SpawnEst: a tool for salmon spawner estimation

Joseph L. Simonis1 and Lucius K. Caldwell2

1DAPPER Stats

2Cramer Fish Sciences

2017-06-12

**INTRODUCTION**

Anadromous Pacific salmon return to freshwater streams to spawn at the completion of their life cycle. Spawning adults represent the contribution to the next generation and are perhaps the most accessible life stage for sampling, and thus are often the target of management practices and legal permitting guidelines (Fried 1994, Quinn and Deriso 1999, Knudsen 2008, Millar *et al*. 2012). It is therefore important that the size of the spawner population be estimated using the most robust (*i.e.*, least biased) and informative method available (Good *et* al. 2007, Gallagher *et al.* 2010).

A standard approach to collecting data on spawner population sizes in the absence of weir data is to engage in repeated ground surveys, where an observer walks a defined length of a stream and counts all live spawners seen (Cousens *et al*. 1982, Bue *et al*. 1998, Holt and Cox 2008). The surveys typically occur at irregular frequencies, and while the goal is for sampling to begin before and continue after the entire run (*i.e.*, counts start with 0 and end with 0), ground survey datasets typically have some runs that were not fully captured by sampling. The goal of the estimation is therefore to translate irregular and incomplete counts of observed spawners into a reliable estimate of total spawner population size, which can then be used to manage the population and set take limits (Hilborn *et al*. 1999). The mathematics underlying this estimation is the subject of this project.

The historical approach to spawner estimation has been to integrate the observed spawner count over time (generating an estimate of observed spawner-days) using the trapezoidal area under the curve (AUC) method (Ames and Phinney 1977, Beidler and Nickelson 1980, English *et al*. 1992, Bue *et al*. 1998). Because the AUC method fails if the data do not begin and end at 0, the counts are amended with terminal 0s, typically placed on dates that are the average steamlife of a spawner away from the actual terminal samples (Bue *et al*. 1998, Hilborn *et al*. 1999). The trapezoidal area is then divided by the average streamlife of a spawner (generating an estimate of observed spawners) and then divided by the detection probability to produce the number of total spawners.

As a result of its formulation, the AUC methodology has substantial statistical and biological shortcomings (including the need to amend the data, assumption of a fixed streamlife and a fixed detection probability, lack of variance in the estimate, and improper specification of the observation model) that render its output likely highly biased and falsely precise. Despite these issues and the publication of subsequent estimators that remedied some of them (Quinn and Gates 1997, Hilborn *et al*. 1999, Su *et* al. 2001, Holt and Cox 2008, Szerlong and Rundio 2008, Millar *et al*. 2016), the AUC method is still commonly used with salmon ground surveys (*e.g.*, Gallagher *et al*. 2010, Souhein *et al*., 2016). This fact is likely due to the historical precedent and ease of application of the AUC. Thus, in order for a proposed estimator to be of actualizable utility in managing fisheries, it needs to overcome the ease of use of the AUC method. To that end, it is the goal of this project to develop a robust and informative estimator that is easy and intuitive to implement.

Here, we develop a flexible estimator that addresses the existing statistical and biological issues with the AUC method and other proposed methods. We base our approach on generalized process-based modeling using continuous time Markov chains (Allen 2011) and particle filtration (Ionides *et al.* 2006, Ionides *et al.* 2015, King *et al.* 2016), facilitating robust likelihood estimation and higher-order analyses (Pawitan 2001, Burnham and Anderson 2002). We wrap our estimator within a graphic user interface (GUI) in order to enable ease of application and comparison with other estimators such as the AUC. We have named this estimator tool SpawnEst and have deployed a limited proof-of-concept version of the GUI at https://dapperstats.shinyapps.io/SpawnEst/.

For purposes of comparison and generalization, we conceive of SpawnEst as a general tool for spawner estimation that includes data visualization and analysis using any of a variety of specific estimators. As a beginning point, the currently deployed SpawnEst application includes an automatic graphic capacity, the standard AUC estimator, and the beginning stages of our pomp model (see below), with pre-loaded data from pink salmon in Irish Creek in 1990 (Hilborn *et al*. 1999). The goal of this stage of our project is to develop a more robust and informative, yet easy-to-use estimator of spawner population size and to incorporate it in the SpawnEst tool alongside the AUC estimator. The subsequent stage of our project will focus on expanding the capacity of SpawnEst to estimate spawner population sizes across multiple sites and multiple years, based explicitly and hierarchically upon the single-site, single-year model.

