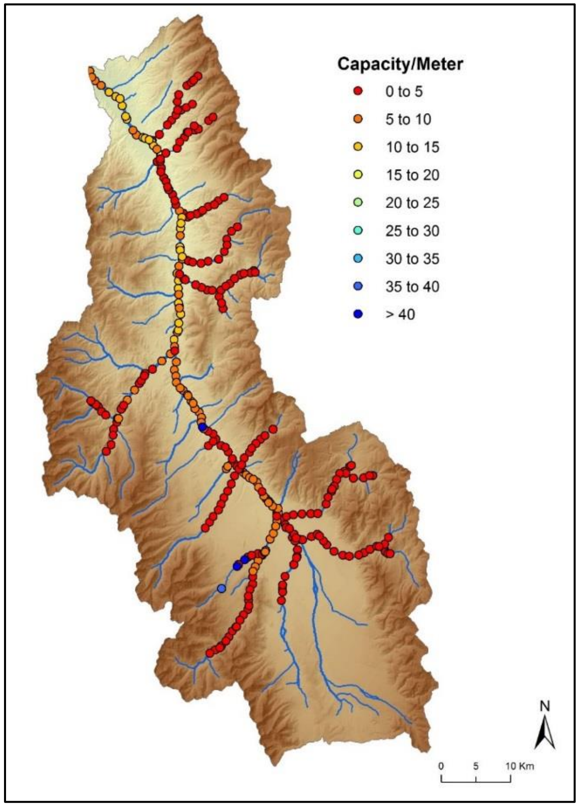
An Introduction to CHaMP Sampling and Data Analysis

Presenter:
Matt Nahorniak

An Introduction to CHaMP Sampling and Data Analysis

- Introduction
- Review of sampling basics
  - Stratified sampling
- Design Based Analysis
- Model Based Analysis
  - Incorporation of sample design in model based analysis
- Example Analyses in R
- Working with CHaMP Statisticians
  - Helping us help you



Introduction

An Introduction to CHaMP Sampling and Analysis

CHaMP data analysis

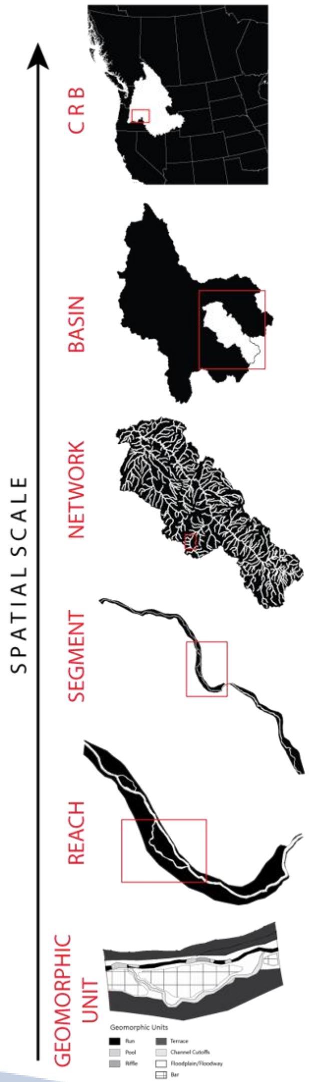
Today's key message:

Sampling design needs to be taken into account during any and all analyses of CHaMP data

Questions for the audience:

- What watershed(s) are of interest to you?
- What are you hoping to estimate?
- What questions are you hoping to answer?
- How else to you hope to use the data?

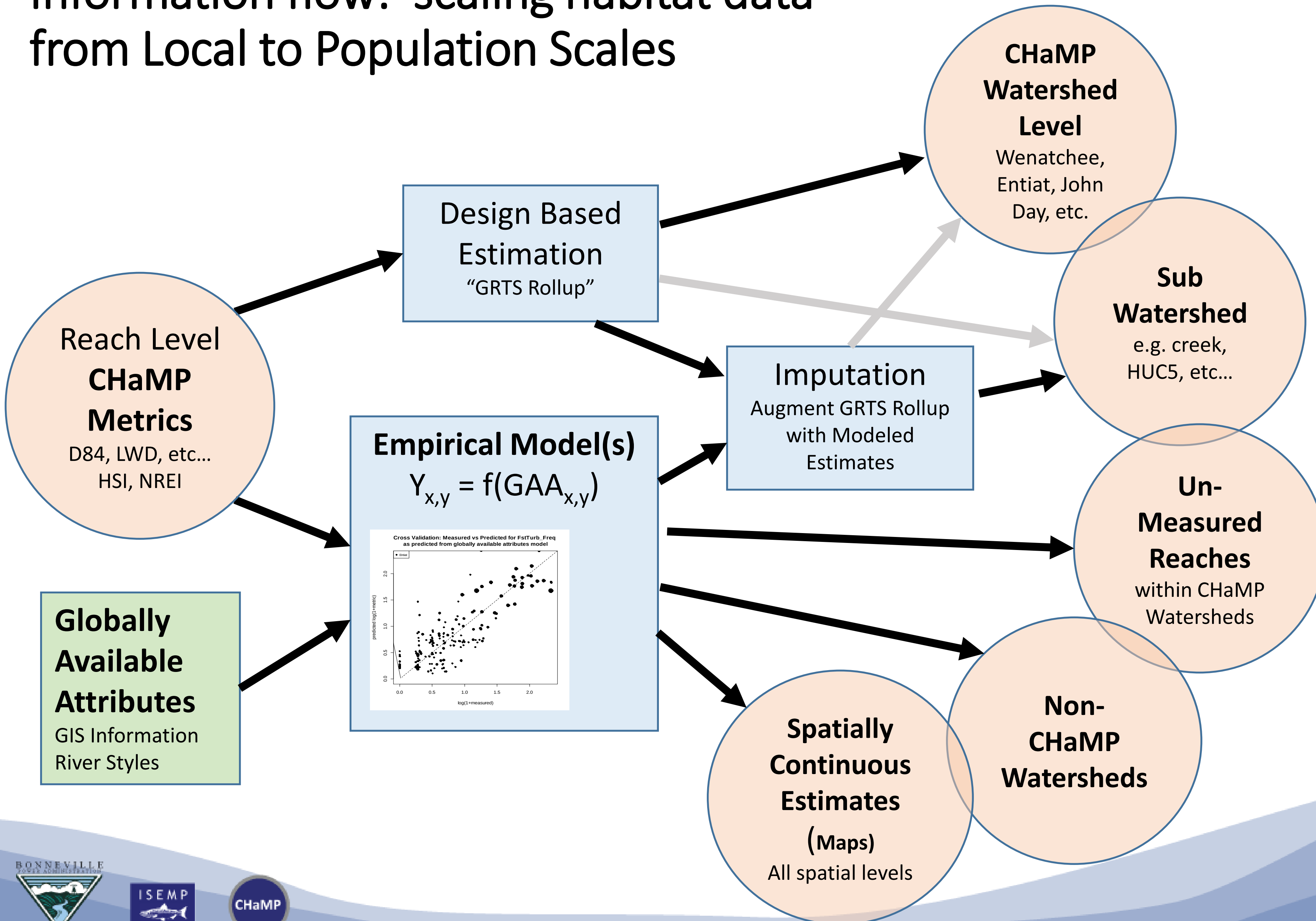
CHaMP Spatial levels of Interest:



- Site (Reach) Level
  - Level of data Collection (CHaMP Sampling)
- Management
  - Reach through CRB
- Fish Biology
  - Geomorphic unit (individual fish)
  - Segment-Network *Populations* of fish
- Data Visualization (Continuous spatial estimates)
  - Network-Basin

BPA has indicated that we need to be able ask (and answer) questions across spatial scales ranging from channel units to the entire Upper Columbia basin.

Information flow: scaling habitat data from Local to Population Scales



Analysis of sampled data: a simple example

An Introduction to CHaMP Sampling and Analysis

I measured 20 sites from a population, as follows:

Question: What is sample average?

Question: What is estimated population average?

Sample Data	
Site #	Value
Am	3
Fa	6
Fb	6
Ak	3
Fd	6
Ef	6
Ch	3
Bf	3
Aa	3
Ab	3
Dh	3
Bn	3
Fl	6
Em	3
Dc	6
Cd	3
Ej	6
Fm	6
Fh	6
Ea	6

Population Map. Bold Boxes indicate sampled sites

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Sample Data

Site #	Value
Am	3
Fa	6
Fb	6
Ak	3
Fd	6
Ef	6
Ch	3
Bf	3
Aa	3
Ab	3
Dh	3
Bn	3
Fl	6
Em	3
Dc	6
Cd	3
Ej	6
Fm	6
Fh	6
Ea	6

Value = 3

Value = 6

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Sample Data

Site #	Value
Am	3
Fa	6
Fb	6
Ak	3
Fd	6
Ef	6
Ch	3
Bf	3
Aa	3
Ab	3
Dh	3
Bn	3
Fl	6
Em	3
Dc	6
Cd	3
Ej	6
Fm	6
Fh	6
Ea	6

Stratum A

59 Units

10 Selected At Random

Stratum B

31 Units

10 Selected At Random

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Sample Data

Site #	Value
Am	3
Fa	6
Fb	6
Ak	3
Fd	6
Ef	6
Ch	3
Bf	3
Aa	3
Ab	3
Dh	3
Bn	3
Fl	6
Em	3
Dc	6
Cd	3
Ej	6
Fm	6
Fh	6
Ea	6

Stratum A

59 Units

10 Selected At Random

Stratum B

31 Units

10 Selected At Random

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Sample Data			
Site #	Value	Stratum	Weight
Am	3	A	5.9
Fa	6	B	3.1
Fb	6	B	3.1
Ak	3	A	5.9
Fd	6	B	3.1
Ef	6	B	3.1
Ch	3	A	5.9
Bf	3	A	5.9
Aa	3	A	5.9
Ab	3	A	5.9
Dh	3	A	5.9
Bn	3	A	5.9
Fl	6	B	3.1
Em	3	A	5.9
Dc	6	B	3.1
Cd	3	A	5.9
Ej	6	B	3.1
Fm	6	B	3.1
Fh	6	B	3.1
Ea	6	B	3.1
Average	4.50		
Weighted Average	4.03		

Introduction to sampling

An Introduction to CHaMP Sampling and Analysis

Introduction to sampling

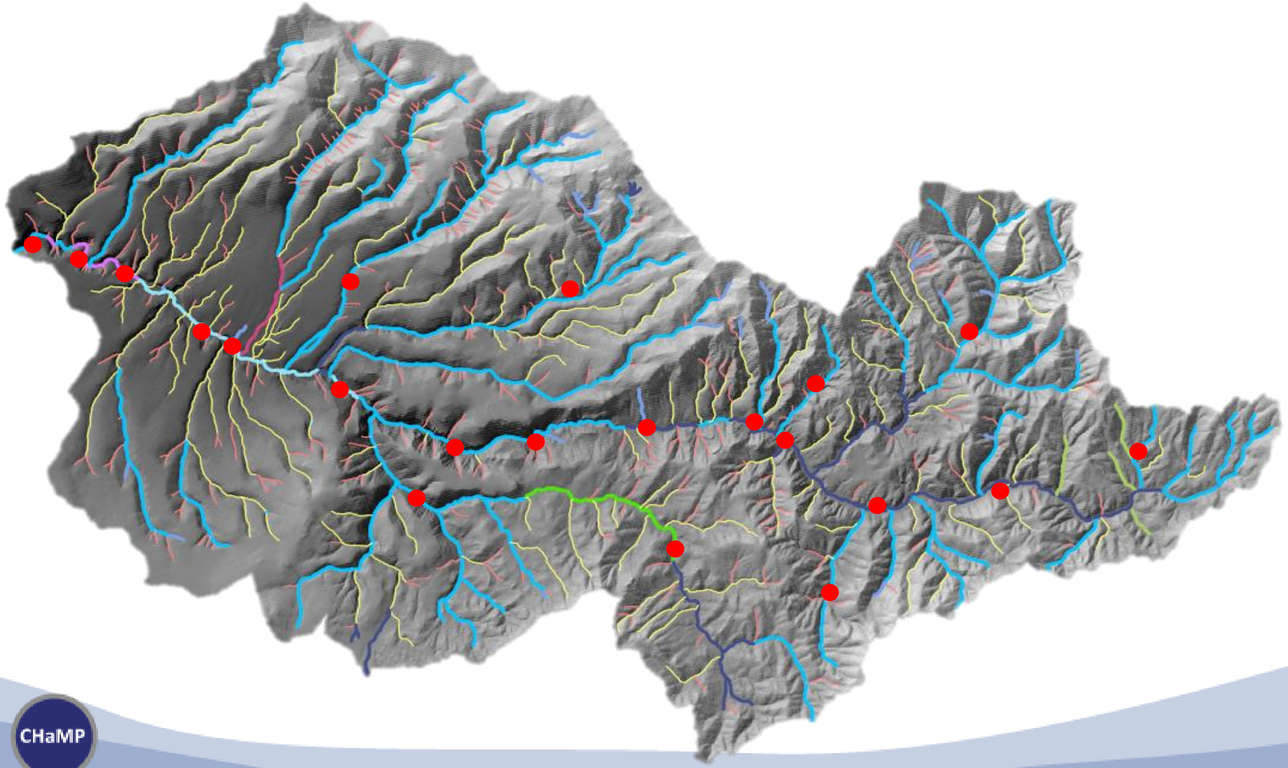
- Question: why do we sample?

Sampling, when done well, enables us to make inference to an entire population, while only measuring directly a subset of that population

Effective sampling of only a tiny fraction of a population can often lead to precise inference regarding a broad population

Introduction to sampling

- Question: What are the tradeoffs to consider between intensive reach level sampling (as in CHaMP) and doing a census (i.e. measure the entire stream network, but with far less information content per length of stream)?



Sampling Terminology

An introduction to CHaMP Sampling and Analysis

Sampling: Definitions

- Target population
  - The resource about which estimates are needed
    - Defined conceptually using written text
    - **Must define what are the elements of the target population.**

Source: http://www.epa.gov/nheerl/arm/designpages/monitdesign/targetpopframe.htm

Introduction to sampling

- Question: What is our target population in CHaMP?

- Question: What are the elements of our population?

Sampling: Definitions

- Sampling Frame
  - A physical representation of the target population
    - It consists of sample units that are potential members of the sample
    - Extent (size) of the frame is obtained by summation
    - Sample Frames almost always are not exact representations of the target population
      - Sample Frame may not include some Target Population elements: Undercoverage
      - Sample Frame may contain non-target elements, e.g., mis-identified sample units: Overcoverage

Source: http://www.epa.gov/nheerl/arm/designpages/monitdesign/targetpopframe.htm

Sampling: Definitions

- Sample
  - The subset of the Sample Frame sample units selected for sampling
    - Probability survey designs used to select the subset
      - One design - GRTS
      - May include stratification, unequal probability selection, panels for surveys over time, etc.

