**Quantile Regression Forest Models to Estimate Juvenile Rearing Capacity for Anadromous Salmonids**

Kevin See1,2, Michael W. Ackerman1, Richard A. Carmichael1, Chris Beasley1

Estimating Life-Stage Specific Habitat Capacity for Anadromous Salmonids using Quantile Random Forest Models

1Biomark and Quantitative Consultants, Inc., 705 South 8th St., Boise, Idaho, 83702, USA

2Corresponding author: [Kevin.See@biomark.com](mailto:kevin@qcinc.org)

**Running Head:** QRF to Estimate Habitat Capacity

# Abstract

# Outline

1. INTRODUCTION
   1. Paragraph 1: Salmon Habitat Restoration
      1. Salmon are depleted
      2. Depletion partially due to loss of tributary habitat
         1. Loss of spawning gravels
         2. Loss of juvenile rearing habitat
            1. Both quantity and quality
      3. Lot of effort toward restoring quantity/quality to historic levels
      4. To restore habitat capacity to support increase abundance
         1. Spawning
         2. Rearing
   2. Paragraph 2: Capacity As A Metric For Salmon Recovery
      1. What is habitat carrying capacity (*K*)?
      2. Why do we think *K* is an important metric to measure for recovery?
      3. A need for tools to estimate capacity
      4. Previous efforts to describe/estimate capacity or habitat quality
         1. Successes
         2. Downfalls
   3. Paragraph 3: Introduce Quantile Regression Forests and Random Forests
      1. Why do these tools provide promise for estimating carrying capacity?
      2. Benefits over previous methods
      3. List all of the benefits of QRF
   4. Paragraph 4: The Data
      1. Introduce Chinook salmon
      2. What is our fish data
      3. What is our habitat data – CHaMP
      4. Pairing of data
      5. Upper Columbia and Snake River Chinook ESUs
   5. Paragraph 5: The Application
      1. Identify habitat characteristics associated with fish abundance/density
      2. Elicit-fish habitat relationships
      3. Use information from fish-habitat relationships and QRF model to predict capacity where habitat data is available (i.e., CHaMP sites)
      4. Extrapolate estimates at CHaMP sites across watersheds using GAA data
   6. Paragraph 6: Synopsis
      1. We provide estimates of carrying capacity for sites and watersheds
      2. Results are validated with spawner-recruit models (Ricker, Beverton-Holt, etc.)
      3. QRF provides promise to guide restoration planning
2. METHODS
   1. Study Site
   2. Data
   3. Habitat Covariate Selection
   4. QRF Model Fit
   5. Predictions at CHaMP Sites (QRF Model Predictions)
      1. All CHaMP sites
   6. Model Extrapolation
      1. Lemhi River
   7. Model Validation
      1. Everywhere validation is possible
3. RESULTS
   1. Habitat Covariate Selection
   2. QRF Model Fit
   3. Predictions at CHaMP Sites (QRF Model Predictions)
      1. All CHaMP sites
   4. Model Extrapolation
      1. Lemhi River
   5. Model Validation
      1. Everywhere validation is possible
4. DISCUSSION
   1. Biological Expectations from QRF Model
      1. Discussion of PDPs
   2. Model Assumptions
   3. Extrapolation Model
      1. Pros:
         1. Can get estimates for watersheds
      2. Cons:
         1. Relies on messy GAA data
         2. CHaMP habitat data is not reproducible, expensive
         3. Correlations between habitat and capacity is messy
   4. The future
      1. Channel unit scale QRF
      2. Improved habitat data collection
         1. Drones
         2. LiDAR
      3. Collect remote sensing habitat data everywhere
   5. Next Steps
   6. Conclusions

Introduction The decline of anadromous Pacific salmonids (*Oncorhynchus* spp.) across the Pacific Northwest, USA has prompted numerous actions aimed at reversing that trend. These actions are often categorized into four H’s – harvest modification, hatchery practices, hydro-system operations, and habitat rehabilitation. Problematically, there is substantial uncertainty regarding the degree of change that can be exerted across and within these categories, and what combination of changes might most cost-effectively and sustainably reduce mortality. Recently released “de-listing” criteria (NOAA 2016) identified adult escapement targets at the population scale, providing a quantitative metric useful for evaluating the magnitude of survival improvements required. These abundance targets provide a benchmark against which habitat rehabilitation actions can be measured. Here we describe an approach for estimating life-stage specific habitat-based carrying capacity that can be used to quantitatively identify the magnitude of tributary habitat rehabilitation necessary to support de-listing. For perhaps the first time, the necessity of tributary habitat restoration can be demonstrated and the magnitude of required change can be placed in context with the other “H’s.”

Pacific salmon (*Oncorhynchus* spp.) species have experienced large declines in abundance throughout much of their range (Good et al. 2005). Declines can be partially attributed to lost or altered habitat, and thus, efforts to recover depleted salmon populations are replete with efforts to rehabilitate habitat used during the freshwater life-stages. Specifically, restoring salmonid carrying capacity through tributary rehabilitation actions has been identified as a key component of recovery efforts for salmon and steelhead (*Oncorhynchus mykiss*) in the Pacific Northwest, USA. Efforts have included increasing and improving existing habitat for both spawning adults and rearing juveniles. However, estimating habitat carrying capacity, both historic and contemporary, for various life-stages of Pacific salmon, as well as identifying important habitat characteristics that influence capacity, has been an ongoing but necessary challenge. Reliable methods to better understand fish-habitat relationships as well as to estimate capacity are necessary to identify those salmon and steelhead life-stages that are limited by habitat capacity, and further, to better direct tributary rehabilitation efforts.

