prior distributions according to the expert view.

Later in this paper we will refer to the model with simple, fixed, observation process as model 1) and to the observation model with expert knowledge and environmental covariate as model 2).

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 the background knowledge he has about the process in question instead of the data he knows and that will be analysed via final model structure. 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 in question. Changes were made in the prior distributions until the expert agreed that the illustration is in line with his views on the topic. However, care was taken to ensure that the priors would not be chosen based on the updated posterior distributions.

Table 1: List of symbols, descriptions and prior distributions

|  |  |  |
| --- | --- | --- |
| Symbol | Description | Prior distribution |
|  | Total annual number of smolts on log scale | **Uniform(7,15)** |
|  | Total annual number of smolts on real scale | - |
|  | Daily proportion of smolts arriving at video site | - |
|  | Expected proportion of smolts arriving at video site | - |
|  | Overdispersion in the arrival process | **Uniform(0.001,100000)** |
|  | Probability to begin migration (if haven’t done so earlier) | - |
|  | Proportion of smolts departing each day | - |
|  | Intercept for prob to depart given temperature | N(-20,1) |
|  | Regression coefficient for prob to depart given temperature | logN(0.6,0.1) |
|  | Standard deviation of probability to depart | logN(0,1) |
|  | Unstandardised proportion for travel time | - |
|  | Standardised proportion for travel time | - |
|  | Mean of log(travel time) | - |
|  | Standard deviation of log(travel time) |  |
|  | Coefficient of variation of log(travel time) | **Uniform(0.001,2)** |
|  | Intercept for expected travel time given flow | logN(0.52,0.07) |
|  | Regression coefficient for expected travel time given flow | logN(-4.6,0.04) |
|  | Coefficient of variation of expected travel time | **Uniform(0.001,1)** |
|  |  |  |
| Data |  |  |
|  | Daily (air) temperature | - |
|  | Daily flow | - |
|  | Daily number of observed smolts | - |
|  |  |  |

Table 2. Examples of the 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?** | * 1. °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 traveling time (in days) in which a smolt can be considered to arrive at the video site?** | 14 days |
| * **What is the minimum travelling 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) flow velocity?** | 90% (30%) |
| * **At what level of flow visibility can be considered optimal?** | 20m3/s |
| * **At what level of flow visibility is considered to be poorest?** | 50-60m3/s |
|  |  |

# Results

Limit the results to answers to the questions posed in the purpose of the work and condense them as comprehensively as possible. Give the findings as nearly as possible in the terms in which the observations or measurements were made so as to avoid confusion between facts and inferences. State noteworthy findings to be noted in each table and figure, and avoid restating in the text what is clear from the captions. Material supplementary to the text can be archived in the report literature or a recognized data depository and referenced in the text (see Supplementary material section)

Models 1 and 2 (with simple vs. realistic observation process, respectively) were fitted using MCMC sampling with JAGS (Just Another Gibbs Sampler, Plummer 2003) software. (describe run length, burnin, about convergence diagnostics, should traces be put to appendix/ supplementary material?). Posterior estimates of model 2 indicate that on average, 30-75% of smolts pass the video site unobserved (Figure 3). Considering the estimated uncertainty, proportion unobserved can be as high as 89% among the years studied. Posterior estimates of daily arrivals are illustrated in figures 4 (model 1) and 5 (model 2). Estimated number of smolts passing the video site during the first 23 days of the study in 2005 (days with the missing data) has mean 1530 and 95% PI [290, 4060] based on model 1 and mean 1880 and 95% PI [320, 5310] based on model 2. As a proportion of the total in 2005, both models estimate an equal share for the missed days with mean 0.09 and 95% PI [0.02, 0.23].

Figures 6 and 7 illustrate the prior and posterior distributions from model 2 for expected processes of departing and travelling. Posterior distributions from model 1 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 6). The expected distribution for travel time to the video site (Figure 7) supports somewhat longer duration than expected a priori both for low and high flow velocities. Note that the figure 7 illustrates the expected travel time in unstandardized form and that the standardized travel time accounts also for variation around these expected values (see equation 6).

In model 2, observation probability depends on the flow velocity. Figure 8 illustrates the prior and posterior distributions of the observation process given flow velocity. Priors and posteriors are close the same, indicating that in this model framework there isn’t available information enabling further learning about the process of observing.

 Figure 3. Annual number of smolts passing the video site. Posterior distributions of model 1 (grey boxplots), model 2 (black boxplots) and count data (black dots).

 Figure 4. Daily number of smolts that pass the video site. Posterior distributions of model 1 (boxplots) and count data (grey dots).

Figure 5. Daily number of smolts that pass the video site. Posterior distributions of model 2 (boxplots) and count data (grey dots).



Figure 6. Expected probability that a smolt begins the migration at given temperature. Prior (grey boxplots) and posterior distributions (black boxplots) from model 2.

 Figure 7. Expected (unstandardized) travel time in days from the point of departure to the video site. Above: Distribution of expected travel time at low flow velocity (10 m3/s). Below: Distribution of expected travel time at high flow velocity (100 m3/s). Prior (grey boxplots) and posterior distributions (black boxplots) from model 2.



Figure 8. The expected observation probability at given flow velocity. Prior (grey boxplots) and posterior distributions (black boxplots) from model 2.

# Discussion

Limit the Discussion to giving the main contributions of the study and interpreting particular findings, comparing them with those of other workers. Emphasis should be maintained on synthesis and interpretation and exposition of broadly applicable generalizations and principles. If these are exceptions or unsettled points, note them and show how the findings agree or contrast with previously published work. Limit speculation to what can be supported with reasonable evidence. End the Discussion with a short summary of the significance of the work and conclusions drawn. If the Discussion is brief and straightforward, it can be combined with the Results section.