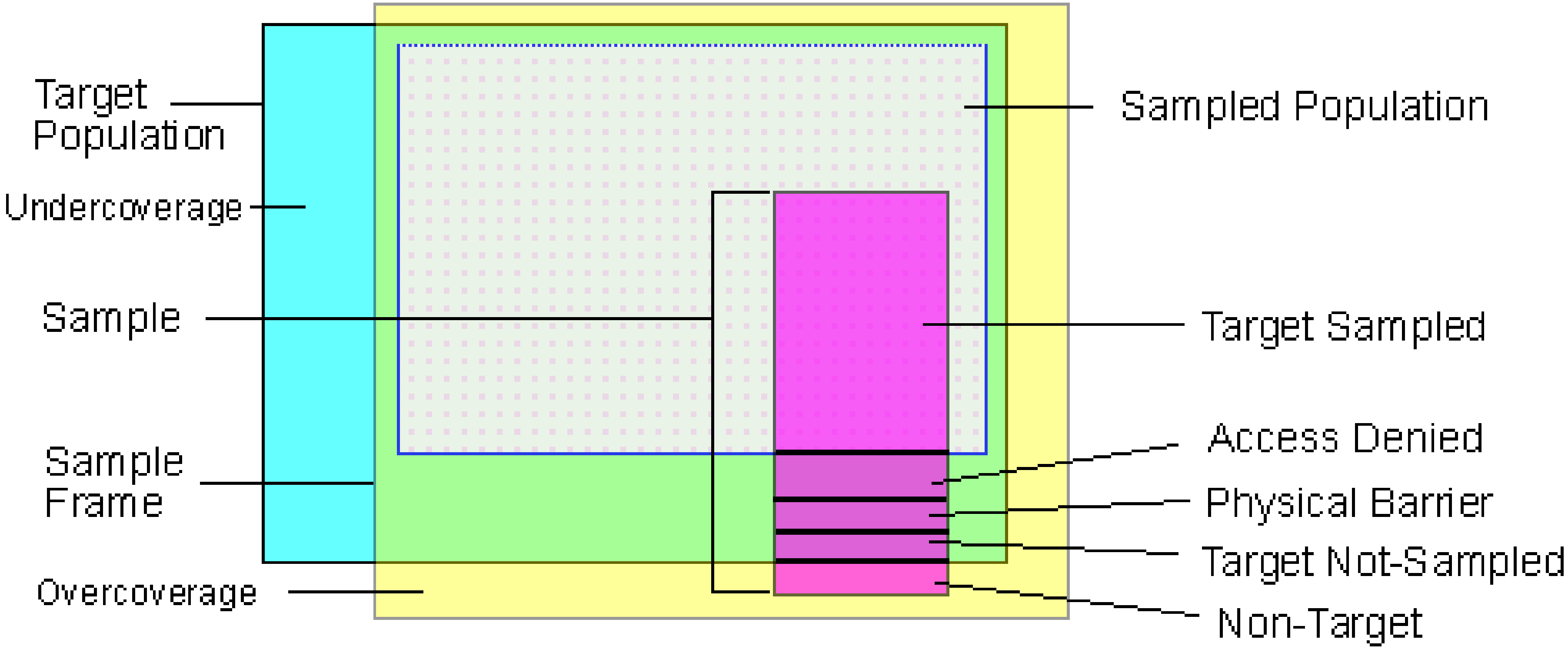
Source: http://www.epa.gov/nheerl/arm/designpages/monitdesign/targetpopframe.htm

Sampling: Definitions

- Sampled Population
  - A conceptual population that is a subset of intersection the Target Population and the Sample Frame
  - excludes portion of the Target Population within the Sample Frame that could not be sampled (conceptually) due to access problems, lost samples, or other reasons a sample could not be collected
  - It doesn't include part of the Sample Frame that is determined to not be elements of the Target Population (Non-Target)

Source: http://www.epa.gov/nheerl/arm/designpages/monitdesign/targetpopframe.htm

Sampling: Terminology

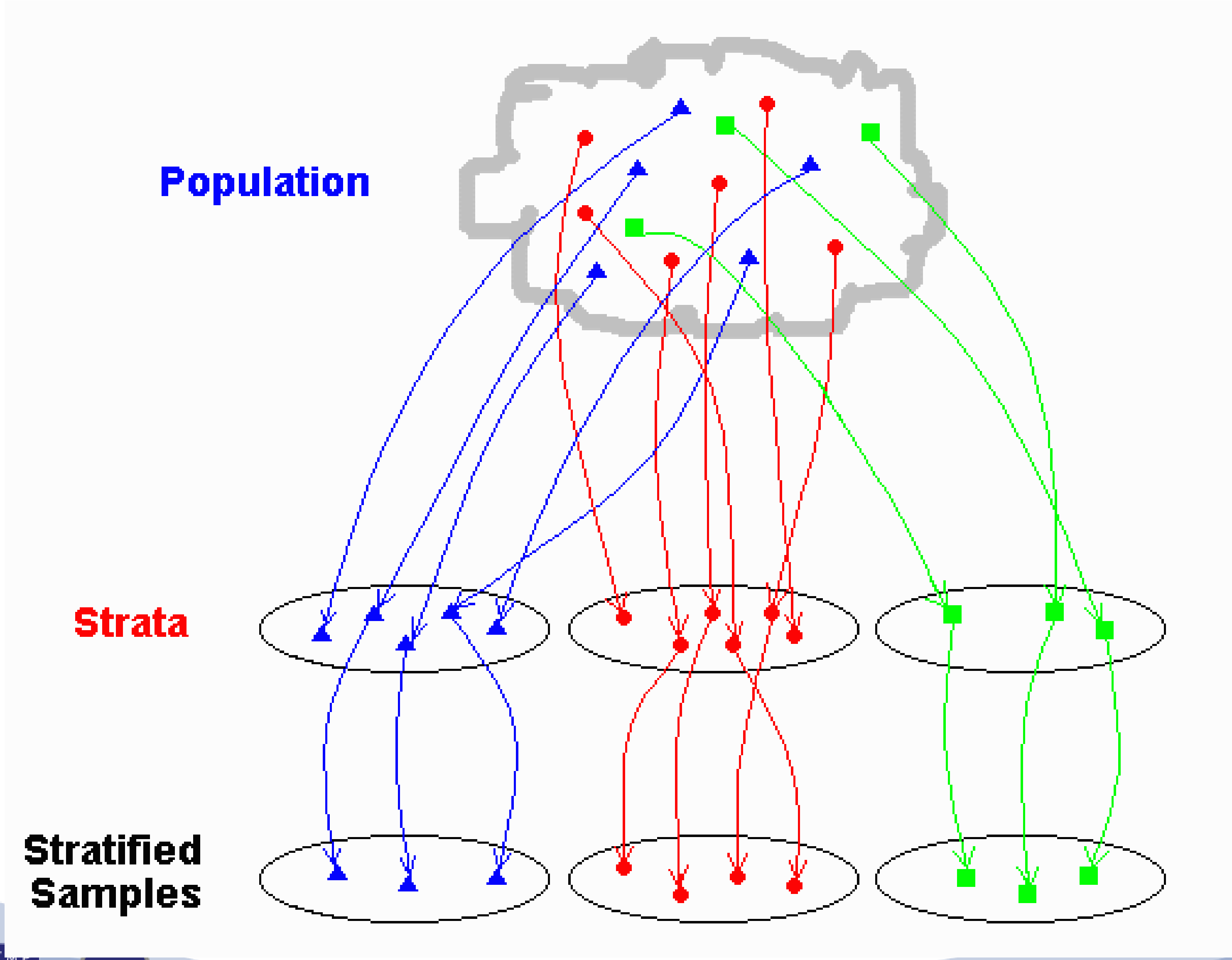


Source: http://www.epa.gov/nheerl/arm/designpages/monitdesign/targetpopframe.htm

Types of Sample Designs

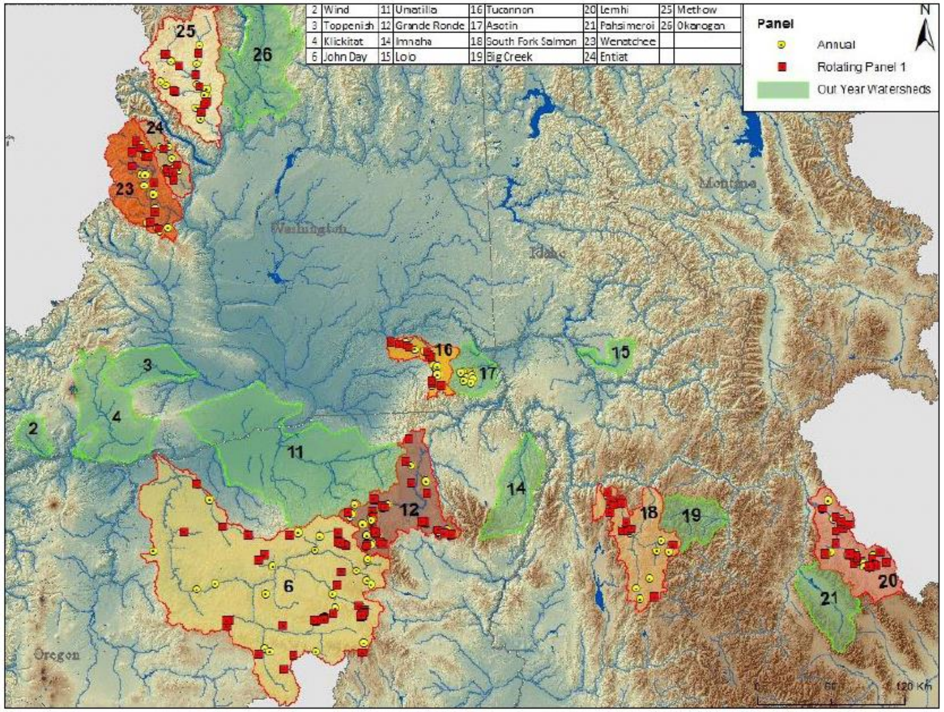
- Simple Random Sampling
  - Every element in a sample frame has equal probability of being selected in the sample
- Stratified random Sampling
  - Sampling frame is divided into strata
  - Each stratum is mutually exclusive
  - Random sampling takes place within stratum
- Cluster Sampling
  - Sample is divided into natural groups or "clusters"
  - SRS sample used to pick subset of clusters
  - Subset of elements selected from within each cluster
- Other…

Stratified Sampling



Type of Sample Designs

- Question: CHaMP selected a subset of watersheds within the Columbia, then a subset of sites within selected watersheds. Is this cluster sampling?



GRTS Sampling

Question: CHaMP uses GRTS Sampling.

What is "GRTS"?

GRTS Sampling

- GRTS Sampling = "Generalized Random Tessellation Stratified" sampling

http://www.epa.gov/nheerl/arm/documents/presents/grts_ss.pdf

Generalized Random Tessellation Stratified sampling (GRTS)

- GRTS is an alternative to random sampling
  - Can be applied to other sampling designs (i.e. uniform probability sampling, stratified sampling, cluster sampling, etc.)
- Achieves a more spatially balanced sample
  - Enables more efficient sampling

GRTS Sampling

- GRTS sampling is considerably more spatially balanced the a true random sample
- Benefit:
  - Spatial balance enables – in many cases – more efficient estimates of natural resource response metrics
    - Sometimes sites adjacent to each other are highly correlated – i.e. the 2nd site doesn't provide much new information not contained by the first
    - "More efficient" = higher precision for the same sample size

Stratified Sampling

- What is a stratum?
  - A group of sites for which, within the stratum, there is equal probability of each site being selected in the sample
    - Strata do not need to be spatially continuous
    - Strata are mutually exclusive
    - All sites must be in a stratum

Stratum A
59 Units
10 Selected At Random

Stratum B
31 Units
10 Selected At Random

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Stratified Sampling

- Question: I want to know the average value for 4 groups of sites as outlined.
- Are these groups of sites "strata"?

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Stratified Sampling

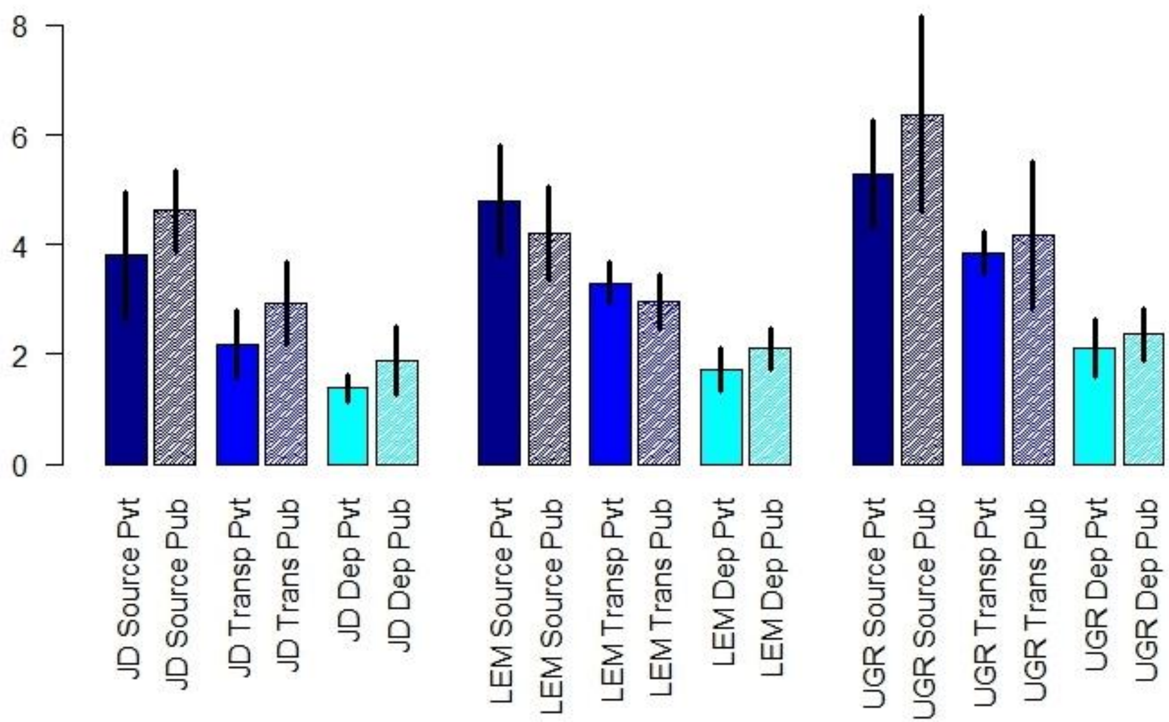
- Question: Why does CHaMP use stratified sampling?

Stratified Sampling

- Why stratified sampling?
  - Desire to ensure at least a minimum sample size in different strata
  - Some strata may be deemed more important than others, and we may choose to sample more densely in those strata
  - Some strata may naturally have higher variance than others. More efficient estimation is possible if sample size is greater in high variability strata than low variability strata
  - Belief that within strata variation is less than strata-strata variation
  - Other?

