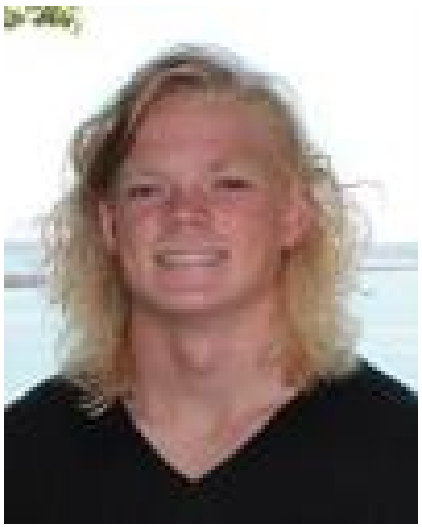
Conformal prediction for timber damage after wildfires

Jerzy Wieczorek
Colby College
jerzy.wieczorek@colby.edu

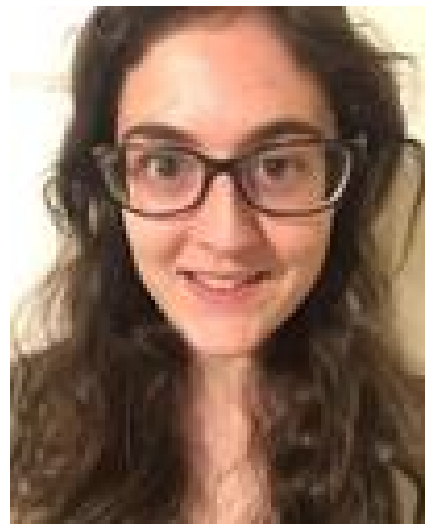
April 25, 2023
Natural Science Division Lunch

Part of a project with Colby College undergraduates



Zach Cody '23

Emily Tan '23



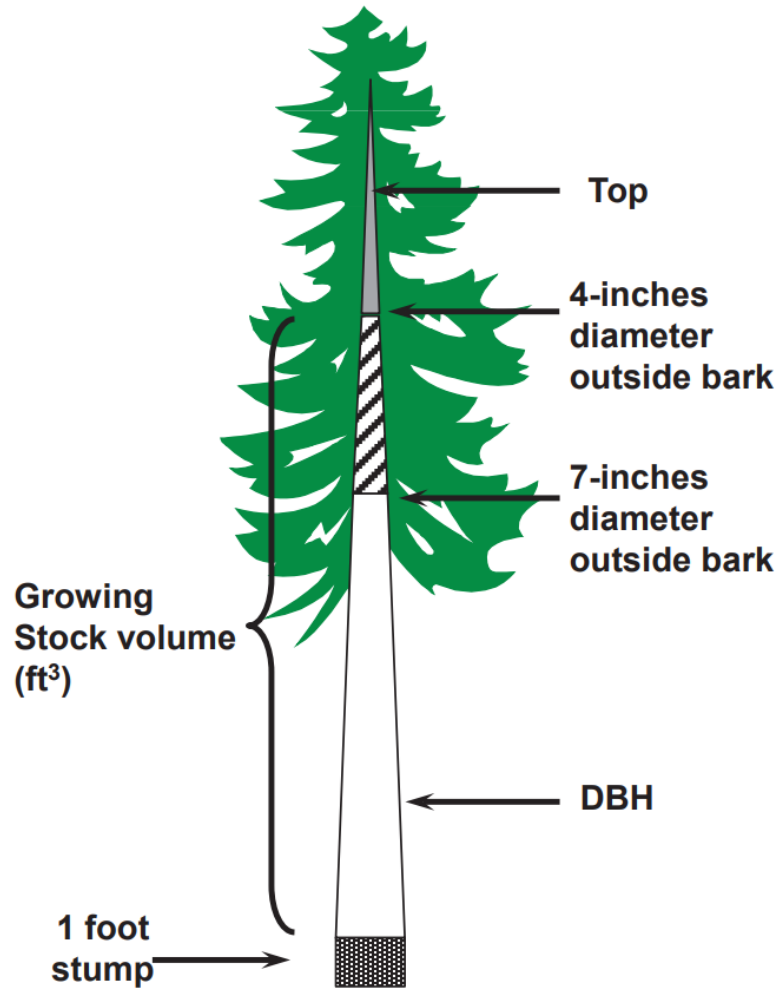
Jackie Chistolini '24J

See their work at:
https://github.com/ColbyStatSvyRsch/FIA-simpop-app

Motivating example: Forest Inventory & Analysis

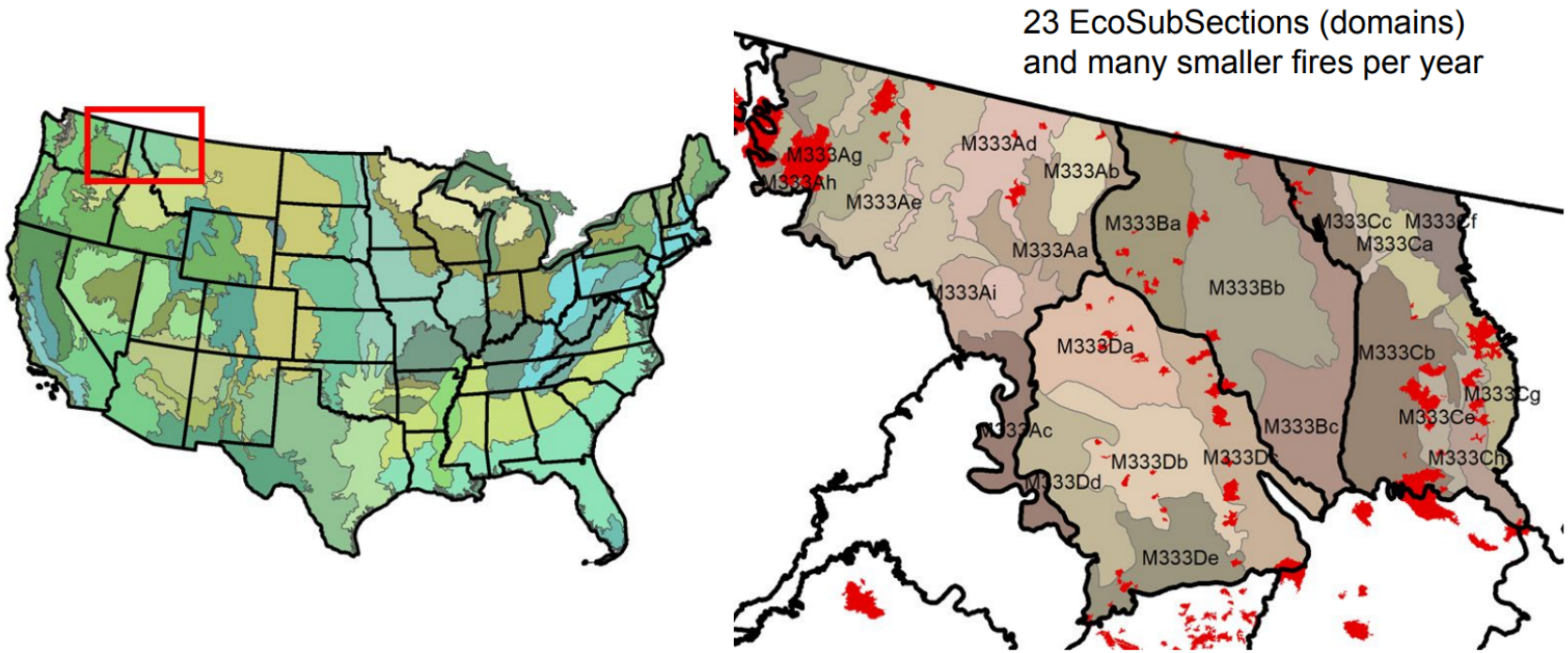
The FIA program of the US Forest Service tracks data on US forests, including wood production.