Fausch et al. (1988) conducted a thorough review of attempts to predict the standing crop of fish from measurable habitat covariates from 1950 to 1985, and found that the vast majority of multiple linear regression models failed to detect a significant fish-habitat signal. Since that review, there has been progress in identifying some fish-habitat relationships for some salmonid species. Nickelson (1992) found that juvenile coho were found in higher densities within pool habitat on the Oregon coast. Sharma and Hilborn (2001) determined that pool and pond densities were good predictors of coho smolt densities in western Washington. Pess et al. (2002) demonstrated that densities of adult spawning coho were higher in forested areas compared to urban or agricultural areas in the Snohomish River watershed. Mossop and Bradford (2006) examined juvenile Chinook in the Yukon river and found positive correlations between the log of fish density and several metrics related to residual pool dimensions and large woody debris abundance as well as a negative correlation between fish density and gradient. Byrant and Woodsmith (2009) found a positive relationship between juvenile coho abundance and large wood at the reach scale, and a negative one between coho and the number of pools. Braun and Reynolds (2011) found positive associations between spawner densities of sockeye in the Fraser River and percent undercuts, large wood and percent pools. However, all of these studies were focused on predicting observed fish densities, not necessarily capacity, and for most of them the predictive extent is limited to a particular watershed. In addition, they all assumed some form of linear fish-habitat relationship, which often results in weak predictive power.

There have been a number of studies that utilized other modeling approaches to elicit fish habitat relationships. Dunham et al. (2002) used a quantile regression approach to show a negative relationship between cutthroat trout densities and the width:depth ratio of a stream for the upper quantiles of trout density. Eastwood et al. (2003) also utilized a quantile regression model to map the potential extent of sole in the English Channel and southern North Sea. Machine learning models such as boosted regression trees and random forests have been used to examine species biomass, diversity and distribution for a number of different species (Pittman et al. 2009, Knuby et al. 2010 and Compton et al. 2012). The results from these studies highlight the importance of using techniques that can accommodate non-linear fish habitat relationships.

Most studies that have investigated fish habitat relationships have focused on predicting a species’ distribution (presence / absence) or the average abundance or density, neither of which can be easily manipulated to predict carrying capacity. Also, many of these studies focus on only one or two measures of habitat. Sweka and Mackey (2010) did estimate carrying capacity of Atlantic salmon parr using a quantile regression approach, but the only habitat covariate they included was cumulative drainage area. Traditionally, carrying capacity for salmonids has been estimated through stock-recruitment curves. However, this requires a long time-series of data with variety in the number of spawners which is not usually available (Cramer and Ackerman, 2009), and it often fails to account for habitat covariates. This manuscript incorporates multiple measures of stream habitat to estimate fish-habitat relationships that encompass the collinear nature of most stream habitat metrics and can be used to predict carrying capacity. Our approach moves across several spatial scales, inferring fish-habitat relationships from detailed, localized habitat data and extrapolating capacity predictions across wide swaths of unsampled locations.

In fisheries it has long been recognized that that biotic and abiotic factors limit productivity within and across life-stages. For the purposes of this paper, we define carrying capacity as the maximum number of individuals that can be supported given the quantity and quality of habitat available at a given life-stage. We assume that higher observed relative densities within a given life stage are a function of habitat quantity and quality. Furthermore, we assert that observed fish density is a poor predictor of habitat capacity owing to both a paucity of individuals, especially for threatened or endangered species, and the existence of unmeasured variables that may serve to limit capacity. To address this, we have developed a model to estimate juvenile rearing capacity for Pacific salmon in wadeable streams based on quantile regression forest (QRF; Meinshausen 2006) models using measurements of fish abundance (and density) and habitat characteristics. QRF models combine the theory and justification of quantile regression modeling (Koenker and Bassett 1978; Cade and Noon 2003) with the flexibility and framework of random forest models (Breiman 2001). They account for unmeasured variables and can be used to describe the entire distribution of predicted fish densities for a given set of habitat conditions, not just the mean expected density. Random forest models have been shown to outperform more standard parametric models in predicting fish-habitat relationships in other contexts (Knudby et al. 2010). Quantile regression forests share many of the benefits of random forest models, such as the ability to capture non-linear relationships between independent and dependent variables, naturally incorporate interactions between covariates, and work with untransformed data while being robust to outliers (Prassad et al. 2006). Quantile regression models have been used in a variety of ecological systems to estimate the effect of limiting factors (Terrell et al. 1996; Cade and Noon 2003).

The fish abundance/density and habitat data used to fit the QRF model presented here were available from nine watersheds within the interior Columbia River Basin (CRB), Pacific Northwest, USA. Within the interior CRB two major runs of Chinook salmon occur, stream-type (i.e., spring/summer run) and ocean-type (i.e., fall run), each characterized by different life history characteristics. Stream-type Chinook salmon enter freshwater earlier in the year, spawn in the upper reaches of a watershed, and the juveniles rear for up to 16 months in the freshwater before entering the ocean as smolts. Ocean-type Chinook salmon enter freshwater later (e.g. fall or winter) spawn lower in the watershed, and the juveniles may only spend between several weeks and six months in freshwater before migrating to the ocean. Here we focus on juvenile stream-type Chinook, in particular the summer rearing period during low flow. Data presented here are from Chinook salmon populations in the Upper Columbia River spring-run and Snake River spring/summer-run Evolutionary Significant Units (ESU). The Upper Columbia Spring-run ESU is listed as endangered under the Endangered Species Act, the Snake River Spring/Summer-run is listed as threatened.

at the site scale We then extrapolated to larger spatial scales and validated our estimates of capacity with those from spawner recruit curves from a number of watersheds around the interior Columbia basin. Moreover, the model provides managers a framework to guide the identification, prioritization, and development of habitat enhancement actions to recover salmon populations.

