Generalization of Hydrographic Features and Automated Metric Assessment Through Bootstrapping

Larry Stanislawski, ATA, Center of Excellence for Geospatial Information Science (CEGIS), USGS, Rolla Missouri

Barbara Buttenfield, Geography, University of Colorado-Boulder

Vijay Samaranayake, Math and Statistics, Missouri University of Science and Technology

Research Assistants:
Ryan Haney, Jeremy Koontz, and **Otto Schnarr III**, USGS
Chris Anderson-Tarver CU-Boulder

13th Workshop of the ICA Commission on Generalization and Multiple Representation, Zurich, Switzerland, September 12-13, 2010

Outline

- CEGIS Generalization project objectives
- National Hydrography Dataset (NHD) generalization workflow
- Generalization of high-resolution NHD for a humid-hilly subbasin in Missouri
- Coefficient of line correspondence (CLC)
- Non-parametric method (bootstrap) to estimate confidence interval for CLC
- Summary Statements

CEGIS Generalization Project

Objective: research and develop automated methods for generalization to support multiple-scale display and delivery of *The National Map* and other USGS geographic data.

Basic steps in generalization research:

1) Establish feature prominence estimates for features within data themes and relations with display scale,
2) Data partitioning to maintain valid local density variations
3) Conditional feature pruning that maintains data model integrity
4) Geometric operations (simplify, merge, aggregate, etc.) that support graphic production at multiple mapping scales
5) Quality assurance to validate results with existing data, and include a statistical confidence level

Generalization Work Flow to Produce a Level of Detail (LoD) from High-Resolution NHD

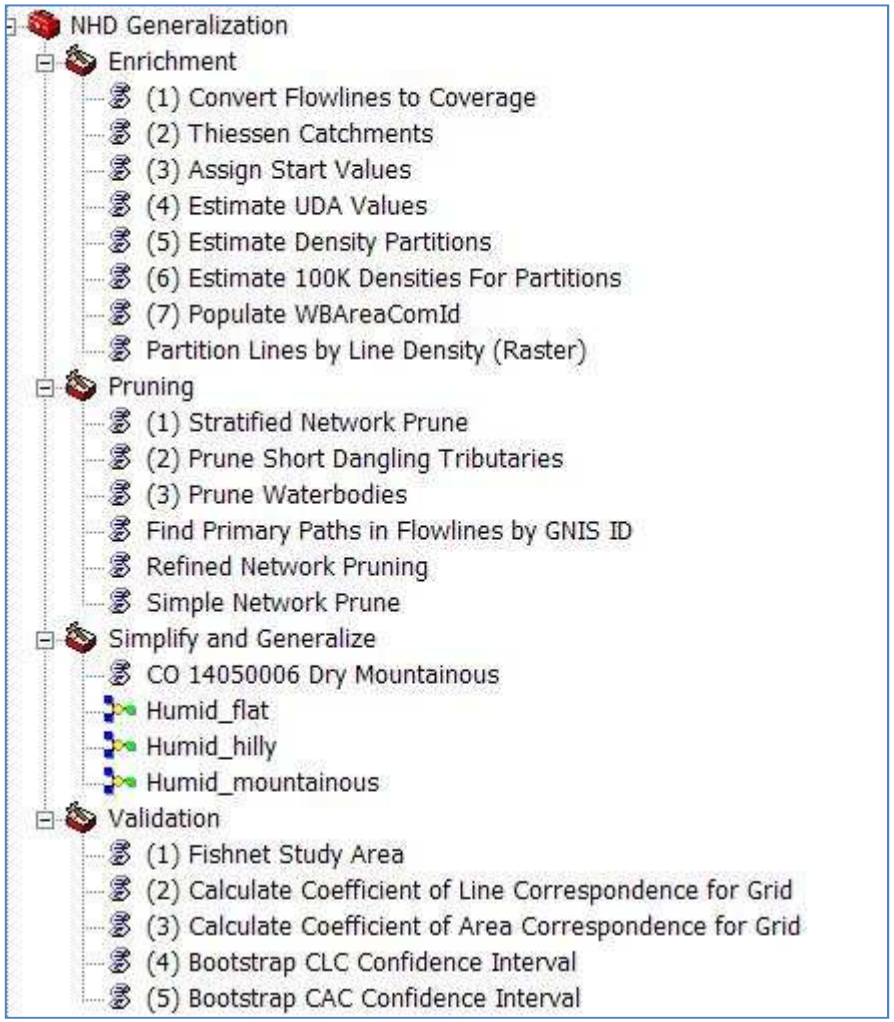
Develop automated methods to:

1) Enrich data with feature prominence estimates (UDA)
2) Prune features from high-resolution NHD layer
3) Simplify or further generalize remaining features for cartographic display (LoD)
4) Validate (refine procedures as needed)

Constraints = methods must preserve:

1) Connectivity (topology) of hydrographic network
2) Local density variations that typify physiographic or climate variations
3) Full reaches
4) Complete attribution
5) For 50K LoD, geometric characteristics (e.g. vertex spacing, curve shape) and feature type categories similar to the 100K