Example: estimating the volume (in board feet) of commercially valuable timber in a forest stand.
To directly measure the timber volume in a stand, need to send field crews for a long day of measurements.
Expensive, slow... and can't be done retroactively after a fire!



Berg et al. (2015)

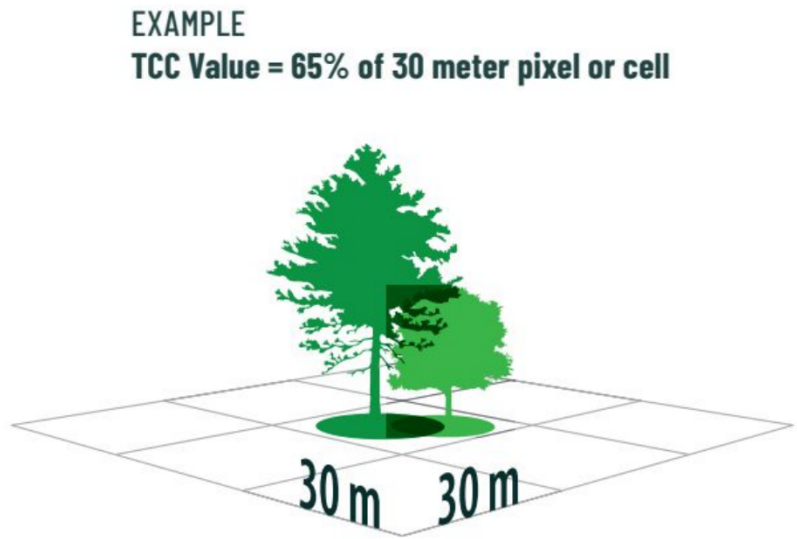
Our region of interest: EcoProvince M333 (with fires)



FIA data sources

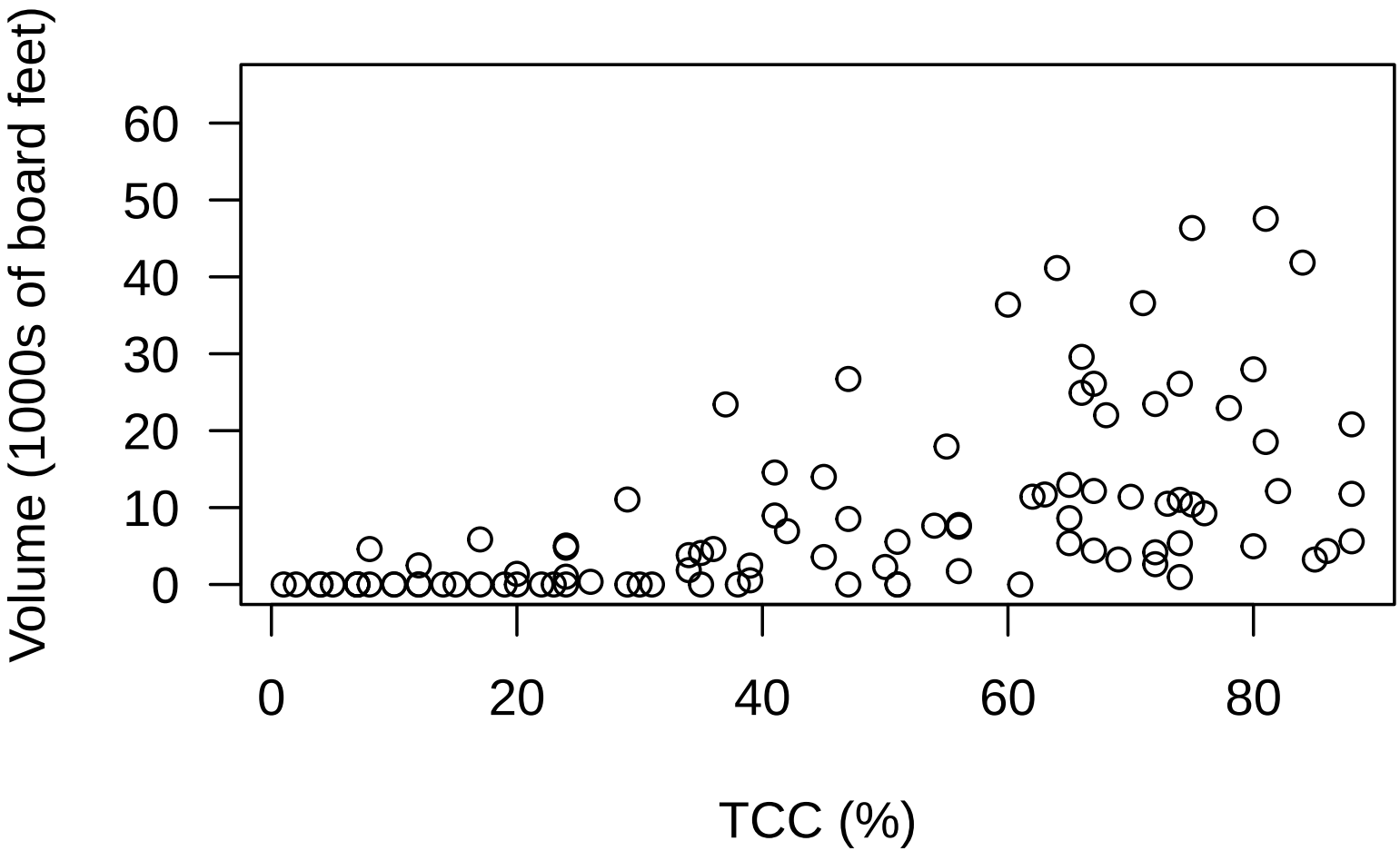
"Survey data": FIA field crews visit a rotating panel of randomly sampled sites, updating data every few years.
Each site fits in an approx 90m by 90m square.
Sampled only once every 6000 acres.

"Auxiliary data": Satellite imagery; precipitation and temperature records; elevation and terrain roughness; etc.
*Known for **every** 90m by 90m pixel in the US.*



TCC = Tree Canopy Cover
USFS (2020)

Simple data example: predict Volume using TCC



FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).

FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Use the models to get estimates for pixels that field crews didn't visit.

FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Use the models to get estimates for pixels that field crews didn't visit.
- **Today's talk:**
  Find **prediction intervals (PIs)** for these estimates.

FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Use the models to get estimates for pixels that field crews didn't visit.
- **Today's talk:**
  Find **prediction intervals (PIs)** for these estimates.
- Use during or after wildfires:

FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Use the models to get estimates for pixels that field crews didn't visit.
- **Today's talk:**
  Find **prediction intervals (PIs)** for these estimates.
- Use during or after wildfires:
  - If the fire reaches here, we expect the volume of timber lost will be between A and B thousands of board feet.

FIA project goals

- Using sampled sites, build statistical models to predict survey variables (eg timber volume) from auxiliary data (eg TCC).
- Use the models to get estimates for pixels that field crews didn't visit.
- **Today's talk:**
  Find **prediction intervals (PIs)** for these estimates.
- Use during or after wildfires:
  - If the fire reaches here, we expect the volume of timber lost will be between A and B thousands of board feet.
  - We estimate that before the fire, there were between A and B thousands of board feet of valuable timber here.

What are prediction intervals (PIs)?

- Point estimate:
  For a pixel with these auxiliary data values
  (TCC is 80%, elevation is 1200m, temperature is. . .),
  we predict it will have timber volume of 17,000 board feet.

What are prediction intervals (PIs)?

- Point estimate:
  For a pixel with these auxiliary data values
  (TCC is 80%, elevation is 1200m, temperature is. . .),
  we predict it will have timber volume of 17,000 board feet.

