**CHaMP Metric Extrapoloation**

Matt Nahorniak, South Fork Research

December 2017

**Introduction**

Fitting models of CHaMP metrics on globally available attributes (GAA) is being done to estimate CHaMP metrics continuously at every point along stream networks of interest. Model based estimates for selected CHaMP metrics are being made both within current CHaMP watersheds, as well to several watersheds where CHaMP data is not being taken (including the South Fork Clearwater, Lower Clearwater, Lolo, and Lochsa). Because we are estimating CHaMP metrics outside of the sampled CHaMP domain, we refer to this exercise as “extrapolation”, although for unsampled sites within the population frame of sampled CHaMP watersheds, a better terms might be “interpolation” or perhaps “imputation”. Nevertheless, we use only the single term “extrapolation” for simplicity.

CHaMP metrics are measured at, typically, 45 points along a stream network within each CHaMP watershed, in a sampling design that includes both annual and 3 year rotating panel sites. The sampling design is optimized for estimation of the distribution (including mean and variance) of CHaMP metrics at the spatial scale of the portions of watersheds accessible to anadromous Chinook or Steelhead. However, importance is also being placed on making spatially continuous estimates of CHaMP metrics at all points along stream networks of interest, throughout the Interior Columbia River Basin (ICRB), rather than simply at CHaMP watershed spatial levels. Statistical models enable generation of maps showing estimated CHaMP distributions in a spatially continuous manner, providing an additional tool for habitat management. In addition, modeled estimates can be used to augment design based estimates, where limited or no sampling data exists, in order to provide inputs to habitat and salmonid life cycle models.

For successful, informative models capable of both predicting CHaMP metrics within CHaMP watersheds and extrapolating CHaMP metrics into non-CHaMP watersheds, a set of GAAs available must be statistically informative to the CHaMP metric being validated. In other words, there must be quantifiable relationships between our GAAs and CHaMP metrics. This is examined through cross validation at various spatial levels.

CHaMP metrics included in extrapolation efforts are shown in Table 1.

**Table 1. CHaMP metrics included in extrapolation process**



**Methods**

Simple multiple regression models have been constructed relating our CHaMP metrics of interest to available globally available attributes (GAA). Because CHaMP data are from a non-uniform probability GRTS [1] sampling designs, it is necessary to take into account sample inclusion probabilities in the fitting of regression models [2]. To do this, we use model assisted regression [3], utilizing the R function *svyglm* from the package *survey* [4]. Sample inclusion probabilities were calculated separately as part of the annual design based estimates of CHaMP metric status and trends for 2011-2016 data. 2017 data were not included in this set of extrapolations (full 2017 data are not yet available as of the time of this report).

GAAs used to model each CHaMP metric (Table 2) were selected using a combination of forward and backward stepwise regression with the R functions *add1* and *drop1*. Minimization of AIC was the criterion used for inclusion, and covariates are include in models if they reduce total model AIC by value greater than about 5. Note that because GAAs selected may be correlated with one another, and because we are dealing with observational data, extreme caution should be made in interpreting model coefficients. Our intent is not to infer meaning from model coefficients, but rather to develop predictive models only. Nevertheless, efforts were made to avoid using GAA’s in the same model that are likely highly correlated or have essentially the same meaning.

Plots of measured vs predicted values for each CHaMP model are produced to visually assess model goodness of fit (Figure 1). Inspection of model residuals is performed to determine the need for a transformation of the metric being modeled, or other potential problems with the model fit (heteroscedasticity, non-linearity, etc.). If transformations are deemed necessary, then the metric being modeled is transformed and the entire process of variable selection and model fitting is repeated.

**Figure 1. Example extrapolation model: measured vs predicted juvenile steelhead wetted usable area per meter (HIS\_St\_Juv\_WUA\_per\_m)**

**Model Validation**

Selecting a subset of model covariates (GAA’s) from a large population of potential candidates invariably leads to an over-estimation of model precision if the modeler relies on r-squared or other residual based goodness of fit metrics. Therefore, additional model validation is necessary to accurately assess goodness of fit and estimate the precision of predicted values.

For this assessment we attempted to validate models using two methods of differing spatial scales. First, we perform leave one out cross validation where one site at a time is removed during the model fitting process, and the model fit from the rest of the data is used to predict the remaining point. This is repeated for all points, such that an error (predicted value minus measured value) is obtained for each point. We then compare measured to predicted values at all points (Figures 2 and 3) and observe the distribution of error. The distribution of error is then used to calculate prediction uncertainties and r-squared values, reflective of the likely accuracy of using the models to predict points within currently sampled CHaMP watersheds.

**Figure 2. Cross validation results, measured vs predicted juvenile Steelhead wetted usable area per meter of stream length (R-squared = .76)**



**Figure 3. Cross validation results, measured vs predicted D50 (R-squared = .53)**



Because we also want to use CHaMP to predict outside of CHaMP watersheds, we need an estimates of error levels likely when doing this. The method for this is to again use a leave-one-out cross validation method, except for this we leave data from an entire watershed out during each iteration of the model fitting stage. The fit model is then used to predict results at all measured points within the left out watershed (Figures 4 and 5). Because an entire watershed is left out of the process, we might expect the performance of the model in cross validation to be somewhat less than during cross validation within measured watersheds. This is the level of model performance we expect in non-CHaMP watersheds if we can assume that CHaMP watersheds are representative of non-CHaMP watersheds into which we predict points. Note that CHaMP watersheds were not selected at random from the broader population of Interior Columbia Basin watersheds, so this may not be a valid assumption.