**MOTIVATING APPLICATION CONTEXT**

Our work is motivated by the need to develop a more robust and informative, yet still user-friendly estimator of salmon spawner population sizes. For the sake of sound science, there is a general need to move away from the standard AUC approach and any analytical methods that require fabricated data in order to work. At the same time, there is desire to retain the simplicity and flexibility of the AUC methodology as well as a need to minimize the end-user learning curve to overcome the computational simplicity of the AUC method.

In the context of salmon adults returning to a stream to spawn before dying, we consider a population that is subject to two processes: immigration and mortality. Following existing models, we implicitly assume that there is no emigration, that all adults die after spawning, and that there are no resident spawners. We recognize that there are situations where these assumptions may not hold, and future articulations of this model will explicitly relax each of those assumptions. For the present model, however, we consider our population as only subject to immigration and mortality.

The rate of immigration (return) for salmon spawners is highly variable through time, often exhibiting multiple pulses on top of a non-normal arrival distribution (Keefer *et al*. 2004, Szerlong and Rundio 2008, Branch and Hilborn 2010, Gard and Bottorff 2014). Mortality rates for spawners are also not constant, but rather tend to decrease over the course of the season, such that salmon arriving later in the season tend to have shorter stream lifespans than those arriving earlier (Dangel and Jones 1988, Perrin and Irvine 1990, Fukushima and Smoker 1997, Su *et al*. 2001). In addition to their temporal variability, both the immigration and mortality rates could be influenced by environmental covariates such as hydrological variables or temperature.

We recognize that the data collected (number of salmon spawners observed) are impacted by the sampling and observation components of the system and therefore only partially speak to the true state of the system (total number of spawners), resulting in additional uncertainty (Hobbs and Hooten 2015). For the purposes of the present model, data on spawner populations are collected from ground surveys that tend to occur at variable (non-fixed) intervals (Bue *et al*. 1998). In addition, while the goal of sampling is to fully cover the run, many datasets do not fully encompass the run of interest (Bue *et al*. 1998). The observation process is governed by the detection probability (a.k.a. searcher efficiency): the rate at which spawners present are actually observed (Hobbs and Hooten 2015). Searcher efficiencies can vary substantially among sites as well as within sites both within and between years (Bue *et al*. 1998) due to hydrological variation, geology, substrate, shading caused by leaf-out, and other factors.

**GENERAL MODEL FRAMEWORK**

The approach we use here is based on continuous time Markov chains that are partially observed (Allen 2011, King *et al*. 2016), here referred to as “partially observed Markov process” (pomp) models. POMP models (a.k.a. state-space models) are a general class of hierarchical mathematical models used to analyze time series of data that have been measured incompletely and with error (Ionides *et al.* 2006, King *et al.* 2016). Such approaches are certainly beneficial in fisheries biology, where time series of population sizes or harvest quantities are often estimated via subsampling methods (Hilborn and Mangel 1997). We first develop a general model framework that can be used to robustly estimate escapement (and its uncertainty) under a wide range of specific data, biological, and management situations. Recognizing the need for applicability, we then develop specific models for use in analyzing spawner ground survey data.

Reflective of our focus on spawning adults returning to streams, we model the population using a simple immigration-death chain (*sensu* Allen 2011), where individuals arrive through a time-varying arrival process and die via a time-varying mortality process . The fish present are sampled according to a time-varying detection probability Fish present, alive, and able to be counted are tracked by the time-varying discrete-valued state variable and the count of fish made at any time is tracked by the time-varying discrete-valued state variable .

We elaborate the chain by defining the infinitesimal increment probabilities, which describe how the number of individuals arrive and die though infinitesimal slices of time. Specifically, we articulate the counting process for arrivals and the counting process for deaths, where is defined for the infinitesimal amount of time (Allen 2011). The probability of one (or more than one) individual passing through each of the counting processes during is

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where (“little o of h”) is Bachmann-Landau notation and can be thought of as a general function that is of smaller order than (which is, itself, infinitesimal) (Allen 2011). Formally, for a sufficiently large , , but . The counting processes define how the number of alive and present spawners () changes through time via the arrival of new spawners and the mortality of already present spawners:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

This general, time-non-homogeneous immigration-death Markov chain corresponds to the following familiar ordinary differential equation in mean-field:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

We consider explicitly the fact that not all present spawners are necessarily counted (*i.e.*, the system is “partially observed”) by hierarchically adding an observation component to the model. Specifically, the number of fish observed at a given time is a binomial sampling process defined by the detection probability () and the number of spawners available for counting (), both of which are time-varying:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

By defining the population and our sampling of it through time-non-homogeneous processes and parameters (, , and ), our model retains its flexibility, which allows us to avoid restrictive assumptions regarding parameters (*e.g.*, normal distribution of arrival times, fixed mortality rate). In its current form, however, the model is still too general to be of utility in specific situations. In subsequent sections, we first articulate and specify the model such that it is functionally useful for estimation of spawner population sizes and then detail the inference machinery.