NHD Generalization Toolbox

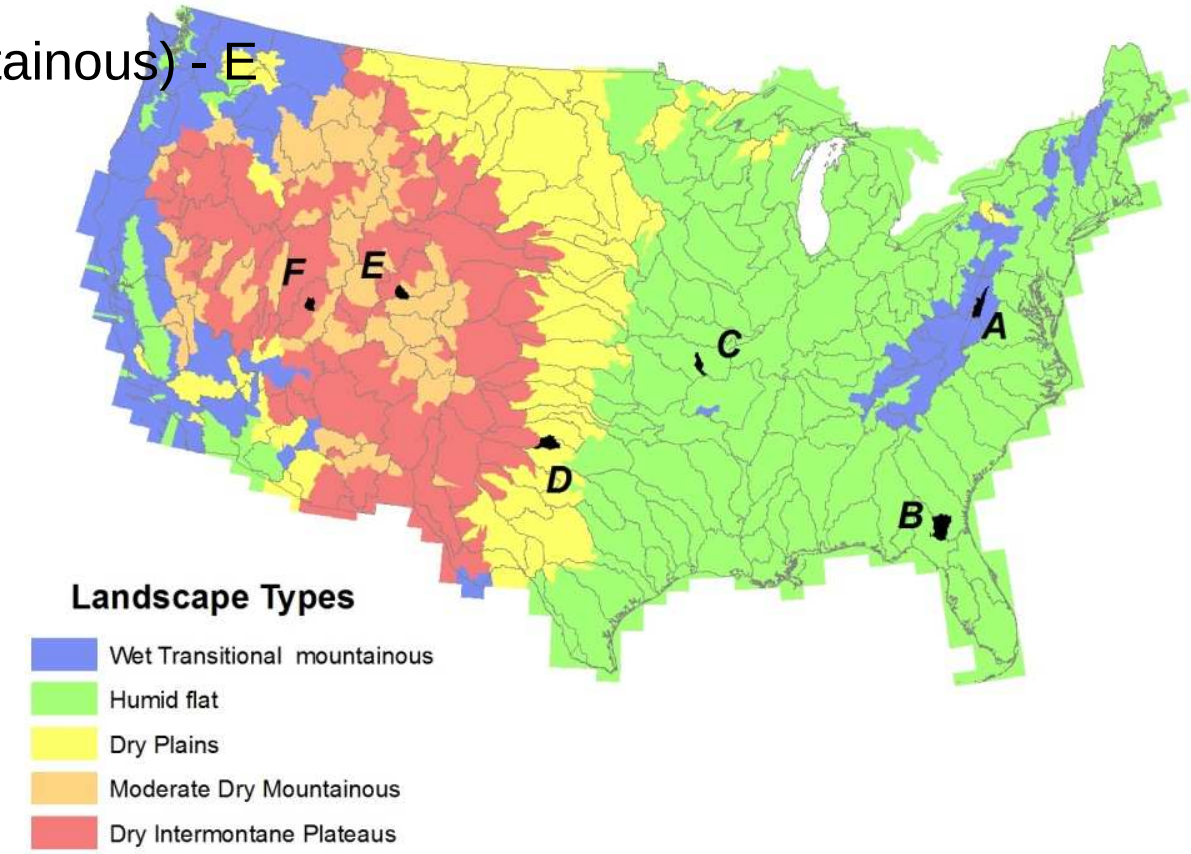


USGS
science for a changing world

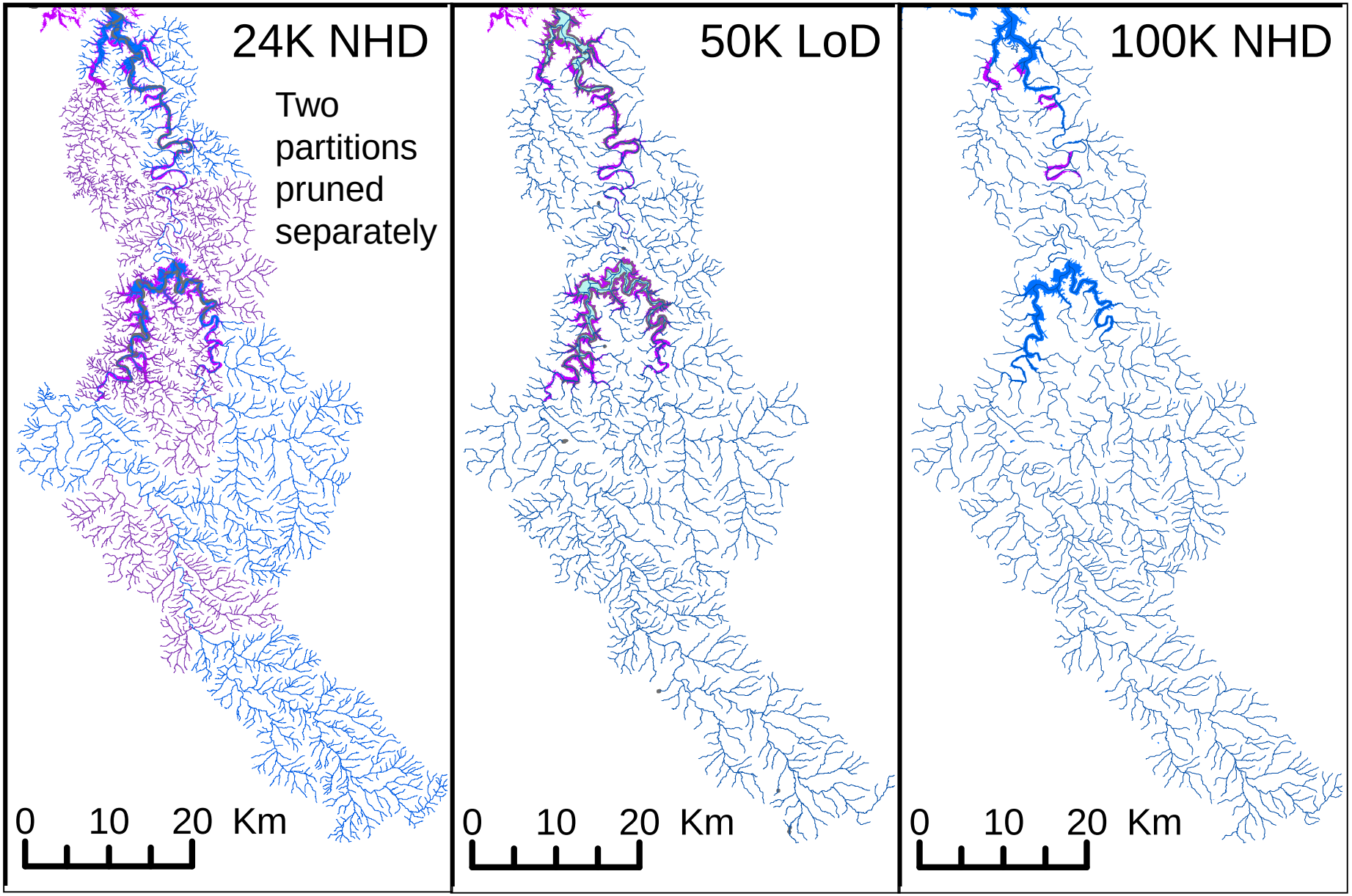
50K LoDs (six subbasins)

50K LoDs processed from high-res NHD

- West Virginia (humid-mountainous) – A
- Florida-Georgia (humid-flat) – B
- **Missouri (humid-hilly) – C**
- Texas (dry-hilly) - D
- Colorado (dry-mountainous) - E
- Utah (dry-flat) - F