- 90% **confidence interval**:
  Among all pixels with these same auxiliary data values,
  we're 90% confident that their **mean** timber volume is
  between 15,000 and 19,000 board feet.

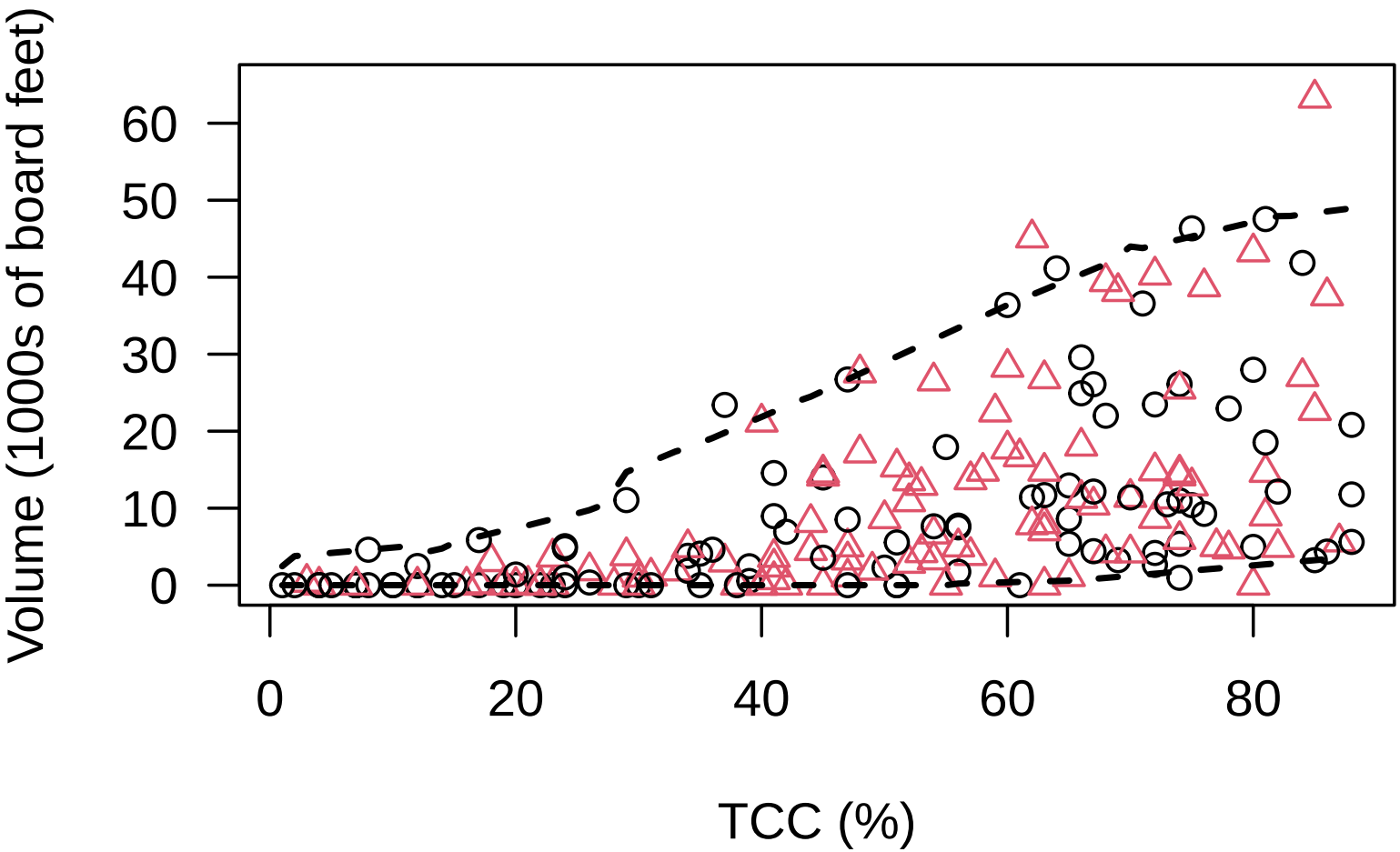
What are prediction intervals (PIs)?

- Point estimate:
  For a pixel with these auxiliary data values
  (TCC is 80%, elevation is 1200m, temperature is. . .),
  we predict it will have timber volume of 17,000 board feet.

- 90% **confidence interval**:
  Among all pixels with these same auxiliary data values,
  we're 90% confident that their **mean** timber volume is
  between 15,000 and 19,000 board feet.

- 90% **prediction interval**:
  For a pixel with these auxiliary data values,
  we're 90% confident that its **individual** timber volume is
  between 4,000 and 30,000 board feet.

Example of 90% PI band for Volume using TCC



How to find prediction intervals (PIs)?

For linear regression, there are known equations for PIs if certain conditions are met (errors are Normally distributed, etc.).

But not guaranteed to work when conditions aren't met;
and not developed for many other statistical models or machine learning algorithms.

Conformal prediction

Conformal prediction is a statistical approach to generating PIs, for nearly **arbitrary predictive models** and **with guaranteed finite-sample coverage**, while making **minimal assumptions** about the data distribution.

Conformal prediction: intuition

- Say we've fit a regression line to n observations. We might wish we could say:

  90% of the Y-values we've seen are no more than $\hat{q}_{.90}$ units away from the regression line.
  (90% of the absolute residuals are no bigger than $\hat{q}_{.90}$.)
  So we're 90% confident that the *next* observation we see will have a Y-value no more than $\hat{q}_{.90}$ away from the regression line.

Conformal prediction: intuition

- Say we've fit a regression line to n observations. We might wish we could say:

  90% of the Y-values we've seen are no more than $\hat{q}_{.90}$ units away from the regression line.
  (90% of the absolute residuals are no bigger than $\hat{q}_{.90}$.)
  So we're 90% confident that the *next* observation we see will have a Y-value no more than $\hat{q}_{.90}$ away from the regression line.

- If this were safe to say, then
  [regression line $\pm$ $\hat{q}_{.90}$] would be a 90% PI band.

  It turns out to be not quite that simple... but almost!

Conformal prediction: intuition

We have a simple random sample of n observations: $Y_1, \ldots, Y_n$.

There is a 90% chance that the next observation, Y_{n+1}, will be below the 90^{th} percentile of the **population**...

So there is **approximately** a 90% chance that Y_{n+1} will be below the 90^{th} percentile of **this sample**:

Let $\hat{q}_{n,0.90} = \lceil 0.9n \rceil$ smallest value in $\{Y_1, \ldots, Y_n\}$,
so then
$\text{Prob}(Y_{n+1} \leq \hat{q}_{n,0.90}) \approx 0.90$

But this is **only approximate**.
How good of an approximation?
Depends on a lot of things about the population!

Conformal prediction: intuition

Quantile lemma:

- Let's just imagine we *already have* the next observation!
  We have a simple random sample of $n+1$ observations:
  $Y_1, \ldots, Y_{n+1}$.
  Since they are in random order, there is a 90% chance that the last observation Y_{n+1} is below their 90^{th} percentile:
  Let $\hat{q}_{n+1,0.90} = \lceil 0.9(n+1) \rceil$ smallest value in $\{Y_1, \ldots, Y_{n+1}\}$,
  so then
  $\text{Prob}(Y_{n+1} \leq \hat{q}_{n+1,0.90}) = 0.90$ **exactly**—no approximation.