**LOCAL SPAWNER ESTIMATION**

Given the wide range of potential run timing distributions, we use a flexible Fourier series to model the arrival rate through time ():

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

where defines the overall mean and and control the seasonal dynamics, allowing for a near-infinite diversity of run timings and distributional shapes as well as total spawner population sizes. The rate defines the mean rate at which spawners arrive through the arrival process .

The translation from to could occur simply via the identity function (*i.e.*, ), but can also be leveraged to incorporate relevant process variability. In many circumstances, a Poisson distribution governed by as its single rate would adequately describe the variability in the data. However, we recognize that the number of spawners arriving to streams are substantially more variable than would be expected under a Poisson distribution. Thus, we model the number of spawners arriving during a time interval (from to ) using a negative binomial distribution governed by the mean rate and with additional variance controlled by the size (dispersion) parameter :

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

where as , .

We consider that mortality occurs through time according to the rate (the inverse of survey life), which we consider to be a simple linear function of time:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

We model the number of spawners that die during a time interval (from to ) using a binomial distribution governed by the number of spawners and

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Finally, the count of fish observed in the stream length when it is sampled at time point using a binomial distribution governed by the number of spawners and the detection rate. For the sake of model simplicity at this early stage, we set the detection rate as constant through time (*i.e.*, for all ), producing the sampling model

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

**MODEL INFERENCE**

The likelihood forms the basis of statistical inference using a wide range of methodologies including frequentist, Bayesian, and information-theoretic frameworks, and therefore is a crucial tool for gaining statistical insight into our spawner model (Pawitan 2001). Given the non-linear, non-Gaussian, stochastic, indirect observation, and time-series natures of POMP models, it is preferable to use simulation-based methods (*i.e.,* Monte Carlo) to estimate likelihoods, and avoid evaluating the probability densities directly (Bretó *et al.* 2009, He *et al.* 2010, Ionides et al. 2015). Indeed, by taking a simulation-based approach, we are retaining the flexibility of our model and facilitating its future evolution organically.

Specifically, we will rely on the particle filter (a.k.a. sequential Monte Carlo, SMC) to numerically estimate the likelihood under given parameter sets (Kitagawa 1987, Doucet *et al.* 2001, Ionides *et al.* 2006). The particle filter is a general Monte Carlo method used for statistical inference of noisy time series in a variety of fields including physical chemistry, signal processing, geolocation, machine learning, bioinformatics, economics, and pharmacokinetics (King *et al.* 2016). While incredibly flexible, simulation-based methods like particle filtration are typically hamstrung by computational requirements due to their need to simulate the model many times in specific ways in order to ensure proper estimation of likelihoods. As a result, most simulation-based inference methods have only been applied to real-world problems within the past few decades, in particular following the advent of multi-core processing (Ionides *et al.* 2006, King et al. 2016). Thus, while the mathematics underneath simulation-based approaches to POMP models are well-defined and robust, their implementation has only been made possible very recently (Doucet *et al.* 2001, Ionides *et al.* 2006, Andrieu *et al.* 2010, Ionides *et al.* 2015, King *et al.* 2016).

Iterated filtering algorithms (Ionides *et al.* 2006, Ionides *et al.* 2015) provide frequentist optimization based on full-information likelihood (rather than likelihoods based on features of the data) for pomp models. Here we use the IF2 algorithm (Ionides *et al.* 2015), a relatively computationally expedient (but still intensive and time-consuming to run) procedure that relies on iteratively perturbing the Bayes map of the model. The IF2 algorithm operates by conducting a particle filter with the parameter values doing a random walk (Ionides *et al.* 2015). At the end of the time series, the collection of parameter vectors is used to initiate a new particle filter with a smaller random walk. This process is repeated, thereby iteratively filtering the data. The decrease in the variance of the random walk over time is referred to as cooling and controls how quickly (or slowly) the search algorithm switches from a more “global” to a more “local” search of the likelihood surface (Ionides *et al.* 2015). Theoretically, the IF2 algorithm converges on the parameter values that maximize the likelihood, and this is supported by empirical evaluations (Ionides *et al.* 2015).

As with any simulation-based approach to statistical inference, we recognize the influence that control parameters (such as the number of particles) and the starting values used for model parameters have on the outcome of our analyses (Ionides *et al*. 2015). To that end, it is crucial to determine a standardized, robust approach to evaluating and accounting for the influence of control parameters and starting values. This is the current stage of the project.