Stratified Sampling – CHaMP Strata

Estimated Mean Fast Turbulent Frequency by Valley Class x Ownership Type



High strata-strata variability

Low within strata variability

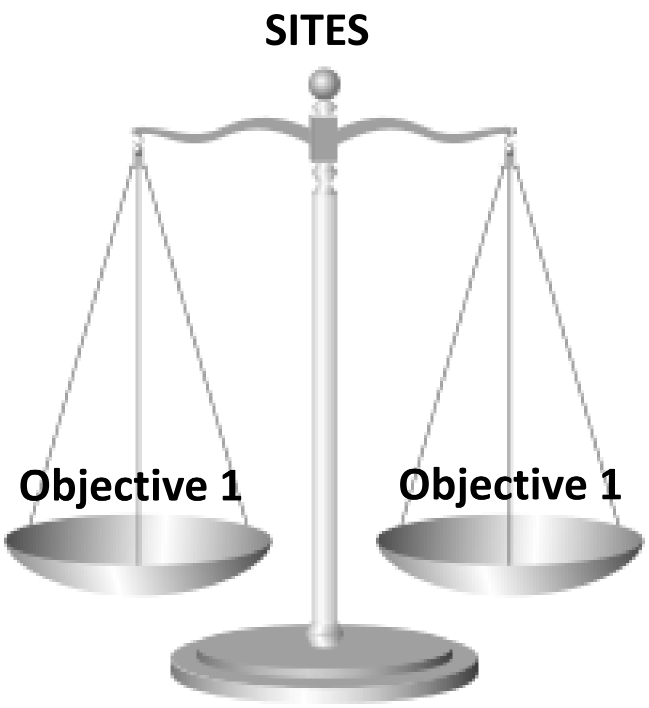
BONNEVILLE POWER ADMINISTRATION ISEMP CHaMP

Stratified Sampling

- Basic steps:
  - Define strata
  - Choose sample size by strata
  - Randomly select sites within each strata
    - Apply GRTS if desired
  - Collect Data
  - Calculate design weights
  - Analyze data taking design weights into account
    - Design or Model based analysis depending on question(s) of interest

Stratified Sampling

- Defining strata requires balancing competing objectives!
- Question: What are some of CHaMP competing objectives?



Stratified Sampling – Define Strata

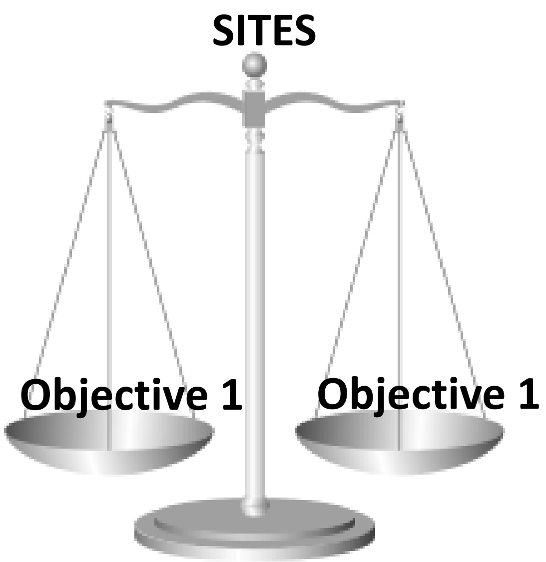
- Typically CHaMP stratifies by Valley Class x Ownership Type
  - Sometimes other strata are defined

Wenatchee Watershed: Total Stream Length (km) by Stratum		
	Public	**Private**
Source	32.85	12.80
Transport	6.67	12.87
Depositional	109.64	98.38
Little Wenatchee	11.15	

Q: How do we make good choices for strata?

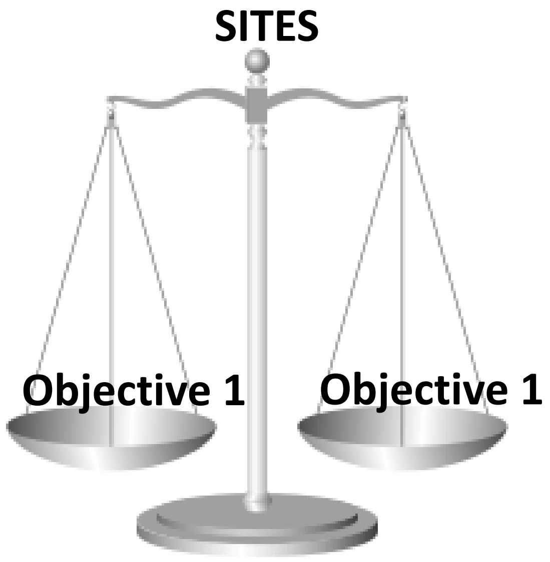
Stratified Sampling – Define Strata

- Strata Considerations
  - Important subpopulations may be give their own stratum
  - Make Stratum-Stratum variability high, within stratum variability low
    - For efficient estimation



Stratified Sampling – Choose sample size by strata

- Sample size per stratum depends on:
  - Total available resources for sampling
  - Relative importance of strata
  - Expected variation by strata
  - Other?



Stratified Sampling – Choose sample size by strata

- Question: Is it important to give larger strata proportionally more samples?

Wenatchee Watershed: Total Stream Length (km) by Stratum		
	Public	**Private**
Source	32.85	12.80
Transport	6.67	12.87
Depositional	109.64	98.38
Little Wenatchee	11.15	

Stratified Sampling – Collect Data

- See “CHaMP Camp”

Calculating Design Weights

- Design weight = $\text{Extent}_{\text{stratum}} / N_{\text{stratum}}$

- A site's design weight represents the total length of stream that it "represents" in the analysis
- The more sites in a stratum, the lower the design weight.

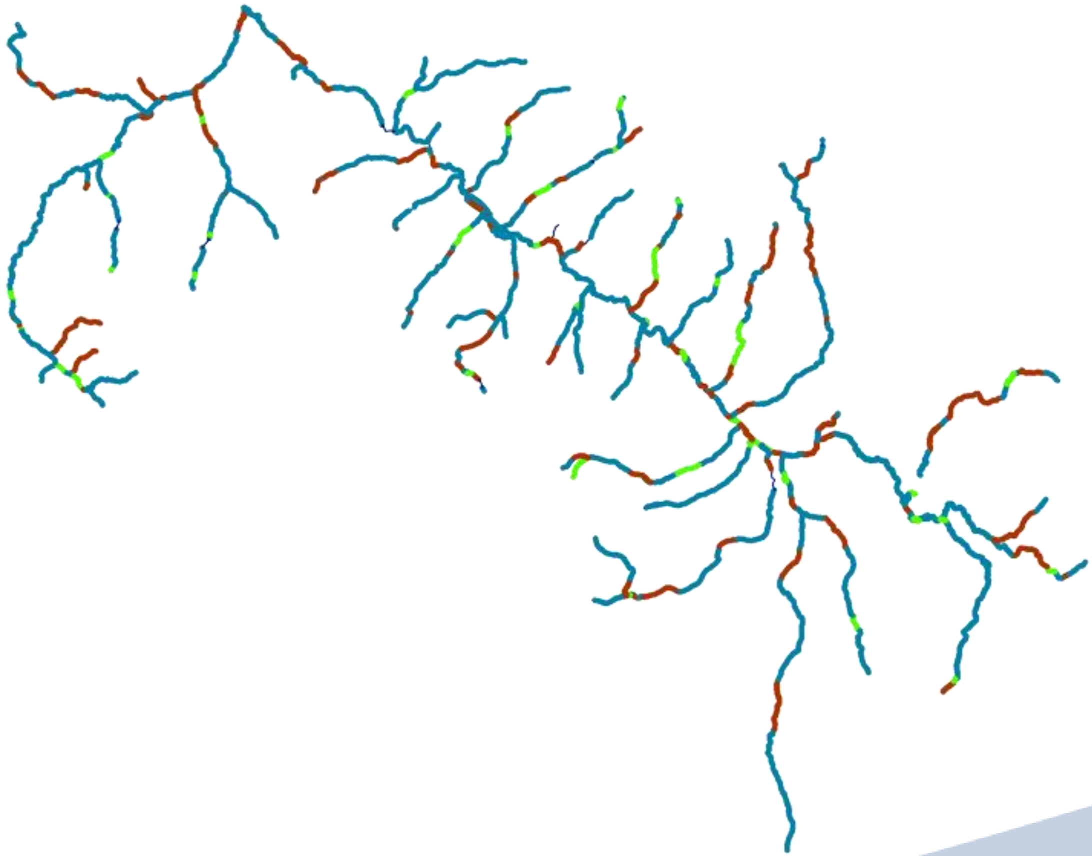
- Question: What is the weight of each of the selected sites in the example to the right?

Stratum A Extent = 59 Units 10 Selected At Random Wgt = 5.9/10 = 5.9 Units			Stratum B Extent = 31 Units 10 Selected At Random Wgt = 31/10 = 3.1 Units		
Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Calculating Design Weights

- Design weight = $\text{Extent}_{\text{stratum}} / N_{\text{stratum}}$

Example Stream Network: Color = Strata



In CHaMP, strata extents are measured in distance (km) of a linear stream network.

Design weights have units! (km of stream distance)

Calculating design weights

- Question
  - What if, during site evaluations, we find that a site is “non-target”?

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Calculating design weights

Answer: We adjust the frame to remove entire portion “represented” by non-target site.

Stratum A	Stratum B
Extent = 59 Units	Extent = 31 Units
10 Selected At Random	10 Selected At Random
Wgt = 5.9/10 = 5.9 Units	Wgt = 31/10 = 3.1 Units

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Stratum A	Stratum B
Extent = 59 Units	Extent = 27.9 Units
10 Selected At Random	9 Selected At Random
Wgt = 5.9/10 = 5.9 Units	Wgt = 27.9/9 = 3.1 Units

Aa	Ba	Ca	Da		**Fa**
Ab	Bb	Cb	Db		**Fb**
Ac	Bc	Cc	**Dc**		Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Calculating design weights

- Question
  - What if, during analysis, we have an "NA" for a given metric?

Stratum A Extent = 59 Units 10 Selected At Random Wgt = 5.9/10 = 5.9 Units			Stratum B Extent = 31 Units 10 Selected At Random Wgt = 31/10 = 3.1 Units		
Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Calculating design weights

Answer: The frame doesn't change. Assuming "missing at random within stratum" we re-calculate stratum sample weights.

Stratum A	Stratum B
Extent = 59 Units	Extent = 31 Units
10 Selected At Random	10 Selected At Random
Wgt = 5.9/10 = 5.9 Units	Wgt = 31/10 = 3.1 Units

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Stratum A	Stratum B
Extent = 59 Units	Extent = 31 Units
10 Selected At Random	9 Selected At Random
Wgt = 5.9/10 = 5.9 Units	Wgt = 31/9 = 3.444 Units

Aa	Ba	Ca	Da	**Ea**	**Fa**
Ab	Bb	Cb	Db	Eb	**Fb**
Ac	Bc	Cc	**Dc**	Ec	Fc
Ad	Bd	**Cd**	Dd	Ed	**Fd**
Ae	Be	Ce	De	Ee	Fe
Af	**Bf**	Cf	Df	**Ef**	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	**Ch**	**Dh**	Eh	**Fh**
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	**Ej**	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	**Fl**
Am	Bm	Cm	Dm	**Em**	**Fm**
An	**Bn**	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Calculating design weights

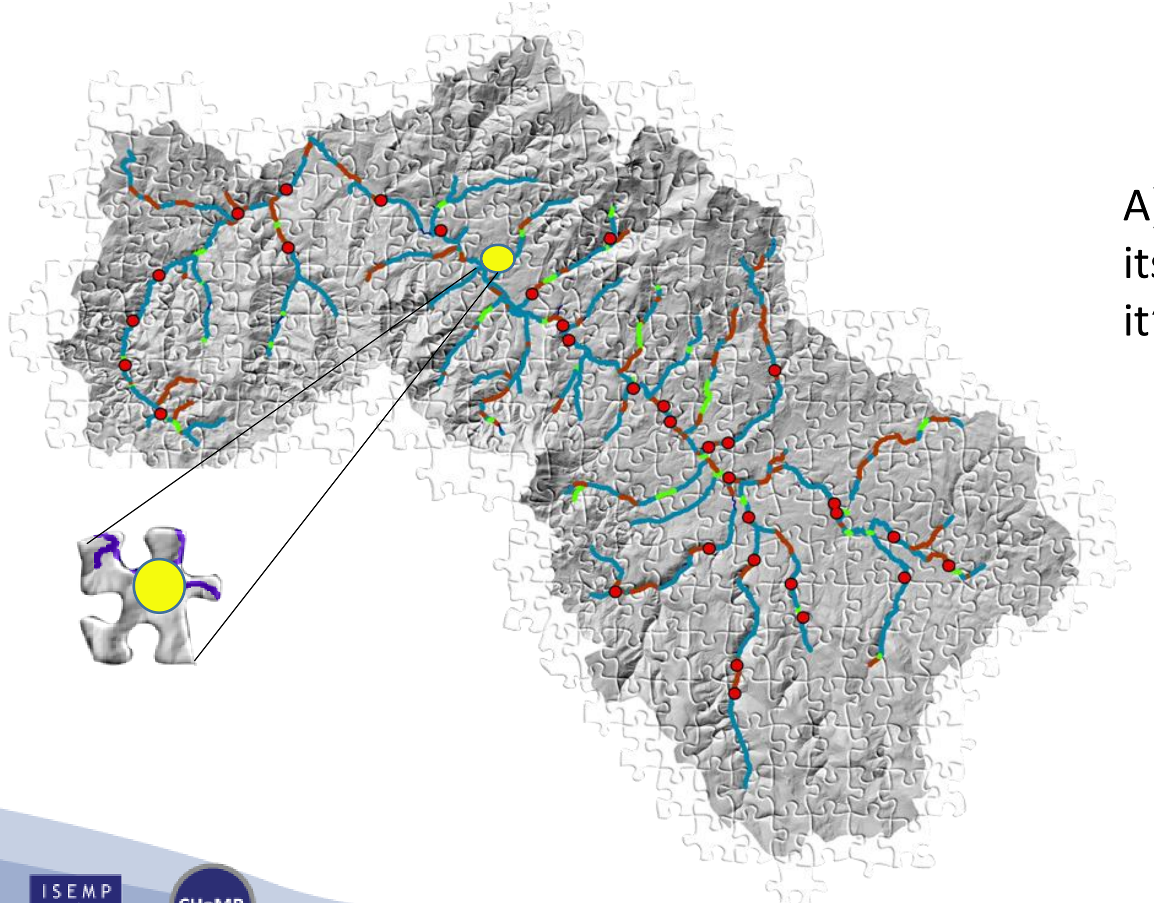
- Why isn't there a big list of weights for every site?

- Note that, in the presence of N/A data, the same site may end up with different weights for different metrics

- We typically calculate weights during the analysis to avoid propagating erroneous weights

Calculating design weights

- Question
  - What if, after sampling design, John Q Manager decides to add site 86753, because it's an interesting site and besides, it right next to the road? I.e. it's an "opportunistic" site?



A). Site will represent only its own length. It will be it's own stratum.

Calculating design weights

- Question
  - What if a landowner decides, at that last minute, that you're not allowed to sample on her property?

Analysis of CHaMP Data

An introduction to CHaMP Sampling and Analysis

Stratified Sampling: Analyze Data

- Two basic types of statistical analyses:
  - **Design based analysis**
    - "Status and Trend" of a finite population
  - **Model based analysis**
    - Suitable for examination of complex relationships between variables
    - Higher risk: requires assumptions about model structure, distributions of residuals and other random effects, etc.

Design Based Inference

- Design based inference
  - Estimations of status or trend of an attribute of a finite population
  - We don't need to assume a distribution for response variable(s)
    - All stochastic elements are controlled by sample design
      - Population is fixed
      - Sample units are selected by probability sample
      - Statistical inference is based on sampling design

Design based inference:
Horvitz-Thompson Estimator

$\overline{y_{HT}}$= Estimate of the population mean

π_i= Probability that the i_{th} population element … be selected in the sample

N = Sample Size

Don’t panic! This is nothing more than a weighted average!

Estimated Population Mean

$$\overline{y}_{HT} \ldots$$

Estimated Population Variance

$$\widehat{\mathrm{Var}}(\overline{y}_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{n}\left(\frac{\cdots}{\pi_i}\right)y_i^2 + \sum_{i=1}^{n}\sum_{i\neq j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{y_i y_j}{\pi_{ij}}\right]$$

Design based inference

- Question: Design based inference is effective for status and trend estimates. What do we mean by "Status and Trend"?