# Methods

## Study Site

Fish and habitat data used in our study were collected from nine watersheds within the interior Columbia River Basin (CRB), Pacific Northwest, USA (Figure 1). The CRB covers greater than 668,000 km2 and drains large portions of Idaho, Oregon, and Washington, and smaller portions of Montana, Nevada, Utah, and Wyoming, as well as the southeastern portion of British Columbia. The habitat data used to populate the QRF model were collected by the Columbia Habitat Monitoring Program (CHaMP; ISEMP/CHaMP 2017) and were downloaded from <https://www.champmonitoring.org>. Data from the following nine CHaMP watersheds were used in this study: Entiat, Grande Ronde (upper), John Day, Lemhi, Methow, Minam (tributary of Grande Ronde), Secesh, Tucannon, and Wenatchee. Juvenile density and abundance data collected at CHaMP survey reaches were graciously provided by a number of projects, including the Integrated Status and Effectiveness Monitoring Project (ISEMP).

## Data

CHaMP sites are 200 m to 500 m reaches within wadeable streams across select basins within the interior Columbia River Basin (CRB) selected based on a spatially balanced Generalized Random Tesselation Stratified (GRTS) sample selection algorithm (Stevens and Olsen 1999, 2004). Habitat data within CHaMP sites are collected using the CHaMP protocol (CHaMP 2016) which calls for field data collection during the low-flow period, typically from June through October. CHaMP habitat data include, but are not limited to, measurements describing channel complexity, channel units, disturbance, fish cover, large woody debris, riparian cover, size (depth, width, discharge), substrate, temperature, and water quality.

Juvenile fish surveys were conducted for sp/sum Chinook salmon parr during the summer low-flow season at many of the same sites surveyed using the CHaMP protocol. Survey methods included mark-recapture, three-pass removal sampling, two-pass removal sampling, and single-pass electrofishing, as well as snorkeling. These data were used to estimate sp/sum Chinook salmon parr abundance at all CHaMP sites where fish survey data were available. Three-pass removal estimates used the Carle-Strub estimator (Carle and Strub 1978), following advice from Hedger et al. (2013). Two-pass removal estimates used the estimator described by Seber (2002). Mark-recapture estimates used Chapman’s modified Lincoln-Peterson estimator (Chapman 1951) and were deemed valid if they met the criteria described in Robson and Regier (1964). These estimates were made using the *removal* function from the *FSA* package (Ogle 2017) or the *closedp.bc* function from the *Rcapture* package (Rivest and Baillargeon 2014) in R software (R Core Team 2015). Snorkel counts were transformed to abundance estimates using paired snorkel-electrofishing sites to calibrate snorkel counts. For sites with invalid estimates or that were sampled with a single electrofishing pass, we developed an estimate of capture probability based on valid estimates, using a binomial generalized linear model that was developed separately for each basin and sampling entity. Possible covariates included the number of fish captured on the first pass, year, site length, and Julian day. After fitting all possible models with those covariates to data with valid abundance estimates, the model with the lowest AICc for each basin and sampling entity was chosen and used to predict abundances based on the number of fish caught on the first pass and any other covariates.

Abundance estimates at all sites were then translated into linear (parr/m) fish densities and density estimates were paired with the associated CHaMP habitat data. For sites that were sampled in multiple years, only the fish and habitat data from the year with the highest observed fish density was retained to avoid possible pseudo-replication, while remaining consistent with our goal of estimating carrying capacity. After removing duplicate samples, our initial dataset contained 218 unique sites with paired fish-habitat data (Table 1).

## Habitat Covariate Selection

A key step in developing a QRF model to predict fish capacities is selecting the habitat covariates to include in the model. Random forest models naturally incorporate interactions between correlated covariates, which is essential since nearly all habitat variables are considered correlated to one degree or another. However, we aimed to avoid overly redundant variables (i.e., variables that measure similar aspects of the habitat). Further, including too many covariates can result in overfitting of the model (e.g., including as many covariates as data points).

The CHaMP protocol produces more than 100 metrics describing the quantity and quality of fish habitat for each survey site. To assist in determining the habitat metrics to include in the QRF model, we used the Maximal Information-Based Nonparametric Exploration (MINE) class of statistics (Reshef et al. 2011) to determine those habitat characteristics (covariates) most highly associated with observed parr densities. We calculated the maximal information coefficient (MIC), using the R package *minerva* (Albanese et al. 2013), to measure the strength of the linear or non-linear association between two variables (Reshef et al. 2011). The MIC value between each of the measured habitat characteristics and the response variable, parr density (fish/m), was used to inform decisions on which habitat covariates to include in the QRF parr capacity model.

Habitat metrics were first grouped into broad categories that included channel unit, complexity, cover, disturbance, riparian, size, substrate, temperature, water quality, and woody debris. Habitat metrics measuring volume and area were scaled to the wetted area of each site. Within each category, metrics were ranked according to their MIC value (Figure 1). Our strategy was to select one or two variables with the highest MIC score within each category so that covariates describe different aspects of rearing habitat (e.g., substrate, temperature, etc.). Additionally, we attempted to avoid covariates that were highly correlated (Figure 2) while including covariates that can be directly influenced by restoration actions or have been shown to impact salmonid juvenile density.

## QRF Model Fit

Using the selected habitat covariates (Table 2), we fit a QRF model to predict habitat rearing capacity for spring/summer Chinook salmon parr, during summer months. QRF models combine the flexibility of random forest models (Breiman 2001) with the ability of quantile regression to extract relationships between quantiles of the data other than the mean (Cade and Noon 2003). Random forests also account for non-linear relationships between the response and predictor variables, and naturally incorporate interactions between the predictor variables, two common features of ecological datasets (Liaw and Wiener 2002). After constructing a random forest, predictions of the mean response can be made by averaging the predictions of all trees, similar to the expected value predictions from a statistical regression model. However, the individual predictions from each tree, viewed collectively, describe the entire distribution of the predicted response. Therefore, the random forest model can be used in the same way as other quantile regression methods to predict any quantile of the response. We fit the QRF models using the *quantregForest* function from the *quantregForest* package (Meinshausen 2016) in R software (R Core Team 2015). The 90th quantile of the predicted distribution was used as a proxy for carrying capacity, following the suggestion of Sweka and Mackey (2010). One reason for the 90th quantile, instead of something higher, is to avoid using predictions that are aimed at the very upper tails of observed fish density, where the variability of predictions may be influenced by sample size issues.