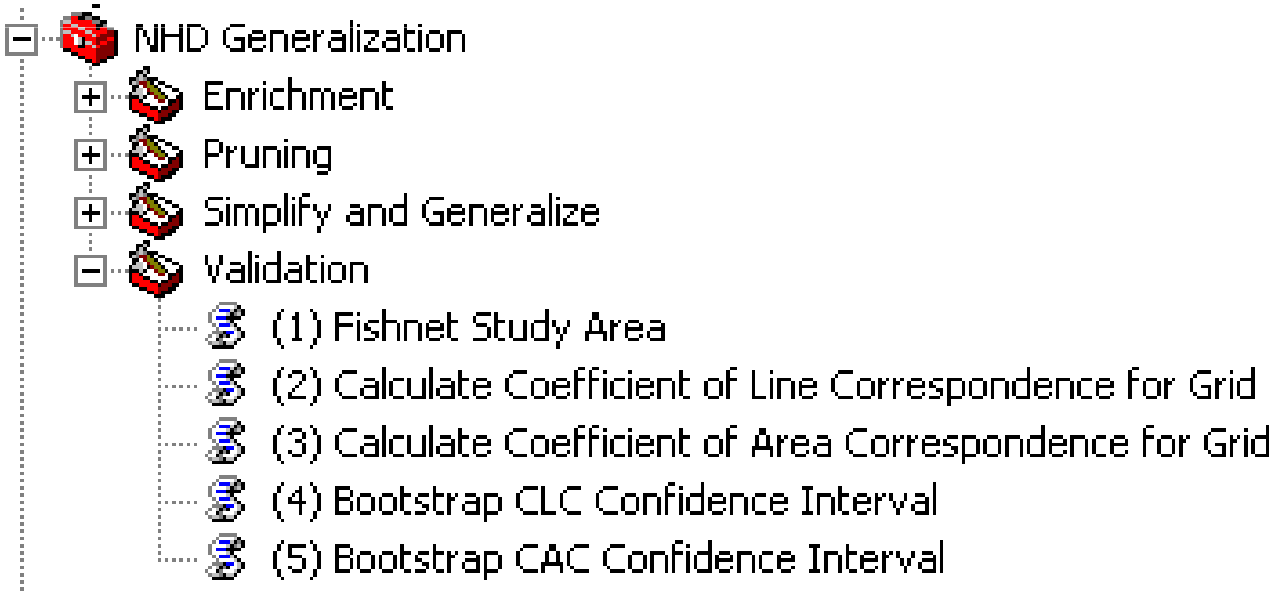
Humid Hilly Missouri Subbasin



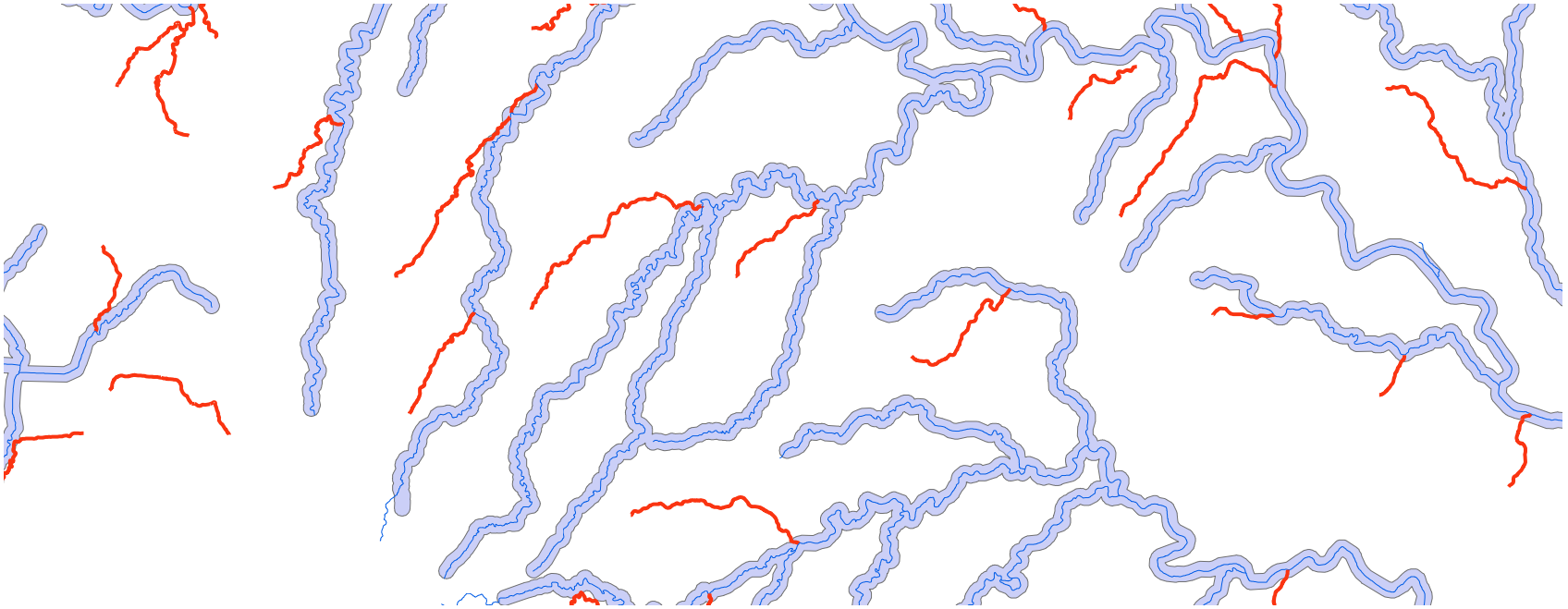
Validation

Automated comparison between generalized data against benchmark (100k NHD)

- Coefficient of Line Correspondence (CLC) and
- Coefficient of Area Correspondence (CAC)



Validation: Commission Errors



Pruned High-Resolution NHD Flowline Network
Commission Error
Match
Buffer of Benchmark Lines