- Question: What does "GRTS Rollup" mean, exactly?

CHaMP default, annual “GRTS Rollups”

- What we estimate for every CHaMP metric:
  - Status, calculated separately for each year
  - Status – Average of all three years
    - Responses are site level averages taken across all measured years*
  - Trend
    - Responses are slope of CHaMP metrics vs year by site*

- By Watershed for all Estimates
  - But we can easily modify this to rollup by any subgroup of your choice.

*Different sites have different number of measurements

Example status and trend results- table

Mean

Standard dev of subgroup

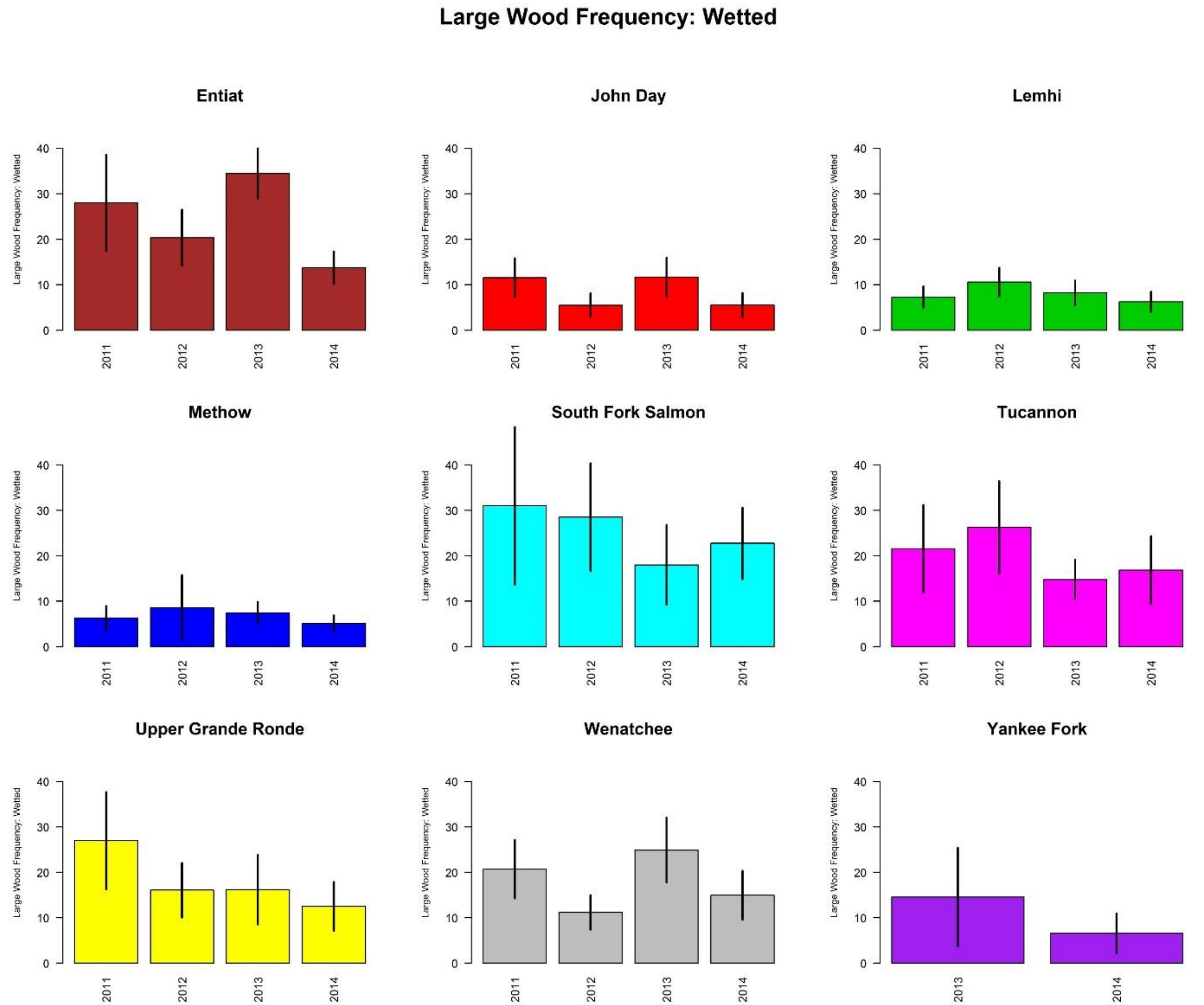
Trend

Metric	Visit.Year	Sub-Population	Number of CHaMP Sites	N	Mean	Std Error of Mean Estimate	Standard Deviation	Median	CV	95 PCT LCB	95 PCT UCB	Trend	Std Error of Trend Estimate	Trend 95 PCT5 LCB	Trend 95 PCT UCB
Fast NonTurbulent Frequency	2011	All.Sites	335	307	0.975	0.163	1.557	0.319	1.596923	0.657	1.294				
Fast NonTurbulent Frequency	2012	All.Sites	326	275	1.02	0.099	1.04	0.658	1.019608	0.825	1.214				
Fast NonTurbulent Frequency	2013	All.Sites	375	323	1.243	0.135	1.471	0.635	1.183427	0.979	1.507				
Fast NonTurbulent Frequency	2014	All.Sites	273	211	1.13	0.132	1.28	0.812	1.132743	0.87	1.39				
Fast NonTurbulent Frequency	Average of All Years	All.Sites	762	590	0.986	0.071	1.147	0.59	1.163286	0.847	1.125	0.062	0.068	-0.071	0.195
Fast NonTurbulent Frequency	2011	Entiat	75	73	0.595	0.288	1.19	0	2	0.03	1.16				
Fast NonTurbulent Frequency	2012	Entiat	55	52	0.532	0.218	0.96	0	1.804511	0.105	0.958				
Fast NonTurbulent Frequency	2013	Entiat	75	72	0.688	0.153	0.949	0.333	1.37936	0.387	0.988				
Fast NonTurbulent Frequency	2014	Entiat	46	15	0.211	0.061	0.304	0	1.440758	0.092	0.33				
Fast NonTurbulent Frequency	Average of All Years	Entiat	100	100	0.678	0.13	1.059	0.201	1.561947	0.423	0.933	0.118	0.033	0.055	0.182
Fast NonTurbulent Frequency	2011	John Day	59	43	0.776	0.162	1.133	0.482	1.460052	0.458	1.094				
Fast NonTurbulent Frequency	2012	John Day	76	36	0.788	0.327	0.95	0.501	1.205584	0.147	1.429				
Fast NonTurbulent Frequency	2013	John Day	65	31	1.369	0.291	1.153	1.357	0.842221	0.8	1.939				
Fast NonTurbulent Frequency	2014	John Day	30	16	2.133	0.641	1.908	0.818	0.894515	0.877	3.389				
Fast NonTurbulent Frequency	Average of All Years	John Day	190	50	1.334	0.21	1.187	0.826	0.889805	0.922	1.746	0.925	0.28	0.375	1.474
Fast NonTurbulent Frequency	2011	Lemhi	42	39	1.1	0.145	1.056	0.681	0.96	0.815	1.384				
Fast NonTurbulent Frequency	2012	Lemhi	49	44	1.239	0.109	0.911	0.937	0.73527	1.026	1.452				
Fast NonTurbulent Frequency	2013	Lemhi	48	45	0.932	0.121	1.062	0.7	1.139485	0.695	1.169				
Fast NonTurbulent Frequency	2014	Lemhi	24	22	0.588	0.107	0.811	0.565	1.379252	0.378	0.797				
Fast NonTurbulent Frequency	Average of All Years	Lemhi	111	100	1.081	0.075	0.946	0.807	0.875116	0.934	1.228	-0.193	0.06	-0.311	-0.076

BONNEVILLE POWER ADMINISTRATION ISEMP CHaMP

Example status and trend results: plots
(Large Wood Frequency: status by watershed x year)

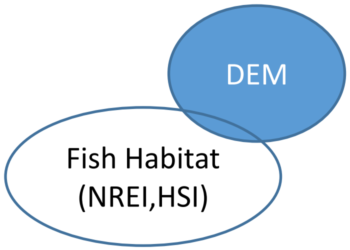
Estimated mean Large Wood Frequency: Wetted (1/m), by watershed x year. Black lines indicates 95% confidence intervals for the mean.

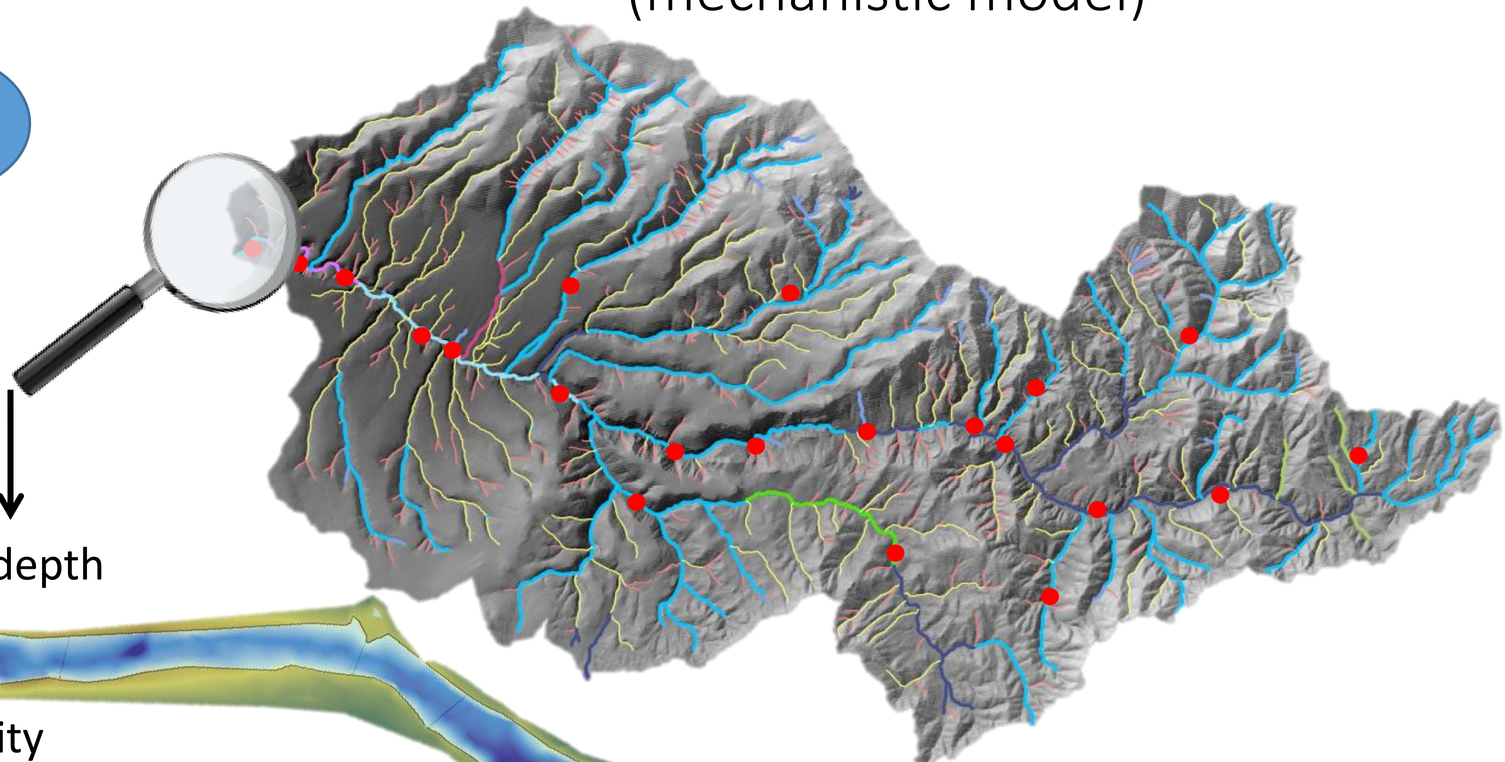


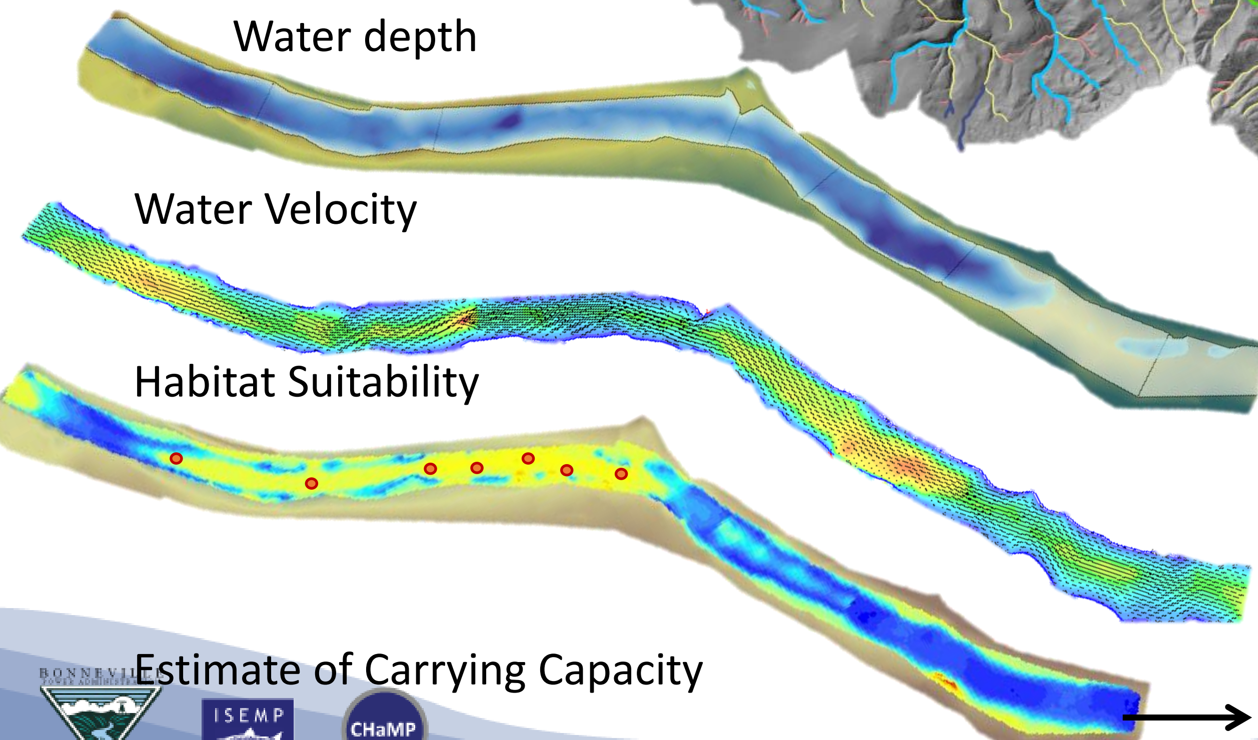
Estimating status and trend ("GRTS Rollups")

- Question: can we do a GRTS rollup on CHaMP "products" such as HSI or NREI?