**USER INTERFACE**

The most up-to-date functional version of the SpawnEst app is deployed via ShinyApps at https://dapperstats.shinyapps.io/SpawnEst/.

**CODE**

SpawnEst is coded in R (R Core Team 2017). The most up-to-date code for SpawnEst is held in a repository on GitHub at https://github.com/DAPPERstats/spawnest.

The user interface is coded using shiny (Chang *et al*. 2017), shinysky (AnalytixWare 2014), and rhandsontable (Owen 2017) packages. Construction and evaluation of the pomp models is done using the pomp package (King *et al*. 2016, King *et al.* 2017).

**FUTURE WORK**

The model presented here is the first step in the development of a broader and more generally applicable spawner estimation tool. We recognize that there is relevant sampling occurring at spatial scales larger than a single stream and temporal scales larger than a single year. A general multi-site and multi-year model is of interest, but builds upon the single-site, single-year formulation, and so we focus here on the local and proximate sampling components.

**REFERENCES**

Allen, L. J. S. 2011. *An introduction to stochastic processes with applications to biology*. CRC Press, New York.

Ames, J. and D. E. Phinney. 1977. 1977 Puget Sound summer-fall chinook methodology: escapement goals, run size forecasts, and in-season run size updates. *Washington Department of Fisheries Technical Reports* **29**. 71 p.

AnalytixWare. 2014. shinysky: A set of shiny UI components/widgets. R package version 0.1.2. https://github.com/AnalytixWare/ShinySky

Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **72**: 269–342.

Beidler, W. M. and T. E. Nickelson. 1980. An evaluation of the Oregon Department of Fish and Wildlife Standard Spawning Fish Survey System for coho salmon. *Oregon Department of Fish and Wildlife Information Report Series* **80-9**. 23 p.

Branch, T. A. and R. Hilborn. 2010. A general model for reconstructing salmon runs. *Canadian Journal of Fisheries and Aquatic Sciences* **67**: 886–904.

Bretó, C., D. He, E. L. Ionides, and A. A. King. 2009. Time series analysis via mechanistic models. *Annals of Applied Statistics* **3**: 319-348.

Bue, B. G., S. M. Fried, S. Sharr, D. G. Sharp, J. A. Wilcock, and H. J. Geiger. 1998. Estimating Salmon Escapement using Area-Under-the-Curve, Aerial Observer Efficiency, and Stream-Life Estimates: The Prince William Sound Pink Salmon Example. *North Pacific Anadromous Fish Commission Bulletin* **1**: 240–250.

Burnham, K. P. and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach* (Second Edition). Springer, New York.

Chang, W., J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. 2017. shiny: web application framework for R. R package version 1.0.3. https://CRAN.R-project.org/package=shiny

Cousens, N. B. F., G. A. Thomas, C. G. Swann, and M. C. Healey. 1982. A review of salmon escapement estimation techniques. *Canadian Technical Report of Fisheries and Aquatic Sciences* **1108**. 130 p.

Dangel, J. R. and J. D. Jones. 1988. Southeast Alaska pink salmon total escapement and stream life studies, 1987. *Alaska Department of Fish and Game, Regional Information Report* **IJ88-24.**

Doucet, A., N. de Freitas, and N. Gordon. 2001. *Sequential Monte Carlo methods in practice*. Springer-Verlag, New York.

English, K. K., R. C. Bocking, and J. R. Irvine. 1992. A robust procedure for estimating salmon escapement based on the area-under-the-curve method. *Canadian Journal of Fisheries and Aquatic Sciences* **49**: 1982–1989.

Fried, S. M. 1994. Pacific salmon spawning escapement goals for the Prince William Sound, Cook Inlet, and Bristol Bay areas of Alaska. Alaska Department of Fish and Game, Division of Commercial Fisheries, Management, and Development. *Special Publication* **8**.

Fukushima, M. and W. W. Smoker. 1997. Determinants of stream life, spawning efficiency, and spawning habitat in pink salmon in the Auke Lake system, Alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **54**: 96–104.

Gallagher, S. P., P. B. Adams, D. W. Wright, and B. W. Collins. 2010. Performance of spawner survey techniques at low abundance levels. *North American Journal of Fisheries Management* **30**: 1086–1097.

Gard, R. and R. L. Bottorff. 2014. A history of sockeye salmon research, Karluk River system, Alaska, 1880-2010. *NOAA Technical Memorandum NMFS-F/SPO* **125**. 413 p.