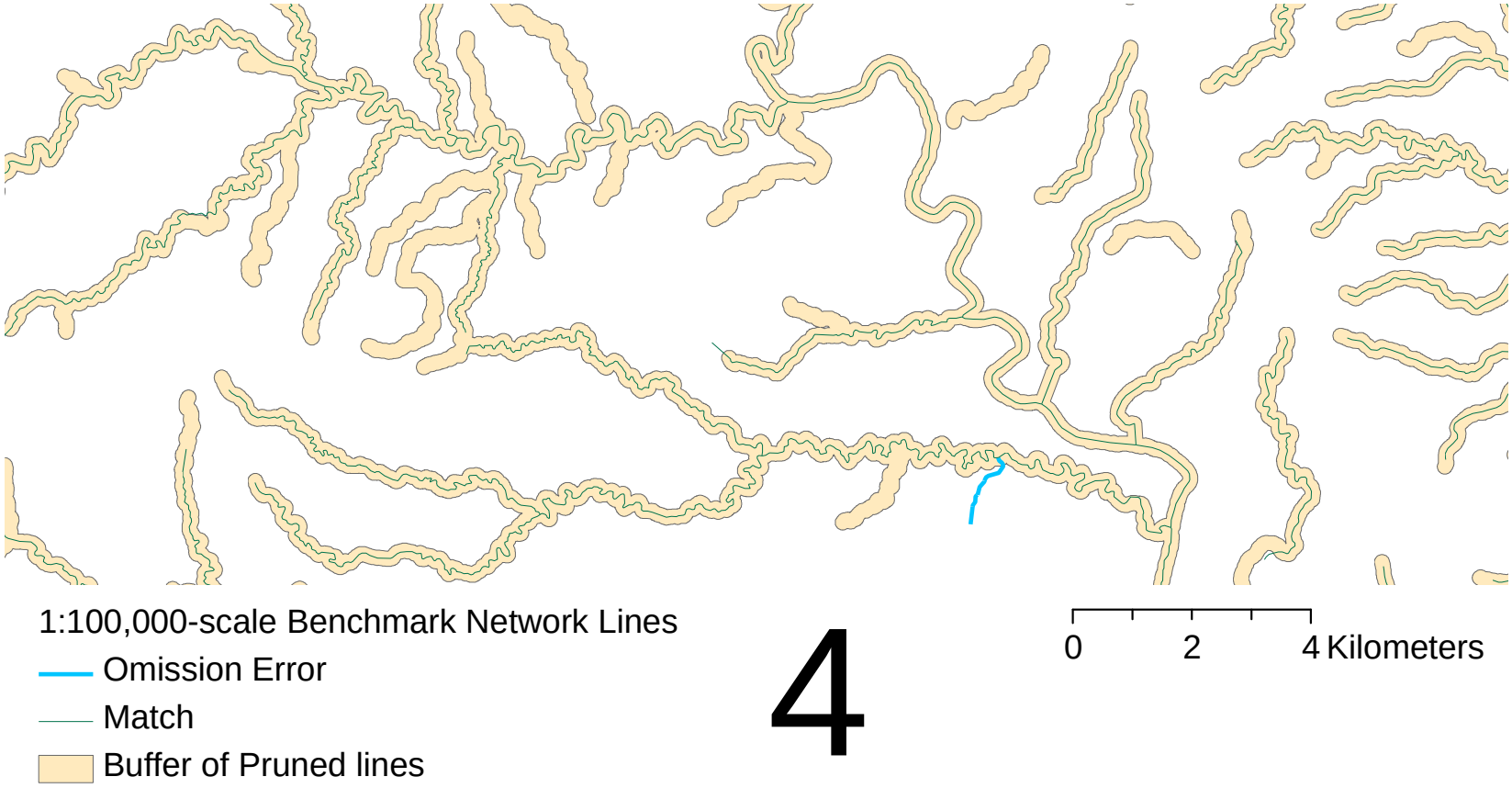
4

0 2 4 Kilometers

Buffer around benchmark lines

Buffer is two times US National Map Accuracy Standards at scale of generalized dataset and scale of benchmark dataset.

Validation: Omission Errors



1:100,000-scale benchmark lines over buffer of pruned lines

Omission error where > 50% confluence-to-confluence feature outside of buffer

Validation
Coefficient of Line Correspondence (CLC)

CLC = M / (O+C+M)

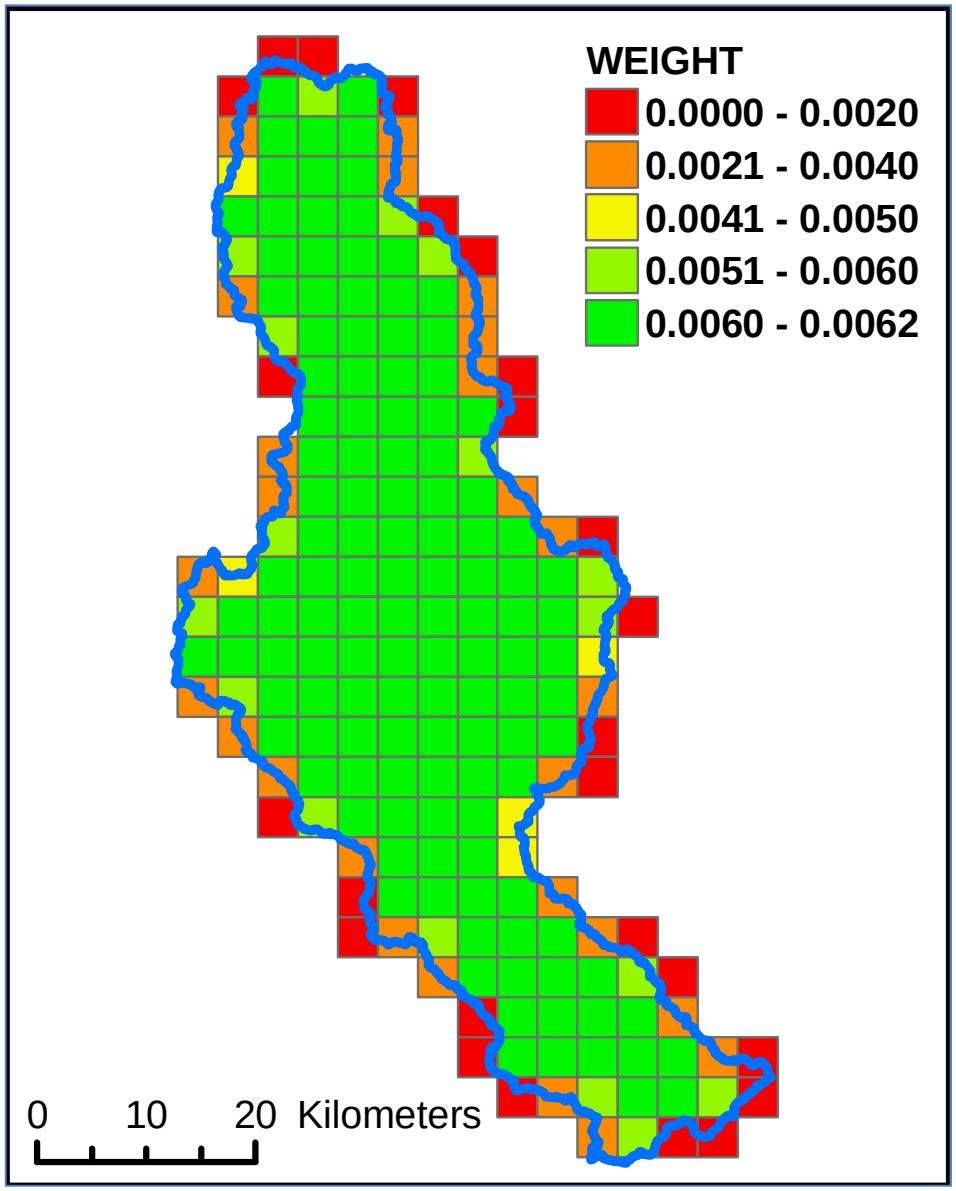
- M = sum of length of matching features from benchmark (1:100,000-scale) dataset
- O = sum of length of omission error features from the benchmark (1:100,000-scale) dataset
- C = sum of length of commission error features from the pruned (high-resolution NHD) line dataset *

Proportion Commission Errors = C / (O+C+M)

Proportion Omission Errors = O / (O+C+M)

* Commission lengths are divided by the benchmark-to-LoD length expansion factor to compensate for higher granularity in LoD representations.