DEM based protocol→ HSI → Potential Redd Abundance (mechanistic model)







CHaMP Surveys collect:

- Substrate
- Topography/Channel Units

e.g. GRTS rollup to get basin level abundance of redds or adults

HSI Design Based Capacity Estimates
"GRTS Rollups"

Watershed	Species	Life Stage	Estimated Capacity (1000's)	LCB95Pct	UCB95Pct	N
Entiat	Chinook	Juvenile	1674.17	1385.37	1962.96	46
Entiat	Chinook	Spawner	14.19	12.28	16.09	44
Entiat	Steelhead	Juvenile	2680.86	2385.45	2976.28	46
Entiat	Steelhead	Spawner	90.91	78.61	103.21	44
Lemhi	Chinook	Juvenile	1416.95	961.78	1872.13	41
Lemhi	Chinook	Spawner	30.08	17.92	42.23	38
Lemhi	Steelhead	Juvenile	1319.13	925.19	1713.06	41
Lemhi	Steelhead	Spawner	42.14	25.19	59.10	38

R-Example: GRTS Rollup for Wetted Width : Depth Ratio, 2014

Introduction to CHaMP sampling and data analysis

Design based Inference

- Question: How do we estimate trend if the sampling design changes from year to year?

Break

Introduction to CHaMP sampling and data analysis

Model based inference

- Question: what is a “model” in statistics

Model based Inference

- Question: Are there some questions for which “design based” analysis is not suitable?

How does measurement noise compare to other sources of variability (site-site, watershed-watershed, year-year

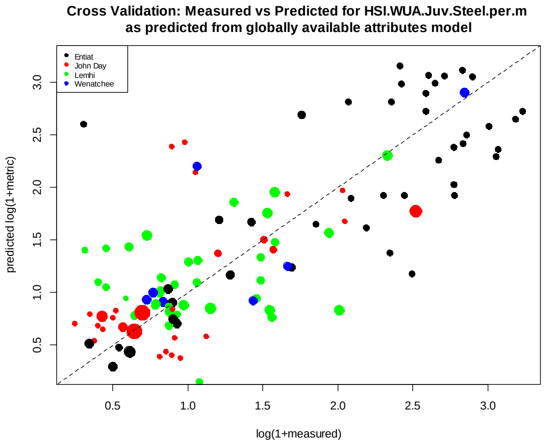
What is the relationship between CHaMP metrics and observed site level juvenile steelhead abundance?

Can we relate CHaMP metrics to globally available attributes to predict CHaMP responses at unmeasured sites?

Model based inference

Model based inference

- Estimate parameters of an assumed statistical model describing the relationship attributes of a population
    - Regression is “model based inference”
- We need to make distributional assumptions about model errors



BONNEVILLE POWER ADMINISTRATION
ISEMP
CHaMP

Model based inference

"All models are wrong. Some models are useful."

- George E. P. Box

Model based inference

- Example of a statistical model:
  - $Y_i = \mu + \beta X_i + e_i$
  - e ~ IID Normal(0, σ^2)

  - μ → the intercept
  - β → slope or "coefficient"
  - e_i → random error
  - Y_i → The measured value at site i (i.e. our CHaMP Metric)
  - X_i → An explanatory variable assessed at site i

- μ and β are unknown, but fixed "truths"
- We cannot know – we can only estimate, the fixed "true" values of μ and β

Model based inference

- What about the error? What is it, really?
  - e ~ IID Normal(0, σ^2)

- What we call "random" may be described as error in the model arising from effects present in nature that our model fails to capture.
  - Mathematically, we model this as "random"

Model based inference

- Question: What does it mean for errors to be "independent, identically distributed"?

Model based inference

- What is this “IID”?
  - IID ➔Independent, identically distributed
    - Independence:
      - $P(A \cap B) = P(A)P(B)$ (Discreet)

        - or -
      - $F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$ (Continuous)

    - All residuals belong to sample distribution
      - Example: e ~ IID Normal(0, σ^2)

Model based inference

- Q: What happens if we violate the I.I.D. assumption in statistical modeling?

Model based inference

- Question: What if the underlying drivers of error are not randomly distributed in a population?
  - Example. Errors in habitat models are driven by geological features that are not quantified
    - Geological features are not randomly distributed throughout space
    - Sampling is not in proportion to the distribution of geological features

Model based inference

- How do we generate data from which we can be certain errors are IID?
  - Even if we don't know a-priori what might drive variation in residuals

  - **We Randomize!**

Model based inference

- **Randomization**
- Simple random sampling:
  - Every individual in a population has equal chance of being selected in a sample
  - SRS is generally the most powerful tool for ensuring IID errors, which in turn yields unbiased parameter estimates
  - IID errors can be ensured by randomization during sample selection (or assignment to treatment groups in a designed experiment)

Model based inference

Simple Random
Sample of 90 Units

Note: This is equivalent to a “stratified” sample with only 1 Strata

All weights are the same

Population
Extent = 90 Units
20 Selected At Random
Wgt = 90/20 = 4.5 Units

Aa	Ba	Ca	Da	Ea	Fa
Ab	Bb	Cb	Db	Eb	Fb
Ac	Bc	Cc	Dc	Ec	Fc
Ad	Bd	Cd	Dd	Ed	Fd
Ae	Be	Ce	De	Ee	Fe
Af	Bf	Cf	Df	Ef	Ff
Ag	Bg	Cg	Dg	Eg	Fg
Ah	Bh	Ch	Dh	Eh	Fh
Ai	Bi	Ci	Di	Ei	Fi
Aj	Bj	Cj	Dj	Ej	Fj
Ak	Bk	Ck	Dk	Ek	Fk
Al	Bl	Cl	Dl	El	Fl
Am	Bm	Cm	Dm	Em	Fm
An	Bn	Cn	Dn	En	Fn
Ao	Bo	Co	Do	Eo	Fo

Model based inference

- Most statistical modeling tools assume that sampled elements from a population behave as if they're from a simple random sample (IID)

- CHaMP elements are not from a stratified sample and tend not to "behave" as they are. They are no, typically, IID.
  - sites in different strata have different probabilities of being included in the sample

Model based analysis

- Question: Are GRTS samples IID?

- No.
  - The sample inclusion probability for site X is less, given that a site near it has been sampled, than it is if that were not true.

- Isn't this a violation of the IID assumption?
  - Yes, but in this case, we don't really care.
  - A nice property of GRTS sampling is that, while technically this assumption is violated, the level of bias in parameter or standard error estimates is negligible

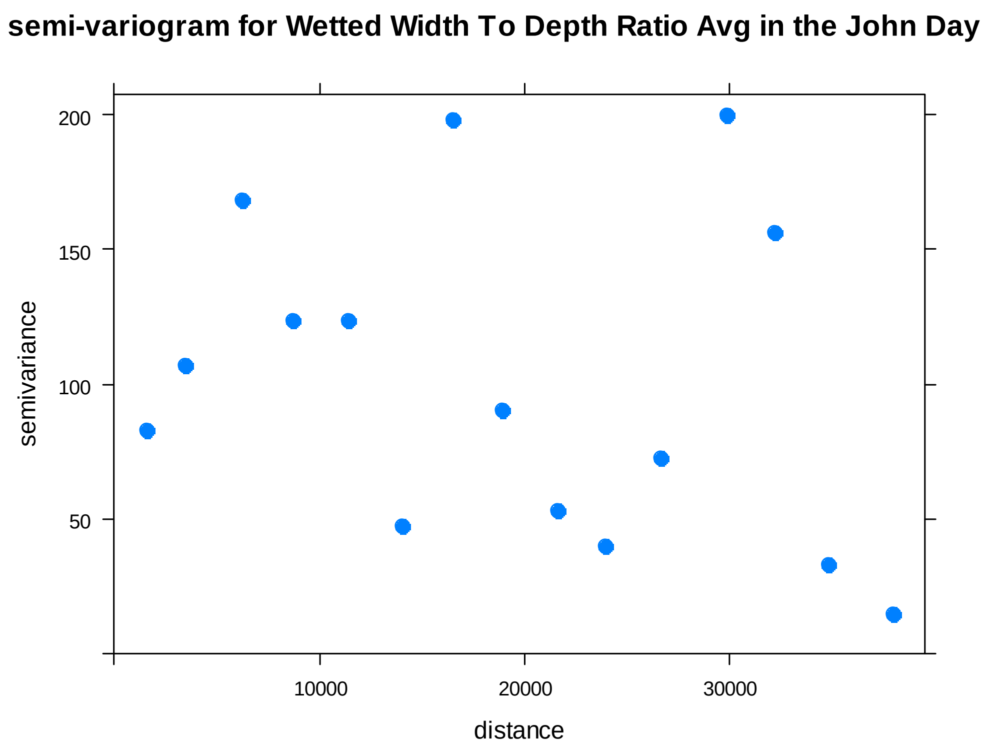
- Question: is there any “cost” associated with GRTS sampling over simple random sampling (within each strata)?

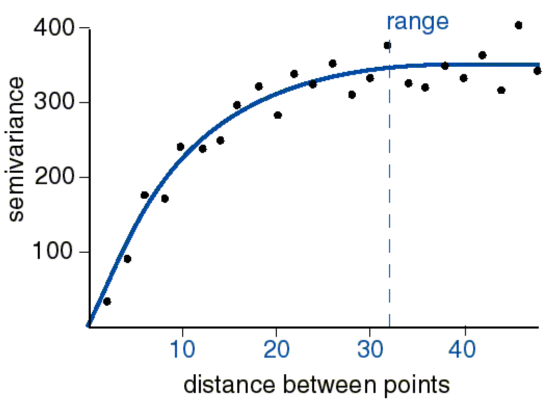
Model based analysis

- Costs of spatial balance:
  - Minor violation of independence assumption in model based analysis
    - This issue ***we can generally ignore***
  - Reduced ability model spatial autocorrelation
    - Modeling spatial autocorrelation effectively requires some data points to be close together
    - Eliminates (or at least reduces) our ability to exploit spatial autocorrelation (i.e. kriging) in extrapolation models

GRTS Sampling

Example: Spatial autocorrelation not detected at the spatial distances between sample points in the John Day ➔





🡐 Ideal Case

Model based inference

Question: What happens if we ignore sampling inclusion probabilities in a model based analysis?

Model based inference with non-uniform probability sampling

- Techniques:
  - Assume errors are independent of sample inclusion probabilities
    - Unfortunately, this is often done without acknowledgement or validation of assumption.
    - **Often WRONG! (But often published)**
  - Include strata as explanatory variable
    - Including it’s interactions with other explanatory variables if you can’t otherwise rule out these interactions
      - Potential high cost in model complexity, df
  - Model Assisted Regression
    - Applicable to many regression techniques
  - Inverse Probability Bootstrapping
    - Applicable to any model based analysis

Model Assisted Regression

- Tool in the R package “survey” enab generalized linear modeling for co lex sur ata
- Function glm:
    - Inputs
    - De
    - Ge based ts (regression trees, quantile reg , SEM, etc.)

We’ll do an example of “model assisted regression” in R. It’s not much more difficult than “regular” regression

Inverse Probability Bootstrapping (IPB)

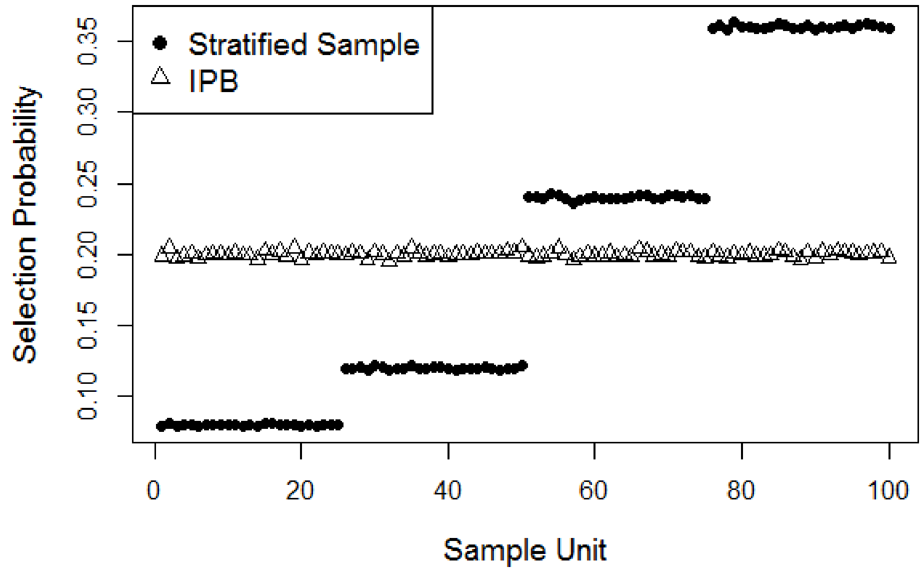
- Re-sample from your sample in such a way that you transform it into (something like) a simple random sample!
  - Resample, with replacement, from original sample using inverse sample inclusion probabilities to transform dataset into uniform sample inclusion probability data
  - Model on IPB re-sample data
  - Iterate and make inference on average across all IPB iterations

- IPB enables use of generalized set of model based tools with complex survey data!

Inverse Probability Bootstrapping (IPB)

- Re-Sample, with replacement, using sample probabilities inversely proportional to initial sampling probabilities

Sample inclusion probabilities for simulated stratified sampling, stratified sampling plus IPB



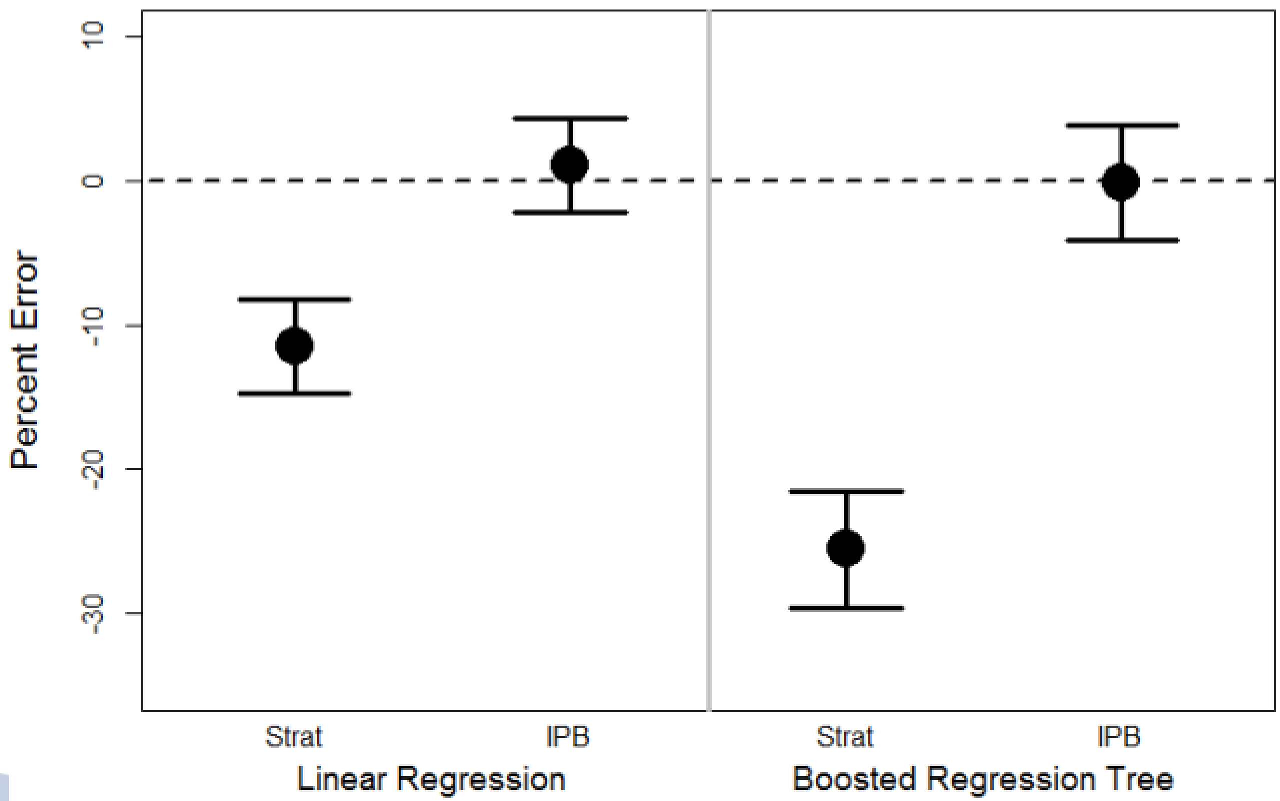
Inverse Probability Bootstrapping (IPB)