Conformal prediction: intuition

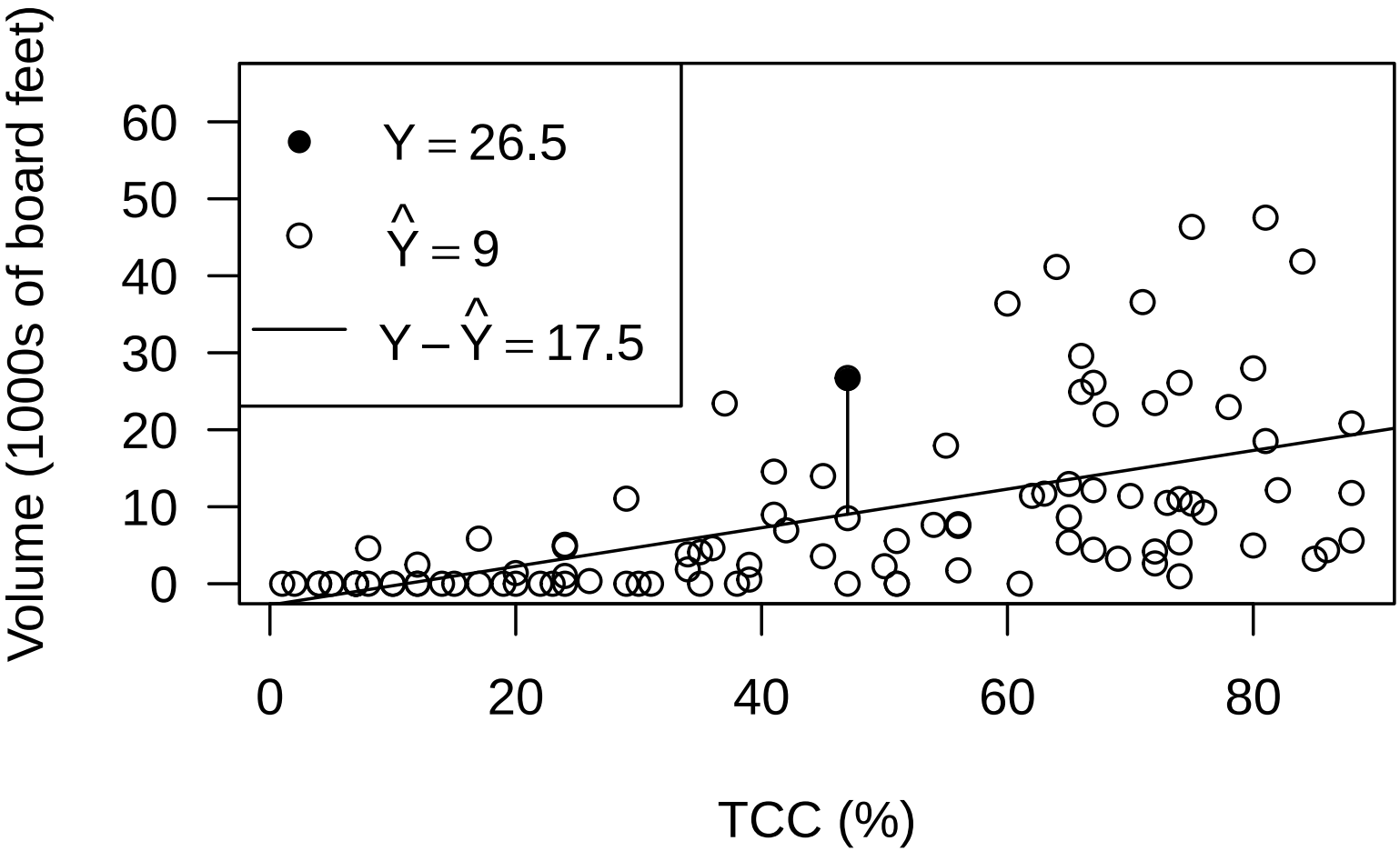
Quantile lemma:

- Let's just imagine we *already have* the next observation!
  We have a simple random sample of $n+1$ observations:
  $Y_1, \ldots, Y_{n+1}$.
  Since they are in random order, there is a 90% chance that the last observation Y_{n+1} is below their 90^{th} percentile:
  Let $\hat{q}_{n+1,0.90} = \lceil 0.9(n+1) \rceil$ smallest value in $\{Y_1, \ldots, Y_{n+1}\}$,
  so then
  $\text{Prob}(Y_{n+1} \leq \hat{q}_{n+1,0.90}) = 0.90$ **exactly**—no approximation.

- If we set aside Y_{n+1} and replace it with ∞, we can find the $\lceil 0.9(n+1) \rceil$ smallest value in $\{Y_1, \ldots, Y_n, \infty\}$.
  Call this value $\hat{q}_{conf,0.90}$.
  Since $\hat{q}_{conf,0.90} \geq \hat{q}_{n+1,0.90}$, we have
  $\text{Prob}(Y_{n+1} \leq \hat{q}_{conf,0.90}) \geq 0.90$ **guaranteed**.

Conformal prediction: intuition

Apply lemma to abs. residuals $|Y - \hat{Y}|$, not to the Ys themselves.
For example:



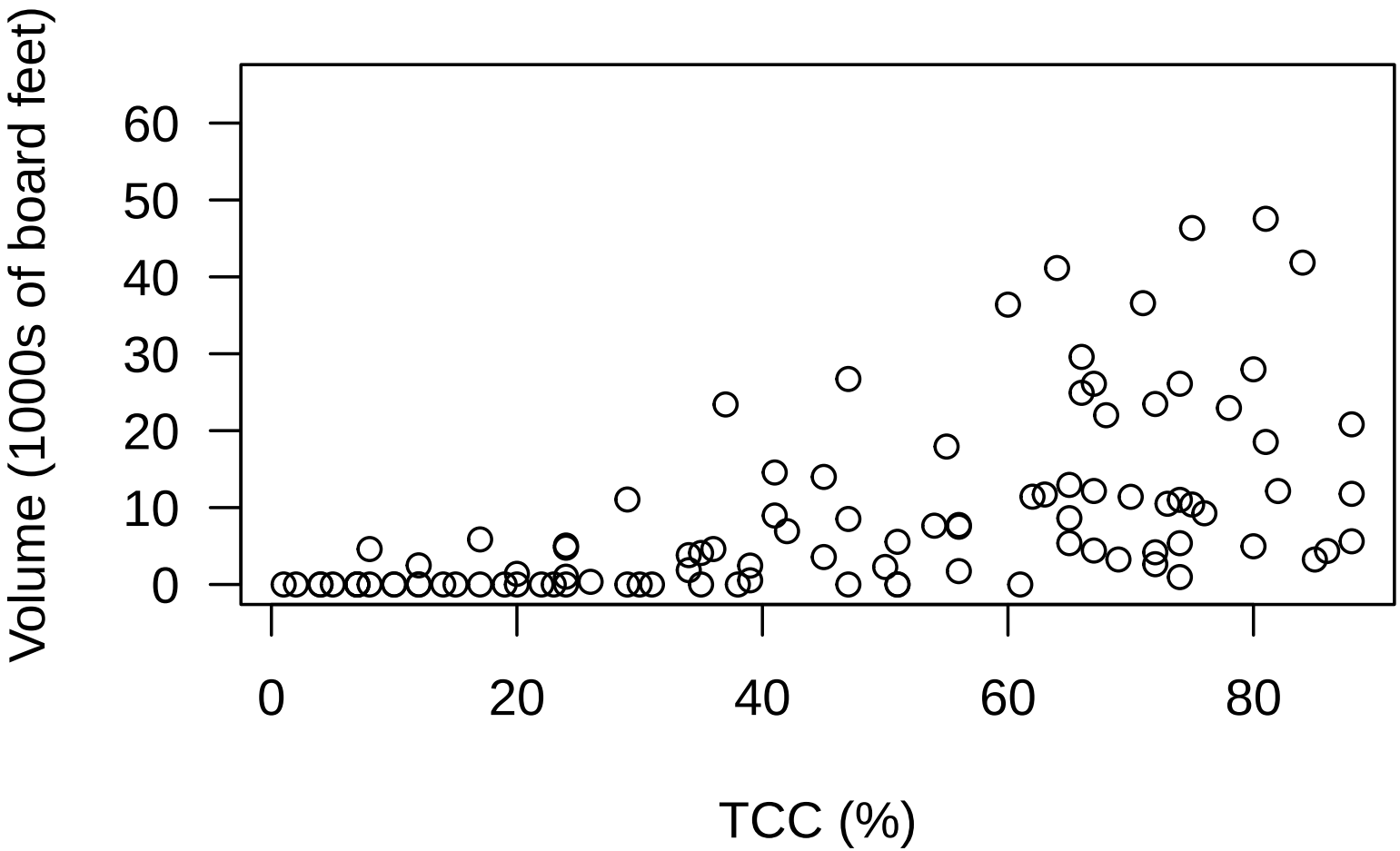
Conformal prediction: intuition

Apply lemma to abs. residuals $|Y - \hat{Y}|$, not to the Ys themselves.