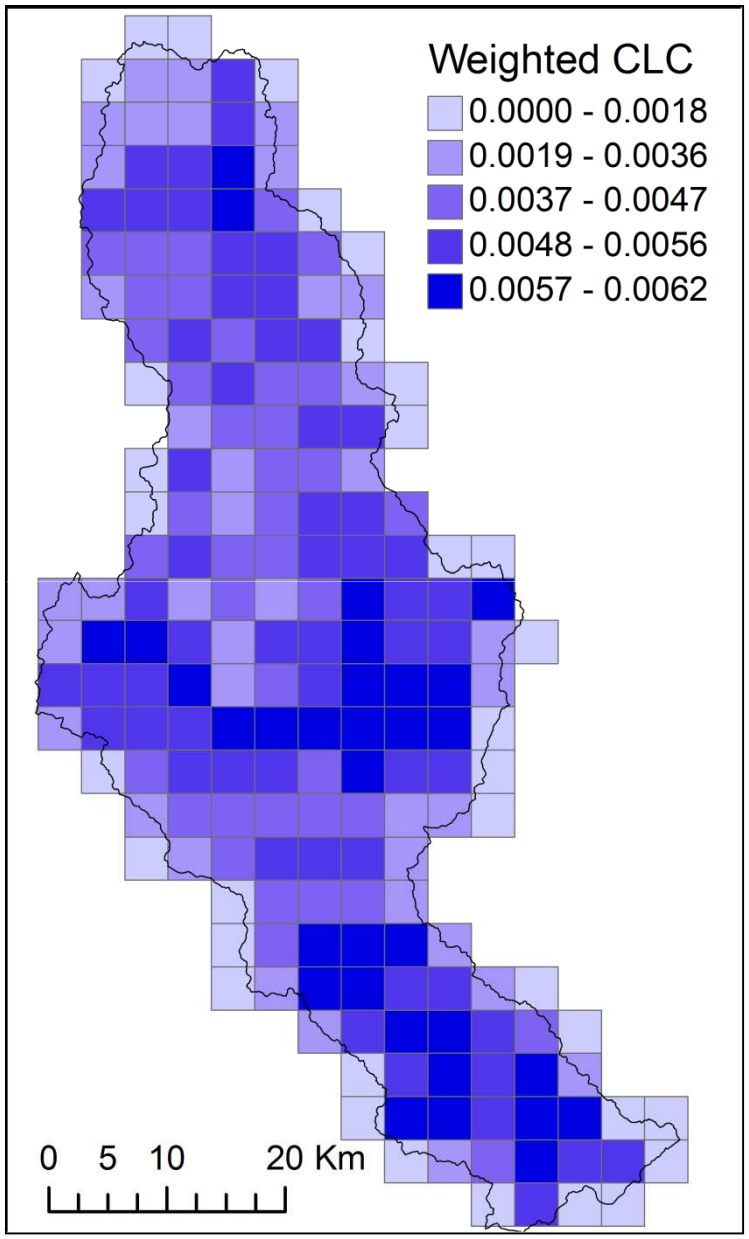
Validation: Weights for CLC and CAC



Comparison of 50-200K Level of Detail (LoD) with 100K NHD.