Parameter estimates for regression of ln(steelhead density, fish/m^2) on selected habitat parameters, for models that: ignore sample inclusion probabilities, and utilize IPB sampling to account for sample inclusion probabilities

	Stratified Sample: Inclusion Probabilities Ignored in Model Fitting Process		Inverse Probability Bootstrap		% Error Due to Ignoring Weights
Parameter	Est. Slope	Std. Error	Est. Slope	Std. Error	
Intercept	-1.60	0.027	-1.49	0.031	-7%
Conductivity	0.13	0.030	0.23	0.023	46%
Site Bankfull Area	-0.35	0.111	-0.68	0.137	49%
Wetted Large Wood Volume By Site	-0.01	0.034	-0.13	0.038	89%
Fast Non-Turbulent Area	-0.05	0.029	-0.09	0.038	39%
Mean Bankfull Width Mean	0.19	0.106	0.52	0.138	64%
Boulders	0.09	0.027	0.13	0.028	36%
Fish Cover Composition LWD	-0.04	0.036	-0.08	0.028	44%
Site Discharge	-0.06	0.031	-0.07	0.043	15%
Fines <2mm	0.06	0.036	0.07	0.036	14%

Inverse Probability Bootstrapping (IPB)

Mean and 95% confidence intervals for cross validation prediction error for regression of steelhead density on independent variables, and boosted regression tree analysis of steelhead density, as a percentage of the mean observed steelhead density at all sites. Models are built on data from stratified sample ignoring sample inclusion probabilities (Srat), and Inverse Probability Bootstrap samples (IPB)



Model based analyses (regression) example in R

An Introduction to CHaMP Sampling and Analysis

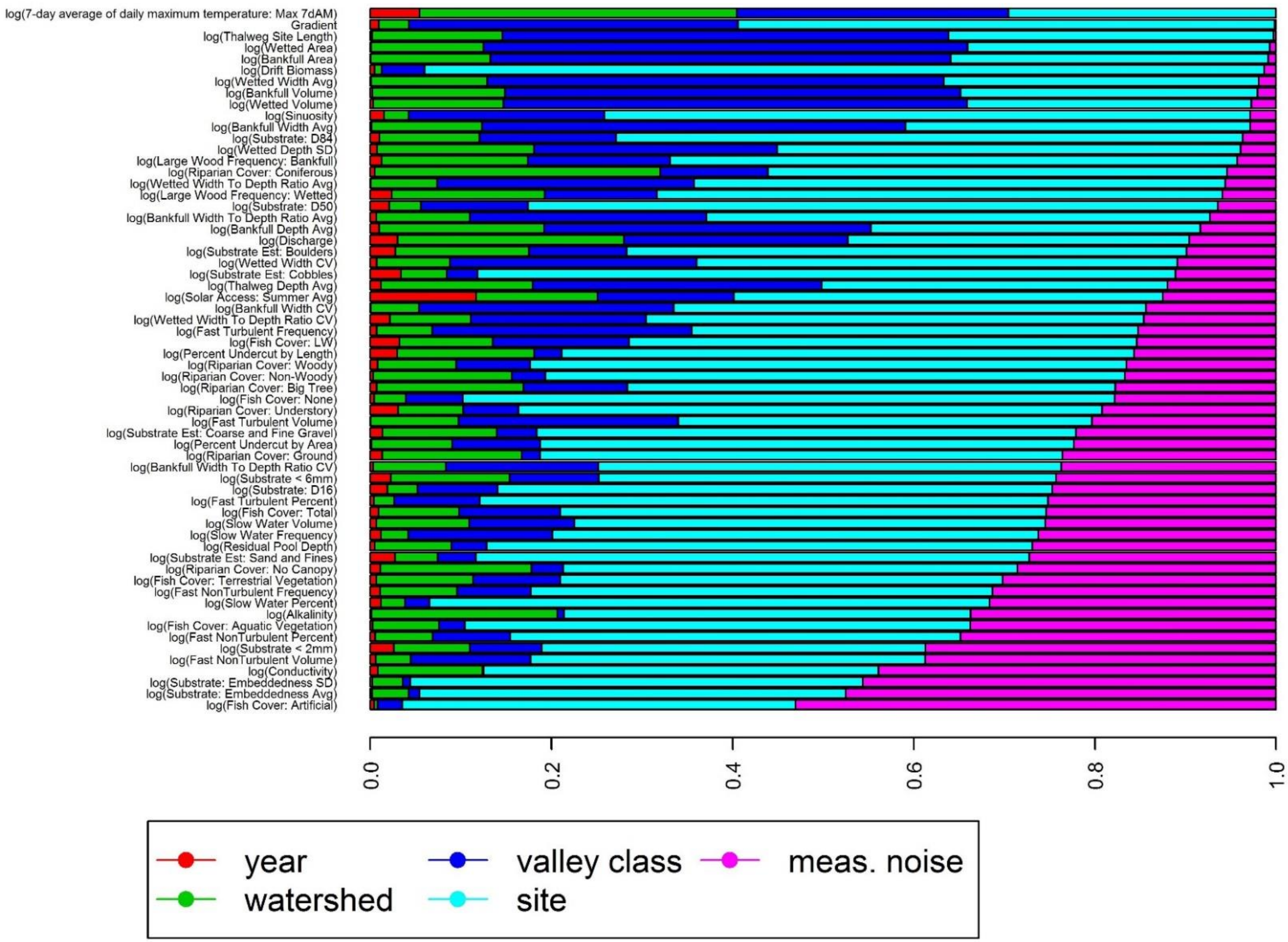
Model based analysis examples

- Variance decomposition
- Modeling HSI vs globally available attributes
- Examining effect of restoration

Model based example:
CHaMP Variance Decomposition

- Objective:
  - Assess relative magnitude of sources of variation for key CHaMP metrics
    - Provide information to feed back into sampling design
    - Assess measurement noise relative to signal
- Methods:
  - Model key CHaMP metrics by Year, Valley Class, Watershed, and Measurement Noise
    - All modeled as random effects
    - lmer function in R
  - Use IPB Bootstrapping used to account for non-uniform sample inclusion probabilities

Model based example: CHaMP Variance Decomposition



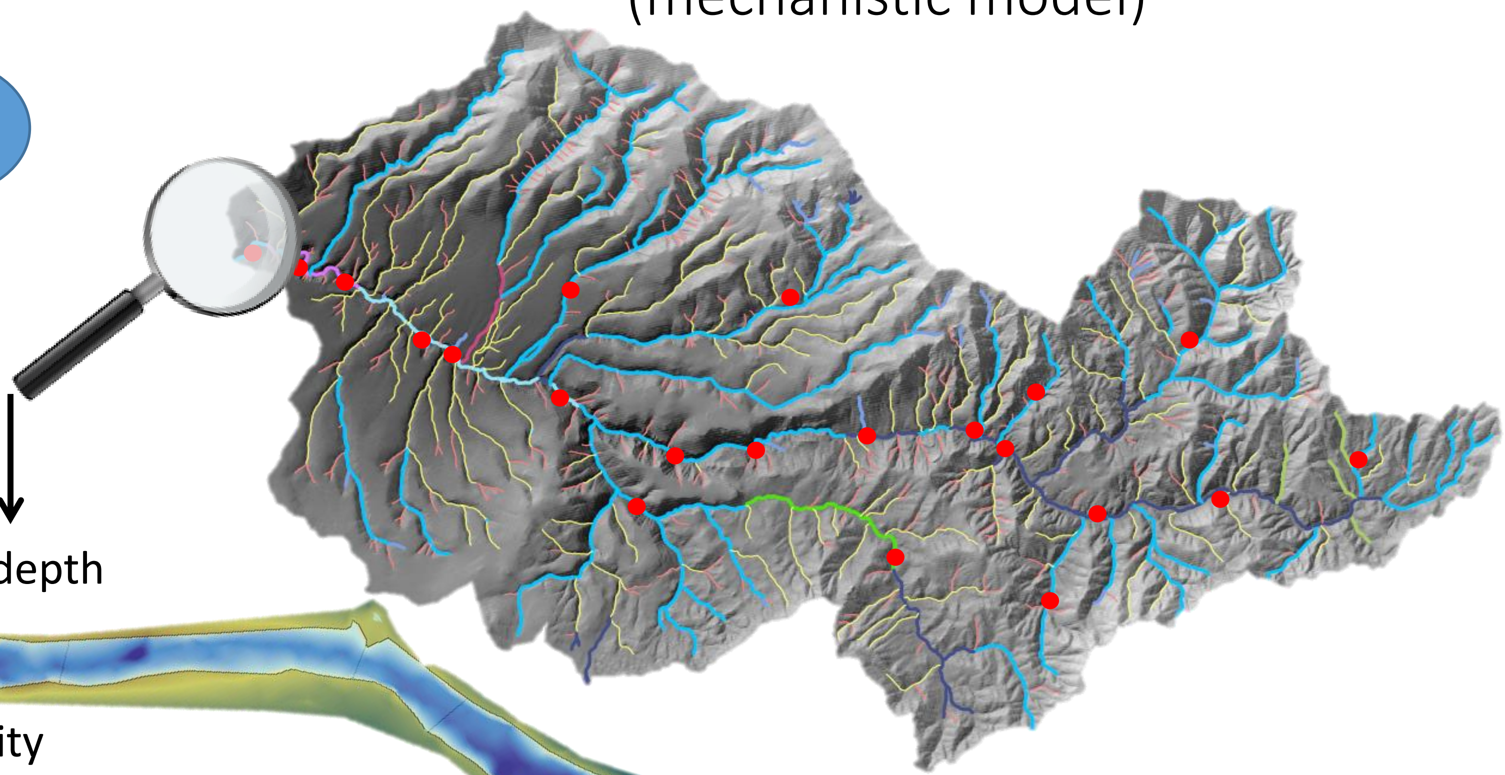
Model based regression example

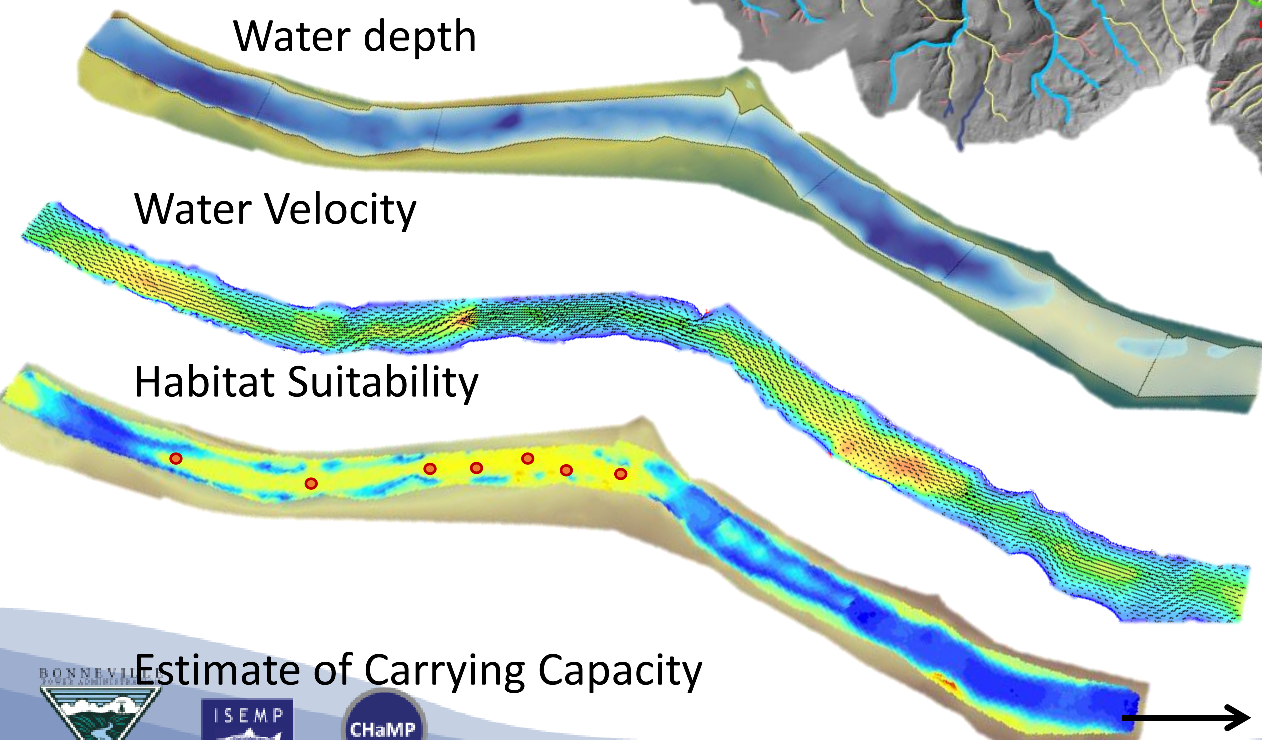
- Regress HSI as a function of globally available attributes

DEM based protocol→ HSI → Juvenile Steelhead Abundance (mechanistic model)

DEM

Fish Habitat (NREI,HSI)





CHaMP Surveys collect:

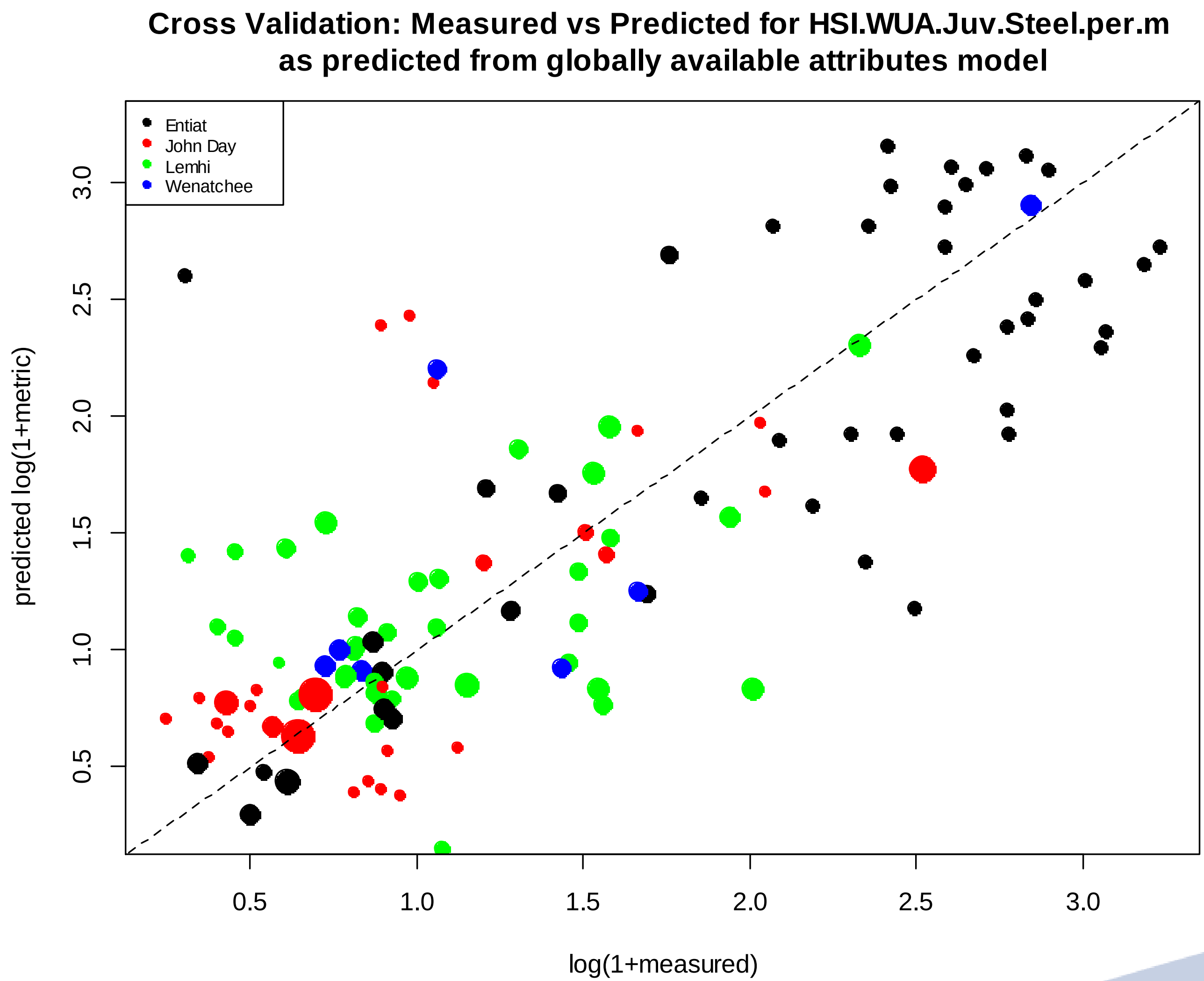
- Substrate
- Topography/Channel Units

e.g. GRTS rollup to get basin level abundance of redds or adults

Empirical Model for HSI Juvenile Steelhead Capacity

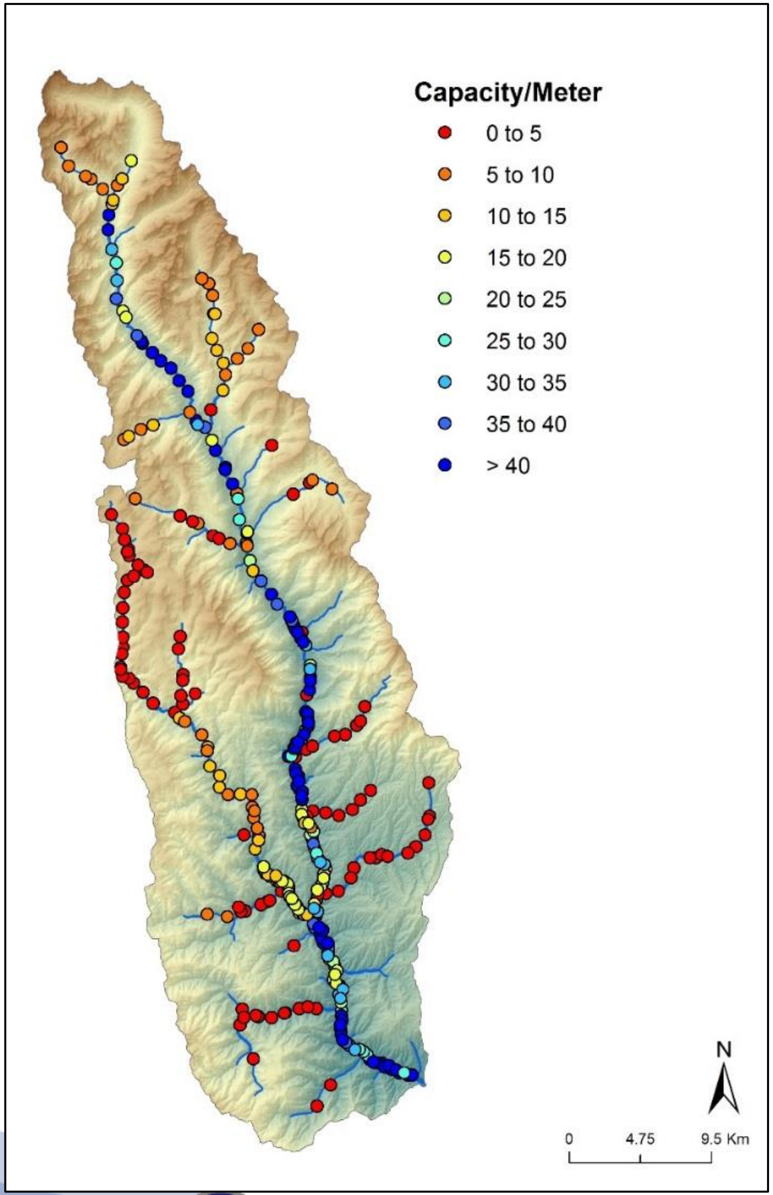
Empirical models relate globally available attributes to HSI and NREI estimates. Modeled Values are used to:

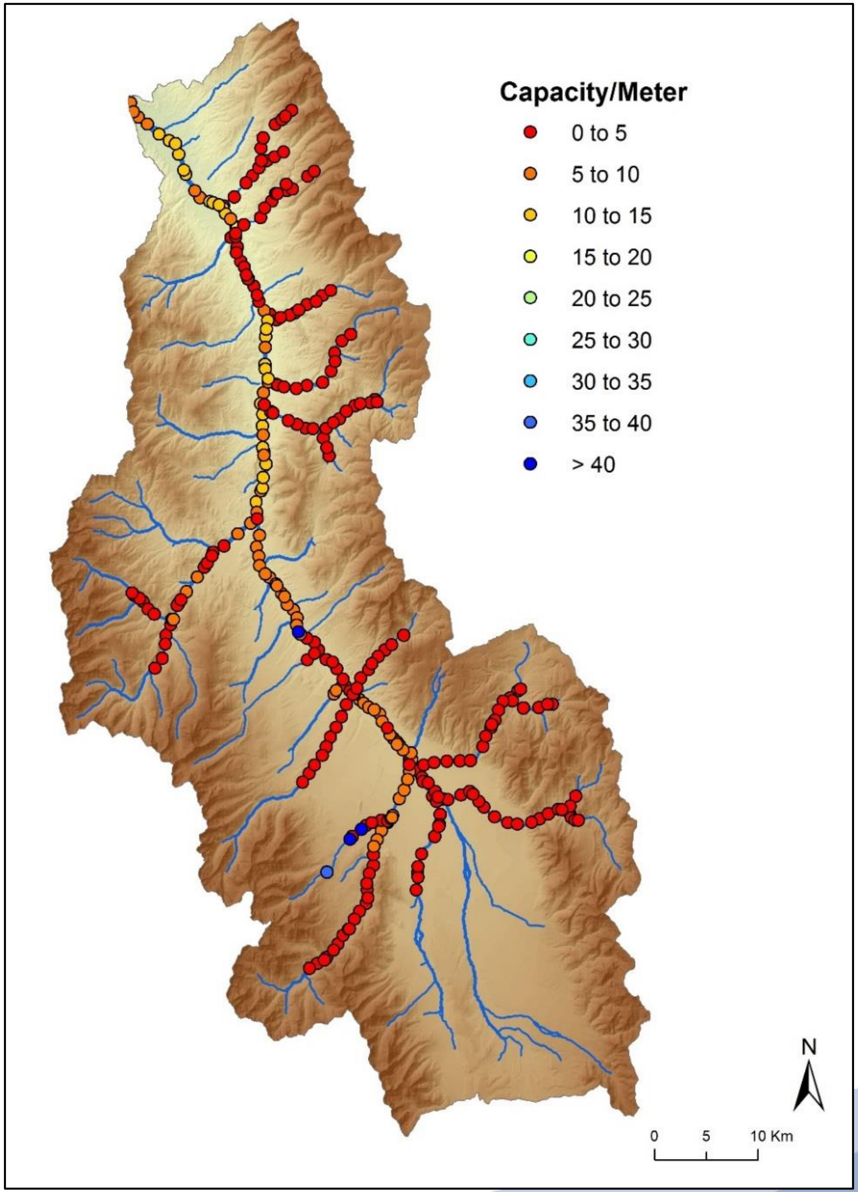
- Generate continuous estimates (maps)
- Estimate capacity in unmeasured regions
- Impute capacity to augment sparsely measured regions



(Dot Size ~ Sample Weight)

HSI: Juvenile Steelhead Capacity per Meter

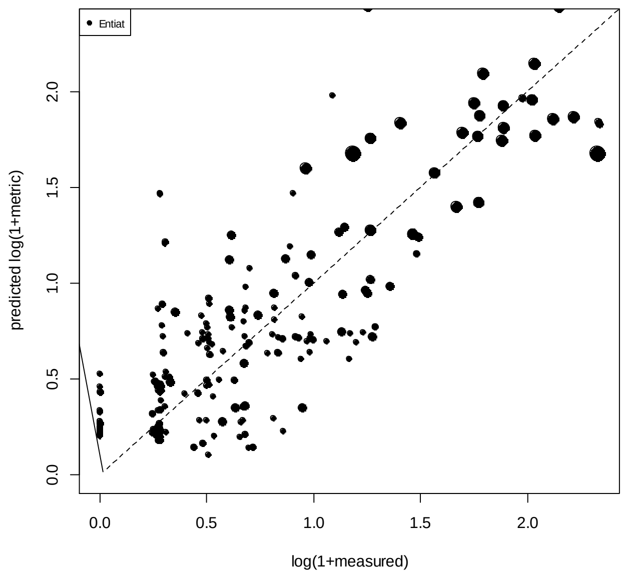


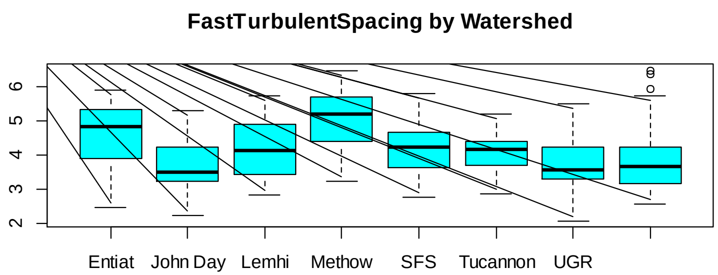


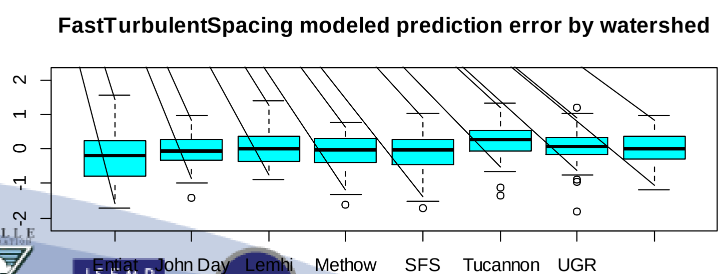
Model based regression example

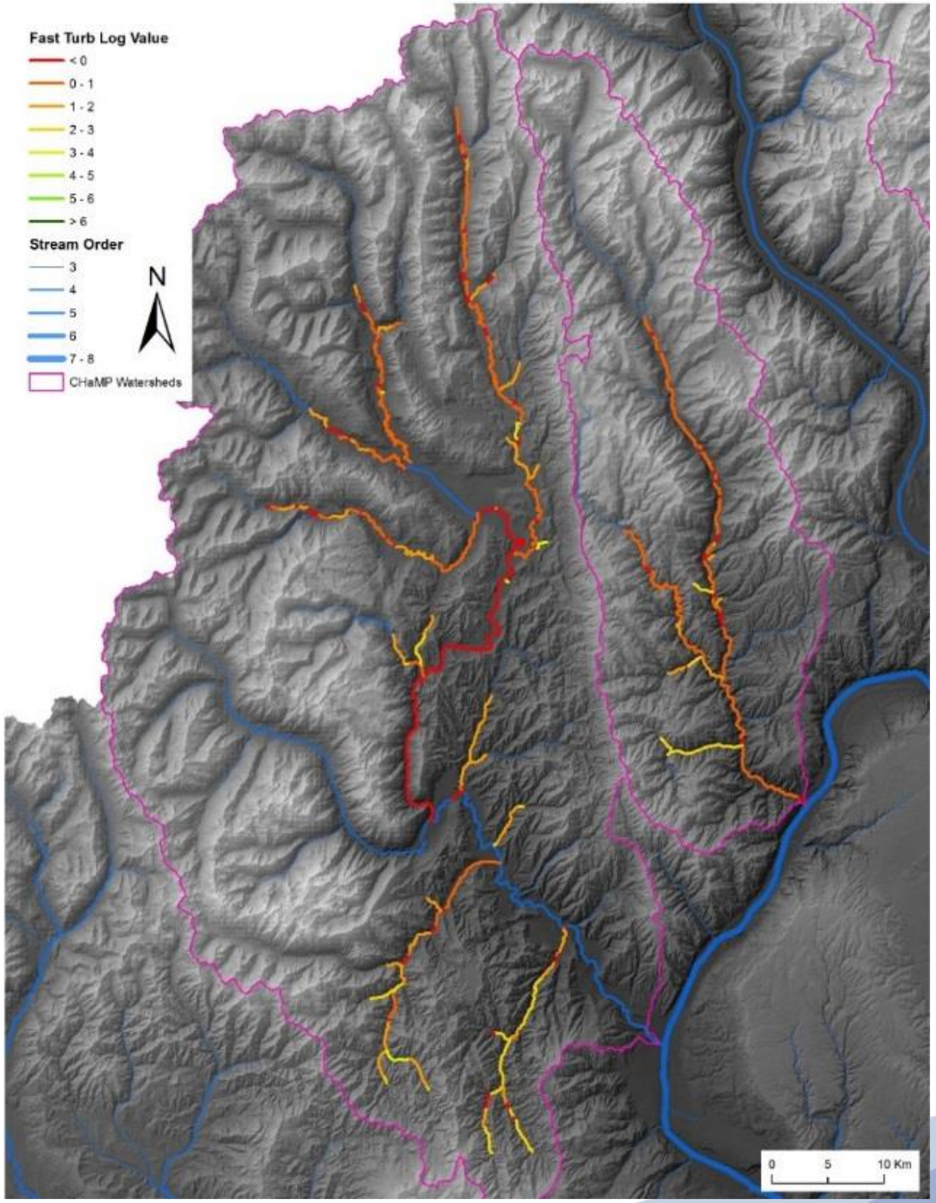
- Regress Fast Turbulent Spacing as a function of globally available attributes
- Predict Fast Turbulent Spacing in Non-CHaMP Watersheds