In our study, we broke down the distinguishable pieces of an arrival model separating those to departing, travelling and observation processes. This distinction made it possible to go beyond assuming that count data of salmon smolts follow mathematical rules as a whole but rather concentrating on realistic rules concerning each sub-process. The model framework aims at biological realism, accounting for expert knowledge in the model formulation (e.g. choices for likelihood functions and prior distributions) and on how environmental covariates may provide information about the timing of the migration. Thus, estimated total uncertainty accounts for the uncertainty in the sub-processes, including the observation process in which changes in the environmental conditions (i.e. level of water/flow velocity) are known to have a large impact on the data.

With the approach chosen, the need to assume a specific shape for the arrival curve is avoided. Past studies (Hilborn et al. 1999, Su et al. 2001, Sethi & Bradley 2016) assume unimodal shape (e.g. normal, skew-normal, student-t) for the arrival curve but when a closer look is taken on arrival time-series, it seems clear that the curves can be imagined to have many different shapes and between years variability may appear.

To estimate reliably the total uncertainty related to fisheries models, it is necessary to construct the models from the perspective of biological realism. We have shown that such is possible for passage count models by considering the biological theory on how environmental covariates, e.g. temperature and flow velocity, may affect the salmon smolts’ arrival to the monitoring site. Furthermore, we claim that the model framework should acknowledge the details of the observation process in question, instead of simply considering the variation as “error” or “noise”.

By considering expert knowledge in our model formulation, we have managed to avoid the need to make assumptions of the mathematical shape of the arrival curve.

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 make more sense biologically than another. We argue further that the Bayesian Model averaging (BMA, Hoeting et al. 1999) is more suitable method to consider model uncertainty compared to DIC. Instead of calculating information criteria separately for a dataset under alternative models, these alternatives can be defined simultaneously under a single model framework and the posterior probabilities for those alternatives calculated (Pulkkinen and Mäntyniemi 2013). Including hierarchical structure over exchangeable datasets and studying which of the alternative models supports the whole dataset the best minimizes the chance that a selected model only fits a set of data by chance.

Rather little attention is paid in past studies to the role of observation models in estimating the total uncertainty. Hillborn et al. 1999 distinguishes process and observation errors, assuming observer efficiency either as known or unchanging in time. They also suggest that weir studies in index streams should be used to estimate observer bias in other streams. Su et al. 2001 discuss the issue of error in the observation process and mainly conclude that such can cause convergence problems in sampling. Sethi and Bradly (2016) assume either normal or negative Binomial models for the observed passage counts, both of which allow for the “true” passage to be greater or smaller than the observed. However, it would seem more realistic to consider the number of individuals observed as a minimum estimate for the true number especially if movement of individuals downstream is negligible. Depending on the approach being chosen, 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.

Furthermore, Su et al. (2001) had 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. However, our model structure is not hierarchical, as we assume that the underlying processes are the same from year to year and 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, exhangeability in biological parameters could be considered between the stocks. In fact, such approach could allow meta-analysis of data from different types of monitoring systems if detailed observation processes were tailored for each system and exhangeability was considered as reasonable assumption for the biological processes.

Our study could be fine-tuned and extended further. Four years of data were chosen to illustrate the approach, but for the needs of stock assessment full time series of data (15 years) will be analysed. Because of a lack of several experts, prior understanding of the processes is based only on expert knowledge of one person, although it would be desirable to combine knowledge from several experts as the views of uncertainty tend to be greater between experts than within them (Uusitalo et al. 2005). 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. Salmon smolts have a tendency to school during their migration (e.g. Bakshtanskiy et al. 1988) and this creates more variation in the monitoring data compared to if those would move independently (Mäntyniemi and Romakkaniemi, 2002). For Utsjoki video survey, daily average school sizes of smolts are recorded. Thus incorporating those into the observation model framework would be relatively simple.

Despite these issues, we claim that our model framework is reasonable and most of all, contains the elements of realism that are required to provide reliable estimates for the strength of total annual smolt run. This approach can be easily applied to similar monitoring surveys, those being, for example, sonar, trap, or weir studies.

* Both the dynamics of the passage and the observation process can be influenced by the environmental conditions and as far as we know, such has not been included in the passage models so far.

# Acknowledgements

Acknowledgements should be written in the third person. We strongly urge authors to limit acknowledgments to those who contributed substantially to scientific and technical aspects of the paper, gave financial support, or improved the quality of the presentation. Avoid acknowledging those whose contribution was clerical only.

# 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.

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.

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.

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, Austrua. ISSN 1609-395X.

Pulkkinen, H., and Mäntyniemi, S. 2013. Maximum survival of eggs as the key parameter of stock–recruit meta-analysis: accounting for parameter and structural uncertainty. Can. J. Fish. Aquat. Sci. 70: 527–533 (2013) dx.doi.org/10.1139/cjfas-2012-0268.

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. dx.doi.org/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(8): 1648–1662. doi:10.1139/f01-099.

Uusitalo, L., Kuikka, S., and Romakkaniemi, A. 2005. Estimation of Atlantic salmon smolt carrying capacity of rivers using expert knowledge. ICES J. Mar. Sci. 62: 708-722. doi:10.1016/j.icesjms.2005.02.005.

# Appendix A: Lognormal approximation for Dirichlet-distribution

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

(A.1) **,**

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

(A.2)

**(A.3)**

**(A.4) and**

**(A.5)**

# Appendix B: JAGS code for the arrival model