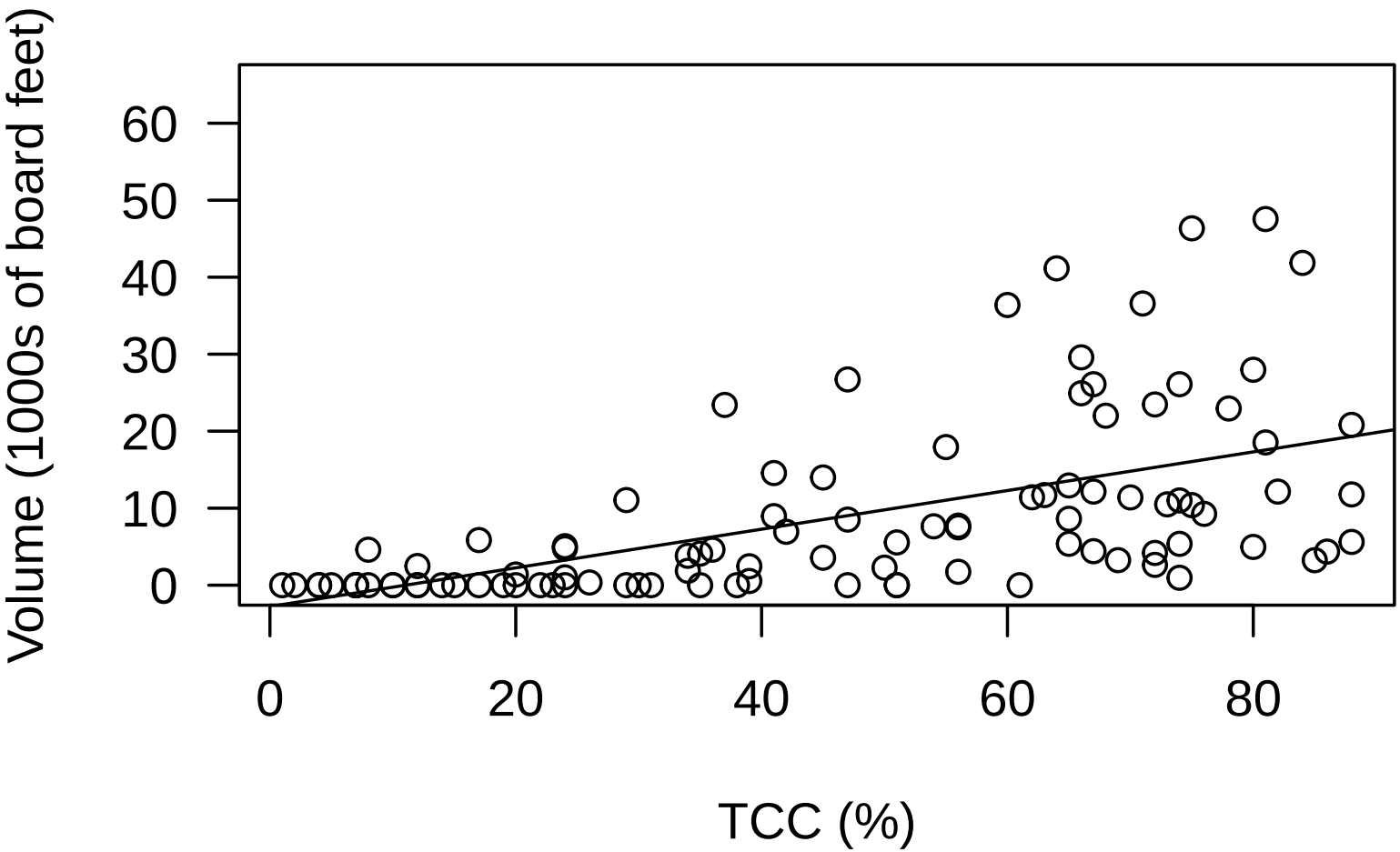
"Split conformal prediction":

- Fit a regression to part of the data ("training sample").
- Using the rest of the data
  ("calibration sample" of size n),
  find the **absolute residuals** $|Y - \hat{Y}|$:
  how far is each Y above/below the regression prediction $\hat{Y}$?
- Calculate $\hat{q}_{conf,.90}$ on the absolute calibration residuals.
- Then [regression line $\pm$ $\hat{q}_{conf,.90}$] is a conformal 90% PI band.
  It is "90% guaranteed" to cover the $(n+1)^{th}$ observation.

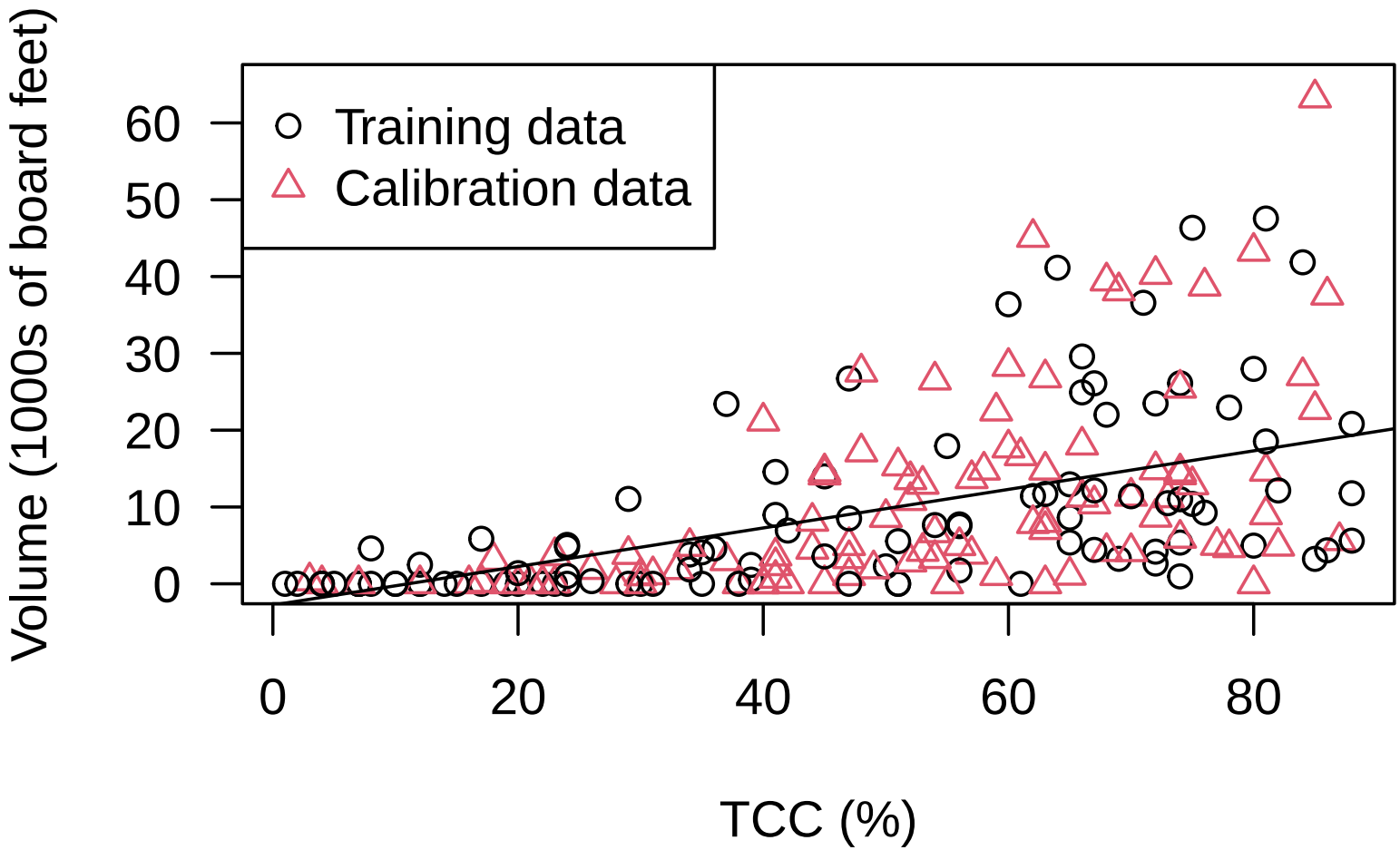
Conformal prediction for Volume using TCC