Pomme de Terre, MO. Subbasin 10290107

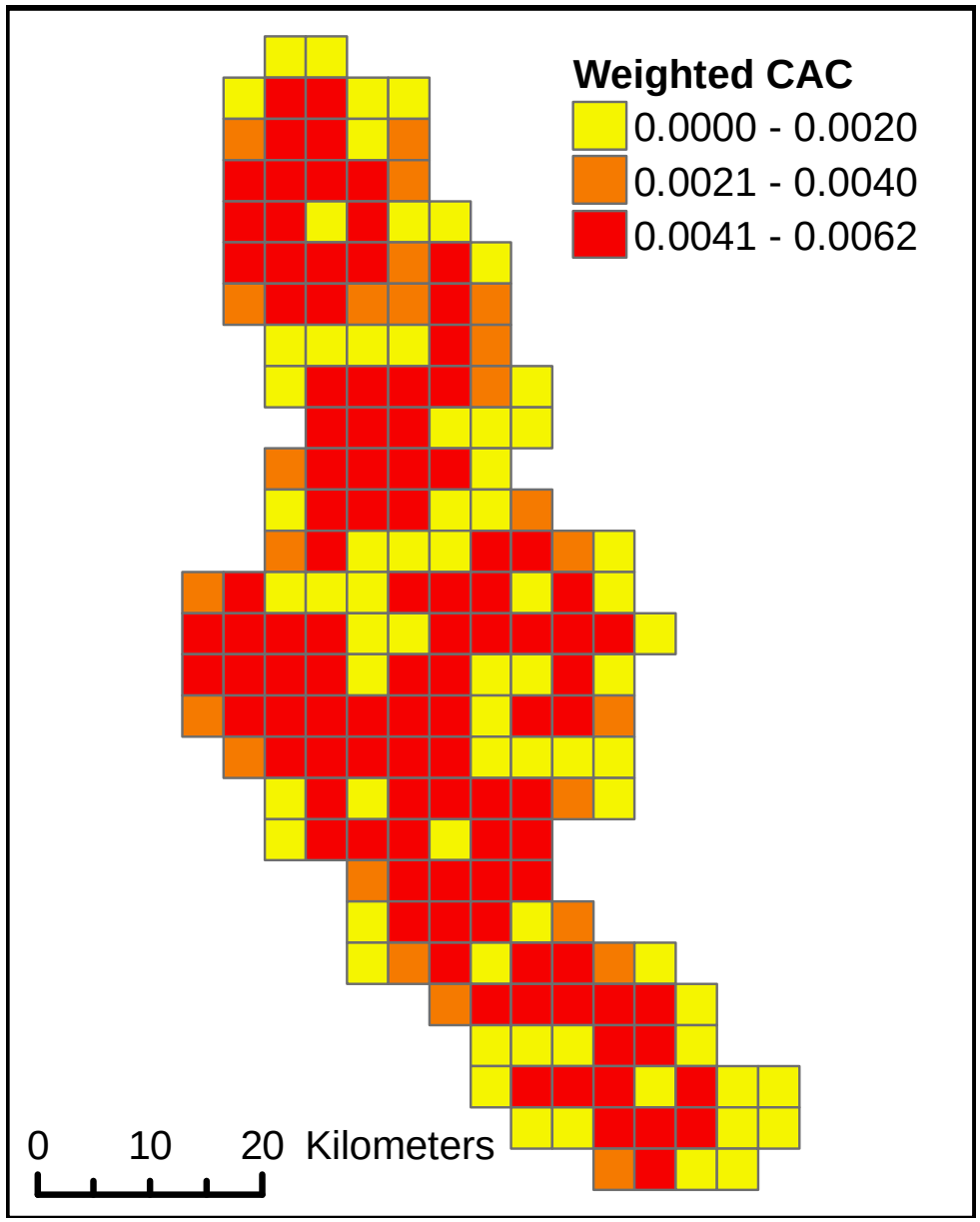
Validation: Weighted CLC



Pomme de Terre, MO.
Subbasin 10290107

Subbasin weighted CLC:
0.8020

Validation: Weighted CAC

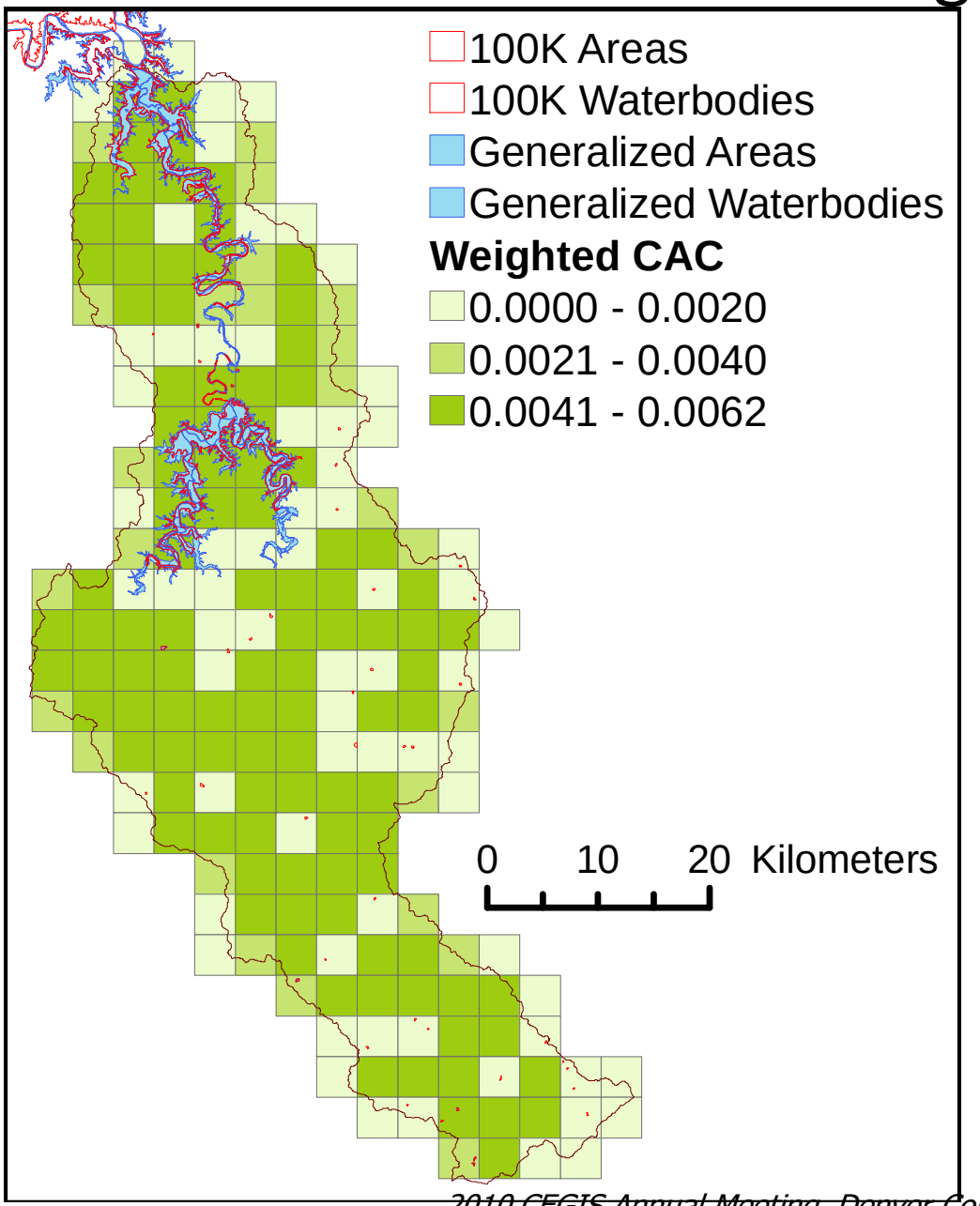


Pomme de Terre, MO.
Subbasin 10290107

Subbasin weighted CAC:
0.72

USGS
science for a changing world

Validation: Weighted CAC



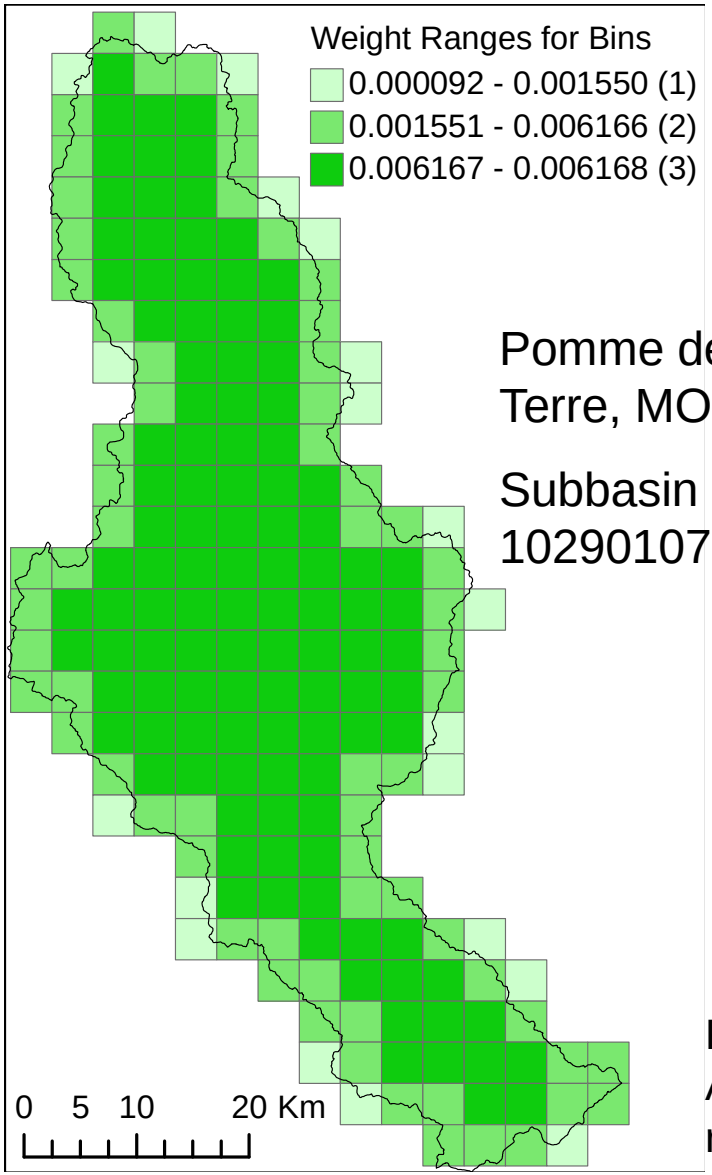
Pomme de Terre, MO.
Subbasin 10290107

Subbasin weighted CAC:
0.72

Many cells do not have any features and are recorded as perfect match. Is this unacceptable bias, perhaps favoring dry areas with few waterbodies?