Summer parr density and habitat data were paired up by site and year. There were some missing values in the habitat dataset. Any site visit with more than three missing covariates was removed; the remaining missing values were imputed using the *missForest* R package (Stekhoven and Buehlmann 2012; Stekhoven 2013). Ultimately, the spring/summer Chinook parr capacity QRF model was fit using 186 records (paired fish-habitat data) and 14 habitat covariates (13.3 data points per covariate) (Table 2).

After model fitting, the QRF model was then be used to predict capacity using measurements of the habitat covariates that were used to fit the model. In our case, this includes all sites within CHaMP basins in the interior Columbia River basin. For CHaMP sites that have been sampled in multiple years, we first calculated the mean for each habitat metric among years to make predictions. In total, we generated 485 predictions of spring/summer Chinook salmon parr capacity, during summer months, for the following basins: Entiat, Grande Ronde (including Minam), John Day, Lemhi, Methow, Secesh, Tucannon, Yankee Fork and Wenatchee.

## Model Extrapolation

Using the QRF model, we made predictions of habitat capacity for juvenile rearing at all CHaMP sites within the interior Columbia River basin. However, fisheries biologists are more often interested in capacity at larger scales (e.g., watershed, population). To accomplish this, we developed two extrapolation models (linear and areal capacity) based on globally available attributes (GAA) to scale our CHaMP site predictions to a series of larger spatial scales; often corresponding to the entirety of tributary habitat utilized by a given population. The GAA data used here was available from the list of GRTS master sample sites that the CHaMP sites were originally selected from. Possible covariates included temperature range, growing degree days, an index of disturbance, the square root of cumulative drainage area, stream power, slope, channel type and watershed. The natural log of the CHaMP site capacity predictions was used as the response variable in a multiple linear regression model that incorporated the design weights of the CHaMP sites using the *svyglm* function from the *survey* package (Lumley 2004, 2016) in R software (R Core Team 2015). Models with all possible combination of GAAs were fit and all the models with cumulative Akaike weights less than 0.95 were model averaged to make predictions of capacity at all master sample sites throughout the interior CRB, generally spaced about one kilometer apart. For sites outside of CHaMP watersheds, only extrapolation models that did not include watershed as a covariate were model averaged to make predictions. Summaries of extrapolation model fit are shown in Table 4.

To summarize capacity at larger scales, the mean linear capacity (e.g., fish/m) of the master sample points along a particular tributary is multiplied by the length of that tributary. We first restricted the master sample points and lengths of streams to those with the domain of spring/summer Chinook, as defined by StreamNet <http://www.streamnet.org> or using expert opinion from local biologists. The capacities of various tributaries could then be summed to estimate capacity at almost any spatial scale. However, for visualization, predictions of areal capacity (fish/m2) were used.

## Model Validation

Spawner-recruit data from several watersheds within the interior CRB were compiled to validate the extrapolated QRF estimates of spring/summer Chinook salmon parr capacity. Some watersheds had direct estimates of parr, while some had estimates of smolts and pre-smolts from rotary screw traps. Using estimates of over-winter survival, estimates of parr were calculated from those watersheds (Table 5). A series of spawner-recruit functions were then fit to this data including Beverton-Hold, Ricker, and hockey stick. Estimates of capacity were made from each of these spawner-recruit curves and compared with QRF estimates of capacity (Table 5, Figure 4).

# Results

## Habitat Covariate Selection

We categorized 145 habitat measurements collected using the CHaMP habitat protocol (CHaMP 2016) into 11 habitat groups, and for each habitat covariate an MIC value was calculated based on the strength of association between the habitat covariate and the response variable, parr density (fish/m). We chose the following 14 CHaMP habitat covariates to fit the QRF model: channel unit frequency, thalweg depth CV, average wetted width to depth ratio, total fish cover, disturbance index, riparian ground cover, average thalweg depth, percent gravel, average substrate size (D50), percent substrate < 6mm, average hourly temperatures in the summer, maximum weekly average temperature, conductivity and frequency of large wood in the wetted channel (Table 2).

## QRF Model

A QRF model was fit using 14 habitat covariates (Table 2) and the *quantregForest* package (Meinshausen 2016) in R (R Core Team 2015) and the 90th quantile of the predicted distribution was used as a proxy for carrying capacity. After model fit, we examined the relative importance of each habitat covariate included in the model (Figure 3). Moreover, QRF models allow one to visually examine the marginal effect of each habitat covariate on the quantile of interest using partial dependence plots (PDP). These plots show the marginal effect of changing a single habitat covariate while maintaining all other covariates at their mean values (Figure 4). However, given that many habitat metrics are somewhat correlated, these marginal effects are not independent of one another making the interpretation of partial dependence plots less straightforward.

After model fitting, the QRF model was used to predict habitat capacity at all CHaMP sites within the interior Columbia River basin.

## Model Extrapolation

Text…

## Model Validation

Text…

# Discussion

Using a quantile regression forest model, we predicted the summer rearing capacity for spring/summer Chinook parr at the reach scale. Quantile regression forests combine the benefits of random forest analyses (e.g. non-linear responses, correlated covariates) with quantile regression, which allows us to estimate how fish capacity, and not just observed fish density, changes with stream habitat characteristics. We then extrapolated those reach-scale estimates to watershed-scale estimates of capacity, which are more in line with the spatial scale at which management actions take place.