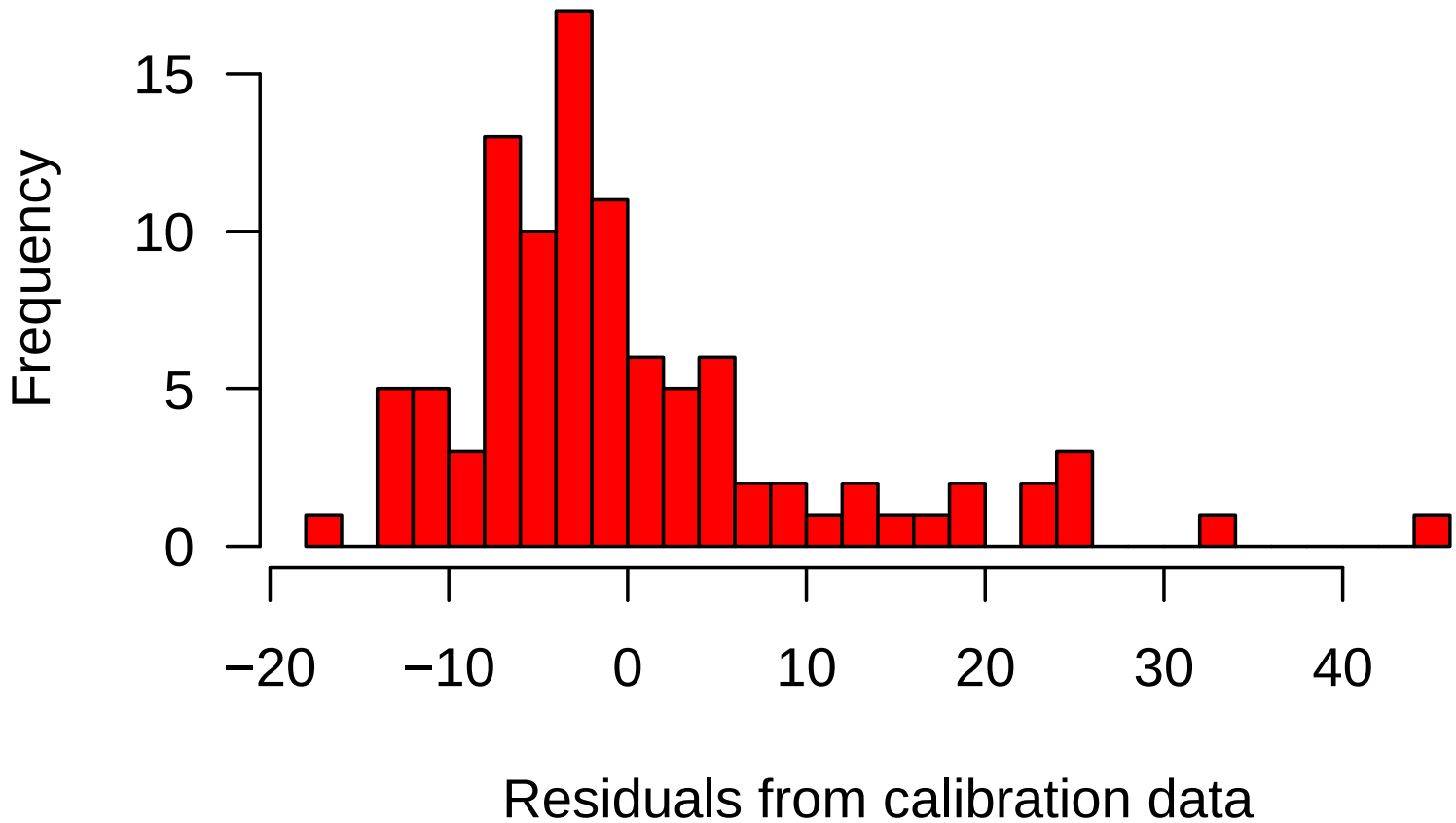
Conformal prediction for Volume using TCC



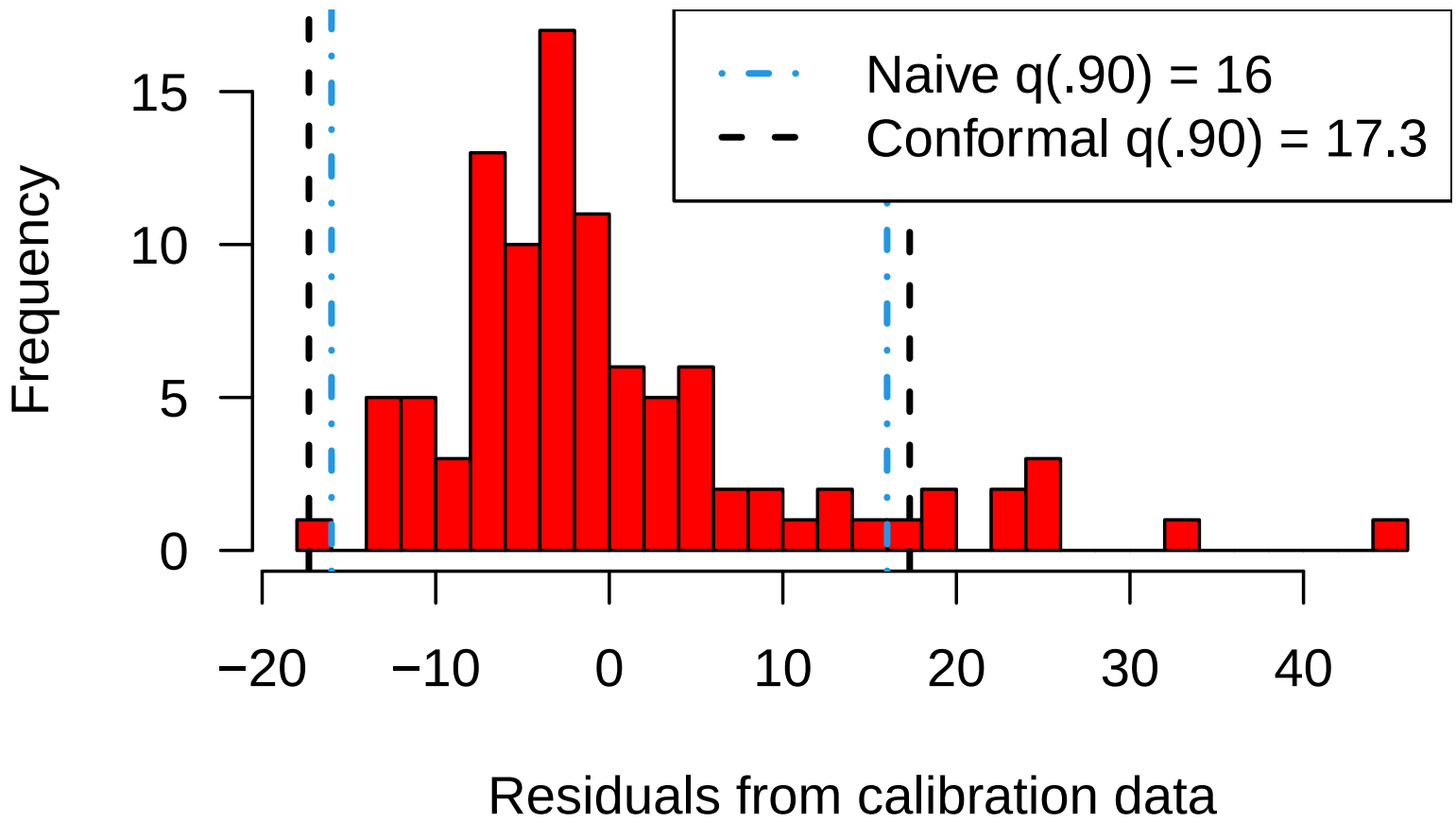
Conformal prediction for Volume using TCC