USGS
science for a changing world

Bootstrap confidence intervals for CLC and CAC



Bootstrapping

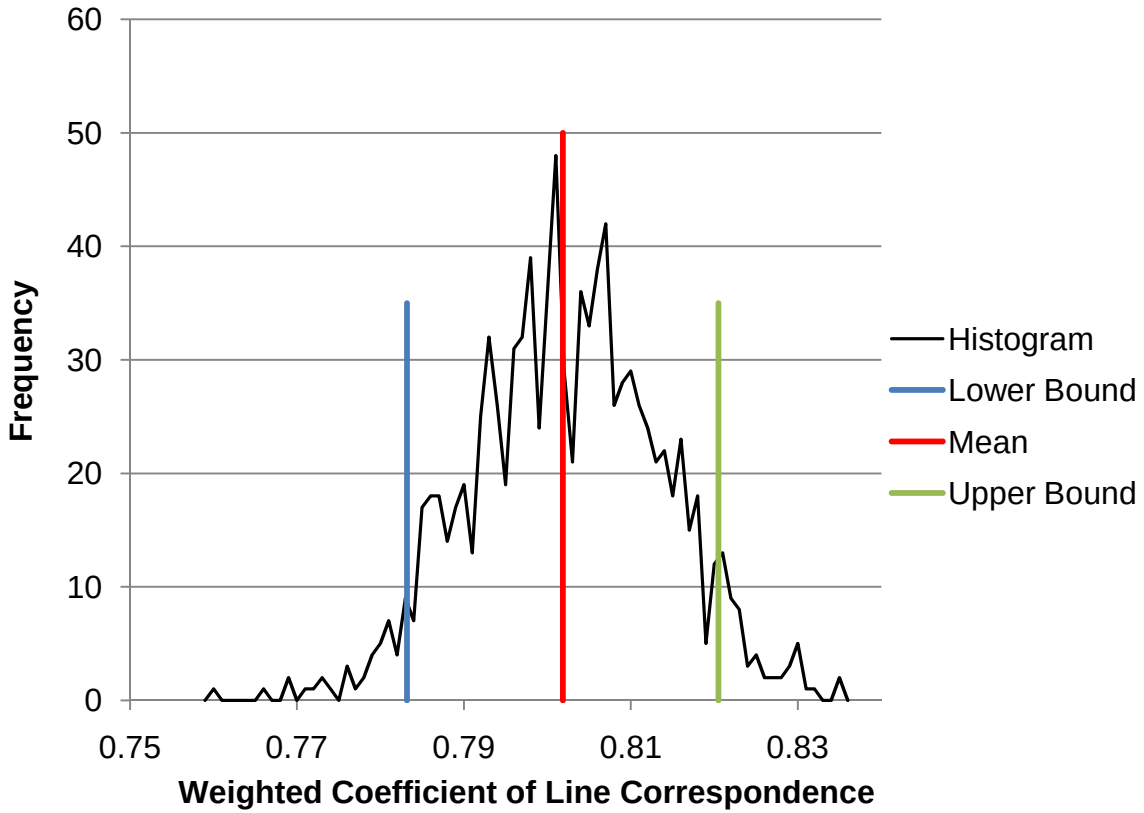
- General approach to statistical inference when underlying distribution is unknown
- Build sampling distribution by resampling the data at hand (Fox 2002)

Bin assignments ensure that weights are distributed in similar proportions during resampling.

Bin 1: lowest 20 weights
Bin 2: cells between bins 1 and 3
Bin 3: cells with highest weight

Fox, J. 2002. *Bootstrapping regression models.* Appendix to an R and S-PLUS companion to applied regression.