Good, T. P., T. J. Beechie, P. McElhany, M. M. McClure, and M. H. Ruckleshaus. 2007. Recovery

planning for Endangered Species Act-listed Pacific salmon: using science to inform goals and strategies. *Fisheries* **32**: 426–440.

He, D. E. L. Ionides, and A. A. King. 2010. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface* **7**: 271-28.

Hilborn, R., B. G. Bue, and S. Sharr. 1999. Estimating spawning escapements from periodic counts: A comparison of methods. *Canadian Journal of Fisheries and Aquatic Sciences* **56**: 888–896.

Hobbs, N. T. and M. B. Hooten. 2015. *Bayesian models: a statistical primer for ecologists*. Princeton University Press, New Jersey.

Holt, K. R. and S. P. Cox. 2008. Evaluation of visual survey methods for monitoring Pacific salmon (*Oncorhynchus* spp.) escapement in relation to conservation guidelines. *Canadian Journal of Fisheries and Aquatic Sciences* **65**: 212–226.

Ionides, E. L., C. Bretó, and A. A. King. 2006. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences, U.S.*A*.* **103**: 18438–18443.

Ionides, E. L., D. Nguyen, Y. Atchadé, S. Stoev, and A.A. King. 2015. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences, U.S.*A. **112**: 719–724.

Keefer, M. L., C. A. Peery, M. A. Jepson, K. R. Tolotti, T. C. Bjornn, and L. C. Stuehrenberg. 2004. Stock-specific migration timing of adult spring-summer Chinook salmon in the Columbia River basin. *North American Journal of Fisheries Management* **24**: 1145–1162.

King, A. A., D. Nguyen, and E. L. Ionides. 2016. Statistical inference for partially observed Markov processes via the R Package pomp. *Journal of Statistical Software*. http://kingaa.github.io/pomp/vignettes/pompjss.pdf

King, A. A., E. L. Ionides, C. M. Breto, S. P. Ellner, M. J. Ferrari, B. E. Kendall, M. Lavine, D. Nguyen, D. C. Reuman, H. Wearing, and S. N. Wood. 2017. pomp: statistical inference for partially observed Markov processes. R package, version 1.12. http://kingaa.github.io/pomp

Kitagawa, G. 1987. Non-Gaussian state-space modeling of non-stationary time series. *Journal of the American Statistical Association* **82**: 1032-1041.

Knudsen, E. E. 2000. Managing Pacific salmon escapements: the gaps between theory and reality. In: E. E. Knudsen, C. R. Steward, D. MacDonald, J. E. Williams, and D. W. Reiser (eds.), *Sustainable Fisheries Management: Balancing the Conservation and Use of Pacific Salmon*. CRC Press, London. pp. 237–272.

Millar, R. B., S. McKechnie, and C. E. Jordan. 2012. Simple estimators of salmonid escapement and its variance using a new area-under-the-curve method. *Canadian Journal of Fisheries and Aquatic Sciences* **69**: 1002–1015.

Jonathan Owen. 2017. rhandsontable: interface to the 'handsontable.js' library. R package version 0.3.4.6. http://jrowen.github.io/rhandsontable/

Pawitan, Y. 2001. *In all likelihood: statistical modeling and inference using likelihood.* Clarendon Press, Oxford.

Perrin, C. J. and J. R. Irvine. 1990. A review of survey life estimates as they apply to the area-under-the curve method for estimating the spawning escapement of Pacific salmon. *Canadian Technical Report Fisheries and Aquatic Sciences*: **1733**.

Quinn, T. J., II and R. Gates. 1997. Estimation of salmon escapement: models with entry, mortality, and stochasticity. *Natural Resource Modeling* **10**: 217–250.

Quinn, T. J., II and R. B. Deriso. 1999. Quantitative fish dynamics. Oxford University Press, New York.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sounhein, B., E. Brown, M. Lewis, and M. Weeber. 2016. Status of Oregon Stocks of coho Salmon, 2015. Oregon Adult Salmonid Inventory & Sampling Project; Western Oregon Research and Monitoring Program; Oregon Department of Fish and Wildlife, Salem, OR. 57 p.

Su, Z., M. D. Adkinson, and B. W. Van Alen. A hierarchical Bayesian model for estimating historical salmon escapement and escapement timing. *Canadian Journal of Fisheries and Aquatic Sciences* **58**: 1648–1662.

Szerlong, R. G. and D. E. Rundio. 2008. A statistical modeling method for estimating mortality and abundance of spawning from a time series of counts. *Canadian Journal of Fisheries and Aquatic Sciences* **65**: 17–26.