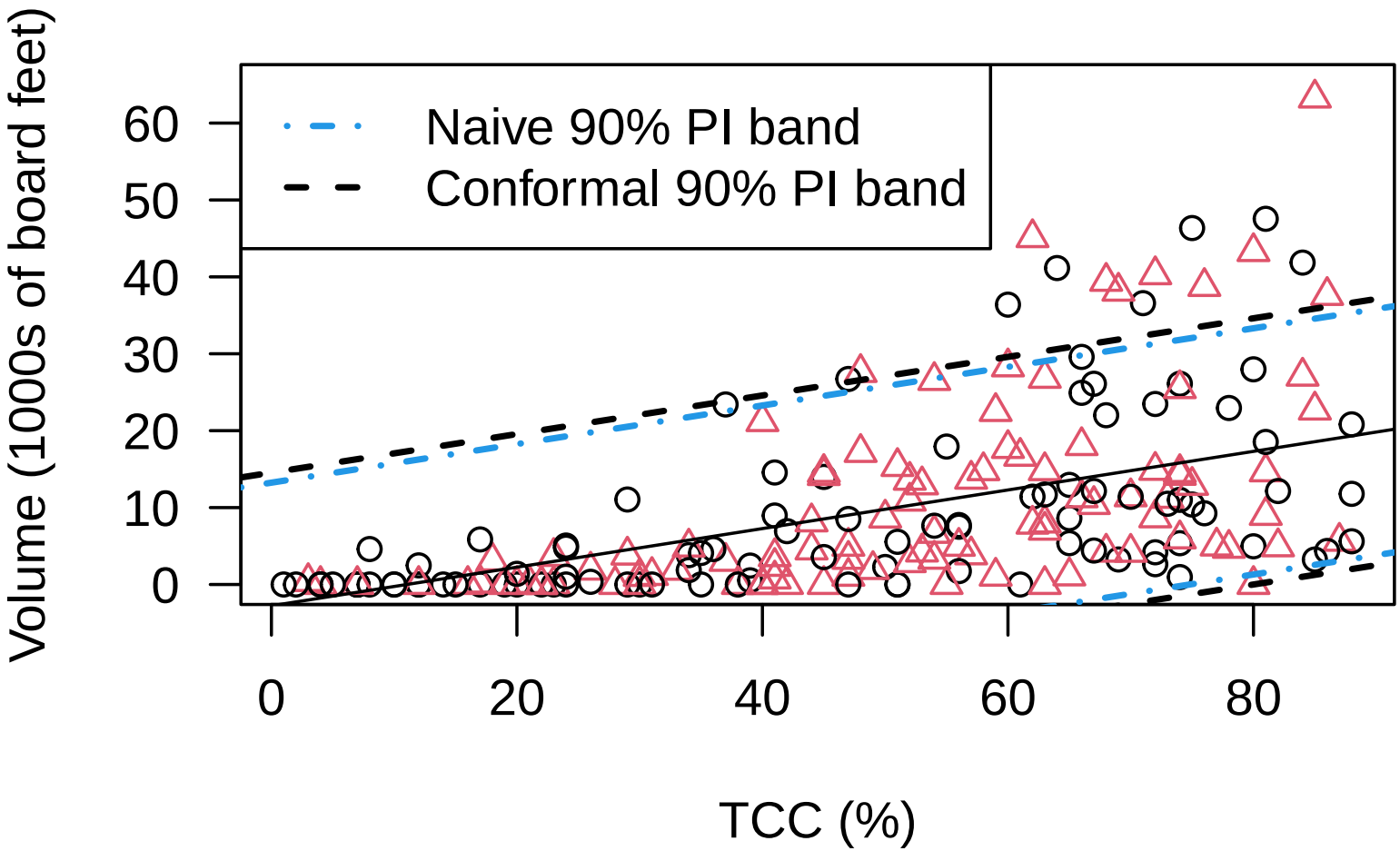
Conformal prediction for Volume using TCC



Conformal prediction for Volume using TCC



Conformal prediction for Volume using TCC



Conformal prediction: what does it guarantee?

"Marginal 90% coverage": Conditional on the training set...
Across all *simple random samples* of size $n+1$, where first n are the calibration set and last is the test observation...
We **guarantee** that the $(n+1)^{th}$ observation will be in the conformal PI built using the first n observations for 90% of such samples.

This may not be what we really want!
Often we want the PI to cover next Y at a particular X, not at a randomly chosen X.

But it's better than nothing. It works "out of the box" for pretty much any predictive algorithm, including ones where statisticians haven't worked out "native" PIs yet.
(...including recent breakthrough for non-SRS data: time series, survey samples, designed experiments, etc.!)