Fast Turbulent Spacing: Measured vs Modeled



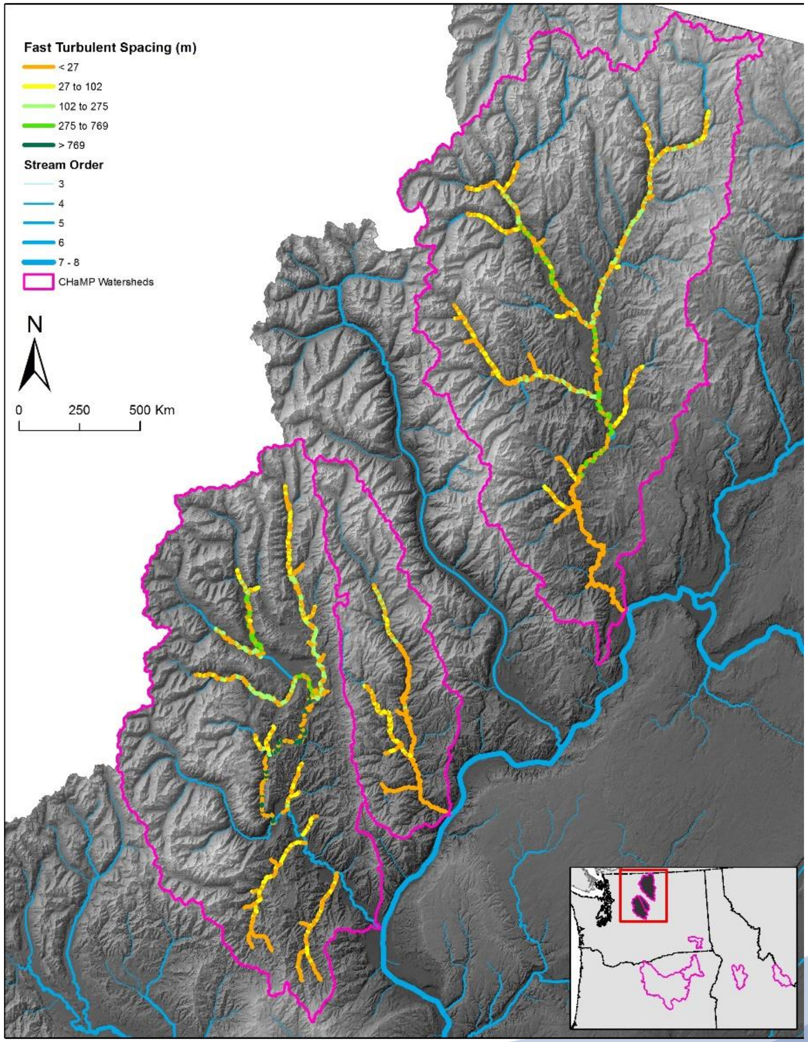






Continuous network estimation extrapolated to non-CHaMP watersheds

- Example: Fast Turbulent Spacing Estimates using "unified" model
  - Unbiased watershed-watershed
  - Less precise within each watershed than watershed-specific models
  - Useful (we hope) for extrapolation into unmeasured watersheds (for which CHaMP watersheds are "representative")



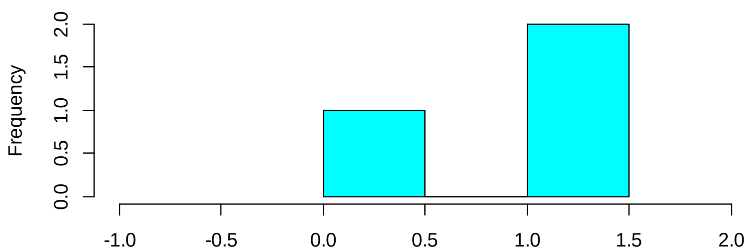
Model based regression example

- Regress Pool Frequency as a function of globally available attributes
- Augment limited CHaMP data with modeled predictions (imputation)

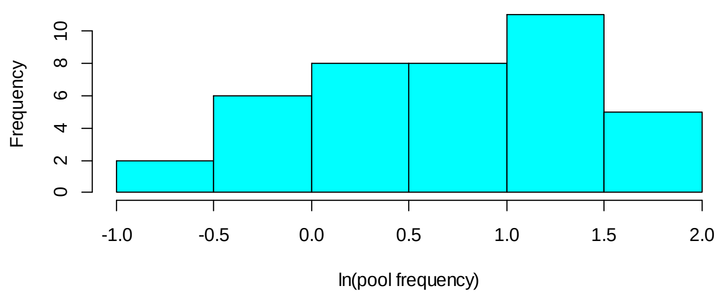
Example: Imputed and Model Based Continuous Estimation of Pool Frequency in the Little Wenatchee

Imputed Estimates

Measured ln(1 + Pool Frequency)



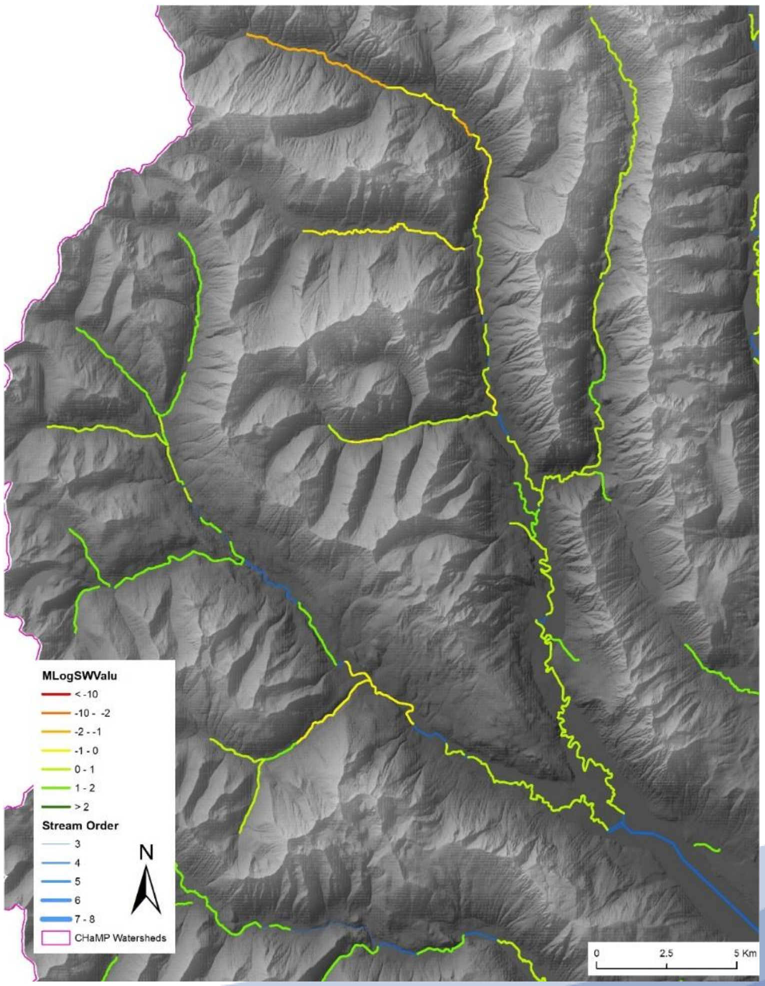
Modeled ln(1 + Pool Frequency)



Estimated Mean (ln 1 + Pool Frequency)

mean	2.50%	median	97.50%
1.284	0.9443	1.286	1.619

Model Based Continuous Estimates



Example: CHaMP Sampling and Choosing an analysis method

An introduction to CHaMP Sampling and Data Analysis

CHaMP Data Analysis

My watershed has treatment sites.

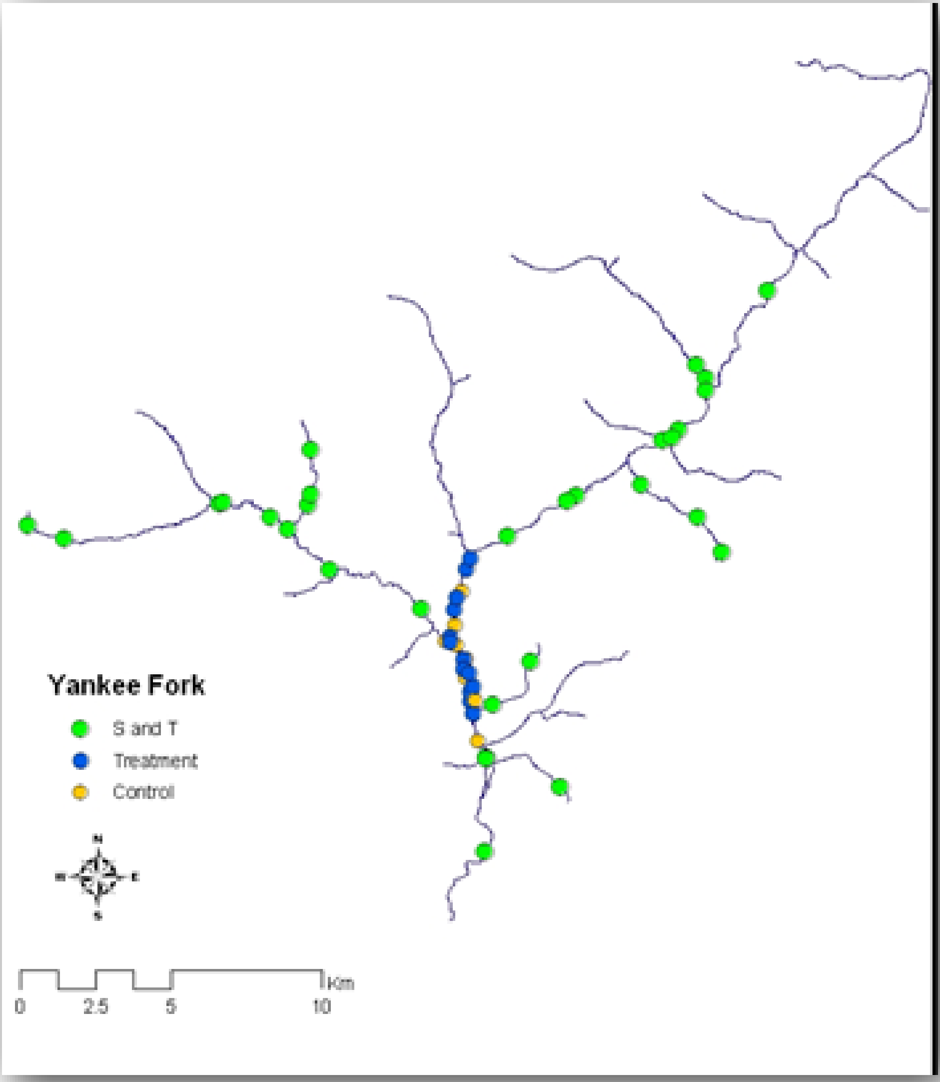
- How should I ensure I acquire data on treated sites, while maintaining a statistically valid (i.e. probabilistic) sampling strategy?
- What sort of analyses should I do?
  - Design based
  - Model based?

Tucannon restoration timeline (partial)

Site_ID	Stream	Sample	Sub_Treat	Sub_Loc	Sub_TrType	Treat2011	Treat2012	Treat2013	Treat2014	Treat2015	Treat2016	Treat2017	Treat2018
CBW05583-007039	Tucannon River	Annual	Control	Upper									
CBW05583-010495	Tucannon River	Annual	Treatment	Upper	LWD, SC					1	1	1	1
CBW05583-018303	Tucannon River	PY2	Treatment	Upper	LWD						1	1	1
CBW05583-038783	Tucannon River	PY1	Control	Upper									
CBW05583-047999	Panjab Creek	PY3	Tributary	Tributary									
CBW05583-051659	Tucannon River	PY2	Control	Upper									
CBW05583-057139	Tucannon River	PY1	Treatment	Upper	LWD, Levee						1	1	1
CBW05583-072139	Tucannon River	PY3	Control	Upper									
CBW05583-079743	Tucannon River	PY3	Control	Upper									
CBW05583-100223	Tucannon River	PY2	Control	Upper									
CBW05583-109611	Pataha Creek	PY3	Tributary	Tributary									
CBW05583-141567	Cummings Creek	PY3	Tributary	Tributary									
CBW05583-141771	Tucannon River	PY2	Control	Upper									
CBW05583-168191	Tucannon River	PY1	Control	Upper									
CBW05583-169855	Tucannon River	PY1	Treatment	Upper	LWD			1	1	1	1	1	1
CBW05583-170443	Tucannon River	Annual	Treatment	Upper	LWD, Levee						1	1	1
CBW05583-178047	Tucannon River	PY1	Control	Upper									
CBW05583-182527	Cummings Creek	PY2	Tributary	Tributary									
CBW05583-196787	Tucannon River	PY2	Control	Upper									
CBW05583-203211	Tucannon River	Annual	Treatment	Upper	Levee (2012), LWD		1	1	1	1	1	1	1
CBW05583-208767	Tucannon River	PY1	Treatment	Upper	LWD					1	1	1	1
CBW05583-212787	Tucannon River	PY1	Control	Upper									
CBW05583-214475	Tucannon River	Annual	Treatment	Upper	LWD, Levee						1	1	1
CBW05583-214911	Tucannon River	PY2	Control	Upper									
CBW05583-222251	Tucannon River	PY1	Control	Lower									
CBW05583-248063	Tucannon River	Annual	Treatment	Upper	LWD, SC					1	1	1	1
CBW05583-256895	Little Tucannon River	PY1	Tributary	Tributary									
CBW05583-274303	Tucannon River	PY3	Control	Upper									
CBW05583-276351	Tucannon River	Annual	Control	Upper									
CBW05583-310143	Panjab Creek	PY1	Tributary	Tributary									
CBW05583-327859	Tucannon River	PY1	Control	Upper									
CBW05583-329599	Cummings Creek	PY2	Tributary	Tributary									
CBW05583-339839	Tucannon River	Annual	Control	Upper									
CBW05583-345983	Tucannon River	PY2	Control	Upper									
CBW05583-353323	Tucannon River	PY3	Control	Lower									
CBW05583-384819	Tucannon River	PY3	Control	Upper									
CBW05583-386091	Tucannon River	Annual	Treatment	Lower	LWD, Levee, SC					1	1	1	1

Combining Status and Trend and Effectiveness Monitoring Designs: Yankee Fork

- Started in 2013
- 2 strata: Status and Trend and Restoration Areas
- Phased restoration with planned before/after sampling resulted in unique 'Step Panel' sampling approach
- Combines AEM and Status and Trend Objectives.
- Status and Trend Sites used as Reference for AEM sites (provides control at different scales)



CHaMP Data Analysis

- Objective: What is the mean LWD by subgroup "treatment" (By Year and Average of all Years)
- Question: Which sort of analysis (design or model based) should I use?

CHaMP Data Analysis

- Objective: What is the effect of restoration on LWD?
- Question: Which sort of analysis (design or model based) should I use?

Outline of model for analysis for estimating the effect of restoration on LWD

- Mixed effects model including:
  - Dependent variable:
    - LWD
  - Independent variables:
    - include:

rs)

Stre

Sub_Treat, Sub_loc, SubTrType

Incorporate sampling design into analysis!

Working with CHaMP Statisticians

Helping us help you

Working with CHaMP Statisticians

- Be clear on analysis objective:
  - Objective, in conjunction with data and sampling design, drives analysis strategy

- Carefully define spatial region(s) of interest
  - Watersheds
  - Subgroups within watersheds

- Be patient. ☺
  - Your statistician may know less about fish biology than you know about design and model based analysis

CHaMP Data Analysis

In Summary:

Sampling design needs to be taken into account during any and all analyses of CHaMP data

Introduction to CHaMP sampling and Analysis

- Acknowledgements
  - Carol Volk
  - Phil Larsen
  - Kevin See
  - Chris Jordan
  - Boyd Bouwes