Bootstrapping Weighted CLC Confidence Interval



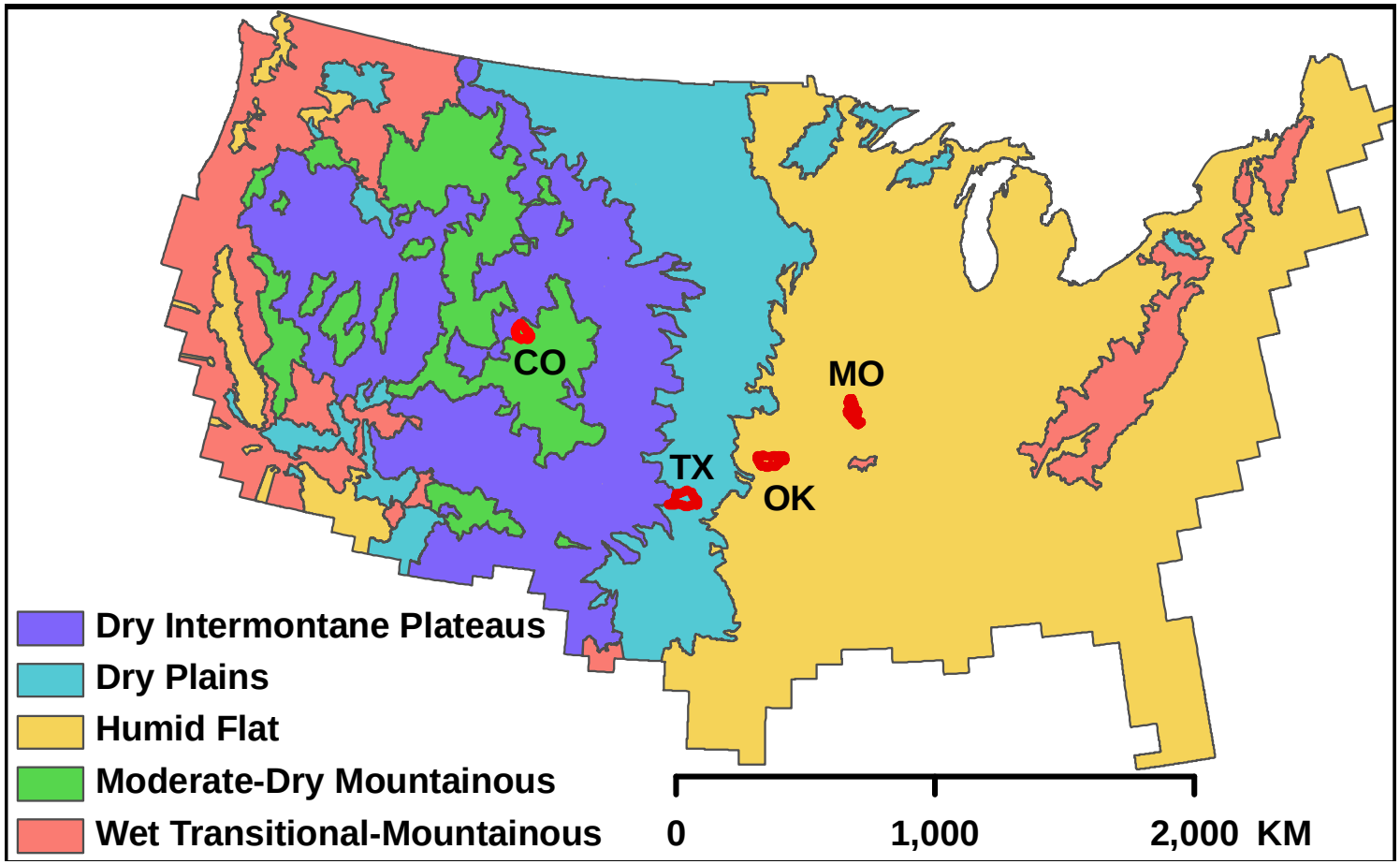
Pomme de Terre, MO. Subbasin 10290107

- 200 cells resampled, with replacement and proportional selection within bins
- Generate 1,000 weighted CLC values for the subbasin.
- Mean and confidence interval determined from the 1,000 weighted CLC values.

Results:
90% Confidence interval: 0.78 to 0.82
Mean: 0.8018, Median: 0.8014, Mode: 0.801 (0.001 bins)
(Mean, median, and mode are equal in normal distribution.)

USGS science for a changing world

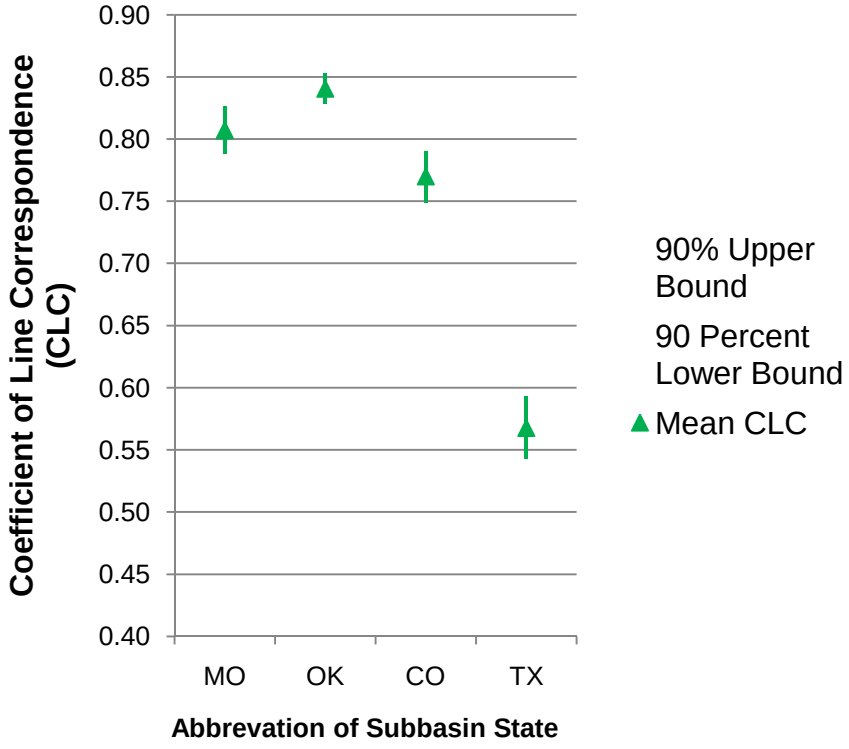
Bootstrapping CLC and CAC Confidence Intervals for four subbasins

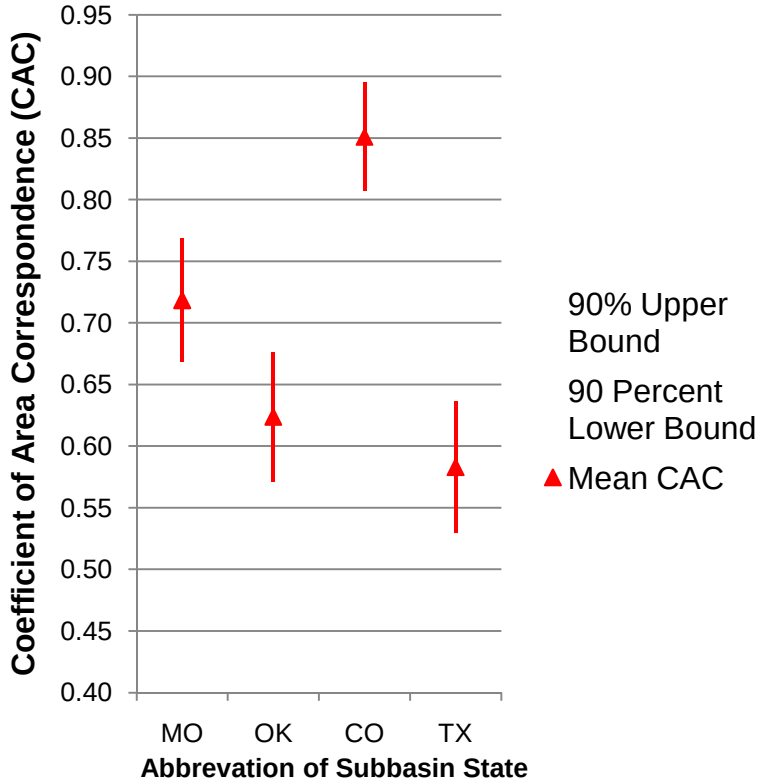


USGS
science for a changing world

Bootstrapping CLC and CAC Confidence Intervals for four subbasins

Compare HR NHD generalized to 50-200K LOD with 100K NHD





USGS science for a changing world

Summary Statements

- CLC and CAC are simple automated methods to compare two different representations of the same line and polygon features, respectively.
- Non-parametric bootstrapping process appears to be valid approach for estimating reliability (confidence interval) of the CLC and CAC.
- CLC and CAC with confidence levels provide a method for comparing generalization results and to test for significant improvements of different generalization alternatives (assists refinement).
- The CLC/CAC are validation tools that will assist the USGS in developing generalization procedures over the various physiographic conditions within the United States.
- CLC/CAC validation methods may be used in constraints for arriving at adequate generalization solutions.