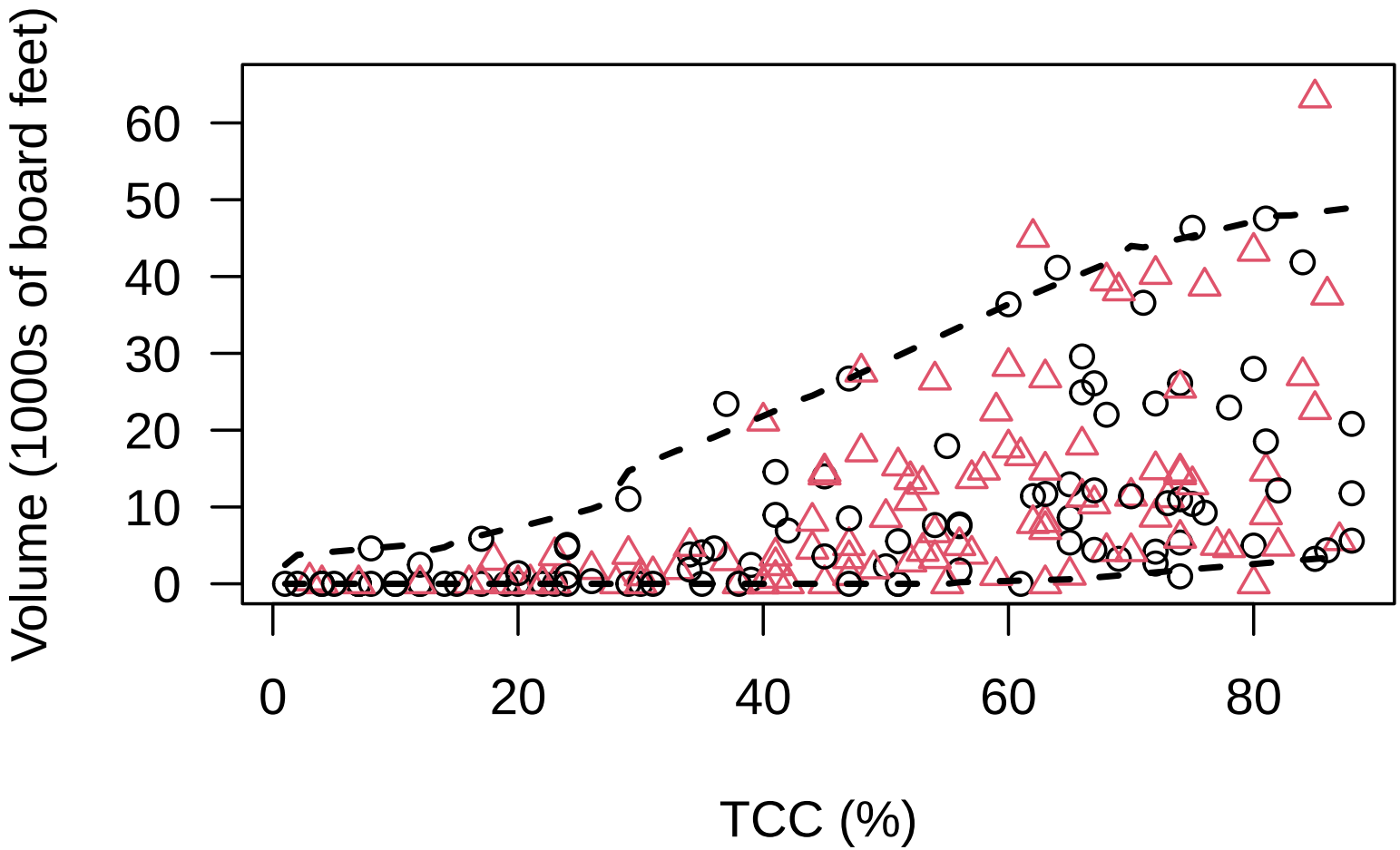
Adaptive-width conformal methods

Constant-width PI band provably has 90% marginal coverage, but scientifically speaking it's not very useful here!
Volumes are close to the line at low TCCs, and far from the line at high TCCs.

"Conformalized quantile regression", Romano et al. (2019):

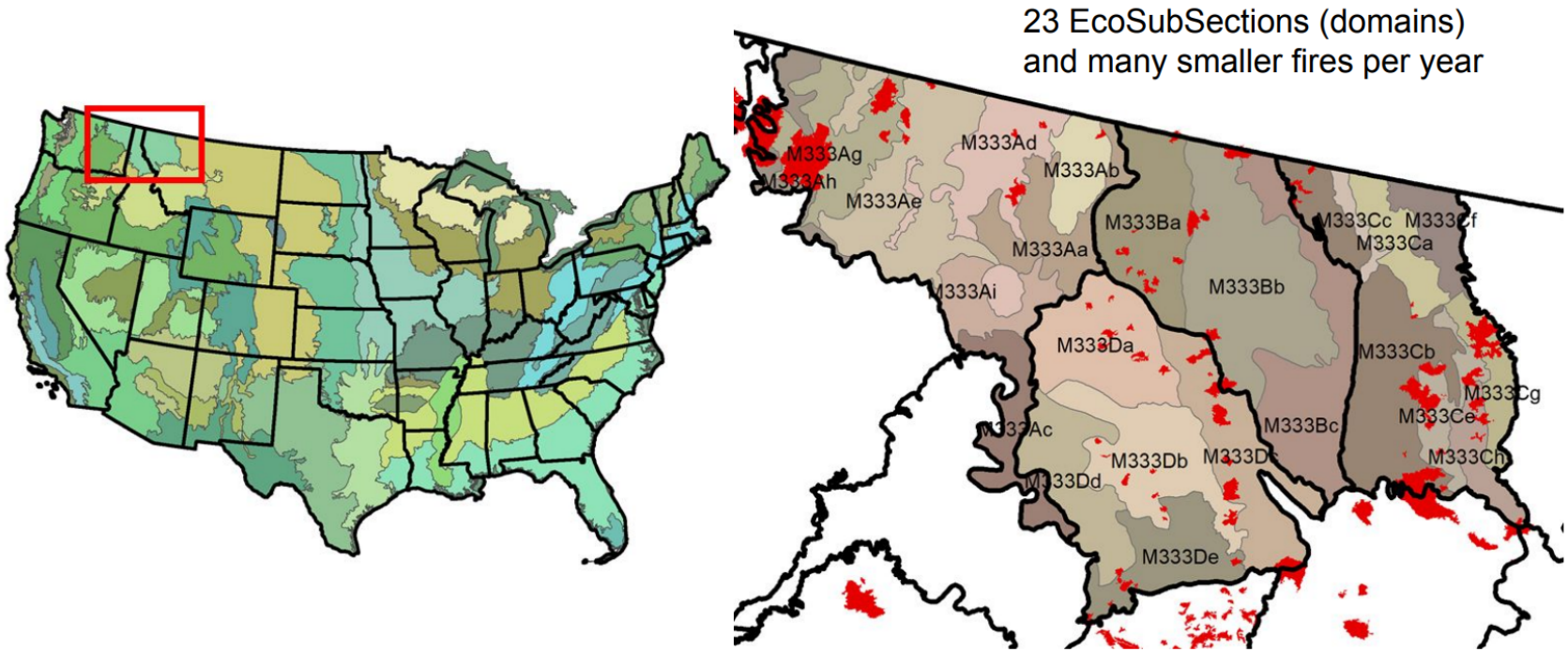
- First fit quantile regression lines that estimate the lower 5% and upper 95% of Y at each X.
- Then adjust these lines in ways that ensure conformal guarantees.
- PI band will be wider at those Xs where Y's spread is wider.

Conformalized Quantile Regression for Volume using TCC



We can find PIs for lost timber in each wildfire pixel

. . . using many more predictor variables, and using models other than linear regression.



Summary

- When building predictive models, make PIs available
- When your model does not have "built-in" PIs, try conformal methods

My current methodological research:

- If data came from a complex survey design, how to account for this when creating conformal PIs?
  Wieczorek (2023), "Design-based conformal prediction" (preprint in revision)
  https://arxiv.org/abs/2303.01422

Please reach out if you are interested in these methods or other statistical collaborations!

Contact: jerzy.wieczorek@colby.edu

Supplemental slides

Why does quantile lemma replace Y_{n+1} with ∞?

Say we have $n+1 = 10$. We hold out the 10th observation Y_{n+1}, and we want $\hat{q}$ such that there's 90% chance that $Y_{n+1} \leq \hat{q}$.

If we knew all 10 observations, their 90th percentile is the 9th smallest = 2nd largest of all 10.

But if we only know the first 9 observations, the 90th percentile of all 10 could be:

- 2nd largest of the first 9 values (if held-out value is small), or
- largest of the first 9 values (if held-out value is larger than any of the others), or
- the held-out value itself (if it's between these other two).

Out of these three options, biggest $\hat{q}$ is "largest of the first 9 values," chosen if we **assume** held-out value is larger than any other. Shorthand: replace held-out value with ∞.
Then $\hat{q}$ *may* be too large, but *certainly* not too small.