**Figure 5. Watershed level cross validation results for D50 (R-squared = .43)**



**Figure 4. Watershed Level Cross Validation Results for HSI Juvenile Steelhead capacity per meter (R-squared = .52)**



**Spatial Autocorrelation**

Note that in our model based estimation process, we do not incorporate spatial autocorrelation correction models. CHaMP sampling is performed using a GRTS [1] design, which generates spatially balanced samples. While these spatially balanced samples generally provide for greater efficiency in the estimation population distribution parameters than conventional random sampling, the samples generally do a poor job of estimating spatial autocorrelation, because sampled sites are rarely close together in space. Thus our samples generally do not have the information required to take advantage of spatial autocorrelation when predicting CHaMP metrics at sites near measured CHaMP sites.

**Results**

Table 2 lists the GAA’s selected for each extrapolation model.

**Table 2: Globally available attributes selected for each extrapolation model**



Table 3 lists the cross validation results, giving r-squared estimates for both within CHaMP watersheds (leave one out cross validation) and for points at non-CHaMP watersheds (leave one watershed out cross validation). The habitat capacity metrics (the various HSI metrics and NREI) generally perform well in cross validation. D50 (the median pebble size) performed fairly, as did sinuosity, although sinuosity is likely difficult to predict in non-CHaMP watersheds based on leave one watershed out cross validation. SubLT2, SubLT6, DpthThlwg\_UF\_CV, LWFreq\_BF, PoolResidDpth performed poorly, indicating a poor ability to predict within a CHaMP watershed, and extremely poor predictive abilities into non-CHaMP watersheds. Simply stated, from our available pool of GAA’s, we did not any that were informative to those metrics.

**Table 3. Cross validation results for leave one out cross validation (indicative of extrapolation accuracy within a CHaMP watershed) and leave watershed out cross validation (indicative of extrapolation accuracy for non CHaMP watersheds).**

|  |  |  |
| --- | --- | --- |
| **CHaMP Metric** | **Leave One Out Cross Validation R-squared estimate** | **Leave Watersheds Out Cross Validation** |
| HSI\_CH\_Juv\_WUA\_per\_m | 0.74 | 0.56 |
| HSI\_CH\_Spawn\_WUA\_per\_m | 0.61 | 0.42 |
| HSI\_ST\_Juv\_WUA\_per\_m | 0.76 | 0.52 |
| HSI\_ST\_Spawn\_WUA\_per\_m | 0.57 | 0.41 |
| NREI.dens.est.fpm | 0.83 | 0.53 |
| SIN | 0.23 | 0.24 |
| SLowerWater\_Pct | 0.14 | 0.00 |
| SubD50 | 0.53 | 0.43 |
| SubLT2 | 0.202 | 0.00 |
| SubLT6 | 0.203 | 0.00 |
| DpthThlwg\_UF\_CV | 0.53 | 0.00 |
| LWFreq\_BF | 0.37 | 0.00 |
| PoolResidDpth | 0.26 | 0.15 |

Based on these models, a series of extrapolations have been done to provide nearly continuous (1 km spacing) estimates within CHaMP watersheds and at a set of non-CHaMP watersheds within the ICRB. Contact South Fork Research for access to these files or for information about continuous maps. These files also contain row level estimates of data quality based on the cross validation results in Table 3. Caution should be used based on data quality estimates and user needs. Also included in the extrapolated estimates are columns indicating whether the estimates are part of the steelhead anadromous domain and / or the chinook anadromous domain. Sites that are part of neither domain are also estimated, although the quality of such estimates is highly suspect as these spatial areas were not part of the population or sampling frames from which data used to build these models were gathered. No estimate of data quality is given for these points, and users should be wary of using these estimates.

Note that the extrapolation models may produce values of zero. These are intended to be zeros – they are not NA values. Occasionally the models will produce negative values, but since the metrics in question cannot be negative, these are set to zero. In addition, we are wary of generating values outside of the observed range of CHaMP metric values, so if an extrapolated value is greater than the greatest measured value observed, it is set at the highest observed value. This prevents the occasional wildly high and unreasonable estimate.

**Discussion**

For some CHaMP metrics, particularly the higher level capacity metrics from habitat suitability index (HSI) and net return on energy investment (NREI) models, we have models that we expect to be effective when used to predict values at sites not sampled by CHaMP within the ICRB region of interest. Other models (D50, DpthThlwg\_UF\_CV) show promise but could benefit from further refinement using more informative GAAs. For some models, however, we have little or no ability to predict outside of measured CHaMP sites and require significantly more informative GAAs to yield valid predictions.

**References**

1. Stevens Jr DL, and Olsen AR (2004). Spatially balanced sampling of natural resources. J Am Stat Assoc, 99:465, 262-278, DOI: 0.1198/016214504000000250
2. Nahorniak M, Larsen DP, Volk C, Jordan CE (2015) Using Inverse Probability Bootstrap Sampling to Eliminate Sample Induced Bias in Model Based Analysis of Unequal Probability Samples. PLoS ONE 10(6): e0131765. doi:10.1371/journal.pone.0131765
3. Lumley T (2004). Analysis of complex survey samples. J Stat Softw 9(1): 1-19.
4. Lumley T (2012). Survey: analysis of complex survey samples. R package version 3.28-2.