## Biological Expectations from QRF model

The results of the QRF parr capacity model for spring/summer Chinook salmon meet many biological expectations. Discussion of biological expectations and PDP plots…

## Model Assumptions

Model assumptions should always be considered when interpreting results; the following assumptions were made when modeling habitat capacity using QRF methods. First, we assume that sites with higher carrying capacity (i.e., ‘better’ habitat) contain more parr during fish surveys. Second, we assume that carrying capacity is a function of habitat. Third, we assume that estimates of parr densities from fish surveys are unbiased. Fourth, we assume that at least some sites surveys in our dataset are at or close to carrying capacity at the site level. However, if this fourth assumption is not met, we will obtain a conservative estimate of carrying capacity (but the framework of the model is not wrong). Moreover, we assume it is other unmeasured factors (not habitat) that are preventing a site from displaying fish densities near carrying capacity (e.g., insufficient adult spawners, competition, predation, etc.). Finally, we assume that the 90th quantile of predictions at each CHaMP site is a reasonable proxy for carrying capacity. With insufficient data, it is difficult to get a well-defined estimate of higher quantiles (i.e., lots of data is necessary to define the tails of a distribution).

Assumptions we don’t need to make. We make no assumptions regarding the shape of fish-habitat relationships. Rather fish-habitat relationships are estimated from the data. Relationships may be positive or negative, linear or non-linear; regardless, no constraints are put on the relationships. Moreover, we make no assumptions regarding correlations among habitat covariates; the random forest framework accounts for possible interactions between habitat metrics, so we don’t need to explicitly place them in the model.

## Extrapolation Model

Text.

## Improving Habitat Data

Discussion of improving habitat data to possibly avoid extrapolation. Collect cheaper/better spatially continuous habitat data.

## Conclusions and Next Steps

Text…

# Acknowledgements

Text…

# References

Albanese, D., M. Filosi, R. Visintainer, G. Jurman, and C. Furlanello. 2012. Minerva and minepy: a C engine for the MINE suite and its R, Python, and MATLAB wrappers. Bioinformatics. 29(3):407-408.

Breiman, L. 2001. Random forests. Machine Learning. 45(1):5-32.

Cade, B.S. and B.R. Noon. 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment. 1:412-420.

Carle, F.L. and M.R. Strub. 1978. A new method for estimating population size from removal data. Biometrics. 34:621-630.

CHaMP (Columbia Habitat Monitoring Program). 2016. Scientific protocol for salmonid habitat surveys within the Columbia Habitat Monitoring Program. Prepared by CHaMP for the Bonneville Power Administration.

Chapman, D.G. 1951. Some properties of the hypergeometric distribution with applications to zoological sample censuses. University of California Press.

Good, T.P., R.S. Waples, and P. Adams (editors). 2005. Updated status of federally listed ESUs of West Coast salmon and steelhead. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-66, 598 p.

Hedger, R.D., E. De Eyto, M. Dillane., O.H. Diserud, K. Hindar, P. McGinnity, R. Poole, and G. Rogan. 2013. Improving abundance estimates from electrofishing removal sampling. Fisheries Research. 137:104-115.

ISEMP/CHaMP. 2017. Integrated Status and Effectiveness Monitoring Program (ISEMP) and Columbia Habitat Monitoring Program (CHaMP) Annual Combined Technical Report, January – December 2016, BPA Projects 2003-017-00 and 2011-006-00, 03 Electronic Pages.

Knudby, A., A. Brenning, and E. LeDrew. 2010. New approaches to modeling fish-habitat relationships. Ecological Modeling. 221(3):503-511.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news 2:18-22.

Lumley, T. 2016. survey: analysis of complex survey samples. R package version 3.31-5.

Lumley, T. 2004. Analysis of complex survey samples. Journal of Statistical Software. 9(1):1-19.

Meinshausen, N. 2016. quantregForest: Quantile Regression Forests. R package version 1.3-5. <https://CRAN.R-project.org/package=quantregForest>

Meinshausen, N. 2006. Quantile regression forests. Journal of Machine Learning Research. 7:983-999.

Ogle, D.H. 2017. FSA: Fisheries Stock Analysis: R package version 0.8.17.

Prasad, A., L. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystem. 9(2):181-199.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reshef, D.N., Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenbacher, and P.C. Sabeti. 2011. Detecting Novel Associations in Large Data Sets. Science. 334:1518-1524.

Rivest, L.P. and S. Baillargeon. 2014. Rcapture: Loglinear models for capture-recapture experiments. R package version 1.4-2.

Robson, D. and H. Regier. 1964. Sample size in Peterson mark-recapture experiments. Transactions of the American Fisheries Society. 93:215-226.

Seber, G. 2002. The estimation of animal abundance and related parameters. Blackburn Press Caldwell, New Jersey.

Stekhoven, D.J. 2013. missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

Stekhoven, D.J. and P. Buehlmann. 2012. MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics. 28(1):112-118.

Stevens, D.L., Jr. and A.R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association. 99(465):262-278.

Stevens, D.L., Jr. and A.R. Olsen. 1999. Spatially restricted surveys over time for aquatic resources. Journal of Agricultural, Biological, and Environmental Statistics. 4:415-428.

Terrell, J.W., B.S. Cade, J. Carpenter, and J.M. Thompson. 1996. Modeling Stream Fish Habitat Limitations from Wedge-Shaped Patterns of Variation in Standing Stock. Transactions of the American Fisheries Society. 125(1):104-117.

Whittier, T., A. Herlihy, C. Jordan, and C. Volk. 2011. Landscape classification of Pacific Northwest hydrologic units based on natural features and human disturbance to support salmonid research and management. NOAA, National Marine Fisheries Service. NOAA Contract #AB1133F10SE2464. Pp. 39.

# Tables

Table 1. The number of unique sites in the initial dataset, by watershed, with paired fish-habitat data used to populate the spring/summer Chinook salmon parr capacity model.

|  |  |
| --- | --- |
| **CHaMP Basin** | **n** |
| Entiat | 36 |
| John Day | 26 |
| Lemhi | 18 |
| Minam | 9 |
| South Fork Salmon | 26 |
| Upper Grande Ronde | 84 |
| Wenatchee | 19 |
| **Total:** | **218** |

Table 2. Habitat metrics and descriptions of metrics included in the QRF model to predict spring/summer Chinook salmon parr capacity (during summer months). Metrics are ranked in order of relative importance.

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Metric | Metric Category | Description |
| 1 | Cumulative Drainage Area | Size | The cumulative land area that drains to a given location/site. |
| 2 | Wetted Width Integrated | Size | Average width of the wetted polygon for a site. |
| 3 | Wetted Depth SD | Complexity | Standard Deviation of water depths within the wetted channel. |
| 4 | Conductivity | Water Quality | Measure of concentration of ionized materials in water, or the ability of water to conduct electrical current. |
| 5 | Summer Hourly Average Temp | Temperature | Average of all hourly temperature measurements collected July 15th - August 31st. |
| 6 | Wetted Width to Depth Ratio Avg | Complexity | Average width to depth ratio of the wetted channel measured from cross-sections. Depths represent an average of depths along each cross-section. |
| 7 | Fish Cover: Total | Cover | Percent of wetted area with the following types of cover: aquatic vegetation, artificial, woody debris, and terrestrial vegetation. |
| 8 | Channel Unit Frequency | Channel Unit | Total number of channel units per 100 meters. |
| 9 | Slow Water Percent | Channel Unit | Number of Slow Water/Pool channel units per 100 meters. |
| 10 | Riparian Cover: No Canopy | Riparian | Percent of riparian canopy devoid of vegetation. |
| 11 | Substrate: D50 | Substrate | Diameter of the 50th percentile particle derived from pebble counts. |
| 12 | Disturbance Index | Disturbance | Disturbance index that includes measures of % urban, % agricultural, % impervious surface, and road density (Whittier et al. 2011). |
| 13 | Substrate: <2 mm | Substrate | Average percentage of pool tail substrates comprised of fine sediment <2 mm. |
| 14 | Large Wood Frequency: Wetted | Wood | Number of large wood pieces per 100 meters within the wetted channel. |

Table 3. Globally available attribute (GAA) habitat covariates used to extrapolate quantile regression forest (QRF) model predictions of spring/summer Chinook parr capacity to a larger scale (e.g., watershed, population). An ‘X’ denotes the GAA was used in the given extrapolation model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covariate** | **Scale** | **Units** | **CHaMP (per m)** | **non-CHaMP (per m)** | **CHaMP (per m2)** | **non-CHaMP (per m2)** |
| Mean Annual Velocity | Reach (2 km) | fps | - | X | X | X |
| Slope | Reach (2 km) | m/m | - | - | X | X |
| Drainage Area (sqrt) | Reach (2 km) | km2 | - | X | X | - |
| Bankfull Width – modeled | Site (300 m) | m | X | X | - | - |
| Stream Power | Reach (2 km) | - | - | - | - | - |
| Bankfull Width | Site (300 m) | m | X | - | X | X |
| Channel Type | Site (300 m) | - | X | X | - | X |
| Temperature Range | Reach (2 km) | °C | - | - | - | X |
| Growing Degree Day | Reach (2 km) | GDU | - | X | X | X |
| Precipitation | Reach (2 km) | cm | - | - | - | X |
| Elevation | Site (300 m) | m | X | X | X | X |
| CHaMP Watershed | Region | - | X | - | X | - |
| Disturbance Class PCA 1 | Watershed (HUC12) | - | X | X | - | X |
| Natural Class PCA 1 | Watershed (HUC12) | - | - | X | X | X |
| Natural Class PCA 2 | Watershed (HUC12) | - | X | - | - | - |

Table 4. Summary of model fit for each of the extrapolation models to extrapolate quantile regression forest (QRF) model predictions of spring/summer Chinook parr capacity to larger scales (e.g., watershed, population).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Response** | **R2** | **Adjusted R2** | |
| CHaMP | fish/m | 0.493 | 0.466 |
| non-CHaMP | fish/m | 0.398 | 0.374 |
| CHaMP | fish/m2 | 0.458 | 0.434 |
| non-CHaMP | fish/m2 | 0.407 | 0.382 |

Table 5. Estimates of parr capacity from both spawner-recruit data (Beverton-Holt, Ricker, hockey stick) and from extrapolated estimates of parr capacity from the quantile regression forest (QRF) model. Numbers in parentheses are blank. ODFW = Oregon Department of Fish and Wildlife; WDFW = Washington Department of Fish and Wildlife; IDFG = Idaho Department of Fish and Game; RST = Rotary Screw Trap; QCI = Quantitative Consultants, Inc.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Watershed** | **n Years** | **Adult Data** | **Adult Source** | **Parr Data** | **Parr Source** | **Beverton-Holt** | **Ricker** | **Hockey Stick** | **QRF** |
| Catherine Creek | 20 | Spawners | ODFW | RST | ODFW | 135,387 (0.269) | 103,021 (0.141) | 99,921 (0.210) | 179,004 (0.039) |
| Chiwawa River | 20 | Spawners | WDFW | Parr Surveys | BioAnalysts | 248,586 (0.240) | 166,139 (0.148) | 174,216 (0.184) | 341,352 (0.059) |
| Grand Ronde River (upper mainstem) | 8 | Spawners | ODFW | RST | ODFW | 171,609 (0.388) | 168,137 (0.298) | 127,052 (0.317) | 212,794 (0.045) |
| Hayden Creek | 8 | Redds | IDFG | RST | IDFG/QCI | 58,493 (0.726) | 37,235 (0.371) | 41,668 (0.535) | 57,651 (0.063) |
| Minam River | 14 | Spawners | NA | RST | NA | NA (NA) | 484,810 (1.444) | 662,806 (1.726) | 262,348 (0.049) |
| Tucannon River | 27 | Redds | WDFW | RST | WDFW | NA (NA) | NA (NA) | 152,809 (0.165) | 380,301 (0.029) |
| upper Lemhi River | 22 | Redds | IDFG | RST | IDFG/QCI | 112,746 (0.784) | 64,303 (0.503) | 72,470 (0.638) | 150,706 (0.061) |

# Figure Captions

Figure 1. Map of the Columbia River Basin showing each of the Columbia Habitat Monitoring Program (CHaMP) watersheds.

Figure 2. Barplot showing the strength of association between each habitat metric and observed densities (fish/m2) of spring/summer Chinook salmon parr, during the summer, facetted by habitat categories. Results were used to determine which the habitat covariates to include in the quantile regression forest (QRF) model to predict carrying capacity.

Figure 3. Pairwise correlations among habitat covariates used to fit the spring/summer Chinook salmon parr capacity quantile regression forest (QRF) model.

Figure 4. Relative importance of each habitat covariate included in the quantile regression forest (QRF) model to predict habitat capacity, during summer months, for spring/summer Chinook salmon parr.

Figure 5. Partial dependence plots for the spring/summer Chinook salmon parr capacity quantile regression forest (QRF) model, depicting how parr capacity shifts as each habitat metric changes, assuming all other habitat metrics remain at their mean values. Tick marks along the X-axis depict observed values, and the sub-basin they came from.

Figure 6. Spawner-recruit data from seven watersheds. Solid lines are the spawner-recruit curve, dashed lines are the estimated capacity, and shaded polygons depict the 95% confidence intervals of capacity. Blue corresponds to Beverton-Holt models, green to Ricker models, yellow to hockey stick models, and red to QRF estimates. The QRF solid curve is a Beverton-Holt model with the capacity parameter fixed to the QRF estimate of capacity.

# Figures

Figure 1

Figure 2

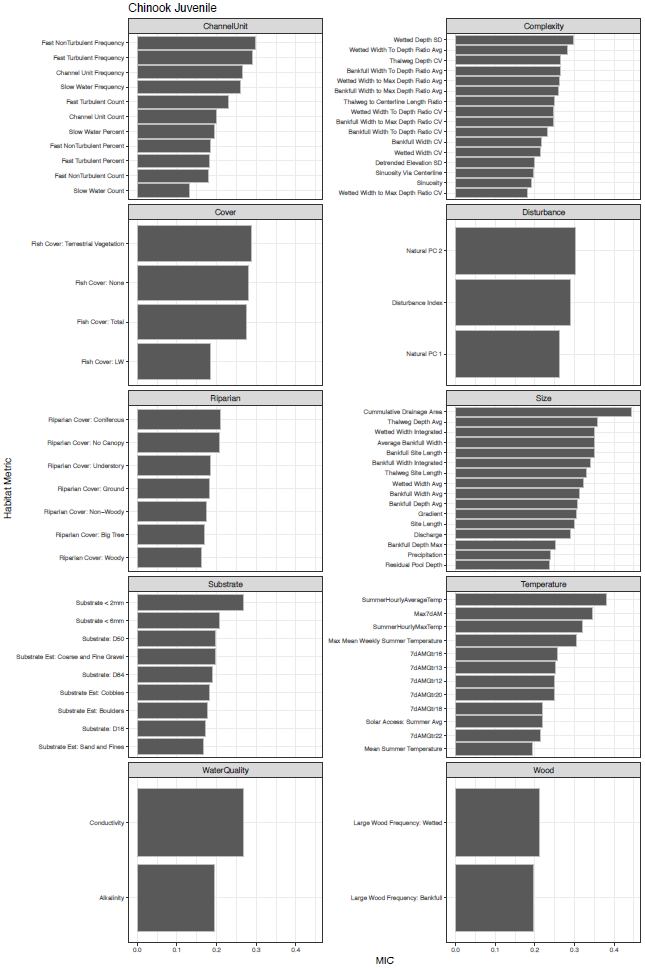


Figure 3

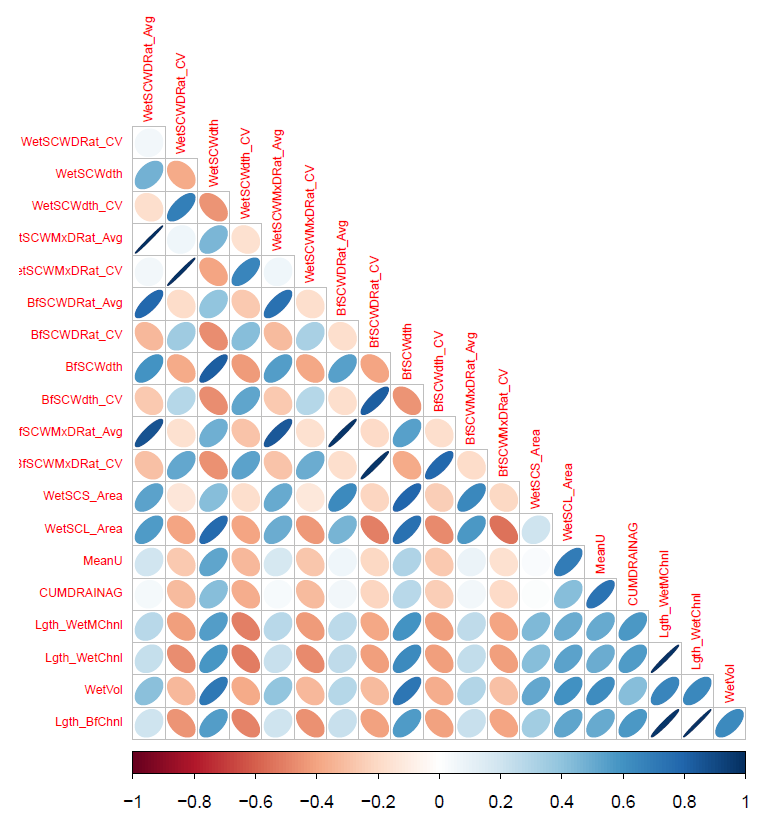


Figure 3

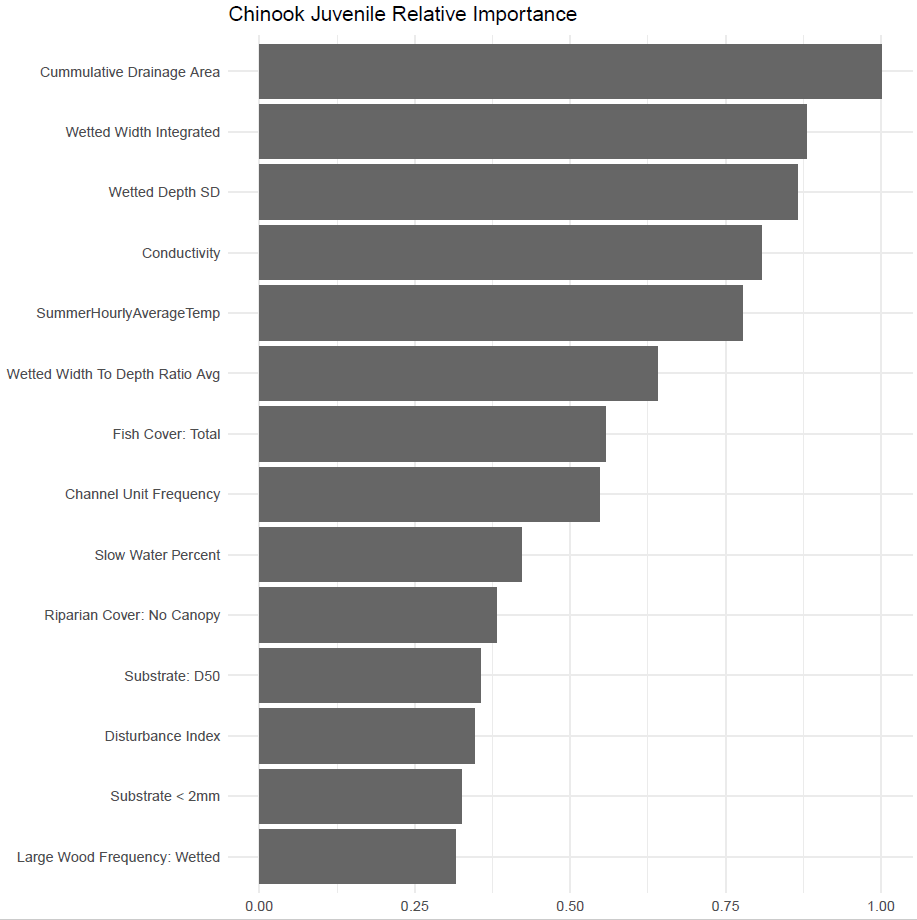


Figure 4

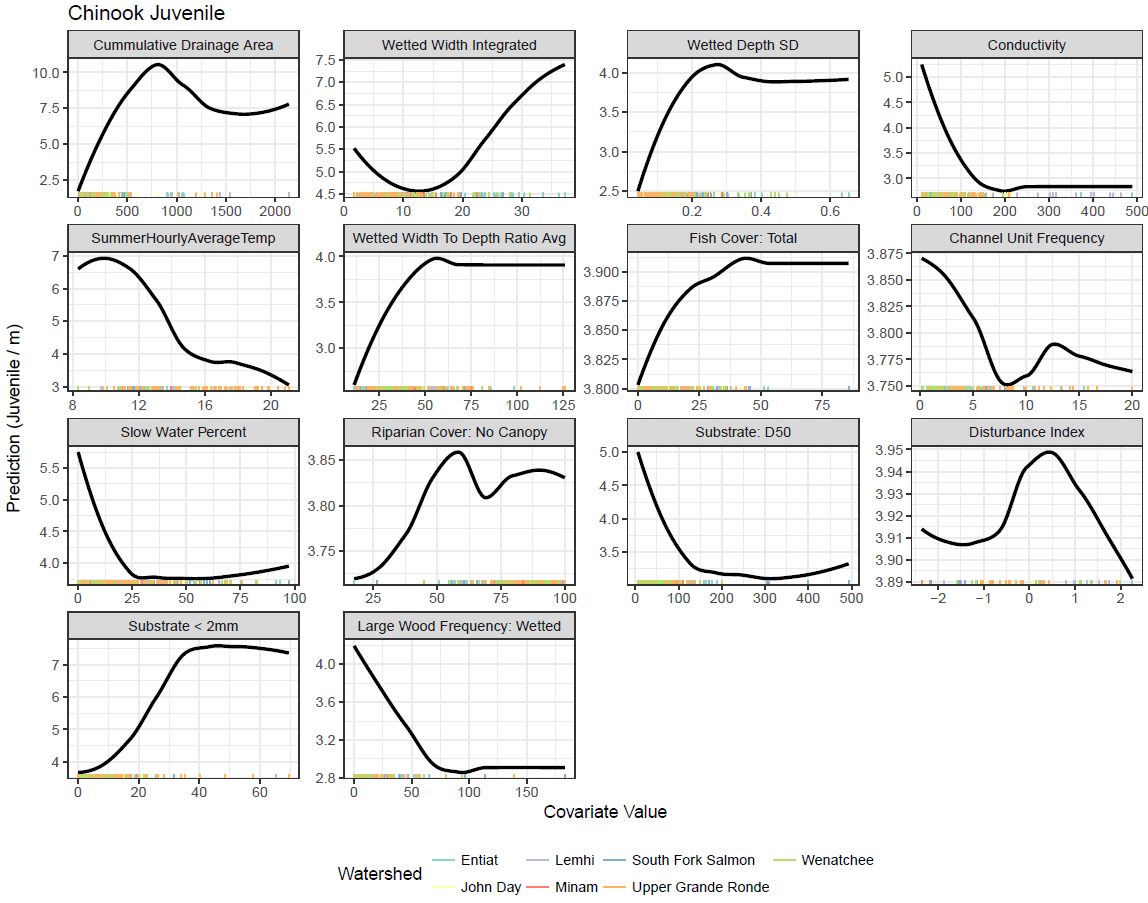


Figure 5

